Universitat d'Alacant
Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior

Building machine translation systems for language pairs with scarce resources

Víctor Manuel Sánchez Cartagena

Memoria presentada para aspirar al grado de
DOCTOR/DOCTORA POR LA UNIVERSIDAD DE ALICANTE
MENCIÓN DE DOCTOR/DOCTORA INTERNACIONAL
DOCTORADO EN APLICACIONES DE LA INFORMÁTICA

Dirigida por
Dr. Felipe Sánchez Martínez
Dr. Juan Antonio Pérez Ortiz

Esta tesis ha sido financiada por la Universidad de Alicante a través del proyecto GRE11-20, el Ministerio de Economía y Competitividad a través de los proyectos TIN2009-14009-C02-01 y TIN2012-32615, la Generalitat Valenciana a través de la beca ACIF/2010/174 y la Unión Europea en virtud del acuerdo con número de concesión PIAP-GA-2012-324414 (Abu-MaTran) del Séptimo Programa Marco FP7/2007-2013

"Si hablas a una persona en una lengua que entiende, las palabras irán a su cabeza.
Si le hablas en su propia lengua, las palabras irán a su corazón"
Nelson Mandela

Agradecimientos

En primer lugar, quiero agradecer a mis directores de tesis, Felipe Sánchez Martínez y Juan Antonio Pérez Ortiz, el apoyo y supervisión que me han brindado durante estos años. Les estoy muy agradecido, en el plano profesional, por haber sabido guiar la investigación por el camino más adecuado, pero también en el personal, por la gran cantidad de tiempo que han dedicado a supervisar mi tesis y por haber sabido animarme y motivarme en los momentos más duros.

El trabajo se ha hecho más llevadero gracias al apoyo de mis compañeros, primero en el Departament de Llenguatges i Sistemes Informàtics y después en Prompsit Language Engineering. Desde los almuerzos en el Don Jamón a los cumpleaños en Prompsit, he pasado grandes momentos con ellos. También estoy muy agradecido a Sergio y Gema, el alma de Prompsit, por haber confiado en mi y permitirme seguir desarrollando mi carrera profesional.

Nunca habría podido finalizar esta tesis sin el apoyo de mi familia, especialmente de mi madre, que ha sido siempre un gran modelo para mi. Gracias también a mis amigos. Siempre he podido contar con ellos, a pesar de que algunos viven a miles de kilómetros de distancia o las circunstancias me han hecho pasar menos tiempo del que me gustaría con los que están más cerca.

Muchas gracias a Cristina. Gracias por la infinita paciencia que ha tenido y su apoyo durante estos años. Gracias por hacerme ver siempre la parte positiva de las cosas. Gracias por estar a mi lado y ayudarme a *ponerle la pajarita* a mi tesis.

Finalmente, me gustaría dar las gracias a todas las instituciones que han financiado mi investigación: la Universidad de Alicante a través del proyecto GRE11-20, el Ministerio de Economía y Competitividad a través de los proyectos TIN2009-14009-C02-01 and TIN2012-32615, la Generalitat Valenciana a través de la beca ACIF/2010/174 y la Unión Europea en virtud del acuerdo con número de concesión PIAP-GA-2012-324414 (Abu-MaTran) del Séptimo Programa Marco FP7/2007-2013.

Víctor Manuel Sánchez Cartagena
Alicante, 20 de abril de 2015

Resumen en español

Introducción

La traducción automática (TA) puede definirse como el proceso llevado a cabo por un sistema informático para traducir un texto escrito en un lenguaje natural, la lengua origen (LO), a otro lenguaje natural, la lengua meta (LM). La TA constituye un reto científico debido a factores como la ambigüedad de los lenguajes naturales, la necesidad de conocimientos sobre el funcionamiento del mundo real para resolver dicha ambigüedad o las divergencias gramaticales existentes entre las distintas lenguas.

Aunque, en general, la traducción de alta calidad está fuera del alcance de los sistemas de TA actuales —con la excepción de lenguas emparentados o dominios específicos, para los cuales sí que es posible realizar TA de alta calidad—, las dos modalidades de uso de la TA descritas a continuación están ampliamente extendidas. En primer lugar, la *diseminación* consiste en usar TA para producir borradores de las traducciones que son posteriormente corregidos manualmente por parte de traductores profesionales. Este proceso, conocido como *posedición*, permite acelerar el proceso de traducción, ya que los traductores profesionales (*poseditores* en este caso) no deben comenzar a traducir desde cero. En segundo lugar, cuando un sistema de TA se usa para la *diseminación*, el objetivo es producir una traducción que permita a un usuario sin conocimientos de la LM hacerse una idea del contenido del texto original en LO. Para cumplir con este objetivo, no es necesario que la traducción resultante sea gramaticalmente correcta ni que el sistema sea capaz de traducir todas las palabras del texto en LO.

Tipos de sistemas de traducción automática

Los sistemas de TA pueden clasificarse según el tipo de conocimiento usado para su construcción. Así, pueden distinguirse principalmente dos tipos de sistemas: basados en reglas y basados en corpus. También existen enfoques híbridos que combinan elementos de ambos tipos de sistemas. Los sistemas de TA basados en reglas utilizan recursos lingüísticos, como diccionarios morfológicos o reglas de transferencia, para llevar a cabo el proceso de traducción. Dichos recursos son normalmente creados a mano por

expertos. Los sistemas basados en corpus, por su parte, emplean grandes colecciones de textos ya traducidos (conocidos como *corpus paralelos*) como fuente de conocimiento.

Los sistemas de TA basados en reglas generalmente realizan el proceso de traducción en tres pasos: primero, analizan el texto en LO para obtener una representación intermedia en LO, que elimina toda la información no relevante para el proceso de traducción y hace explícita aquella que sí que lo es. A continuación, la representación intermedia en LO se transfiere a una representación intermedia en LM. La traducción final se genera a partir de la representación intermedia en LM. De entre los distintos tipos de sistemas de TA basados en reglas, esta tesis doctoral se centra en los sistemas de TA de transferencia morfológica avanzada o sintáctica superficial. Estos sistemas no realizan un análisis sintáctico completo de las oraciones a traducir en LO; la representación intermedia que emplean consta de una secuencia de *formas léxicas* (cada forma léxica está formada por el lema, la categoría léxica y la información de flexión una palabra) y se obtiene tras un análisis morfológico. En particular, los nuevos métodos presentados en esta tesis han sido evaluados en la plataforma de TA por transferencia morfológica avanzada Apertium (Forcada et al., 2011).

Respecto a la TA basada en corpus, actualmente el enfoque más popular dentro de este grupo es la TA estadística. Para cada oración s en LO que deber ser traducida, los sistemas estadísticos buscan la oración en LM $\hat{t}$ que maximiza la probabilidad $p(t|s)$. Dicha probabilidad se obtiene normalmente a partir de la combinación de modelos estadísticos estimados a partir de corpus paralelos (*modelo de traducción*) y de corpus monolingües en LM (*modelo de LM* o simplemente *modelo de lengua*). Mientras que el modelo de traducción indica cómo es de probable que s y t sean traducción mutua, el modelo de lengua indica la probabilidad de que t sea una oración correcta en LM. Dentro de la TA estadística, esta tesis se centra en los sistemas estadísticos basados en segmentos, que son los más populares. El modelo de traducción de un sistema basado en segmentos, que recibe el nombre de *tabla de segmentos*, está formado por parejas de segmentos (cada pareja consta de un segmento en LO y su traducción en LM) y sus correspondientes probabilidades. Las distintas hipótesis de traducción t que son evaluadas para hallar $\hat{t}$ se crean dividiendo la oración s en segmentos y combinando sus traducciones según la tabla de segmentos.

Dado que los dos principales tipos de sistemas de TA que acaban de ser descritos abordan el proceso de traducción de manera completamente diferente, los tipos de errores cometidos por estos sistemas son también muy diferentes. Por una parte, los sistemas estadísticos, gracias al modelo de lengua, producen traducciones más naturales en LM. Gracias a que la información empleada en el proceso de traducción ha sido extraída de millones de oraciones ya traducidas, los sistemas estadísticos tratan mejor con las expresiones que no deben ser traducidas literalmente, así como con palabras con múltiples traducciones. Por otra parte, los sistemas de TA basados en reglas producen traducciones más mecánicas y repetitivas que, sin embargo, son más fieles al texto en LO y facilitan su posedición. Además, los sistema basados en reglas no sufren el problema de la dispersión de datos que normalmente se da en sistemas estadísticos. Como éstos

últimos operan sobre formas superficiales (palabras tal y como se encuentran los textos, sin ningún tipo de análisis), todas las formas flexionadas de cada palabra deben estar presentes en los corpus usados en el entrenamiento (y preferiblemente con diferentes contextos) para que el sistema sea capaz de traducirlas correctamente. Esta condición puede ser muy difícil de cumplir para idiomas con un alto grado de flexión, como alemán o español.

Recursos necesarios para construir sistemas de traducción automática

A la hora de crear un nuevo sistema de TA para un par de lenguas, la decisión de emplear TA basada en reglas o TA basada en corpus está condicionada, además de por los errores cometidos por cada tipo de sistema, por los recursos disponibles para el par de lenguas en cuestión.

Los sistemas de TA estadísticos pueden ser creados de manera automática (sin necesidad de conocimientos lingüísticos aportados por humanos) mientras haya corpus disponibles. Esta característica, junto a la cada vez mayor disponibilidad de corpus paralelos y de capacidad de cómputo, ha contribuido a su popularización. Generalmente, los corpus paralelos necesarios para construir el modelo de traducción son más difíciles de encontrar que los corpus monolingües en LM empleados para construir el modelo de lengua. Dado que son necesarias millones de palabras en cada lengua para obtener un sistema de TA estadístico competitivo, la TA estadística no es adecuada para lenguas con pocos hablantes (suele ser difícil encontrar corpus paralelos) y para la traducción de textos de dominios restringidos para los que no existen corpus paralelos suficientemente grandes.

La construcción de sistemas basados en reglas, por su parte, implica un esfuerzo considerable en la creación de los recursos lingüísticos. Normalmente, este esfuerzo sólo puede ser llevado a cabo por personas con grandes conocimientos sobre la gramática y morfología de los lenguas implicadas, así como sobre el formato en el que los datos lingüísticos son codificados en el sistema concreto de TA basado en reglas. Sin embargo, la TA basada en reglas es la alternativa más razonable para construir sistemas de TA cuando no hay corpus paralelos disponibles, como ocurre para muchos de los pares de lenguas del proyecto Apertium: bretón–francés, islandés–inglés, kazajo–tártaro, etc.

Respecto a los recursos lingüísticos concretos que son necesarios para construir un sistema de TA basado en reglas de transferencia morfológica avanzada (en particular, Apertium), cada paso del proceso (análisis, transferencia, generación), emplea un tipo de recurso distinto. En el paso de análisis, Apertium emplea un analizador morfológico para obtener una secuencia de formas léxicas en LO a partir de cada oración en LO a traducir. Cuando más de una forma léxica puede corresponder a una forma superficial, un desambiguador léxico categorial se encarga de elegir la forma léxica más adecuada (Sánchez-Martínez et al., 2008). Las equivalencias entre formas superficiales

y formas léxicas en LO están codificadas en un diccionario morfológico monolingüe en LO. El proceso de transferencia se divide en transferencia léxica y transferencia estructural. En la transferencia léxica, la traducción a la LM de cada forma léxica en LO se obtiene de un diccionario bilingüe, mientras que en la transferencia estructural, una serie de reglas realizan las operaciones necesarias para obtener una traducción correcta en LM (concordancias, reordenamientos, etc.) cuando la traducción palabra por palabra no es suficiente. Cada regla procesa una secuencia de formas léxicas de la oración; las reglas se aplican de manera voraz, de izquierda a derecha, y cuando más de una regla puede ser aplicada en un punto de la oración, se elige la más larga. Finalmente, en el paso de generación, el generador morfológico produce una secuencia de formas superficiales en LM a partir de la representación intermedia en LM (secuencia de formas léxicas en LM). Las equivalencias entre formas superficiales y formas léxicas en LM están contenidas en el diccionario morfológico monolingüe en LM. Las reglas de transferencia son el recurso que requiere un conocimiento lingüístico más profundo del par de lenguas.

En resumen, a la hora de crear un nuevo sistema de TA, es posible que no exista un corpus paralelo lo suficientemente grande para el par de lenguas o dominio deseado. Además, la creación de los recursos lingüísticos necesarios es muy costosa y puede que los expertos que deben llevar a cabo esa tarea no estén disponibles si alguna de las lenguas implicadas cuenta con un número reducido de hablantes. En esta tesis doctoral, se presentan tres nuevos métodos que facilitan la creación de sistemas de TA cuando los recursos necesarios (corpus paralelos y recursos lingüísticos) son escasos. En particular, en el capítulo 2 se describe un nuevo método para inferir automáticamente un conjunto de reglas de transferencia morfológica avanzada a partir de un corpus paralelo muy pequeño y de los diccionarios de un sistema de TA basado en reglas. En el capítulo 3, se describe una nueva estrategia de hibridación que permite combinar los recursos lingüísticos de un sistema de TA basado en reglas por transferencia morfológica avanzada con corpus paralelos para construir sistemas de TA estadísticos más potentes, lo que conduce a un mejor aprovechamiento de los recursos disponibles. El capítulo 4, por último, se presenta un nuevo método que permite a hablantes de una lengua sin grandes conocimientos lingüísticos insertar entradas en los diccionarios morfológicos monolingües a usar para la TA.

Inferencia automática de reglas de transferencia morfológica avanzada

El nuevo método de inferencia de reglas de transferencia morfológica avanzada a partir de corpus paralelos muy pequeños y diccionarios morfológicos descrito en el capítulo 2 facilita enormemente la creación de sistemas de TA basados en reglas, pues las reglas de transferencia son el recurso que requiere un conocimiento lingüístico más profundo y por tanto, el que requiere de personal más especializado. Las reglas inferidas por este

método son compatibles con el formato usado por Apertium (Forcada et al., 2011) y pueden ser editadas manualmente.

Este nuevo método está inspirado en el trabajo de Sánchez-Martínez y Forcada (2009). Ambos métodos usan una extensión del formalismo de *plantillas de alineamiento* (Och and Ney, 2004) para codificar las reglas, aunque existen múltiples diferencias entre ambos enfoques.

En primer lugar, el formalismo definido por Sánchez-Martínez y Forcada (2009) es menos expresivo que el nuevo formalismo presentado en esta tesis. El formalismo empleado por Sánchez-Martínez y Forcada (2009) no es capaz de representar reglas que son aplicadas independientemente de los atributos morfológicos de las palabras que casan con ellas. En consecuencia, normalmente es necesario inferir múltiples reglas para poder tratar con el mismo fenómeno lingüístico. Por ejemplo, los adjetivos en inglés se sitúan antes del nombre al que acompañan, mientras que en español normalmente se sitúan después. Para traducir correctamente la secuencia nombre–adjetivo de español a inglés, el método de Sánchez-Martínez y Forcada (2009) necesita inferir 4 reglas distintas: una para cada combinación de género y número, mientras que el nuevo método descrito en el capítulo 2 necesita una sola regla. En segundo lugar, el nuevo método presentado en esta tesis emplea un algoritmo de aprendizaje mucho más potente: la inferencia de reglas se aborda como un problema de minimización. Mediante este novedoso enfoque, se resuelven los posibles conflictos entre reglas y se evita la sobregeneralización que podría ser causada por el uso de un formalismo más potente. Éste es el primer método de aprendizaje de reglas que plantea la inferencia como un problema de minimización similar al problema del conjunto de cobertura en lugar de usar un enfoque voraz. En tercer y último lugar, el método de Sánchez-Martínez y Forcada (2009) genera reglas que, al ser aplicadas por el sistema de TA, impiden que otras reglas más útiles e importantes se apliquen. Este problema surge porque su método genera reglas a partir de todos los segmentos bilingües extraídos del corpus paralelo, incluso a partir de aquellos con una secuencia de categorías léxicas en LO que no necesitan ser procesadas conjuntamente (por ejemplo, sustantivo seguido de conjunción y determinante). El nuevo método descrito en esta tesis no sufre este problema porque tiene en cuenta la política seguida por el sistema de TA para elegir las reglas a aplicar (de izquierda a derecha, y dando prioridad a las reglas más largas) y no genera reglas que puedan penalizar la calidad de la traducción resultante.

El proceso de aprendizaje de reglas consta de los pasos que se describen a continuación. Primero, se analiza el corpus paralelo para obtener una secuencia de formas léxicas en LO y una secuencia de formas léxicas en LM, y se calculan los pares de formas léxicas que están alineados, tal y como se hace durante el entrenamiento de un sistema de TA estadístico basado en segmentos. Después, se extraen segmentos bilingües consistentes con los alineamientos, también de manera similar a como se hace en sistemas estadísticos. Además, se descartan aquellos segmentos bilingües que no son compatibles con el diccionario bilingüe del sistema de TA basado en reglas donde las reglas inferidas serán integradas. A continuación, se generan múltiples reglas a partir

de cada segmento bilingüe. Todas las reglas generadas cumplen la siguiente condición respecto al segmento bilingüe a partir del cual han sido generadas: al ser aplicadas al segmento en LO, generan el segmento en LM. Cada regla generada tiene un grado de generalización distinto: algunas reglas casan únicamente con las formas léxicas del segmento en LO a partir del cual han sido generadas, otras se aplican independientemente del lema, otras independientemente del lema y la información de flexión, etc. De esta gran cantidad de reglas, se escoge el subconjunto de cardinalidad mínima que permita *reproducir* el conjunto de segmentos bilingües extraídos del corpus paralelo, es decir, que para cada segmento bilingüe, la aplicación de la regla más específica disponible a su segmento en LO dé como resultado su segmento en LM. Este problema de minimización se reescribe como un sistema de inecuaciones lineales y se resuelve mediante métodos de programación lineal entera (Garfinkel and Nemhauser, 1972).

Para eliminar las reglas que impiden que otras reglas más importantes se apliquen (tercera limitación del método de Sánchez-Martínez y Forcada (2009) descrita anteriormente), se detectan, para cada oración en LO del corpus paralelo, aquellos subsegmentos que deben traducirse con reglas para obtener la máxima similitud con la oración correspondiente en LM: son los segmentos clave. Una vez obtenida esta información, se eliminan las reglas que normalmente impiden que segmentos clave casen con la regla necesaria. Finalmente, también se eliminan reglas redundantes: aquellas cuya labor puede ser llevada a cabo por una combinación de reglas más cortas.

Para evaluar este nuevo algoritmo de aprendizaje de reglas, se han inferido reglas a partir de corpus paralelos pequeños de distintos tamaños (hasta 5 000 oraciones como tamaño máximo) para 5 pares de lenguas distintos: español↔catalán, inglés↔español, y bretón–francés. Tanto los diccionarios morfológicos como el sistema de TA en el que se han integrado las reglas pertenecen al proyecto Apertium (Forcada et al., 2011). Los resultados demuestran que la calidad de la traducción obtenida por este nuevo algoritmo supera al método de Sánchez-Martínez y Forcada (2009) en todos los escenarios evaluados. Además, gracias al formalismo más expresivo, el número de reglas obtenidas es mucho menor, lo que facilita su posterior modificación y mejora. Adicionalmente, cuando las lenguas están emparentadas (en el caso del par español↔catalán), unos pocos cientos de oraciones paralelas son suficientes para obtener reglas de calidad aceptable.

La generación de reglas con diferentes niveles de generalización y la complejidad computacional del problema de minimización que debe ser resuelto limitan el tamaño de los corpus que pueden emplearse. Sin embargo, en los experimentos se ha comprobado que buena parte de dicha complejidad viene dada por la extracción de reglas que se ejecutan independientemente de los atributos morfológicos de las palabras que casan. La extracción de ese tipo de reglas únicamente aumenta la calidad de la traducción obtenida cuando el corpus paralelo es muy pequeño (menor de 1 000 oraciones). Como consecuencia de este descubrimiento, se han repetido parte de los experimentos desactivando la generación de este tipo de reglas y con corpus paralelos más grandes (hasta 25 000 oraciones). Las reglas resultantes han sido capaces de superar la calidad en la

traducción proporcionada por las reglas escritas a mano del proyecto Apertium para los pares de idiomas español–inglés y bretón–francés.

En la siguiente publicación se puede encontrar una descripción más detallada del nuevo algoritmo de inferencia de reglas y de los resultados de su evaluación:

- Sánchez-Cartagena, V. M., Pérez-Ortiz, J. A. y Sánchez-Martínez, F. (2015). A generalised alignment template formalism and its application to the inference of shallow-transfer machine translation rules from scarce bilingual corpora. *Computer Speech & Language*, 32(1):46–90. Hybrid Machine Translation: integration of linguistics and statistics (`http://www.dlsi.ua.es/~vmsanchez/pub/sanchez15.pdf`)

Integración de reglas de transferencia morfológica avanzada en sistemas de traducción automática estadísticos

Incluso si un corpus paralelo mayor que los usados para la inferencia de reglas está disponible, el sistema de TA estadístico construido a partir del mismo puede presentar importantes limitaciones. En primer lugar, el problema de la dispersión de datos comentado anteriormente hace que obtener suficientes segmentos bilingües para poder traducir correctamente todas las formas flexionadas cuando alguna de las lenguas (LO o LM) presenta un elevado grado de flexión requiera una cantidad de corpus considerable, que podría no existir para determinados pares de lenguas. Y en segundo lugar, los corpus paralelos disponibles podrían no pertenecer al mismo dominio que los textos que necesitan ser traducidos con el sistema de TA estadístico resultante. En este caso, la calidad de la traducción obtenida sería más baja de lo deseable.

Una posible solución para estos problemas es la *hibridación*: si también existe un sistema de TA basado en reglas para el mismo par de lenguas, éste puede ser combinado con el sistema de TA estadístico para mitigar sus limitaciones. Así, en el capítulo 3 se describe una nueva estrategia de hibridación consistente en la inserción de información lingüística procedente de un sistema de TA basado en reglas por transferencia morfológica avanzada en la tabla de segmentos de un sistema de TA estadístico basado en segmentos.

Incluso si las reglas de transferencia estructural todavía no han sido creadas, éstas pueden ser inferidas automáticamente a partir del mismo corpus paralelo empleado para entrenar el sistema estadístico con la ayuda del método descrito en el capítulo 2. De este modo, se produce un mejor aprovechamiento de los recursos disponibles (diccionarios y corpus paralelo) que si se emplearan los métodos existentes en la literatura (Schwenk et al., 2009), que consisten simplemente en añadir los diccionarios

a la tabla de segmentos. Combinando la estrategia de inferencia de reglas con la de hibridación, el sistema resultante es capaz de generalizar el conocimiento presente en el corpus paralelo a secuencias de palabras que no aparecen en el corpus pero comparten categoría léxica o atributos morfológicos con aquellas que sí que aparecen.

La estrategia presentada en el capítulo 3 es la primera técnica de hibridación diseñada específicamente para integrar información lingüística proveniente de un sistema de TA basado en reglas por transferencia morfológica avanzada en un sistema de TA estadístico basado en segmentos. A parte de esta nueva estrategia, el único método existente para tal propósito es la estrategia diseñada por Eisele et al. (2008). Sin embargo, la esta última es una estrategia general que se puede aplicar para enriquecer un sistema de TA estadístico a partir de cualquier otro sistema de TA sin utilizar ningún tipo de información sobre su funcionamiento interno (el sistema de TA empleado para enriquecer el sistema estadístico se trata como una *caja negra*). En la estrategia de Eisele et al. (2008), las oraciones que deben ser traducidas con el sistema híbrido se traducen primero con el sistema de TA basado en reglas, se construye una nueva tabla de segmentos a partir del corpus paralelo resultante (*corpus sintético*) y se concatena a la tabla de segmentos obtenida inicialmente a partir del corpus paralelo de entrenamiento. La estrategia seguida por Eisele et al. (2008) presenta ciertas limitaciones que son resueltas por el nuevo método de hibridación que se presenta en esta tesis gracias a que el nuevo método saca partido del funcionamiento interno del sistema de TA basado en reglas. En primer lugar, cuando se emplea el método de Eisele et al. (2008), se insertan en la tabla de segmentos pares de segmentos que no son traducción mutua, como consecuencia de la mala calidad de los alineamientos entre palabras en el corpus sintético. Y en segundo lugar, el método de Eisele et al. (2008) es incapaz de encontrar un equilibrio adecuado entre las probabilidades de los pares de segmentos extraídos del corpus de entrenamiento y del corpus sintético.

La nueva estrategia de hibridación presentada en el capítulo 3 funciona en dos pasos. Primero, se generan una serie de de segmentos bilingües sintéticos a partir de los datos lingüísticos del sistema de TA basado en reglas por transferencia morfológica avanzada (Apertium). Después, se crea una tabla de segmentos combinando los segmentos bilingües extraídos a partir del corpus de entrenamiento y los sintéticos. Los segmentos bilingües sintéticos se generan a partir del diccionario bilingüe y las reglas de transferencia del sistema de TA basado en reglas. Para generar segmentos sintéticos a partir del diccionario bilingüe, se identifican todas las formas superficiales en LO que el sistema es capaz de analizar (con la ayuda del diccionario monolingüe en LO), y se traducen sus formas léxicas correspondientes con el diccionario bilingüe. Para generar segmentos sintéticos a partir de las reglas de transferencia, se identifican en el texto a ser traducido con el sistema híbrido todos los segmentos que casan con reglas de transferencia, y se traducen con las reglas correspondientes. Una vez generados los pares de segmentos sintéticos, éstos y los segmentos extraídos del corpus paralelo se juntan en una sola lista, a partir de la cual se crea la tabla de segmentos. A las puntuaciones empleadas normalmente en TA estadística se les añade una puntuación binaria que

especifica si cada segmento paralelo viene del corpus de entrenamiento o del sistema de TA basado en reglas.

Para evaluar la nueva estrategia de hibridación, se han construido sistemas híbridos a partir de corpus paralelos y monolingües de distintos tamaños y de los datos lingüísticos del sistema Apertium y se han comparado con sistemas de TA puros, tanto basados en reglas como estadísticos, obtenidos a partir de los mismos recursos. La evaluación (automática) se ha llevado a cabo para los pares de lenguas inglés↔español y bretón–francés. Mientras que para bretón–francés el corpus de evaluación tiene el mismo origen que el corpus de entrenamiento, para el caso de inglés↔español se han empleado dos corpus de evaluación distintos: un corpus del mismo tipo que el corpus de entrenamiento (actas del parlamento europeo), y un corpus procedente de un dominio diferente (noticias).

Los resultados muestran que los sistemas híbridos creados con la nueva estrategia son capaces de superar a los sistemas de TA puros construidos a partir de los mismos recursos. Los sistemas híbridos son especialmente útiles cuando se traducen textos de un dominio distinto al del corpus de entrenamiento empleado o cuando dicho corpus de entrenamiento es pequeño. Además, los experimentos confirman que la nueva estrategia supera sistemáticamente al método de Eisele et al. (2008). Un sistema construido siguiendo la estrategia de hibridación presentada en esta tesis fue uno de los sistemas ganadores en la evaluación humana de la tarea de traducción del *Workshop on Statistical Machine Translation* de 2011 (Callison-Burch et al., 2011) para el par de lenguas español–inglés. De este modo, la efectividad del método ha quedado confirmada tanto por una evaluación automática como por una evaluación humana.

Respecto al uso de reglas inferidas automáticamente con el algoritmo descrito en el capítulo 2, los resultados muestran que, cuando los sistemas híbridos emplean dichas reglas, la calidad de la traducción obtenida puede alcanzar la de sistemas híbridos que emplean reglas del proyecto Apertium creadas a mano. En todo caso, siempre se produce una mejora respecto a un sistema de TA estadístico enriquecido únicamente con diccionarios. Finalmente, y como cabe esperar, los experimentos demuestran que, cuanto mayor es el corpus monolingüe en LM empleado para estimar el modelo de lengua, menor es la mejora aportada por el sistema híbrido respecto a un sistema de TA estadístico puro.

La nueva estrategia para integrar reglas de transferencia morfológica avanzada en sistemas de traducción automática estadísticos se describe en profundidad en la siguiente publicación:

- Sánchez-Cartagena, V. M., Sánchez-Martínez, F. y Pérez-Ortiz, J. A. (2011b). Integrating shallow-transfer rules into phrase-based statistical machine translation. En *Proceedings of the Machine Translation Summit XIII*, pp. 562–569, Xiamen, China (`http://www.dlsi.ua.es/~vmsanchez/pub/sanchez11b.pdf`)

La construcción del sistema basado en la nueva estrategia de hibridación que resultó ganador en la evaluación humana de la tarea de traducción del *Workshop on Statistical Machine Translation* de 2011 (Callison-Burch et al., 2011) se explica en la siguiente publicación:

- Sánchez-Cartagena, V. M., Sánchez-Martínez, F. y Pérez-Ortiz, J. A. (2011c). The Universitat d'Alacant hybrid machine translation system for WMT 2011. En *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 457–463, Edinburgh, Scotland (`http://www.dlsi.ua.es/~vmsanchez/pub/sanchez11c.pdf`)

Inserción de entradas en diccionarios morfológicos por parte de usuarios no expertos

La creación de diccionarios morfológicos consume una gran parte del tiempo de desarrollo de un sistema de TA basado en reglas si éstos no pueden ser reutilizados a partir de otros sistemas de TA o aplicaciones de procesamiento del lenguaje natural (Tyers, 2010). Además, los métodos de inferencia automática de reglas e hibridación descritos respectivamente en los capítulos 2 y 3 precisan de diccionarios morfológicos. Si se permitiera a usuarios sin grandes conocimientos lingüísticos ni sobre sistemas de TA basados en reglas participar en la creación de diccionarios morfológicos, se podría acelerar y abaratar el desarrollo de nuevos sistemas de TA basados en reglas. Con este objetivo, se ha desarrollado un nuevo método que permite a usuarios no expertos insertar nuevas entradas en diccionarios morfológicos monolingües, descrito en el capítulo 4.

Este nuevo método está pensado para ser empleado cuando un usuario de un sistema de TA basado en reglas desea traducir un texto que contiene palabras desconocidas (que no están presentes en el diccionario morfológico en LO). El método se emplea para permitir al usuario insertarlas y que, de este modo, el sistema sea capaz de analizarlas. Si el usuario es bilingüe, puede proporcionar la traducción a LM de cada palabra desconocida y el método puede emplearse también para introducirlas en el diccionario morfológico en LM. Una vez se han insertado las entradas en ambos diccionarios morfológicos monolingües, la entrada correspondiente en el diccionario bilingüe puede insertarse automáticamente.

Este nuevo método funciona a base de preguntar iterativamente al usuario "*es X una forma válida de la palabra W?*" siendo W la palabra desconocida encontrada en el texto a traducir y X una nueva palabra formada tras realizar cambios en la flexión de W. Por ejemplo, supongamos que el usuario desea traducir al español la oración en inglés *Many of those policies remain largely unimplemented* y el sistema de TA no contiene la forma superficial *policies* en su diccionario morfológico en LO. El sistema preguntaría al usuario si *policy* y *policying* son formas válidas de *policies*. Si el usuario

fuese bilingüe (imaginemos que quiere poseditar la traducción ofrecida por el sistema de TA), especificaría que la traducción al español de *policies* es *medidas*, y el sistema le preguntaría si *medida* o *medidaba* son formas válidas.

El nuevo método presentado en el capítulo 4 emplea los paradigmas de flexión existentes en el diccionario morfológico para insertar la nueva entrada. El método selecciona la raíz de palabra y elige el paradigma de flexión más adecuado. Como resultado, además de la forma superficial desconocida (*policies*: *policy, nombre, plural* en el ejemplo anterior), todas sus formas flexionadas se añadirán también al diccionario (en el ejemplo anterior, *policy*: *policy, nombre, singular*). La raíz es el prefijo de la palabra a insertar que es común a todas sus formas flexionadas. Los paradigmas de flexión se utilizan en los diccionarios de los sistemas de TA para agrupar regularidades en la flexión de las palabras: un paradigma se define normalmente como un conjunto de sufijos (que se concatenan a la raíz de la palabra para construir sus diferentes formas flexionadas) y la información morfológica correspondiente a cada uno de ellos.

El proceso seguido para insertar la entrada en el diccionario monolingüe es el siguiente. Primero, se examinan los paradigmas de flexión presentes en el diccionario para seleccionar aquellos que son compatibles con la forma superficial a insertar. Los paradigmas compatibles son aquellos que comparten un sufijo con ella. Después, a cada paradigma se le asigna una puntuación que indica cómo de probable es que dicho paradigma sea el más adecuado para la palabra a insertar. Esta puntuación se obtiene, mediante un modelo oculto de Markov (Rabiner, 1989), a partir de un corpus monolingüe y de la oración del texto que el usuario desea traducir en la que la palabra a insertar aparece. Si el usuario está introduciendo una palabra en LM y acaba de introducir su equivalente en LO, el paradigma asignado a la palabra en LO se emplea para obtener una puntuación más precisa, pues los paradigmas en LO y LM están fuertemente correlacionados. La puntuación ayuda a elegir qué formas superficiales deberán ser validadas por el usuario. Dichas formas se obtienen flexionando la palabra a insertar con los diferentes paradigmas compatibles. Si el usuario acepta una forma superficial, todos los paradigmas que no sean capaces de generarla son descartados. Si el usuario rechaza una forma superficial, todos los paradigmas que la generan son descartados. El proceso se repite iterativamente hasta que sólo queda un paradigma compatible, que constituye la solución. Es deseable que la cantidad de preguntas necesarias para elegir un paradigma sea lo más baja posible. Para ello, las formas superficiales a validar se deciden mediante un árbol de decisión construido con una variante del algoritmo ID3 (Quinlan, 1986) que tiene en cuenta tanto la puntuación de cada paradigma como el número de paradigmas que se descartarían con cada respuesta del usuario.

Para evaluar este nuevo método, se ha seleccionado un pequeño subconjunto de 150 entradas del diccionario monolingüe español del proyecto Apertium, se ha elegido la forma superficial más común para cada una de ellas, y se ha pedido a un grupo de usuarios no expertos que las inserten. El resultado muestra que los usuarios no expertos son perfectamente capaces de contestar a las preguntas propuestas por el sistema y que el paradigma correcto es elegido en cerca del 90 % de los casos. Además,

también se ha realizado una evaluación automática más exhaustiva empleando miles de entradas extraídas del historial de cambios del mismo diccionario monolingüe español. En esta evaluación, se ha asumido que el usuario contesta correctamente a las preguntas planteadas por el sistema y se ha contado el número de preguntas que son necesarias como media. Los resultados confirman que unas pocas preguntas (no más de 6) son suficientes para poder elegir el paradigma correcto.

Cabe destacar, sin embargo, que cuando existen distintos paradigmas candidatos que generan el mismo conjunto de formas superficiales (pero con diferente información morfológica asociada) no es posible que el usuario elija el más adecuado mediante la validación de formas flexionadas. Para solventar esta limitación, múltiples líneas de investigación pueden seguirse en el futuro. Por una parte, y como demuestran algunos experimentos preliminares, es posible emplear la información contenida en un corpus monolingüe y un modelo de lengua similar al empleado por sistemas de TA estadísticos (pero basado en información morfológica en lugar de en formas superficiales) para elegir el paradigma con la información morfológica más verosímil. También sería posible plantear al usuario preguntas más sofisticadas: por ejemplo, podrían mostrársele oraciones en las que la palabra a insertar actúa con diferentes categorías léxicas o atributos morfológicos y elegir el paradigma final en función de las oraciones consideradas correctas por el usuario.

Este nuevo método para permitir a usuarios no expertos insertar nuevas entradas en diccionarios morfológicos se describe en las siguientes publicaciones:

- Esplà-Gomis, M., Sánchez-Cartagena, V. M. y Pérez-Ortiz, J. A. (2011a). Enlarging monolingual dictionaries for machine translation with active learning and non-expert users. En *Proceedings of Recent Advances in Natural Language Processing*, pp. 339–346, Hissar, Bulgaria (`http://www.dlsi.ua.es/~vmsanchez/pub/espla11a.pdf`)

- Esplà-Gomis, M., Sánchez-Cartagena, V. M., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Forcada, M. L. y Carrasco, R. C. (2014). An efficient method to assist non-expert users in extending dictionaries by assigning stems and inflectional paradigms to unknown words. En *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pp. 19–26, Dubrovnik, Croatia (`http://www.dlsi.ua.es/~vmsanchez/pub/espla14.pdf`)

Discusión

En conclusión, en esta tesis doctoral se han presentado tres nuevos métodos que facilitan la creación de sistemas de TA cuando los recursos normalmente empleados en su construcción (corpus paralelos y recursos lingüísticos, como reglas y diccionarios) son escasos: un algoritmo de inferencia de reglas de transferencia morfológica avanzada a

partir de corpus paralelos muy pequeños, una estrategia de hibridación entre TA basada en reglas y TA estadística y un sistema que permite a usuarios no expertos insertar entradas en diccionarios morfológicos monolingües.

El nuevo algoritmo para la inferencia automática de reglas de transferencia constituye una alternativa barata (en términos de los recursos humanos que son necesarios) y eficaz para construir sistemas de TA cuando únicamente hay disponibles diccionarios y un corpus paralelo muy pequeño. Su gran poder de generalización le permite crear reglas de transferencia de alta calidad y fácilmente editables a partir de corpus paralelos que contienen únicamente unos millares de palabras en cada idioma. La adopción de este algoritmo contribuirá a facilitar la creación de reglas de transferencia para nuevos pares de lenguas en sistemas como Apertium, reduciendo así el tiempo total necesario para construir nuevos sistemas de TA.

La alta capacidad de generalización del algoritmo de inferencia de reglas puede contribuir también a la mejora de sistemas de TA estadísticos si se combina con la estrategia de hibridación descrita en el capítulo 3. Esta combinación constituye una manera novedosa y no explorada previamente de emplear diccionarios morfológicos para mejorar sistemas de TA estadísticos. De acuerdo con los resultados de los experimentos llevados a cabo, se puede concluir que la combinación de ambos métodos contribuye a mitigar el problema de dispersión de datos que normalmente sufren los sistemas de TA estadísticos cuando deben traducir entre idiomas con un alto grado de flexión morfológica y reduce el tamaño de los corpus necesarios para construir sistemas de TA estadísticos.

Respecto al método para permitir a usuarios no expertos insertar entradas en diccionarios morfológicos monolingües, éste permitirá ahorrar costes en el desarrollo de dichos diccionarios y, por extensión, de nuevos sistemas de TA basados en reglas. A pesar de la limitación que afecta a los paradigmas de flexión que generan las mismas formas superficiales, este nuevo método podría contribuir al ahorro de costes en su estado actual si los usuarios expertos únicamente intervienen al final del proceso para decidir entre los (pocos) paradigmas que generan las mismas formas superficiales, mientras que los usuarios no expertos llevan a cabo el resto del trabajo.

Finalmente, cabe destacar que la implementación de todos los métodos descritos en esta tesis ha sido liberada bajo licencia GNU GPL (los paquetes de software se describen en el Apéndice B). Esta decisión trae consigo varias ventajas. Por una parte, asegura la reproducibilidad de los resultados presentados y facilita que la comunidad científica continúe la investigación llevada a cabo. Y por otra parte, permite que realmente se construyan sistemas de TA para pares de lenguas con recursos escasos sin tener que desarrollar software adicional.

Preface

Machine translation (MT) is the process carried out by a computer in order to automatically translate a text in a natural language, the source language, into another natural language, the target language. According to the kind of knowledge used in their development, machine translation systems may be said to be corpus-based or rule-based. Corpus-based approaches use large collections of parallel texts (*parallel corpora*) as the main source of knowledge —statistical machine translation (SMT) being the leading corpus-based approach—, while rule-based MT (RBMT) systems use linguistic resources such as dictionaries and structural transfer rules. As SMT systems need relatively big corpora in order to be competitive, they are unsuitable for language pairs for which the required amount of parallel data is not available (less-resourced language pairs). Thus, RBMT becomes the approach of choice. However, building RBMT systems usually implies a considerable investment in the development of the linguistic resources, some of which can only be developed by trained experts.

During my work as a predoctoral researcher at the Departament de Llenguatges i Sistemes Informàtics at Universitat d'Alacant, I got involved in the development of the Apertium free/open-source RBMT platform and I witnessed the great effort done by the community grown around Apertium in order to create linguistic resources and RBMT systems for less-resourced language pairs. I noticed that, if the development of the resources was done in a more easy and efficient way, probably more users would be motivated to contribute to the project and Apertium-based MT systems for new language pairs would be more often created. Consequently, I started to research on how to ease the development of linguistic resources for Apertium. Afterwards, when I moved to the language technology company Prompsit Language Engineering, I understood that easing the building of MT resources could reduce the time needed for deploying working systems and thus increase customer satisfaction and improve competitiveness. On that basis, three new methods that ease the building of MT systems when resources (both parallel corpora and RBMT linguistic data) are scarce are presented in this dissertation. Empirical proofs of their successful application for building MT systems for different language pairs are also provided. These methods will help the Apertium community to build MT systems for less-resourced language pairs and they will also be useful in order to speed up the development of new MT systems in the language technology industry.

Firstly, a new method that uses scarce parallel corpora (barely a few hundreds of parallel sentences) and existing morphological dictionaries to automatically infer a set of shallow-transfer rules to be integrated into an RBMT system is described. This new method avoids the need for human experts to hand-craft these rules and overcomes many relevant limitations of previous rule inference approaches. Namely, it is able to achieve a high degree of generalisation over the linguistic phenomena observed in the training corpus, and it is able to select the proper subset of rules which ensure the most appropriate segmentation of the input sentences to be translated. In addition, this new rule inference approach is the first one in which the conflicts between candidate rules that arise during the inference process are resolved by choosing the most appropriate ones according to a global minimisation function rather than proceeding in a pairwise greedy fashion. Experiments conducted using five different language pairs with Apertium show that translation quality significantly improves when compared to previous approaches and it is close to that obtained using hand-crafted rules. Moreover, the resulting number of rules is considerably smaller, which eases human revision and maintenance. The adoption of the rule inference approach presented in this dissertation will hopefully contribute towards making the development of transfer rules for new language pairs in MT systems like Apertium a much more cost-effective process.

Secondly, a new hybridisation strategy aimed at integrating shallow-transfer rules and dictionaries from RBMT into phrase-based SMT is presented. The new hybridisation strategy, which is specific for shallow-transfer RBMT, addresses the main limitations of existing strategies for integrating RBMT resources into SMT; namely, the presence of alignment errors in the phrase pairs obtained from the RBMT system and the inability to find an adequate balance between the weight of the phrase pairs extracted from the parallel corpus and those obtained from the RBMT system. The experiments performed confirm that the new approach delivers higher translation quality than existing ones, and that shallow-transfer rules are specially useful when the parallel corpus available for training is small or when translating out-of-domain texts that are well covered by the shallow-transfer RBMT system. Indeed, a system built by following this hybridisation approach was one of the winners of the pairwise manual evaluation of the Workshop on Statistical Machine Translation 2011 shared translation task for the Spanish–English language pair. In addition, the shallow-transfer rules that are integrated into a phrase-based SMT system can be automatically inferred with the new rule inference approach also presented in this dissertation. The experiments carried out show that the translation quality achieved by hybrid systems built with automatically inferred rules is often similar to that obtained with hand-crafted rules. The combination of the hybridisation strategy and the rule inference algorithm presented in this dissertation will contribute to alleviate the data sparseness problem suffered by SMT systems, since the resulting hybrid system is able to generalise the translation knowledge contained in the parallel corpus to sequences of words that have not been observed in the corpus.

Thirdly, as the two aforementioned approaches need morphological dictionaries in order to be applied, and with the aim of easing the task of building them, a novel approach that allows non-expert users to insert new entries in monolingual morphological dictionaries is presented. The scenario considered is that of non-expert users of an RBMT system who have to introduce into its dictionaries the words found in an input text that are unknown to the system, so that it can subsequently correctly translate them. Given a source-language surface form (i.e., a word as it is found in running texts, without any kind of analysis) to be inserted, the proposed strategy iteratively asks the users (average speakers of a language) polar questions to validate whether certain inflected forms of the word to be inserted are correct. The new approach uses the answers of the users and the existing inflection paradigms in the monolingual dictionary in order to automatically insert the corresponding entry. An inflection paradigm may be defined as a collection of suffixes and their corresponding morphological information; they are commonly used in RBMT systems to group regularities in the inflection of a set of words. In addition, a monolingual corpus, a hidden Markov model and a binary decision tree built with the ID3 algorithm are used to reduce the number of polar questions that need to be asked for gathering all the necessary information for the insertion of the entry. The experiments carried out show that non-expert users are able to successfully answer the polar questions in most cases, and that the approach is efficient: only 5–6 questions on average were needed in order to insert a set of words selected from the revision history of a real Apertium monolingual dictionary. If the user is bilingual and provides the translation of the inserted source-language word, the process is repeated to insert the corresponding entry in the target-language monolingual dictionary. In this case, the information about the source-language entry already inserted and the correlation between morphological features in both languages is used to further increase the efficiency of the approach. Once the entries have been inserted in both monolingual morphological dictionaries, the corresponding entry in the bilingual dictionary can be inserted automatically. The adoption of this approach will contribute to save costs in the development of new dictionaries for RBMT and other natural language processing applications.

The implementation of all the methods described in this dissertation has been released under the GNU GPL license. The release of the tools has two main advantages. On the one hand, it ensures the reproducibility of the results presented in this dissertation and makes it easier for the scientific community to continue the research, either by following the future research lines proposed in this dissertation, or by starting new ones. On the other hand, it permits the effective achievement of the main objective of the research carried out: easing the construction of MT system for language pairs with scarce resources.

This thesis has been possible thanks to the ideas and constant supervision of Dr. Felipe Sánchez-Martínez and Dr. Juan Antonio Pérez-Ortiz from the Departament de Llenguatges i Sistemes Informàtics at Universitat d'Alacant. Suggestions by Dr. Mikel L. Forcada and Dr. Rafael C. Carrasco have also been very useful, especially

during the development of the approach for allowing non-expert users to insert entries in morphological dictionaries. The work and ideas by Miquel Esplà-Gomis have also been very valuable.

Structure of this dissertation

This dissertation is structured in 5 chapters and 2 appendices, organised as follows:

Chapter 1 introduces the most important concepts and definitions about machine translation, explains the motivation for the development of the new approaches and describes the related work that can be found in the literature.

Chapter 2 presents the new method that uses scarce parallel corpora and existing morphological dictionaries to automatically infer a set of shallow-transfer rules to be integrated into an RBMT system.

Chapter 3 describes the new hybridisation strategy aimed at integrating shallow-transfer rules and dictionaries from RBMT into phrase-based SMT.

Chapter 4 explains the novel approach that allows non-expert users to insert new entries in monolingual morphological dictionaries.

Chapter 5 summarises the main contributions in this dissertation and outlines some future research lines.

Appendix A explains in detail the Apertium shallow-transfer RBMT platform, that has been used as the reference shallow-transfer RBMT system in which all the approaches presented in this dissertation have been evaluated.

Appendix B lists the open-source tools released as part of this thesis.

Publications

Some parts of this dissertation have been published in journals and peer-reviewed conference or workshop proceedings. The list below shows them in chronological order. The chapter to which each publication is related is shown in brackets.

- Sánchez-Cartagena, V. M., Sánchez-Martínez, F., and Pérez-Ortiz, J. A. (2011c). The Universitat d'Alacant hybrid machine translation system for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 457–463, Edinburgh, Scotland (`http://www.dlsi.ua.es/~vmsanchez/pub/sanchez11c.pdf`). [**Chapter 3**]

- Sánchez-Cartagena, V. M., Sánchez-Martínez, F., and Pérez-Ortiz, J. A. (2011a). Enriching a Statistical Machine Translation System Trained on Small Parallel Corpora with Rule-Based Bilingual Phrases. In *Proceedings of Recent Advances in Natural Language Processing*, pages 90–96, Hissar, Bulgaria (`http://www.dlsi.ua.es/~vmsanchez/pub/sanchez11a.pdf`). [**Chapter 3**]

- Esplà-Gomis, M., Sánchez-Cartagena, V. M., and Pérez-Ortiz, J. A. (2011a). Enlarging monolingual dictionaries for machine translation with active learning and non-expert users. In *Proceedings of Recent Advances in Natural Language Processing*, pages 339–346, Hissar, Bulgaria (`http://www.dlsi.ua.es/~vmsanchez/pub/espla11a.pdf`). [**Chapter 4**]

- Sánchez-Cartagena, V. M., Sánchez-Martínez, F., and Pérez-Ortiz, J. A. (2011b). Integrating shallow-transfer rules into phrase-based statistical machine translation. In *Proceedings of the Machine Translation Summit XIII*, pages 562–569, Xiamen, China (`http://www.dlsi.ua.es/~vmsanchez/pub/sanchez11b.pdf`). [**Chapter 3**]

- Esplà-Gomis, M., Sánchez-Cartagena, V. M., and Pérez-Ortiz, J. A. (2011b). Multimodal Building of Monolingual Dictionaries for Machine Translation by Non-Expert Users. In *Proceedings of the Machine Translation Summit XIII*, pages 147–154, Xiamen, China (`http://www.dlsi.ua.es/~vmsanchez/pub/espla11b.pdf`). [**Chapter 4**]

- Sánchez-Cartagena, V. M., Esplá-Gomis, M., and Pérez-Ortiz, J. A. (2012). Source-Language Dictionaries Help Non-Expert Users to Enlarge Target-Language Dictionaries for Machine Translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 3422–3429, Istanbul, Turkey (`http://www.dlsi.ua.es/~vmsanchez/pub/sanchez12a.pdf`). [**Chapter 4**]

- Sánchez-Cartagena, V. M., Sánchez-Martínez, F., and Pérez-Ortiz, J. A. (2012b). An open-source toolkit for integrating shallow-transfer rules into phrase-based satistical machine translation. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 41–54, Gothenburg, Sweden (`http://www.dlsi.ua.es/~vmsanchez/pub/sanchez12b.pdf`). [**Chapter 3**]

- Sánchez-Cartagena, V. M., Esplà-Gomis, M., Sànchez-Martínez, F., and Pérez-Ortiz, J. A. (2012). Choosing the correct paradigm for unknown words in rule-based machine translation systems. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 27–39, Gothenburg, Sweden (`http://www.dlsi.ua.es/~vmsanchez/pub/sanchez12c.pdf`). [**Chapter 4**]

- Esplà-Gomis, M., Sánchez-Cartagena, V. M., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Forcada, M. L., and Carrasco, R. C. (2014). An efficient method to assist non-expert users in extending dictionaries by assigning stems and inflectional paradigms to unknown words. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 19–26, Dubrovnik, Croatia (`http://www.dlsi.ua.es/~vmsanchez/pub/espla14.pdf`). [**Chapter 4**]

- Sánchez-Cartagena, V. M., Pérez-Ortiz, J. A., and Sánchez-Martínez, F. (2014). The UA-Prompsit hybrid machine translation system for the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 178–185, Baltimore, MD, USA (`http://www.dlsi.ua.es/~vmsanchez/pub/sanchez14.pdf`). [**Chapters 2 and 3**]

- Rubino, R., Toral, A., Sánchez-Cartagena, V. M., Ferrández-Tordera, J., Ortiz-Rojas, S., Ramírez-Sánchez, G., Sánchez-Martínez, F., and Way, A. (2014). Abu-MaTran at WMT 2014 translation task: Two-step data selection and RBMT-style synthetic rules. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 171–177, Baltimore, MD, USA (`http://www.dlsi.ua.es/~vmsanchez/pub/rubino14.pdf`). [**Chapters 2 and 3**]

- Sánchez-Cartagena, V. M., Pérez-Ortiz, J. A., and Sánchez-Martínez, F. (2015). A generalised alignment template formalism and its application to the inference of shallow-transfer machine translation rules from scarce bilingual corpora. *Computer Speech & Language*, 32(1):46–90. Hybrid Machine Translation: integration of linguistics and statistics (`http://www.dlsi.ua.es/~vmsanchez/pub/sanchez15.pdf`). [**Chapter 2**]

I have also published other papers in peer-reviewed conference proceedings that, although not directly related to the work presented in this dissertation, are linked to the Apertium open-source shallow-transfer MT engine:

- Sánchez-Cartagena, V. M. and Pérez-Ortiz, J. A. (2009). An open-source highly scalable web service architecture for the Apertium machine translation engine. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 51–58 (`http://www.dlsi.ua.es/~vmsanchez/pub/sanchez09.pdf`).

- Sánchez-Cartagena, V. M. and Pérez-Ortiz, J. A. (2010a). ScaleMT: a Free/Open-Source Framework for Building Scalable Machine Translation Web Services. *The Prague Bulletin of Mathematical Linguistics*, 93:97–106 (Presented at the Fourth Machine Translation Marathon in Dublin, Ireland; `http://www.dlsi.ua.es/~vmsanchez/pub/sanchez10a.pdf`).

- Sánchez-Cartagena, V. M. and Pérez-Ortiz, J. A. (2010b). Tradubi: Open-source social translation for the apertium machine translation platform. *The*

Prague Bulletin of Mathematical Linguistics, 93:47–56 (Presented at the Fourth Machine Translation Marathon in Dublin, Ireland; `http://www.dlsi.ua.es/~vmsanchez/pub/sanchez10b.pdf`).

- Ivars-Ribes, X. and Sánchez-Cartagena, V. M. (2011). A Widely Used Machine Translation Service and its Migration to a Free/Open-Source Solution: the Case of Softcatalà. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 61–68, Barcelona, Spain (`http://www.dlsi.ua.es/~vmsanchez/pub/ivars11.pdf`).

- Détrez, G., Sánchez-Cartagena, V. M., and Ranta, A. (2014). Sharing Resources Between Free/Open-Source Rule-based Machine Translation Systems: Grammatical Framework and Apertium. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 4394–4400, Reykjavik, Iceland (`http://www.dlsi.ua.es/~vmsanchez/pub/detrez14.pdf`).

Contents

Chapter 1

Introduction

Machine translation is the process carried out by a computer to translate a text in a natural language into another natural language. Machine translation systems are built from already translated texts, which can be difficult to find for some languages, or from linguistic resources, such as dictionaries and translation rules, which are usually created by trained experts in a highly time-consuming task. This dissertation describes three new approaches aimed at easing the development of machine translation systems when the aforementioned resources (both translated texts and linguistic resources) are scarce. In this chapter, the most important concepts and definitions about machine translation are introduced, the motivation for the development of the new approaches is explained, and the related work that can be found in the literature is also described.

1.1 Machine translation

Machine translation (MT) may be defined as the process carried out by a computer to translate a text in a natural language, the *source language* (SL), into another language, the *target language* (TL). It can be seen as a sub-field of artificial intelligence (Wilks, 1978). MT is a very active research field since the 1950s (Weaver, 1955).

Translation is difficult for computers and poses a research challenge because of the following factors (Arnold, 2003; Costa-Jussà, 2012):

- Natural languages are ambiguous. The same word may have different meanings: *book* does not have the same meaning in the phrase *read a book* as in the phrase *book a flight*.

- Knowledge of the real world is needed in order to solve the ambiguity of a natural language: for example, in order to translate the sentence *I will bring my bike*

tomorrow if it looks nice in the morning, one must be aware that *it* refers to the weather because people usually ride bikes when the weather is nice and bikes do not usually change their aspect in the mornings.

- The same meaning can be expressed in many different ways. MT systems must be able to deal with all the possible ways of expressing the same meaning in order to be able to translate as many input constructions as possible. They must also choose the most appropriate way of expressing the meaning. For instance, the Spanish words *coche*, *carro* and *automóvil* mean *car* in English. An MT system that translates from Spanish to English must be able to translate the three Spanish words as *car*. Moreover, an MT system that translates from English to Spanish must choose the most appropriate translation of *car*, which depends on multiple factors such as the geographical variant of the TL or the degree of formality of the text. In a formal environment, the most appropriate translation is *automóvil*; otherwise, *coche* is commonly used in Spain while *carro* is more often used in America.

- The grammar of the languages involved can be totally different. Thus, expressing the meaning of an SL sentence in the TL involves adapting it to the particular grammatical rules of the TL. The resulting sentence can have a totally different structure from its original structure in the SL. For instance, when translating a possessive phrase built with the Saxon genitive from English to Spanish, the position of the noun phrase that acts as owner is moved after the noun phrase that acts as the owned element: *Bob and Alice's house* is translated into Spanish as *la casa de Bob y Alice*.

- As a consequence of the factors mentioned above, translation needs a huge amount of human knowledge which must be described or learned from corpora, and coded in a computer-usable form.

As a result of these challenges, high-quality MT for general texts has not been achieved yet, except for the translation between closely related languages. However, when the text to be translated belongs to a limited domain, the difficulty of the aforementioned challenges is reduced, and high-quality MT systems can be built (Koehn, 2010, Sec. 1.3.1). Examples of texts from limited domains than can be translated by a computer with high-quality are weather forecasts, summaries of sports events, and rail or flight information, among many others. For instance, when translating flight information, it is clear that most of the times *book* means *reserve*, and the set of possible sentences to be translated is sufficiently restricted that the human knowledge needed to translate them can be easily encoded in a computer-usable form.

The fact that high-quality MT is not generally available for broad domains does not mean that it is useless. Hutchins (2010) defines two main uses for MT:

- The need for high-quality translation that will be published, such as the multilingual documentation in large corporations. This use is called *dissemination*.

The output of MT systems can save time and costs in this case by providing draft translations that are manually edited before publication in a process called *postediting*.

- Often, a perfect translation is not needed for a particular person, but a quick, grammatically incorrect TL text that captures the meaning of the SL text is enough. This is the case, for instance, when a person who is browsing a foreign website needs a general view of its content. This use is called *assimilation*.

Another criterion that can be followed for the classification of MT systems is the kind of knowledge used to build them. On that basis, two main paradigms can be found: rule-based MT (RBMT) and corpus-based MT. On the one hand, RBMT systems (Hutchins and Somers, 1992) use linguistic resources usually hand-crafted by human experts, such as morphological dictionaries or transfer rules, that describe the translation process. Corpus-based approaches, meanwhile, use large collections of parallel texts as the source of knowledge. A *parallel text* is a text that is placed alongside its translation into another language; a collection of parallel texts is usually referred to as a *parallel corpus*. Statistical MT (SMT) (Koehn, 2010) is currently the leading paradigm in corpus-based MT. Hybrid approaches that combine features from both paradigms are also possible (Thurmair, 2009; Costa-Jussà and Farrús, 2014), and will be described in depth later in this chapter.

1.1.1 Rule-based machine translation

Rule-based approaches to MT perform the translation using the linguistic information created by human experts in three steps:

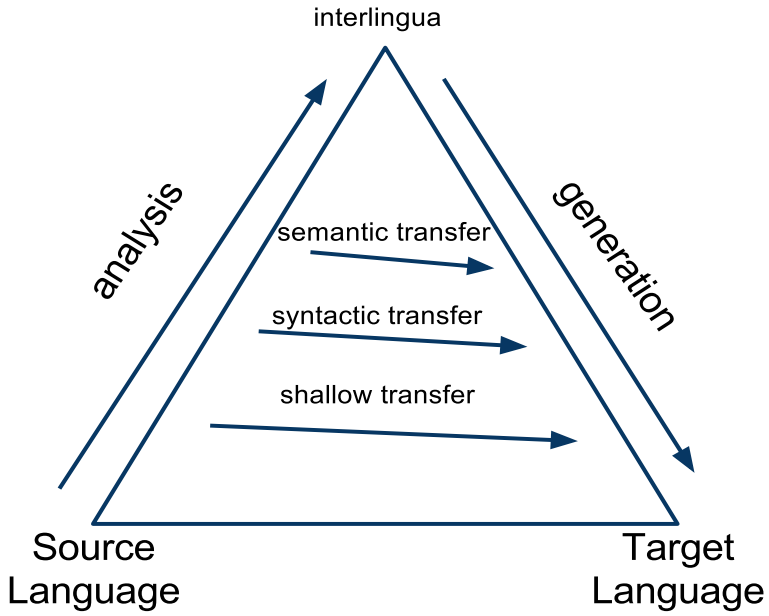
1. *Analysis* of the SL text to build an SL intermediate representation (SL IR). This representation removes from the SL text the features which are not relevant to the translation process and makes the relevant ones more explicit.

2. *Transfer* from that SL IR to a TL intermediate representation (TL IR).

3. *Generation* of the final translation from the TL IR.

Traditionally, RBMT architectures have been classified according to the degree of abstraction of the intermediate representations they use. Figure 1.1 shows the *Vauquois triangle*, that summarises this taxonomy (Vauquois, 1968). The horizontal arrows represent the transfer effort, while the vertical ones represent the effort needed for analysis and generation. Architectures with a more abstract IR are located higher in the triangle. The more abstract the IRs, the simpler the transfer step. However, highly abstract IRs are difficult to design and involve a more difficult analysis step that may not be worth the effort.

Figure 1.1: Vauquois triangle: comparison of RBMT paradigms.



Depending on the nature of this IR, RBMT systems can be said to be *interlingua* based or *transfer* based. Interlingua-based RBMT systems use a language-independent IR; this makes analysis and generation difficult —and almost impossible for broad domains— but avoids the need for transfer (they are represented at the top of the triangle). Transfer-based RBMT systems use language-dependent IRs and include a transfer stage which transforms the SL IR into the TL IR by applying lexical and structural transfer rules. Since they are language-dependent, the IRs used in transfer-based RBMT are much easier to develop than those used by interlingua-based systems, thus making transfer-based RBMT the leading approach in RBMT.

Transfer-based RBMT systems can in turn be classified according to the complexity of the IR used into *shallow-transfer*, *syntactic-transfer*, and *semantic-transfer* RBMT systems. In this dissertation, the main focus is on *shallow-transfer* RBMT systems, which are those that perform a shallow syntactic analysis of the SL, i.e. they do not perform full syntactic parsing and do not build a parse tree. This signifies that the IR they use is as simple as a sequence of *lexical forms* (lemma, lexical category and morphological inflection information) of the words to be translated.

The Apertium platform (Forcada et al., 2011) has been used as the reference shallow-transfer RBMT system in which all the approaches presented in this dissertation have been evaluated. The Apertium platform (and shallow-transfer RBMT in general) has been chosen because of the following reasons:

- Apertium is a free/open-source project, thus it can be easily modified in order to include the new methods described in this dissertation.
- A great community of users has grown around Apertium because writing linguistic resources for Apertium is relatively easy (Forcada et al., 2011). The fact that Apertium uses shallow-transfer rules significantly contributes to ease the development of the resources, since Apertium developers are relieved from dealing from the recursivity of grammars.
- That community, which includes many speakers of less-resourced languages,[1] will hopefully take advantage of the scientific contributions presented in this dissertation and will contribute to their dissemination. As a result, the development of MT systems for less-resourced languages will be boosted.

Apertium uses a morphological analyser to obtain SL lexical forms from the SL sentence to be translated. A part-of-speech tagger (Sánchez-Martínez et al., 2008) solves part-of-speech ambiguities when more than one lexical form can correspond to the same word (*surface form*[2]) in the input. The transfer step is split into two parts: lexical transfer, in which the individual translation of each word is obtained, and structural transfer, that performs the required operations (agreements, reorderings, etc.) when word-for-word translation is not enough to produce a correct TL text. Structural transfer is performed by shallow-transfer rules. Rules are applied to sequences of lexical forms in the SL and produce TL lexical forms as their output. Rules operate exclusively on the lexical forms they have matched, regardless of the SL lexical forms matched by other rules, and are applied in a greedy, left-to-right, longest match fashion. Finally, the morphological generator module generates a sequence of surface forms from the TL IR. A more detailed description of the Apertium shallow-transfer RBMT platform can be found in Appendix A.

1.1.2 Statistical machine translation

The SMT approach works by building statistical models from parallel texts, and then finding, for each SL sentence $s = (s_1, s_2, ..., s_n)$ to be translated, the TL sentence $\hat{t} = (t_1, t_2, ..., t_m)$ that maximises the conditional probability $p(t|s)$ according to those models. In the first SMT approaches (Brown et al., 1993b), the Bayes rule was applied to compute $p(t|s)$ as follows:

$$\hat{t} = \arg\max_t \{p(t|s)\} = \arg\max_t \left\{ \frac{p(s|t) \cdot p(t)}{p(s)} \right\} = \arg\max_t \{p(s|t) \cdot p(t)\}$$

This transformation is known as the *noisy channel* model. On the one hand, $p(t)$ is obtained from a TL model, i.e. a model which tells us how likely it is that the sentence

[1] Languages for which the amount of linguistic resources and corpora available is scarce.

[2] In the remainder of this dissertation, the term surface form will be used to refer to words as they are found in running texts.

t occurs in the TL (Goodman and Chen, 1998). TL models are usually estimated from plain texts in the TL. On the other hand, $p(s|t)$ expresses how likely it is that s is the translation of t and it is obtained from an *inverse translation model*. In the first SMT approaches, the inverse translation model was a word-based translation model (Brown et al., 1993b) estimated from a parallel corpus. The model stores information about the probability of each SL word being the translation of a TL word. Therefore, the value of $p(s|t)$ can be computed by combining the values of these word-to-word translation probabilities for the different words of s and t.

Since the MT challenges described at the beginning of Section 1.1 cannot be overcome by translating words in isolation, several improvements to the word-based translation model have been developed: phrase-based models (Koehn et al., 2003), the use of alignment templates (Och and Ney, 2004), n-gram-based models (Marino et al., 2006), factored models (Koehn and Hoang, 2007), or hierarchical phrase-based models (Chiang, 2007); being phrase-based SMT one of the most popular ones (Costa-Jussà and Farrús, 2014, Sec. 3).

The translation model in phrase-based SMT assigns probabilities to pairs of word sequences, which are called *phrases*,[3] instead of to pairs of words. In order to find the TL sentence $\hat{t}$ that maximises $p(t|s)$, multiple translation hypotheses are generated in the following way: the SL sentence is segmented into phrases, each SL phrase is translated according to the phrase pairs extracted from the parallel corpus, and the translations obtained are assembled together (not necessarily in the order in which the SL phrases are found in the SL sentence). The set of phrase pairs extracted from the parallel corpus and their probabilities is usually called *phrase table* or *phrase translation model*. In order to extract phrase pairs from the parallel corpus, first the mappings between words in the SL and TL sentences, called *word alignments*, are computed using the aforementioned word-based translation models (Och and Ney, 2003). Then, phrase pairs are extracted using word alignments as anchors (Koehn, 2010, sec. 5.2.3). Unlike the word-based translation model used in the first SMT approaches (Brown et al., 1993b), the phrase-based translation model takes into account local context information. For instance, according to the phrase-based translation model, the most probable English translation of the Spanish phrase *estación de tren* would be *train station*, while the most probable translation of *estación de esquí* would be *ski resort*, as long as enough sentences containing these particular translations appear in the parallel corpus.

Additionally, in state-of-the-art SMT systems, the noisy channel model is no longer used to score the translation hypotheses; instead, a log-linear combination of different models, including the previously presented TL model and phrased-based translation

[3]Note that, in the SMT jargon, *phrase* is used to refer to any sequence of words, that does not necessarily constitute a syntactic constituent. For instance, an English–Spanish phrase-based SMT system would assign a relatively high probability to the phrase pair *my cat is very – mi gato es muy*.

model (Koehn, 2010, Sec. 5.3.1) is used:

$$\hat{t} = \arg\max_{t} \left\{ \exp\left(\sum_{m=1}^{M} \lambda_m \cdot h_m(s,t) \right) \right\}$$

where λ_m is the weight of model m, and $h_m(s,t)$ is the score assigned to the pair of sentences (s,t) by model m, also known as *feature function*.[4] Non-probabilistic models can be included in the log-linear combination: for instance, the number of words in the translation hypothesis (also known as word penalty) is a commonly used feature function. When m is a probabilistic model, $h_m(s,t)$ is computed as the logarithm of the probability of the pair of sentences (s,t) under m. State-of-the-art phrase-based SMT systems (Koehn, 2010, Sec. 5.3) usually include the following feature functions: source-to-target and target-to-source phrase translation probabilities, source-to-target and target-to-source word translation probabilities (known as lexical weightings; their purpose is acting as a back-off when scoring phrase pairs with low frequency), word penalty (number of words of the translation hypothesis), phrase penalty (number of phrases in which the SL sentence is split), reordering probability (probability of order of the TL phrases obtained from the phrase table being changed when they are assembled in order to build the translation hypothesis), and probability of the translation hypothesis according to the TL model.

In the SMT literature, the process of building the previously presented models from a parallel corpus is known as *training*, while the search of the TL sentence $\hat{t}$ with maximum probability according to the model combination is called *decoding*. Decoding is an NP-complete problem (Knight, 1999), so that heuristic decoding algorithms have to be used. The set of different weights λ_m is optimised on a small parallel corpus called *development corpus* through a process called *tuning* (Och, 2003; Watanabe et al., 2007).

1.1.3 Differences between machine translation paradigms

Given the fact that the two main paradigms described above approach the MT problem in two completely different ways, the type of translation errors produced by the systems built by following each of these paradigms are also very different.

On the one hand, SMT systems, thanks to the use of the TL model, tend to produce more fluent translations. In addition, as the translation knowledge is extracted from big amounts of already translated sentences, SMT systems deal better with expressions that should not be translated literally. Moreover, according to the error analysis performed by Dugast et al. (2008), SMT systems perform better lexical selections[5] and remove more effectively function words that are not needed in the TL.

[4] Some models, such as the TL model, only take into account t.

[5] Lexical selection is the task of choosing, given an SL word with multiple translations into the TL with the same lexical category, the most adequate translation.

On the other hand, RBMT systems perform translations that are more mechanical and repetitive, and sound less natural in the TL. However, this fact also makes the translation more faithful to the source text (Forcada et al., 2011). The language model in SMT systems improves fluency but can sometimes cause that the translation, although correct in the TL, does not convey exactly the meaning of the SL sentence. Nevertheless, repetitive translations ease the work of posteditors (Way, 2010, Sec. 3.4), as does the *terminological consistency* shown by RBMT systems. This means that, for the same SL word, they provide a reduced set of translations across the TL text produced by the system. Moreover, Dugast et al. (2008) also confirmed that RBMT systems do not suffer from the *data sparseness* problem usually found in SMT systems. Phrase-based SMT systems work with surface forms, that is, words as they are found in running texts, without any kind of analysis. Thus, in order to be able to correctly translate from or to highly inflectional languages, such as Spanish or German, all the inflected word forms[6] of a word (with their appropriate context, if necessary), must be observed in the training corpus. Given the complexity of some languages (for instance, most verbs in Spanish have 53 different inflected word forms), observing all of them in the training corpora is not always possible. Another advantage of RBMT systems (in particular, of those that perform a full syntactic analysis of the SL sentences) is the ability to perform long distance reorderings for language pairs with very different grammatical structures. For instance, English sentences usually contain the subject followed by the main verb and the objects, while in Japanese the common order is subject–object–verb. When translating from English to Japanese, the verb must be moved at the end of the sentence, after all the objects. This operation is difficult to perform with the phrase pairs a phrase-based SMT system relies on.[7]

1.1.4 Resources needed for machine translation

The type of translation errors made is not the only factor that must be taken into account when choosing the MT paradigm for a new system. Different paradigms need different types of resources (namely, human workforce or parallel texts). When creating a new MT system, the availability of the resources limits the kind of MT system that can be built.

On the one hand, SMT systems can be built without any human intervention from a parallel corpus (from which the translation model is built) and a TL monolingual corpus (from which the TL model is obtained). This fact has significantly contributed to their popularity, together with the growing availability of parallel corpora and the

[6]Inflection may be defined as the modification of a word to express grammatical categories such as tense, person, number, gender, etc. An *inflected word form* or simply *word form* is a form of a word obtained by inflection. For instance, the word form *cats* is the result of the inflection of the word *cat* for the plural number.

[7]The reordering model helps to correctly translate expressions that involve a long distance reordering but, according to Dugast et al. (2008), an RBMT system that performs a full syntactic analysis is more effective in this task.

increase in the power of CPUs and available memory. For instance, thanks to the publishing of the minutes of the European Parliament (Koehn, 2005), SMT systems that translate between most of the official languages of the European Union can be built automatically.

In general, the bottleneck of data availability for SMT is the parallel corpus needed to build the translation model, since monolingual data is usually easier to gather, and can even be borrowed from the web (Buck et al., 2014). Nevertheless, the bigger the monolingual corpus used for language modelling, the better (Koehn, 2010, Sec. 7.4). Concerning parallel corpora, SMT systems need an amount of parallel data in the order of millions of words in each language in order to be competitive.[8] These data requirements make SMT systems unsuitable for less-resourced language pairs and for the translation in the restricted domains for which the amount of data available is small.

On the other hand, building an RBMT system usually implies a considerable investment in the development of the linguistic resources, some of which can only be developed by trained experts who master certain skills: deep knowledge of the language or languages involved; advanced knowledge about linguistics and morphology; and expertise in the format used by the particular RBMT system. However, as the availability of a parallel corpus is not a requirement for RBMT systems, the RBMT approach is the approach of choice when building MT systems for the translation between less-resourced language pairs (e.g. Breton–French, Icelandic–English, Kazakh–Tatar) for which large parallel corpora are not readily available.[9]

Shallow-transfer RBMT systems, which are those relevant for this thesis, need monolingual morphological dictionaries for the analysis and generation steps, bilingual dictionaries for lexical transfer and shallow-transfer rules for structural transfer. The SL morphological dictionary contains mappings between SL surface forms and SL

[8]The exact size of the parallel corpus for which an SMT system becomes more competitive than an RBMT system depends on multiple factors, such as the effort spent in the development of the RBMT system, the domain of the parallel corpus (in comparison with the domain of the texts that will need to be translated with the resulting SMT system), the size and domain of the monolingual corpus available, and the language pair involved. In some of the experiments reported in Chapter 3 (Section 3.3.2), a phrase-based SMT needs more than one million words (more than 50 000 sentences) in each language to outperform the Apertium RBMT system (Forcada et al., 2011), despite the fact that the degree of translation quality provided by Apertium is below that offered by other RBMT systems (Forcada et al., 2011, Table 5).

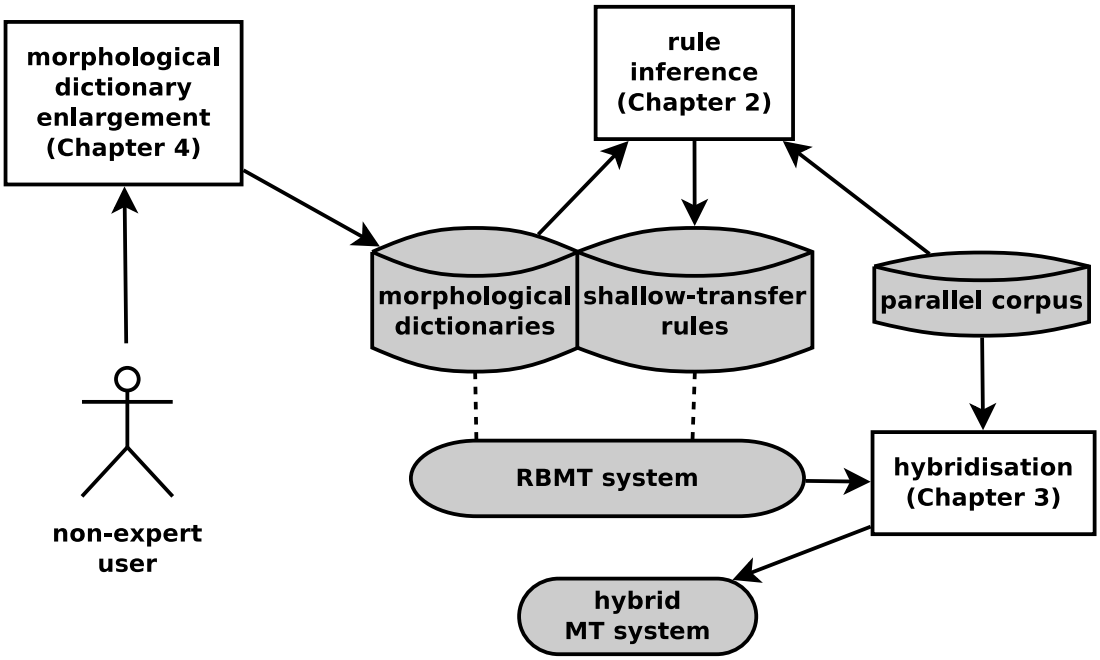
[9]Recently, some authors have proposed a different alternative for building MT systems when a parallel corpus is not available (Post et al., 2012; Zaidan and Callison-Burch, 2011): asking non-expert bilingual speakers to create it through a crowdsourcing platform (Wang et al., 2013) and building an SMT system from the parallel corpus. However, this approach presents some disadvantages when compared to the construction of RBMT resources. Namely, the amount of text that needs to be translated is huge due to the aforementioned data sparseness problem, and the parallel texts are not as easy to reuse for other language pairs or different NLP tasks as the RBMT linguistic data. When there is a small corpus available, character-based SMT models can be used to translate words that cannot be found in the training corpus (Vilar et al., 2007; Tiedemann, 2009). However, this option is only feasible when the languages involved are closely related.

lexical forms, that make up the SL IR. Similarly, the TL morphological dictionary contains mappings between TL lexical forms and TL surface forms, and it is used during the generation step (both of them are also called monolingual dictionaries). Bilingual dictionaries contain mappings between SL lexical forms and TL lexical forms. Shallow-transfer rules require a deeper knowledge of the languages involved, since they encode the operations to be performed when word-for-word translation is not correct, as it has been previously described in Section 1.1.1. Examples of the type of operations encoded in shallow-transfer rules can be found in the descriptions made by the developers of the linguistic data for some language pairs of the Apertium project (Tyers, 2010; Peradin and Tyers, 2012) and in Appendix A. Rules are not as easy to be reused between different systems as dictionaries, as shown by the fact that the Apertium systems for North Sámi–Norwegian (Trosterud and Unhammer, 2012), Icelandic–English (Brandt et al., 2011) and Breton–French (Tyers, 2010) were built by reusing dictionaries from different sources, but rules had to be developed from scratch. Additionally, transfer rules are hard to extend or modify because of the complex relations between them (Arnold et al., 1993, Sec. 4.2). When the matching criteria of a rule are modified, it may prevent other necessary rules from being applied.

1.2 Problems addressed: building machine translation systems for language pairs with scarce resources

As it has been pointed out in the previous section, the availability of resources is a limiting factor for the creation of new MT systems: parallel corpora big enough for a certain language pair or domain may not be available; and the human effort needed in order to build RBMT systems may be too high and bilingual people with enough level of expertise in the languages involved may be difficult to find for languages with a small number of speakers. In this dissertation, three new methods that ease the building of MT systems when resources (both parallel corpora and RBMT linguistic data) are scarce are described. In Chapter 2, a new method for inferring shallow-transfer rules from very small parallel corpora is presented, Chapter 3 describes a new approach for combining RBMT linguistic information with parallel corpora for building SMT systems, thus making the most of the available resources. In Chapter 4, a novel method that lowers the entry barrier for developing linguistic resources for MT is described: it allows average speakers of a language to insert entries into the morphological dictionaries of RBMT systems. This new approach eases the application of the methods for the automatic inference of shallow-transfer rules and hybridisation described in chapters 2 and 4 respectively, since they make use of morphological dictionaries. Figure 1.2 summarises the relationship between the three main approaches presented in this dissertation, the resources they use, and the results they produce.

Figure 1.2: Relationship between the three main approaches presented in this dissertation, represented as boxes with a white background, and the resources they use and the results they produce, represented as the elements with a darker background. The method presented in Chapter 4 produces morphological dictionary entries with the help of non-expert users, while the approach discussed in Chapter 2 produces shallow-transfer rules from a small parallel corpus and the morphological dictionaries of an RBMT system. The hybridisation approach described in Chapter 3 produces a hybrid MT system from a parallel corpus and an RBMT system. The shallow-transfer rules and morphological dictionaries of the RBMT system can be respectively obtained with the methods described in chapters 2 and 4.



1.2.1 Automatic inference of shallow-transfer rules

The first of the new approaches for easing the construction of MT systems faces the situation in which the size of the parallel corpus that is available for a particular language pair is not sufficiently large to train a competitive SMT system (recall that, usually, millions of words in each language are needed), being RBMT the only way to go. However, the cost and slow development cycles of rule-based MT systems may also pose a strong limitation. As it has been previously stated, shallow-transfer rules encode the information needed in order to deal with the grammatical divergences between languages, and usually they can only be developed by trained linguists. On the contrary, dictionaries require a less expert knowledge and are easier to reuse from other sources. On this basis, Chapter 2 presents a new approach with which to automatically learn shallow-transfer MT rules from very small parallel corpora (barely a few hundreds of sentences) and existing RBMT dictionaries. The inferred rules are compatible with the formalism used by Apertium (Forcada et al., 2011) to code shallow-transfer rules. They can be easily modified by human experts and they can co-exist with hand-crafted rules. In this manner, the new approach presented in Chapter 2 reduces the difficulty of creating an RBMT system by relieving the developers from creating the part that requires the deepest linguistic knowledge: the structural transfer rules. At the same time the inferred rules can be manually improved if trained linguists become available.

This new approach is inspired by the work by Sánchez-Martínez and Forcada (2009), uses alignment templates,[10] like those initially used in SMT (Och and Ney, 2004) and overcomes the main limitations of their method (see Section 2.2.1 for a thorough description of these limitations). Namely, the approach by Sánchez-Martínez and Forcada (2009) uses a rule formalism with low expressiveness that prevents the rules obtained from performing a strong generalisation over the linguistic phenomena observed in the training corpus; and generates rules that often prevent the application of other, more convenient rules. The approach in Chapter 2 is able to achieve a higher degree of generalisation over the linguistic phenomena observed in the training corpus than the approach by Sánchez-Martínez and Forcada (2009) and it is able to select the proper subset of rules which ensure the most appropriate segmentation of the input sentences to be translated. It overcomes these limitations thanks to the fact that it formalises the rule learning problem as a minimisation problem and treats conflicts between rules at a global level. It is the first rule inference approach that addresses rule inference in that way.

The experiments performed show that the new method generally outperforms the approach by Sánchez-Martínez and Forcada (2009) and that the resulting number of rules is considerably smaller, which eases human revision and maintenance. When the languages involved in the translation are closely related (e.g. Spanish↔Catalan), a

[10] Alignment templates are a generalised version of the phrase pairs used in phrase-based SMT (Koehn, 2010) which use word classes rather than words and also include word alignment information.

few hundred parallel sentence have proved to be sufficient to obtain a set of competitive transfer rules. Experiments also show that, for slightly bigger corpora (a few tens of thousands of sentences; hundreds of thousands of words), the new approach reaches, and in some cases surpasses, the translation quality achieved by hand-crafted rules. The adoption of this new rule inference approach will make it possible to create RBMT shallow-transfer rules even when expert developers who know all the grammatical divergences between the languages involved are not available, thus making the development of the linguistic data for new language pairs in MT systems like Apertium a much more cost-effective process.

1.2.2 Integration of shallow-transfer rules into phrase-based statistical machine translation

Even when a parallel corpus that is bigger than those used for rule inference is available —corpora used for rule inference contain a few hundreds or thousands of sentences in each language, as already said in the previous section— and a monolingual corpus from which to learn a TL model is available too, the resulting phrase-based SMT system may still present some limitations:

- The aforementioned data sparseness problem: collecting enough phrase pairs for covering all the inflected word forms in context for highly inflected languages would require a massive amount of parallel corpora, which may not exist for many language pairs.
- The available parallel corpus may belong to a different domain from that of the texts that will be translated with the resulting system, thus degrading the quality of the translations obtained.

One potential solution for these problems is *hybridisation*: if an RBMT system is also available, it can be combined with the SMT system in order to mitigate these limitations.[11] In Chapter 3, a new hybridisation strategy that consists of the insertion of linguistic information from a shallow-transfer RBMT system into the phrase table of a phrase-based SMT system is presented.

Since when a word is inserted into an RBMT dictionary, all its inflected word forms are included,[12] the use of RBMT dictionaries allows the SMT system to translate many of the words found in input texts (obviously, this fact depends on the effort invested in building the dictionaries). Moreover, shallow-transfer rules allow the SMT decoder to choose phrase pairs that go beyond the word-for-word translation of words from the RBMT dictionaries. In addition, the data from a general-purpose RBMT system can

[11] In order to mitigate the second limitation, the dictionaries of the RBMT system must cover the domain of the texts to be translated with the MT system.

[12] This is possible thanks to the use of inflection paradigms, described in Section 1.2.3.

help to reduce the bias of an SMT system trained on domain-specific corpora if it is to be used in a general domain.

Even if the rules from the RBMT system have not yet been created, they can be automatically inferred from the same parallel corpus from which the SMT system is trained with the help of the rule inference approach presented in Chapter 2. This strategy makes a better use of a parallel corpus and RBMT dictionaries than existing approaches (Schwenk et al., 2009), in which the dictionaries are simply added to the phrase table. By combining the rule inference algorithm with the hybridisation approach, the resulting SMT system is able to generalise the translation knowledge contained in the parallel corpus to sequences of words that have not been observed in the corpus, but share lexical category or morphological inflection information with the words observed. The combination of the automatic inference of shallow-transfer rules and the new hybridisation strategy is also discussed in Chapter 3.

Costa-Jussà and Fonollosa (2015) classifies hybrid MT architectures in those that integrate RBMT elements in an SMT system and those that integrate SMT elements in RBMT systems. The former strategy has been followed in the new hybrid approach presented in Chapter 3 because of the following reasons:

- A hybrid system driven by SMT can take full advantage of the TL model in order to improve the fluency of the output.
- In shallow-transfer RBMT, the SL sentence, after being analysed, is split into *chunks*. Each chunk is translated by a shallow-transfer rule and the results are concatenated in order to build the TL sentence. This process is similar to the process carried out by an SMT decoder, that builds translation hypotheses by segmenting SL sentence into phrases and translating each SL phrase according to the phrase table. The fact that both systems work with flat subsegments of the sentence eases the integration of chunks from RBMT into the SMT phrase table. The RBMT chunks can thus be scored by all the feature functions used by the phrase-based SMT system.
- When an RBMT system is combined with a phrase-based SMT system in a hybrid architecture, in addition to mitigating the aforementioned data sparseness problem, some other advantages over a pure SMT system can be achieved. For instance, the ordering of the elements of the sentence can also be improved when SMT elements are integrated in an RBMT architecture and the RBMT system performs a full syntactic analysis, as happens in the hybrid approach by Labaka et al. (2014). However, shallow-transfer RBMT systems do not perform such an analysis, so it is not worth following the option of integrating SMT elements in an RBMT architecture in the approach presented in Chapter 3.

The approach presented in Chapter 3 is the first hybrid approach specifically designed to integrate linguistic data from shallow-transfer RBMT into a phrase-based

SMT system. Besides this novel approach, the only existing strategy for enriching an SMT system with RBMT resources that can be applied to shallow-transfer RBMT (in fact, it can be applied to any MT system) is that by Eisele et al. (2008). It involves integrating the output of an RBMT system into a phrase-based SMT system without using information from the inner workings of the RBMT system: the sentences to be translated with the hybrid system are first translated with the RBMT system and a new phrase table is built from the resulting parallel corpus. It is afterwards appended to the phrase table obtained from the training parallel corpus. However, their method presents some limitations that are overcome by the approach in Chapter 3, which does take advantage of the way in which the RBMT system performs the translation of the SL sentences. The main limitations of the method by Eisele et al. (2008) are the following: the insertion in the phrase table of the SMT system of phrase pairs extracted from the RBMT data that are not mutual translation as a consequence of the bad quality of the word alignments in the parallel corpus obtained from the RBMT system, and the inadequate balance between the probabilities of phrase pairs extracted from the parallel corpus and those extracted from the RBMT data.

The experiments reported in Chapter 3 confirm that the new hybrid approach outperforms the strategy developed by Eisele et al. (2008). Moreover, they also show that when the hybrid system is built with automatically inferred rules, it is able to reach the translation quality that would be achieved by a hybrid system built with hand-crafted rules. This hybridisation approach, together with the aforementioned rule inference algorithm, will contribute to alleviate the data sparseness problem suffered by SMT systems when highly inflectional languages are involved and reduce the corpora size requirements for building SMT systems.

A system built with the hybridisation approach described in Chapter 3 (Sánchez-Cartagena et al., 2011c) that used hand-crafted rules from the Apertium project (Forcada et al., 2011) was one of the winners[13] in the pairwise manual evaluation of the WMT 2011 shared translation task (Callison-Burch et al., 2011) for the Spanish–English language pair.

1.2.3 Insertion of entries in morphological dictionaries by non-expert users

As it has been mentioned above, building an RBMT system from scratch requires a great investment in skilled labour. Finding a person who possesses all these skills can be difficult, especially for less-resourced languages. Although transfer rules are the component that requires the deepest linguistic knowledge, morphological dictionaries end up consuming most of the development time (Tyers, 2010) if they cannot be reused from other systems. Moreover, the methods for automatic shallow-transfer rule inference and hybridisation between RBMT and SMT outlined above rely on morphological

[13]No other system was found statistically significantly better using the sign test at $p \leq 0.10$.

dictionaries. As a word needs to be properly analysed before it can match a rule, the higher the coverage of the dictionaries, the higher the impact the rule inference and hybridisation approaches will have.

Given the amount of time usually invested in the creation of the morphological dictionaries of an RBMT system, methods that ease the contribution of a broad group of non-expert users can significantly speed up their creation and reduce the development costs. Chapter 4 presents a novel method that eases the enlargement of monolingual morphological dictionaries. The scenario considered is that of non-expert users who have to introduce into the monolingual dictionaries of an RBMT system the words found in an input text that are unknown to the system (not present in the SL monolingual morphological dictionary), so that it can subsequently correctly translate them. The non-expert users to which this method is addressed are average speakers who may not know the difference between, for example, an adjective and a noun, and who will not be required to learn any of the aspects relating to the encoding of the entries in the dictionary (see Section A.3.1 for an example of an entry encoded in an Apertium morphological dictionariy). If the user is bilingual and provides also a translation of the unknown SL word, the method can also be used to insert it in the TL morphological dictionary and, once both monolingual entries are inserted, the corresponding bilingual dictionary entry can be inserted without further human intervention. The whole system works under the assumption that average speakers of a language can correctly answer the polar question "*is the surface form X a valid inflected word form of word W?*" or, worded in a more plain language, "*is X a valid form of word W?*". As will be explained in Section 1.3.3, other authors (Font-Llitjós, 2007; McShane et al., 2002) have already proved that non-expert users can be really helpful to reduce the costs of building an MT system.

Let us illustrate the interaction between the user who wants to insert an entry and the approach described in Chapter 4 with an example. Suppose that the user is translating the English sentence *Many of those policies remain largely unimplemented* into Spanish with an MT system, and its SL monolingual dictionary does not contain an analysis for *policies*. The system would ask the user to validate whether *policy* and *policying* (the criterion followed for choosing these words will be explained in Chapter 4) are correct forms of *policies*. If the user is bilingual (she may want to postedit the translations and thus use the MT system for dissemination), she would state that *medidas* is the translation into Spanish of *policies*, and she would be asked to validate whether *medida* or *medidaba* are correct forms of *medidas*.

The objective of this approach is to obtain a system which can be used not only to add a particular unknown surface form (for example, *policies*) to the dictionary, but also to assist in discovering an appropriate *stem* and a suitable *inflection paradigm* so that all the inflected word forms of the unknown word and their associated morphological inflection information (such as *policies*: *policy* `NOUN-number:plural` or *policy*: *policy* `NOUN-number:singular`) can be inserted as well in one go. The stem is the part of a word that is common to all its inflected word forms. Inflection paradigms are commonly

used in RBMT systems in order to group regularities in the inflection of a set of words;[14] a paradigm is usually defined as a collection of suffixes and their corresponding morphological information; e.g., the paradigm assigned to many common English verbs indicates that by adding the suffix *-ing* to the stem, the gerund is obtained; by adding the suffix *-ed*, the past is obtained; etc. The approach presented in Chapter 4 assumes that the paradigms for all possible words in the language are already included in the dictionary.[15]

Users are expected to be motivated to contribute to the task if the system is deployed in an online MT engine because they will notice an immediate improvement in the MT system just by answering a few polar questions. This minimal interaction with the system and the involvement of regular speakers make the approach specially suitable for less-resourced languages. Moreover, a rise in the number of potential contributors can be expected due the constant growth in the number of users of online MT systems (Gaspari and Hutchins, 2007). This novel approach could also be applied in other scenarios. For instance, other non-expert individuals, who are not users of the MT system, can be recruited through crowdsourcing platforms (Wang et al., 2013) to collaboratively perform the task of inserting new entries in morphological dictionaries. Moreover, linguists themselves can also benefit from the approach because validating inflected word forms of words is usually faster than choosing among a list of paradigms.[16]

The experiments performed for the Spanish language with real users showed that they were able to insert entries in the dictionary with a high success rate and that the approach is very efficient: only 5–6 questions on average were needed in order to insert a set of words selected from the revision history of a real Apertium monolingual dictionary. This approach will thus contribute to save costs in the development of new RBMT systems.

1.3 Related approaches

In this section, the most relevant works in the literature related to the three new approaches presented in chapters 2, 3 and 4 are described and their differences with these novel approaches are highlighted.

[14] Paradigms ease dictionary management by reducing the quantity of information that needs to be stored, and by simplifying revision and validation because of the explicit encoding of regularities in the dictionary.

[15] Automatic acquisition of paradigms from monolingual corpora has already been explored (Monson, 2009), but this task is out of the scope of this work.

[16] A Spanish morphological dictionary may contain more than 500 inflection paradigms.

1.3.1 Automatic inference of shallow-transfer rules

Chapter 2 presents a novel approach for the automatic inference of shallow-transfer rules from scarce parallel corpora. The most relevant rule inference approaches related to it can be classified according to whether the rules are designed to be eventually used in an RBMT system (these approaches are discussed in Section 1.3.1.1) or in a corpus-based one (discussed in Section 1.3.1.2).

1.3.1.1 Rule-based machine translation

There have been other attempts to automatically learn structural transfer rules for RBMT. As it has been pointed out previously, the most similar related approach is that by Sánchez-Martínez and Forcada (2009). It is also aimed at learning shallow-transfer rules from a small parallel corpus and the morphological dictionaries of an RBMT system, but the learning algorithm and rule formalism employed are different from those used by the new approach presented in Chapter 2. The rule formalism defined by Sánchez-Martínez and Forcada (2009) is less expressive: it is not able to encode rules that are applied regardless of the morphological inflection information attributes of the words they match. This often results in having to learn several rules in order to describe the translation of the same linguistic phenomenon.[17] The inference algorithm is also very different. The new approach in Chapter 2 formalises rule inference as a minimisation problem in order to properly solve conflicts between rules and avoid the potential overgeneralisation that can arise because of the use of a more powerful formalism. Instead, Sánchez-Martínez and Forcada (2009) simply choose the most frequent rule when dealing with conflicts. Finally, the method by Sánchez-Martínez and Forcada (2009) generates rules that usually prevent the application of other, more convenient rules, when they are used in the Apertium RBMT platform, while the approach in Chapter 2 explicitly takes into account the interaction between rules when the RBMT engine chooses which rules to apply and avoids the generation of rules that harm translation quality. A more detailed description of the approach by Sánchez-Martínez and Forcada (2009) can be found in Section 2.2.

Probst (2005) also developed a method with which to learn transfer rules from a reduced set of bilingual segments. These segments are obtained by asking a group of bilingual annotators to translate a controlled, parsed corpus containing examples of all the relevant grammatical structures in the SL. Word alignments are also provided by the bilingual annotators. The transfer rules learnt follow a hierarchical formalism similar to that used in the early METAL system (Hutchins and Somers, 1992).

[17] For instance, adjectives in English are placed before the noun, whereas in Spanish they are usually placed after the noun. In order to properly translate a noun followed by an adjective from Spanish to English, 4 different rules encoded in the formalism defined by Sánchez-Martínez and Forcada (2009) are needed: one for each possible combination of gender and number in Spanish.

The main differences between the approach by Probst (2005) and the new approach presented in Chapter 2 are the following. First, her method learns hierarchical syntactic rules, whereas the new method presented in this dissertation learns flat, shallow-transfer rules. Second, her method uses a corpus whose SL side needs to be parsed, whereas the new method does not use information about the syntactic constituents; in addition to this, the approach in Chapter 2 learns how to automatically segment the text into chunks for their translation. Third, the alignments between the words that her approach uses are human-annotated, whereas the approach in Chapter 2 obtains the alignments automatically through the use of statistical methods, which forces it to tolerate alignment errors, especially when the training corpus is very small. Fourth, the strategy applied in order to generalise bilingual phrase pairs to rules is clearly different. Her initial approach (Probst et al., 2002) consists of selecting the minimum set of rules which correctly translates the set of bilingual phrase pairs by following a greedy strategy based on merging pairs of rules, while the approach in Chapter 2 selects the minimum set of rules by using a global strategy based on integer linear programming that is able to find the optimal solution. In her latest approach (Probst, 2005), a two-step procedure is followed for the generalisation problem. Her system first learns the context-free backbone of the rules, that is, how the terminal symbols of the grammar (which represent lexical categories) are grouped together and with non-terminal symbols to generate other non-terminal symbols. Value and agreement constraints are then obtained. Rules initially contain only value constraints, i.e., they are only applied to words with the same morphological feature values as the examples from which the rules have been extracted. Agreement constraints, which replace value constraints and generalise the values of the morphological attributes in the learning examples, are then inferred by considering the frequency of the different values that each morphological inflection attribute happens to have in the examples used for learning. Font-Llitjós (2007) approached the automatic inference of the same kind of hierarchical rules from a completely different source of bilingual information: postedits performed by users of the MT system. In addition, the rule inference/refinement is performed incrementally.

Varga and Yokoyama (2009) also developed a method for the automatic inference of transfer rules from small parallel corpora. The differences with the method in Chapter 2 are similar to those that have been just described for the method by Probst (2005). Namely, the rules inferred by Varga and Yokoyama (2009) are also hierarchical syntactic rules that need a parsed parallel corpus in order to be obtained. However, since the TL side of the parallel corpus is not provided by bilingual translators, the alignments between words in both languages are not available, and an existing bilingual dictionary is used in order to obtain them. In the new method presented in this dissertation, they are obtained by statistical methods, although the bilingual dictionary is used to filter the results. Moreover, conflicts between rules are not treated by Varga and Yokoyama (2009), but rules are assigned a probability based on the number of examples from which they have been inferred. Thus, rules must be used in a hybrid RBMT with

statistical components, while the rules inferred by the approach in Chapter 2 can be used in an RBMT system in which a single sentence is obtained after the transfer step.

Caseli et al. (2006) present a method in which shallow-transfer rules and bilingual dictionaries are learnt from a parallel corpus. With regard to the shallow-transfer rule inference, these rules are learnt from a set of bilingual phrase pairs obtained after aligning the words in the SL and TL sentences by means of statistical methods in a way similar to that used by the approach in Chapter 2. After obtaining the bilingual phrase pairs, rules are inferred from them and those containing complementary information are joined in order to reduce their number. For instance, if rule a is applied to masculine nouns and rule b is applied to feminine nouns, a new rule c is created, which is applied to both masculine and feminine nouns. In a final step, conflicts between rules are avoided in a greedy fashion by specialising the rules, either by including more morphological inflection attributes as a condition for their application, or by lexicalising some of the lexical forms they match. If a rule cannot be further specialised, the most frequent one is retained.

The approach by Caseli et al. (2006) principally differs from that presented in this dissertation as regards the way in which bilingual phrase pairs are generalised to obtain rules. On the one hand, their approach does not generalise unseen linguistic features, that is, if a rule is learnt from bilingual phrase pairs containing only masculine nouns, it will never be applied to feminine nouns. The method in Chapter 2, meanwhile, generates rules that generalise morphological inflection values not seen in the training set thanks to its more expressive formalism, provided that these rules are able to correctly reproduce the bilingual phrase pairs from which they are learnt. This is a great advantage when the size of the training corpus is very small. On the other hand, the new minimisation approach considers all the possible alternatives when dealing with conflicts between rules matching the same sequence of lexical categories, rather than doing so in a greedy manner. With regard to the way in which the rules learnt affect the segmentation into chunks of the SL sentences to be translated, Caseli et al. (2006) do not confront the problem that long rules may prevent the application of shorter and more accurate ones; they merely select the sequences of SL and TL lexical categories for which rules will be generated based on their frequencies in the parallel corpus. Finally, the approach by Caseli et al. (2006) is designed to infer shallow-transfer rules and bilingual dictionaries at the same time, and it cannot infer rules to be used together with existing, high-quality, bilingual dictionaries. The approach in Chapter 2, on the contrary, works with an existing bilingual dictionary.

1.3.1.2 Corpus-based machine translation approaches

There have also been attempts to learn linguistic resources which are not used in RBMT, but are in fact similar to structural transfer rules. For instance, in the

example-based MT (EBMT) framework (Carl and Way, 2003),[18] some researchers have dealt with the problem of inferring a kind of translation rules called *translation templates* (Kaji et al., 1992; Brown, 1999; Carl, 1999; Cicekli and Güvenir, 2001). A translation template can be defined as a bilingual pair of sentences in which corresponding units (words or phrases) are coupled and replaced with variables. Liu and Zong (2004) summarise different translation template acquisition methods. Other approaches with which to learn structural transformations in the EBMT framework include, among others, the acquisition of transfer mappings from bilingual corpora (Menezes and Richardson, 2003) and the induction of probabilistic translation grammars from syntactically-parsed parallel sentences (Carl, 2001).

There are multiple differences between the new approach described in Chapter 2 and those applied in the EBMT framework. For instance, EBMT translation templates can be hierarchical and consequently more than one translation template may be applied to a given SL segment, while shallow-transfer rules are flat. In addition, the way in which transfer rules are usually applied in shallow-transfer RBMT and translation templates are applied in EBMT are also different: whereas in shallow-transfer RBMT rules are applied in a left-to-right, longest match, greedy manner and an SL lexical form can only be processed by a single transfer rule, in EBMT different translation templates, not necessarily nested ones, can match the same SL word, i.e. EBMT allows the overlapped matching of translation templates. Another distinguishing feature is that the approach in Chapter 2 is mainly based on lexical forms consisting of lemma, lexical category and morphological inflection information. Most of the approaches based on EBMT translation templates, on the contrary, use variables which, even though they may be linguistically motivated (e.g. NOUN, VERB, NP, PP), do not include lower level morphological inflection attributes (e.g. gender, number, person, case) whose values can be obtained through references to the matching segment in the SL sentence or to a bilingual dictionary.

Finally, in the SMT framework, the use of alignment templates (Och, 2002; Och and Ney, 2004) can be seen as an integration of translation rules into statistical translation models, since an alignment template is a generalisation or an abstraction of the transformations to be applied when translating SL into TL by using word classes. Hierarchical SMT systems (Chiang, 2007), in which hierarchical statistical translation rules are learnt from parallel corpora, are also moderately similar to the approach in Chapter 2, particularly when the rules have many different non-terminal symbols (Zollmann and Vogel, 2011). The differences are again that the shallow-transfer rules in the approach in Chapter 2 are flat, less structured and non-hierarchical. In addition, the application of shallow-transfer rules is not statistically driven.

[18] A corpus-based approach to MT.

1.3.2 Integration of shallow-transfer rules into phrase-based statistical machine translation

Chapter 3 presents a new hybrid approach in which linguistic resources from shallow-transfer RBMT are integrated into a phrase-based SMT system in order to overcome, among other limitations, the data sparseness problem usually suffered by SMT systems when one of the languages involved is highly inflected. The related approaches can be split into those that integrate RBMT elements in an SMT system and those that follow a different hybrid approach (either the integration of SMT elements in the RBMT architecture or the combination of the outputs of the systems involved without changing their architecture). Approaches in the first group can in turn be split into two groups according to whether the linguistic information is directly extracted from an existing RBMT system (Section 1.3.2.1) or inferred from the training parallel corpus from which the SMT models are estimated (Section 1.3.2.2). Approaches in the second group are summarised in Section 1.3.2.3.

1.3.2.1 Integrating hand-crafted linguistic knowledge into statistical machine translation

Concerning the approaches in which elements from existing RBMT systems are integrated in phrase-based SMT, they can be split according to the type of RBMT information used: in some approaches, only the RBMT dictionaries are integrated into an SMT system while other approaches take advantage of all the RBMT linguistic data, including transfer rules.

Bilingual dictionaries are the most reused resource from RBMT; they have been added to SMT systems since its early days (Brown et al., 1993a). One of the simplest strategies, which has already been put into practice with the Apertium bilingual dictionaries (Tyers, 2009), consists of adding the dictionary entries directly to the training parallel corpus. In addition to the obvious increase in lexical coverage, Schwenk et al. (2009) state that the quality of the alignments obtained is also improved when the words in the bilingual dictionary appear in other sentences of the parallel corpus. However, it is not guaranteed that, following this strategy, multi-word expressions from the bilingual dictionary that appear in the SL sentences are translated as such because they may be split into smaller units by the phrase-extraction algorithm. Dictionaries have also been added to SMT systems together with other rule-based enhancements, as in the work by Popovic and Ney (2006), who propose combining dictionaries with the use of hand-crafted rules to reorder the SL sentences to match the structure of the TL.

With regard to approaches that take advantage of the full RBMT system, Eisele et al. (2008) present a strategy based on the augmentation of the phrase table to include information provided by an RBMT system, as it has been advanced in Section 1.2.2. In

their approach, the sentences to be translated by the hybrid system are first translated with an RBMT system and a small phrase table is obtained from the resulting parallel corpus. Phrase pairs are extracted following the usual procedure in SMT (Koehn, 2010, sec. 5.2.3), which generates the set of all possible phrase pairs that are consistent with the word alignments. Word alignments are computed using an alignment model previously built from a large parallel corpus. Finally, the RBMT-generated phrase table is directly added to the original one. The new approach presented in this dissertation improves two aspects of the hybridisation strategy by Eisele et al. (2008), briefly outlined in Section 1.2.2:

- It ensures that the phrase pairs obtained from the shallow-transfer RBMT system are mutual translations by taking advantage of how the RBMT system uses dictionaries and shallow-transfer rules to segment the input. In contrast, the approach by Eisele et al. (2008) relies on word alignment models that may be inaccurate if the large parallel corpus does not share the domain with the text to be translated.

- It takes into account two factors not considered by Eisele et al. (2008) when computing the scores of the phrase pairs generated from the RBMT system. Firstly, phrase pairs whose SL phrase is usually translated in the same way in the parallel corpus and by the RBMT system should receive higher scores than those that are translated differently. Secondly, the frequency of an SL phrase in the parallel corpus should influence the reliability of the phrase pairs that contain it when they are in conflict with the phrase pairs generated from the RBMT system: the higher the frequency of an SL phrase in the training parallel corpus, the higher the reliability of the phrase pairs that contain it.

Another interesting approach is that by Enache et al. (2012), in which an interlingua RBMT system developed for the limited domain of patent translation is integrated into a phrase-based SMT architecture. Synthetic phrase pairs are generated from chunks extracted from the SL sentences that can be parsed by the RBMT system.[19] The same philosophy is behind the new hybrid approach presented in Chapter 3, in which synthetic phrase pairs are generated from the chunks matched by shallow-transfer rules. However, significant differences exist in the method used for scoring the phrase pairs generated from the RBMT system. In the method by Enache et al. (2012), a single value, defined in advance, is assigned to the source-to-target and target-to-source phrase translation probabilities and the source-to-target and target-to-source lexical weightings of all the synthetic phrase pairs, i.e. all the synthetic phrase pairs are equiprobable. Consequently, the relative weight of the synthetic phrase pairs is not optimised in the tuning step of the SMT training process. In the new method presented in Chapter 3, however, a more sophisticated scoring scheme is followed. The relative

[19] A parse tree may not be obtained from the sentences that do not follow usual structure in the restricted domain. This happens for 66.7% of the sentences in their test set.

weight of the synthetic phrase pairs is optimised during the tuning process, phrases translated in the same way in the parallel corpus and by the RBMT system receive higher scores and the lexical weightings of the synthetic phrase pairs are computed based on the same principles as in SMT: taking into account the translations of the individual words that make up the phrases.

Finally, Rosa et al. (2012) created a set of rules that are applied to the output of an SMT system in order to fix its most common errors. The main difference with the approach in Chapter 3 lies in the fact that, although these rules are similar to transfer rules, they operate only on the TL side and that a syntactic analysis is performed before applying them.

1.3.2.2 Adding morphological information to statistical machine translation

The hybridisation approach in Chapter 3 can be combined with the rule inference approach described in Chapter 2 in order to integrate a set of structural transfer rules inferred from the very same training parallel corpus used in the SMT system, thereby extending the SMT models with new linguistic information. Since shallow-transfer rules operate on lexical forms made of lemma, lexical category and morphological inflection information, the combination of the approaches presented in chapters 2 and 3 can be seen as a novel way of extending phrase-based SMT, which commonly works with surface forms, with morphological features.

In this manner, the approach presented in this dissertation is related to factored translation models (Koehn and Hoang, 2007). They are an extension of phrase-based SMT in which each word is replaced by a set of factors that represent lemma, lexical category, morphological inflection information, etc. A phrase-based translation model is inferred for lemmas and an independent one for lexical categories and morphology. A word-based generation model, that can be inferred from additional TL monolingual data, maps combinations of TL lemmas, lexical category and morphological inflection information to TL inflected word forms. These are the main differences between the factored models and the new hybrid approach presented in this dissertation:

- In factored models, the translation of lemmas and morphological information is completely independent. Since both types of translations are combined in order to generate the final sequence of surface forms, a combinatorial explosion is likely to be produced (too many combinations of lemmas and morphological information need to be scored by the generation model). As all the options cannot be explored, correct translation hypotheses may be pruned (Bojar and Hajič, 2008; Graham and van Genabith, 2010). Moreover, idiomatic translations that do not follow the general morphological rules of the TL may be assigned a very low probability by the translation model, even though they would have a high probability in a phrase table built from surface forms. That strategy differs from that

followed by the hybrid system resulting from the combination of the approaches presented in chapters 2 and 3, in which translation hypotheses are built from the surface-form-based models usually employed in phrase-based SMT. Thus, the probability distribution of the sequences of surface forms in the training corpus is taken into account. The complexity of dealing with translations of lemmas and morphological inflection information is moved from decoding time to training time, when the rule inference algorithm from Chapter 2 deals with it. Note also that Graham and van Genabith (2010) proposed an strategy for partially mitigating the issues caused by the fact that factored models treat lemmas and morphological information as totally independent elements: the extraction from the training parallel corpus of factored templates, which are phrases that will not be decomposed in lemma and morphological information for translation.

- The new hybrid approach works with existing bilingual dictionaries, while factored models do not use bilingual dictionaries at all. As a consequence of the use of bilingual dictionaries, the way in which morphological inflection information is translated from the SL into the TL is different in both approaches. In factored models, the probability of the morphological inflection factors of a TL translation hypothesis depends solely on the morphological inflection factors of the SL sentence. In the transfer rules used by the hybrid approach described in this dissertation, however, the morphological inflection attributes of TL words can be obtained either from SL words or from their translation according to the bilingual dictionary. This fact makes the formalism more expressive and eases the treatment of certain linguistic phenomena. Consider, for instance, the case in which there is a morphological inflection attribute that only exists in the TL (such as gender when translating from English to Spanish or French). In the hybrid approach presented in this dissertation, the structural transfer rule for gender and number agreement between a noun and an adjective would assign the gender of the translation into the TL according to the bilingual dictionary of the SL noun to the TL noun and the TL adjective. This type of rule can be inferred from a very small parallel corpus. In factored models, however, the translation model would presumably assign similar probabilities to TL noun-adjective sequences with both genders, and the success of the agreement would depend on whether the TL model has enough information to properly score the different candidate sequences of surface forms in the TL.

Other relevant approaches in which morphological attributes are integrated into the translation model of an SMT system can be found in the literature. Green and DeNero (2012), for instance, defined a new feature function that models agreement (e.g. nouns, adjectives and determiners in the same noun phrase must agree in gender and number), while the factored language models (Kirchhoff and Yang, 2005) assign probabilities to TL sentences depending on their sequences of word forms and morphological features, among other factors. These approaches differ from the new strategy presented in this dissertation mainly in the fact that they do not perform a generalisation that enriches

the translation model with translations of sequences of SL words unseen in the training corpus.

Riezler and Maxwell III (2006) went further and also added syntactic information to SMT. They developed a hybrid RBMT-SMT system which works as follows. The SL sentence is parsed with a lexical functional grammar (Riezler et al., 2002) in order to obtain an SL IR. Then the SL IR is transferred into the TL IR by applying a set of probabilistic rules obtained from a parallel corpus. Each rule contains a set of scores inspired by those present in the phrase table of a phrase-based SMT system. Finally, the TL sentence is generated from the TL IR. Since an SL sentence can be parsed in many different ways and many different TL IRs can be generated by applying different rules, a TL model is also used in addition to the aforementioned phrase-table-like features. The approach by Riezler and Maxwell III (2006) combines the different feature functions by means of a log-linear model, and their weights are optimised by means of minimum error rate training (Och, 2003) as in SMT. The results showed that the grammar used was not able to completely parse half of the sentences of the test set used (partial parse trees were obtained from these sentences, but their resulting translation was much worse than the translation of fully parsed sentences), and considering only the sentences that could be fully parsed, there was not a statistically significant improvement over a phrase-based SMT system trained from the same data. However, a human evaluation showed an improvement on the grammaticality of the translations. The main differences with the new approach presented in Chapter 3 are the following: first, the approach by Riezler and Maxwell III (2006) does not use existing bilingual dictionaries; second, it uses syntactic information, that allows the system to perform a deeper linguistic analysis at the expense of not being able to fully parse some input sentences (resulting in a translation performance drop), while the approach described in this dissertation works with lexical categories and morphological inflection information and is more robust against ungrammatical input.

1.3.2.3 Other hybrid architectures

In addition to the hybrid approaches that integrate elements from RBMT into the SMT architecture, there are hybrid approaches that integrate statistical elements in the RBMT architecture, and strategies in which the architecture of the systems combined is not changed, but their outputs are simply combined.

Regarding the enhancement of an RBMT architecture with statistical elements, it is worth noting that RBMT systems often use statistical methods for part-of-speech tagging (Cutting et al., 1992) and parsing (Federmann and Hunsicker, 2011). Besides these components, other elements from SMT have been integrated into RBMT causing deeper changes in the RBMT architecture. For instance, in the transfer step multiple hypotheses can be generated, and then the most probable one may be chosen according to a TL model (Lavie, 2008; Carl, 2007). Another option is using phrase pairs instead of transfer rules in the transfer step, but keep using the RBMT analysis and generation

modules (Crego, 2014). The approach by Riezler and Maxwell III (2006), discussed previously, also uses a TL model in order to choose among translations generated by applying rules, but it integrates more elements from SMT, such as the feature functions usually encoded in an SMT phrase table.

Another option consists of taking advantage of the full syntactic analysis performed by the RBMT system in order to create the structure of the TL sentence, and then insert phrase translations obtained from a phrase table such as those used by SMT systems in some nodes of the parse tree that acts as the TL IR (Labaka et al., 2014). As in SMT, the final translation is that with the maximum probability according to a TL model and to the phrase table from which the phrases inserted in the tree have been obtained. However, phrase reordering is not allowed, since the structure of the TL sentences is guided by the parse tree. This set-up has also been followed in systems proposed by other authors (Federmann et al., 2010; Zbib et al., 2012). The automatic inference of transfer rules (discussed in the previous section) or bilingual dictionaries (Eberle, 2014, HyghTra project[20]) from parallel corpora can also be considered as a method for integrating statistical elements in RBMT.

Different MT systems can also be combined without changing their internal architecture (i.e. considering them as a *black box*) by simply combining the translations they produce. The simplest option, that is called *system selection*, consists of, for each sentence, choosing the best candidate from the translations provided by the different systems (Hildebrand and Vogel, 2008). The decision is based on different features, such as the probability of the candidate translations according to a TL model and the degree of agreement between the words and n-grams in the different translations. System selection can also be performed before translating the SL sentence with the different systems. In that case, only those features whose value can be obtained from the SL sentence itself are used (Sánchez-Martínez, 2011). A more sophisticated way of combining the translations provided by the different systems consists of building synthetic sentences by combining segments from the candidate translations. This approach is called *system combination*. Most of the system combination approaches that can be found in the literature are inspired by the concept of *confusion network*, already used in speech recognition (Fiscus, 1997). A confusion network is a directed graph in which nodes are organized as a sequence and arcs represent words and their probability. For each sentence, a confusion network is built from the different candidate translations and the path with maximum probability constitutes the final translation. Different approaches have been devised for building confusion networks from the outputs of a set of MT systems: Levenshtein distance (Bangalore et al., 2001), word alignment models (Matusov et al., 2006), MT evaluation metrics (Rosti et al., 2007), etc.

MT engines can also be combined in a serial way by following the statistical postedition architecture (Simard et al., 2007). In that set-up, the SL sentences to be translated by the hybrid system are first translated into the TL by an RBMT system, and the

[20] `http://www.hyghtra.eu/`

result is translated from the TL to a corrected version of the TL by means of an SMT system. The SL sentences of the training corpus from which the SMT system is built are obtained by translating the original SL sentences of a parallel corpus with the RBMT engine, while the TL sentences are those from the original parallel corpus. They can also be obtained by asking users to postedit the output of the RBMT system.

The main differences between the approaches that treat the systems to be combined as a black box just listed and the new method presented in Chapter 3 are the following. First, the approaches listed are designed to take advantage of the capacity of RBMT to perform long-distance reorderings (either by means of a confusion network or thanks to the translation performed by the RBMT system in the statistical postedition approach), but shallow-transfer RBMT systems, to which the new approach is addressed, are not able to perform such reorderings. Second, the approach in Chapter 3 treats the RBMT system as a white box and takes advantage of how dictionaries and rules are used in the transfer step in order to achieve a better integration with SMT. Third and last, the RBMT transfer rules in the approach in Chapter 3 can be automatically inferred by following the method described in Chapter 2.

1.3.3 Insertion of entries in morphological dictionaries by non-expert users

A novel approach that allows average speakers of a language to insert entries into RBMT morphological dictionaries is described in Chapter 4. The most relevant and related approaches can be split into three groups: strategies in which users are assisted in the insertion of new entries into the dictionaries as in the approach in Chapter 4 (outlined in Section 1.3.3.1); approaches in which the insertion of new entries is performed automatically, mainly with the help of monolingual corpora (described in Section 1.3.3.2); and approaches in which users are assisted in the creation of full MT systems (Section 1.3.3.3).

1.3.3.1 Human-assisted creation of morphological dictionaries

Existing approaches in which users are assisted in the addition of entries to morphological dictionaries are mainly addressed to experienced users. For instance, the *smart paradigms* devised by Détrez and Ranta (2012) help users to obtain the right inflection paradigm for a new word to be inserted into an RBMT morphological dictionary. A smart paradigm is a function that returns the most appropriate paradigm for a word given its lexical category, some of its word forms and, in some cases, some morphological inflection information. There are two important differences with the new approach presented in Chapter 4: firstly, smart paradigms are created exclusively by human experts; and secondly, users of smart paradigms need to have some linguistic background. For instance, an expert could decide that in order to correctly choose the

inflection paradigm of most verbs in French the infinitive and the first person plural present indicative forms are needed; dictionary developers must then provide these two forms when inserting a new verb. Bartůšková and Sedláček (2002) also present a tool for semi-automatic assignment of words to declension patterns; their system is based on a decision tree with a question in every node. Their proposal works only for nouns and is aimed at experts because of the technical nature of the questions. Moreover, decision trees must be manually created in advance, while the approach in Chapter 4 automatically creates the decision trees that guide the questions asked to non-expert users. The proposal by Kaufmann and Pfister (2010), which is also addressed to experts, helps users to insert new entries into a morphological dictionary by showing them the most probable stems and morphosyntactic features. Probabilities are obtained by applying machine learning techniques using a set of features derived from a monolingual corpus and the existing entries in the dictionary. Unlike that approach, the new strategy presented in Section 4 does not require the users to know what a stem is and it does not show them any probability.

1.3.3.2 Automatic creation of morphological dictionaries

As regards the automatic acquisition of morphological resources for MT, Lindén et al. (2009) propose choosing the most appropriate paradigm for a given word by simply counting the frequency in a monolingual corpus of the different word forms resulting from the inflection of the new word with each candidate paradigm. Cholakov and Van Noord (2009) follow an analogous approach in which frequencies are obtained after querying a web search engine. Forsberg et al. (2006) also developed a similar method in which some manual constraints are introduced in order to improve the precision of the results. Šnajder (2013) defines a more sophisticated approach: he poses the choice of the most appropriate paradigm for a given word as a machine learning problem. Given the values of a set of features extracted from a monolingual corpus and from the orthographic properties of the word forms generated by each inflection paradigm, each paradigm is classified as correct/incorrect by a *support vector machine* classifier (Suykens and Vandewalle, 1999). Morphological inflection paradigms, which are needed for inserting new entries in morphological dictionaries following either the approaches which have just been named or the approach in Chapter 4 can also be automatically inferred from monolingual corpora (Monson, 2009; Eskander et al., 2013).

A closely related problem to the automatic acquisition of morphological resources is that of *morphological guessing*, that consists of determining the most appropriate morphological analysis for an unknown word, without necessarily choosing an inflection paradigm for inserting it into a morphological dictionary. Morphological guessers often use rules that depend on the language and guess the morphological information on the basis of the suffix of the unknown word. Morphological guessers for many different languages, such as French (Chanod and Tapanainen, 1999), Czech (Hlaváčová, 2001) or German (Nakov et al., 2003), can be found in the literature. Other authors address the

analysis of unknown words from a syntactic point of view: their objective is determining the syntactic category of an unknown word in a sentence so that a full syntactic analysis of the sentence can be obtained. To that end, Thomforde and Steedman (2011) propose a method that determines the category of the unknown word on the basis of the analysis of the remaining words of the sentence and a model learned from fully parsed sentences.

1.3.3.3 Building full machine translation systems from non-expert users

Two of the most prominent works in literature in relation to the elicitation of knowledge to build or improve full RBMT systems are those by Font-Llitjós (2007) and McShane et al. (2002). The former, already mentioned in Section 1.3.1.1, proposes a strategy for improving both transfer rules and dictionaries by analysing the postediting process performed by a non-expert user through a dedicated interface. McShane et al. (2002) designed a framework to elicit linguistic knowledge from informants who are not trained linguists and used this information in order to build MT systems which translate into English; their system provides users with a lot of information about different linguistic phenomena to ease the elicitation task. The two aforementioned approaches are addressed to a transfer-based RBMT system in which multiple translations are generated during the transfer step and a TL model decides which is the most adequate translation. Conversely, the new method in Chapter 4 aims at easing the acquisition of morphological resources for RBMT systems that do not contain a TL model, and thus are notably more sensitive to erroneous linguistic information. The approach in Chapter 4 has been designed with the aim of relieving users from acquiring linguistic skills, so that the collaboration is easier.

Non-expert users can also help in the creation of SMT systems. For instance, Ambati et al. (2010) propose asking non-expert bilingual informants to translate SL sentences in order to create a parallel corpora from which an SMT is eventually built; users interact through a crowdsourcing platform (Wang et al., 2013). An efficient active learning strategy (Haffari et al., 2009) is critical in this scenario: the SL sentences to be translated by the users should be those that, when included in the parallel corpus, cause a big performance boost in the resulting SMT system.

Chapter 2

Inferring shallow-transfer rules from small parallel corpora

This chapter describes a new method that uses scarce parallel corpora (barely a few hundreds of parallel sentences) and existing morphological dictionaries to automatically infer a set of shallow-transfer rules to be integrated into a rule-based machine translation system. This new method avoids the need for human experts to handcraft these rules and overcomes many relevant limitations of previous approaches. Namely, it is able to achieve a higher degree of generalisation over the linguistic phenomena observed in the training corpus and it is able to select the proper subset of rules which ensure the most appropriate segmentation of the input sentences to be translated with them. The method presented in this chapter is the first approach in which conflicts between rules are resolved by choosing the most appropriate ones according to a global minimisation function rather than proceeding in a pairwise greedy fashion. Experiments conducted using five different language pairs with the free/open-source rule-based machine translation platform Apertium show that translation quality significantly improves when compared to previous approaches, and it can even surpass that obtained using hand-crafted rules. Moreover, the resulting number of rules is considerably small, which eases human revision and maintenance.

2.1 Introduction

As it has been pointed out in the introduction, although SMT is currently the leading MT paradigm, its application is limited by the availability of parallel corpora. When parallel corpora big enough to build a competitive SMT system is not available, RBMT systems are the only option, although the cost in terms of time spent by trained linguists of developing an RBMT system from scratch can be prohibitively high. Transfer rules constitute the resource from RBMT that requires the deepest linguistic knowledge in order to be created, since they encode the operations to be carried out in order to

deal with the grammatical divergences between the languages. In this chapter, a new approach with which to automatically learn shallow-transfer MT rules from very small parallel corpora is presented. The inferred rules are compatible with the formalism used by Apertium (Forcada et al., 2011) and they can be easily modified by human experts, as well as co-exist with hand-crafted rules. The new approach is inspired by the method by Sánchez-Martínez and Forcada (2009), uses an extension of the alignment template (AT) formalism (Och and Ney, 2004) to encode rules, and overcomes the main limitations of the method by Sánchez-Martínez and Forcada (2009) (see Section 2.2.1 for a thorough description of these limitations). The experiments carried out confirm that this new approach outperforms that by Sánchez-Martínez and Forcada (2009) and that the translation quality achieved with the automatically inferred rules is generally close to that obtained with hand-crafted rules. Moreover, for some language pairs, the automatically inferred rules are even able to outperform the hand-crafted ones.

The remainder of the chapter is organised as follows. The following section presents a brief description of the approach by Sánchez-Martínez and Forcada (2009), stressing its main limitations and summarising how they are overcome in this approach. Sections 2.3 and 2.4, respectively, introduce the AT formalism used in this approach and the method employed to extract rules encoded as ATs. The experiments conducted to test this new approach are presented in Section 2.5, whereas the results obtained are reported and discussed in Section 2.6. Finally, the chapter ends with some concluding remarks.

2.2 Previous approach

The approach by Sánchez-Martínez and Forcada (2009) is based on the AT approach (Och, 2002; Och and Ney, 2004) initially proposed in the context of SMT. An AT performs a generalisation of bilingual phrase pairs by using word classes rather than the words themselves. Sánchez-Martínez and Forcada (2009) adapted the AT approach for their application in RBMT by extending the ATs with a set of restrictions in order to control their application as shallow-transfer rules. An *extended* AT (henceforth, EAT) is defined as a tuple $z = (S, T, A, R)$, consisting of a sequence S of SL word classes, the corresponding sequence T of TL word classes, a set A of pairs of word class indexes (i, j) with the alignment information between the word classes in the two sequences, and a set R of restrictions over the TL inflection information that the words to translate need to meet. These restrictions prevent, for example, the application of an AT that produces a TL masculine noun from an SL feminine noun to an SL noun whose translation, according to the bilingual dictionary of the system, does not have a masculine gender.

The method by Sánchez-Martínez and Forcada (2009) needs a human-designed set of *lexicalised units*. This set is made up of both the SL and TL lexical forms (usually corresponding to closed lexical categories) involved in lexical changes and which should

not be generalised. EATs are then learnt from a parallel corpus by performing the following steps:

1. Analyse and convert both sides of the parallel corpus into the IR used by the RBMT system to be used; in the case of Apertium, sequences of lexical forms.

2. Apply classical statistical, word-translation models (Brown et al., 1993b; Vogel et al., 1996) in order to obtain word alignments in both translation directions, and then symmetrise the alignments obtained using the refined intersection method proposed by Och and Ney (2003).

3. Extract bilingual phrase pairs that are compatible with the set of alignments (Koehn, 2010, Sec. 5.2.3).

4. Remove those bilingual phrase pairs that cannot be reproduced by the RBMT system in which the transfer rules will be used because, according to the bilingual dictionary, the translation equivalent of at least one lexical form not present in the set of lexicalised units differs from that observed in the bilingual phrase.

5. Replace lexical forms with word classes. Word classes represent the lexical category and morphological inflection information of the corresponding words. They are obtained by removing the lemma from each lexical form that is not present in the set of lexicalised units provided by the user.

6. Infer the set of restrictions R by looking up in the bilingual dictionary the lexical forms that do not belong to the set of lexicalised units.

The resulting set of EATs is then used to generate structural shallow-transfer rules after removing those EATs whose frequency is below a threshold that is empirically determined on a development parallel corpus.

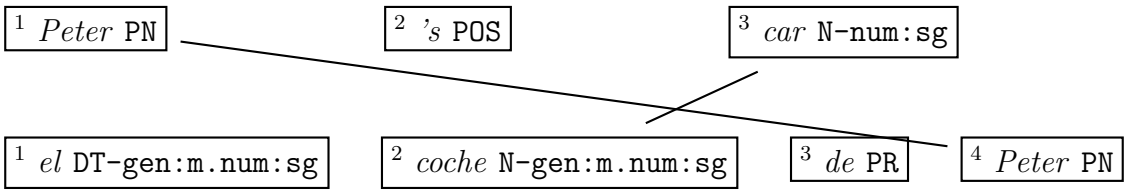
During translation, when an EAT matches a sequence of SL lexical forms, the actions that need to be performed in order to build each TL lexical form of the output depend on the type of word class:

- if the TL word class includes a lemma (e.g. *de* `PR`, the Spanish preposition *de*), because the corresponding lexical form belongs to the set of lexicalised units, it is introduced unchanged.

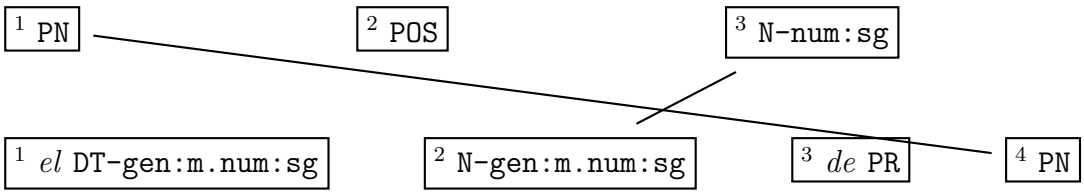
- if the TL word class does not include a lemma (e.g. `N-gen:m.num:sg`; noun, masculine, singular), the lemma to be included in the TL lexical form is obtained by looking up in the bilingual dictionary the SL lexical form that is matched by the SL word class to which the TL word class is aligned. The TL lemma is then accompanied by the morphological inflection attributes in the TL word class.

Figure 2.1: English–Spanish bilingual phrase pair p and EAT z obtained with the method devised by Sánchez-Martínez and Forcada (2009). To obtain z, the lexical forms in p are replaced with word classes. These word classes are obtained by removing the lemma from the lexical forms, with the exception of those in the set of lexicalised units provided by the user (in the example, prepositions and determiners). Restrictions r_1 and r_2 are empty, whereas r_3 forces the EAT to be applied only to those SL nouns that are masculine in the TL. `PN`, `POS`, `N`, `DT` and `PR` stand for proper noun, possessive ending, noun, determiner and preposition, respectively. `gen:m` indicates that the gender of the word is masculine, and `num:sg` that its number is singular. Lines between word classes or lexical forms represent alignments; lemmas appear in italics. With this EAT, the translation into Spanish of the English phrase *Fran's pen* would be *el bolígrafo de Fran*.

Bilingual phrase pair p:



EAT z:



$r_1 = \{\}; r_2 = \{\}; r_3 = \{\texttt{gen:m}\}$

Figure 2.1 shows an English–Spanish bilingual phrase pair and the EAT obtained from it. This EAT matches any proper noun (`PN`) followed by a possessive ending (`POS`), and a singular (`num:sg`) noun (`N`). As an output, this EAT produces a masculine (`gen:m`) singular determiner (`DT`) with the lemma *el*, a masculine singular noun whose lemma is obtained by looking up in the bilingual dictionary the lemma of the noun that is matched in the SL, a preposition (`PR`) with the lemma *de*, and a proper noun whose lemma is retrieved from the bilingual dictionary by looking up the proper noun that is matched in the SL. Restriction r_3 prevents the EAT from being applied when the noun is not masculine in the TL, which would produce a TL translation with no gender agreement between the determiner and the noun. With this EAT, the translation into Spanish of the English phrase *Fran's pen*, with SL IR w_1 =*Fran* `PN`, w_2 ='*s* `POS`, w_3 =*pen* `N-num:sg`, would be *el bolígrafo de Fran*, with TL IR w'_1 =*el* `DT-gen:m.num:sg`, w'_2 =*bolígrafo* `N-gen:m.num:sg`, w'_3 =*de* `PR`, w'_4 =*Fran* `NP`.

2.2.1 Main limitations

Although the method by Sánchez-Martínez and Forcada (2009) infers shallow-transfer rules capable of producing translations that are close to those produced with hand-crafted rules and provides better results than SMT systems trained on the same parallel corpus extended with the bilingual dictionary of the RBMT system (Sánchez-Martínez and Forcada, 2009, Sec. 5.2.1), it has three main limitations which are described below. The first two limitations are inherent to the expressiveness of the EATs they use, whereas the third one is a limitation of the aforementioned authors' learning algorithm.

First limitation: partial generalisation. The definition of word classes is not sufficiently flexible to permit EATs with different generalisation levels. The most general word classes are obtained by removing the lemma from the lexical forms and they therefore take into account the lexical category and all the inflection information (e.g. gender, number, verb tense, person, case, etc.). This often results in having to learn several EATs in order to describe the translation of the same linguistic phenomenon. For instance, adjectives in English are placed before the noun, whereas in Spanish they are usually placed after the noun. In order to properly translate a noun followed by an adjective from Spanish to English, Sánchez-Martínez and Forcada (2009) need to learn the four EATs shown in Figure 2.2; these EATs only differ in the morphological inflection information (gender and number) of the lexical forms they match. Note that if more general word classes were used, so that all adjectives (or nouns) were assigned to the same word class regardless of the inflection information, the noun–adjective reordering could be described with the EAT shown in Figure 2.3. In that Figure, the morphological inflection attributes `gen:*` and `num:*` in the SL word classes mean that they match any gender and number, respectively. The morphological inflection attribute `num:`$\$^1_s$ in a TL word class means that the TL lexical form produced as a

translation takes the value of the attribute with the same name in the first SL lexical form matched (more information on the word classes used is provided in Section 2.3).

In general, the new rule inference algorithm presented in this chapter solves the partial generalisation limitation by using word classes with different levels of generalisation and exploiting the information contained in the bilingual phrase pairs to decide, in a context-dependent manner, the generalisation level of the EATs, that is, the morphological inflection attributes that contain the wildcard value (*) that matches any possible value. In the new approach, multiple EATs, with different levels of generalisation, are generated from each bilingual phrase pair. The set of EATs to be used —and therefore the appropriate generalisation level to be used to describe the translation of the different linguistic phenomena found in the training corpus— is then automatically determined by selecting the minimum number of EATs that are needed to reproduce the bilingual phrase pairs from which the EATs are obtained. In order to deal with the complexity of choosing that minimum set of EATs when working with all the bilingual phrase pairs extracted from the corpus, the problem is posed as an optimisation problem by defining a set of inequations which are solved using integer linear programming methods (Garfinkel and Nemhauser, 1972). This approach is the first in literature (see Section 1.3.1) in which the problem of automatically inferring transfer rules is reduced to finding the optimal value of a minimisation problem.

Being able to use word classes with different generalisation levels implies that the new method needs fewer examples to learn common structural transformations between the SL and the TL. In addition, having more general EATs makes it easier for linguists to revise the inferred rules.

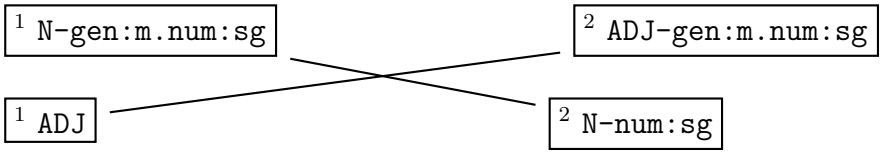
Second limitation: no context-dependent lexicalisations. The way in which word classes are defined by Sánchez-Martínez and Forcada (2009), that is, by using a set of lexicalised units, not only prevents better generalisations from being created, as explained above, but also prevents context-dependent lexicalisation from taking place. Context-dependent lexicalisation would permit a different treatment to be given to those words that, in a given context, are not properly translated by more general EATs. For instance, in Spanish some adjectives —called *prepositive* adjectives[1]— are usually placed before the noun, e.g. *gran hombre*,[2] instead of after the noun as usual. In order to properly translate the English adjective–noun construction into Spanish when the Spanish equivalent of the English adjective is a prepositive adjective, prepositive adjectives need to be lexicalised. In the approach by Sánchez-Martínez and Forcada (2009) this would require knowing the set of the most frequent prepositive adjectives in Spanish in advance, adding them to the set of lexicalised units, and learning EATs like those shown in Figure 2.4 for the adjective *great*. Note that z_1 from Figure 2.4

[1] Although only a reduced set of adjectives are always prepositive in Spanish, all adjectives can be used prepositively in poetry and literature. Postpositive adjectives are unusual in English, but they can be found in phrases such as *body politic*, *queen consort* or *time immemorial*.

[2] Translated into English as *great man*.

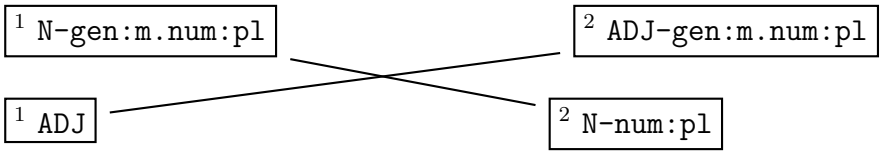
Figure 2.2: Set of EATs needed by Sánchez-Martínez and Forcada (2009) to codify the noun–adjective reordering when translating Spanish into English. z_1 will be used to translate *tren viejo* into *old train*; z_2 will be used to translate *trenes viejos* into *old trains*; z_3 will be used to translate *locomotora vieja* into *old locomotive*; z_3 will be used to translate *locomotoras viejas* into *old locomotives*.

z_1:



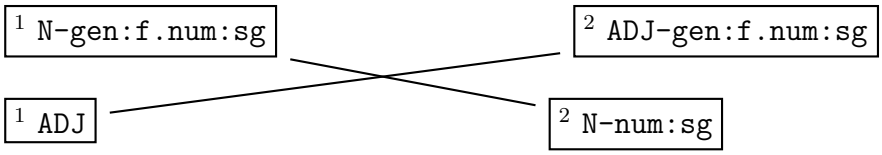
$r_1 = \{\}, r_2 = \{\}$

z_2:



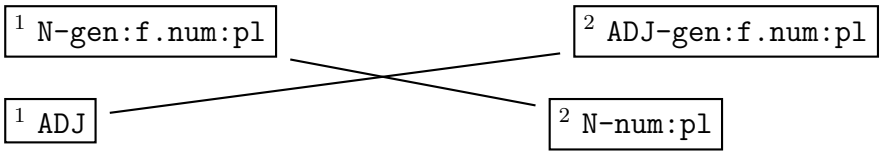
$r_1 = \{\}, r_2 = \{\}$

z_3:



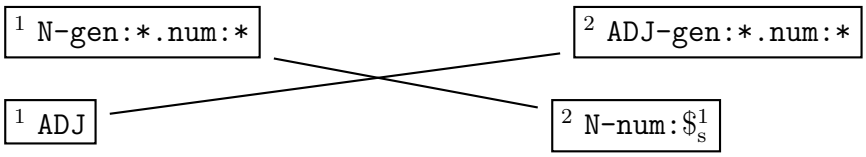
$r_1 = \{\}, r_2 = \{\}$

z_4:



$r_1 = \{\}, r_2 = \{\}$

Figure 2.3: EAT learnt by the approach presented in this chapter in order to codify the noun–adjective reordering when translating Spanish into English.



$r_1 = \{\}, r_2 = \{\}$

is an exception to the general rule used to translate the adjective–noun constructions because it does not perform any reordering, as opposed to the EAT in Figure 2.5.[3] Note also that the translation rule encoded in z_2 from Figure 2.4 is equivalent to the general rule used to translate a determiner, followed by a singular noun, the verb to *be* in the past tense, 3rd person, singular, and a (predicative) adjective. It is therefore clear that the lexicalisation in z_2 is not needed and performing such a lexicalisation leads Sánchez-Martínez and Forcada (2009) to generate more EATs than are really necessary, some of which may be useless.

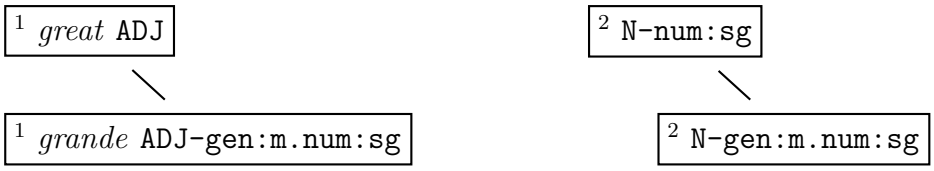
The new approach overcomes this limitation because the different generalisation levels explored for each word class include EATs in which the lemma of the lexical forms is kept unchanged. Then the approach outlined above to select the minimum number of EATs that are needed to reproduce the bilingual phrase pairs from which EATs are obtained is followed. Consequently, these lexicalisations are only used when they are needed to encode an exception to a more general translation rule.

Third limitation: rules preventing the application of more convenient rules. Finally, Sánchez-Martínez and Forcada (2009) do not apply any method with which to discard those EATs that force SL lexical forms that should be processed together by the same rule —because they are involved in the same linguistic phenomenon— to be dealt with by different rules. This is a common situation when the bilingual phrase pairs, from which the EATs are obtained, are extracted by following the standard method in SMT (Koehn, 2010, Sec. 5.2.3), which is likely to separate words that should be processed together into different phrases. This is a problem in shallow-transfer RBMT because an SL lexical form can only be translated by a single rule. In the specific case of Apertium, the SL sentence to be translated is divided into chunks so that each chunk is matched and translated by a single rule in a left-to-right, longest-match fashion. Apertium starts from the first SL lexical form in the sentence, selects the longest applicable rule, applies it to the matched chunk, prints the result, and starts the process again from the next (unmatched) SL lexical form in the sentence. If no

[3]It is assumed that if several EATs match the same sequence of SL lexical categories, the most specific EAT is applied. It is also assumed that the dictionaries of the RBMT system do not contain any information indicating whether or not an adjective is prepositive.

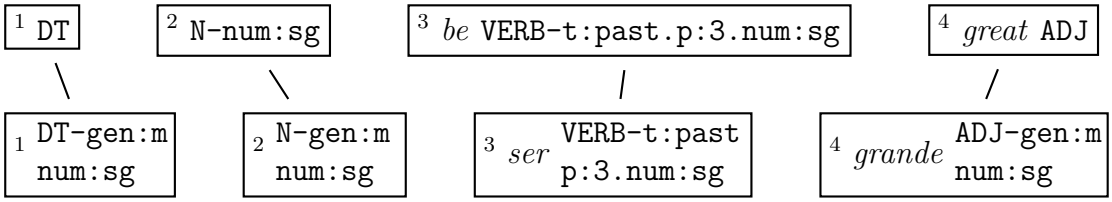
Figure 2.4: EATs learnt by Sánchez-Martínez and Forcada (2009) to translate the English adjective–noun construction into Spanish (z_1) when the adjective is *great*, and to translate this same adjective when it is preceded by a determiner, followed by a singular noun, and the verb to *be* in the past tense, 3rd person, singular (z_2). Note that this requires the adjective *great* to be added to the set of lexicalised units which do not have to be generalised.

z_1:



$r_1 = \{\}, r_2 = \{\texttt{gen:m}\}$

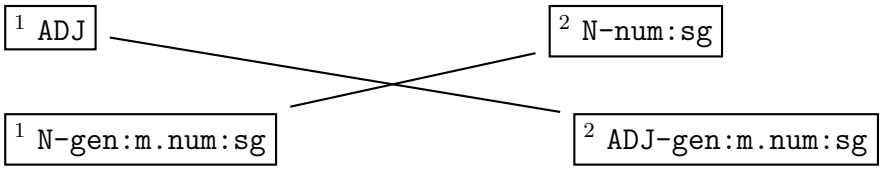
z_2:



$r_1 = \{\}, r_2 = \{\texttt{gen:m}\}$

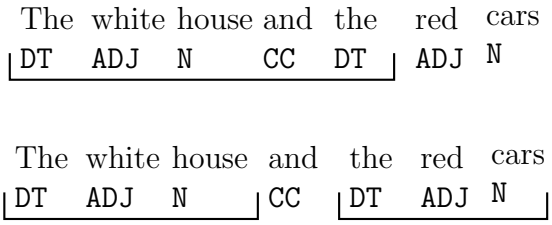
Figure 2.5: EAT learnt by Sánchez-Martínez and Forcada (2009) to translate the English adjective–noun construction into Spanish when the noun is singular and its translation into Spanish is masculine.

z_1:



$r_1 = \{\}, r_2 = \{\texttt{gen:m}\}$

Figure 2.6: Example of the application of shallow-transfer rules in Apertium. The rule that matches a determiner–adjective–noun–conjunction–determiner construction is applied at the top and, as a result, the last two words of the sentence are translated in isolation. The resulting translation into Spanish is *La casa blanca y el rojo coches*. Note that *el* is a masculine singular definite determiner that should be plural (i.e. *los*) in order to agree with the noun *coches*, and that the adjective *rojo* and the noun *coches* should get reordered and agree in gender and number. At the bottom, the last three words of the sentence are translated by a rule that matches a determiner–adjective–noun construction and performs the reordering and the gender and number agreement between the matched words. The resulting translation is *La casa blanca y los coches rojos*.



rule can be matched, the corresponding SL lexical form is translated in isolation and the process starts again with the next one.

Not discarding the EATs that perform a segmentation of the SL sentence that is unsuitable for a shallow-transfer RBMT system may result in not applying other EATs that would perform a correct translation, despite having been learnt from the parallel corpus. This is illustrated in Figure 2.6, in which an EAT that matches a determiner, an adjective, a noun, a conjunction (`CC`) and another determiner is applied (top), rather than applying an EAT that matches a determiner, an adjective and a noun twice (bottom). In the first case, the determiner after the conjunction is not processed together with the noun and the adjective it has to agree with in gender and number, which may result in an incorrect translation into the TL.

This last problem is tackled in the new approach by retaining in the final set of EATs only those that make the translation of the SL side of the training parallel corpus sufficiently close to its TL side when the EATs to be used are selected in a left-to-right longest match manner, as the RBMT engine will do. This is done by following a greedy approach in order to identify the set of sequences of lexical categories that should be translated by the same rule, and by removing those EATs that produce the same translation as a set of shorter EATs would produce.

2.3 Generalised alignment templates

This section describes the notation that will be used in the remainder of the chapter along with the improvement made to the EAT formalism used by Sánchez-Martínez

and Forcada (2009) in order to be able to learn more general EATs which we shall refer to as *generalised alignment templates* (henceforth, GATs).

As stated in the introduction, shallow-transfer RBMT systems use as IR sequences of lexical forms in both languages. Recall that the translation process in shallow-transfer RBMT is as follows: first the SL IR is obtained from the SL text, usually with the help of a monolingual dictionary and a part-of-speech tagger; then the SL IR is converted into a TL IR by applying shallow-transfer rules (in this case encoded as GATs) and using a bilingual dictionary; finally the TL text is generated from the TL IR with the help of a TL monolingual dictionary.

A lexical form w, e.g. *car* `N-gen:`ϵ`.num:sg`, consists of:

- a lemma $\lambda(w)$, e.g. $\lambda(w) =$*car*,
- a lexical category[4] $\rho(w)$, e.g. $\rho(w) =$`N` (noun),
- a set of morphological inflection attributes $\alpha(w)$, e.g. $\alpha(w) = \{$`gen`,`num`$\}$ (gender and number), and
- their values $\upsilon(w, a)$, e.g. $\upsilon(w,$`num`$) =$`sg` (singular).

Some morphological inflection attributes may be assigned an empty value (ϵ) because they do not apply to that language; in the example above, $\upsilon(w,$`gen`$) = \epsilon$ because nouns do not have a gender in English. This is done for convenience so that lexical forms have the same morphological inflection attributes in the two languages involved in the translation. It is worth noting that the functions described above can equally be applied to lexical forms and word classes; what is more, $\alpha(\cdot)$ and $\upsilon(\cdot)$ can also be applied to restrictions.[5]

SL lexical forms are translated into TL lexical forms by looking them up in a bilingual dictionary. An SL lexical form may have more than one equivalent in the TL; in these cases, a lexical selection module (Tyers et al., 2012; Tyers, 2013) is responsible for selecting the most appropriate translation given the SL context prior to the execution of the structural transfer module. The result of translating an SL lexical form w into a TL lexical form is referred to as $\tau(w)$ throughout this chapter.

In order to be able to learn GATs, the following special values for the morphological inflection attributes have been introduced in the new approach described in this chapter:

[4]Without loss of generality, ρ could also be used to represent lexical subcategories.

[5]Note that a restriction merely consists of a set of restricted morphological inflection attributes and their values.

- The wildcard $*$ value in the morphological inflection attribute of an SL word class signifies that it matches any value. Hence, a GAT $z = (S, T, A, R)$,[6] with $S = (s_1, s_2, \cdots, s_n)$, and $R = (r_1, r_2, \cdots, r_n)$, *matches* a sequence of SL lexical forms $W = (w_1, w_2, \cdots, w_n)$ only if W and S have the same length and every SL lexical form $w_i \in W$ meets the following conditions:

  - either its lemma equals the lemma in the SL word class s_i, or s_i has no lemma (because it has been generalised):

$$\lambda(w_i) = \lambda(s_i) \ \vee \ \lambda(s_i) = \epsilon;$$

  - its lexical category equals the lexical category in the SL word class s_i:

$$\rho(w_i) = \rho(s_i);$$

  - either the value of the morphological inflection attributes of w_i equal those in s_i, or the value of the corresponding morphological inflection attributes in s_i contain wildcards:

$$\forall a \in \alpha(s_i) : \upsilon(w_i, a) = \upsilon(s_i, a) \ \vee \ \upsilon(s_i, a) = *;$$

  - the value of the morphological inflection attributes specified in the restrictions r_i are equal to those in the TL lexical form obtained by looking up the SL lexical form w_i in the bilingual dictionary:[7]

$$\forall a \in \alpha(r_i), \upsilon(\tau(w_i), a) = \upsilon(r_i, a).$$

- The SL reference $\$_s^j$ as the value of an attribute a of a TL word class t_i means that the TL lexical form w'_i produced as a translation takes the value of the corresponding attribute from the j-th SL lexical form matched by the GAT:

$$\upsilon(w'_i, a) \leftarrow \upsilon(w_j, a).$$

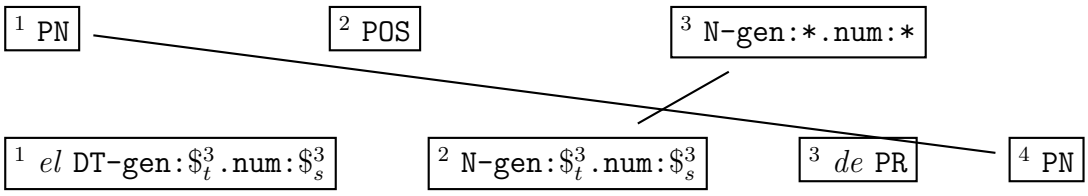
- The TL reference $\$_t^j$ as the value of an attribute a of a TL word class t_i means that a takes the value from the corresponding morphological inflection attribute in the TL lexical form obtained after translating the j-th SL lexical form by looking it up in the bilingual dictionary:

$$\upsilon(w'_i, a) \leftarrow \upsilon(\tau(w_j), a).$$

[6]The elements (S, T, A, R) have the same meaning as in EATs: S is a sequence of SL word classes, T is a sequence of TL word classes, A is a set of pairs of word class indexes (i, j) with the alignment information between the word classes in S and T, and R is a set of restrictions over the TL morphological inflection information of the lexical forms matching the GAT.

[7]The bilingual dictionary provides the translation of the lemma and also of the lexical category and morphological inflection attributes of the lexical form. For instance, the bilingual dictionary provides the gender that a noun must have when it is translated into Spanish.

Figure 2.7: GAT for the translation of the English Saxon genitive construction into Spanish. Compare with the EAT learnt by Sánchez-Martínez and Forcada (2009) (see Figure 2.1).



$r_1 = \{\}, r_2 = \{\}, r_3 = \{\}$

It is worth noting that, even though the translation of most linguistic phenomena can be encoded using only TL references ($\$_t^j$), there are situations, such as that described below, in which SL references ($\$_s^j$) are needed. Consider the translation into English of the Spanish phrase *es guapa*,[8] with SL IR $w_1 =$ *ser* `VERB-t:present.p:3.num:sg`, $w_2 =$ *guapa* `ADJ-gen:f.num:sg`. As the Spanish phrase contains no personal pronoun, the GAT that must be applied for its translation has to resort to the gender of the adjective in the SL to determine which pronoun, *she* or *he*, needs to be used, and an SL reference therefore needs to be used.

Apart from the changes explained above, GATs are applied to translation in the same way in which the EATs of Sánchez-Martínez and Forcada (2009) are applied. The following example illustrates how the new attribute values $\$_s^j$ and $\$_t^j$ are used during translation. The GAT shown in Figure 2.7 encodes the translation of the English Saxon genitive construction —proper noun + possessive ending + noun— into Spanish. The wildcard attribute in the number makes it match both singular and plural nouns; the SL and TL reference values propagate the gender and number of the noun to the determiner. When translating the SL (English) phrase *Mary's family*, with the SL IR $w_1 =$ *Mary* `PN`, $w_2 =$ *'s* `POS`, $w_3 =$ *family* `N-gen:`ϵ`.num:sg` with the GAT in Figure 2.7, the four TL lexical forms $w'_1 \cdots w'_4$ produced as output are obtained as follows. The lemmas of the first and third TL lexical forms are taken from the GAT: $\lambda(w'_1) = \lambda(t_1) =$ *el* and $\lambda(w'_3) = \lambda(t_3) =$ *de*. The lemmas of the other two TL lexical forms are obtained by looking up the SL lexical forms aligned with them in the bilingual dictionary: $\lambda(w'_2) = \lambda(\tau(w_3)) =$ *familia*, $\lambda(w'_4) = \lambda(\tau(w_1)) =$ *Mary*. The lexical categories are taken from the TL word classes: $\rho(w'_1) = \rho(t_1) =$ `DT`, $\rho(w'_2) = \rho(t_2) =$ `N`, $\rho(w'_3) = \rho(t_3) =$ `PR`, and $\rho(w'_4) = \rho(t_4) =$ `PN`. The morphological inflection attributes `gen` (gender) of the first and second TL lexical forms take their values from the corresponding attribute in the translation of the third SL lexical form (TL reference): $\upsilon(w'_1, \texttt{gen}) = \upsilon(w'_2, \texttt{gen}) = \upsilon(\tau(w_3), \texttt{gen}) =$ `f`. The morphological inflection attributes `num` (number) of these same TL lexical forms take their value from the corresponding attribute in the third SL lexical form (SL reference): $\upsilon(w'_1, \texttt{num}) = \upsilon(w'_2, \texttt{num}) = \upsilon(w_3, \texttt{num}) =$ `sg`. The resulting sequence of TL (Spanish)

[8] Translated into English as *She is beautiful*.

lexical forms is w'_1 = *el* `DT-gen:f.num:sg`, w'_2 = *familia* `N-gen:f.num:sg`, w'_3 = *de* `PR`, w'_4 = *Mary* `PN`, which after morphological generation leads to *la familia de Mary*.

2.4 Inference of shallow-transfer rules

The complete process followed to obtain shallow-transfer rules from a parallel corpus consists of the steps described in the remainder of this section and summarised in Figure 2.8. First, word alignments and bilingual phrase pairs are obtained from the parallel corpus (Section 2.4.1). Multiple GATs, each one with a different level of generalisation, are then inferred from each of the bilingual phrase pairs obtained. This is done by using different sets of wildcard and reference attributes, and also with different lexicalised words (Section 2.4.2); these GATs, encoded with the formalism described in Section 2.3 do not suffer from the partial generalisation issue described in Section 2.2.1. After filtering certain GATs to deal with the noise present in the corpus and to prevent overgeneralisations (Section 2.4.3), the GATs with the most appropriate lexicalised words and wildcard and reference attributes are automatically selected by finding the minimum set of GATs needed to correctly reproduce all the bilingual phrase pairs obtained from the corpus (Section 2.4.4). With this minimisation process, conflicts between GATs are removed and GATs with lexicalised word classes are selected only when they are strictly necessary in the context in which they appear; the second limitation described in Section 2.2.1 is therefore overcome. Any GATs that cause deficient chunking of the input are then discarded (Section 2.4.5) in order to get over the third limitation described in Section 2.2.1. Finally, the GATs selected are converted into the Apertium rule format, although they could be converted into the format used by any other shallow-transfer RBMT system.

2.4.1 Obtaining word alignments and bilingual phrase pairs

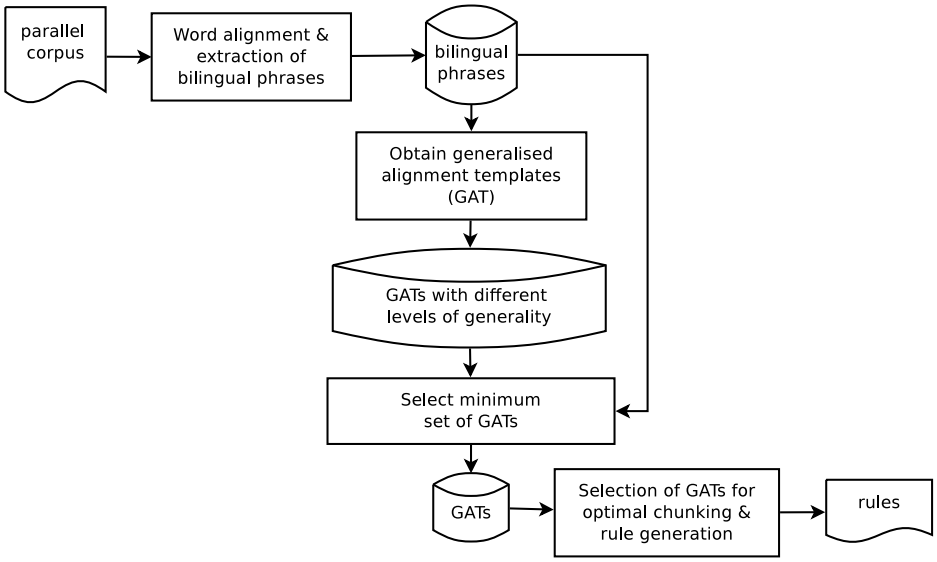
Word alignments and bilingual phrase pairs are obtained using the state-of-the-art method in order to obtain bilingual phrases pairs for their use in SMT (Koehn, 2010). This method, which was also followed by Sánchez-Martínez and Forcada (2009), consists of the following steps:

1. Morphologically analyse both sides of the parallel corpus and solve the part-of-speech ambiguities in order to obtain sequences of lexical forms in both languages.

2. Train IBM models 1, 3 and 4 (Brown et al., 1993b), and the HMM alignment model (Vogel et al., 1996), for 5 iterations by means of GIZA++ for both translations directions (Och and Ney, 2003).[9]

[9] `http://code.google.com/p/giza-pp/`

Figure 2.8: Steps followed to obtain a set of generalised alignment templates (GAT) from a parallel corpus.



3. Compute the Viterbi alignment according to these models for both translation directions.

4. Symmetrise the two sets of Viterbi alignments using the refined intersection method proposed by Och and Ney (2003) to obtain word-aligned sentence pairs.

5. Extract bilingual phrase pairs that are consistent with the alignments (Koehn, 2010, Sec. 5.2.3).

6. Discard bilingual phrase pairs not suitable for rule inference.

The following criteria, also followed by Sánchez-Martínez and Forcada (2009), are applied in order to identify which are the bilingual phrase pairs suitable for rule inference:

- Bilingual phrase pairs containing either unknown words or punctuation marks are removed. On the one hand, bilingual phrase pairs containing unknown words are not useful unless a morphological guesser is used; on the other hand, it is assumed that punctuation marks do not provide relevant information from the point of view of the structural transference.

- Bilingual phrase pairs whose first or last word on either side (SL and TL) are left unaligned are discarded because there is no evidence that they are actually part of the translation of the segment in the opposite language, and using them could result in incorrect GATs.

- Bilingual phrase pairs that are not consistent with the bilingual dictionary are also discarded to avoid unnecessary lexicalisations (see below).

2.4.1.1 Bilingual phrase pairs consistent with the bilingual dictionary

As it has been previously pointed out, the approach described in this chapter generates a set of GATs which correctly reproduces all the bilingual phrase pairs. When the translation of a word in a bilingual phrase pair does not appear as an equivalent in the bilingual dictionary, the GATs obtained from it need to be lexicalised, i.e. its lemma cannot be removed from the corresponding word classes (the process for obtaining GATs from bilingual phrases will be described in Section 2.4.2). If a bilingual phrase pair consists of a *free* translation or contains translation equivalents that are different to those in the Apertium dictionaries, an unnecessary lexicalisation may occur. To avoid these unnecessary lexicalisations while allowing the method to learn common lexical changes between the SL and the TL, the set of bilingual phrase pairs obtained is filtered.[10] Those bilingual phrase pairs for which one of the following conditions is not met for all SL and TL lexical forms are discarded:

1. If the lexical form w is an *open-class* lexical form, i.e. it belongs to an open-class lexical category,[11] it must be either aligned with an open-class lexical form in the other language that appears in the bilingual dictionary as an equivalent for w, or otherwise not aligned with any open-class lexical form. If it is a *closed-class* lexical form it may be aligned with any lexical form. This filtering is based on the assumption that open-class words carry the meaning of the sentence while the role of closed-class words is to provide grammatical information.

2. If an (open-class) lexical form w does not meet the previous condition, it must be single-aligned to an open-class lexical form in the other language that meets the first condition. This second condition is based on the assumption that here the open-class lexical form that does not meet the first condition might be working as an auxiliary particle, and does not therefore convey any meaning.

[10] In the method by Sánchez-Martínez and Forcada (2009), the filtering is much simpler and consists of discarding those phrase pairs with at least one lexical form aligned with a lexical form in the other language that does not appear in the bilingual dictionary as its equivalent and that does not belong to the set of lexicalised units provided by the user. In the experiments reported in Section 2.5 the sets of lexicalised units used were as follows: for the Spanish↔Catalan language pair, the set of lexicalised units originally defined by Sánchez-Martínez and Forcada (2009) was used; for the rest of language pairs the set of closed-class lexical forms was used instead.

[11] Nouns, adjectives, adverbs, and verbs are among the set of open-class lexical categories; whereas determiners, pronouns and prepositions are considered to be closed-class lexical categories.

2.4.2 Extracting generalised alignment templates from bilingual phrase pairs

From each bilingual phrase p, many different GATs that correctly *reproduce* it —when applied to the SL phrase in p, the corresponding TL phrase is obtained— are generated, although not all of them will eventually be used for rule generation. The selection of GATs to be used for rule generation is described in sections 2.4.4 and 2.4.5.

Given a bilingual phrase pair p, the generation of GATs from it can be described as the initial generation of the most specific GAT, $\beta(p)$ (Section 2.4.2.1), and the chained application of 3 different functions (σ_1, σ_2 and σ_3), each one of which takes the set of GATs produced by the previous one as input and generates a new set of GATs from each GAT in the input set. Function σ_1 removes lemmas from word classes (Section 2.4.2.2), function σ_2 introduces wildcards, SL and TL references and removes restrictions (Section 2.4.2.3), and function σ_3 ensures that each non-lexicalised TL word class is aligned with at most one SL word class (Section 2.4.2.4).

2.4.2.1 Obtaining the initial generalised alignment template (β)

The initial GAT $z = \beta(p) = (S, T, A, R)$ created from a bilingual phrase pair p is the most specific GAT that can be obtained from it, and therefore only matches the SL phrase in p.

Let $p = (W, W', A')$ be a bilingual phrase pair with a sequence of SL lexical forms $W = (w_1, w_2, \cdots, w_n)$, a sequence of TL lexical forms $W' = (w'_1, w'_2, \cdots, w'_m)$ and alignment information $A' = \{(i, j) : i \in [1, n] \wedge j \in [1, m]\}$. Each SL word class $s_i \in S$ in GAT z has the same lemma, lexical category and morphological inflection attribute values as the corresponding SL lexical form w_i in p, i.e. $\forall i \in [1, n], s_i \leftarrow w_i$. The same applies to the TL word classes: $\forall i \in [1, m], t_i \leftarrow w'_i$. The alignment information A in z is also copied from the bilingual phrase pair: $A \leftarrow A'$. Finally, restrictions R are obtained by looking up each SL lexical form in the bilingual phrase pair in the bilingual dictionary as follows:

$$\forall w_i \in W, \alpha(r_i) \leftarrow \alpha(w_i), \text{ and}$$

$$\forall r_i, \forall a \in \alpha(r_i), \upsilon(r_i, a) \leftarrow \upsilon(\tau(w_i), a).$$

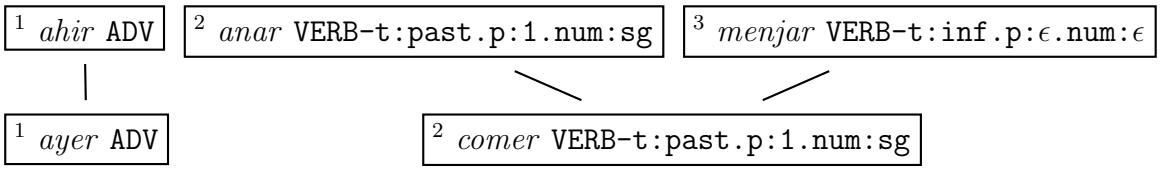
Figure 2.9 shows a bilingual phrase pair p and the initial GAT z obtained from it. Note that the restrictions limit the morphological attribute values to those in the bilingual dictionary.

2.4.2.2 Removing lemmas (σ_1)

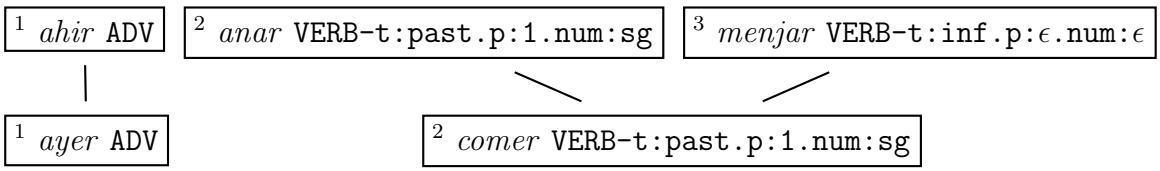
The next step as regards obtaining more general GATs is to remove from each initial GAT the lemma from some of the SL and TL lexical forms that are related according

Figure 2.9: Catalan–Spanish bilingual phrase pair p and initial GAT z obtained from it with function β (see Section 2.4.2.1).

p :



z:



$r_1 = \{\}, r_2 = \{\texttt{t} : \texttt{past}, \texttt{p} : \texttt{1}, \texttt{num} : \texttt{sg}\}, r_3 = \{\texttt{t} : \texttt{inf}, \texttt{p} : \epsilon, \texttt{num} : \epsilon\}$

to the bilingual dictionary. Recall that, during translation, when a TL word class does not contain a lemma, the lemma of the TL lexical form produced is obtained by looking up the SL lexical form to which it is aligned in the bilingual dictionary.

Function σ_1 generates a new GAT for each of the possible subsets of the set E with the positions of the SL word classes from which the lemma can be removed. Given an input GAT $z = (S, T, A, R)$, with $S = (s_1, s_2, \cdots, s_n)$ and $T = (t_1, t_2, \cdots, t_m)$, the set E is obtained by first computing for each SL word class s_i the set D_i of the TL word classes aligned to it whose lemmas are related according to the bilingual dictionary:

$$D_i = \{t_j : (i, j) \in A \ \wedge \ \lambda(\tau(s_i)) = \lambda(t_j)\};$$

and then including in E the positions of the SL word classes whose lemmas, according to the bilingual dictionary, are related to at least one TL word class:

$$E = \{i : D_i \neq \emptyset\}.$$

For each possible subset $F \in \mathcal{P}(E)$,[12] σ_1 generates a new GAT z'. Each new GAT $z' = (S', T', A, R)$ is a copy of z in which the lemmas have been removed from the SL word classes whose positions are specified in F, and from the TL word classes aligned with them whose lemmas are related according to the bilingual dictionary:

$$\forall i \in F, \lambda(s'_i) \leftarrow \epsilon \text{ and}$$

$$\forall i \in F, \forall j : (i, j) \in A \ \wedge \ \lambda(\tau(s_i)) = \lambda(t_j), \lambda(t'_j) \leftarrow \epsilon.$$

[12] $\mathcal{P}(E)$ is the power set of E.

As the empty set $\emptyset$ is always contained in $\mathcal{P}(E)$, the initial (non-generalised) GAT is always contained in the output of σ_1 (identity transformation).

Figure 2.10 shows the result of applying σ_1 to the GAT shown in Figure 2.9 (z_0). In this example, the number of GATs to be generated is 4 and $E = \{1, 3\}$ because, according to the bilingual dictionary, the first SL lemma is translated as the first TL lemma, and the third SL lemma is translated as the second TL lemma.

2.4.2.3 Introducing wildcards and references in the morphological inflection attributes (σ_2)

The use of wildcards and SL and TL references in the morphological inflection attributes allows the translation rules to be generalised to words with different values in their morphological attributes. This allows, for example, general reordering rules, like that presented in Figure 2.3, to be learnt, which are usually independent of the gender and number of the words involved.

Function σ_2 generates a set of GATs for each input GAT z by introducing wildcards in some of the morphological inflection attributes of the SL word classes and references in the counterpart morphological attributes of the TL word classes. It also removes the restrictions associated with the attributes of the SL word classes whose values have been replaced with a wildcard.

For each input GAT $z = (S, T, A, R)$, it is first necessary to obtain the set of candidate attributes C which are allowed to contain wildcards in the SL and references in the TL, and then the sets $M_{j,a}$ of possible SL references and TL references for each TL word class t_j and morphological inflection attribute $a \in C$.

A morphological attribute a is present in C only if, for each TL word class $t_j \in T$ with $a \in \alpha(t_j)$, it contains an empty value ($\upsilon(t_j, a) = \epsilon$) or the non-empty value it contains can be obtained with an SL reference ($\exists i : \upsilon(s_i, a) = \upsilon(t_j, a)$) or with a TL reference ($\exists i : \upsilon(r_i, a) = \upsilon(t_j, a)$):

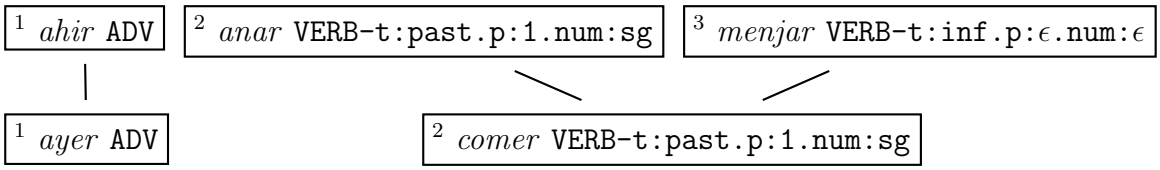
$$\begin{aligned} C = \quad & \{a : \upsilon(t_j, a) = \epsilon \ \vee \ (\exists i : \upsilon(s_i, a) = \upsilon(t_j, a)) \ \vee \\ & (\exists i : \upsilon(r_i, a) = \upsilon(t_j, a)) \ \forall t_j \in T : a \in \alpha(t_j)\}; \end{aligned}$$

Note that the restrictions are used to check whether an attribute value can be obtained with a TL reference, since their values have been obtained from the bilingual dictionary.

The sets $M_{j,a}$ of possible SL references and TL references for each TL word class t_j and morphological inflection attribute $a \in \alpha(t_j) \cap C$ are computed using Algorithm 1. This algorithm proceeds as follows. If attribute a can be obtained with a reference to an SL word class to which t_j is aligned, the corresponding reference is added to $M_{j,a}$. If not, the algorithm adds references to other SL word classes from which attribute a can be obtained to $M_{j,a}$. In either case, the SL references are only included in $M_{j,a}$ if TL references cannot be used. $M_{j,a,1}$ represents the TL reference attributes to SL word

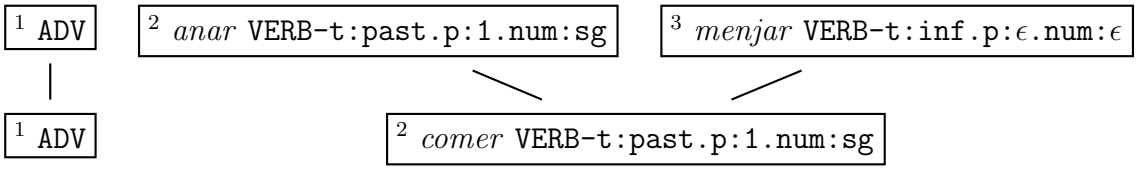
Figure 2.10: Set of GATs generated by σ_1 from the GAT in Figure 2.9 (z_0). For each GAT, the set $F \in \mathcal{P}(E)$ used to remove the lemmas is provided; $E = \{1, 3\}$ (see Section 2.4.2.2). Note that, according to the bilingual dictionary, the translation into Spanish of a lexical form whose lemma is *menjar* is a lexical form whose lemma is *comer*, while the translation of a lexical form whose lemma is *anar* is a lexical form whose lemma is *ir*, which is not part of any TL word class in z_0.

$z_0 = z_1$:



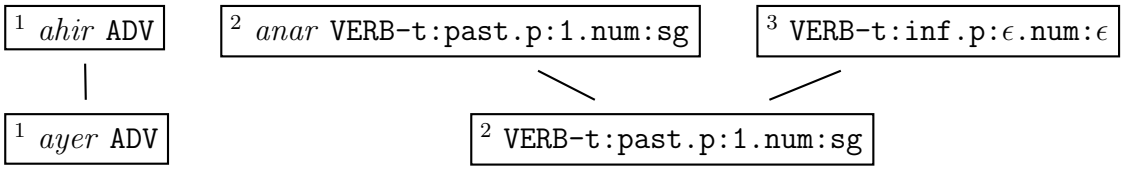
$r_1 = \{\}, r_2 = \{\texttt{t} : \texttt{past}, \texttt{p} : \texttt{1}, \texttt{num} : \texttt{sg}\}, r_3 = \{\texttt{t} : \texttt{inf}, \texttt{p} : \epsilon, \texttt{num} : \epsilon\}$; $F = \{\}$

z_2:



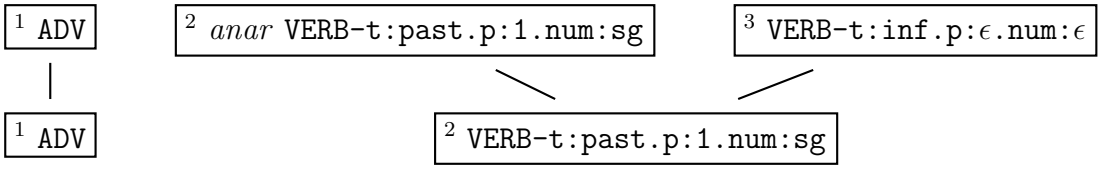
$r_1 = \{\}, r_2 = \{\texttt{t} : \texttt{past}, \texttt{p} : \texttt{1}, \texttt{num} : \texttt{sg}\}, r_3 = \{\texttt{t} : \texttt{inf}, \texttt{p} : \epsilon, \texttt{num} : \epsilon\}$; $F = \{1\}$

z_3:



$r_1 = \{\}, r_2 = \{\texttt{t} : \texttt{past}, \texttt{p} : \texttt{1}, \texttt{num} : \texttt{sg}\}, r_3 = \{\texttt{t} : \texttt{inf}, \texttt{p} : \epsilon, \texttt{num} : \epsilon\}$; $F = \{3\}$

z_4:



$r_1 = \{\}, r_2 = \{\texttt{t} : \texttt{past}, \texttt{p} : \texttt{1}, \texttt{num} : \texttt{sg}\}, r_3 = \{\texttt{t} : \texttt{inf}, \texttt{p} : \epsilon, \texttt{num} : \epsilon\}$; $F = \{1, 3\}$

Algorithm 1 Algorithm that computes the set of possible SL and TL reference values that a given morphological inflection attribute a of a TL word class t_j can have.

$M_{j,a,1} \leftarrow \{\$_t^i : (i,j) \in A \ \wedge \ \upsilon(r_i,a) = \upsilon(t_j,a)\}$
$M_{j,a} \leftarrow M_{j,a,1} \cup \{\$_s^i : (i,j) \in A \ \wedge \ \upsilon(s_i,a) = \upsilon(t_j,a) \ \wedge \ \$_t^i \notin M_{j,a,1}\}$
if $M_{j,a} = \emptyset$ **then**
  $M_{j,a,2} \leftarrow \{\$_t^i : \upsilon(r_i,a) = \upsilon(t_j,a)\}$
  $M_{j,a} \leftarrow M_{j,a,2} \cup \{\$_s^i : \upsilon(s_i,a) = \upsilon(t_j,a) \ \wedge \ \$_t^i \notin M_{j,a,2}\}$
end if
return $M_{j,a}$

classes to which t_j is aligned that give the value of the attribute a in t_j as a result, while $M_{j,a,2}$ contains the TL reference attributes to SL word classes to which t_j is not aligned and give the value of that attribute a as a result.

Finally, a set of GATs G_L is then obtained for each possible set of attributes $L \in \mathcal{P}(C)$, thus permitting GATs with different generalisation levels to be built: the more attributes in L, the more general the resulting GATs. As occurs with σ_1, the empty set $\emptyset$ is always contained in L, and every input GAT is therefore also part of the result of applying σ_2 to it.

All the GATs in G_L share the same sequence of SL word classes S', set of restrictions R' and alignment information A'; they only differ in the sequence of TL word classes. S' is a copy of the original sequence of SL word classes S in which the value of the morphological inflection attributes in L has been replaced with a wildcard:

$$\forall s_i \in S, \forall a \in \alpha(s_i) : a \in L, \upsilon(s'_i, a) \leftarrow *;$$

R' is a copy of the original set of restrictions R in which the attributes in S whose values have been replaced with a wildcard in S' have been removed:

$$\forall s_i \in S, \forall a \in \alpha(s_i) : a \in L, r_i \leftarrow r_i - \{a\};$$

and A' is a copy of the original alignment information A.

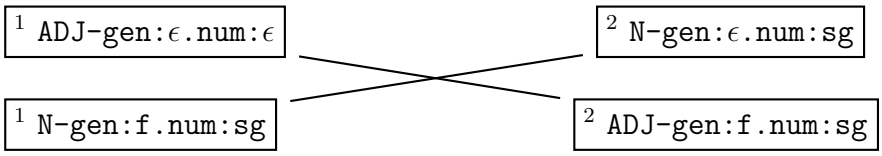
The different sequences of TL word classes to be generated, one for each GAT in G_L, differ as regards the attribute values that need to be used. These values are obtained as the Cartesian product $N = \prod_{t_j \in T} \prod_{a \in \alpha(t_j)} \omega(t_j, a)$, where $\omega(t_j, a)$ equals a set with the original attribute value if attribute a will not be assigned a reference, or otherwise a set with the references to be used:

$$\omega(t_j, a) = \begin{cases} M_{j,a} & \text{if } a \in L \wedge |M_{j,a}| > 0 \\ \{\upsilon(t_j, a)\} & \text{otherwise.} \end{cases}$$

Finally, a GAT is created for each element $n \in N$. The sequence of TL word classes T' of each new GAT is a copy of the original sequence of TL word classes T in which the values of the attributes have been replaced with those in n.

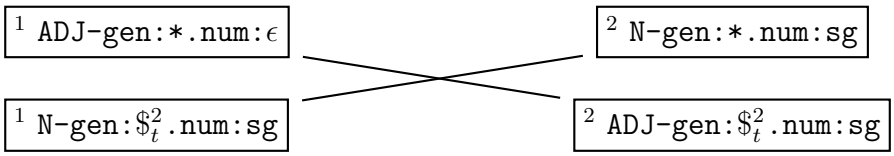
Figure 2.11: GAT codifying the reordering and gender and number agreement rule when translating a singular noun preceded by an adjective from English to Spanish (z_0). The noun is feminine in Spanish. The set of GATs (z_1–z_4) resulting from the application of σ_2 to z_0 are shown. For each GAT, the set L used to introduce wildcards in the SL and references in the TL are provided; $C = \{\texttt{gen}, \texttt{num}\}$, $M_{1,\texttt{gen}} = \{\$_t^2\}$, $M_{1,\texttt{num}} = \{\$_t^2\}$, $M_{2,\texttt{gen}} = \{\$_t^2\}$, $M_{2,\texttt{num}} = \{\$_t^2\}$ (see Section 2.4.2.3). The minimisation process described in Section 2.4.4 will be responsible for removing the redundancy present in this set of rules.

$z_0 = z_1$:



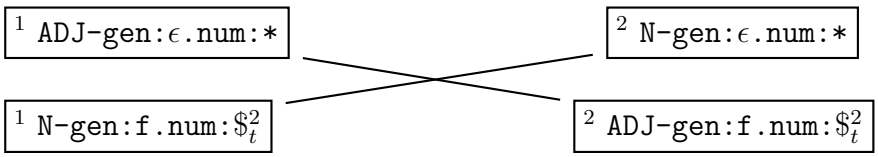
$r_1 = \{\texttt{gen} : \epsilon, \texttt{num} : \epsilon\}, r_2 = \{\texttt{gen} : \texttt{f}, \texttt{num} : \texttt{sg}\}$; $L = \{\}$

z_2:



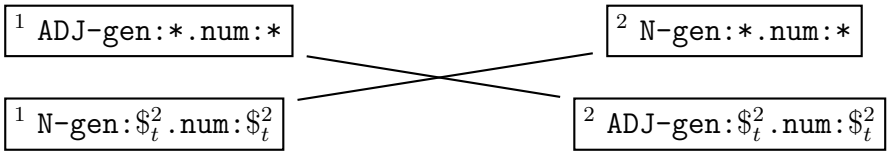
$r_1 = \{\texttt{num} : \epsilon\}, r_2 = \{\texttt{num} : \texttt{sg}\}$; $L = \{\texttt{gen}\}$

z_3:



$r_1 = \{\texttt{gen} : \epsilon\}, r_2 = \{\texttt{gen} : \texttt{f}\}$; $L = \{\texttt{num}\}$

z_4:



$r_1 = \{\}, r_2 = \{\}$; $L = \{\text{gen}, \text{num}\}$

Figure 2.11 shows the four GATs (z_1–z_4) generated by σ_2 for the input GAT z_0 from the same figure. These GATs codify the reordering and gender and number agreement rule that must be applied for the English–Spanish translation of an adjective followed by a noun. The set of morphological inflection attributes that can be assigned a wildcard in the SL, and a reference in the TL is $C = \{\texttt{gen}, \texttt{num}\}$; wildcards are permitted in the `num` (number) attribute because its value can be obtained by using an SL reference or a TL reference (in this case using both types of references) for both TL word classes; wildcards are permitted in the `gen` (gender) attributes because its value can be obtained using a TL reference. The sets of possible reference values to be used are $M_{1,\texttt{gen}} = \{\$_t^2\}$, $M_{1,\texttt{num}} = \{\$_t^2\}$, $M_{2,\texttt{gen}} = \{\$_t^2\}$ and $M_{2,\texttt{num}} = \{\$_t^2\}$, $\mathcal{P}(C) = \{\{\}, \{\texttt{gen}\}, \{\texttt{num}\}, \{\texttt{gen}, \texttt{num}\}\}$.

2.4.2.4 Removing alignments (σ_3)

For a GAT to be useful in shallow-transfer RBMT, every non-lexicalised TL word class must be aligned with at most one SL word class: that from which, at translation time, the TL lemma will be obtained by looking up the SL lexical form matched in the bilingual dictionary.

Function σ_3 removes those alignments that would render z not applicable in shallow-transfer RBMT from each input GAT $z = (S, T, A, R)$. This is done by first obtaining the set with the positions of the non-lexicalised TL word classes V:

$$V = \{j : t_j \in T \ \wedge \ \lambda(t_j) = \epsilon\}.$$

Then, for each TL word-class position $j \in V$, the set of possible alignments X_j is computed by considering the bilingual phrase pair (W, W') from which the GAT z was obtained and ensuring that it can be reproduced using the selected alignment points:

$$X_j = \{(i, j) : (i, j) \in A \ \wedge \ \lambda(\tau(w_i)) = \lambda(w'_j)\}.$$

Finally, all the possible subsets of A that ensure that the original bilingual phrase pair can be reproduced from z are calculated as the Cartesian product $Y = \prod_{j \in V} X_j$, and σ_3 generates an alternative GAT $z_y = (S, T, A_y, R)$ for each element $y \in Y$, where A_y stands for the subset of A that contains exactly the elements from the tuple y.

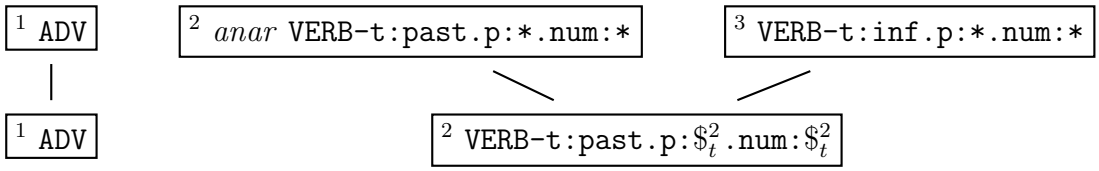
Figure 2.12 shows an input GAT z_0 and the GAT z_1 produced from it by σ_3. The valid alignment points for each TL word class are $X_1 = \{(1, 1)\}$ and $X_2 = \{(3, 2)\}$; the Cartesian product $Y = \{((1, 1), (3, 2))\}$ consists of a single element, which generates GAT z_1.

2.4.3 Filtering unreliable generalised alignment templates

Once a set of GATs has been generated from each bilingual phrase pair, a filtering of the GATs obtained must be carried out in order to discard those that are very

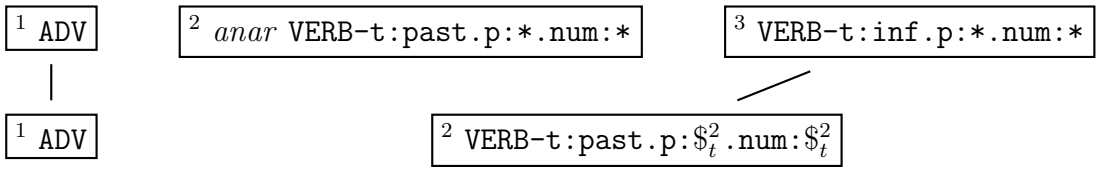
Figure 2.12: One of the GATs obtained from the bilingual phrase pair p in Figure 2.9 (z_0) and the GAT obtained from it (z_1) by function σ_3 (see Section 2.4.2.4).

z_0:



$r_1 = \{\}$, $r_2 = \{\texttt{t} : \texttt{past}\}$, $r_3 = \{\texttt{t} : \texttt{inf}\}$

z_1:



$r_1 = \{\}$, $r_2 = \{\texttt{t} : \texttt{past}\}$, $r_3 = \{\texttt{t} : \texttt{inf}\}$

infrequent or are not able to reproduce a large proportion of the bilingual phrase pairs they match. This may occur as a result of either the noise present in the training parallel corpus or overgeneralisations.

Given the set of bilingual phrase pairs P extracted from the parallel corpus (see Section 2.4.1) and a GAT $z \in Z$, the set of GATs obtained from P (see Section 2.4.2), $\mathcal{M}(z) \subseteq P$ is defined as the set of bilingual phrase pairs that are matched by z.[13] Some of these bilingual phrase pairs, $\mathcal{G}(z) \subseteq \mathcal{M}(z)$, are correctly translated by z —when applied to their SL side, their TL side is obtained— while others, $\mathcal{B}(z) = \mathcal{M}(z) - \mathcal{G}(z)$, are not.

The filtering consists of discarding, on the one hand, those GATs z whose number of correctly reproduced bilingual phrase pairs $\mathcal{G}(z)$ is below a threshold θ; and on the other hand, those GATs for which the ratio of bilingual phrase pairs correctly reproduced and matched to the total number of bilingual phrase pairs matched is below a threshold δ. Any GATs that encode very infrequent linguistic transformations, along with those that overgeneralise, are thus avoided. The number of matching and correctly reproduced bilingual phrase pairs is calculated by considering the frequency in the training parallel corpus of each bilingual phrase pair. A GAT z is thus discarded if

$$\mathcal{Q}(\mathcal{G}(z)) < \theta \ \vee \ \frac{\mathcal{Q}(\mathcal{G}(z))}{\mathcal{Q}(\mathcal{M}(z))} < \delta,$$

[13] A GAT matches a bilingual phrase pair if the SL word classes match the sequence of SL lexical forms and all restrictions are met (see Section 2.3 for more details).

where $\mathcal{Q}$ is the aggregated frequency of a set of bilingual phrase pairs:

$$\mathcal{Q}(P) = \sum_{p \in P} \text{count}(p),$$

and count(p) is the absolute frequency in the list of phrase pairs extracted from the parallel corpus (see Section 2.4.1) of the bilingual phrase pair p.

2.4.4 Choosing the most appropriate generalised alignment templates

The objective of this approach is to obtain a set of GATs that is able to correctly translate at least the set of bilingual phrase pairs extracted from the training parallel corpus. What is more, the GATs in that set must be as general as possible in order to extend the linguistic knowledge obtained from the corpus to unseen input texts. This objective is achieved by selecting the minimum amount of GATs needed to correctly reproduce all the bilingual phrase pairs. Since the more general the GATs, the higher the amount of bilingual phrase pairs they match and (hopefully) reproduce, if the amount of GATs is minimised, the most general ones that are able to reproduce the bilingual phrase pairs in the training corpus are selected.

Unlike the other approaches used to automatically learn shallow-transfer rules from parallel corpora (Sánchez-Martínez and Forcada, 2009; Caseli et al., 2006; Probst et al., 2002), here all the bilingual phrase pairs are considered together when checking their reproducibility by the set of GATs obtained. Conflicting rules are thus treated at a global level, while previous approaches treat them locally.

2.4.4.1 Minimisation problem definition

To define the minimisation problem, GATs need to be ordered according to their level of specificity. A GAT $z = (S, T, A, R)$ is said to be more specific than another GAT $z' = (S', T', A', R')$ if it has any component —either a lemma, a morphological inflection attribute or a restriction— that takes into account more fine-grained information than

z':[14]

$$\begin{aligned}\text{more\_specific}(z, z') \iff \quad & |S| = |S'| \ \wedge \ \forall s_i \in S, \\ & (\rho(s_i) = \rho(s'_i) \ \wedge \\ & (\lambda(s_i) = \lambda(s'_i) \ \vee \ \lambda(s'_i) = \epsilon) \ \wedge \\ & \forall a \in \alpha(s_i), (\upsilon(s_i, a) = \upsilon(s'_i, a) \ \vee \ \upsilon(s'_i, a) = *) \ \wedge \\ & \forall a \in \alpha(r_i), (\upsilon(r_i, a) = \upsilon(r'_i, a) \ \vee \ a \notin r'_i)) \ \wedge \\ & (\exists s_i \in S : s_i \neq s'_i \vee \exists r_i \in R : r_i \neq r'_i).\end{aligned}$$

On the basis of the set of bilingual phrase pairs P, the set of GATs Z and their relation of specificity defined by the function more_specific$(\cdot)$, the minimum set of GAT $O \subseteq Z$ is chosen subject to the following constraints:

$\mathcal{C}_1$: Each bilingual phrase pair is correctly reproduced by at least one GAT that is part of the solution:

$$\bigcup_{z_i \in O} \mathcal{G}(z_i) = P$$

$\mathcal{C}_2$: If a GAT z_i that is part of the solution incorrectly reproduces the TL part of a bilingual phrase pair p, there is another GAT z_j that is part of the solution, is more specific than z_i and correctly reproduces the TL part of p:

$$\forall z_i \in O, \forall p \in \mathcal{B}(z_i), \exists z_j \in O : \text{more\_specific}(z_j, z_i) \ \wedge \ p \in \mathcal{G}(z_j)$$

In practice, constraint $\mathcal{C}_1$ needs to be relaxed because, as a result of the filtering method described above (see Section 2.4.3), there may not be a subset $O \subset Z$ satisfying it, i.e., the minimisation problem may not have a solution because it is impossible to reproduce all the bilingual phrase pairs regardless of the set of GATs chosen. This occurs when the highly lexicalised GATs that would be needed to reproduce certain bilingual phrase pairs have been removed and there is a conflict between the less specific GATs that are able to reproduce them. When this happens, the set of bilingual phrase pairs P is replaced by its subset $P_O \subset P$ that maximises $\sum_{p \in P_O} \text{count}(p)$ and makes the minimisation problem solvable, i.e., that permits finding a set of GATs that meets the constraints $\mathcal{C}_1$ and $\mathcal{C}_2$.

There may also be multiple solutions to the minimisation problem, i.e., different sets of GATs with the same (minimum) size may satisfy the two constraints above. In this case, the set of GATs containing the most general GATs is chosen. This is

[14] Another option would be to compare the sets of bilingual phrase pairs matched by each GAT and consider z as more specific than z' if $\mathcal{M}(z) \subset \mathcal{M}(z')$. However, when the training corpus is small (only a few hundred sentences), it may occur that z and z' match the same set of bilingual phrase pairs in spite of z' being more general than z because it has the potential to match more sequences of lexical forms when translating new texts.

done by defining a function spec_level(z)[15] that accounts for the level of specificity of a GAT z (see below), computing the aggregated level of specificity of the possible solutions to the minimisation problem, $\sum_{z \in O} \text{spec\_level}(z)$, and choosing the set with the smaller aggregated level of specificity as the solution. The level of specificity of a GAT z is simply obtained by counting the number of lexicalised words and the number of morphological inflection attributes in the SL word classes with non-wildcard values:

$$\text{spec\_level}(z) = \gamma_1 |\{s_i : s_i \in S \wedge \lambda(s_i) \neq \epsilon\}| + \gamma_2 \sum_{s_i \in S} |\{a : \upsilon(s_i, a) \neq *\}| + 1.$$

The first two terms in the equation above are assigned a weight so that lexicalised word classes have a higher impact on the final result than the morphological inflection attributes with non-wildcard values. This is achieved by making $\gamma_2 = 1$ and γ_1 higher than the highest possible value of the second term, that is, $\gamma_1 = \sum_{s_i \in S} |\alpha(s_i)| + 1$, since, in practice, a different minimisation subproblem is solved for each sequence of SL lexical categories (see below). The third term is added for convenience, to prevent spec_level(z) from returning a null value.

The minimisation problem defined is similar to the well-known *set covering* problem (Garey and Johnson, 1979), which is NP-hard (Korte and Vygen, 2012, Sec. 15.7). Despite its complexity, it can be solved in a reasonable amount of time when the quantity of bilingual phrase pairs and GATs is relatively small —a common situation when the amount of training parallel corpora is scarce— by splitting the problem into independent sub-problems: one for each different sequence of the SL lexical categories. The resolution of each minimisation subproblem is described next.

2.4.4.2 Solving the minimisation problem with integer linear programming

Each minimisation sub-problem is formulated as an integer linear programming problem (Garfinkel and Nemhauser, 1972) which allows the use of existing methods to solve it. This kind of problems involves the optimisation (maximisation or minimisation) of a linear objective function subject to linear inequality constraints. In the experiments, the state-of-the-art *branch and cut* approach (Xu et al., 2009) has been followed. An integer linear programming problem has the following general form:

- optimise $\sum_{i=1}^{n} c_i x_i$
- subject to m constraints: $\sum_{i=1}^{n} a_{ij} x_i \geq b_j$ with $j = 1, ..., m$
- where $x_i \in \mathbb{Z} \ \ \forall i \in [1, n]$.

[15] Note the difference between more_specific($\cdot$) and spec_level($\cdot$). more_specific($\cdot$) defines a strict partial order in which two GATs are related if, and only if, the set of bilingual phrase pairs matched by one of them is a subset of the set of bilingual phrase pairs matched by the other. This makes the solution of the minimisation problem look like a hierarchy with general rules and more specific rules fixing the cases not correctly translated with the general ones. Contrarily, spec_level($\cdot$) simply permits selecting from among different solutions with the same amount of GATs.

In order to reformulate the minimisation problem defined in the previous section using integer linear programming inequations, two sets of integer variables are defined: X and Y. The set of integer variables X is associated with the GATs in the set of GATs Z obtained after performing the filtering described in Section 2.4.3 such that $x_i \in X$ equals 1 if the GAT $z_i \in Z$ is part of the solution set O, zero otherwise. The set of integer variables Y is associated with the bilingual phrase pairs $p_j \in P$ so that y_j equals 1 if $p_j \notin P_O$, i.e. if it has been removed to make the minimisation problem solvable.

The function to be minimised (optimised) is defined as:

$$\left(\sum_{i=1}^{|X|} x_i\right) + \left(\sum_{i=1}^{|X|} x_i \cdot \frac{\text{spec\_level}(z_i)}{\max(\{\text{spec\_level}(z_j) : z_j \in Z\})} \cdot \frac{1}{|Z|}\right) + \left(\sum_{j=1}^{|P|} y_j \cdot \text{count}(p_j) \cdot T\right)$$

where spec_level(z_i) computes the level of specificity of GAT z_i (see the equation on Section 2.4.4.1), count(p_j) is the frequency of the bilingual phrase pair p_j in the parallel corpus, and T is a penalty whose value is set to $|Z| + 2$ (see below).

The first term in the equation above counts the number of GATs in Z that are part of the solution set O. The second term is introduced to discriminate between different solution sets with the same number of GATs (recall that, when there are multiple solution sets with the same (minimum) number of GATs, that with the lowest aggregated specificity level is chosen). Here $\frac{1}{|Z|}$ is introduced to ensure that the second term only discriminates between different solutions sets with the same number of GATs and that it does not promote solution sets with a large amount of GATs but a low level of specificity.[16]

The third term counts the number of occurrences in the training corpus of the bilingual phrase pairs that need to be discarded to make the minimisation problem solvable. Here a penalty T is introduced to ensure that only the minimum amount of bilingual phrase pairs needed to make the minimisation problem solvable are removed, i.e. not included in P_O; otherwise we could be removing bilingual phrase pairs not because they cannot be reproduced but because by removing them, the GATs reproducing them could also be removed, thus reducing the size of O. The value of T is set to $|Z| + 2$ because it is the lowest possible value which guarantees that this term is greater than the sum of the first and the second terms when deciding whether or not to remove a bilingual phrase that can be correctly reproduced and by removing it the amount of GATs is also reduced.[17]

The inequations representing the constraints to the minimisation problem are as follows:

[16]Note that the value of the second term is always in the range $[0, 1]$.

[17]Note that the sum of the first two terms of the expression is always less than or equal to $|Z| + 1$, and the third term is always greater than or equal to T.

- $\mathcal{C}_1$: There may exist at least one GAT in O that reproduces each bilingual phrase pair in P_O:

$$\forall_{p \in P} \sum_{i:p_k \in \mathcal{G}(z_i)} x_i + y_k \geq 1$$

  Note that, in order to check this constraint, the loop iterates over all bilingual phrase pairs P, not over P_O, and that if a bilingual phrase pair p_k is not in P_O, the constraint is met because $y_k = 1$.

- $\mathcal{C}_2$: For each bilingual phrase pair in P_O matched but not correctly reproduced by a GAT z_i, either z_i is not part of the solution or there is at least one more specific GAT that is part of the solution and correctly reproduces it:

$$\forall_{i \in [1,|Z|]} \forall_{p_k \in \mathcal{B}(z_i)} \sum_{j:j \neq i \wedge p \in \mathcal{G}(z_j)} \Lambda_{ji} x_j + y_k \geq x_i$$

  where Λ_{ji} maps the output of the function more_specific (see Section 2.4.4 on page 56) to values 0 or 1:

$$\Lambda_{ji} = \begin{cases} 1 & \text{if more\_specific}(z_j, z_i) \\ 0 & \text{otherwise.} \end{cases}$$

  As before, if p_k is not part of P_O the constraint is met because $y_k = 1$.

2.4.5 Optimising rules for chunking

The problem of selecting the minimum set of GATs that are needed to reproduce all the bilingual phrase pairs obtained from the training parallel corpus has been independently solved for each sequence of SL lexical categories. However, several GATs are used in the translation of an SL sentence, and each one translates a different sequence of SL lexical categories. The segmentation of the input SL sentences into chunks (sequences of SL lexical forms) is done by the GATs to be applied, which are chosen by the engine in a greedy, left-to-right, longest match fashion. It is therefore necessary to avoid the situation of having lexical forms that should be processed together —because they are involved in the same linguistic phenomenon— being assigned to different chunks.

This section describes the process carried out in order to select the subset of the set of GATs obtained after solving the minimisation problem that ensures that the text to be translated will be chunked in the most convenient way. The sequences of SL lexical categories that GATs must contain in order to be part of the final solution are selected; to do this, a greedy approach is followed. It attempts to maximise the similarity between the TL side of the training parallel corpus and the result of translating its SL side using GATs in the same way as the RBMT engine would do. The method first identifies the minimum set of SL text segments (*key segments*) in the training corpus which need to be translated by a rule to obtain the highest similarity. Afterwards, the sequences of SL categories that ensure that the maximum number of key segments get translated properly are selected.

Identifying key segments Let $\mathcal{K}$ be the set containing all the possible sets of text segments in the SL sentences of the training corpus,[18] and $\mathcal{K}^\star \subseteq \mathcal{K}$ be the set of sets of text segments that maximise the *similarity* between the TL side of the training corpus and the translation obtained by translating each text segment in $K \in \mathcal{K}^\star$ with the most specific GAT available (as the RBMT engine would do) and the rest of the SL words in the training corpus word for word by looking them up in the bilingual dictionary. Here, similarity may be computed by using any standard MT evaluation measure.

The set of key text segments $\mathcal{I}$ is one of the sets in $\mathcal{K}^\star$. As $\mathcal{K}^\star$ may contain more than a single set, $\mathcal{I}$ is chosen so that it satisfies the following two conditions:

1. $\mathcal{I}$ is one of the sets with the fewest and shortest segments, i.e., with the minimum number of words covered by segments:

$$\mathcal{I} \in \arg\min_{K \in \mathcal{K}^\star} \sum_{k \in K} |k|,$$

   where $|x|$ denotes the number of words of text segment x.

2. $\mathcal{I}$ is one of the sets with the minimum average segment length:

$$\mathcal{I} \in \arg\min_{K \in \mathcal{K}^\star} \frac{\sum_{k \in K} |k|}{|K|}$$

These two conditions give priority to short text segments, and therefore to short GATs, over longer ones, in addition to the use of as few GATs as possible. If more than one set satisfies these two conditions, $\mathcal{I}$ is chosen at random from among them.

As exploring the whole set $\mathcal{K}$ is computationally infeasible, in practice $\mathcal{I}$ is obtained by processing one parallel sentence at a time and following a dynamic programming approach similar to the *beam search* approach used for decoding in SMT (Koehn, 2004a).[19]

Note that when computing the set of key text segments $\mathcal{I}$, two text segments consisting of the same sequence of words are considered different if they appear in different positions in the corpus. This is also applicable to the description provided as follows.

Selecting the sequences of lexical categories The sequences of lexical categories that GATs must contain in order to be part of the final solution are chosen from among

[18] Note that the text segments in $K \in \mathcal{K}$ do not overlap and do not necessarily cover all the words in the corpus.

[19] Note that, despite the fact that the key text segments are computed independently for each sentence, it is highly unlikely that the addition of a new sentence substantially affects the solution because all the key segments are considered together when selecting the sequence of lexical categories for which rules will be generated.

the set $\mathcal{L}$ with the candidate sequences of lexical categories, which are in turn obtained from the words in the set of key text segments:

$$\mathcal{L} = \bigcup_{g \in \mathcal{I}} \{(\rho(w_i))_{i=1}^{|g|}\}.$$

For each sequence $l \in \mathcal{L}$, a score seq_qa(l) is computed. This score measures the impact on the translation quality of having rules matching the sequence of lexical categories l:

$$\text{seq\_qa}(l) = \frac{|\text{key\_seg\_ok}(l)|}{|\text{key\_seg\_ok}(l)| + |\text{key\_seg\_broken}(l)|},$$

where key_seg_ok(l) is the set of key text segments correctly translated by a rule matching the sequence of lexical categories l; and key_seg_broken(l) is the set of key text segments not correctly translated by a rule matching l plus the set of key text segments whose words are not translated together by the same rule as a consequence of having a rule matching l.

On the one hand, key_seg_ok(l) is defined as:

$$\begin{aligned}\text{key\_seg\_ok}(l) = \{g : g \in \mathcal{I} \ \wedge \ (((\rho(w_i))_{i=1}^{|g|} = l) \ \vee \\ (\exists g' : g \in \text{seg}(g') \ \wedge \ (\rho(w_i))_{i=1}^{|g'|} = l \ \wedge \ \exists K \in \mathcal{K}^\star : g' \in K))\},\end{aligned}$$

where seg(x) is the set of all possible (sub)segments of text segment x. The text segments returned by key_seg_ok(l) are the key text segments ($g \in \mathcal{I}$) with a sequence of lexical categories l ($(\rho(w_i))_{i=1}^{|g|} = l$), and the key text segments contained in longer segments ($\exists g' : g \in \text{seg}(g')$) with a sequence of lexical categories l ($(\rho(w_i))_{i=1}^{|g'|} = l$) and correctly translated by a GAT ($\exists K \in \mathcal{K}^\star : g' \in K$).

On the other hand, key_seg_broken(l) is defined as:

$$\begin{aligned}\text{key\_seg\_broken}(l) = \{g : g \in \mathcal{I} \ \wedge \ ((\exists g' : g \in \text{seg}(g') \ \wedge \ (\rho(w_i))_{i=1}^{|g'|} = l \ \wedge \\ \nexists K \in \mathcal{K}^\star : g' \in K \ \wedge \ \exists z \in O : \text{match}(z, g')) \ \vee \\ (\exists g'' : (\rho(w_i))_{i=1}^{|g''|} = l \ \wedge \ \exists z \in O : \text{match}(z, g'') \ \wedge \\ \text{start}(g'') < \text{start}(g) \ \wedge \ \text{end}(g'') < \text{end}(g) \ \wedge \ \text{end}(g'') \geq \text{start}(g)))\}\end{aligned}$$

where start(x) and end(y) refer to the position in the corpus of the first word of text segment x and the last word of text segment y, respectively; match(z, x) equals true if the GAT z matches the sequence of SL lexical forms of text segment x, otherwise zero. The text segments returned by key_seg_broken(l) are the key text segments ($g \in \mathcal{I}$) contained in longer segments ($\exists g' : g \in \text{seg}(g')$) with a sequence of lexical categories l ($(\rho(w_i))_{i=1}^{|g|} = l$), matched by at least one GAT ($\exists z \in O : \text{match}(z, g')$) and not correctly

translated by any of the GATs matching it ($\nexists K \in \mathcal{K}^{\star} : g' \in K$). It also returns the key text segments which are intersected on the left by another text segment g'' with a sequence of lexical categories l ($\exists g'' : (\rho(w_i))_{i=1}^{|g''|} = l$) and matched by at least one GAT ($\exists z \in O : \text{match}(z, g'')$). Note that any text segment intersecting on the left with a key text segment g and matched by a GAT prevents the words in g from being translated together by the same GAT. This happens, for instance, in the example presented at the top of Figure 2.6 (see page 40) for the sentence *The white house and the red cars*: the GAT applied to the chunk *The white house and the* prevents the words in the chunk *the red cars* from being processed together by the GAT that would perform the gender and number agreement that is needed to produce a correct translation of that sentence into Spanish.

A subset of the set of GATs O obtained as a result of the minimisation step described in Section 2.4.4 is then selected as follows:

$$O_{\text{sel}} = \{z = (S, T, A, R) : z \in O \wedge (\rho(s))_{s \in S} \in \{l : l \in \mathcal{L} \wedge \text{seq\_qa}(l) \geq \mu\}\}$$

Where μ is a threshold whose value is automatically determined by trying all its possible values[20] and choosing that which maximises the *similarity* of the TL side of the training corpus and the translation obtained when its SL sentences are translated with the set of GAT O_{sel}. Note that not all GATs in O_{sel} will eventually be used to generate shallow-transfer rules, since some of them may be discarded as a result of the next step.

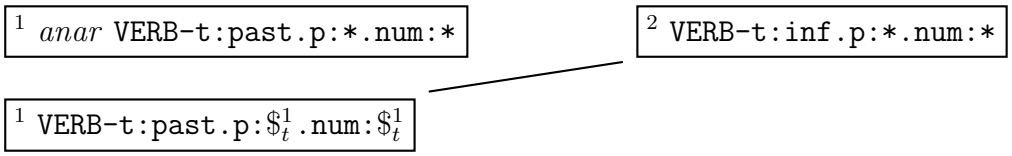
Removing redundant generalised alignment templates The number of GATs can be further reduced without decreasing the translation performance by removing those GATs which produce the same translations that a set of shorter GATs would produce. Let us suppose that GAT z produces the translation W' when applied to the SL segment W. It often occurs that, when removing z from the set of GATs of the RBMT system, the engine still produces W' when translating W. This may occur because the RBMT system splits W into two or more chunks and the translation of these chunks by the matching GATs yields W', because the word for word translation of W produces W' as a result, or because of a combination of these two reasons. If this occurs for all the SL segments that match z, then z can be safely removed from the set of GATs from which rules will be generated because it is redundant, i.e., the information in z is already contained in other GATs. Removing these longer GATs has actually improved the translation performance. Since long GATs are learnt from fewer examples and the useless ones are removed, shorter, more reliable GATs are applied.

In order to detect and remove these redundant GATs, the following process is carried out. First, the GATs in O_{sel} are sorted in order of decreasing length, while GATs of

[20] Actually, all the possible vales of μ do not need to be tested since those that generate the same set O_{sel} will produce the same result.

Figure 2.13: GAT encoding the translation from Catalan into Spanish of the verb *anar* in the past tense followed by a verb in infinitive mood.

z:



$r_1 = \{\texttt{t} : \texttt{past}\}$, $r_2 = \{\texttt{t} : \texttt{inf}\}$

the same length are sorted by increasing level of specificity.[21] For each GAT z, the bilingual phrase pairs correctly reproduced by it, $\mathcal{G}(z)$, are then collected, and each bilingual phrase pair $p \in \mathcal{G}(z)$ is checked in order to ascertain whether or not, when translating the SL side of p with the set of GAT $O_{\text{sel}} - \{z\}$, its TL side is obtained. If this requirement is met for all $p \in \mathcal{G}(z)$, z is definitively removed from O_{sel}, i.e., $O_{\text{sel}} \leftarrow O_{\text{sel}} - \{z\}$. It is therefore possible to guarantee that, after removing redundant GATs, the TL side of each bilingual phrase pair can be safely reproduced with the GATs that remain in O_{sel}.

For example, the Catalan–Spanish GAT z_1 in Figure 2.12 (see page 54) could be safely removed from the set of GATs O_{sel} if the GAT in Figure 2.13 is also part of O_{sel} because the presence of an adverb before the Catalan verb *anar* does not change the way in which the verb *anar* in the past tense followed by a verb in infinitive mood is translated. All the bilingual phrase pairs matching z_1 in Figure 2.13 can thus be reproduced by translating the adverb in isolation, i.e., by looking it up in the bilingual dictionary, and applying the GAT in Figure 2.13 to the other two lexical forms.[22]

2.4.6 Generation of Apertium shallow-transfer rules

Finally, the GATs resulting from the application of all of the above steps are converted into the rule format of the Apertium RBMT engine so that they can be used in real-world translation tasks. A list containing all the GATs and compatible with the strict partial order defined by the function more_specific(·) is built by means of a topological sorting algorithm (Kahn, 1962). This list contains the resulting GATs sorted in

[21] The sorting is based on the function more_specific, defined in Section 2.4.4. GATs of the same length are arranged in a list compatible with the strict partial order defined by the function more_specific(·) by means of a topological sorting algorithm (Kahn, 1962).

[22] The proportion of GATs discarded because they can be replaced by shorter ones varies across language pairs and training corpus sizes. Generally, larger training corpora and more distant language pairs involve fewer GATs discarded. For instance, in the experiments described in Section 2.5, 75% of the Catalan–Spanish GATs with 5 SL lexical forms were discarded when the training corpus contained 250 sentences, while the proportion dropped to 41% for the training corpus with 5 000 sentences. When Spanish–English rules were inferred from the training corpus with 5 000 sentences, only 14% of the GATs with 5 SL lexical forms were discarded.

decreasing order of specificity and is used when generating the rules so that the most specific GAT is always applied when different GATs match the same input sequence of lexical forms. Figure 2.14 shows a fragment of an Apertium shallow-transfer rule that encodes the structural transformation provided by the GAT shown in Figure 2.13. The reader is referred to Section A.3.2.1 for the details of this conversion.

2.5 Experimental settings

The new rule inference approach has been evaluated by comparing the number of GATs extracted and the resulting translation quality with those obtained by (a) following the method proposed by Sánchez-Martínez and Forcada (2009), (b) using hand-crafted rules, and (c) using no rules at all (word-for-word translation). In order to assess the contribution of the different methods that are used in the new approach in order to improve translation quality, the impact of the following changes has also been evaluated: (d) wildcards and reference values are not used when creating word classes (i.e. the function σ_2 described in Section 2.4.2.3 returns the input set of GATs unchanged); and (e) the approach by Sánchez-Martínez and Forcada (2009) benefits from the method described in Section 2.4.5 when selecting the final set of rules, which ensures a convenient chunking of the input. The translation performance of the combination of hand-crafted rules and rules automatically inferred with the approach presented in this chapter has been also tested.

The evaluation covers a wide variety of language pairs: pairs in which the two languages involved in the translation belong to the same language family (Spanish↔Catalan; the arrows mean that both translation directions are followed), in addition to pairs in which the languages belong to different language families (English↔Spanish and Breton–French).

The Spanish↔Catalan training corpus consists of parallel sentences extracted from the newspaper *El Periodico de Catalunya*,[23] which is published in both languages; the test corpus consists of sentences randomly chosen from the *Revista Consumer Eroski* parallel corpus (Alcázar, 2005), which contains product reviews. The English–Spanish rules have been inferred from the Europarl Parallel Corpus (Koehn, 2005) version 7, a collection of minutes from the European Parliament, and they have been evaluated with the *newstest2013* corpus, a set of parallel sentences extracted from pieces of news and released as part of the shared translation task of the eighth Workshop on Statistical Machine Translation (Bojar et al., 2013). The Breton–French training and test corpora have both been randomly extracted from the collection compiled by Tyers (2009) from a heterogeneous set of sources, including software localisation and tourism.

In order to evaluate the impact of the size of the training corpus on the quality of the resulting translations, subsets containing 100, 250, 500, 1 000, 2 500 and 5 000

[23] `http://www.elperiodico.com/`

Figure 2.14: Fragment of an Apertium shallow-transfer rule that encodes the structural transformation provided by the GAT shown in Figure 2.13 for the translation of the Catalan verb *anar* in the past tense followed by a verb in infinitive mood into Spanish. The first `equal` element checks whether the lemma of the first matching SL lexical form is *anar*, while the two `equal` elements after it check the values of SL morphological inflection attributes. The last two `equal` elements check whether the TL restrictions of the GAT are met. Finally, the `out` XML element defines the output of the rule. The first `clip` element indicates that the lemma is obtained after translating the second matching lexical form with the bilingual dictionary. The remaining `clip` elements in the same line represent the TL reference attributes of the GAT, while the `lit-tag` element is used to explicitly define the lexical category and the value of the *tense* attribute. As GATs that match the same sequence of SL lexical categories are grouped together in the same Apertium shallow-transfer rule, the verification of the lexical category of the matching lexical forms is not performed by this fragment of code. A more detailed description of the process followed for encoding GATs as Apertium shallow-transfer rules can be found in Section A.3.2.1.

```
...
<when>
  <test><and>
    <equal>
      <clip pos="1" side="sl" part="lemma" />
      <lit v="anar"/>
    </equal>
    <equal>
      <clip pos="1" side="sl" part="tense" />
      <lit-tag v="past"/>
    </equal>
    <equal>
       <clip pos="2" side="sl" part="tense" />
       <lit-tag v="inf"/>
    </equal>
    <equal>
      <clip pos="1" side="tl" part="tense" />
      <lit-tag v="past"/>
    </equal>
    <equal>
      <clip pos="2" side="tl" part="tense" />
      <lit-tag v="inf"/>
    </equal>
  </and></test>
  <out>
    <lu><clip pos="2" side="tl" part="lemma"/><lit-tag v="verb.past"/
       ><clip pos="2" side="tl" part="person"/><clip pos="2" side="tl
       " part="number"/></lu>
  </out>
</when>
...
```

sentences have been randomly extracted from each training corpus in a such a way that all the sentences in the smaller subsets are contained in the bigger ones.[24] For English↔Spanish and Breton–French, two additional subsets containing 10 000 and 25 000 sentences respectively have also been used to evaluate the new approach when no generalisation of the morphological inflection attributes is performed (i.e. no wildcard and reference values are used). Each corpus subset has then been split into two parts: the largest one, containing $\frac{4}{5}$ of the sentences, has been used as the actual training subset from which GATs are extracted, whereas the remaining sentences have been used as the development set to determine the threshold values to be used with each method (see below).[25] Table 2.1 provides the number of sentences in the training and development corpora, the number of words and the size of the vocabulary for each language pair and corpus size; Table 3.1 provides these data for the different test sets used for evaluation.[26]

Regarding the threshold used by each method, Sánchez-Martínez and Forcada (2009) use a threshold to discard the EATs that reproduce a number of bilingual phrase pairs below its value; this threshold is obtained as the integer value between 1 and 10 which maximises the BLEU score (Papineni et al., 2002) in the development corpus. The new approach uses two different thresholds, θ and δ, as described in Section 2.4.3. The value of δ (used to discard those GATs with an inadequate ratio of bilingual phrase pairs correctly reproduced to the total number of bilingual phrase pairs matched) has been chosen by trying all the values in the range $[0, 1]$ at increments of 0.05 and selecting the value that maximises the BLEU score in the development set. With regard to θ (used to discard GATs that reproduce a small number of bilingual phrase pairs), different values have been used, one for each different minimisation subproblem (one subproblem per sequence of SL lexical categories), to ensure that the number of input GATs to each of the different minimisation subproblems is below 1 000; in any case a minimum value of 2 has been established for θ to discard those GATs that are only able to reproduce a single bilingual phrase pair. This is done to make the minimisation problem computationally feasible. For the experiments where wildcards and reference values are not used, the value of δ has been optimised in the same way, while the value of θ has been always set to 2 because the computational complexity of the minimisation problem is much smaller.

With respect to the similarity measure used to optimise the rules for chunking and select the sequences of lexical categories for which rules will eventually be generated (see Section 2.4.5), the metric used is BLEU (Papineni et al., 2002) with the smoothing

[24] As it is usually done in SMT, only sentences containing at most 45 words have been chosen in order to prevent GIZA++ from truncating long sentences.

[25] For the subsets containing 25 000 sentences, the training part contains 23 000 sentences, while the development section contains the remaining 2 000 sentences.

[26] For a given language pair, the same test set has been used to evaluate the systems built with the different sizes of the training corpus.

Table 2.1: Number of sentences, number of words, and vocabulary size of the training and development corpora for each language pair and corpus size. These corpora are divided into training (4/5 of the sentences) and development (1/5 of the sentences). If a corpus contains 25 000 sentences, its training part is assigned 23 000 sentences and its development section contains the remaining 2 000 sentences.

training + development	Spanish		Catalan	
# sentences	# words	# vocabulary	# words	# vocabulary
100	1 539	789	1 597	798
250	3 830	1 684	3 969	1 685
500	7 697	2 985	7 939	2 946
1 000	15 136	5 062	15 576	4 959
2 500	37 301	9 783	38 470	9 580
5 000	73 637	15 315	75 981	14 933

(a) Spanish↔Catalan

training + development	English		Spanish	
# sentences	# words	# vocabulary	# words	# vocabulary
100	2 145	913	2 151	945
250	5 460	1 868	5 672	1 992
500	11 228	3 016	11 704	3 342
1 000	22 447	4 653	23 292	5 209
2 500	56 003	7 756	57 961	8 984
5 000	113 290	11 045	117 051	13 197
10 000	227 088	15 329	234 854	18 723
25, 000	571 364	22 703	589 400	28 927

(b) English↔Spanish

training + development	Breton		French	
# sentences	# words	# vocabulary	# words	# vocabulary
100	1 319	714	1 520	760
250	3 451	1 537	3 768	1 621
500	6 937	2, 623	7 565	2 833
1 000	14 456	4 364	15 863	4 915
2 500	35 335	7 846	37 931	9 038
5 000	69 500	11 880	75 427	14 008
10 000	141 838	17 741	153 455	20 985
25 000	354 417	28 828	387 354	34 117

(c) Breton–French

Table 2.2: Number of sentences, words, and size of the vocabulary of the test set used for evaluation for each language pair.

Language pair	# sentences	SL		TL	
		# words	# voc	# words	# voc
English–Spanish	3 000	62 873	10 867	67 762	12 400
Spanish–Catalan	3 000	76 794	13 414	78 089	13 130
Breton–French	3 000	41 800	8 824	45 278	10 211

implemented by the National Institute of Standards and Technology (NIST)[27] to avoid null values when it is used at the sentence level.

All the experiments have been carried out with the translation engine[28] and linguistic data[29] of the rule-based MT system Apertium (Forcada et al., 2011). A software package which implements the pipeline for the inference of shallow-transfer rules as described in Section 2.4 has been released (see Appendix B.1). However, some external tools have also been used, namely, the minimisation subproblems have been solved with the integer linear programming Cbc solver,[30] while word alignment and bilingual phrase pair extraction were carried out by using the Giza++ toolkit (Och and Ney, 2003) and the phrase extraction implementation in the Moses statistical MT system (Koehn et al., 2007), respectively. It is worth noting that the Apertium bilingual dictionary was added to the corpus before word alignment and removed it afterwards. This has improved word alignment when the amount of parallel sentences is scarce.

Two heuristics have been added to the method presented in Section 2.4. First, in order to limit the number of minimisation subproblems to be solved, an additional condition has been added to the set of requirements a bilingual phrase must meet for being used for rule inference (described in Section 2.4.1): the maximum number of words allowed for the SL and TL side of a bilingual phrase pair is 5. Moreover, the heuristic described below (Section 2.5.1) helps to further reduce the number of input GATs to each minimisation subproblem.

Figure 2.15 shows the number of bilingual phrase pairs obtained from the different training corpora after applying the filtering criteria described in Section 2.4.1 plus the additional filtering on phrase length. These bilingual phrase pairs have then been used to infer GATs by following the remainder of steps described in Section 2.4. The figure also depicts the proportion of bilingual phrase pairs discarded as a result of the filtering.

[27] MTeval utility version 13; `ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13.pl`.

[28] More specifically, revision 47871 of the Subversion repository at `https://svn.code.sf.net/p/apertium/svn/trunk/apertium`.

[29] Repository for English–Spanish: `https://svn.code.sf.net/p/apertium/svn/trunk/apertium-en-es`, revision 41294; Spanish–Catalan: `https://svn.code.sf.net/p/apertium/svn/trunk/apertium-es-ca`, revision 34111; Breton–French: `https://svn.code.sf.net/p/apertium/svn/trunk/apertium-br-fr`, revision 28674; Chinese–Spanish: `https://svn.code.sf.net/p/apertium/svn/incubator/apertium-zho-spa`, revision 49858.

[30] `https://projects.coin-or.org/Cbc`, version 2.7.

Note that the number of bilingual phrase pairs discarded for Spanish↔Catalan is much smaller than for the other language pairs. This is because Spanish and Catalan are closely-related languages with less lexical translation ambiguity, which signifies that more translation equivalents in the bilingual phrase pairs match those in the bilingual dictionary. In addition, Breton–French is the language pair with the highest proportion of discarded bilingual phrase pairs because its dictionaries have a low coverage, as shown in Figure 2.16.

2.5.1 Reducing the number of input generalised alignment templates to the minimisation subproblems

As explained in Section 2.4.2.3, in order to introduce wildcards and SL and TL references in the morphological inflection attributes of a GAT $z = (S, T, A, R)$, σ_2 considers the power set $\mathcal{P}(C)$ of the set C with the attributes that can be generalised in z. This could lead to a situation in which the minimisation subproblems are unsolvable in a reasonable amount of time as a result of the combinatorial explosion that occurs when generating GATs for the translation between highly inflected languages, such as some of those in the experimental settings followed in this chapter (e.g. Spanish, Catalan). In order to reduce the amount of GATs generated by σ_2 , only one subset $H_C \subset \mathcal{P}(C)$ has been considered for each GAT; this subset is defined as:

$$H_C = \{H : H \in \mathcal{P}(C) \ \wedge \ (\forall s_i \in S, \nexists a, a' : a \in H \ \wedge \ a' \notin H \ \wedge \ \text{rank}(\rho(s_i), a) < \text{rank}(\rho(s_i), a')\};$$

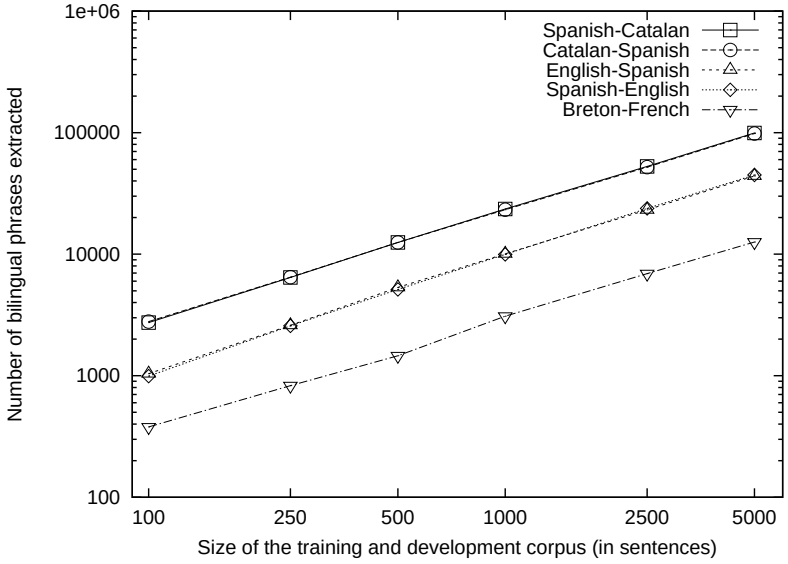
where $\text{rank}(c, a)$ returns the position of the morphological inflection attribute a in the list, ordered in decreasing order of *specificity*, of the morphological inflection attributes associated with the lexical category c. An attribute a is considered to be more specific than another attribute a' if it is applicable to a smaller number of lexical categories, e.g. the attribute verb tense is more specific than the attribute number because it can only be applied to verbs, whereas number can be applied to verbs, nouns, pronouns and (in some languages) adjectives and determiners. Therefore, for a lexical category c (e.g. verb) an attribute a (e.g. verb tense) is generalised only if the more general attributes of c (e.g. number and person) are also generalised.[31]

2.6 Results and discussion

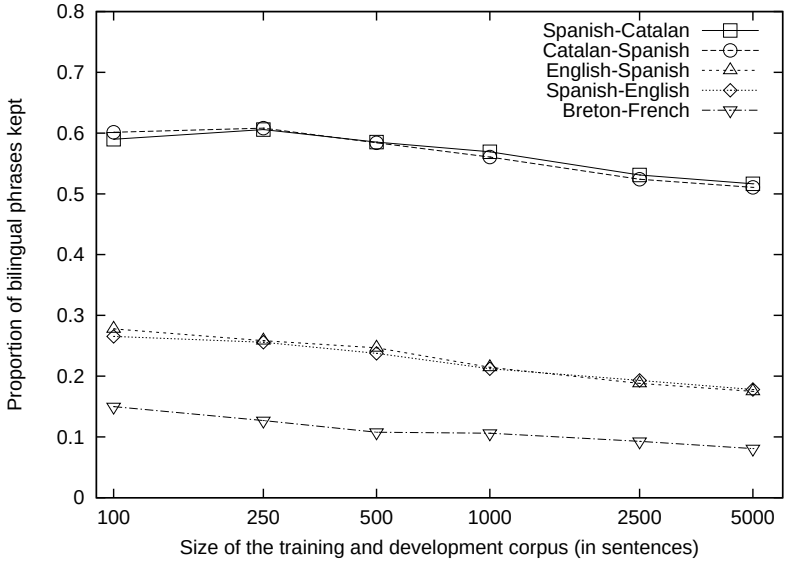
The translation quality achieved by the rules inferred when they are used with the Apertium RBMT engine, and the exact number of ATs obtained with each approach, are presented in figures 2.17–2.21. Translation quality has been estimated using the

[31] In practice H_C does not need to be explicitly calculated because the ordering provided by $\text{rank}(\cdot)$ matches that used to codify the morphological inflection attributes in the Apertium dictionaries.

Figure 2.15: Number of bilingual phrases obtained from the training corpora after applying the filtering criteria defined in Section 2.4.1 (top) and proportion of the bilingual phrases with length 5 or lower initially extracted from the parallel corpus that are kept after the filtering (bottom). The extraction of bilingual phrases is the first step of the rule inference algorithm described in Section 2.4.

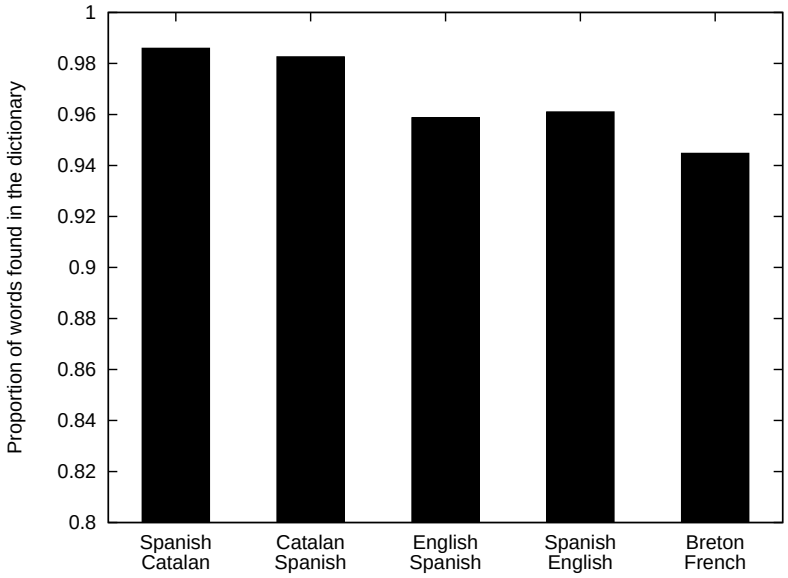


(a) Number of bilingual phrases obtained after filtering.



(b) Proportion of bilingual phrases kept after filtering.

Figure 2.16: Proportion of words in the test set for which there is at least one analysis in the Apertium dictionary, for the different language pairs.

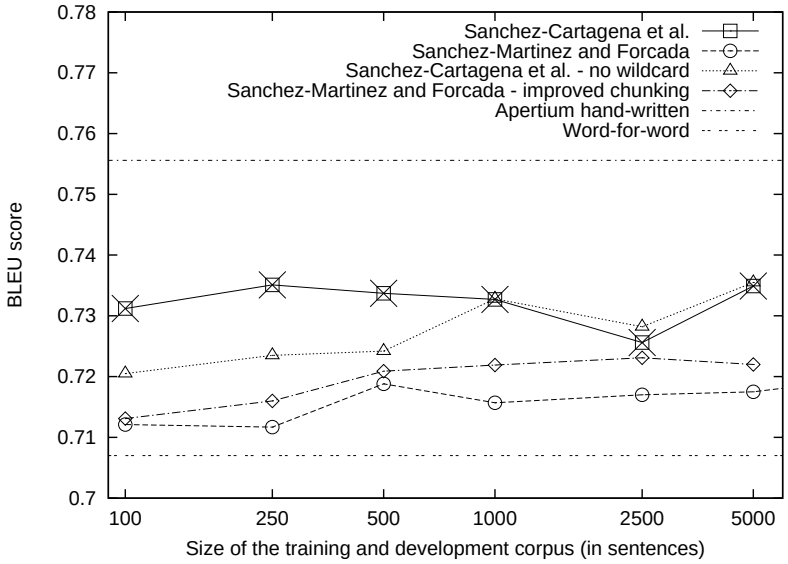


automatic evaluation metrics BLEU (Papineni et al., 2002), TER (Snover et al., 2006) (the figures represent 1-TER) and METEOR (Banerjee and Lavie, 2005). It has also been tested whether the new approach outperforms the approach proposed by Sánchez-Martínez and Forcada (2009) (henceforth, baseline approach) by a statistically significant margin through the use of paired bootstrap resampling (Koehn, 2004b) with each evaluation metric and test set ($p \leq 0.05$, 1 000 iterations); if the difference between the two approaches is statistically significant, a diagonal cross is placed on top of the points that represent the results of the approach that performs best. The figures also show the coverage provided by the rules, i.e., the proportion of words in each test set that have been translated using an AT, and the time spent on the inference of the ATs from the bilingual phrase pairs.[32]

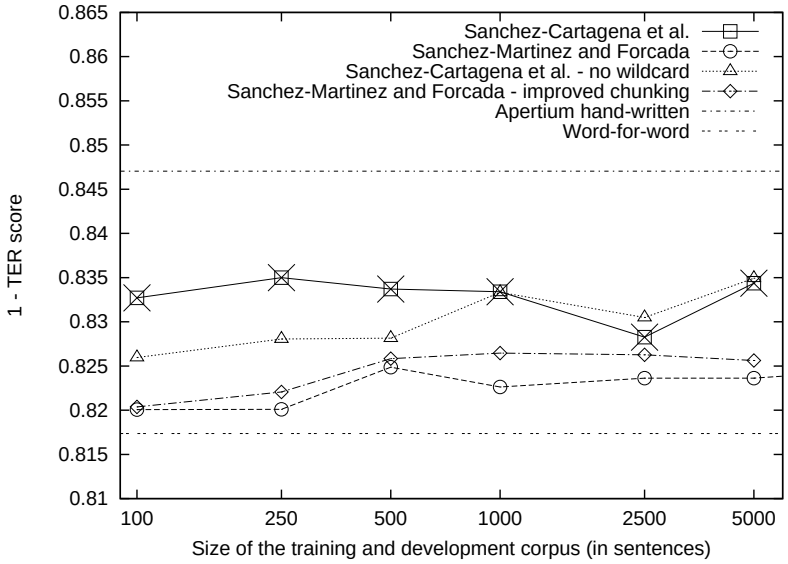
The results show that, overall, the new approach (Sánchez-Cartagena et al.) outperforms the baseline approach (Sánchez-Martínez and Forcada) by a statistically significant margin ($p \leq 0.05$) for all language pairs and automatic evaluation metrics. As expected, the translation quality of both approaches lies between the translation quality achieved by a word-for-word translation and a translation performed using hand-crafted rules. The new approach achieves results close to those obtained with hand-crafted rules and, in the case of Breton–French, it even outperforms the use of

[32]For the new approach presented in this chapter, the time is computed as the sum of the processes described in Sections 2.4.2 and 2.4.4 for the best threshold δ, since they constitute the most time-consuming part of the rule inference pipeline. For the approach by Sánchez-Martínez and Forcada (2009), the time reported is that spent on the generation of the final set of EATs from the set of bilingual phrases for the best threshold θ. The experiments have been executed in a computing cluster with 26 computing nodes with a hexacore Intel Xeon X5660 CPU each one. Times displayed are the sum of the times of the different parallel jobs.

Figure 2.17: Translation quality, number of alignment templates inferred, coverage (proportion of words in the test set translated by an alignment template) and computing time required to infer alignment templates from the different systems evaluated for the Spanish–Catalan language pair. A diagonal cross over a square point indicates that the new approach outperforms the baseline approach proposed by Sánchez-Martínez and Forcada (2009) by a statistically significant margin ($p \leq 0.05$). If the cross is over a circle, the baseline outperforms the new approach.

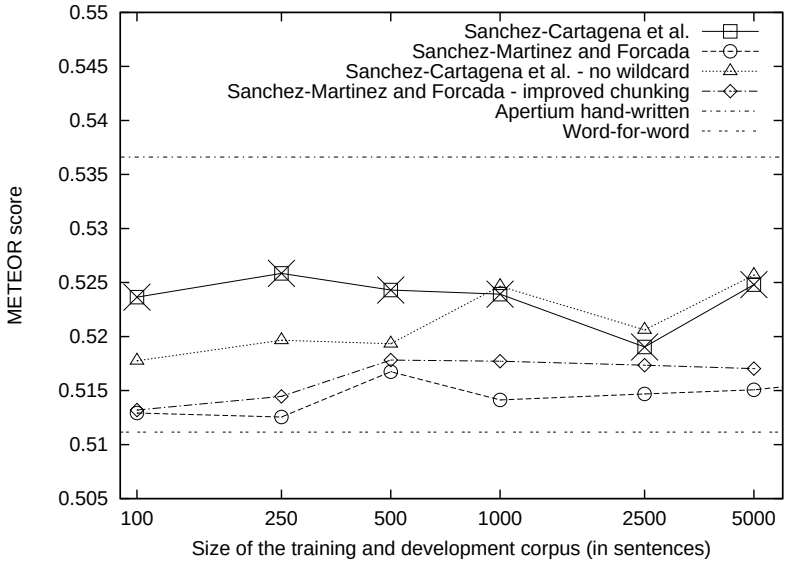


(a) Translation quality measured using BLEU.

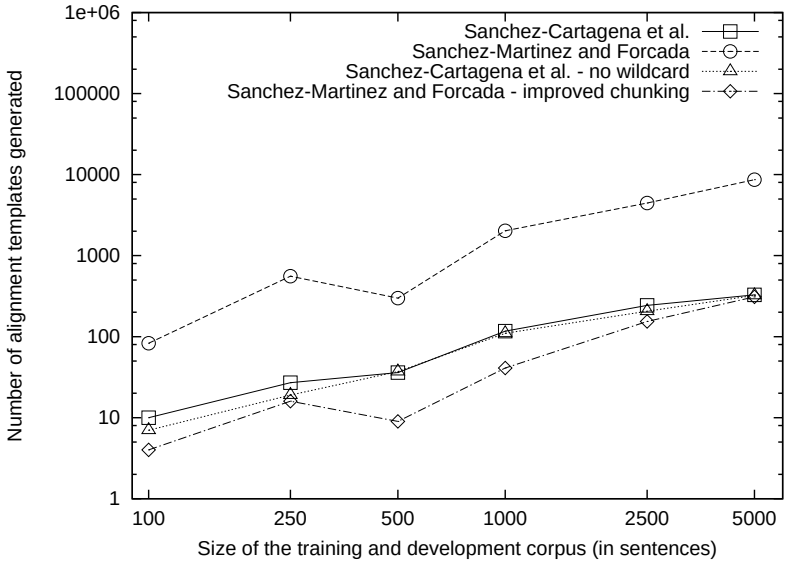


(b) Translation quality measured using TER.

(continued in next page)

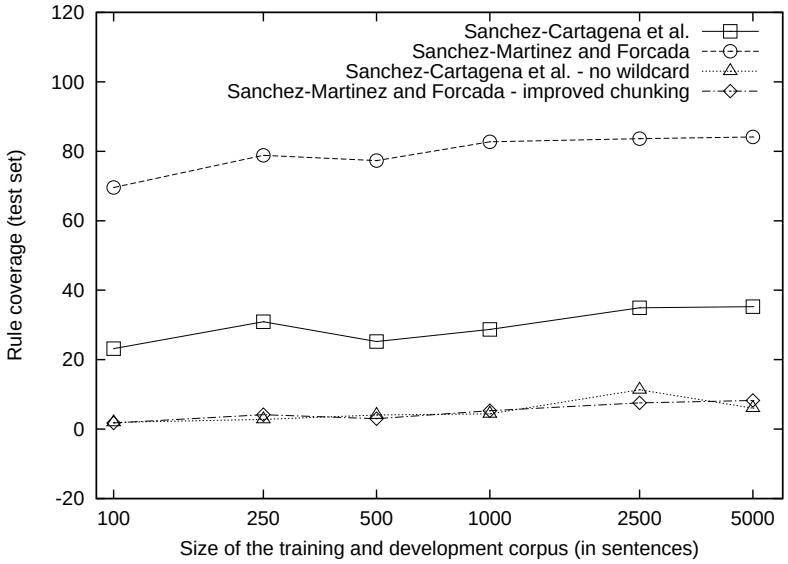


(c) Translation quality measured using METEOR.

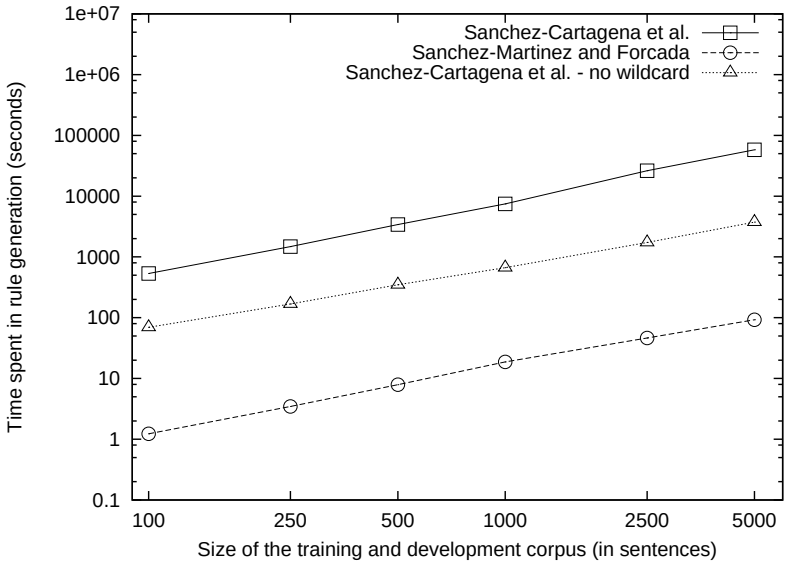


(d) Number of alignment templates inferred.

(continued in next page)

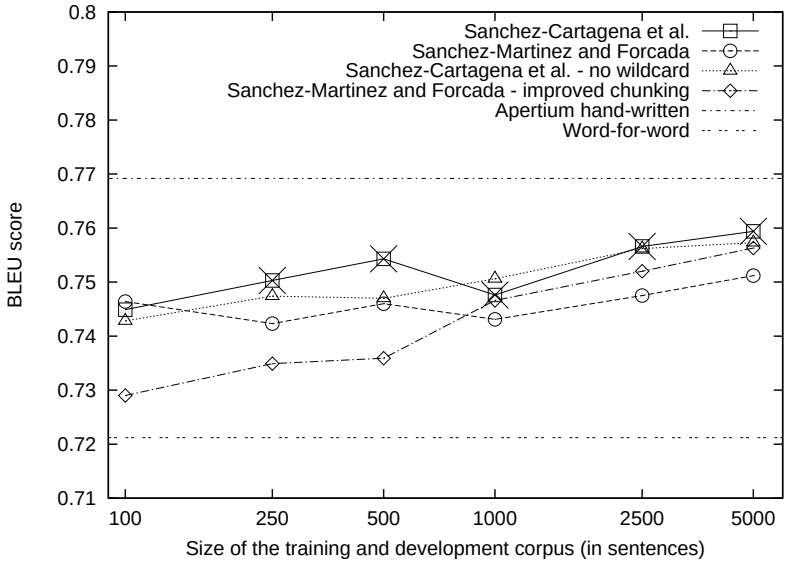


(e) Proportion of words from the test set translated by an alignment template.

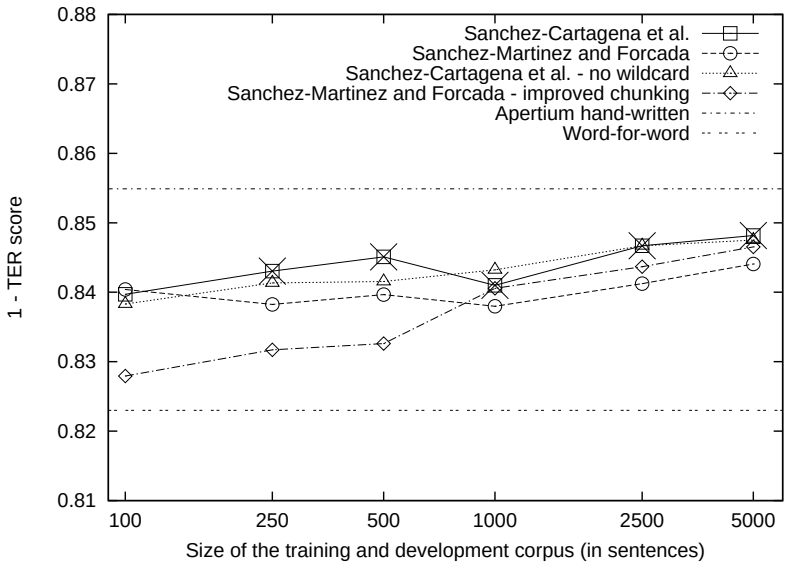


(f) Computing time required to infer alignment templates.

Figure 2.18: Translation quality, number of alignment templates inferred, coverage (proportion of words in the test set translated by an alignment template) and computing time required to infer alignment templates from the different systems evaluated for the Catalan–Spanish language pair. A diagonal cross over a square point indicates that the new approach outperforms the baseline approach proposed by Sánchez-Martínez and Forcada (2009) by a statistically significant margin ($p \leq 0.05$). If the cross is over a circle, the baseline outperforms the new approach.

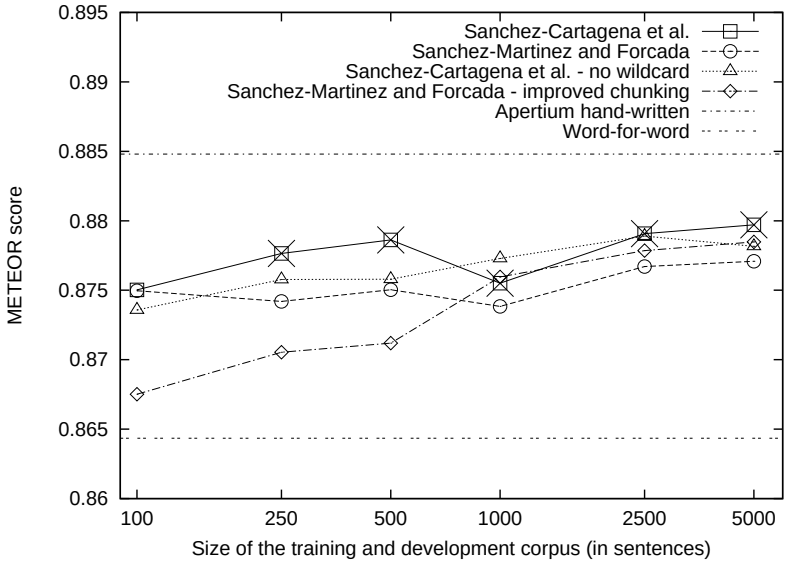


(a) Translation quality measured using BLEU.

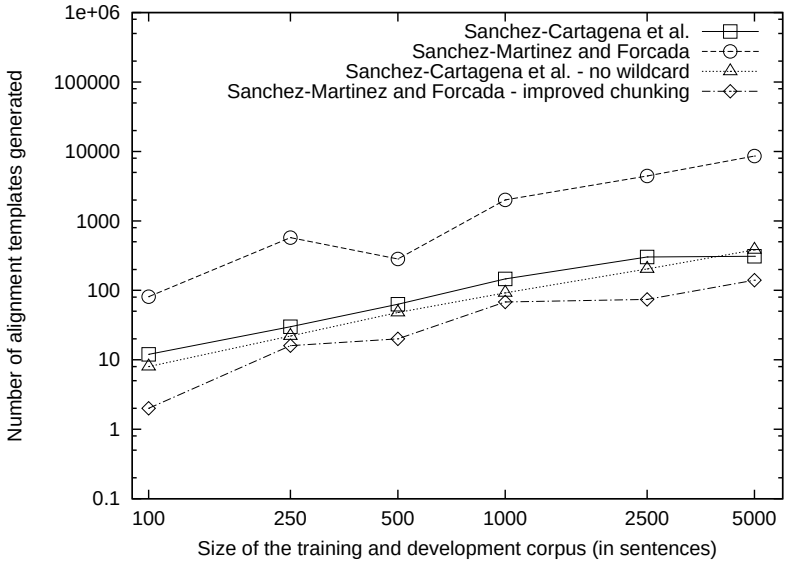


(b) Translation quality measured using TER.

(continued in next page)

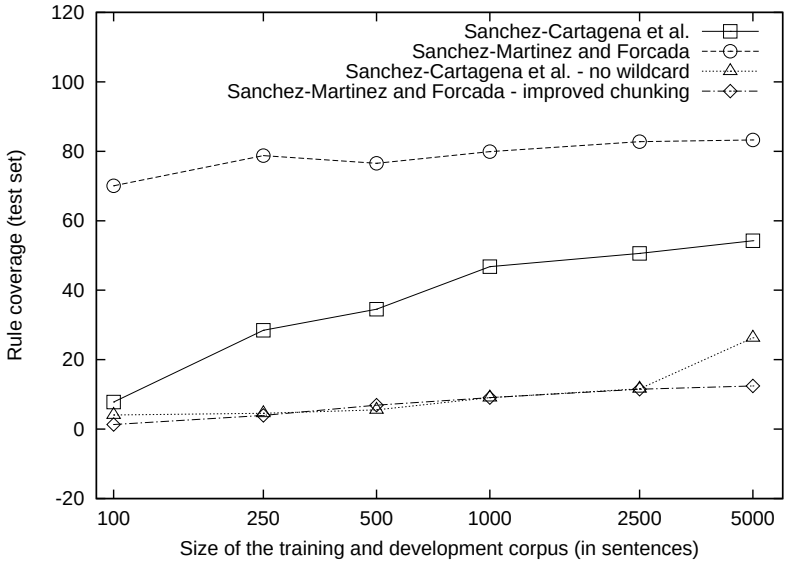


(c) Translation quality measured using METEOR.

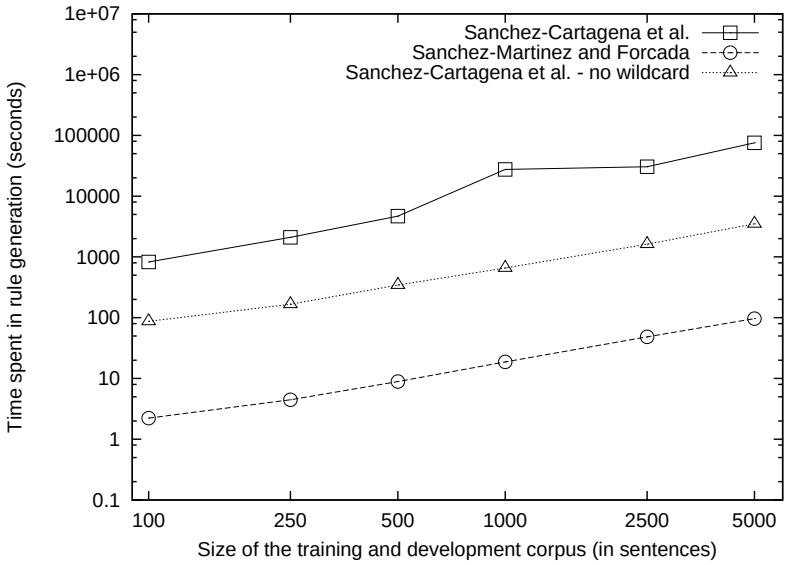


(d) Number of alignment templates inferred.

(continued in next page)

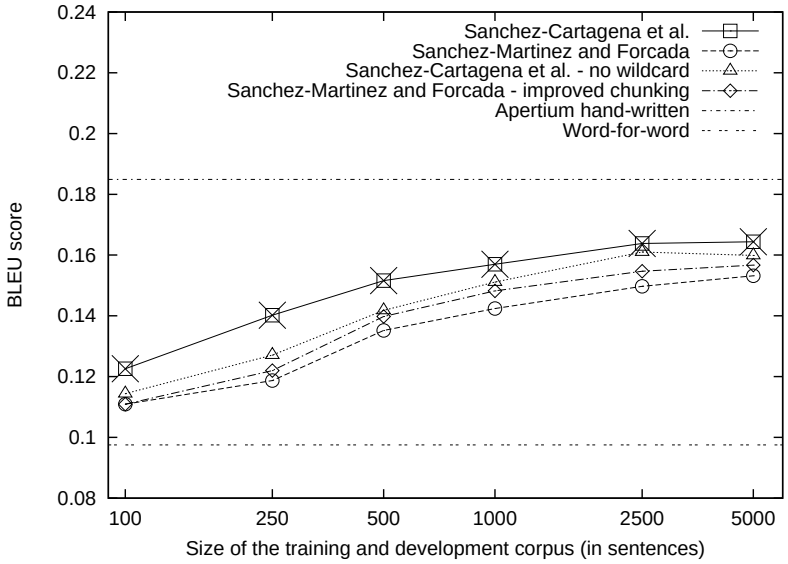


(e) Proportion of words from the test set translated by an alignment template.

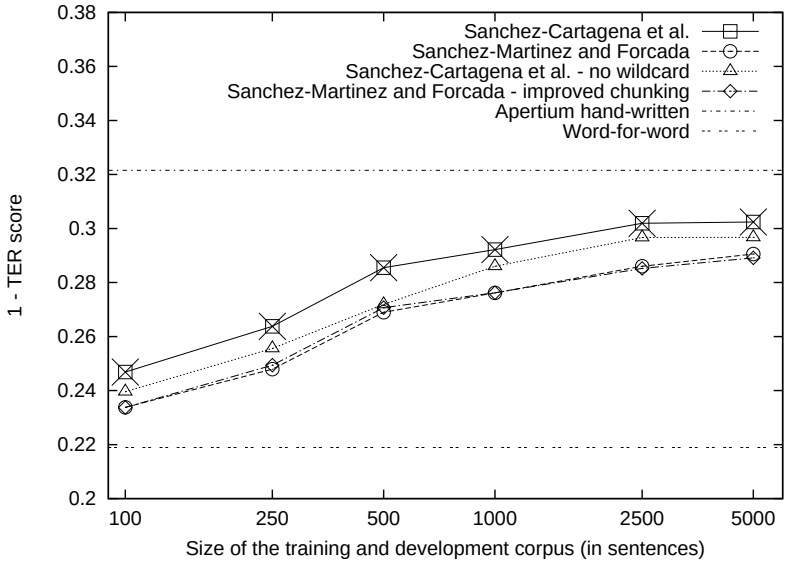


(f) Computing time required to infer alignment templates.

Figure 2.19: Translation quality, number of alignment templates inferred, coverage (proportion of words in the test set translated by an alignment template) and computing time required to infer alignment templates from the different systems evaluated for the English–Spanish language pair. A diagonal cross over a square point indicates that the new approach outperforms the baseline approach proposed by Sánchez-Martínez and Forcada (2009) by a statistically significant margin ($p \leq 0.05$). If the cross is over a circle, the baseline outperforms the new approach.

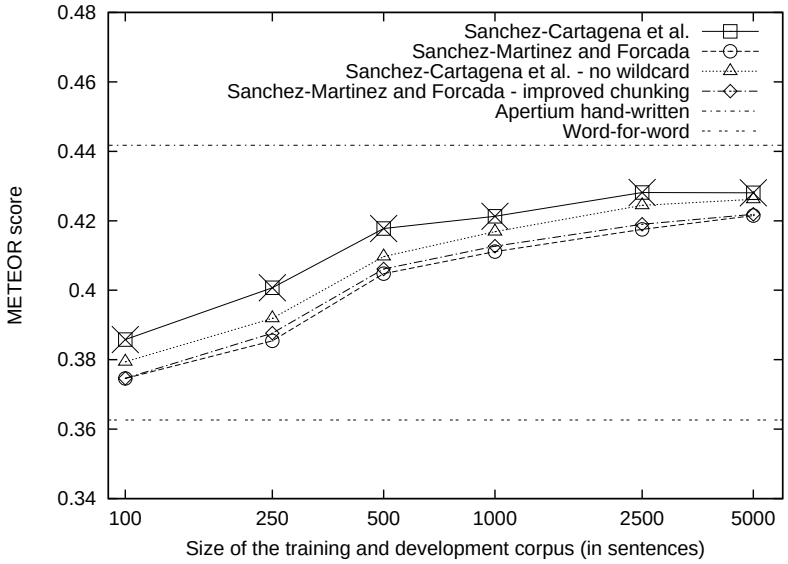


(a) Translation quality measured using BLEU.

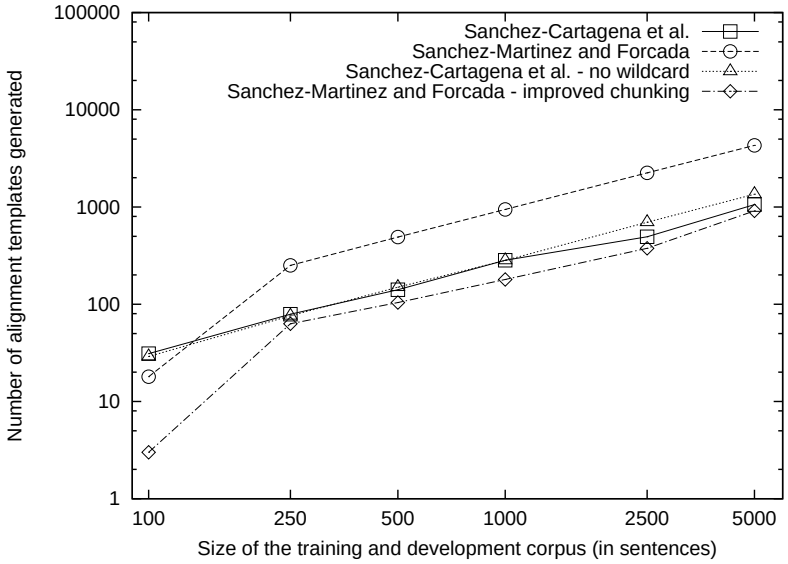


(b) Translation quality measured using TER.

(continued in next page)

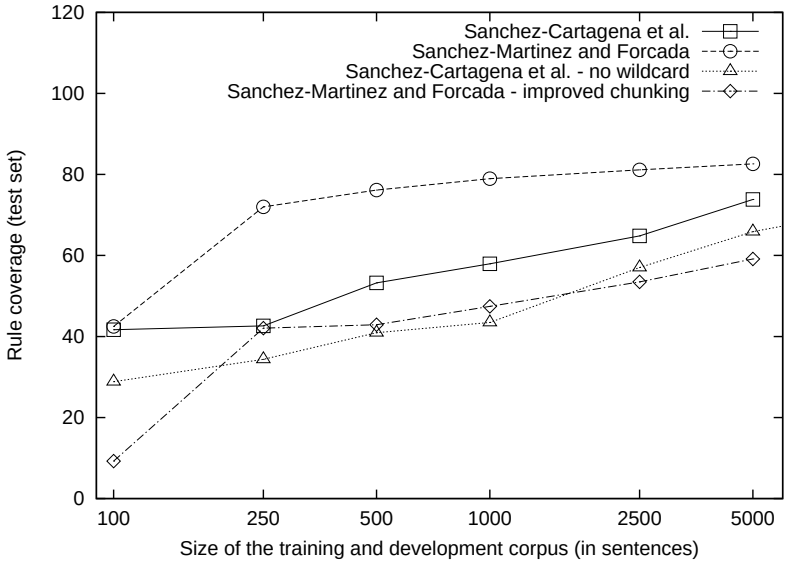


(c) Translation quality measured using METEOR.

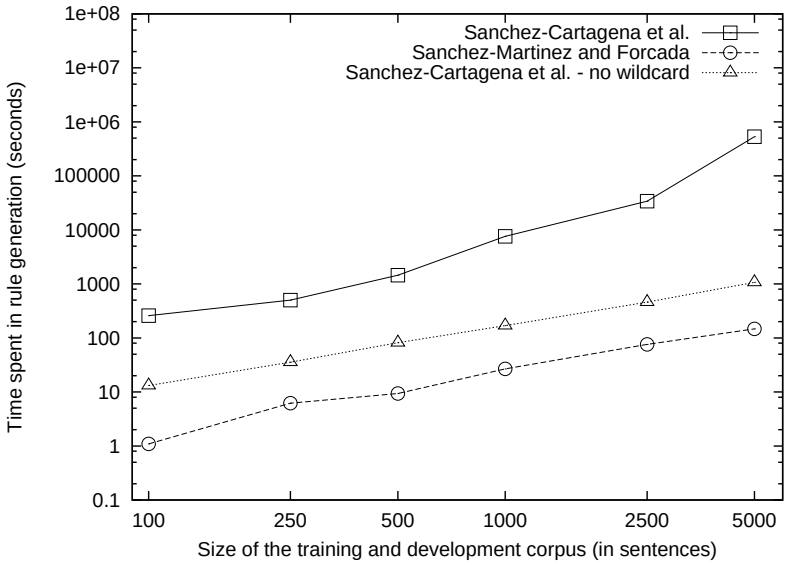


(d) Number of alignment templates inferred.

(continued in next page)

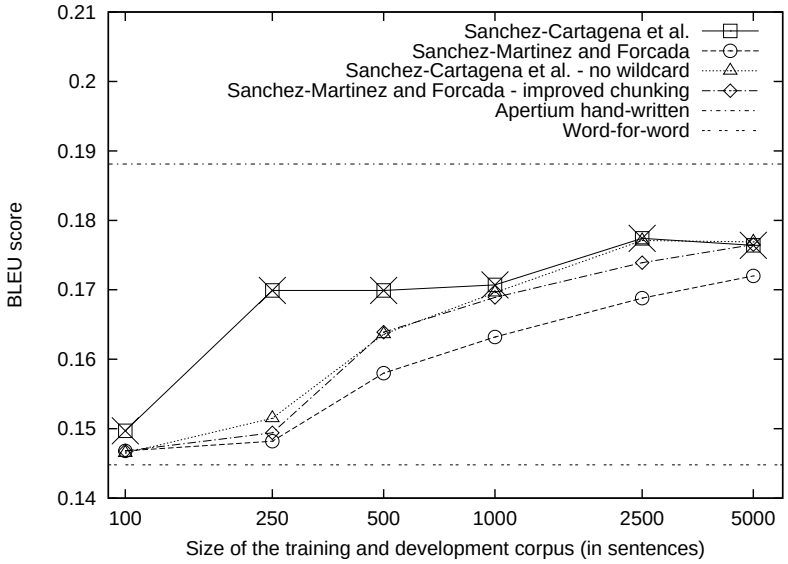


(e) Proportion of words from the test set translated by an alignment template.

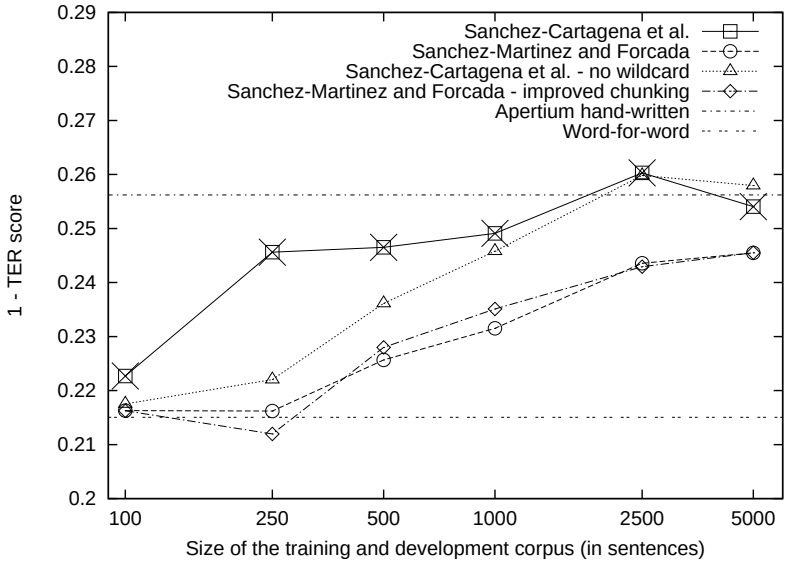


(f) Computing time required to infer alignment templates.

Figure 2.20: Translation quality, number of alignment templates inferred, coverage (proportion of words in the test set translated by an alignment template) and computing time required to infer alignment templates from the different systems evaluated for the Spanish–English language pair. A diagonal cross over a square point indicates that the new approach outperforms the baseline approach proposed by Sánchez-Martínez and Forcada (2009) by a statistically significant margin ($p \leq 0.05$). If the cross is over a circle, the baseline outperforms the new approach.

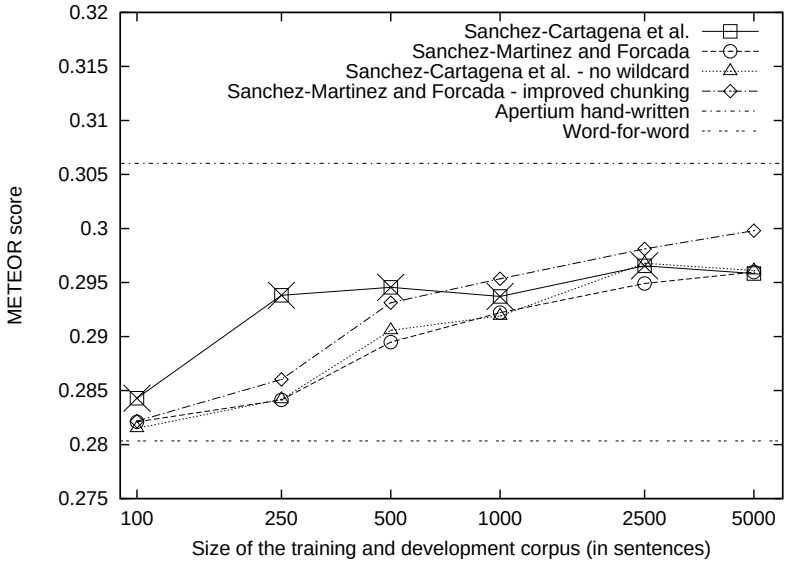


(a) Translation quality measured using BLEU.

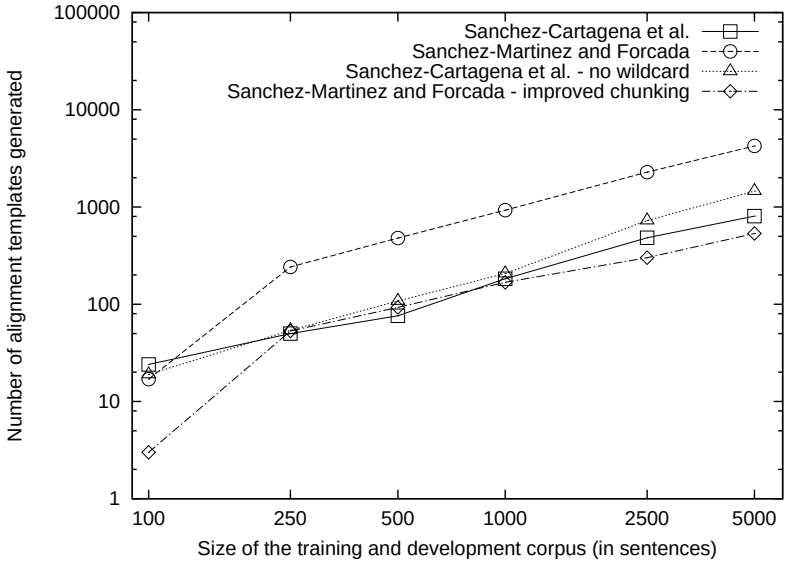


(b) Translation quality measured using TER.

(continued in next page)

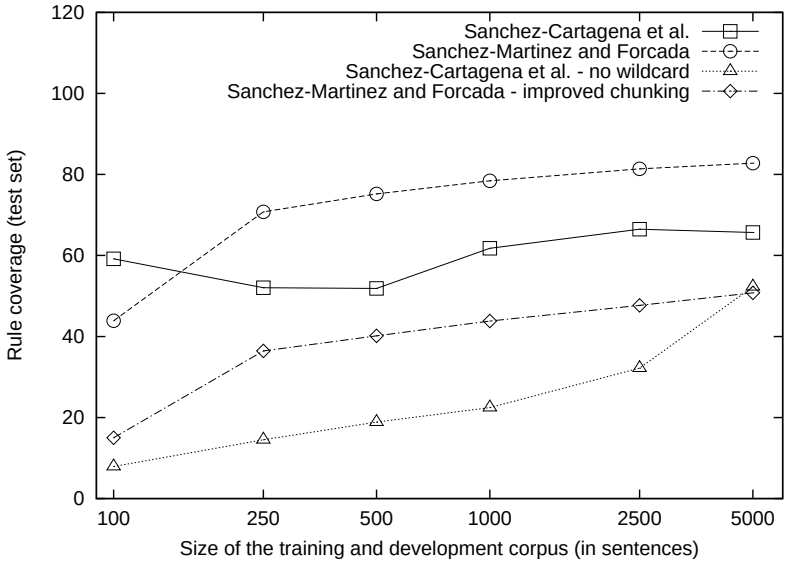


(c) Translation quality measured using METEOR.

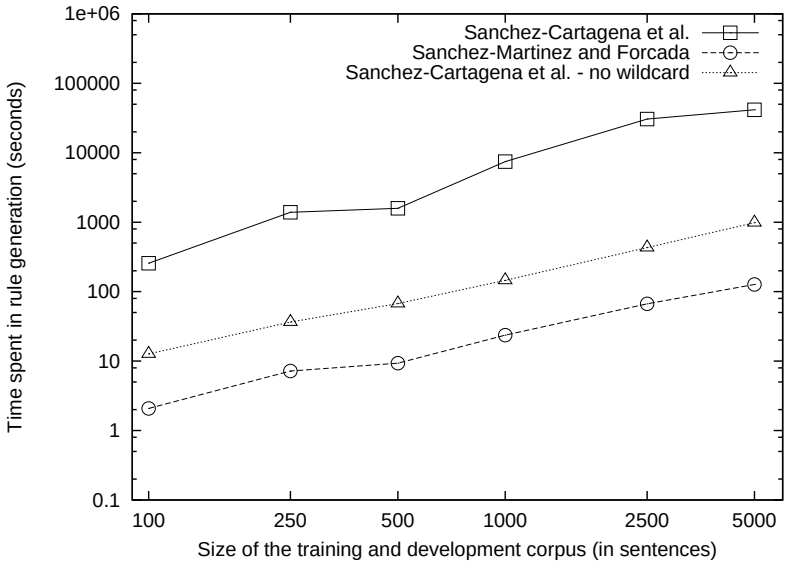


(d) Number of alignment templates inferred.

(continued in next page)

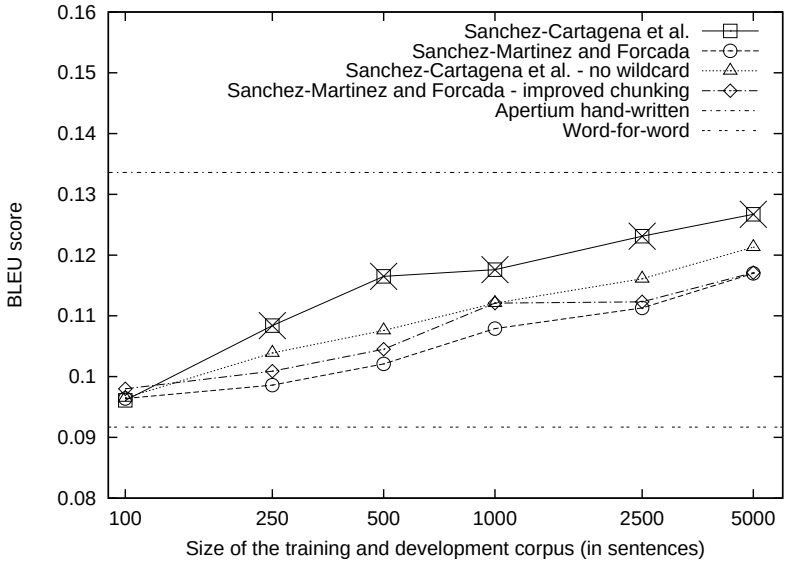


(e) Proportion of words from the test set translated by an alignment template.

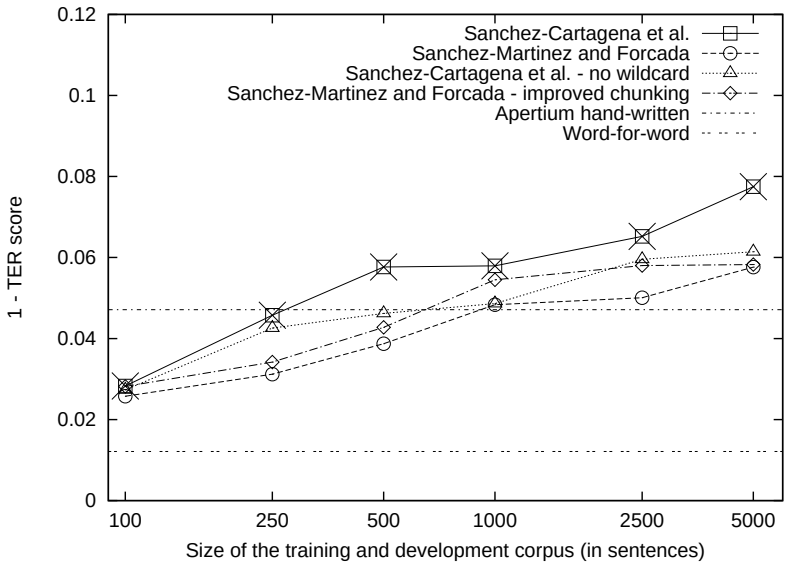


(f) Computing time required to infer alignment templates.

Figure 2.21: Translation quality, number of alignment templates inferred, coverage (proportion of words in the test set translated by an alignment template) and computing time required to infer alignment templates from the different systems evaluated for the Breton–French language pair. A diagonal cross over a square point indicates that the new approach outperforms the baseline approach proposed by Sánchez-Martínez and Forcada (2009) by a statistically significant margin ($p \leq 0.05$). If the cross is over a circle, the baseline outperforms the new approach.

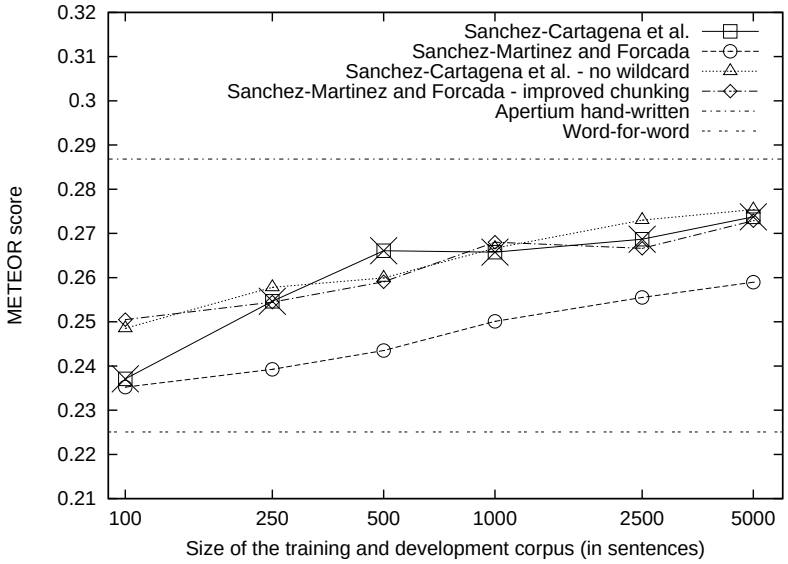


(a) Translation quality measured using BLEU.

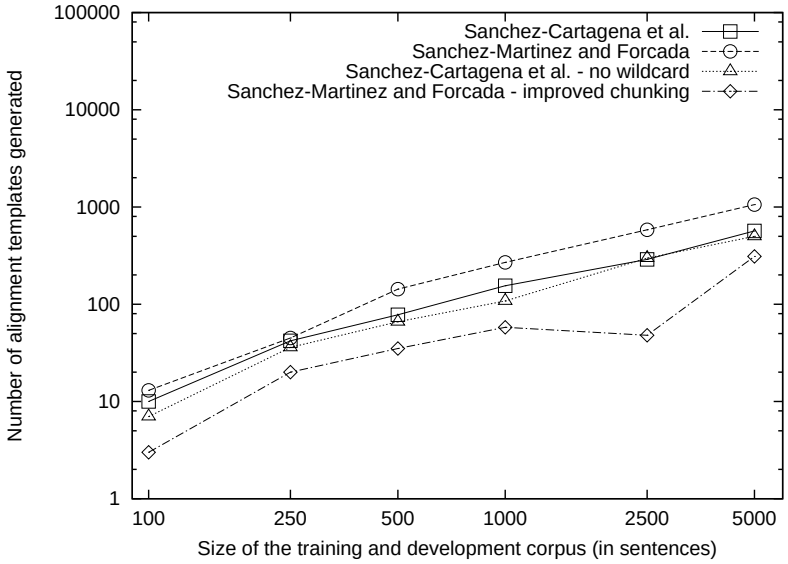


(b) Translation quality measured using TER.

(continued in next page)

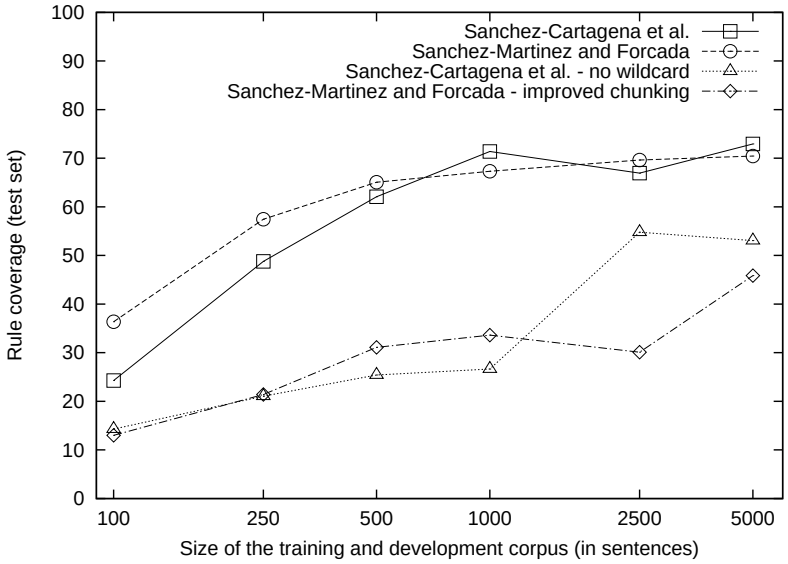


(c) Translation quality measured using METEOR.

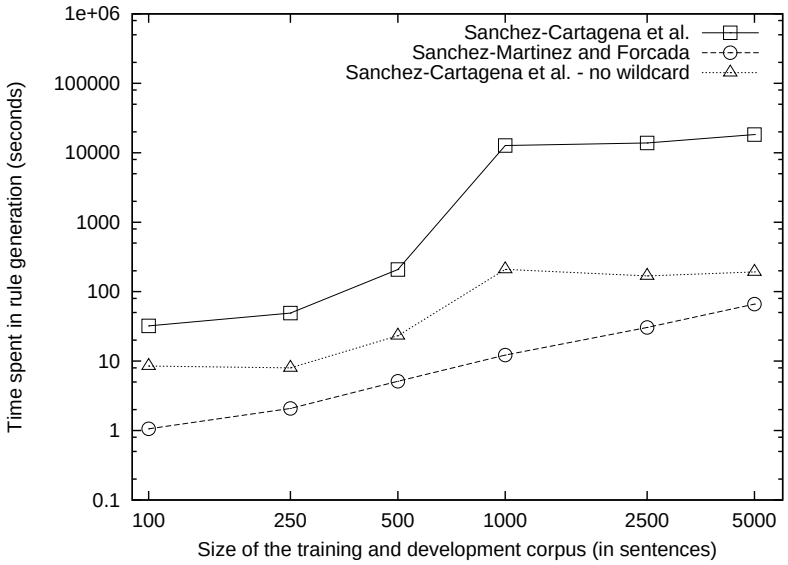


(d) Number of alignment templates inferred.

(continued in next page)



(e) Proportion of words from the test set translated by an alignment template.



(f) Computing time required to infer alignment templates.

hand-crafted rules when translation quality is evaluated using TER (Figure 2.21). This may be explained by the fact that the Breton–French hand-crafted rules are less mature since less work seems to have been carried out for their development.[33]

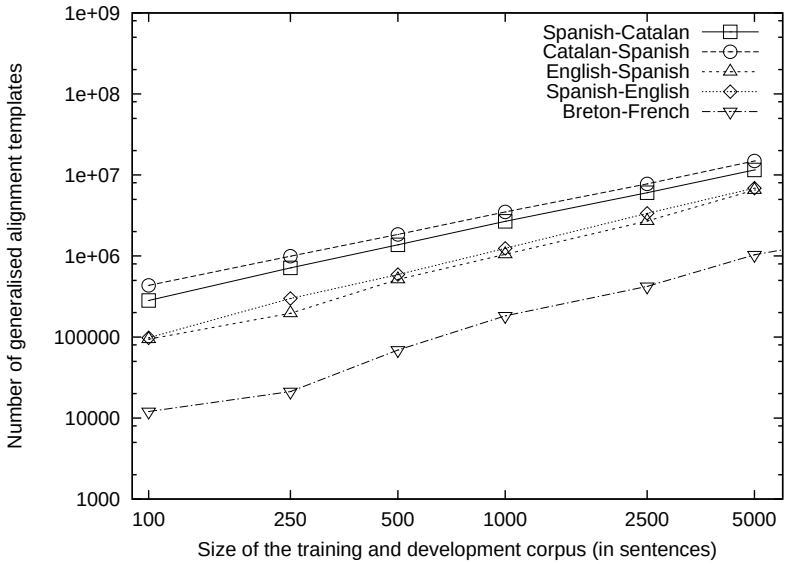
In general, the translation quality achieved by the rules inferred with the new approach grows with the size of the training corpus for the language pairs which are not closely related, namely English↔Spanish and Breton–French. Systems built with the baseline approach also follow this pattern. In the case of closely-related languages, e.g. Spanish↔Catalan, the translation quality grows with the amount of corpora used for training at a slower pace and with some fluctuations (Catalan–Spanish) or does not grow at all (Spanish–Catalan). These results suggest that a few hundred parallel sentences are sufficient to infer useful shallow-transfer rules for closely-related language pairs, since it would appear that no clear improvement is obtained by increasing the size of the training corpus. The drop in translation quality detected by the three metrics for Spanish–Catalan when the training corpus contains $2,500$ sentence pairs is caused by an inadequate value of δ: the value which optimizes the BLEU score in the development set appears to cause a drop in performance in the test set.

The difference in performance between the new approach and the baseline is reduced as the amount of corpora used for training grows, mainly because the effect of generalising the morphological inflection attributes is stronger when the corpus is very small (see below). However, the effect of the filtering based on the threshold θ performed to reduce the amount of input GATs to each minimisation subproblem should also be considered. Figure 2.22 shows the average proportion of GATs retained after applying the filtering based on the threshold θ (described at the end of Section 2.5) for the different language pairs and training corpus sizes. The proportion of GATs retained after the filtering starts to decrease at a faster pace when the size of the training corpora exceeds 1 000 sentences. This decrease is less sharp for the Breton–French language pair because the Breton–French dictionaries have a lower coverage (see Figure 2.16)[34] and the amount of bilingual phrases extracted is consequently lower when compared to the other language pairs (see Figure 2.15). Contrarily, the most pronounced decrease occurs in both directions of the English↔Spanish pair. Even though English↔Spanish is not the language pair for which the highest amount of bilingual phrase pairs are extracted, it is the pair for which a greater amount of GATs are discarded in order to meet the limit of 1 000 input GATs per minimisation subproblem. This may be explained by the fact that English and Spanish are more distant languages than Spanish and Catalan (which are closely related), for which more bilingual phrase pairs are extracted.

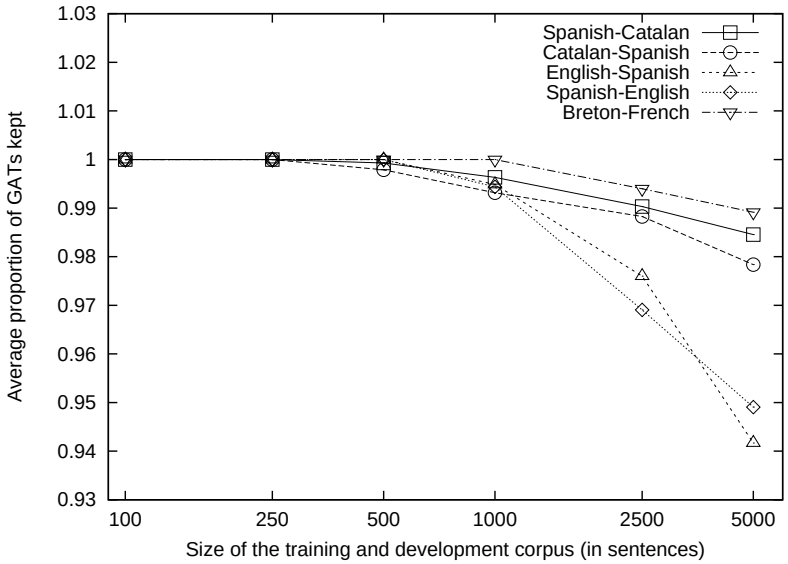
[33]This conclusion is drawn from the date when the first commit affecting the files containing the rules in each language pair was made in the Apertium Subversion repository.

[34]Coverage here is defined as the proportion of surface forms (running words) for which there is at least one possible analysis in the dictionaries being used; note that this does not mean that the correct analysis is returned.

Figure 2.22: For each language pair, number of GATs initially generated from the set of bilingual phrases (top), and average proportion of GATs retained after applying the filtering based on the threshold θ and described at the end of Section 2.5 (bottom). The values reported correspond to the filtering performed on the GATs obtained with a value of $\delta = 0$, and a value of θ automatically chosen for each minimisation subproblem to limit the number of input GATs to 1 000. GATs that do not reproduce at least 2 bilingual phrases have been excluded from the computation of the proportion, since they are always discarded (see Section 2.5).



(a) Number of GATs initially generated from the set of bilingual phrases extracted from the parallel corpus.



(b) Average proportion of GATs retained after applying the filtering based on the threshold θ.

A comparison between the performance of the approach described in this chapter and the alternative approach that does not generalise the morphological inflection attributes (Sánchez-Cartagena et at. - no wildcard, figures 2.17–2.21) shows that translation quality grows at a similar rate for both approaches when the corpus size is above 1 000 sentences. Recall that when no generalisation of the morphological inflection attributes is performed, no pruning takes place because θ is set to 2 for all the minimisation subproblems. These results suggest that the pruning based on θ has little impact on translation quality since, otherwise, a bigger drop in translation quality would occur.

With respect to the number of GATs eventually included in the rules, also shown in figures 2.17–2.21 for the different corpus sizes evaluated, for most of the language pairs, the number of GATs inferred is one order of magnitude (and in some cases almost two) lower than the amount of EATs obtained with the baseline approach. The greater expressiveness of the new formalism with regard to the baseline approach, and the selection of GATs used to optimise the chunking of the sentences to be translated have led to this reduction in the number of GATs. This reduction is expected to alleviate the effort needed to manually edit the set of inferred rules, if it is necessary to do so. An analysis of the coverage of the test set with the GATs obtained with the different approaches shows another advantage of selecting the GATs to optimise the chunking and the removal of redundant GATs: the new approach achieves better translation quality by applying fewer rules, i.e., only the words which actually need to be processed together are covered by GATs.

As regards the relative impact on the translation quality of the different improvements in comparison to the method proposed by Sánchez-Martínez and Forcada (2009) presented in this chapter, it can be observed that the generalisation of morphological inflection attributes with wildcards and reference values brings a clear advantage, but in general, only when the training corpus is really scarce (less than 1 000 sentences). As mentioned previously, the difference between the complete rule inference algorithm described in this chapter and the variant that does not generalise the morphological inflection attributes (Sánchez-Cartagena et al. - no wildcard) disappears or becomes very small for most of the language pairs when the corpus size exceeds 1 000 sentences. It can therefore be concluded that the overhead brought by the generalisation of morphological inflection attributes is justified and, when its computational cost starts to be prohibitively high (see the computing time required to infer alignment templates in figures 2.17–2.21(f)), the improvement in translation quality that can be expected is really small.

It is also worth comparing the results obtained using the alternative approach that does not generalise the morphological inflection attributes (Sánchez-Cartagena et al. - no wildcard) with those obtained using the approach proposed by Sánchez-Martínez and Forcada (2009) with improved chunking (Sánchez-Martínez and Forcada - improved chunking), that are also depicted in figures 2.17–2.21. A higher translation quality is generally obtained with the first approach. An analysis of the rules inferred by both

systems confirms that GATs with more appropriate lexicalised word classes can be obtained by following the strategy presented in this chapter. Moreover, it has been detected that the input sentences are not chunked in the most convenient way when the rules are inferred with the approach proposed by Sánchez-Martínez and Forcada (2009), even when it is complemented with the strategy aimed at improving chunking (see Section 2.4.5); this fact is especially relevant in the Spanish↔Catalan language pairs. These results suggest that the method described in Section 2.4.5 loses effectiveness when it is not applied to the result of the global minimisation problem.

Given that the positive impact of generalising the morphological inflection attributes is only remarkable for small corpora, disabling it permits scaling up the new approach to bigger corpora. In particular, it has been evaluated with two more subsets of the training corpora that contain 10 000 and 25 000 sentences, respectively. The only language pairs used in this evaluation were the English↔Spanish and Breton–French language pairs, since they are those for which the experimental results described previously suggest that translation quality may continue growing at a fast pace with the size of the corpus.

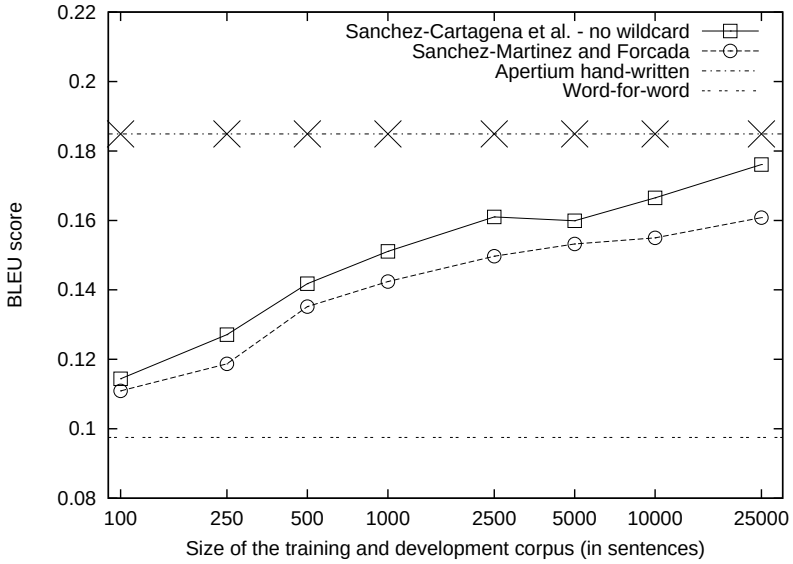
Figures 2.23–2.25 show the translation quality achieved by the rules inferred by the approach described in this chapter when no generalisation of the morphological inflection attributes is performed (i.e. without wildcards and reference values) for the aforementioned language pairs and with larger corpora (the results obtained with small corpora are shown for comparison).[35] The performance of the method proposed by Sánchez-Martínez and Forcada (2009) and the hand-crafted rules is also presented. It can be observed that the translation quality achieved by the new approach keeps growing when the size of the corpus is increased, and it still generally outperforms the approach by Sánchez-Martínez and Forcada (2009). Furthermore, as shown in Figure 2.24, the Spanish–English rules obtained outperform the hand-crafted rules for the biggest corpus size by a statistically significant margin,[36] according to two of the three evaluation metrics (a diagonal cross is placed on top of the points that represent the results of the new approach if they are statistically significantly better than the hand-crafted rules, and also over the points that represent the hand-crafted rules if they are statistically significantly better than the new approach). Notice that the translation quality for English–Spanish and Breton–French also continues to grow.

Finally, the translation quality (as measured by BLEU (Papineni et al., 2002); the rest of metrics behave in a similar way) achieved by the combination of the hand-crafted rules in the Apertium project and those inferred by the new approach is depicted in Figure 2.26. Since the objective is assessing whether the linguistic information contained in the inferred rules is complementary to that in the hand-crafted ones or

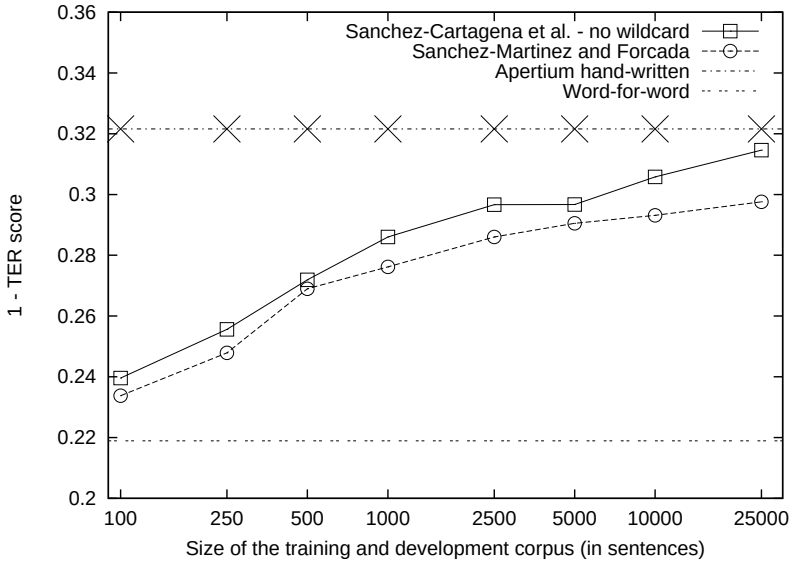
[35]The optimisation of the parameter μ described in Section 2.4.5 for the sets of rules inferred from 10 000 and 25 000 sentences has been performed by means of a ternary search instead of an exhaustive search in order to speed up the process.

[36]Statistically significance margins have been computed with paired bootstrap resampling (Koehn, 2004b) ($p \leq 0.05$, 1 000 iterations).

Figure 2.23: Translation quality of the different systems evaluated for the English–Spanish language pair with larger corpora subsets than those used in the primary evaluation. A diagonal cross over a square point indicates that the new approach outperforms the hand-crafted rules by a statistically significant margin ($p \leq 0.05$). A diagonal cross over the top horizontal line means that the hand-crafted rules outperform the new approach by a statistically significant margin ($p \leq 0.05$).

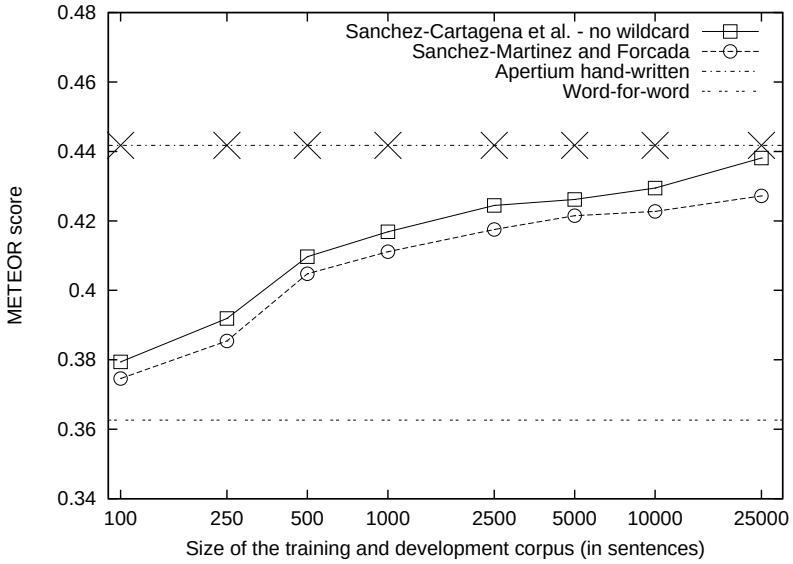


(a) Translation quality measured using BLEU.



(b) Translation quality measured using TER.

(continued in next page)



(c) Translation quality measured using METEOR.

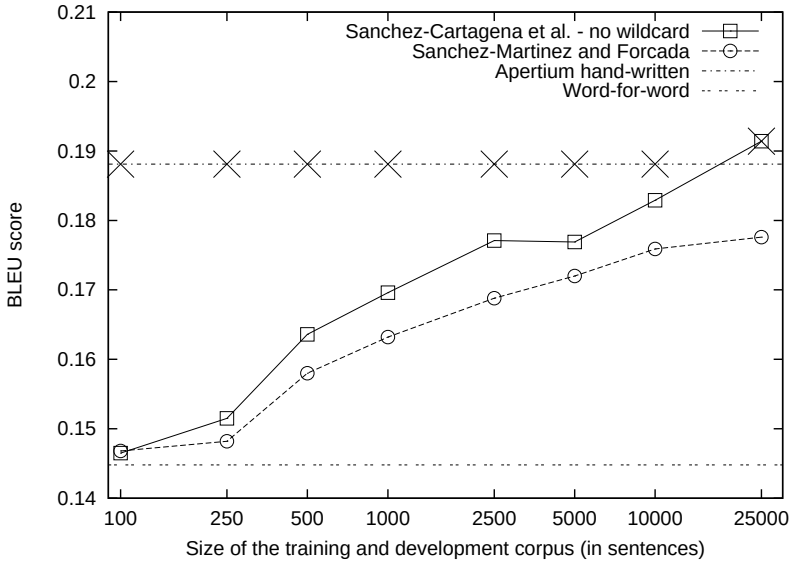
not, they have been combined in such a way that, when the longest matched text segment matches multiple rules, the most specific hand-crafted rule is applied. If there is no hand-crafted rule to apply, the most specific inferred one is used.[37] The results show that the combination does not improve performance. On the contrary, in most cases it causes a degradation of the translation quality originally achieved using the hand-crafted rules. These results suggest that the linguistic information inferred by the new approach has already been encoded by the experts who wrote the rules. It is also worth considering that when an inferred rule matches a segment that is not matched by any hand-crafted rule and translates it, it may prevent a hand-crafted rule from being applied afterwards owing to the greedy rule matching mechanism followed by the Apertium engine. Thus, it may be worth considering in the future the application of the method described in Section 2.4.5 for optimising chunking to the combination of hand-crafted and inferred rules. Nevertheless, the strategy for rule combination that might be most profitable is the use of the approach described in this chapter to infer a set of rules that are then improved or edited by human experts.

2.7 Concluding remarks

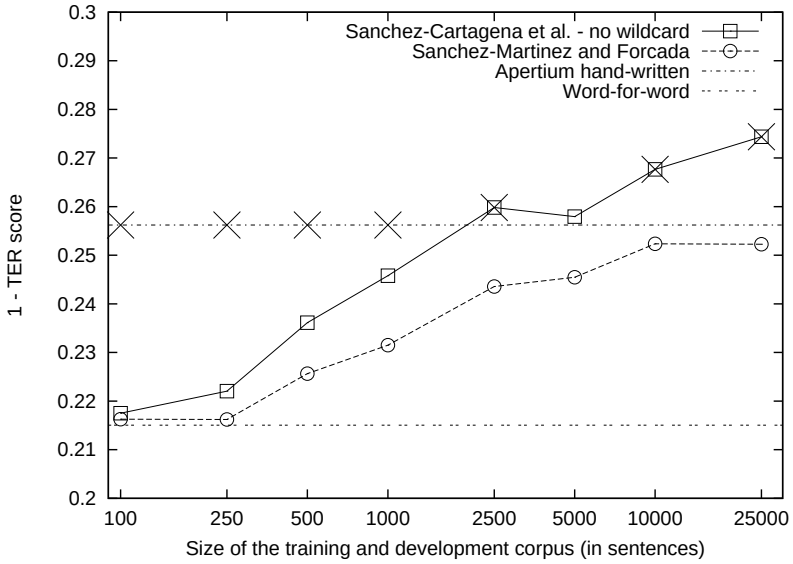
A new alignment-template-based formalism and a language-independent algorithm for the automatic inference of shallow-transfer rules to be used in rule-based MT have been described in this chapter. This new approach has been evaluated with five

[37]Note that, due to the way in which the hand-crafted rules are encoded, it is not possible to treat hand-crafted rules and automatically-inferred ones as a single set and sort them together according to their specificity level.

Figure 2.24: Translation quality of the different systems evaluated for the Spanish–English language pair with larger corpora subsets than those used in the primary evaluation. A diagonal cross over a square point indicates that the new approach outperforms the hand-crafted rules by a statistically significant margin ($p \leq 0.05$). A diagonal cross over the top horizontal line means that the hand-crafted rules outperform the new approach by a statistically significant margin ($p \leq 0.05$).

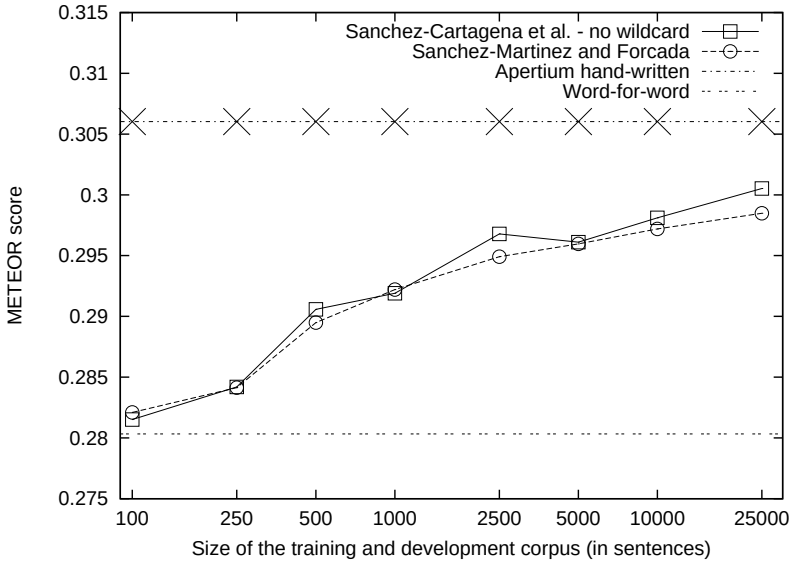


(a) Translation quality measured using BLEU.



(b) Translation quality measured using TER.

(continued in next page)



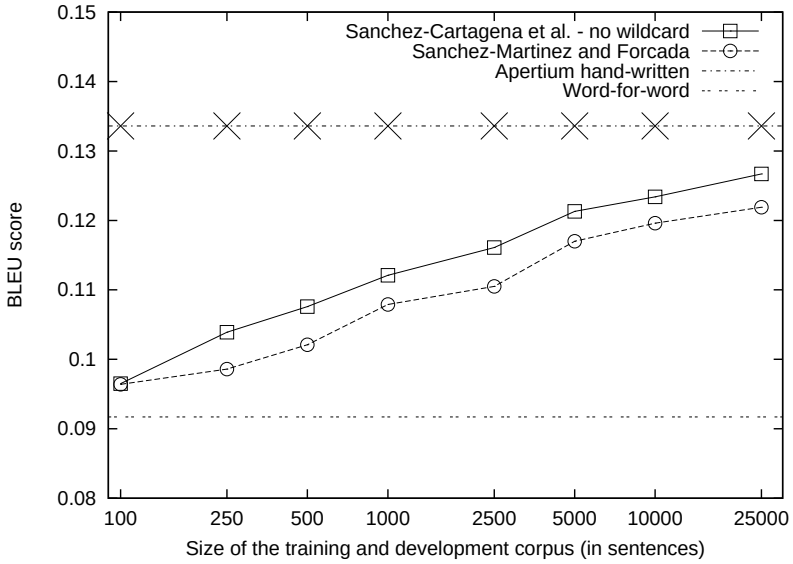
(c) Translation quality measured using METEOR.

different language pairs and with parallel corpora of different sizes. The evaluation performed shows that, in almost all cases and by a statistically significant margin ($p \leq 0.05$), the new method outperforms the previous alignment-template-based approach by Sánchez-Martínez and Forcada (2009). In addition, when the languages involved in the translation are closely-related (e.g. Spanish↔Catalan), a few hundred parallel sentence have proved to be sufficient to obtain a set of competitive transfer rules, since the addition of more parallel sentences does not result in great improvements to the translation quality. What is more, this translation quality is close to that obtained with hand-crafted rules.

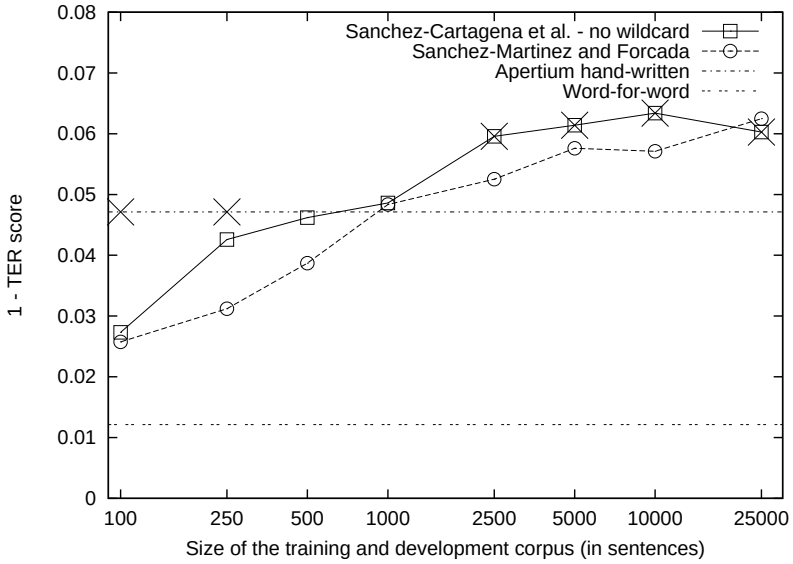
The new approach overcomes many relevant limitations of the previous work, principally those related to the inability to find the appropriate generalisation level for the alignment templates and to select the proper subset of alignment templates which ensures an adequate chunking of the input sentences. Furthermore, the amount of rules inferred by the new approach is much smaller than that of the baseline, and this has a positive impact on the possible manual refinement of the resulting rules, since having fewer and more expressive rules eases editing them. In addition, the approach presented in this chapter is the first to resolve the conflicts between the inferred rules at a global level by choosing the most appropriate rules according to a global minimisation function rather than by following a pairwise greedy approach. This global minimisation function also allows it to automatically determine the appropriate level of generalisation of the GATs to be eventually used for rule generation.

The combinatorial explosion in the generation of GATs with different levels of generalisation and the computational complexity involved in solving the minimisation problem has limited the experiments conducted to very small parallel corpora, and forced us to introduce some heuristics in order to limit the number of GATs to be

Figure 2.25: Translation quality of the different systems evaluated for the Breton–French language pair with larger corpora subsets that those used in the primary evaluation. A diagonal cross over a square point indicates that the new approach outperforms the hand-crafted rules by a statistically significant margin ($p \leq 0.05$). A diagonal cross over the top horizontal line means that the hand-crafted rules outperform the new approach by a statistically significant margin ($p \leq 0.05$).

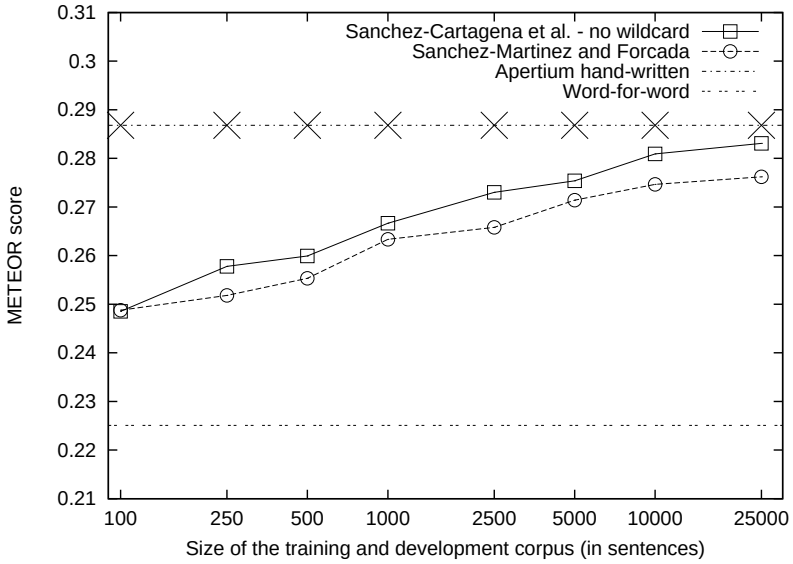


(a) Translation quality measured using BLEU.



(b) Translation quality measured using TER.

(continued in next page)



(c) Translation quality measured using METEOR.

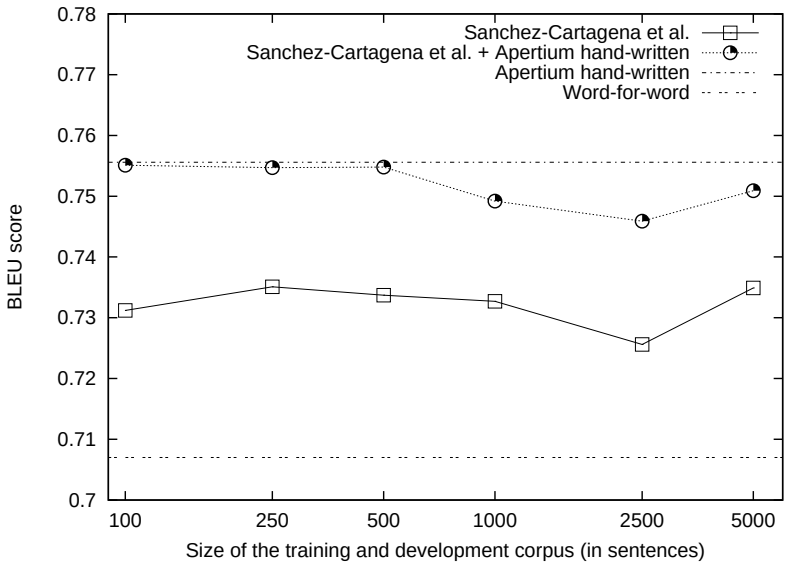
considered during the minimisation. It is, however, when the amount of parallel corpora is scarce that the method achieves the greatest improvement when compared to the baseline approach. In addition, disabling the generalisation of morphological inflection attributes with wildcards and reference values has permitted scaling the new approach to bigger corpora and reach, and in some cases surpass, the translation quality of hand-crafted rules.

In summary, the new algorithm for the inference of shallow-transfer rules presented in this chapter is a cost-effective approach for building MT systems when only dictionaries (monolingual and bilingual) and a small parallel corpus are available.[38] Its generalisation power allows it to create high-quality transfer rules, which can be easily edited by humans, from parallel corpora that contain a few thousands of words in each language. Its adoption will hopefully contribute towards easing development of transfer rules for new language pairs in MT systems like Apertium, thus reducing the total time necessary to deploy working systems. Moreover, this rule inference algorithm can also be used to improve the degree of generalisation an SMT system can perform over the training corpus, as shown in Chapter 3.

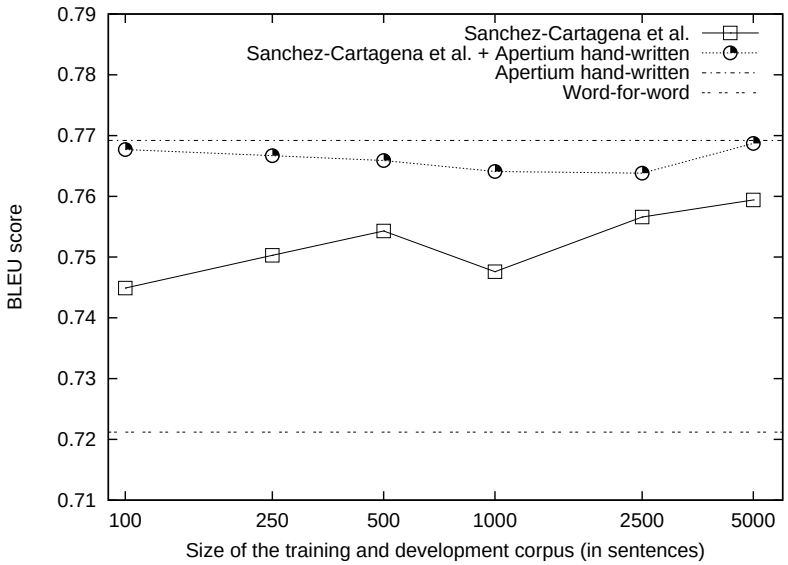
Another interesting extension of the approach presented in this chapter, which could be explored in the future, would be a method with which to select the most informative sentences from a monolingual corpus that should be manually translated in order to obtain a parallel corpus for rule inference. It will be discussed in Section 5.2, together

[38] According to Sánchez-Martínez and Forcada (2009), for the parallel corpus sizes considered in this chapter, a rule-based MT system with rules inferred from the parallel corpus is able to outperform SMT system trained on the same parallel corpus (without additional data for training the language model), even when it is complemented with the entries from the bilingual dictionary of the RBMT system.

Figure 2.26: Translation quality, as measured by BLEU score, of the combination of the rules inferred by the approach described in this chapter with the hand-crafted rules from the Apertium project. The scores achieved by the hand-crafted rules alone, the rules obtained by the approach described in this chapter alone, and word-for-word translation are also depicted.

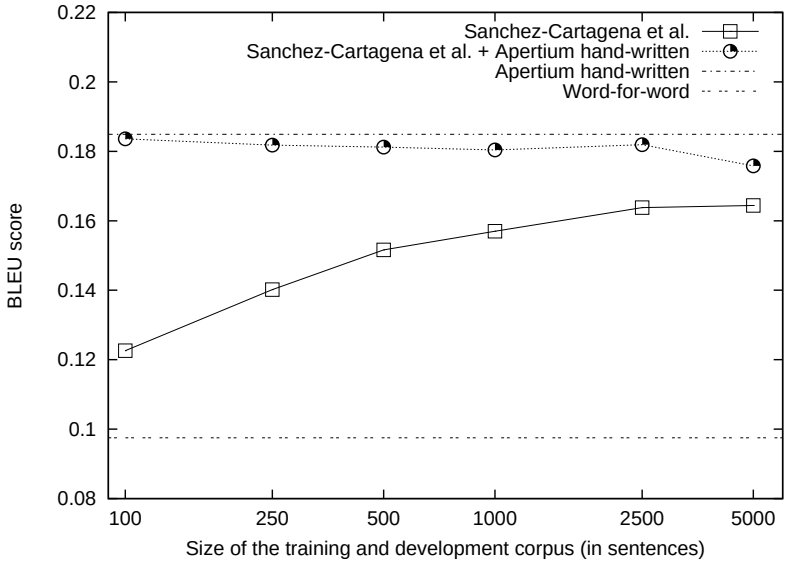


(a) Spanish–Catalan.

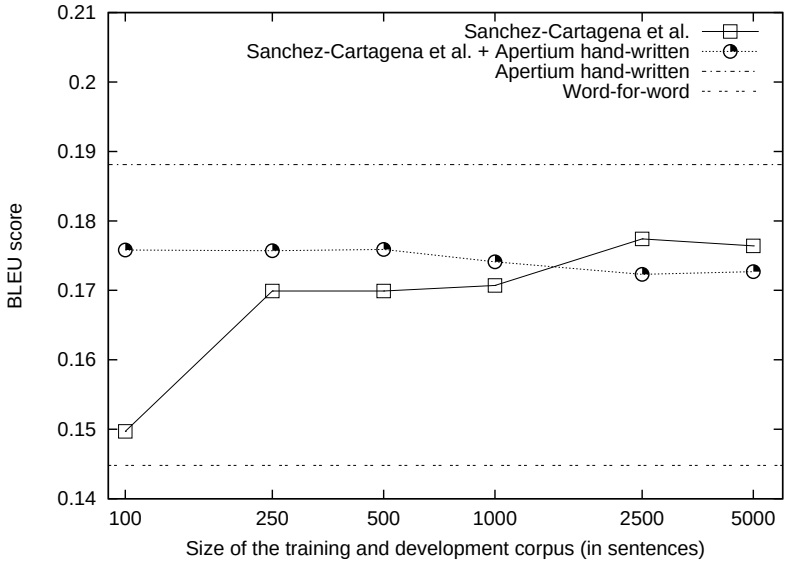


(b) Catalan–Spanish.

(continued in next page)

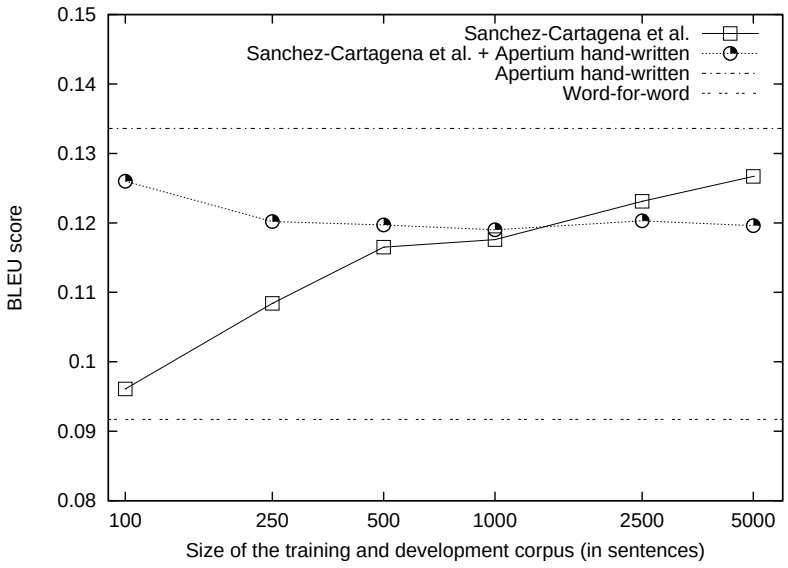


(c) English–Spanish.



(d) Spanish–English.

(continued in next page)



(e) Breton–French.

with alternative approaches for some of the steps of the rule learning procedure that could further improve the results obtained.

Chapter 3

Integrating shallow-transfer rules into statistical machine translation

In this chapter, a new hybridisation strategy aimed at integrating the linguistic resources (rules and dictionaries) from a shallow-transfer rule-based machine translation system into phrase-based statistical machine translation is presented. The new strategy takes advantage of how the linguistic resources are used by the rule-based system to segment the source-language sentences to be translated and overcomes the limitations of the existing general approach that treats the rule-based machine translation system as a black box; namely, the extraction from the rule-based system of phrase pairs that are not mutual translation and the inability to find an adequate balance between the weight of the phrase pairs extracted from the parallel corpus and those obtained from the rule-based system. The experiments performed confirm that the new approach delivers a higher translation quality than the existing general approach, and that shallow-transfer rules are specially useful when the parallel corpus available for training is small or when translating out-of-domain texts that are well covered by the shallow-transfer rule-based machine translation dictionaries. When this approach is combined with the rule inference algorithm presented in Chapter 2, a significant boost in translation quality over a baseline statistical machine translation system is obtained. In that case, the only hand-crafted resource needed is the set of dictionaries commonly used by a rule-based machine translation system. The translation quality achieved by hybrid systems built with automatically inferred rules reaches that obtained by hybrid systems that contain hand-crafted rules.

3.1 Introduction

As it has already been pointed out in the introductory chapter, SMT can be combined with RBMT in order to mitigate some of the SMT limitations such as the data sparseness caused by highly-inflected languages, or simply to increase the amount of data available for building the SMT system.

This chapter presents a new strategy aimed at enriching phrase-based SMT models with linguistic knowledge from shallow-transfer RBMT. Although the enrichment of phrase-based SMT models with RBMT linguistic data has already been explored by other authors (see Section 1.3.2.1), the approach presented in this chapter is the first one that has been specifically designed for its use with shallow-transfer RBMT and that takes advantage of the way in which the linguistic resources are used by the rule-based system (henceforth, it treats the RBMT system as a *white box*). It overcomes the limitations of the more general hybridisation method devised by Eisele et al. (2008) (described in Section 3.1.1). Note that, since no other hybridisation strategy focusing on shallow-transfer RBMT can be found in the literature, the approach by Eisele et al. (2008) is considered the reference approach in this chapter. It is the only existing approach for enriching an SMT system with RBMT resources that can be applied to shallow-transfer RBMT (in fact, it can be applied to any MT system). Moreover, the shallow-transfer rule inference algorithm presented in the previous chapter permits the application of this hybrid approach to language pairs for which hand-crafted shallow-transfer rules are not available using, in this case, the rules automatically inferred from a small fragment of the parallel corpus.

The rest of the chapter is organised as follows: the remainder of this section describes the limitations of the general hybridisation strategy by Eisele et al. (2008); after that, Section 3.2 describes the new hybridisation strategy and a set of different alternatives for scoring the phrase pairs generated from the linguistic data in the RBMT system. Then, three different sets of experiments are described in order to:

- Evaluate the new hybridisation strategy (Section 3.3). Results that confirm that it outperforms the previous strategy by Eisele et al. (2008), are also reported and discussed in the same section.

- Assess whether the automatically inferred rules can replace hand-crafted ones in the hybrid system (Section 3.4). The experiments confirm that automatically inferred rules can achieve a translation quality similar to that of hand-crafted rules in some cases.

- Study the impact of the size of the language model on the improvement brought by the hybrid approach (Section 3.5). As expected, results show that the impact of the RBMT data included in the SMT system decreases as the size of the monolingual corpus from which the language model is obtained grows.

The chapter ends with some concluding remarks.

3.1.1 Limitations of the general hybridisation approach for integrating a black-box MT system into the SMT architecture

The hybridisation approach defined by Eisele et al. (2008) treats the RBMT system from which the linguistic knowledge is extracted and integrated into the SMT models as a black box, i.e. it does not use information from the inner workings of the RBMT system. The sentences to be translated by the hybrid system developed by Eisele et al. (2008) are first translated with the RBMT system (actually, multiple MT systems can be used) and a phrase table is obtained from the resulting parallel corpus (from now on, *synthetic corpus*). Phrase pairs are extracted by following the usual procedure carried out in phrase-based SMT (Koehn, 2010, sec. 5.2.3), which generates the set of all possible phrase pairs that are consistent with the word alignments. Since the synthetic corpus may be small, word alignments are computed using an alignment model previously built from the parallel corpus from which the original SMT phrase table was extracted (usually much larger than the synthetic corpus). Finally, the synthetic phrase table is directly added to the original one.

The approach by Eisele et al. (2008) presents the following limitations, which are overcome by the new hybrid approach described in this chapter:

Deficient segment alignment. When phrase pairs are extracted from the synthetic corpus through the usual procedure followed in phrase-based SMT (Koehn, 2010, sec. 5.2.3), word alignments are used as anchors. Unaligned words are included in multiple phrase pairs since there is no evidence about their correspondence in the other language, and phrase pairs made solely of unaligned words are not extracted. If word alignments are incorrect, phrase pairs that are not mutual translation may be extracted and other correct phrase pairs present in the parallel sentence may not be obtained.[1] The less reliable the word alignments are, the more severe this problem becomes.

The word alignment of the synthetic corpus obtained in the approach by Eisele et al. (2008) may be unreliable due to a vocabulary mismatch between the synthetic

[1]For instance, consider the following segment of an English–Spanish parallel sentence: *Barcelona City Council – Ayuntamiento de Barcelona*. If the only word alignment between these two segments was a link between *Barcelona* in both languages, incorrect phrase pairs such as *Barcelona City Council – Barcelona* would be extracted, whereas the correct phrase pair *City Council – Ayuntamiento* would not be extracted.

corpus itself and the alignment models inferred from the training corpus.[2] This limitation becomes more evident when the test corpus does not share the domain with the training corpus, which is actually when the data from the RBMT system is more useful. Aligning the synthetic corpus with an alignment model learnt from the concatenation of the synthetic corpus and the training corpus could be a solution, but it would be computationally too expensive, since the process would have to be carried out each time a new text is translated with the resulting hybrid system.[3]

Relying on word alignments is a reasonable strategy when extracting phrase pairs from a parallel corpus for which the process followed to obtain its TL side from its SL side is unknown. However, when that process is known because an RBMT system has been used to translate its SL side, a more precise phrase extraction mechanism can be followed. In the new approach presented in this chapter, phrase pairs are extracted from the SL sentences to be translated by the hybrid system by taking advantage of how the RBMT system uses dictionaries and shallow-transfer rules to segment the SL sentences.

Inadequate balance between the scores of phrase pairs extracted from the parallel corpus and those extracted from the RBMT system. The probabilities derived by Eisele et al. (2008) and contained in the resulting phrase table are not consistent because they have been independently estimated from two different corpora. In their computation, some important factors are not taken into account. Firstly, if an SL phrase is translated in the same way in the training parallel corpus and by the RBMT system, the probability of the corresponding phrase pair is not increased over phrase pairs for which the translations of the SL phrase according to the two bilingual resources being combined differ. Secondly, when the translations of the SL phrase differ, its frequency in the training parallel corpus should be taken into account when scoring the corresponding phrase pairs in order to avoid the noise that may be introduced by low-frequency SL phrases in the parallel corpus. A phrase pair extracted from the training parallel corpus whose SL phrase appears only once is less reliable and should receive a lower score than a phrase pair whose SL phrase appears 10 000 times. These limitations, which are described in more detail in Section 3.2.2, are overcome in the new hybridisation approach presented in this chapter by following a more sophisticated scoring scheme in which the synthetic phrase pairs and those obtained from the training parallel corpus are scored together by relative frequency as it is usually done in SMT (Koehn, 2010, sec. 5.2.5) and a new binary feature function is added to the phrase table.

[2]Alignment models do not have information about words in the test corpus that are not present in the training corpus, thus these words are not aligned and it is likely that phrase pairs that are not mutual translation are extracted from them.

[3]For instance, building word alignment models from the English–Spanish parallel corpus with 600 000 sentences described in Section 3.3.1 took around 6 hours in an AMD Opteron 2 Ghz processor.

3.2 Enhancement of phrase-based statistical machine translation with shallow-transfer linguistic resources

As already mentioned, the structural transfer module of a shallow-transfer RBMT system (in particular, the Apertium RBMT platform (Forcada et al., 2011), which has been used in the experiments) detects sequences of lexical forms —consisting of lemma, lexical category and morphological inflection information— which need to be treated together to prevent them from being wrongly translated.

If the RBMT system is treated as a white box, the correspondence between the SL segments of a sentence to be translated and their translations with the RBMT system can be computed without relying on statistical word alignments. In fact, it is not even necessary to translate the whole sentence with the RBMT system. The individual translation according to the bilingual dictionary of each word, and the translation of each segment that matches a shallow-transfer rule constitute the minimum set of bilingual phrases that ensures that all the linguistic information from the RBMT system has been extracted. Another advantage of this method versus the approach by Eisele et al. (2008) lies in the fact that rules that match a segment of the SL sentence but would not applied by the shallow-transfer RBMT system because of its greedy operating mode are also taken into account.[4] Thus, the new hybrid strategy first generates these synthetic phrase pairs from the RBMT linguistic data and then integrates them into the phrase-based SMT models without further decomposition.

In the remainder of this section, the generation of these phrase pairs from the linguistic resources of the Apertium RBMT project is first presented in more detail and, afterwards, different methods to integrate them into the phrase-based SMT models and properly score them are presented and discussed.

[4]Consider, for instance, that the English sentence *I visited Bob and Alice's dog was sleeping* is to be translated into Spanish with the Apertium shallow-transfer RBMT system. Let us suppose that the following segments of the sentence match a shallow-transfer rule: *I visited* matches a rule that removes the personal pronoun (it can be omitted in Spanish), adds the corresponding preposition and generates *visité a*; *Bob and Alice's dog* matches a rule that processes the Saxon genitive, adds the preposition and determiner needed in Spanish and generates *el perro de Bob y Alice*; and *Alice's dog* also matches a rule that processes the Saxon genitive when the noun phrase acting as owner contains a single proper noun, and generates *el perro de Alice*. When the RBMT engine chooses the rules to be applied in a left-to-right, longest match fashion, it produces *visité al perro de Bob y Alice estaba durmiendo*, that means *I visited Bob's dog and Alice was sleeping*. The right translation, *visité a Bob y el perro de Alice estaba durmiendo*, can be obtained if the rule that matches *Alice's dog* is applied. If the method by Eisele et al. (2008) is applied in order to build a hybrid system, the phrase pairs from the correct translation *I visited Bob – visité a Bob* and *Alice's dog was sleeping – el perro de Alice estaba durmiendo* will not be available in the phrase table of the hybrid system.

3.2.1 Synthetic phrase pair generation

The process followed to generate synthetic phrase pairs depends on the linguistic resource from which they are generated: either the bilingual dictionary or the set of shallow-transfer rules.

In order to generate bilingual phrase pairs from the bilingual dictionary, all the SL surface forms that can be analysed by the shallow-transfer RBMT system and their corresponding SL IR are listed; then, each SL IR obtained in the previous step is translated with the bilingual dictionary in order to obtain its corresponding TL IR; finally, the generation module of the RBMT system is run so as to produce the corresponding TL word form(s).[5] For instance, for the generation of phrase pairs from the English–Spanish bilingual dictionary in the Apertium RBMT system, mappings between SL surface forms and lexical forms such as *houses* – *house* `N-num:pl` and *however* – *however* `ADV` are generated. They are then translated into the TL by the bilingual dictionary: the resulting phrase pairs are *houses* – *casas* and *however* – *sin embargo*. Since dictionaries may contain multi-word units, the phrase pairs generated may contain more than one word in both (SL and TL) sides. Note that, unlike in the method by Eisele et al. (2008), the sentences to be translated are not used. Thus, the generation of phrase pairs from the bilingual dictionary only needs to be performed once, instead of being done each time a new text is to be translated with the hybrid system.

Bilingual phrase pairs which match structural transfer rules are generated in a similar way. First, the SL sentences to be translated are analysed in order to get their SL IR, and then the sequences of lexical forms that match a structural transfer rule are passed through the rest of the RBMT pipeline to get their translations. If a sequence of SL lexical forms is covered by more than one structural transfer rule, they will be used to generate as many bilingual phrase pairs as different rules it matches. This differs from the way in which Apertium translates, since in these cases only the longest rule would be applied.

Let the English sentence *My little dogs run fast* be one of the sentences to be translated into Spanish. It would be analysed by Apertium as the following sequence of lexical forms: *my* `POSP-p:1.num:pl`, *little* `ADJ`, *dog* `N-num:pl`, *run* `VERB-t:inf`, *fast* `ADV`.[6] If the RBMT system only contained two rules, one that performs the swapping and number and gender agreement between an adjective and the noun after it, and

[5] If the TL IR contains missing values for morphological inflection attributes, a different TL phrase for each possible value of the attribute is generated. For instance, from the mapping between the SL word form *beautiful* and the SL lexical form *beautiful* `ADJ-num:sg` (assuming that phrase pairs for the English–Spanish language pair are being generated), two phrase pairs are generated: *beautiful* – *bonito* and *beautiful* – *bonita*. Since adjectives have gender in Spanish, in the first phrase the translation of *beautiful* has been labelled with masculine gender, while in the second phrase pair the feminine gender has been used.

[6] The meanings of the abbreviations used to represent lexical categories are: *POSP* = possessive pronoun; *ADJ* = adjective; *N* = common noun; *VERB* = verb; and *ADV* = adverb. Regarding

another one that matches a determiner followed by an adjective and a noun, swaps the adjective and the noun and makes the three words to agree in gender and number, the segments *little* `ADJ` *dog* `N-num:pl` and *my* `POSP-p:1.num:pl` *little* `ADJ` *dog* `N-num:pl` would be used to generate bilingual phrase pairs. As a result, the following phrase pairs would be obtained: *little dogs – perros pequeños* and *my little dogs – mis perros pequeños*.

Note that, unlike the generation of bilingual phrases from the bilingual dictionary, the generation of bilingual phrase pairs from the shallow-transfer rules is guided by the text to be translated. It has been decided to do it in this way in order to avoid meaningless phrases and to make the approach computationally feasible. Consider, for instance, the rule which is triggered by a determiner followed by an adjective and a noun in English. Generating all the possible phrase pairs virtually matching this rule would involve combining all the determiners in the dictionary with all the adjectives and all the nouns, causing the generation of many meaningless phrases, such as *the wireless boy – el niño inalámbrico*.

All the phrase pairs generated from the bilingual dictionary and from the shallow-transfer rules are assigned a frequency of 1 since they have not been generated from an actual parallel corpus. Their frequencies will be used for scoring them, as described in the next section.

3.2.2 Scoring the synthetic phrase pairs

In order to allow the phrase-based SMT decoder to generate hypotheses that contain the synthetic phrase pairs generated from the RBMT data, the synthetic phrase pairs must be added to the phrase translation table (translation model). Phrase-based SMT systems usually attach 4 scores (Koehn, 2010, Sec. 5.3) to every phrase pair in the phrase translation table: source-to-target and target-to-source phrase translation probabilities and source-to-target and target-to-source lexical weightings. The source-to-target translation probability $\phi(t|s)$ of a phrase pair (s,t) is usually computed as shown in the equation below. The function $\text{count}(\cdot)$ stands for the frequency of a phrase pair in the list of phrase pairs extracted from the training parallel corpus.

$$\phi(t|s) = \frac{\text{count}(s,t)}{\sum_{t_i} \text{count}(s,t_i)}$$

The purpose of lexical weightings is acting as a back-off when scoring phrase pairs with low frequency (Koehn, 2010, Sec. 5.3.3). The lexical weighting score of a phrase pair is usually computed as the product of the lexical translation probability of each source word and the target word with which it is aligned. Lexical translation probabilities are obtained from a lexical translation model estimated by maximum likelihood

morphological inflection information, `p:1` means *first person*, `num:pl` means *plural number* and `t:inf` means *infinitive mood*.

from all the word alignments of the parallel corpus. For more details, the reader is referred to Section 5.3.3 of the book on SMT by Koehn (2010).

The values of these four scores for the synthetic phrase pairs generated from the shallow-transfer RBMT data can be calculated in different ways, which can also affect the scores of the phrase pairs extracted from the original training corpus. The strategy for scoring the new phrase pairs should allow the tuning step of the SMT training process to adapt the relative relevance of both types of phrase pairs (extracted from the training corpus and synthetic) to the type of texts to be translated with the hybrid system. In addition, there are other desirable features of the method followed for scoring both synthetic and corpus-extracted phrase pairs, already outlined in Section 3.1.1. Firstly, agreements between the translation of a certain SL phrase according to the corpus and the RBMT system should increase the probability of the corresponding phrase pair (i.e. probabilities of both types of phrase pairs should not be computed independently); and secondly, when considering whether to use a translation extracted from the RBMT system or from the parallel corpus for a given SL phrase, the frequency in the parallel corpus of the SL phrase itself should influence the reliability of its corpus-extracted translations.

Finally, it is also desirable that the addition of the synthetic phrase pairs to the statistical models does not involve a big computational effort, since it is executed once for each text to be translated by the hybrid system.

In this section, a new method for integrating the set of synthetic phrase pairs obtained from the RBMT data in the SMT system that meets all the aforementioned requirements is described (Section 3.2.2.3). The remainder of this section contains, in addition to the new method, the description of other phrase scoring approaches that can be found in the literature and their limitations.[7] All the strategies presented below have been evaluated as it will be described in Section 3.3.

3.2.2.1 Creating an additional phrase table

A simple strategy for integrating the synthetic phrase pairs in the SMT system is putting them in a different phrase table, as Koehn and Schroeder (2007) propose in the context of domain adaptation. When the decoder builds translation hypotheses, it looks for phrase pairs in both phrase tables. If the same phrase pair is found in both phrase tables, one instance from each phrase table is used to build translation hypotheses.[8] For that reason, some authors refer to this approach as *alternative decoding paths*.

[7] Methods in which the relevance of the phrase tables being combined must be defined in advance (i.e., there is a primary and a secondary phrase table), such as *fill-up* (Bisazza et al., 2011), are not described in this section and have not been evaluated. As it has been pointed out previously, the strategy for scoring the new phrase pairs should allow the tuning step of the SMT training process to adapt the relative relevance of both types of phrase pairs (extracted from the training corpus and synthetic) to the type of texts to be translated with the hybrid system.

[8] It may have a different probability in each phrase table.

Each score[9] in each phrase table receives a different weight during the tuning process, which should help the hybrid system to obtain the appropriate relative weighting of both sources of phrase pairs.

When this scoring strategy is used for integrating the synthetic phrase pairs generated as described in Section 3.2.1 into the SMT models (the experiments carried out in order to evaluate this strategy are described in Section 3.3), the synthetic phrase table contains exclusively the synthetic phrase pairs generated from the RBMT resources and their phrase translation probabilities are computed also by relative frequency (see the equation at the beginning of Section 3.2.2), as it is done with the phrase pairs extracted from the parallel corpus. When building the synthetic phrase table, the function count$(\cdot)$ represents the frequency of a phrase pair in the list of phrase pairs generated from the RBMT system, which equals to 1 for all the phrase pairs.

To compute the lexical weighting scores of the synthetic phrase pairs, a set of word alignments for each phrase pair and a lexical translation model are needed. The lexical translation model is estimated from a synthetic corpus generated only from the RBMT bilingual dictionary as described in Section 3.2.1. Since synthetic phrase pairs are not extracted through the usual procedure followed in phrase-based SMT (Koehn, 2010, sec. 5.2.3), the statistical word alignments generated by that procedure are not available. Instead, the word alignments needed to compute the lexical weightings of each phrase pair are obtained by tracing back the operations carried out by the RBMT engine.[10] The lexical weighting scores of each phrase pair are computed from the word alignments obtained from the RBMT engine and the lexical translation model as described by Koehn (2010, Sec. 5.3.3).

Since both phrase tables are computed in a totally independent way, the phrase translation probabilities of the phrase pairs which appear in both phrase tables are not increased over phrase pairs that appear only in one of the phrase tables. Consider, for instance, that the SL phrase a has two different translations according to the RBMT system: b and c. The source-to-target translation probability in the synthetic phrase table for the resulting phrase pairs would be $\phi_{\text{synth}}(b|a) = 0.5$ and $\phi_{\text{synth}}(c|a) = 0.5$. Let us also suppose that, after extracting phrase pairs from the parallel corpus, the phrase

[9]Source-to-target and target-to-source phrase translation probabilities and source-to-target and target-to-source lexical weightings.

[10]The Apertium engine keeps track, on each step of its translation pipeline described in Appendix A, of the input word from which each output word has been obtained. Then, the path starting from each input SL surface form is followed in order to obtain the TL surface form aligned with it. An exception is made when a step of the pipeline converts an input word into multiple output words or vice-versa. In that case, the involved words remain unaligned. This is done to avoid the generation too many word alignments that could be incorrect. Let us suppose that the Spanish sentence *Por otra parte mis amigos americanos han decidido venir* is translated into English as *On the other hand my American friends have decided to come* by Apertium. The Spanish phrase *Por otra parte* is analysed by Apertium as a single lexical form. After being translated into English, it produces the segment *on the other hand* in the generation step. If the exception was not made, the SL word *por* would be aligned with the four TL words *on*, *the*, *other* and *hand* and the SL words *otra* and *parte* would also be aligned with same set of TL words.

pairs (a, b) and (a, d) have the same frequency, and there are not other phrase pairs with a as a source. The resulting source-to-target phrase translation probabilities would be $\phi_{\text{corpus}}(b|a) = 0.5$ and $\phi_{\text{corpus}}(d|a) = 0.5$. Although there is evidence that suggests that b is a more likely translation than c and d, the three translations have the same probability.

3.2.2.2 Phrase table linear interpolation

As an alternative to the previous method, once the two phrase tables have been built, they can be linearly interpolated into a single one following the approach by Sennrich (2012, Sec. 2.1). For each phrase pair (s, t) in the resulting phrase table, each one of the four scores attached to it is obtained as the linear interpolation of the value of that score for the phrase pair (s, t) in the corpus-extracted phrase table and in the synthetic phrase table. For instance, the source-to-target phrase translation probability is computed as shown in the equation below, being $\text{count}_{\text{synth}}(\cdot)$ the frequency of a phrase pair in the list of phrase pairs generated from the RBMT system, $\text{count}_{\text{corpus}}(\cdot)$ the frequency of a phrase pair in the list of phrase pairs extracted from the training parallel corpus and λ_{corpus} and λ_{synth} the weights for both phrase tables; obviously $\lambda_{\text{corpus}} + \lambda_{\text{synth}} = 1$. The weights are optimised by perplexity minimisation on a phrase table built from a development set (Sennrich, 2012, Sec. 2.4).

$$\phi(t|s) = \lambda_{\text{corpus}} \frac{\text{count}_{\text{corpus}}(s, t)}{\sum_{t_i} \text{count}_{\text{corpus}}(s, t_i)} + \lambda_{\text{synth}} \frac{\text{count}_{\text{synth}}(s, t)}{\sum_{t_i} \text{count}_{\text{synth}}(s, t_i)}$$

This combination method, unlike the one that uses two independent phrase tables with independent decoding paths, increases the probability of phrase pairs that are obtained both from the training parallel corpus and from the RBMT system over those phrase pairs that are only present in one of the phrase tables. For the phrase pairs (a, b), (a, c) and (a, d) mentioned above, the resulting probabilities would be $\phi(b|a) = 0.5\lambda_{\text{synth}} + 0.5\lambda_{\text{corpus}} = 0.5$; $\phi(c|a) = 0.5\lambda_{\text{synth}}$; and $\phi(d|a) = 0.5\lambda_{\text{corpus}}$.

However, this method does not use the frequency of the SL phrases in the training parallel corpus, which could help the decoder to select the most appropriate phrase pair. If the SL phrase x is found only once in the training parallel corpus, and it is aligned with y, but its unique translation according to the RBMT system is z, the source-to-target phrase translation probabilities of both phrase pairs would be, respectively, $\phi(y|x) = \lambda_{\text{corpus}}$ and $\phi(z|x) = \lambda_{\text{synth}}$. If x were found 10 000 times in the training parallel corpus, and always translated as y, the probabilities would be exactly the same (since the weights λ_{corpus} and λ_{synth} are the same for all the phrase pairs). However, the phrase pair (x, y) is much more reliable when it is found in the training corpus 10 000 times than when it is found only once. Probabilities in the resulting phrase table should reflect this difference in order to make it easier for the decoder to choose the most appropriate phrase pair.

3.2.2.3 New strategy: directly expanding the phrase table

In order to take into account the absolute frequency of the different phrases in the training parallel corpus, a new scoring method is presented in this section. This new strategy consists of joining the phrase pairs extracted from the parallel corpus and the synthetic phrase pairs, and then calculating phrase translation probabilities by relative frequency as usual (Koehn, 2010, sec. 5.2.5). Thus, the source-to-target phrase translation probabilities of the resulting phrase table are computed as follows:

$$\phi(t|s) = \frac{\text{count}_{\text{corpus}}(s,t) + \text{count}_{\text{synth}}(s,t)}{\sum_{t_i} \text{count}_{\text{corpus}}(s,t_i) + \text{count}_{\text{synth}}(s,t_i)}$$

Since $\text{count}_{\text{synth}}(\cdot) = 1$ for all the synthetic phrase pairs, when a synthetic phrase pair shares its SL side with a corpus-extracted phrase pair, the source-to-target phrase translation probability of the synthetic phrase pair may be too small when compared with the phrase pair extracted from the training corpus.[11] Depending on the texts to be translated with the hybrid system, it may be desirable that a synthetic phrase pair has a higher phrase translation probability than a corpus-extracted phrase pair with the same SL side. In order to adapt their relative weight to the texts to be translated, an additional boolean score to flag synthetic phrase pairs is added to the phrase table.[12]

The lexical weighting scores (Koehn, 2010, Sec. 5.3.3) of the phrase table built with this combination method are obtained as follows. The same lexical translation model is used to compute the lexical weighting scores for both corpus-extracted and synthetic phrase pairs. The model (actually, one model for source-to-target and another model for target-to-source lexical weightings) is obtained from the concatenation of the training parallel corpus and the synthetic phrase pairs generated from the RBMT bilingual dictionary. Then, the lexical weighting scores are computed from the word alignments as described by Koehn (2010, Sec. 5.3.3). The alignments for the phrase

[11] The same applies to phrase pairs that share their TL for the target-to-source phrase translation probability.

[12] In order to take into account the absolute frequencies in the parallel corpora from which the two phrase tables to be combined have been obtained, Sennrich (2012, Sec. 4.2) also defined the *weighted counts* interpolation method, which is similar to the new strategy presented in this chapter. There are two main differences between both approaches. Firstly, in order to adapt the weight of both types of phrases to the texts to be translated, in the *weighted counts* approach the frequency of each phrase pair is multiplied by a factor before building the phrase table. Depending on the origin of the phrase pair, a different factor is used. On the contrary, in the new strategy presented in this chapter, a binary feature function is added to the phrase table. And secondly, in the *weighted counts* the factors that determine the relative weight of each type of phrase pair are optimised by perplexity minimisation on a phrase table built from a development set (Sennrich, 2012, Sec. 2.4) in isolation, i.e. with no connection to the rest of elements present in the log-linear model. In the new strategy, on the contrary, the weight of the binary feature function is optimised together with the rest of elements of the log-linear model during the minimum error rate training (Och, 2003) process. Given the poor results obtained by the phrase table interpolation method —in which the weights are also optimised by perplexity minimisation— in the experiments reported in Section 3.3.2, the *weighted counts* has not been included in the experimental setup.

pairs extracted from the training corpus are those obtained by statistical methods as usual (Koehn, 2010, Sec. 5.2.1), while the word alignments for the synthetic phrase pairs are obtained by tracing back the operations carried out in the different steps of Apertium, in the same way as it is done when using a different phrase table for the synthetic phrase pairs (Section 3.2.2.1).

3.2.2.4 Augmenting the training corpus

Finally, the simplest approach involves appending the RBMT-generated phrase pairs to the training corpus and running the usual phrase-based SMT training algorithm. Unlike in the previous approaches, this improves the alignments of the original training corpus and enriches the lexicalised reordering model (Koehn, 2010, Sec. 5.4.2), in addition to the phrase table. The phrase extraction algorithm (Koehn, 2010, sec. 5.2.3), however, may split the resulting bilingual phrase pairs into smaller units which may cause multi-word expressions not to be translated in the same way as they appear in the RBMT bilingual dictionary.

Although this strategy is not feasible in a real-world environment because of the computational cost of word aligning the whole training corpus for each document to be translated,[13] it is worth evaluating it because it is the only strategy that enriches the data from which the lexicalised reordering model is obtained.

3.3 Evaluating the new hybridisation strategy when using hand-crafted transfer rules

In this section, a set of experiments aimed at evaluating the feasibility of the new hybridisation strategy described in Section 3.2 is presented. In the experiments, it is used for integrating hand-crafted linguistic resources from the Apertium RBMT project into an SMT system. For different language pairs, training corpus sizes and domains, the translation quality achieved by a baseline SMT system, the RBMT system from which the data is extracted, hybrid systems that use the phrase scoring alternatives described in Section 3.2.2, and a hybrid system built by following the approach by Eisele et al. (2008) are compared. First, the experimental setup is described (Section 3.3.1), and then the results are presented and discussed (Section 3.3.2).

[13] Recall that a different set of synthetic phrase pairs is generated for each SL text to be translated by the hybrid system.

3.3.1 Experimental setup

The RBMT–SMT hybridisation approach has been evaluated on the language pairs Breton–French and English↔Spanish,[14] with different training corpus sizes in each case. While the Breton–French language pair suffers from resource scarceness —there are only around 60 000 parallel sentences available (Tyers, 2009; Tiedemann, 2012)—, English↔Spanish have been chosen because they have a wide range of parallel corpora available, which permits performing both in-domain and out-of-domain evaluations. Moreover, as Spanish presents a higher degree of inflection than English, the results from both directions of the English↔Spanish language pair allow us to evaluate in detail the impact of inflection in the hybrid strategy.

The translation model of the SMT systems for English↔Spanish has been trained from the Europarl parallel corpus (Koehn, 2005) version 5,[15] collected from the proceedings of the European Parliament. The TL language model has been trained on the same corpus. In both cases, the Q4/2000 portion has not been used for training, since it was set aside for evaluation purposes. Different subsets of the parallel corpus with different number of sentences have been used to build systems; however, in all cases the language model has been trained on the whole TL side of the Europarl corpus. These subsets have been chosen randomly but in such a way that larger corpora include the sentences in the smaller ones. The sizes of the different subcorpora are 10 000, 40 000, 160 000, 600 000 sentences, and the whole corpus (1 272 260 sentences).

Regarding Breton–French, the translation model has been built using the only freely-available parallel corpus for such language pair (Tyers, 2009; Tiedemann, 2012), which contains short sentences from the tourism and computer localisation domains. Different training corpus sizes have been used too, namely 10 000, 25 000 and 54 196 parallel sentences. The latter corresponds to the whole corpus except for the subsets reserved for development and testing. As in the English↔Spanish pair, sentences have been randomly chosen but in such as way that larger corpora include the sentences in the smaller ones. The TL model has been learned from a monolingual corpus built by concatenating the target side of the whole parallel training corpus and the French Europarl corpus provided for the WMT 2011 shared translation task.[16]

Although there are more monolingual corpora available for the target languages included in the evaluation setup, they have not been used in the experiments described in this section. The experiments are focused on evaluating the impact of the RBMT data in the SMT translation model. By learning the TL model from a monolingual corpus that do not exceeds the size of the biggest parallel corpus used in the experiments, the risk that a huge language model shadows the impact of the RBMT data on

[14] The symbol ↔ means that the evaluation has been performed in both translation directions: from English to Spanish and from Spanish to English.

[15] `http://www.statmt.org/europarl/archives.html#v5`

[16] `http://www.statmt.org/wmt11/translation-task.html`

the SMT translation model is reduced. Nevertheless, Section 3.5 presents a set of experiments that have been performed with bigger monolingual corpora, like those used in the aforementioned WMT shared translation task. Note also that, in a real-world environment, the size of the TL model may need to be limited if the hybrid MT system is required to have a reduced memory footprint because it is going to be executed in a handheld device.

Breton–French systems have been tuned using 3 000 parallel sentences randomly chosen from the available parallel corpus and evaluated using another randomly chosen subset of the same size. Both subsets do not intersect and were removed from the training section of the corpus. Note that only an in-domain evaluation can be performed for this language pair. Regarding English↔Spanish, both in-domain and out-of-domain evaluations have been carried out. The former has been performed by tuning the systems with 2 000 parallel sentences randomly chosen from the Q4/2000 portion of Europarl v5 corpus (Koehn, 2005) and evaluating them with 2 000 random parallel sentences from the same portion of the corpus; special care has been taken to avoid the overlapping between the test and development sets. The out-of-domain evaluation has been performed by using the *newstest2008* set for development and the *newstest2010* test for testing; both sets belong to the news domain and are distributed as part of the WMT 2010 shared translation task.[17] Table 3.1 summarises the data about the corpora used in the experiments. Sentences that contain more than 40 tokens have been removed from all the parallel corpora, as it is customary, in order to avoid problems with the word alignment tool GIZA++ (Och and Ney, 2003).[18]

All the experiments have been carried out with the free/open-source phrase-based SMT system Moses[19] (Koehn et al., 2007) together with the SRILM language modelling toolkit (Stolcke, 2002), which has been used to train a 5-gram language model using interpolated Kneser-Ney discounting (Goodman and Chen, 1998). Word alignments have been computed by means of GIZA++ (Och and Ney, 2003). The weights of the different feature functions have been optimised by means of minimum error rate training (Och, 2003). The parallel corpora have been lowercased prior to training, as well as the test sets used for evaluating the systems.

The hand-crafted shallow-transfer rules and dictionaries have been borrowed from the Apertium project (Forcada et al., 2011). In particular, the engine and the linguistic resources for English–Spanish, and Breton–French have been downloaded from the Apertium Subversion repository.[20] The Apertium linguistic data contains 326 228 entries in the English–Spanish bilingual dictionary, 284 English–Spanish shallow-transfer

[17] `http://www.statmt.org/wmt10/translation-task.html`

[18] They have been also removed from the development and test sets in order to ensure that the approach by Eisele et al. (2008) is able to extract all the needed phrase pairs. Recall that this method needs to align the sentences in the test set with their RBMT translations.

[19] Release 2.1, downloaded from `https://github.com/moses-smt/mosesdecoder/tree/RELEASE-2.1`.

[20] Revisions 24177, 22150 and 28674, respectively.

Table 3.1: Number of sentences, words, and size of the vocabulary of the training, development and test sets used in the experiments.

Corpus	#sentences	SL		TL	
		# words	# voc	# words	# voc
Language model (English)	1 650 152	-	-	45 712 294	110 018
Language model (Spanish)	1 650 152	-	-	47 734 244	165 896
training	10 000	209 562	11 561	216 187	15 884
	40 000	836 194	20 883	862 789	30 583
	160 000	3 341 577	36 798	3 452 067	55 584
	600 000	12 546 758	61 654	12 971 035	94 315
	1 272 260	26 595 542	82 585	27 496 270	125 813
Europarl development	2 000	42 642	5 157	43 348	6 411
Europarl testing	2 000	42 114	5 080	42 661	6 289
newstest2012 development	1 732	34 878	6 209	36 410	7 085
newstest2013 testing	2 215	48 367	7 701	50 745	9 277

(a) English↔Spanish

Corpus	#sentences	SL		TL	
		# words	# voc	# words	# voc
Language model (French)	2 041 625	-	-	60 356 583	155 028
training	10 000	146 255	16 711	146 556	17 588
	25 000	365 856	27 606	369 396	28 333
	54 196	795 045	41 157	801 780	40 279
development	3 000	44 586	8 340	45 086	8 907
testing	3 000	44 586	8 340	45 086	8 907

(b) Breton–French

rules and 138 Spanish–English shallow-transfer rules. Regarding Breton–French, the bilingual dictionary contains 21 593 entries and there are 254 shallow-transfer rules.[21]

For each language pair, domain, and training corpus size, the following systems have been built and evaluated:

- *baseline*: a standard phrase-based SMT system.
- *Apertium*: the Apertium shallow-transfer RBMT engine, from which the dictionaries and transfer rules have been borrowed.
- *extended-phrase*: the hybrid system described in Section 3.2 following the new strategy for scoring the phrase pairs generated from the RBMT data described in Section 3.2.2.3. This strategy involves adding the synthetic phrase pairs (a single instance of each phrase pair is added) to the list of phrase pairs extracted from the training parallel corpus (with the frequency observed in the corpus) and scoring the phrase pairs by relative frequency.
- *extended-phrase-dict*: the same as above, but using only the dictionaries of the RBMT system (without shallow-transfer rules). The comparison between this system and *extended-phrase* allows us to evaluate the impact of the use of shallow-transfer rules.
- *two-phrase-tables*: the hybrid system described in Section 3.2 following the strategy for scoring the synthetic phrase pairs based on two independent phrase tables (Koehn and Schroeder, 2007), described in Section 3.2.2.1.
- *interpolation*: the hybrid system described in Section 3.2 following the strategy for scoring the synthetic phrase pairs based on the linear interpolation of two phrase tables (Sennrich, 2012, Sec. 2.1), described in Section 3.2.2.2. The interpolation weights have been obtained by perplexity minimisation on a phrase table built from the development set.
- *extended-corpus*: the hybrid system described in Section 3.2 following the strategy for scoring the synthetic phrase pairs which simply involves adding the synthetic phrase pairs to the training corpus (Section 3.2.2.4).
- *Eisele*: the approach by Eisele et al. (2008), using the alignment model learned from the training corpus to get the word alignments between the SL sentences and the RBMT-translated sentences.

[21] The transfer step is split by Apertium in three steps for the language pairs evaluated in this experimental setup, and each step works with its own set of rules (see Appendix A for a description of the different types of shallow-transfer rules in Apertium). Specifically, the Apertium linguistic data contains 216 chunker rules, 60 interchunk rules, and 7 postchunk rules for English–Spanish; 106 chunker rules, 31 interchunk rules, and 7 postchunk rules for Spanish–English; and 169 chunker rules, 79 interchunk rules and 6 postchunk rules for Breton–French.

3.3.2 Results and discussion

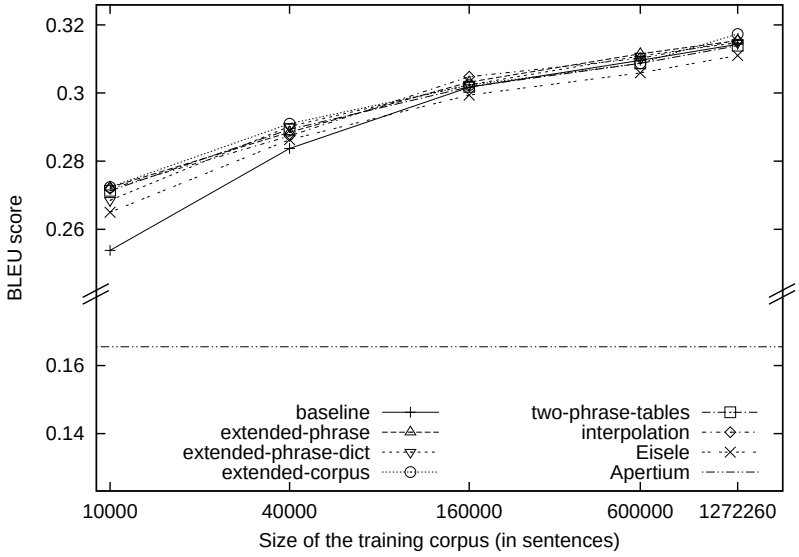
Figures 3.1–3.5 show the BLEU (Papineni et al., 2002), TER (Snover et al., 2006) (the figures represent 1-TER as the value of TER is inversely proportional to translation quality) and METEOR (Banerjee and Lavie, 2005) automatic evaluation scores for the systems evaluated. In addition, the statistical significance of the difference between the hybridisation approach *extended-phrase* described in Section 3.2.2.3, and the other systems has been computed with paired bootstrap resampling (Koehn, 2004b) ($p \leq 0.05$; 1 000 iterations).[22] The table appended to each set of figures contains the results of the pair-wise comparison. Each cell represents the reference system with which the approach *extended-phrase* is compared and the training corpus size, and it contains the results for the three evaluation metrics: BLEU(B), TER(T) and METEOR(M). An arrow pointing upwards (⇑) means that *extended-phrase* outperforms the reference system, an arrow pointing downwards (⇓) means that the reference system outperforms *extended-phrase*, and an equal sign (=) means that the difference between both systems is not statistically significant.

These results show that the new hybrid approach described in Section 3.2 (*extended-phrase*) outperforms by a statistically significant margin both the RBMT and the baseline phrase-based SMT system in different scenarios. Namely, when translating out-of-domain texts (texts whose domain is different from the domain of the parallel corpus used; this happens for all training corpus sizes and language pairs) and when translating in-domain texts with an SMT system trained on a relatively small parallel corpus. Thus, it is confirmed that shallow-transfer RBMT and SMT systems can be combined in a hybrid system that outperforms both of them.

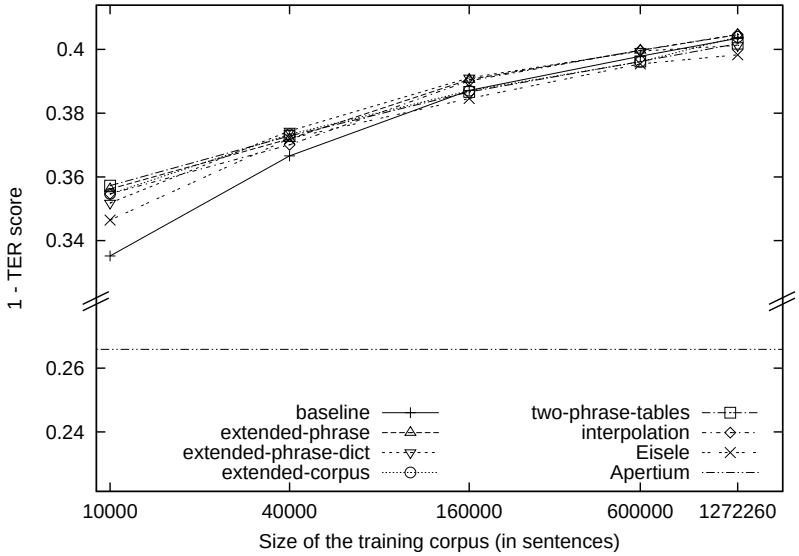
Concerning the differences observed in the results between in-domain and out-of-domain evaluation, it is important to remark that, for English↔Spanish, the out-of-domain development and test sets come from a general (news) domain and the RBMT data has been developed bearing in mind the translation of general texts (mainly news). In this case, Apertium-generated phrases which contain hand-crafted knowledge from a general domain cover sequences of words in the input text which are not covered, or are sparsely found, in the original training corpora. Contrarily, the in-domain tests reveal that, as soon as the phrase-based SMT system is able to learn some reliable information from the parallel corpus, the synthetic RBMT phrases become useless because the in-domain test sets come from the specialised domain of parliament speeches. For Breton–French, given the small size of the corpus available, the hybrid approach outperforms both pure RBMT and SMT approaches in all the experiments performed.

[22] Only the strategy *extended-phrase* is compared with the other systems because it is expected to achieve the highest translation quality among the different hybrid approaches, as in theory it overcomes most of the limitations of the other approaches (see Section 3.2.2).

Figure 3.1: Automatic evaluation scores obtained by the baseline phrase-based SMT system, Apertium, the hybrid approaches described in Section 3.2.2, and the hybrid approach by Eisele et al. (2008) for the English–Spanish language pair (in-domain evaluation). The table represents the results of the paired bootstrap resampling comparison (Koehn, 2004b) with the system *extended-phrase* (described in Section 3.2.2.3).

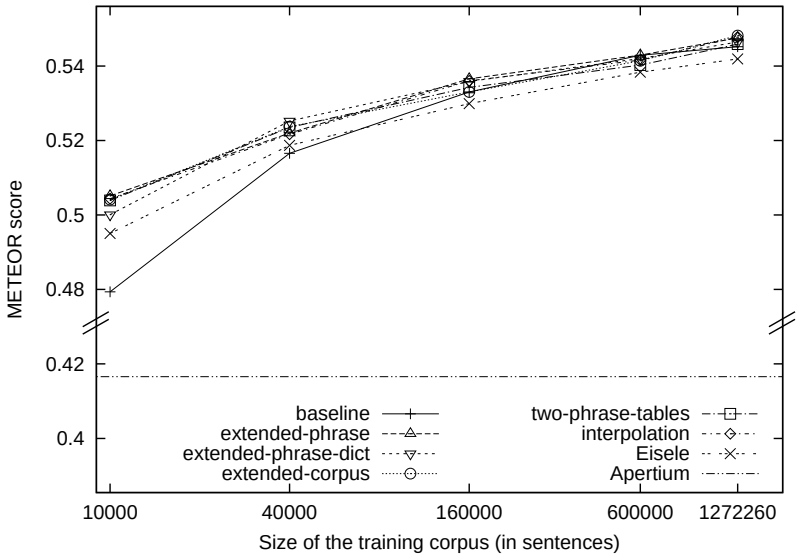


(a) BLEU scores.



(b) 1−TER scores.

(continued in next page)

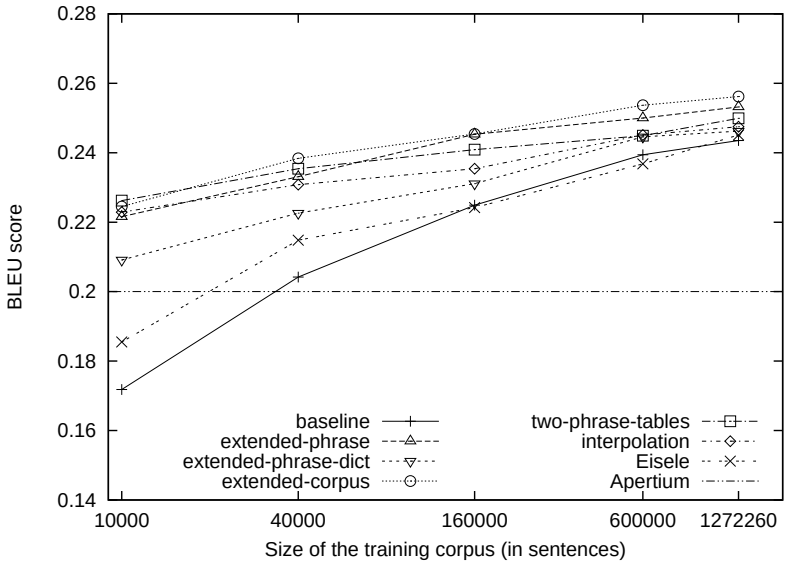


(c) METEOR scores.

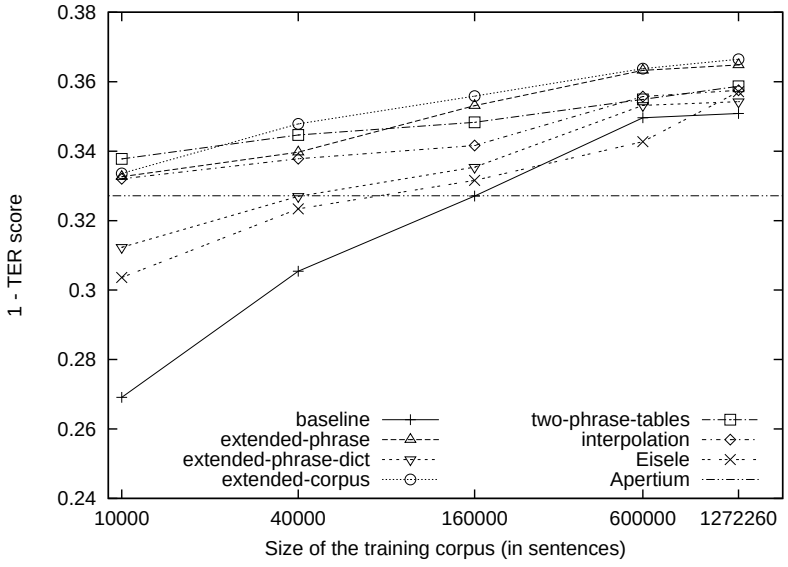
system	10 000	40 000	160 000	600 000	1 272 260
metric	B T M	B T M	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	= ⇑ ⇑	= = =	= = ⇑
Apertium	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Eisele	⇑ ⇑ ⇑	= = ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	⇑ ⇑ ⇑	= = ⇓	= = =	= = =	= ⇑ =
extended-corpus	= = =	= = =	= ⇑ ⇑	⇑ ⇑ =	= = =
two-phrase-tables	= = =	= = =	= ⇑ ⇑	= ⇑ ⇑	= ⇑ =
interpolation	= = =	= = =	= = =	= = =	= = =

(d) Results of the paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between the *extended-phrase* hybridisation strategy and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU(B), TER(T) and METEOR(M). An arrow pointing upwards means that *extended-phrase* outperforms the reference method by a statistically significant margin, an arrow pointing downwards means the opposite and an equal sign means that there are not statistically significant differences between the systems.

Figure 3.2: Automatic evaluation scores obtained by the baseline phrase-based SMT system, Apertium, the hybrid approaches described in Section 3.2.2, and the hybrid approach by Eisele et al. (2008) for the English–Spanish language pair (out-of-domain evaluation). The table represents the results of the paired bootstrap resampling comparison (Koehn, 2004b) with the system *extended-phrase* (described in Section 3.2.2.3).

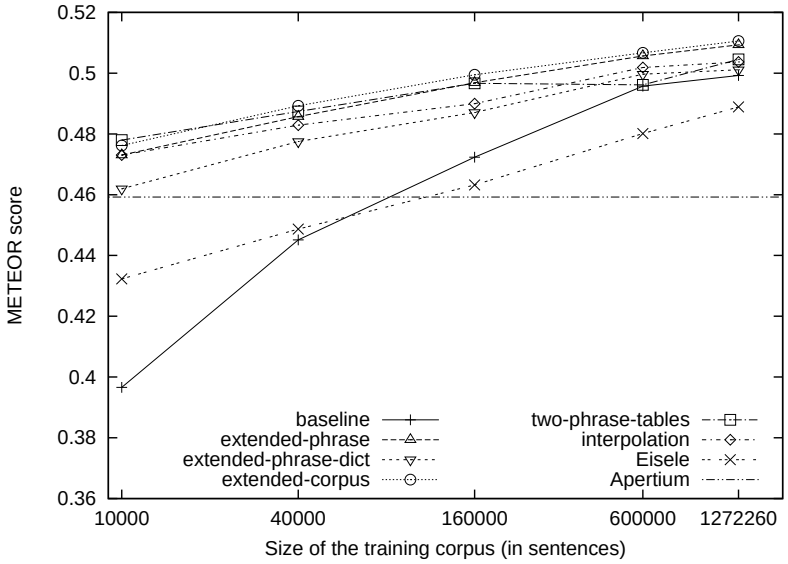


(a) BLEU scores.



(b) 1−TER scores.

(continued in next page)

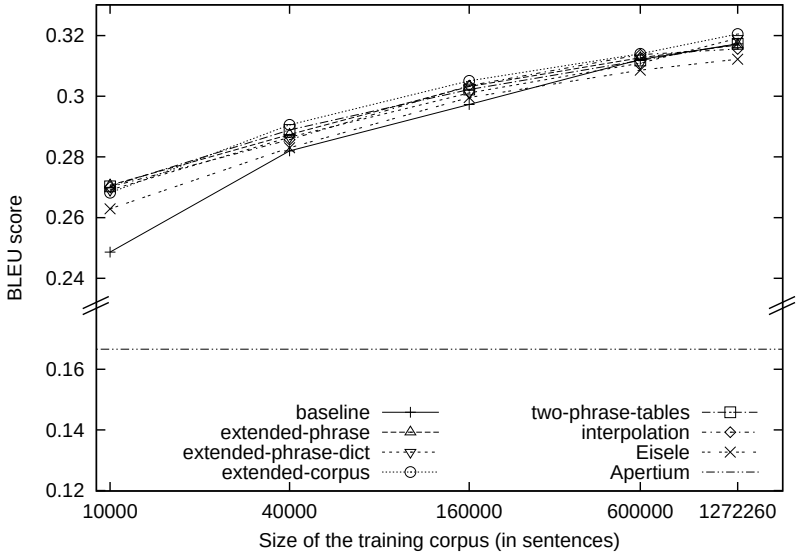


(c) METEOR scores.

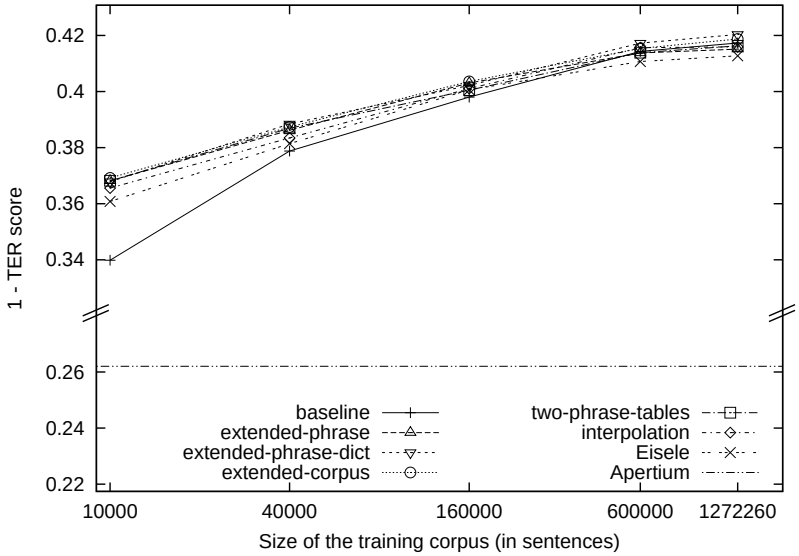
system	10 000	40 000	160 000	600 000	1 272 260
metric	B T M	B T M	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Apertium	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Eisele	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-corpus	= = ⇓	⇓ ⇓ ⇓	= = ⇓	⇓ = =	⇓ = =
two-phrase-tables	⇓ ⇓ ⇓	= ⇓ =	⇑ ⇑ =	⇑ ⇑ ⇑	⇑ ⇑ ⇑
interpolation	= = =	= = ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑

(d) Results of the paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between the *extended-phrase* hybridisation strategy and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU(B), TER(T) and METEOR(M). An arrow pointing upwards means that *extended-phrase* outperforms the reference method by a statistically significant margin, an arrow pointing downwards means the opposite and an equal sign means that there are not statistically significant differences between the systems.

Figure 3.3: Automatic evaluation scores obtained by the baseline phrase-based SMT system, Apertium, the hybrid approaches described in Section 3.2.2, and the hybrid approach by Eisele et al. (2008) for the Spanish–English language pair (in-domain evaluation). The table represents the results of the paired bootstrap resampling comparison (Koehn, 2004b) with the system *extended-phrase* (described in Section 3.2.2.3).

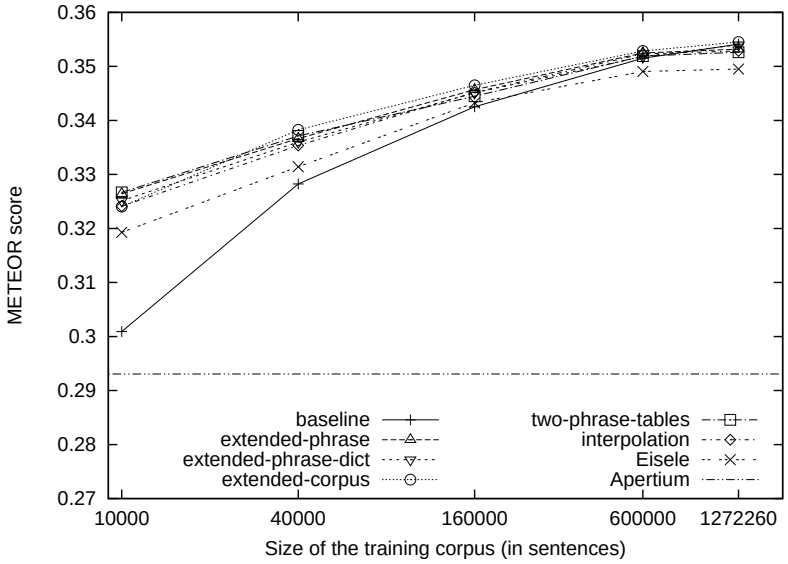


(a) BLEU scores.



(b) 1−TER scores.

(continued in next page)

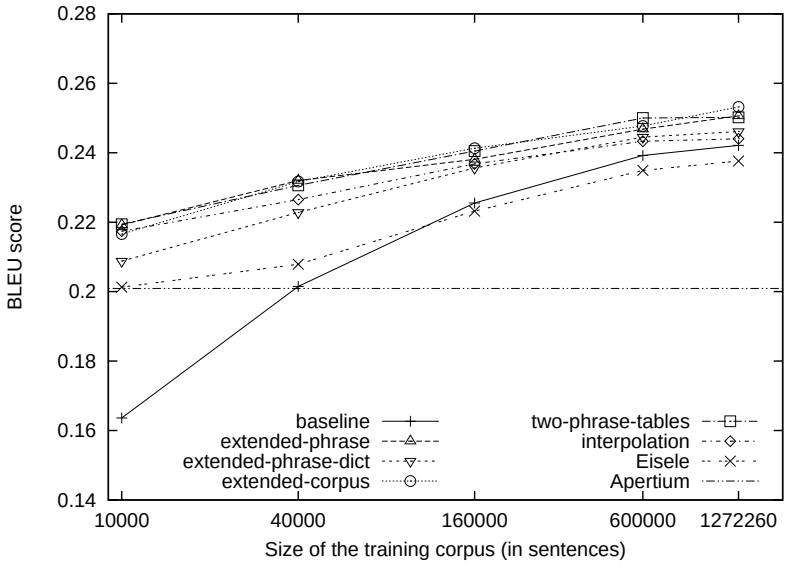


(c) METEOR scores.

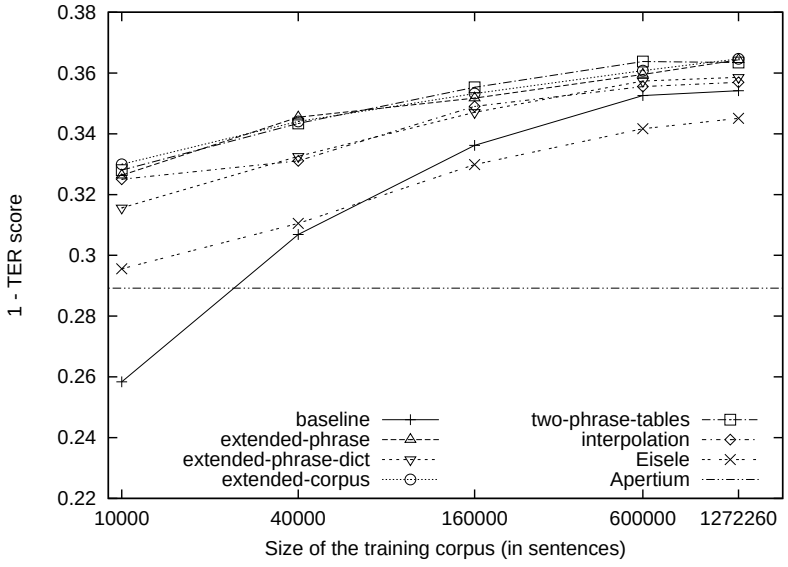
system	10 000	40 000	160 000	600 000	1 272 260
metric	B T M	B T M	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	= = =	= = =
Apertium	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Eisele	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ = ⇑	⇑ = ⇑	⇑ = ⇑
extended-phrase-dict	= = ⇑	= = =	= = =	= ⇓ =	= ⇓ =
extended-corpus	= = ⇑	⇓ = ⇓	= = =	= = =	⇓ ⇓ ⇓
two-phrase-tables	= = =	= = =	= = ⇑	= = =	= = =
interpolation	= = ⇑	= = =	= = =	= = =	= = =

(d) Results of the paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between the *extended-phrase* hybridisation strategy and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU(B), TER(T) and METEOR(M). An arrow pointing upwards means that *extended-phrase* outperforms the reference method by a statistically significant margin, an arrow pointing downwards means the opposite and an equal sign means that there are not statistically significant differences between the systems.

Figure 3.4: Automatic evaluation scores obtained by the baseline phrase-based SMT system, Apertium, the hybrid approaches described in Section 3.2.2, and the hybrid approach by Eisele et al. (2008) for the Spanish–English language pair (out-of-domain evaluation). The table represents the results of the paired bootstrap resampling comparison (Koehn, 2004b) with the system *extended-phrase* (described in Section 3.2.2.3).

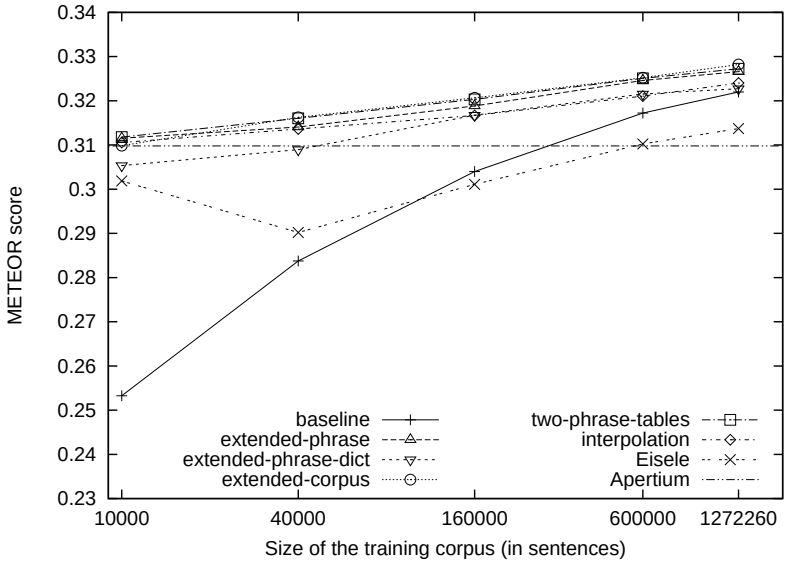


(a) BLEU scores.



(b) 1−TER scores.

(continued in next page)

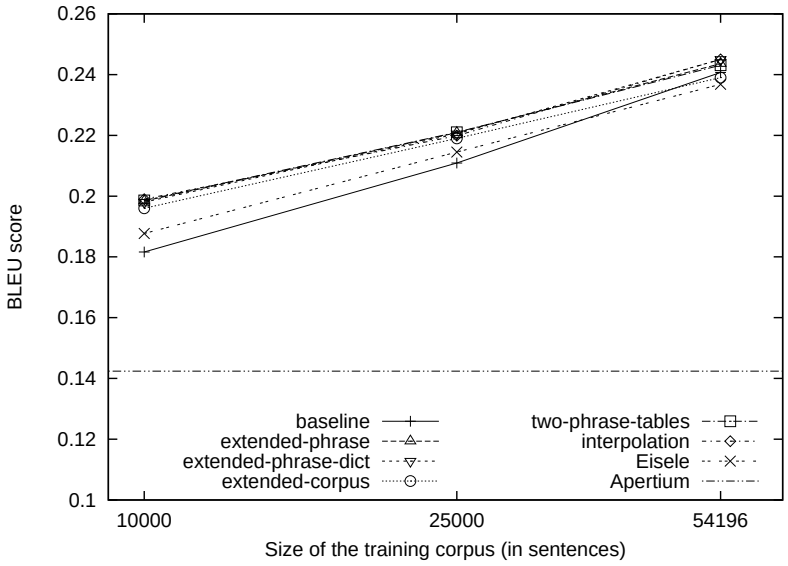


(c) METEOR scores.

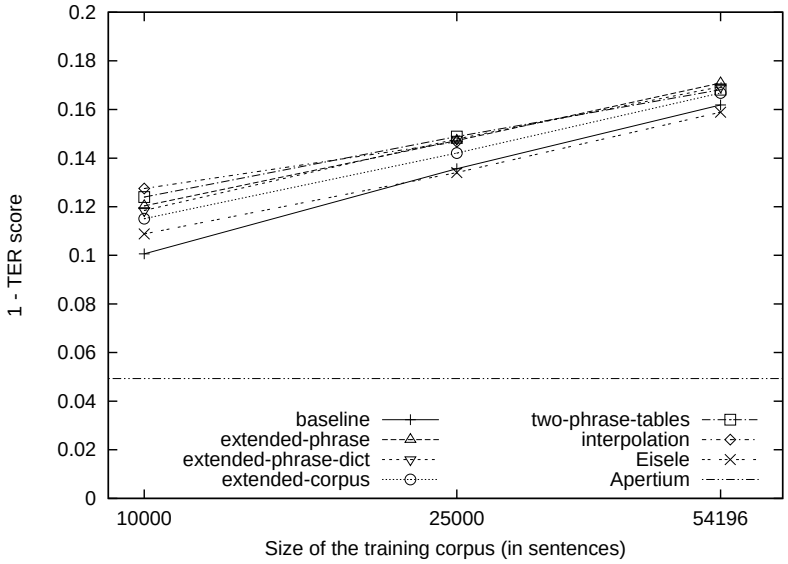
system	10 000	40 000	160 000	600 000	1 272 260
metric	B T M	B T M	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Apertium	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Eisele	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	= = ⇑	⇑ ⇑ ⇑
extended-corpus	= ⇓ ⇑	= = ⇓	⇓ = ⇓	= = =	= = ⇓
two-phrase-tables	= = =	= = ⇓	= ⇓ ⇓	⇓ ⇓ =	= = =
interpolation	= = =	⇑ ⇑ =	= = ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑

(d) Results of the paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between the *extended-phrase* hybridisation strategy and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU(B), TER(T) and METEOR(M). An arrow pointing upwards means that *extended-phrase* outperforms the reference method by a statistically significant margin, an arrow pointing downwards means the opposite and an equal sign means that there are not statistically significant differences between the systems.

Figure 3.5: Automatic evaluation scores obtained by the baseline phrase-based SMT system, Apertium, the hybrid approaches described in Section 3.2.2, and the hybrid approach by Eisele et al. (2008) for the Breton–French language pair (in-domain evaluation). The table represents the results of the paired bootstrap resampling comparison (Koehn, 2004b) with the system *extended-phrase* (described in Section 3.2.2.3).

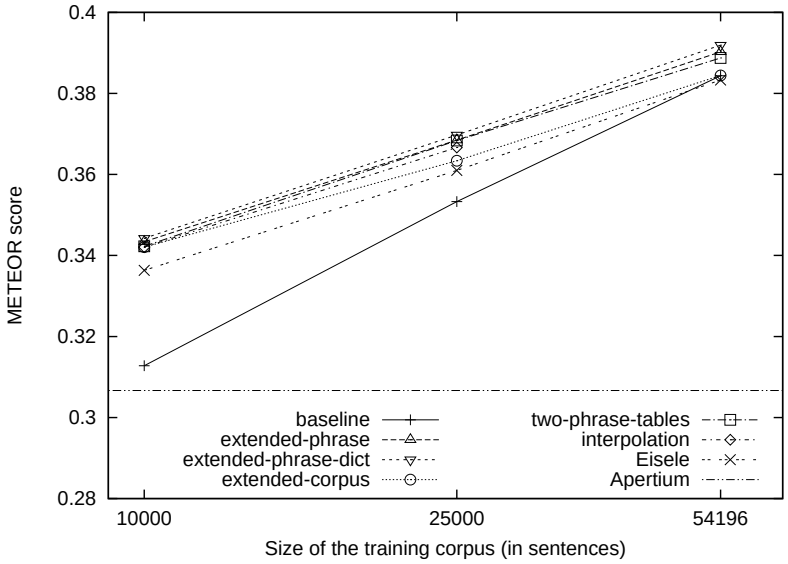


(a) BLEU scores.



(b) 1−TER scores.

(continued in next page)

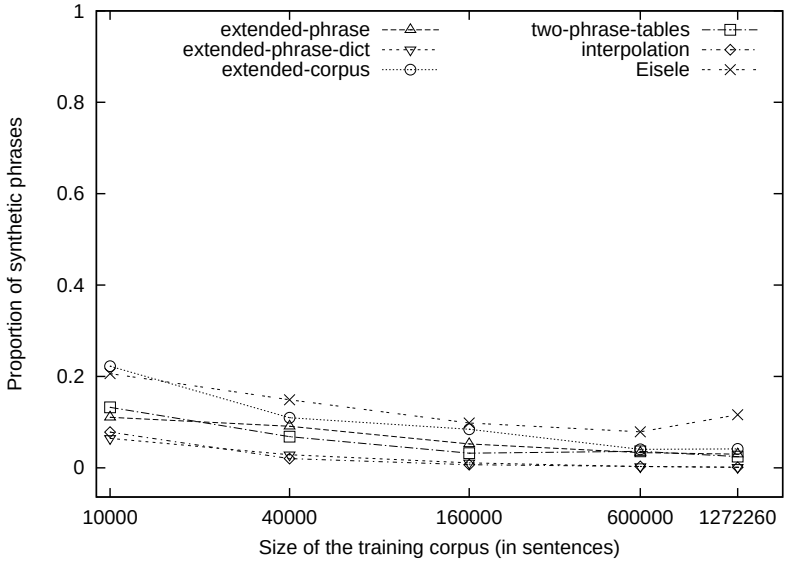


(c) METEOR scores.

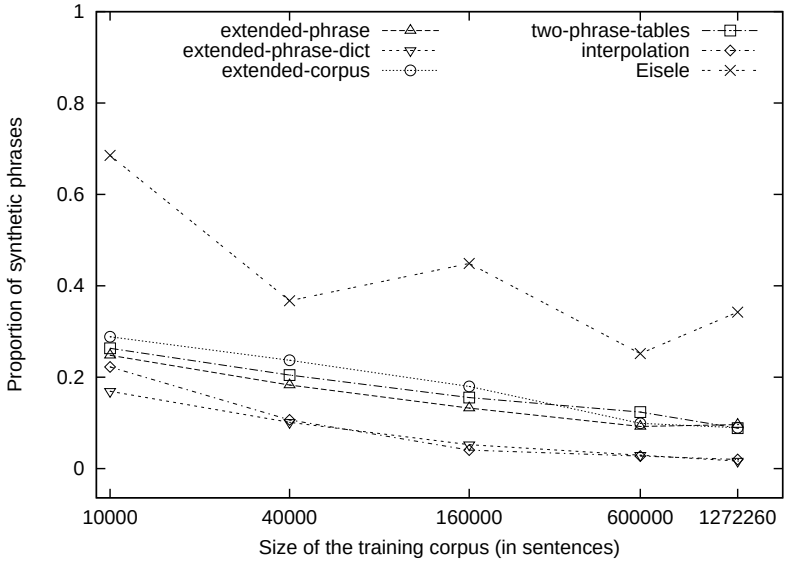
system	10 000			25 000			54 196		
metric	B	T	M	B	T	M	B	T	M
baseline	⇑	⇑	⇑	⇑	⇑	⇑	=	⇑	⇑
Apertium	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑
Eisele	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑
extended-phrase-dict	=	=	=	=	=	=	=	=	=
extended-corpus	=	⇑	=	=	⇑	⇑	⇑	=	⇑
two-phrase-tables	=	⇓	=	=	=	=	=	=	=
interpolation	=	⇓	=	=	=	⇑	=	=	=

(d) Results of the paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between the *extended-phrase* hybridisation strategy and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU(B), TER(T) and METEOR(M). An arrow pointing upwards means that *extended-phrase* outperforms the reference method by a statistically significant margin, an arrow pointing downwards means the opposite and an equal sign means that there are not statistically significant differences between the systems.

Figure 3.6: Proportion of phrase pairs generated from the RBMT data chosen by the decoder when translating the test set for the different hybrid approaches described in Section 3.2.2, and the hybrid approach by Eisele et al. (2008) for the English–Spanish language pair.

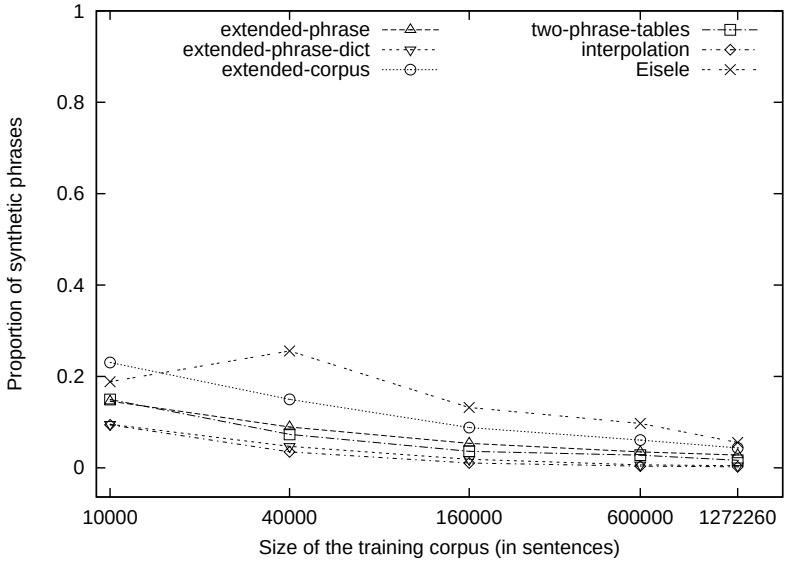


(a) In-domain evaluation.

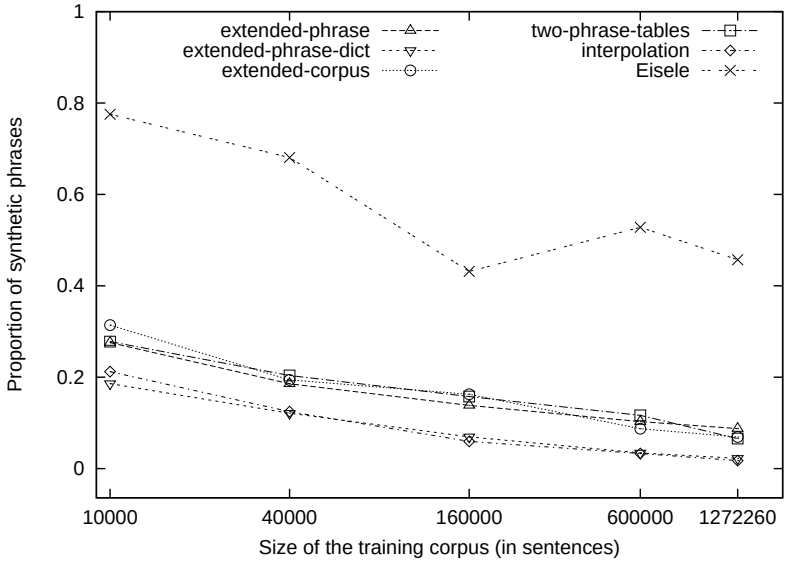


(b) Out-of-domain evaluation.

Figure 3.7: Proportion of phrase pairs generated from the RBMT data chosen by the decoder when translating the test set for the different hybrid approaches described in Section 3.2.2, and the hybrid approach by Eisele et al. (2008) for the Spanish–English language pair.

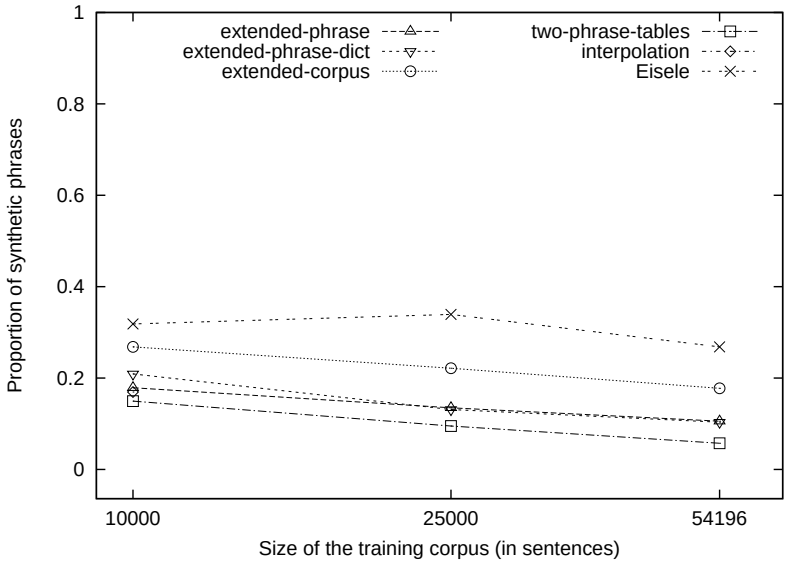


(a) In-domain evaluation.



(b) Out-of-domain evaluation.

Figure 3.8: Proportion of phrase pairs generated from the RBMT data chosen by the decoder when translating the test set for the different hybrid approaches described in Section 3.2.2, and the hybrid approach by Eisele et al. (2008) for the Breton–French language pair.

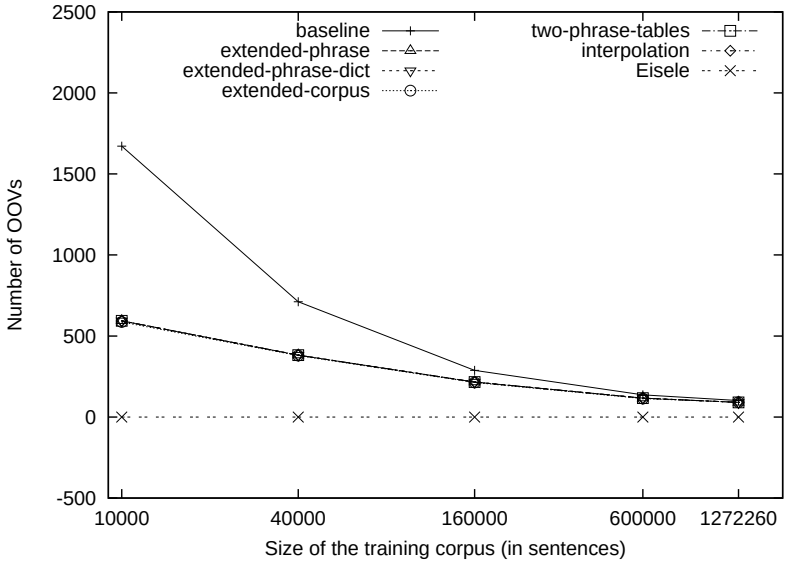


An analysis of the proportion of synthetic phrase pairs included by the decoder in the final translation[23] for the different evaluation scenarios, depicted in figures 3.6–3.8, confirms the reason of the differences between in-domain and out-of-domain evaluations. For all the English↔Spanish training corpus sizes and hybrid systems, the proportion of synthetic phrases is higher in the out-of-domain evaluation. The fact that the proportion of synthetic phrase pairs used drops faster in the in-domain evaluation is also remarkable.

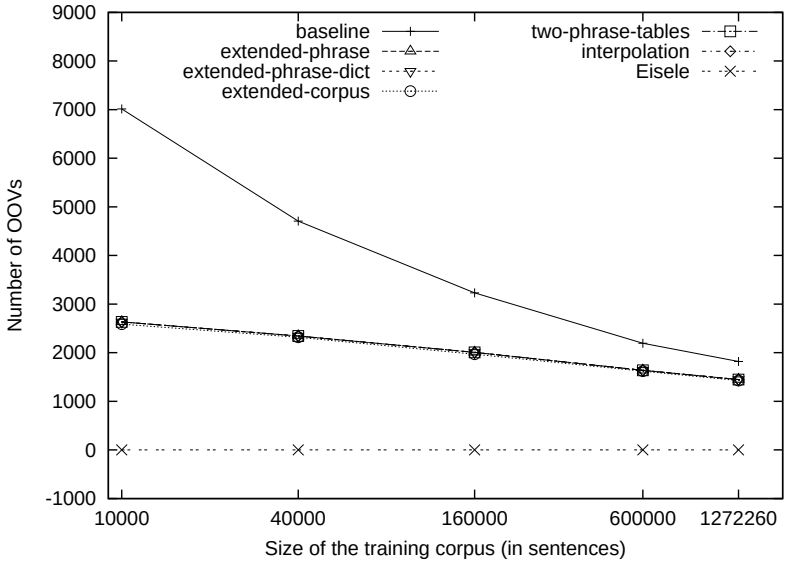
Regarding the difference between the hybrid systems enriched with all the RBMT resources (*extended-phrase*) and those only including the dictionary (*extended-phrase-dict*), some patterns can be detected. For English↔Spanish, the impact of the shallow-transfer rules is higher when translating out-of-domain texts and decreases as the training corpus grows. Their impact is therefore higher when the decoder chooses a high proportion of Apertium phrases, according to figures 3.6 and 3.7. Moreover, the systems including shallow-transfer rules outperform their counterparts which only include the dictionary by a wider margin when translating out-of-domain texts from English to Spanish than the other way round. As Spanish morphology is richer, transfer rules help to perform more agreement operations when translating into Spanish. On the contrary, when Spanish is the SL, one of the main limitations suffered by the baseline SMT system is the high number of out-of-vocabulary (OOV) words, which is already mitigated by integrating the dictionaries into the phrase table with the

[23]If a synthetic phrase pair has also been obtained from the parallel corpus, it is not considered as synthetic in figures 3.6–3.8.

Figure 3.9: Number of out-of-vocabulary words found in the SL side of the test set for the baseline phrase-based SMT system, the different hybrid approaches described in Section 3.2.2, and the hybrid approach by Eisele et al. (2008) for the English–Spanish language pair.

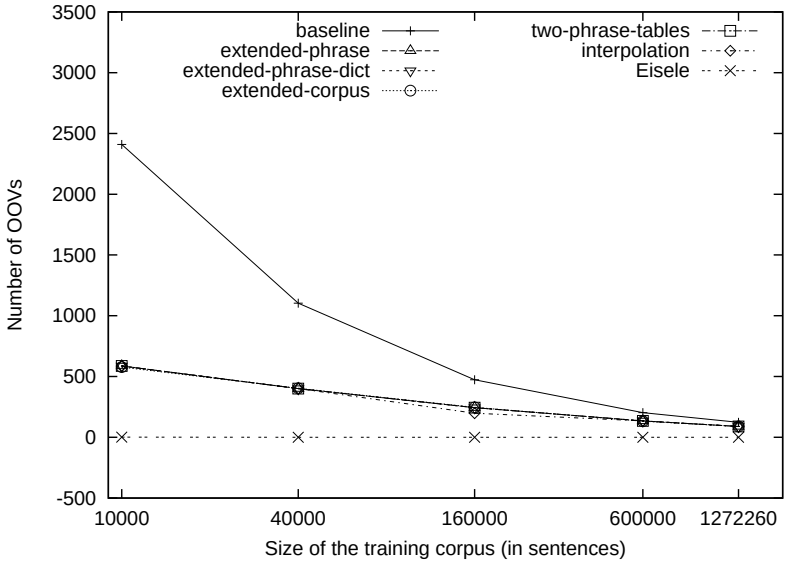


(a) In-domain evaluation.

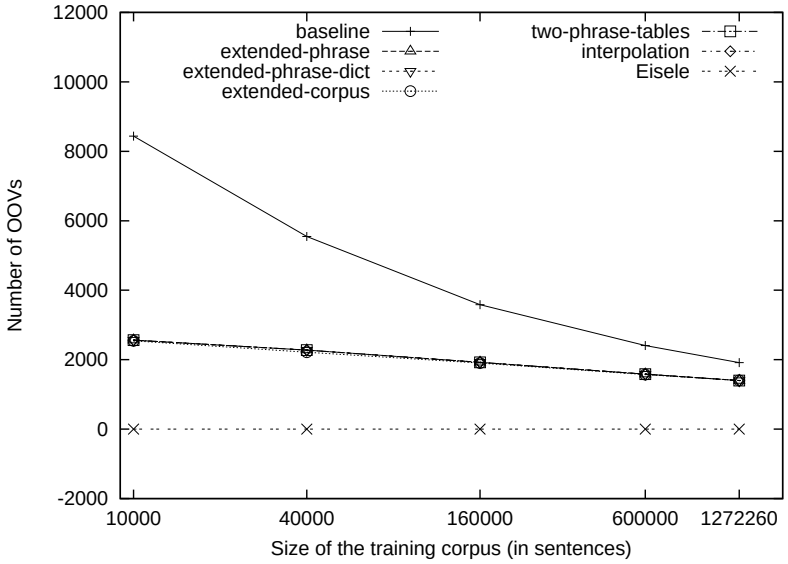


(b) Out-of-domain evaluation.

Figure 3.10: Number of out-of-vocabulary words found in the SL side of the test set for the baseline phrase-based SMT system, the different hybrid approaches described in Section 3.2.2, and the hybrid approach by Eisele et al. (2008) for the Spanish–English language pair.

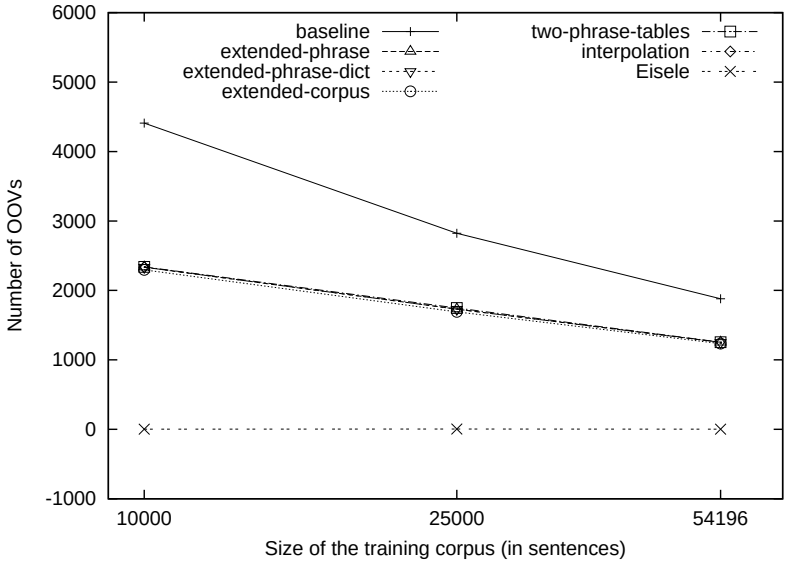


(a) In-domain evaluation.



(b) Out-of-domain evaluation.

Figure 3.11: Number of out-of-vocabulary words found in the SL side of the test set for the baseline phrase-based SMT system, the different hybrid approaches described in Section 3.2.2, and the hybrid approach by Eisele et al. (2008) for the Breton–French language pair.



extended-phrase-dict approach. Figures 3.9–3.11 show the number of OOVs found in the test set for the different evaluation setups.[24] It can be clearly observed that it is much higher for the baseline system when the SL is Spanish than when the SL is English. Consequently, the reduction in the amount of OOVs when adding the RBMT dictionaries is also higher in the first case.

On the contrary, the positive impact of the rules is very limited in the English↔ Spanish in-domain evaluation, where a statistically significant improvement over the hybrid system enriched only with dictionaries (according to the three evaluation metrics) can only be observed for the smallest English–Spanish training corpus. In fact, for a few training corpus sizes the inclusion of the shallow-transfer rules in the hybrid system produces a statistically significant drop in translation quality, according to one of the three evaluation metrics (METEOR in the case of English–Spanish in-domain evaluation and TER for Spanish–English). As it has been pointed out previously, when the training parallel corpus belongs to the same domain as the test corpus, phrase pairs extracted from the training corpus are likely to contain more accurate and fluent translations compared to the mechanical and regular translations provided by the RBMT shallow-transfer rules. One possible explanation to the fact that the degradation caused by the rules is only measured by TER or METEOR is the way in which the MERT tuning (Och, 2003) works. It uses BLEU as its evaluation metric and thus the weight of

[24]In the approach by Eisele et al. (2008) the number of out-of-vocabulary words is always 0 because the phrase table of the hybrid system has been enriched from a synthetic corpus obtained by translating the test set with the RBMT system. The phrase table contains entries even for SL words that are unknown to the RBMT system. These SL words are not translated by the RBMT system, but simply printed without any change.

the feature function which tags whether a phrase pairs comes from the parallel corpus or from the RBMT system is set so that the inclusion of shallow-transfer rules does not penalise translation quality as measured by BLEU.

With regard to Breton–French, the impact of the shallow-transfer rules is also limited: the difference between the hybrid system enriched with shallow-transfer rules and the system enriched only with dictionaries is not statistically significant for any of the training corpus subsets evaluated. The reason is probably that the sentences from the test set do not have a complex grammatical structure: the average sentence length is about 9 words (Tyers, 2009) and it contains many sentences that are simply noun phrases. Another possible cause, already outlined in the previous chapter (Section 2.6), is the fact that the quality of the Breton–French shallow-transfer rules may be lower than the rules used for other language pairs.

As regards the different phrase scoring approaches defined in Section 3.2.2, some differences can be found between them. The most remarkable differences can be observed in the evaluation setups where the inclusion of synthetic phrase pairs has a great impact, that is, in English↔Spanish out-of-domain evaluations. Firstly, the *interpolation* strategy is frequently outperformed by other strategies, and the hybrid systems built with it usually choose a relatively small proportion of synthetic phrases. In theory, it should outperform the *two-phrase-tables* strategy because it assigns higher probabilities to synthetic phrase pairs that are also found in the training parallel corpus, but actually the *two-phrase-tables* approach generally achieves a higher translation quality. A possible cause of this result may be the fact that, while in the *interpolation* method the relative weights of the two types of phrase pairs are optimised so as to minimise the perplexity on a set of phrase pairs extracted from a development corpus, in the *two-phrase-tables* strategy the relative weights are optimised so as to maximise translation quality by the minimum error rate training (Och, 2003) algorithm. In the latter strategy, the interaction of the phrase pairs with the rest of elements of the SMT system is taken into account during the optimisation process. Nevertheless, additional experiments aimed at deeply evaluating the impact of the method used for optimising the relative weight of both types of phrase pairs will need to be carried out. Concerning the *extended-corpus* strategy, it does not consistently outperform the other strategies, probably because the synthetic bilingual phrase pairs were too short to clearly improve the reordering model. Anyway, as already said, this strategy could not be used in a real-world system because of the high computational cost of aligning together the synthetic phrase pairs and the training corpus for every document to be translated. Finally, the *two-phrase-tables* strategy is systematically outperformed by the new *extended-phrase* strategy in the experiments carried out with the English–Spanish language pair, although for the reverse language pair, the *two-phrase-tables* wins more often. However, the difference between both is much larger and the three evaluation metrics agree more often in English–Spanish. These results suggest that, at least in the evaluation scenario where the shallow-transfer rules have the highest

impact, the new phrase scoring strategy defined in Section 3.2.2.3 is able to achieve a better balance between the two sources of phrase pairs.

Finally, the hybridisation strategy defined in Section 3.2 (with the phrase scoring strategy defined in Section 3.2.2.3) outperforms the approach by Eisele et al. (2008) for all language pairs, training corpus sizes and domains. The biggest difference between both approaches is observed when small corpora are used for training. As it has been anticipated in Section 3.1.1, under such circumstances, no reliable alignment models can be learned from the training corpus and therefore no reliable phrases pairs can be obtained from the input text and its rule-based translation. The approach presented in this chapter, contrarily, is not affected by this issue because it does not rely on word alignments in order to generate phrase pairs from the RBMT system. In addition, there is a significant difference even when the training corpus is relatively big (more than one million parallel sentences). This fact, together with the high proportion of synthetic phrase pairs used when compared to the other hybrid approaches (see figures 3.6–3.8), suggests that the method for scoring phrase pairs followed by Eisele et al. (2008), that simply consists of concatenating the phrase table obtained from the training parallel corpus and that obtained from the RBMT system, penalises the translation quality achieved by their approach.

3.4 Evaluating the new hybridisation strategy with automatically inferred rules

As it has been empirically proved in the previous section, shallow-transfer rules can improve the performance of phrase-based SMT. However, a considerable human effort and a high level of linguistic knowledge are needed to create them. In order to reduce the degree of human effort required to achieve such improvement, the algorithm presented in Chapter 2 can be used to infer a set of shallow-transfer rules from the training parallel corpus from which the SMT models are built, and this set of rules, together with the RBMT dictionaries, can be used to enlarge the phrase table as described previously in this chapter. This way, a significant boost in translation quality could be achieved with the sole addition of RBMT dictionaries. In this section, a set of experiments aimed at assessing the viability of this approach is presented.

3.4.1 Experimental setup

There are a couple of considerations to take into account when inferring a set of shallow-transfer rules to be integrated into the SMT system. Firstly, the experiments described in the previous chapter concluded that the generalisation of generalised alignment templates (GATs) to combinations of values of morphological inflection attributes not

observed in the training corpus, which is one cause of the vast complexity of the minimisation problems, brings a significant translation quality boost only when the training corpus is very small (below 1 000 parallel sentences; see Section 2.6). Given the fact that the parallel corpus sizes for which an SMT system starts to be competitive are much bigger, the generalisation of morphological inflection attributes can be skipped when inferring shallow-transfer rules to be integrated in SMT. Moreover, preliminary experiments showed that, even disabling the generalisation to non-observed combinations of values of morphological inflection attributes, the global minimisation algorithm still needs a huge amount of processing time in order to infer a set of rules from a parallel corpus that contains hundreds of thousands of sentences. For that reason, the amount of data from which the shallow-transfer rules are inferred has been limited to 160 000 sentences in the experiments carried out in this section.

Secondly, the rule inference algorithm chooses the rules to be generated so as to ensure that, when they are applied by a shallow-transfer RBMT system in a greedy, left-to-right, longest-match way, the groups of words which need to be processed together are translated with the same rule (Section 2.4.5). Since, in principle, the SMT decoder splits the input sentences in all possible ways, it may not be necessary to optimize the rules for chunking. In this manner, shallow-transfer rules for all the sequences of SL lexical categories present in the corpus would be generated.

With these considerations in mind, the experiments have been carried out as follows. For the same language pairs, corpora and RBMT dictionaries used in the previous section, two new systems have been built:

- *extended-phrase-learned*: in this system, the rule inference algorithm described in Chapter 2 has been applied on the training corpus. The selection of sequences of SL lexical categories for which rules are generated and the removal of redundant rules (described in Section 2.4.5) have not been performed. The rules inferred, together with the dictionaries, have been used to enrich the SMT system following the hybridisation strategy described in Section 3.2. Because of the time complexity of the minimisation problem to be solved in the rule inference process approach (see Section 2.4.4), only the first 160 000 sentences of the training corpus have been used for rule inference in those cases in which the corpus was larger than 160 000 sentences. In other words, the systems built from 160 000, 600 000, and the whole set of parallel sentences use exactly the same set of shallow-transfer rules.[25]

- *extended-phrase-learned-chunking*: it is a variant of the previous system in which the whole rule inference algorithm has been applied, including the optimisation

[25]In addition, the part of the training corpus used for rule inference has been split into two parts: the first 4/5 of the corpus has been used for actual rule inference, while the last 1/5 has been employed as a development corpus in order to optimise the threshold δ, as in the experiments described in the previous chapter. For training corpora bigger than 10 000 sentences, only 2 000 sentences have been used for optimising δ, while the remaining part of the corpus has been used for rule inference.

of rules for chunking in RBMT (described in Section 2.4.5). It has only been built for the smallest training corpus size (10 000 sentences), since the purpose of this system is to assess if there is a drop in translation performance when disabling the optimisation of rules for chunking. In addition, since the computing time of the optimisation of rules for chunking also grows fast with the size of the training corpus, it would be prohibitive to run the optimisation of rules for chunking with bigger corpora.

All these systems have been compared with a pure SMT baseline built from the same data, a hybrid system built from the Apertium hand-crafted rules and dictionaries, a hybrid system built with the same strategy but only from the Apertium dictionaries, and the RBMT system Apertium with rules inferred from the training corpus. In all the hybrid systems, the scoring method described in Section 3.2.2.3 has been used, since this is the scoring method that proved to perform better in the experiments described in the previous section.

3.4.2 Results and discussion

Table 3.2 depicts the BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and METEOR (Banerjee and Lavie, 2005) automatic evaluation scores obtained by the two hybrid systems with rules automatically inferred for the different language pairs and domains and only for the training corpora that contain 10 000 sentences. The number of GATs inferred with each method is also displayed. The statistical significance of the difference between the values of the automatic evaluation scores obtained by both systems has been computed with paired bootstrap resampling (Koehn, 2004b) ($p \leq 0.05$; 1 000 iterations). Scores in bold for a system mean that it outperforms the other system by a statistically significant margin.

The results show that there are not consistent differences between the systems whose rules have been optimised for chunking and the systems whose rules have not: statistically significant differences can only be found only for some of the evaluation metrics. In the Spanish–English language pair, optimising rules for chunking brings a tiny improvement, while in English–Spanish, the effect is the opposite. Since the impact of the rules is higher in out-of-domain evaluation, the effect of the optimisation is also more noticeable in this scenario.

The optimisation of rules for chunking affects the resulting hybrid system in two ways. On the one hand, it prevents the inclusion in the phrase table of multiple noisy phrase pairs that were generated from shallow-transfer rules that match sequences of lexical categories that do not need to be processed together for translating between the languages involved.[26] Due to the fact that the decoder cannot evaluate all the

[26]For instance, the rule inference without optimisation for chunking from the English–Spanish parallel corpus that contains 10 000 sentences produced an English–Spanish shallow-transfer rule that

Table 3.2: Results of the automatic evaluation carried out for the hybrid systems in which the shallow-transfer rules have been inferred from the training corpus. Automatic evaluation scores for systems in which the rules have been optimised for chunking (value *yes* in the row labelled as *chunking*) and for systems in which they have not been optimised (value *no* in the same row) are shown. If there is a statistically significant difference between both options (according to paired bootstrap resampling; $p \leq 0.05$; 1 000 iterations), the score of the winning option is shown in bold. The experiments have been carried out with a subset of the training corpus that contains 10 000 sentences.

Metric	# GATs		BLEU		TER		METEOR	
chunking	yes	no	yes	no	yes	no	yes	no
en-es (in-domain)	2 124	2 781	0.2708	0.2710	0.6458	0.6464	0.5034	0.5033
en-es (out-of-domain)			0.2151	0.2161	0.6809	0.6785	0.4655	**0.4677**
es-en (in-domain)	1 993	5 481	0.2687	0.2672	0.6351	0.6323	0.3245	0.3245
es-en (out-of-domain)			**0.2151**	0.2111	**0.6763**	0.6877	0.3079	0.3070
br-fr (in-domain)	1 081	1 875	0.1972	0.1986	0.8807	0.8796	0.3422	0.3423

translation hypotheses, these useless phrase pairs may prevent other, more important phrase pairs from being included in the final translation. It may also happen that the language model does not have enough information to properly score the synthetic phrase pairs built from these noisy rules. From this point of view, the optimisation of rules for chunking should have a positive impact in translation quality. On the other hand, since an SMT system does not perform a greedy segmentation of the input sentence, some of the rules discarded during optimisation for chunking in RBMT may still be useful if they are included in an SMT system. Rules that would prevent the application of a more important rule by the RBMT engine do not prevent the application of that rule in the hybrid system because, in principle, all the possible segmentations are taken into account.[27] In the light of the results, it seems that the former is more relevant for Spanish–English, while the latter has a higher positive impact for English–Spanish. Since Spanish is morphologically more complex, more rules are needed to correctly perform agreements, and probably more rules discarded during optimisation for chunking were useful. Nevertheless, these differences remain to be studied more deeply.

matches a singular noun followed by a preposition and the determiner *a*. As result, the rule produces the translation of the SL noun according to the bilingual dictionary followed by the translation of the SL preposition according to the bilingual dictionary and the Spanish masculine singular determiner *uno*. The phrase pairs generated by this rule are often incorrect because the gender and number of the determiner depend on a noun that is not matched by the rule.

[27]Rules that prevent the application of more important rules in some sentences may not prevent it in other sentences. When there are more sentences in which the impact of the rule is negative than sentences in which the impact is positive, the rule is generally discarded during the optimisation for chunking (see Section 2.4.5). However, higher translation quality would be achieved if the rule were applied only in the sentences in which it has a positive impact. The more flexible way of applying the rules in the hybrid system may make this possible.

The comparison between the results obtained by the hybrid approach that contains automatically inferred rules created without optimisation for chunking (*extended-phrase-learned*) and those obtained by the other approaches being evaluated is presented in figures 3.12–3.16. They show the BLEU (Papineni et al., 2002), TER (Snover et al., 2006) (the figures represent again 1-TER) and METEOR (Banerjee and Lavie, 2005) automatic evaluation scores for the different systems evaluated. In addition, the statistical significance[28] of the difference between the hybridisation approach *extended-phrase-learned* and the other systems is also presented in a table, in the same way as depicted the previous section.

Firstly, when new hybrid approach with automatically inferred rules is compared with the SMT baseline and the pure RBMT systems, it behaves in the same way as when hand-crafted rules are used: it outperforms both types of baselines when the training corpus is small or an out-of-domain evaluation is performed.

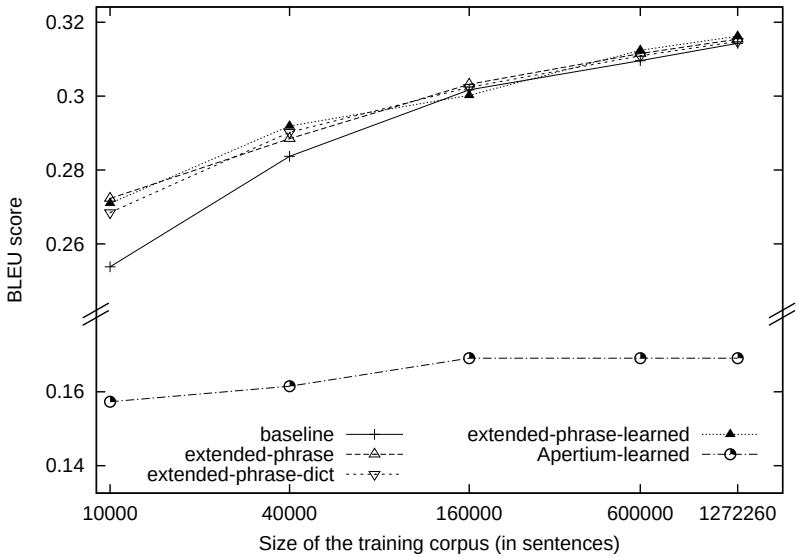
If the comparison is performed with the hybrid system that uses only dictionaries, it can be observed that the new hybrid approach also outperforms the dictionary-based approach almost in the same cases as the hybrid approach with hand-crafted rules (out-of-domain evaluation and in-domain evaluation only with the smallest parallel corpus size, although the three evaluation metrics do not agree in the latter case). In other words, with the automatic inference of shallow-transfer rules, a statistically significant improvement over the hybrid approach that uses only dictionaries has been achieved without using any additional linguistic resource.

Moreover, in some cases there is not a statistically significant difference between the hand-crafted and the automatically inferred rules when they are used in the hybrid system. This happens, for instance, in the English–Spanish out-of-domain evaluation when the training corpus contains 600 000 pairs of sentences. A translation quality similar to that obtained with hand-crafted rules has therefore been reached without the intervention of the human experts who usually create them. In the remainder of situations where the hybrid system with hand-crafted rules outperforms the hybrid system with dictionaries, the translation quality achieved by the hybrid system with inferred rules (*extended-phrase-learned*) lies between them.

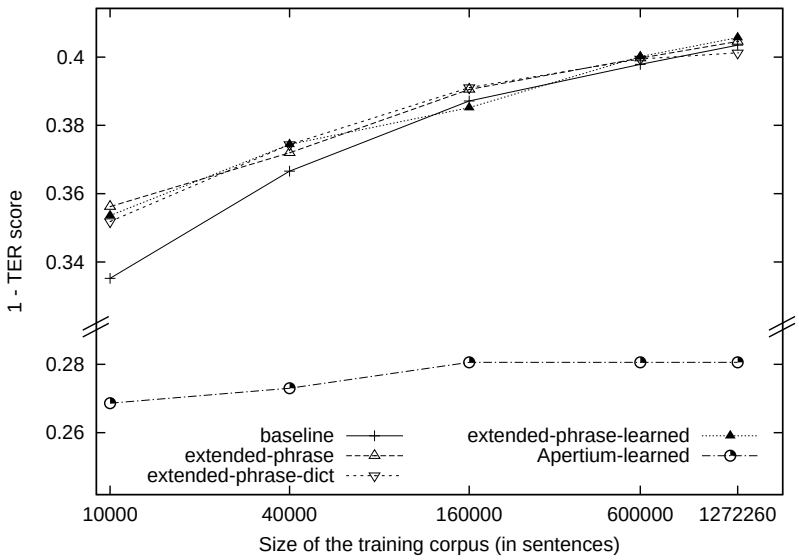
In addition, it is worth noting that the translation quality of the approach *extended-phrase-learned* does not drop (compared to the other hybrid systems) when the size of the training corpus exceeds 160 000 parallel sentences and the full training corpus is not used for rule inference. In fact, under these circumstances (600 000 parallel sentences) it can be observed that there are not significant differences between the use of automatically learned rules and hand-crafted rules in hybrid systems (for English–Spanish, out-of-domain evaluation). That observation is probably related to the fact that the translation performance of the automatically inferred rules grows very slowly with the size of the training corpus, and the rules obtained from bigger parallel corpora would probably be similar to those obtained from the fragment of 160 000 sentences.

[28] Again obtained through paired bootstrap resampling (Koehn, 2004b) ($p \leq 0.05$; 1 000 iterations).

Figure 3.12: Automatic evaluation scores obtained by the baseline phrase-based SMT system, Apertium with learned rules (*Apertium-learned*), and the new hybrid approach described in Section 3.2.2 using hand-crafted shallow-transfer rules (*extended-phrase*), a set of rules inferred from the training corpus (*extended-phrase-learned*), and no rules at all (*extended-phrase-dict*) for the English–Spanish language pair (in-domain evaluation). The table represents the results of the paired bootstrap resampling comparison (Koehn, 2004b) with the new hybrid approach using automatically inferred rules.

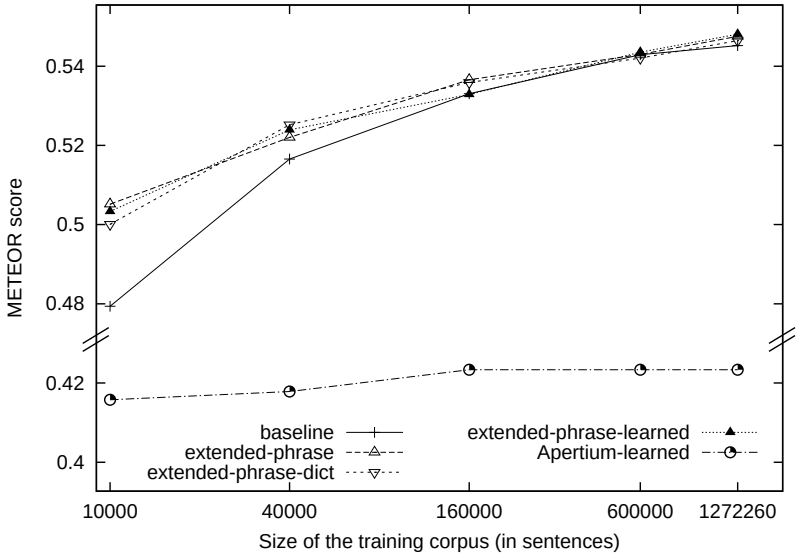


(a) BLEU scores.



(b) 1−TER scores.

(continued in next page)

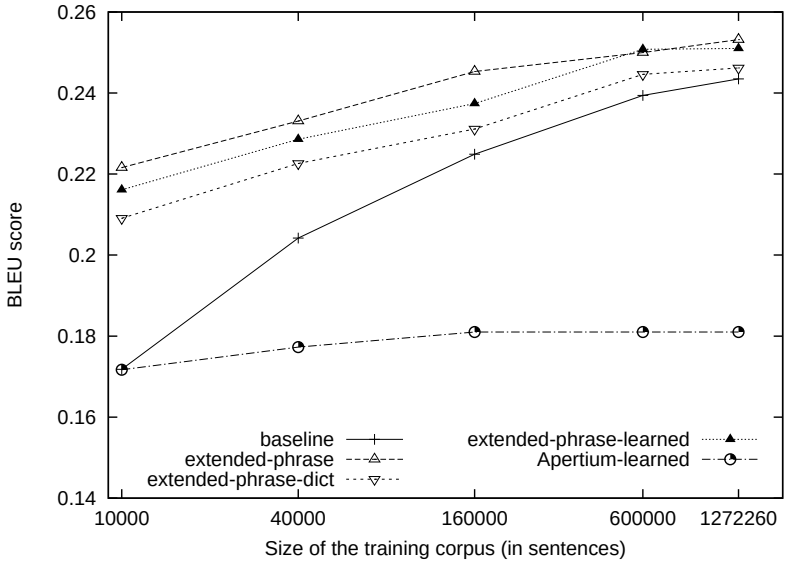


(c) METEOR scores.

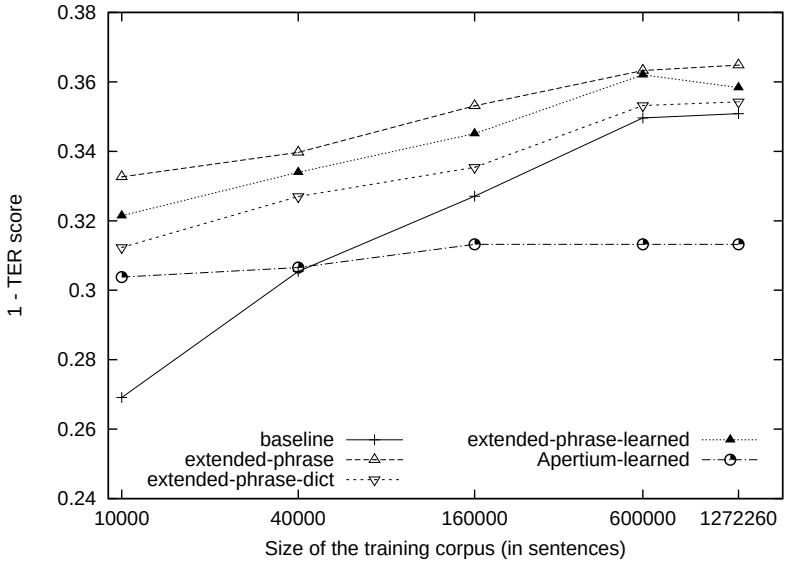
system	10 000	40 000	160 000	600 000	1 272 260
metric	B T M	B T M	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	= = =	= = =	= = ⇑
Apertium-learned	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	= = ⇑	= = =	= ⇓ ⇓	= = =	= ⇑ =
extended-phrase	= = =	⇑ = =	= ⇓ ⇓	= = =	= = =

(d) Results of the paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between the *extended-phrase-learned* system and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU(B), TER(T) and METEOR(M). An arrow pointing upwards means that *extended-phrase-learned* outperforms the reference method by a statistically significant margin, an arrow pointing downwards means the opposite and an equal sign means that there are not statistically significant differences between the systems.

Figure 3.13: Automatic evaluation scores obtained by the baseline phrase-based SMT system, Apertium with learned rules (*Apertium-learned*), and the new hybrid approach described in Section 3.2.2 using hand-crafted shallow-transfer rules (*extended-phrase*), a set of rules inferred from the training corpus (*extended-phrase-learned*), and no rules at all (*extended-phrase-dict*) for the English–Spanish language pair (out-of-domain evaluation). The table represents the results of the paired bootstrap resampling comparison (Koehn, 2004b) with the new hybrid approach using automatically inferred rules.

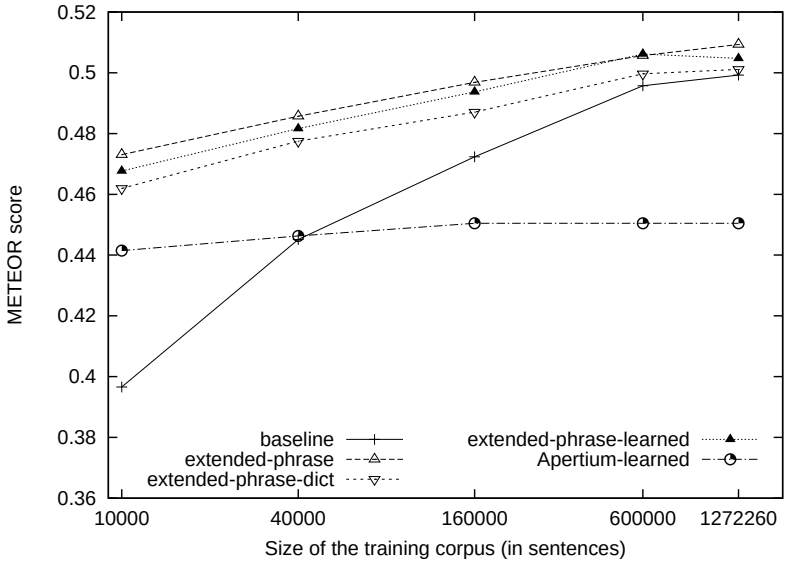


(a) BLEU scores.



(b) 1−TER scores.

(continued in next page)

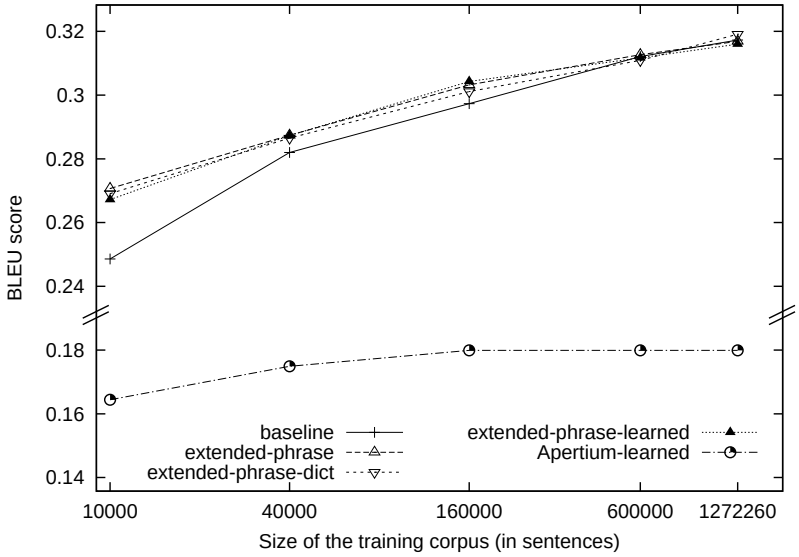


(c) METEOR scores.

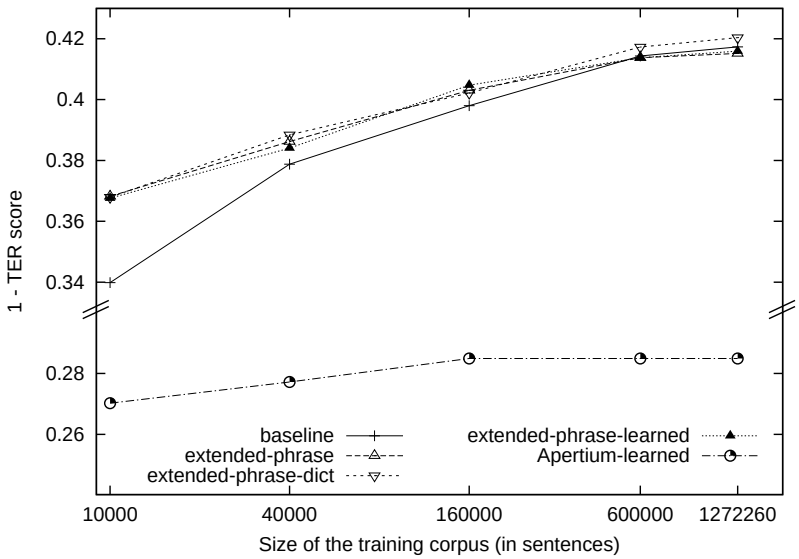
system	10 000	40 000	160 000	600 000	1 272 260
metric	B T M	B T M	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Apertium-learned	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase	⇓ ⇓ ⇓	⇓ ⇓ ⇓	⇓ ⇓ ⇓	= = =	= ⇓ ⇓

(d) Results of the paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between the *extended-phrase-learned* system and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU(B), TER(T) and METEOR(M). An arrow pointing upwards means that *extended-phrase-learned* outperforms the reference method by a statistically significant margin, an arrow pointing downwards means the opposite and an equal sign means that there are not statistically significant differences between the systems.

Figure 3.14: Automatic evaluation scores obtained by the baseline phrase-based SMT system, Apertium with learned rules (*Apertium-learned*), and the new hybrid approach described in Section 3.2.2 using hand-crafted shallow-transfer rules (*extended-phrase*), a set of rules inferred from the training corpus (*extended-phrase-learned*), and no rules at all (*extended-phrase-dict*) for the Spanish–English language pair (in-domain evaluation). The table represents the results of the paired bootstrap resampling comparison (Koehn, 2004b) with the new hybrid approach using automatically inferred rules.

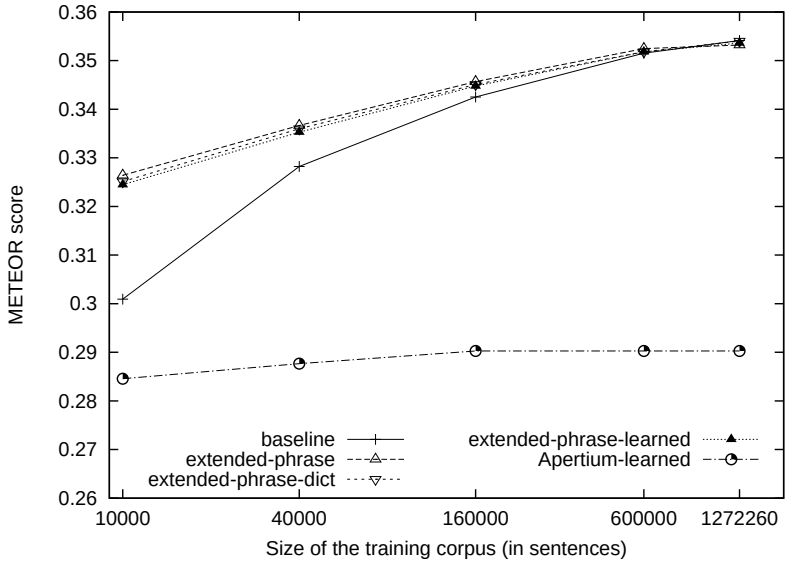


(a) BLEU scores.



(b) 1−TER scores.

(continued in next page)

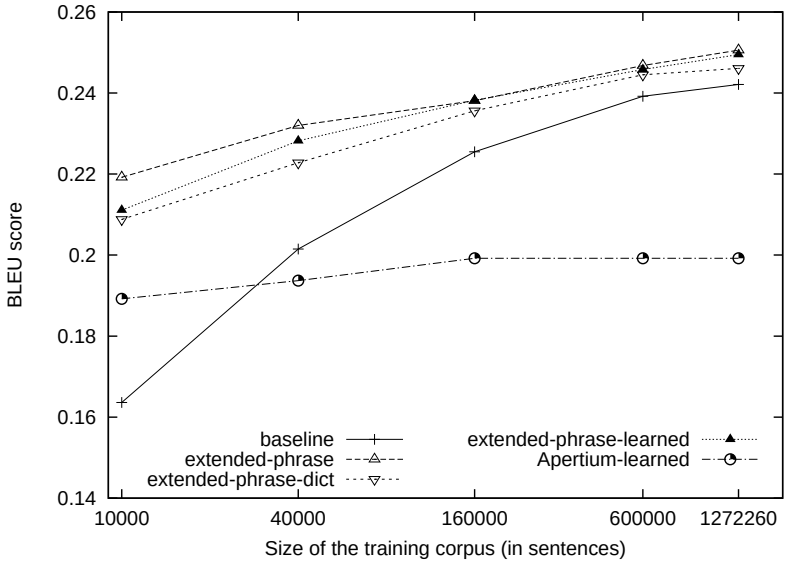


(c) METEOR scores.

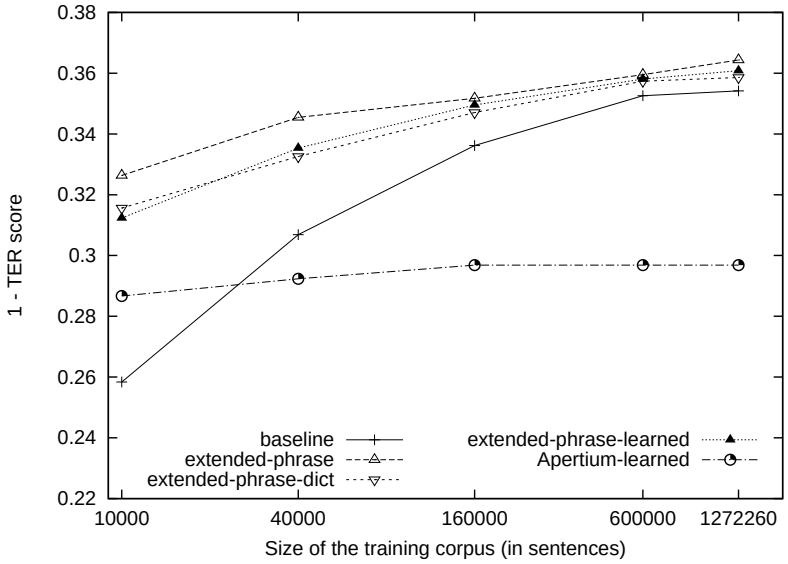
system	10 000			40 000			160 000			600 000			1 272 260		
metric	B	T	M	B	T	M	B	T	M	B	T	M	B	T	M
baseline	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑	=	=	=	=	=	=
Apertium-learned	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑
extended-phrase-dict	=	=	=	=	⇓	=	⇑	=	=	=	⇓	=	⇓	⇓	=
extended-phrase	⇓	=	⇓	=	=	⇓	=	=	=	=	=	=	=	=	=

(d) Results of the paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between the *extended-phrase-learned* system and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU(B), TER(T) and METEOR(M). An arrow pointing upwards means that *extended-phrase-learned* outperforms the reference method by a statistically significant margin, an arrow pointing downwards means the opposite and an equal sign means that there are not statistically significant differences between the systems.

Figure 3.15: Automatic evaluation scores obtained by the baseline phrase-based SMT system, Apertium with learned rules (*Apertium-learned*), and the new hybrid approach described in Section 3.2.2 using hand-crafted shallow-transfer rules (*extended-phrase*), a set of rules inferred from the training corpus (*extended-phrase-learned*), and no rules at all (*extended-phrase-dict*) for the Spanish–English language pair (out-of-domain evaluation). The table represents the results of the paired bootstrap resampling comparison (Koehn, 2004b) with the new hybrid approach using automatically inferred rules.

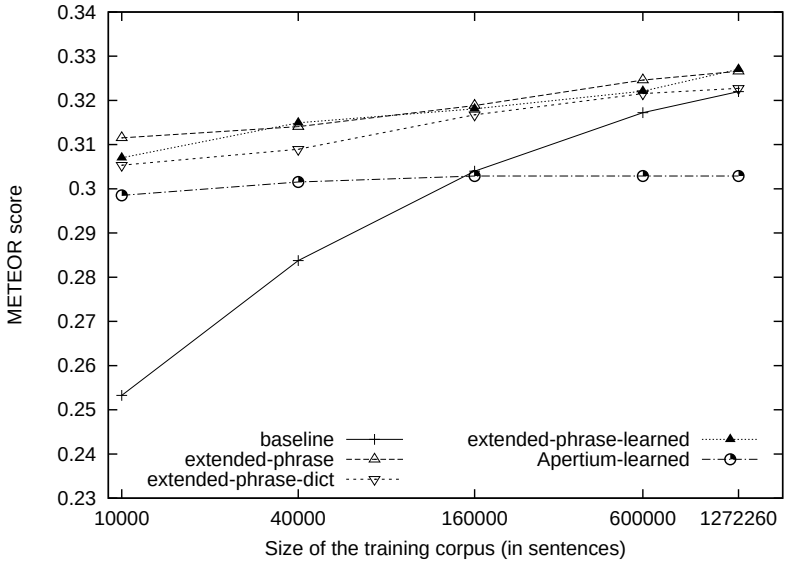


(a) BLEU scores.



(b) 1−TER scores.

(continued in next page)

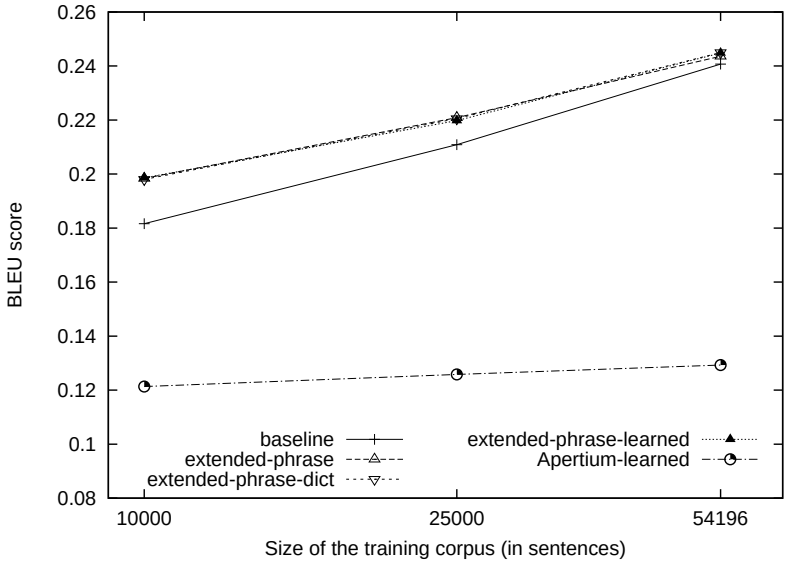


(c) METEOR scores.

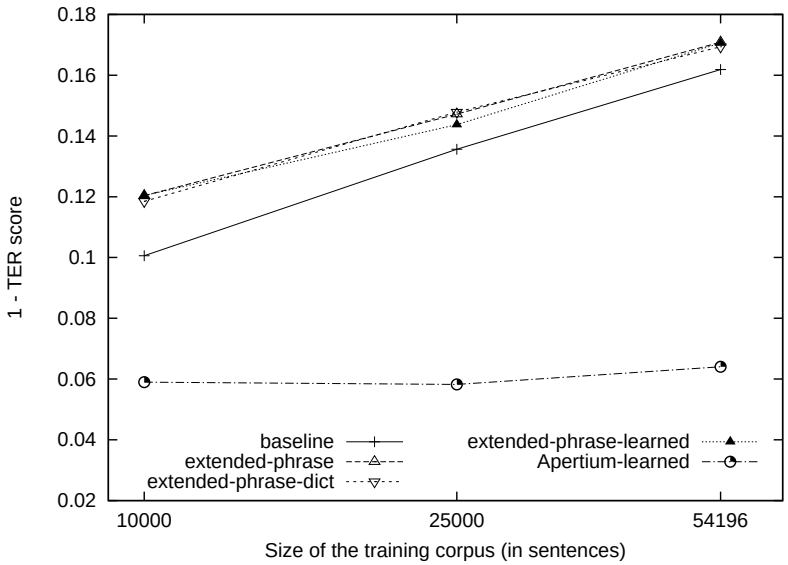
system	10 000	40 000	160 000	600 000	1 272 260
metric	B T M	B T M	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Apertium-learned	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	= ⇓ ⇑	⇑ = ⇑	= = ⇑	= = =	⇑ = ⇑
extended-phrase	⇓ ⇓ ⇓	⇓ ⇓ =	= = =	= = ⇓	= ⇓ =

(d) Results of the paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between the *extended-phrase-learned* system and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU(B), TER(T) and METEOR(M). An arrow pointing upwards means that *extended-phrase-learned* outperforms the reference method by a statistically significant margin, an arrow pointing downwards means the opposite and an equal sign means that there are not statistically significant differences between the systems.

Figure 3.16: Automatic evaluation scores obtained by the baseline phrase-based SMT system, Apertium with learned rules (*Apertium-learned*), and the new hybrid approach described in Section 3.2.2 using hand-crafted shallow-transfer rules (*extended-phrase*), a set of rules inferred from the training corpus (*extended-phrase-learned*), and no rules at all (*extended-phrase-dict*) for the Breton–French language pair (in-domain evaluation). The table represents the results of the paired bootstrap resampling comparison (Koehn, 2004b) with the new hybrid approach using automatically inferred rules.

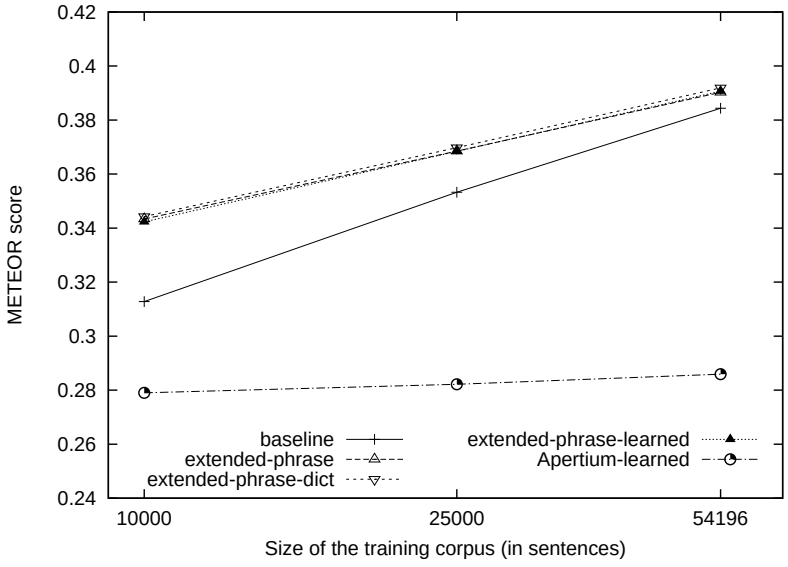


(a) BLEU scores.



(b) 1−TER scores.

(continued in next page)



(c) METEOR scores.

system	10 000			25 000			54 196		
metric	B	T	M	B	T	M	B	T	M
baseline	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑
Apertium-learned	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑	⇑
extended-phrase-dict	=	=	⇓	=	⇓	=	=	=	=
extended-phrase	=	=	=	=	⇓	=	=	=	=

(d) Results of the paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between the *extended-phrase-learned* system and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU(B), TER(T) and METEOR(M). An arrow pointing upwards means that *extended-phrase-learned* outperforms the reference method by a statistically significant margin, an arrow pointing downwards means the opposite and an equal sign means that there are not statistically significant differences between the systems.

Nevertheless, the exact impact of the proportion of the training corpus used for rule inference for different training corpus sizes, language pairs and domains, as well as the effect of the number of generalised alignment templates inferred, also deserves further research.

It is also worth remarking the difference between the hand-crafted and the automatically inferred rules when they are used in an RBMT system: for the three evaluation metrics considered, it is considerably bigger than the difference between the hybrid system enriched with hand-crafted rules and that enriched with automatically inferred rules (compare, for instance, the results depicted in figures 3.2 and 3.13). In the two RBMT systems, the translation is completely led by the shallow-transfer rules, and the possible errors encoded in the automatically inferred rules have a direct impact on the output. Moreover, the inferred rules have not been optimised for chunking, and sequences of words that need to be treated together may not be processed by the same rules. On the contrary, when the inferred rules are integrated in the hybrid system, the language model can assign low scores to hypotheses built from incorrect rules.

3.5 Measuring the impact of the language model size

The experiments carried out in the two previous sections have proved that shallow-transfer rules can be successfully used to build hybrid SMT/RBMT systems and that they can be automatically inferred from the training corpus and still bring a clear advantage over hybrid systems that only use dictionaries. In those experiments, the TL model was estimated from a relatively small monolingual corpus (it was in the same order of magnitude as the biggest parallel corpus used) in order to better assess the impact of the shallow-transfer rules in the translation model.

In this section, a set of experiments is performed with a language model estimated from a monolingual corpus much larger than the TL side of the training parallel corpus. This scenario is more realistic since, when building a phrase-based SMT system from all the available resources, it is usually easier to find TL monolingual corpora than parallel corpora. The objective is to confirm whether the conclusions drawn with relatively small language models also apply to bigger ones.

3.5.1 Experimental setup

The experiments carried out with a big language model only comprise the out-of-domain evaluation for the English–Spanish and Spanish–English language pairs, since these are the scenarios where the gain in translation quality provided by the shallow-transfer rules is generally statistically significant.

Table 3.3: Number of sentences, words, and size of the vocabulary of the monolingual corpora used to train the language models used in the experiments described in Section 3.5. Compare the size of this corpus with that of the corpora shown in Table 3.1.

Corpus	# sentences	TL	
		# words	# voc
Language model (English)	6 228 203	153 394 585	648 379
Language model (Spanish)	6 228 203	182 229 011	643 856

(a) English↔Spanish

The language models (both for English and Spanish) have been estimated from the concatenation of the monolingual text used in the previous experiments (Europarl; described in Section 3.3.1) and the *News Crawl* monolingual corpus provided for the WMT 2011 shared translation task.[29] Note also that this additional monolingual corpus shares domain with the out-of-domain test and development corpora. Given the fact that the English *News Crawl* monolingual corpus is much bigger than the Spanish one, only a subset of it (4 578 051 sentences) has been used, so that both monolingual corpora have the same size. Thus, the possible differences observed in the results between the different language pairs will be caused by the features of the languages involved and not by the amount of data employed. The number of words and vocabulary size of the monolingual corpora are presented in Table 3.3.

The remainder of the experimental setup is the same as in the previous section. The following systems have been evaluated with the big language models: the baseline SMT system, the hybrid approach described in Section 3.2 with the hand-crafted rules from the Apertium project (*extended-phrase*), the same hybrid approach, but using with automatically inferred rules (*extended-phrase-learned*; the same restrictions over the corpus used for rule inference as in the previous section have been applied) and the hybrid approach in Section 3.2 using only dictionaries as RBMT linguistic resource (*extended-phrase-dict*).

3.5.2 Results and discussion

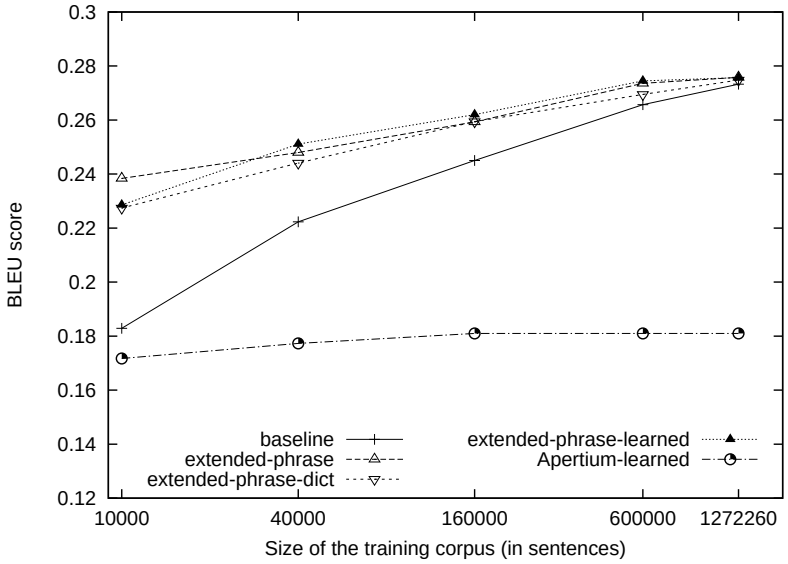
Figures 3.17 and 3.18 show the BLEU, TER (the figures represent 1-TER) and METEOR automatic evaluation scores for the different systems evaluated. The statistical significance[30] of the difference between the *extended-phrase-learned* system and the other systems is also presented in a table, in the same way as it has been depicted the previous sections.

Results show that a bigger language model reduces the difference between the baseline SMT system and the hybrid system and the positive impact of the shallow-transfer

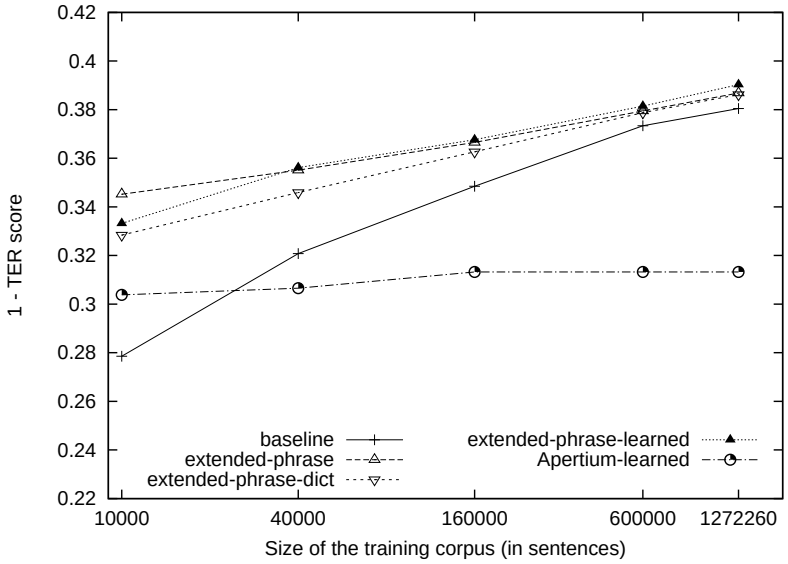
[29] `http://www.statmt.org/wmt11/translation-task.html`

[30] Paired bootstrap resampling (Koehn, 2004b) ($p \leq 0.05$; 1 000 iterations).

Figure 3.17: Automatic evaluation scores obtained by the baseline phrase-based SMT system, Apertium with learned rules (*Apertium-learned*), and the new hybrid approach described in Section 3.2.2 using hand-crafted shallow-transfer rules (*extended-phrase*), a set of rules inferred from the training corpus (*extended-phrase-learned*), and no rules at all (*extended-phrase-dict*) for the English–Spanish language pair (out-of-domain evaluation). The TL model has been estimated from a monolingual corpus that is much larger than the training parallel corpus.

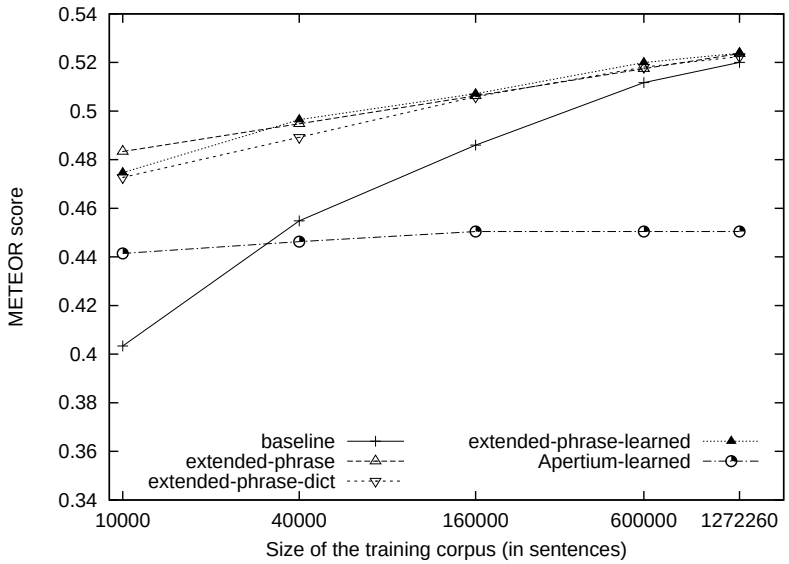


(a) BLEU scores.



(b) 1−TER scores.

(continued in next page)

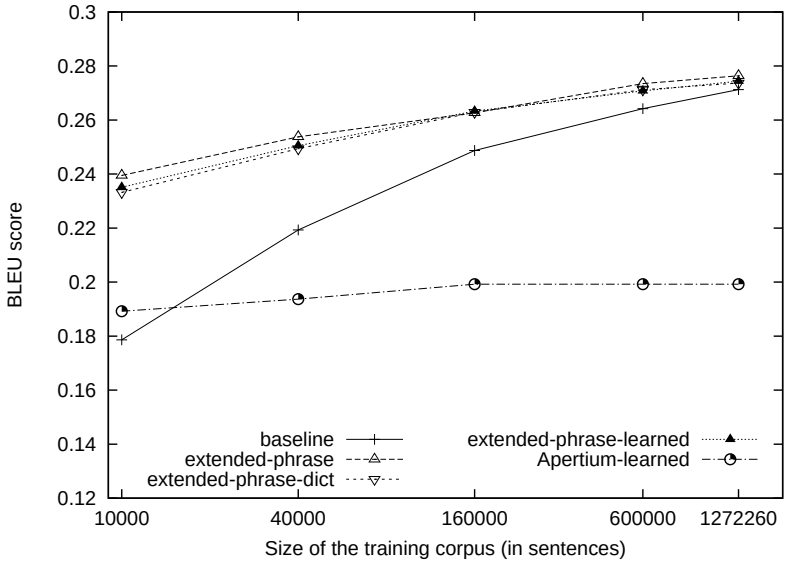


(c) METEOR scores.

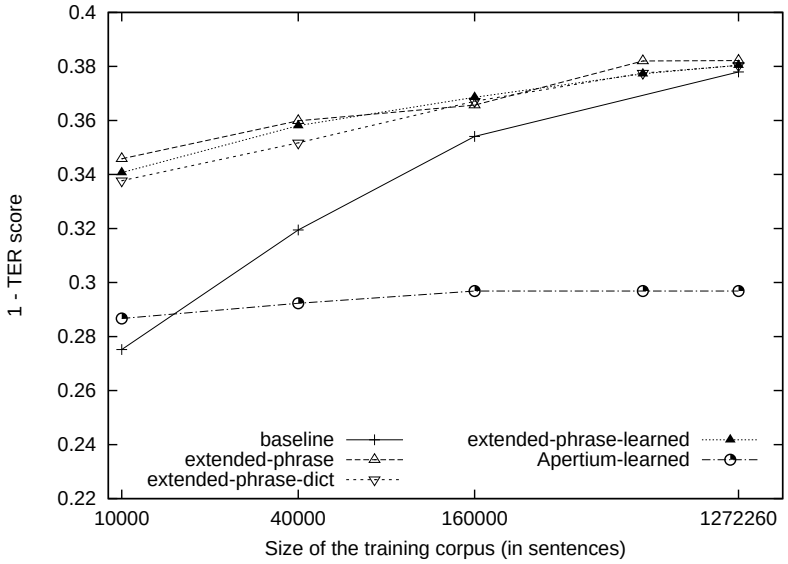
system	10 000	40 000	160 000	600 000	1 272 260
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	= ⇑ ⇑
Apertium-learned	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	= ⇑ =	⇑ ⇑ ⇑	= ⇑ =	⇑ = =	= ⇑ =
extended-phrase	⇓ ⇓ ⇓	= = =	= = =	= = ⇑	= ⇑ =

(d) Results of the paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between the *extended-phrase-learned* system and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU(B), TER(T) and METEOR(M). An arrow pointing upwards means that *extended-phrase-learned* outperforms the reference method by a statistically significant margin, an arrow pointing downwards means the opposite and an equal sign means that there are not statistically significant differences between the systems.

Figure 3.18: Automatic evaluation scores obtained by the baseline phrase-based SMT system, Apertium with learned rules (*Apertium-learned*), and the new hybrid approach described in Section 3.2.2 using hand-crafted shallow-transfer rules (*extended-phrase*), a set of rules inferred from the training corpus (*extended-phrase-learned*), and no rules at all (*extended-phrase-dict*) for the Spanish–English language pair (out-of-domain evaluation). The TL model has been estimated from a monolingual corpus that is much larger than the training parallel corpus.

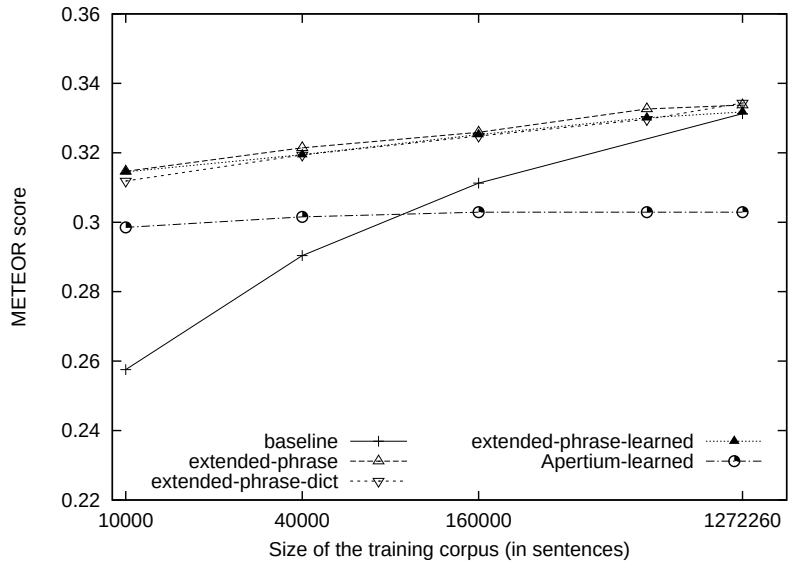


(a) BLEU scores.



(b) 1−TER scores.

(continued in next page)



(c) METEOR scores.

system	10 000	40 000	160 000	600 000	1 272 260
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ = =
Apertium-learned	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	= ⇑ ⇑	= ⇑ =	= = =	= = =	= = ⇓
extended-phrase	⇓ ⇓ =	⇓ = ⇓	= ⇑ =	⇓ ⇓ ⇓	= = ⇓

(d) Results of the paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between the *extended-phrase-learned* system and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU(B), TER(T) and METEOR(M). An arrow pointing upwards means that *extended-phrase-learned* outperforms the reference method by a statistically significant margin, an arrow pointing downwards means the opposite and an equal sign means that there are not statistically significant differences between the systems.

rules. Even when the phrase table does not contain the most appropriate translations of the phrases from the SL sentences, or contains them, but they are scored with a low probability, a powerful language model can help the decoder to generate better translations from shorter phrases or compensate the low translation probabilities of the most adequate phrases by assigning high scores to the hypotheses that contain them. Note also that, when synthetic phrase pairs are generated from the RBMT dictionaries, all the inflection variants of the words are generated, and that makes easier for the RBMT system to generate hypotheses with correct agreements between the TL inflection features. These hypotheses will be assigned a high score by the language model as long as the right sequences of TL surface forms are observed in the TL monolingual corpus.

As in the experiments presented previously in this chapter, the differences between the baseline SMT system and the hybrid systems and the positive impact of the shallow-transfer rules are reduced as the size of the training parallel corpus grows. For the biggest training corpus size, the scores of the different hybrid systems and the baseline SMT system are similar, and statistically significant differences between them are only found for some of the evaluation metrics. As in previous experiments, it can also be observed that when Spanish is the TL the difference between the hybrid system built only with dictionaries and that built also with shallow-transfer rules is bigger than when Spanish is the SL.

Another interesting observation is the fact that, for English–Spanish, the translation performance achieved by the hybrid system with inferred rules improves when compared with the hybrid system with hand-crafted rules as the size of language model is increased. For the smaller language model (Figure 3.13), the former system (*extended-phrase-learned*) is not able to outperform the latter (*extended-phrase*) by a statistically significant margin for any training corpus. The latter system wins in most cases, although there is a tie when the training corpus contains 600 000 sentences. For the bigger language model, on the contrary, there is a tie in most cases, although for some training corpus size/evaluation metric combinations the hybrid system with learned rules wins. These results suggest that a more powerful language model is able to find a better segmentation of the SL sentences and use the synthetic phrase pairs, that have been generated from rules that match all the sequences of SL lexical categories observed in the training corpus, in a more effective way. For Spanish–English, on the contrary, this improvement cannot be observed. This observation is compatible with the results shown at the beginning of Section 3.4.2: there is a small drop in translation quality for Spanish–English and a small improvement for English–Spanish when the optimisation of rules for chunking is disabled. In other words, for English–Spanish, allowing the decoder to choose among more synthetic phrase pairs helps improving translation quality, and a more powerful language model helps to choose the most appropriate ones.

3.6 Conclusions

In this chapter, a new hybridisation approach aimed at enriching phrase-based SMT models with the linguistic data from a shallow-transfer RBMT system has been presented. It has been confirmed that data from shallow-transfer RBMT can improve phrase-based SMT systems and that the resulting hybrid system outperforms both pure SMT and RBMT systems built from the same data. This new hybridisation approach overcomes the limitations of the general-purpose strategy aimed at improving phrase-based SMT models with data from other non-SMT systems developed by Eisele et al. (2008) thanks to the fact that it takes advantage of the way in which the shallow-transfer RBMT system uses its linguistic resources to segment SL sentences. The experiments carried have shown that the new hybrid approach outperforms the strategy by Eisele et al. (2008) by a statistically significant margin in a wide range of situations. In fact, a system built with the hybridisation approach described in this chapter (Sánchez-Cartagena et al., 2011c) was one of the winners[31] in the pairwise manual evaluation of the WMT 2011 shared translation task (Callison-Burch et al., 2011) for the Spanish–English language pair. The effectiveness of the new hybrid approach is thereby confirmed by both automatic and human evaluations.

Moreover, it has been proved that the rule inference algorithm presented in Chapter 2 can be successfully combined with the new hybrid approach. Thus, a hybrid system can be built using dictionaries as the only hand-crafted linguistic resource. At the same time, an improvement of translation quality is achieved as if hand-crafted shallow-transfer rules were used. The hybrid system with automatically inferred rules is able to reach the translation quality achieved by a hybrid system with hand-crafted rules and, even when it does not reach it, the automatically inferred rules often bring a clear improvement over a hybrid system that only uses dictionaries to enrich the SMT models. Additionally, the need for a human expert for writing the rules, who may not always be available, is avoided.

According to the results obtained, the hybrid approach presented in this chapter is especially recommended when the training parallel corpus (for the translation model) and monolingual corpus (for the language model) have a moderate size and when the domain of the training corpus is different from the domain of the texts to be translated by the hybrid system. In the experiments described previously in this chapter, when a parallel corpus that contains around 26 000 000 words and a monolingual corpus with around 182 000 000 words are used to train the systems, the difference between the hybrid system built with automatically inferred rules and the baseline SMT system becomes not statistically significant for some of the evaluation metrics. The use of moderate-sized training corpora may be necessary in order to limit the size of the phrase table and the TL model when the hybrid system is required to have a limited memory footprint because it must be executed in a mobile device. Moreover, the hybrid approach presented in this chapter can also be safely applied in other scenarios, since

[31] No other system was found statistically significantly better using the sign test at $p \leq 0.10$.

drops in translation quality in comparison with a phrase-based SMT baseline have not been detected.

Finally, the improvement observed in the hybrid system for English–Spanish when using a language model trained from a relatively big monolingual corpus suggests that, with a big enough language model, the decoder may be able to deal better with the big amount of translation hypotheses generated by the synthetic phrase pairs. However, the same behaviour has not been observed for Spanish–English. In fact, for that pair, the high number of translation hypotheses generated when disabling the optimisation of rules for chunking seems to slightly degrade translation quality. In the light of the results obtained with the large language models, Section 5.2 discusses some enhancements to the process of optimising of rules for chunking that would produce rules more suitable to be integrated in an SMT system. A further reduction in the time complexity of the rule inference algorithm when the inferred rules are integrated in an SMT system is another future research line. That reduction would permit inferring shallow-transfer rules from larger parallel corpora.

Chapter 4

Assisting non-expert users in extending morphological dictionaries

A novel approach that allows non-expert users to insert new entries in monolingual morphological dictionaries, such as those used in rule-based machine translation, is presented in this chapter. The scenario considered is that of non-expert users of a rule-based machine translation system who introduce into its monolingual dictionaries the words found in an input text that are unknown to the system. Given a source-language surface form to be inserted (i.e. a word as it is found in running texts), the proposed strategy iteratively asks the user (an average speaker of the source language) to validate whether certain variations in the inflection of the word are correct. Users are asked the polar question *Is word* X *a valid form of the word to be inserted?*. This approach uses the answers of the users and the existing inflection paradigms in the monolingual dictionary to automatically insert the corresponding entry, which involves determining the stem of the word and choosing an inflection paradigm. A monolingual corpus, a hidden Markov model and a binary decision tree are used to reduce the number of polar questions that need to be asked for inserting an entry. The experiments carried out show that non-expert users are able to successfully answer the polar questions in most cases, and that the ID3 algorithm increases the efficiency of the approach —as compared to an heuristic approach that iteratively asks the user to validate an inflected word form generated from the most likely paradigm according to the monolingual corpus— and the robustness against possible erroneous information extracted from the corpus. If the user is bilingual and provides the translation of the inserted source-language word, the process is repeated to insert the corresponding entry in the target-language monolingual dictionary. Afterwards, the corresponding entry in the bilingual dictionary can be inserted automatically.

4.1 Introduction

When the resources needed for building an MT system are scarce, in addition to the automatic inference of linguistic resources with approaches such as those discussed in Chapter 2, the path of knowledge elicitation from people without strong linguistic and computer skills can be followed as a cheap alternative to the recruitment of experts. As already pointed out in the introductory chapter, average speakers of a language can insert new entries in monolingual dictionaries if they are asked questions in an appropriate way.

In this chapter, a novel strategy for allowing non-expert users to insert entries in monolingual morphological dictionaries such as those used in RBMT is presented. The scenario considered, which has been described in more detail in Chapter 1, is the following: when a user asks the MT system in which the knowledge elicitation approach is integrated to translate a sentence that contains a word that is not present in the SL monolingual dictionary, the user is asked to insert the new word in the dictionary. If she also knows its translation into the TL, she is also asked to insert it in the TL monolingual dictionary.

The objective of the approach presented in this chapter is to obtain a system which can be used to add the unknown word in a sentence (for example, *wants*) to the dictionary, together with its lemma and its different inflected forms and their associated morphological inflection information (such as *wants* : *want* `VERB-tense:present.person:3` or *wanting* : *want* `VERB-tense:gerund`). The proposed method operates under the following assumptions: inflection paradigms are already contained in the monolingual dictionaries and average speakers of a language can correctly answer the polar question: *Is word X a valid form of the word to be inserted?*. Recall that inflection paradigms are commonly used in RBMT systems in order to group regularities in the inflection of a set of words: a paradigm is usually defined as a collection of suffixes and their corresponding morphological information.

The rest of the chapter is organised as follows. Section 4.2 formalises the proposed strategy to allow non-expert users to insert new entries in RBMT dictionaries. It has two critical components that strongly influence the number of polar questions that need to be asked to the user in order to properly insert the dictionary entry: *feasibility score* and *querying algorithm*. Each paradigm[1] is assigned a feasibility score that represents how likely is that it is the right paradigm for the word to be inserted, and the querying algorithm decides which inflected word forms are validated by the user according to the feasibility score of each paradigm, among other factors. In Section 4.3, a set of experiments that confirms that average speakers of a language can successfully answer the aforementioned polar questions is described. The feasibility score and querying algorithm used in this set of experiments are based on intuitive heuristics. Afterwards, two improvements to the strategy are presented and evaluated: exploiting correlation

[1] Actually, each stem/paradigm pair has a different feasibility score, as explained in Section 4.2.1.

between the SL and the TL entries in existing dictionaries in order to reduce the number of polar question asked (Section 4.4); and a probabilistic reformulation based on binary decision trees and hidden Markov models (Rabiner, 1989, HMMs) of the feasibility score and querying algorithm used in the first set of experiments (Section 4.5). Finally, some conclusions are drawn.

4.2 Knowledge elicitation approach

This section formalises the process carried out in order to select the most appropriate polar questions to be asked to the user and build the corresponding monolingual dictionary entry from the answers provided. As explained previously, the knowledge elicitation process described in this section has been devised to aid non-expert users in inserting in the SL monolingual dictionary the unknown words present in the SL sentences they want to translate. If the user who interacts with the system is bilingual and provides a translation of the SL word, the knowledge elicitation process can be executed again in order to insert the corresponding entry in the TL monolingual dictionary. Afterwards, the corresponding entry in the bilingual dictionary can be inserted automatically.

4.2.1 Paradigm selection

As it has been pointed out in Chapter 1, the set of morphological inflection paradigms is an essential part of RBMT monolingual dictionaries. Paradigms are commonly introduced in RBMT systems in order to group regularities in the inflection of a set of words; a paradigm is usually defined as a collection of suffixes and their corresponding morphological information; e.g., the paradigm assigned to many common English verbs indicates that by adding the suffix *-ing* to the stem,[2] the gerund is obtained; by adding the suffix *-ed*, the past is obtained; etc. Adding a new entry to a monolingual dictionary therefore implies determining the stem of the new word and a suitable inflection paradigm among those defined by the MT system for the corresponding language. In this work, it is assumed that the paradigms for all the possible words in the language are already included in the dictionary. This is a common situation in the development of RBMT systems. A paradigm is created when the first set of words associated with it is added to the dictionary. Thus, it is very likely that the most important paradigms (those that generate regular inflections) are already defined in the first steps of the development of a dictionary.

[2]The stem is the part of a word that is common to all its inflected forms.

Let $P = \{p_i\}$ be the set of paradigms in a monolingual dictionary. Each paradigm p_i consists of a set of pairs (f_{ij}, m_{ij}), where f_{ij} is a suffix[3] which is appended to the stems to build new word forms[4], and m_{ij} is the corresponding morphological information. Given a *stem/paradigm* combination $c = t/p_i$ composed of a stem t and a paradigm p_i, the *expansion* $I(c)$ is the set of possible word forms resulting from appending each of the suffixes in p_i to t. For instance, an English dictionary may contain the stem *want-* assigned to a paradigm with suffixes $p_i = \{$*-ϵ*, *-s*, *-ed*, *-ing*$\}$ (hereinafter the morphological inflection information contained in p_i is omitted and only suffixes are shown); the expansion $I(\text{want}/p_i)$ consists of the set of word forms *want*, *wants*, *wanted* and *wanting*.

Given a new word form w to be added to a monolingual dictionary, the objective is to find both the stem $t \in \Pr(w)$[5] and the paradigm p_i so that $I(\text{t}/p_i)$ is the set of word forms which contains all (and only) the correct forms of the unknown word w. To that end, first the set $L = \{t_i/p_i : \exists_{f \in p_i} t_i f = w\}$ that contains all the stem/paradigm combinations compatible with w is built. This can be efficiently determined by using a *generalised suffix tree* (McCreight, 1976) containing all the possible suffixes included in the paradigms in P.

We illustrate this with an example. Consider a simple dictionary with only four paradigms: $p_1 = \{$*-ϵ*, *-s*$\}$; $p_2 = \{$*-y*, *-ies*$\}$; $p_3 = \{$*-y*, *-ies*, *-ied*, *-ying*$\}$; and $p_4 = \{$*-a*, *-um*$\}$. Let us assume that the new word form to be inserted into the dictionary is *w=policies* (plural form of the noun *policy*); the compatible stem/paradigm combinations which will be contained in the set L after this stage are: c_1=*policies*/p_1; c_2=*policie*/p_1; c_3=*polic*/p_2; and c_4=*polic*/p_3. The suffix tree built to efficiently obtain them is depicted in Figure 4.1.

4.2.2 Asking polar questions to the user

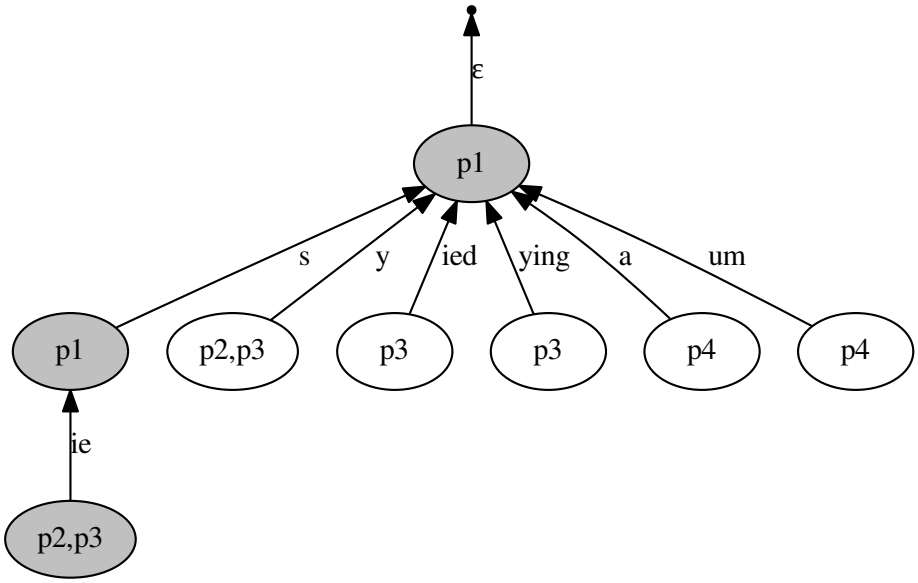
Once the list L of compatible stem/paradigm pairs has been obtained, the user must decide which element in L is the correct one for the word form w to be inserted into the dictionary, that is, which of the candidates $\{c_n \in L\}$ contains exactly all the valid forms of w in its expansion $I(c_n)$. To that end, the user is simply iteratively asked to confirm whether a word form in $I(c_n)$ is a valid form of w. After each question, the candidate stem/paradigm pairs from L not compatible with the answer are discarded, and the process continues until there is only a stem/paradigm pair in L, which will be the one assigned to the new entry in the dictionary. A monolingual corpus can be used

[3]Although this approach has been developed for languages that generate word forms by adding suffixes to the stems (for example, Romance languages), it could be adapted to inflectional languages based on different ways of adding morphemes.

[4]Recall that, in this dissertation, the term *surface form* is used to refer to words as they are found in running texts, such as *houses*, *policy* and *identifying*. A *word form* or *inflected word form* is a surface form generated by combining an inflection paradigm with a stem.

[5]$\Pr(w)$ is the set of all possible prefixes of w.

Figure 4.1: Suffix tree built to efficiently determine the candidate stem/paradigm pairs compatible with a surface form to be inserted into a monolingual dictionary. The word to be inserted is *policies*. Nodes that represent a compatible stem/paradigm pair are shadowed. The paradigms contain the following suffixes: p_1 = {-ϵ, -*s*}; p_2 = {-*y*, -*ies*}; p_3 = {-*y*, -*ies*, -*ied*, -*ying*}; and p_4 = {-*a*, -*um*}.



to obtain an a-priori estimation of the candidate pairs that are more likely to be the correct ones and guide the questions based on that information.

More formally, the process comprises the following steps:

1. *Paradigm scoring.* A feasibility score is computed for each compatible stem/paradigm pair $c_n \in L$ using a large monolingual corpus C. Following the previous example, the word forms for the different candidates would be: $I(c_1)$={*policies*, *policiess*}; $I(c_2)$={*policie*, *policies*}; $I(c_3)$={*policy*, *policies*}; and $I(c_4)$={*policy*, *policies*, *policied*, *policying*}. A large English monolingual corpus would probably contain evidence that suggests that c_3 is the most likely candidate, and therefore it would probably obtain the highest feasibility score. Different systems can be used to score paradigms: an heuristic approach that simply accounts for the number of word forms found in the monolingual corpus is first presented (Section 4.3.1), then a more sophisticated approach based on hidden Markov models is described (Section 4.5.1).

2. *Selection of word forms.* The best candidate is chosen from L by asking the user whether some word forms w' for some of the compatible stem/paradigm pairs $c_n \in L$ are correct forms of w or not. The questions asked to the user are polar questions: the only possible answers are yes/no. The following actions are triggered depending on the user's answer to the question about w':

- if it is accepted, all $c_n \in L$ for which $w' \notin I(c_n)$ are removed from L
- if it is rejected, all $c_n \in L$ for which $w' \in I(c_n)$ are removed from L

This process is repeated until $|L| = 1$. The remaining candidate in L will be the stem/paradigm pair inserted into the dictionary. The criterion followed to choose, in each step, the word form w' presented to the user (from now on, the querying algorithm) should ensure that as few questions as possible are asked to the user before arriving to the solution, thus increasing the *efficiency* of the approach.

The querying algorithm uses the feasibility score described above and the membership relation between the different word forms and the candidate stem/paradigm pairs as sources of information to reduce the number of questions. As happens with the paradigm scoring strategy, an heuristic querying algorithm initially designed for the assessment of the feasibility of this strategy is first presented (Section 4.3.1), and later in this chapter a refinement based on decision trees is described (Section 4.5.2).

3. *Equivalent paradigms.* When more than one paradigm provides exactly the same set of suffixes but with different morphological inflection information, no additional polar question can be asked in order to discriminate between them and, consequently, the iterative querying process described above never reaches the situation in which only a single candidate stem/paradigm pair remains in L. For example, in Spanish many adjectives such as *alto* and nouns such as *gato* are inflected identically. Therefore, two paradigms that produce the same collection of suffixes {*-o* (masculine, singular), *-a* (feminine, singular), *-os* (masculine, plural), *-as* (feminine, plural)} but with different morphological inflection information are defined in the monolingual dictionary (the stem *alt-* is assigned to the inflection paradigm whose lexical category is *adjective*, while *gat-* is assigned to the paradigm whose lexical category is *noun*). This issue also affects paradigms with the same lexical category: *abeja* and *abismo* are nouns that are inflected identically; however, *abeja* is feminine, whereas *abismo* is masculine. When adding unknown words such as *gato* or *abeja*, no polar question can consequently be asked in order to discriminate between both paradigms. The solution to this issue is out of the scope of this dissertation. Future research lines aimed at solving these problems are described in Section 5.2, including some preliminary work (Sánchez-Cartagena et al., 2012a) on an automatic process based on an n-gram language model of lexical categories and morphological inflection information. In all the experiments presented in this chapter, when all the candidate stem/paradigm pairs in L generate the same set of word forms, the process is finished. If one of the candidates in L is the stem/paradigm pair used as a reference, the result of inserting the word is considered successful. In the experiments described in Section 4.5.3, for around 87% of words in the test sets, the final solution contained more than one candidate stem/ paradigm pair in L.

4.3 Viability of the approach

This section presents a set of experiments that confirm that entries can be successfully inserted with high success rate and that non-expert users are able to correctly answer the polar questions. In this set of experiments, heuristic approaches for the computation of the feasibility score and the querying algorithm have been followed. They are described (Section 4.3.1) prior to the presentation of the experimental setup and the discussion of the results (Section 4.3.2). Note that more sophisticated alternatives based on well-known probabilistic models are described and evaluated in Section 4.5.

4.3.1 Heuristic approaches

4.3.1.1 Heuristic feasibility score

A feasibility score is assigned to each stem/paradigm candidate $c_n \in L$ using a large monolingual corpus C. This score should be higher for the candidates c_n that are more likely to be the correct one, according to the evidence found in the monolingual corpus. The more accurate the feasibility score, the fewer questions that will need to be asked by the querying algorithm.

The heuristic feasibility score used in the experiments presented in this section has been defined under the assumption that, the more word forms in $I(c_n)$ found in the monolingual corpus C, the most likely is that the stem/paradigm candidate c_n is the most appropriate one. Thus, one possible way to compute the score is

$$\text{Score}(c_n) = \frac{\sum_{\forall w' \in I(c_n)} \text{Appear}_C(w')}{\sqrt{|I(c_n)|}},$$

where $\text{Appear}_C(w')$ is a function that returns 1 when the inflected form w' appears in the corpus C and 0 otherwise, and I is the expansion function as defined before. The square root term is used to avoid very low scores for large paradigms which include many suffixes.[6]

One potential problem with the previous formula is that all the inflected word forms in $I(c_n)$ are taken into account, including those that, although morphologically correct, are not very frequent in the language and, consequently, in the corpus.[7] In order to

[6]Preliminary experiments were carried out in order to find the most adequate value of the exponent to be used in the denominator. A set of words from the Spanish monolingual dictionary in the Apertium RBMT system were chosen and inserted by following the strategy described in Section 4.2. It was found out that the value that maximised the number of words of the test set for which the best paradigm was assigned the highest feasibility score was between 0.4 and 0.6, hence 0.5 (square root) is used.

[7]For instance, the paradigms for verbs in the Apertium Spanish monolingual dictionary include multiple combinations of enclitic pronouns. Some of the enclitic pronoun combinations are seldom

overcome this, Score(c_n) is redefined as

$$\text{Score}(c_n) = \frac{\sum_{\forall w' \in I'_C(c_n)} \text{Appear}_C(w')}{\sqrt{|I'_C(c_n)|}},$$

where $I'_C(c_n)$ is the difference set

$$I'_C(c_n) = I(c_n) - \text{Unusual}_C(c_n).$$

The function $\text{Unusual}_C(c_n)$ uses the words in the dictionary already assigned to p_i as a reference to obtain which of the inflected word forms generated by p_i are not frequent in the corpus C. Let $T(p_i)$ be a function retrieving the set of stems in the dictionary assigned to the paradigm p_i. For each of the suffixes f_{ij} in each paradigm p_i, the following ratio is computed:

$$\text{Ratio}(f_{ij}, p_i) = \frac{\sum_{\forall t \in T(p_i)} \text{Appear}_C(tf_{ij})}{|T(p_i)|},$$

and the set $\text{Unusual}_C(c_n)$ is built by concatenating the stem t to all the suffixes f_{ij} with $\text{Ratio}(f_{ij}, p_i)$ below a given threshold Θ.

Recall that the expansions of the candidate stem/paradigm pairs from the example in Section 4.2.2 are $I(c_1)$={*policies*, *policiess*}; $I(c_2)$={*policie*, *policies*}; $I(c_3)$={*policy*, *policies*}; and $I(c_4)$={*policy*, *policies*, *policied*, *policying*}. Using a large monolingual English corpus C, word forms *policies* and *policy* will be easily found; the other inflected word forms (*policie*, *policiess*, *policied* and *policying*) will not be found. To simplify the example, assume that $\text{Unusual}_C(c_n) = \emptyset$ for all the candidates; the resulting scores will be: Score(c_1)=0.71, Score(c_2)=0.71, Score(c_3)=1.41, Score(c_4)=1.

4.3.1.2 Heuristic querying algorithm

The querying algorithm chooses, in each step of the iterative querying process described in Section 4.2.2, the word form that the user will have to validate. The algorithm should select the word forms in such a way that the cardinality $|L| = 1$ is reached after asking as few questions as possible. To that end, the algorithm uses the feasibility score in order to predict the answer of the user, and the membership relation between the different word forms and the candidate stem/paradigm pairs in order to reduce the size of the set of candidates L as fast as possible.

The heuristic querying algorithm treats L as a list and sorts it in descending order by Score(c_n). The membership relation is defined by the function $G(w', L)$, that returns the number of candidates $c_i \in L$ for which $w' \in I(c_i)$. Depending on the elements

used and may not have semantic sense for most of the words assigned to the paradigm. For instance, the concatenation of the third-person plural form of a verb in imperative mood, the reflexive enclitic pronoun *se* and the personal enclitic pronoun *los* (e.g. *viajénselos*) is rarely used in Spanish.

contained in L, the criterion followed by the algorithm in order to select the word form to be queried can be either *confirmation* or *discarding*. The algorithm starts in confirmation mode.

Confirmation mode. In this mode, it is assumed that all the word forms generated by the candidate with highest score c_1 in the list L are correct. Consequently, the user is asked about the word form $w' \in I(c_n)$ with the lowest value for $G(w', L)$ because, if it is accepted, a significant part of the paradigms in L will be removed from the list. The algorithm keeps working in confirmation mode until one of these conditions is met:

- Only one single candidate remains in L and the process stops.
- All the word forms $w' \in I(c_1)$ are generated by all the remaining candidates in L. In this situation, if a word form $w' \in I(c_1)$ is accepted by the user, the list L will remain unchanged. If it is discarded, L will become empty. The algorithm moves to discarding mode in order to break this lockout.

Discarding mode. In this mode, the system has accepted c_1 as a possible solution, but it needs to check whether any of the remaining candidates in L is more suitable. Therefore, the system asks the user about those word forms $w' \notin I(c_1)$ with the highest possible value for $G(w', L)$. This process is repeated until one of these two conditions is met:

- Only c_1 remains in L and the process stops.
- A word form $w' \notin I(c_1)$ is accepted. This means that some of the other candidates is better than c_n.

If the second situation holds, the system removes c_1 from L and goes back to the confirmation mode.

In both modes, if there are multiple word forms with the same value for $G(w', L)$, the system chooses the one with higher $\text{Ratio}(f_{ij}, p_i)$, that is, the most usual in C and, consequently, the most likely to be familiar to the user.

4.3.2 Experimental setup: querying real users

In order to confirm that average speakers of a language can successfully use the strategy described in Section 4.2 to insert new entries in a monolingual dictionary, a group of 4 human users (computer engineers without advanced linguistic knowledge) was chosen and asked to add a set of words to a monolingual dictionary in the Apertium RBMT platform (Forcada et al., 2011).

4.3.2.1 Data

The Apertium Spanish monolingual dictionary from the language pair Spanish–Catalan[8] was chosen as the dictionary in which the new entries were to be inserted. A Spanish Wikipedia dump[9] was chosen as the monolingual corpus used to compute the feasibility scores. The value of the threshold Θ used to compute the set $\text{Unusual}_C(c_n)$ was set to 0.1.[10] In order to build the test set, that is, the word forms that the users must insert into the monolingual dictionary, the process described below was carried out.

Firstly, the paradigms meeting the following restrictions were selected:

- The lexical information they encode includes an open lexical category.[11] When creating the monolingual dictionary for a given language, words from closed lexical categories constitute a small set which is usually inserted by expert users in the early stages of the development.
- When removing the 5 most frequent words assigned to them for being included in the test set, at least one word form of one of the remaining words can be found in the monolingual corpus. This is needed to properly compute $\text{Unusual}_C(c_n)$ when the words in the test set are inserted.
- They have at least six words assigned in the dictionary (necessary condition for meeting the previous one).

From the set of paradigms fulfilling the previous conditions, the 30 paradigms assigned to the 30 entries whose inflected word forms have the highest aggregated frequency in the monolingual corpus were chosen.[12] From each of them, the 5 most common entries[13] were extracted and a global set with 150 entries was built. To ensure that users were familiar with the word forms to be added, the most frequent word form in the monolingual corpus was chosen from each of the aforementioned entries in order to obtain the final test set. These 150 word forms were divided into 4 subsets, one of them for each of the four non-expert users, introducing some redundancy which permitted the computation of inter-annotator and intra-annotator agreements. Each subset contained 50 word forms and was built as follows:

[8]Revision 33900 in the Apertium Subversion repository: `https://apertium.svn.sourceforge.net/svnroot/apertium/trunk/apertium-es-ca`

[9]`http://dumps.wikimedia.org/eswiki/20110114/eswiki-20110114-pages-articles.xml.bz2`

[10]Preliminary experiments showed that when this value of the threshold is used the most infrequent inflected forms (like the aforementioned unusual combinations of enclitic pronouns in Spanish) are not taken into account in the computation of Score(cn).

[11]Verb, noun, adjective, adverb or interjection.

[12]Repeated paradigms are not allowed in the list; for instance, if the first two entries with the highest frequency belong to the same paradigm, the place of the second one is taken by another paradigm.

[13]The 5 entries with the highest sum of the frequencies in the monolingual corpus of the word forms resulting from their expansion.

- 30 word forms were extracted from the global set, and they were not shared with any other user. Additionally, 5 word forms randomly chosen from the set of 30 word forms were included twice, in order to compute intra-annotator agreement.
- The remaining 15 word forms were used to compute inter-annotator agreement. Each pair of users was assigned 5 word forms extracted from the global set. Since 4 users took part in the evaluation, each user was paired with 3 other users.

A sentence randomly chosen from the monolingual corpus containing the word form to be classified was also shown to the user. As it has been previously stated, the strategy presented in this chapter is meant to be applied when a user of an MT system wants to translate a sentence that contains a word that is not present in its dictionaries. In the proposed evaluation scenario, the sentence helps to ease the classification of homographs.[14]

4.3.2.2 Evaluation metrics

In order to estimate the reliability of the results, pair-wise inter-annotator agreement for each pair of annotators and intra-annotator agreement for each annotator were computed using Cohen's kappa (Cohen, 1960). In addition, the following evaluation scores about the performance of the strategy for inserting entries into the dictionary were calculated:

- Success rate: percentage of words from the test set that were tagged with the paradigm originally assigned to them in the monolingual dictionary. Recall that, as pointed out in Section 4.2.2, when the solution found contains multiple paradigms that share the same inflected word forms, the result is considered correct if one of them is the one originally assigned to the word in the monolingual dictionary.
- *Average precision and recall*: precision (P) and recall (R) were computed as

$$P(c, c') = \frac{|I(c) \cap I(c')|}{|I(c)|} \qquad R(c, c') = \frac{|I(c) \cap I(c')|}{|I(c')|},$$

  where c is the stem/paradigm pair obtained and c' is the stem/paradigm pair originally found in the dictionary. This metric is intended to assess the similarity between the chosen paradigm and the one originally present in the dictionary.
- Average position of the correct candidate in the initial sorted list L of stem/paradigm pairs. This metric gives an estimation of the accuracy of the feasibility score.

[14] Words written in the same way but with different grammatical features. For instance, *books* can be either a plural noun or a third-person, singular verb.

Table 4.1: Inter-annotator agreement computed using Cohen's kappa in the experiments involving the heuristic feasibility score and querying algorithm described in Sections 4.3.1.1 and 4.3.1.2, respectively.

User pair	κ
A-B	0.76
A-C	0.74
A-D	0.71
B-C	0.76
B-D	1.00
C-D	1.00
average	0.83

- Average number of questions asked to the user for each word.

These metrics (except for the last one, since it needs human interaction in order to be calculated) were also computed for a non-interactive baseline in which the chosen paradigm was that with the highest feasibility score, so as to assess whether the sole feasibility score computed from a monolingual corpus is enough to correctly choose the right paradigm.

4.3.3 Results and discussion

Before describing and discussing the results obtained in the experiments with real users, it is worth assessing their reliability by analysing the values observed for the pair-wise inter-annotator agreement of each pair of users (shown in Table 4.1) and the intra-annotator agreement of each user (depicted in Table 4.2). According to Cohen (1960), a kappa value between 0.6 and 0.8 is usually interpreted as *good agreement*, and when it ranges between 0.8 and 1.0 it is usually stated that there is a *very good agreement* between annotators. Since all the values obtained fall in one of these ranges, it can be concluded that each of the 4 users was quite consistent (high intra-annotator agreement) and that they agreed in their answers (high inter-annotator agreement), which ensures the confidence on the remainder of the results.

Table 4.3 shows the value of the five evaluation metrics for the interactive framework with the heuristics defined above, and the non-interactive baseline method that consists of just choosing the candidate stem/paradigm pair with the highest feasibility score. Confidence intervals were estimated with 95% statistical significance with a *t-test*. Results show a quite high success rate, that confirms that users are able to correctly answer to the polar questions asked by the system. The difference with the non-interactive baseline is remarkable: the feasibility score is not as accurate as the users' answers for correctly assigning paradigms to new words. The recall of the baseline,

Table 4.2: Intra-annotator agreement computed using Cohen's kappa in the experiments involving the heuristic feasibility score and querying algorithm described in Sections 4.3.1.1 and 4.3.1.2, respectively.

User	κ
A	1.00
B	0.73
C	1.00
D	1.00
average	0.93

Table 4.3: Success rate, precision (P), recall (R), position of the right paradigm in the initial sorted list of candidates and average number of questions asked to the users (confidence intervals for $p \leq 0.05$) when inserting entries in the Apertium Spanish monolingual dictionary using the approach presented in this chapter, and a non-interactive baseline in which the stem/paradigm pair with highest feasibility score is chosen.

System	success rate	P	R	initial position in L	# questions
non-interactive	$16\% \pm 5$	$49\% \pm 5$	$88\% \pm 3$		-
interactive	$88\% \pm 5$	$94\% \pm 3$	$95\% \pm 3$	12.0 ± 2.0	6.1 ± 0.7

however, is relatively high, which suggests that the paradigms chosen by the non-interactive approach generate many of the right surface forms, but also many incorrect ones. In addition, the values for precision and recall (95%; higher than success rate) for the interactive approach suggest that those words which were assigned to incorrect paradigms, were assigned to paradigms that share many word forms with the right one.

These results (Sánchez-Cartagena et al., 2012a) are compatible with those obtained in a previous evaluation of the same approach (Esplà-Gomis et al., 2011a), where the words from the test were extracted from 166 different paradigms (a couple of words were randomly chosen from each paradigm), 10 non-expert users took part, and annotator agreement metrics were not computed. In those experiments, the value of average precision and recall was slightly behind 90%, although the recall of the non-interactive baseline was much lower than the one observed in these experiments.

Some of the most common mistakes made by users were related to verbs and superlative adjectives. Spanish morphological rules allow multiple concatenations of enclitic pronouns at the end of verbs. In many occasions, users rejected forms of verbs with too many enclitic pronouns or for which some concrete enclitics had no semantic sense (for instance, *viájasela*). This happens because, in order to reduce the number of possible paradigms, dictionaries in Apertium can assign some words to existing paradigms which are a superset of the correct one; since the included semantically incorrect word forms will never occur in a text to be translated, this, in principle, may be safely done.[15] Regarding superlative adjectives, Apertium contains paradigms for adjectives which have superlative form and for those which do not have it. Users often accepted the superlative form of an adjective which, according to the Apertium dictionary, does not have it (such as *ultimísimo*).[16]

4.4 Exploiting correlation between source and target languages to improve the feasibility score

As it has been stated at the beginning of this chapter, if the user who inserted the SL unknown word in the SL monolingual dictionary also knows its translation into the TL, the mechanism described in Section 4.2 and evaluated in the previous section can be used to allow her to insert its translation into the TL monolingual dictionary.[17] However, the information obtained by identifying the most appropriate SL paradigm can be used to improve the feasibility score of the candidate TL paradigms. In this section, an enhancement to the feasibility score based on the correlation between SL

[15]These types of paradigms also motivated the introduction of the $\text{Unusual}_C(c_n)$ function in the computation of the heuristic feasibility score (see Section 4.3.1.1).

[16]Despite being correct, the word form *ultimísimo* was not generated by the paradigm assigned to the adjective *último* in the Apertium Spanish monolingual dictionary.

[17]Once a word has been inserted into both monolingual dictionaries, the insertion of the corresponding entry in the bilingual dictionary can be done automatically.

and TL paradigms in the entries already existing in the dictionaries is presented. The enhancement is first described in Section 4.4.1 and its evaluation is discussed in Section 4.4.2.

4.4.1 Identifying the correlation between paradigms

Bilingual dictionaries in RBMT allow one to establish the relationships between the lexical forms in the two languages involved in the translation. The paradigm corresponding to each lexical form in each entry included in the bilingual dictionary can be easily obtained from the word entries in the monolingual dictionaries. In the resulting relationship between SL and TL paradigms, it can be observed that, usually, only a reduced set of TL paradigms correspond to an SL paradigm and only a smaller subset of them appear in a relatively high amount of entries associated to that SL paradigm. This observation suggests that knowing the paradigm of an SL word may help to choose the best paradigm for its TL counterpart.

In order to statistically confirm the observed close relationship between SL and TL paradigms, the conditional probability of p_i^{TL} being the most appropriate paradigm for a TL word once it is known that the paradigm of its SL equivalent is p_j^{SL} can be estimated by maximum likelihood as follows:

$$p(p_i^{TL}|p_j^{SL}) = \frac{\mathrm{count}(p_j^{\mathrm{SL}}, p_i^{\mathrm{TL}})}{\mathrm{count}_{\mathrm{SL}}(p_j^{\mathrm{SL}})}$$

where $\mathrm{count}(p_j^{\mathrm{SL}}, p_i^{\mathrm{TL}})$ is the number of entries in the bilingual dictionary whose SL paradigm is p_j^{SL} and whose TL paradigm is p_i^{TL}, and $\mathrm{count}_{\mathrm{SL}}(p_j^{\mathrm{SL}})$ is the number of entries whose SL paradigm is p_j^{SL}.

The conditional entropy $H(p^{\mathrm{TL}}|p_j^{\mathrm{SL}})$ defines the uncertainty of the random variable that represents the possible correct paradigm of a TL word (p^{TL}) once it is known that the paradigm of its SL equivalent is p_j^{SL}. If T is the set of TL paradigms, it is computed as:

$$H(p^{\mathrm{TL}}|p_j^{\mathrm{SL}}) = -\sum_{p_k^{TL} \in T} p(p_k^{\mathrm{TL}}|p_j^{\mathrm{SL}}) \cdot \log_2 p(p_k^{\mathrm{TL}}|p_j^{\mathrm{SL}})$$

This entropy (measured in bits) has been computed for all the SL paradigms of the Apertium Catalan–Spanish and English–Spanish dictionaries, and the resulting histograms are shown in Figures 4.2 and 4.3, respectively.

In both cases, the TL paradigms corresponding to the translation of most SL paradigms present a value of entropy under 0.5,[18] which confirms the strong correlation between the paradigms. Note that the proportion of SL paradigms whose related

[18] An entropy of 0.5 corresponds to an uncertainty between that of a random variable with only one possible outcome (entropy 0: the value of the random variable that represents the TL paradigm would be completely determined by the SL paradigm) and that of a random variable with two equally likely outputs (entropy 1).

Figure 4.2: Histogram that represents the value of the conditional entropy (see Section 4.4.1) of the random variable that represents the TL paradigm assigned to a word given the paradigm of its SL equivalent (according to the bilingual dictionary) for the Apertium Catalan–Spanish dictionaries. The total number of paradigms in the Catalan (SL) monolingual dictionary is 417.

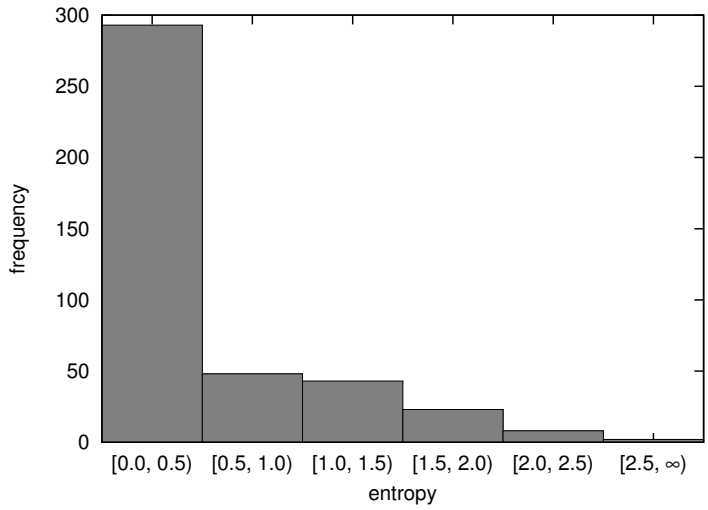
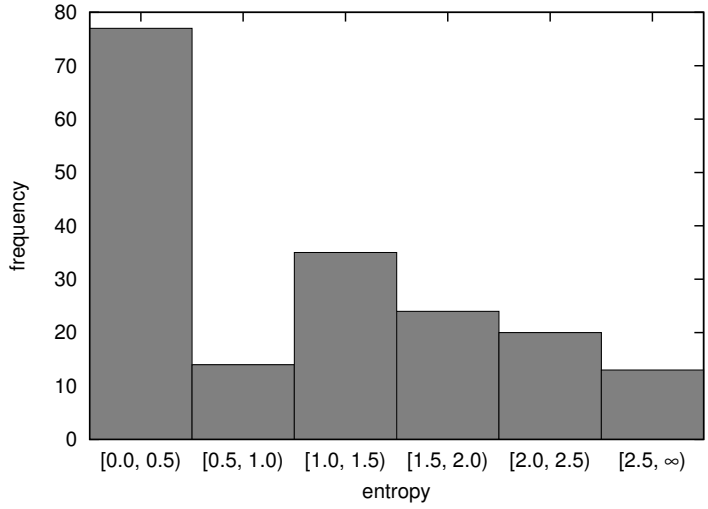


Figure 4.3: Histogram that represents the value of the conditional entropy (see Section 4.4.1) of the random variable that represents the TL paradigm assigned to a word given the paradigm of its SL equivalent (according to the bilingual dictionary) for the Apertium English–Spanish dictionaries. The total number of paradigms in the English (SL) monolingual dictionary is 183.



TL paradigm has an entropy under 0.5 is higher in the Catalan–Spanish dictionaries, probably because these languages are more closely related than English–Spanish.

Thus, when a word form is being inserted into the TL monolingual dictionary and the paradigm assigned to its SL equivalent is known, candidate paradigms which have a strong correlation with the known SL paradigm should have a higher feasibility score, since there is evidence that suggest that they could be the most appropriate paradigms. This evidence can be be taken into account by simply multiplying the original feasibility score of each TL candidate stem/paradigm p_i^{TL} by the conditional probability $p(p_i^{TL}|p_j^{SL})$, where p_j^{SL} is the paradigm of the SL equivalent.[19]

4.4.2 Evaluating the enhancement of the feasibility score

In order to asses the positive impact of enhancing the feasibility score with the SL paradigm information, two types of experiments were carried out. First, the users who took part in the experiments described in Section 4.3.2 were asked to repeat the insertion of the words into the monolingual dictionary with the enhanced feasibility score. An automatic evaluation in which only the position of the correct stem/paradigm pair in the sorted list of candidates L was evaluated was also performed.

4.4.2.1 Human evaluation

The experiments from the previous section were repeated with Spanish acting as the TL and assuming that the Catalan translation (Catalan acted as the SL) of each word from the test set was already introduced and updating the feasibility scores as described in Section 4.4.1. The Catalan monolingual dictionary and the Catalan–Spanish bilingual dictionary used were also borrowed from the Apertium project.[20] The intra- and inter-annotator agreement scores were very similar to those reported in the first round of experiments. The remaining scores are depicted in Table 4.4. As in the previous evaluation, confidence intervals were estimated with 95% statistical significance with a *t-test*.

Although neither the success rate nor the precision and recall are improved when the information of the SL paradigm is included, a statistically significant improvement in the position of the correct paradigm in the list of candidate stem/paradigms and in the number of questions asked to the users is obtained. This confirms that the information provided by the SL paradigm is valuable, at least for closely-related languages such as Spanish and Catalan. These results also suggest that there is a correlation between the

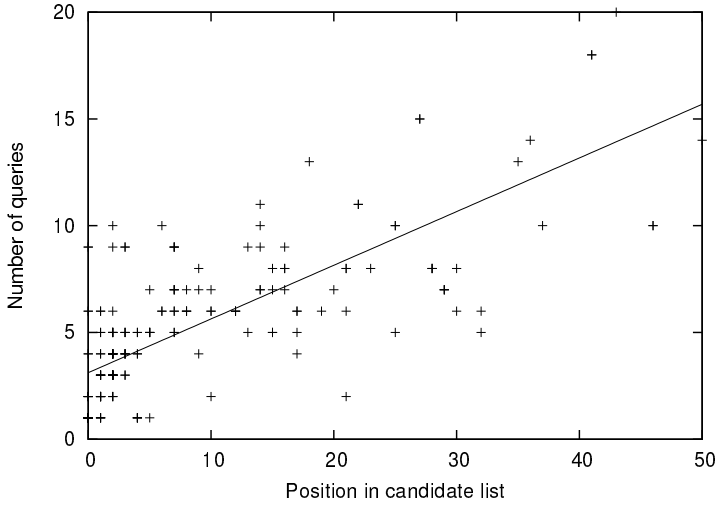
[19] In the experiments carried out in order to evaluate this enhancement, a simple smoothing has been applied: when the value of one of the two factors is zero, it is replaced by the lowest non-zero value among all the candidate stem/paradigms divided by 10.

[20] Revision 33900 in the Apertium Subversion repository: `https://apertium.svn.sourceforge.net/svnroot/apertium/trunk/apertium-es-ca`

Table 4.4: Success rate, precision (P), recall (R), position of the right stem/paradigm pair in the initial sorted list of candidates and average number of questions asked to the users (confidence intervals for $p \leq 0.05$) when inserting entries into the Apertium Spanish monolingual dictionary. The heuristic feasibility score described in Section 4.3.1.1 (*interactive*) is compared with an enhancement that exploits the correlation between SL and TL paradigms, as explained in Section 4.4.1 (*+ SL paradigm*).

System	success rate	P	R	initial position in L	# questions
interactive	$88\% \pm 5$	$94\% \pm 3$	$95\% \pm 3$	12.0 ± 2.0	6.1 ± 0.7
+ SL paradigm	$87\% \pm 5$	$94\% \pm 2$	$98\% \pm 2$	2.3 ± 0.7	4.4 ± 0.4

Figure 4.4: Correlation, obtained by least squares linear regression, between the position of the right candidate stem/paradigm in the sorted list of candidates and the number of questions asked to the user. The value of Pearson's r is 0.8377.



quality of the feasibility score, measured as the average position of the right candidate in the list L, and the number of questions asked to the users. The most reliable the feasibility score, the most often the user will accept a word form that causes the discarding of the maximum number of candidates. In order to formally confirm this correlation, the function that maps the position of the right candidate in the list L to the number of questions asked to the user has been obtained using the least squares linear regression method from the results of the experiments described in this section and in Section 4.3.2. The function, plotted in Figure 4.4 together with all the data points extracted from the results of the experiments, confirms this correlation.

4.4.2.2 Automatic evaluation

Results of the human evaluation have shown a significant correlation between the position of the correct candidate stem/paradigm pair in the list L and the number of questions asked to the users. Since the first metric can be computed without human interaction, the detected correlation permits automatically estimating the improvement in the efficiency of the approach brought by the SL paradigm information with a much bigger test set. Accordingly, the position of the correct stem/paradigm pair in L when inserting an entry into the Spanish monolingual dictionary has been computed with a test set bigger than that used in the experiments carried out with human interaction. In order to detect whether the positive impact of considering the SL paradigm depends on how related are the languages involved, the experiment has covered two different language pairs: English–Spanish[21] and Catalan–Spanish.[22]

In this automatic evaluation, the test set was built as follows. First, all the paradigms from open lexical categories that contain at least two entries were selected. Then, for each entry of each selected paradigm, the most frequent inflected word form in the monolingual corpus used to compute the feasibility score was added to the test set. For each word form in the test set, its corresponding word entry was temporarily removed from the dictionary, and the position of the correct stem/paradigm candidate was computed using the heuristic feasibility score defined in Section 4.3.1 (labelled as *baseline* in Table 4.5) and, afterwards, with the enhanced one described in this section (labelled as *feasibility score using SL paradigm*). Confidence intervals were estimated with 95% statistical significance with a *t-test*.

One of the most important elements of the linguistic information encoded by an inflection paradigm is the lexical category of the words that belong to it. It is also interesting to assess whether an improvement in the accuracy of the feasibility score when using the information of the SL paradigm is caused only by the correlation between the lexical categories in SL and TL, or the rest of information provided by the paradigms is also useful. To do so in the context of automatic evaluation, a third strategy to calculate the position of the right stem/paradigm pair in L (labelled as *enhanced feasibility score using lexical category* in Table 4.5) has been added to the experimental setup: a modification of the enhanced approach in which the conditional probability $p(p_i^{TL}|p_j^{SL})$, calculated from the relative frequency of paradigms in the bilingual dictionary, is computed assuming that, for a given language, all the paradigms that generate the same lexical category are grouped into a single one. In other words, the conditional probability $p(p_i^{TL}|p_j^{SL})$ described in Section 4.4.1 is the same for all the candidate paradigms with the same lexical category.

[21] Revision 36247 in the Apertium Subversion repository: `https://svn.code.sf.net/p/apertium/svn/trunk/apertium-en-es`.

[22] Revision 33900 in the Apertium Subversion repository: `https://svn.code.sf.net/p/apertium/svn/trunk/apertium-es-ca`.

Table 4.5: Average position of the right paradigm in the initial sorted list of candidates (confidence intervals for $p \leq 0.05$) when inserting, using the methods described in Section 4.4.2.2, each entry in the Spanish monolingual dictionary of the Catalan–Spanish language pair, and each entry in the Spanish monolingual dictionary of the English–Spanish language pair. These experiments have been carried out without human interaction.

Language pair	System	initial position in L
Catalan–Spanish	baseline	21.9 ± 0.2
	feasibility score using lexical category	15.1 ± 0.2
	feasibility score using SL paradigm	13.2 ± 0.2
English–Spanish	baseline	26.1 ± 0.3
	feasibility score using lexical category	21.0 ± 0.3
	feasibility score using SL paradigm	21.1 ± 0.3

As shown in Table 4.5, when the evaluation is extended to the whole dictionary, significant improvement in the position of the right candidate in L still occurs, even in the case of a less related language pair such as English–Spanish. However, for English–Spanish, the only information from the SL paradigm that helps to classify the TL word is the lexical category. This can be explained by the fact that closely related languages share the inflection scheme (for instance, in Spanish and Catalan nouns have gender and number) and most words keep their inflection features when they are translated (for instance, most masculine nouns whose plural is built by appending *-s* in Catalan, are also masculine and their plural is built in the same way in Spanish). On the contrary, inflection schemes are different in less related languages (such as English and Spanish) and, therefore, the inflection information encoded in paradigms is not useful. Nevertheless, a deeper analysis that includes other language pairs should be carried out in order to clarify the impact of lexical category and morphological inflection information in the enhancement of the feasibility score.

4.5 Probabilistic alternatives

Once the experiments with (real) non-expert users have proved the feasibility of the strategy for inserting entries into RBMT morphological dictionaries that is being described in this chapter, this section presents more rigorous and principled alternatives for the feasibility score and the querying algorithm which are meant to replace the intuitive heuristics defined in Section 4.3.1. In particular, hidden Markov models (Rabiner, 1989) are used for computing the feasibility score, while binary decision trees are the foundation of the new querying algorithm. First, these two new alternatives are described and, afterwards, the experiments carried out in order to properly evaluate them are reported (Section 4.5.3), and their results are presented and discussed (Section 4.5.4).

4.5.1 Paradigm scoring with hidden Markov models

The heuristic approach for computing the feasibility score of a candidate stem/paradigm pair described in Section 4.3.1 is based on the proportion of word forms found in a monolingual corpus. That kind of score, however, can be misleading under certain circumstances:

- When the word forms generated by the correct paradigm cannot be found in the corpus.
- Even when they are found in the monolingual corpus, it does not necessarily mean that they are correct forms of the word to be inserted into the dictionary. Consider, for instance, that the homograph word *complete*, that has been found in the sentence *I needed complete silence*, is to be inserted into the monolingual dictionary. Clearly, a paradigm that generates, among others, the word forms *completed* and *completing* (which would signify that their lexical category is *verb*) should not obtain a high feasibility score. However, the heuristic scoring method described in Section 4.3.1 would assign a high feasibility score to that candidate paradigm because most of the words forms generated can be found in a monolingual corpus.

These two limitations of the heuristic feasibility score can be addressed by taking into account the sentence (context) in which the word to be inserted is found. To that end, a solution based on first-order hidden Markov models (Rabiner, 1989, HMMs) is proposed in this section. A first-order HMM is a statistical model in which the system being modelled emits a sequence of observable outputs. Each time it emits an output, its internal state can change. The probability of emitting each observable output depends on the state, which cannot be observed (it is *hidden*), while the current state depends only on the previous one.[23] More formally, a first-order HMM is defined as $\lambda = (\Gamma, \Sigma, A, B, \pi)$, where Γ is the set of states, Σ is the set of observable outputs, A is the $|\Gamma|\times|\Gamma|$ matrix of state-to-state transition probabilities, B is the $|\Gamma|\times|\Sigma|$ matrix with the probability of each observable output $\sigma \in \Sigma$ being emitted from each state $\gamma \in \Gamma$, and the vector π, with dimensionality $|\Gamma|$, defines the initial probability of each state. The parameters (A, B, π) of an HMM can be estimated from a sequence of observable outputs, so that the probability of observing the sequence given the parameters is maximised. Once they are estimated, the most probable sequence of states given a new sequence of observable outputs and the model parameters can be obtained. Moreover, the process can be constrained through the definition of *allowed states*: for each observable output in a sequence, its possible set of states can be defined in advance and included in the training data. The emission of an observable output

[23]The order of an HMM determines the length of the sequence of previous states on which the current state depends. In a first-order HMM, the current state depends only on the previous one, while in a second-order HMM, the current state depends on the two previous states.

from a state that does not belong to its set of allowed states will be an impossible event for the HMM.

The HMM used for computing feasibility scores models sequences of words. States represent paradigms and observable outputs are obtained from the word forms. Thus, Γ is built from the set of all the paradigms in the dictionary and Σ is obtained as the union of the suffixes produced by all these paradigms: $\Sigma = \bigcup_{p_i \in P} \bigcup_{(f_{ij}, m_{ij}) \in p_i} f_{ij}$. The parameters (A, B) of the HMM[24] are obtained from an untagged text corpus in a way very similar to how they are obtained when HMMs are used for part-of-speech tagging (Cutting et al., 1992), and trained by means of the Baum-Welch expectation-maximisation algorithm (Baum, 1972). More precisely, the sequence of observable outputs and sets of allowed states from which the parameters are optimised is built from a monolingual untagged text corpus C as described below (Table 4.6 depicts an example sentence and the training data extracted from it):

- The entries of the monolingual dictionary in which new words will be inserted are expanded in order to obtain the set F of all possible word forms. Let D be the set of entries already present in the dictionary, the set of all possible word forms is computed as $F = \bigcup_{t/p_i \in D} I(t/p_i)$.
- The set of allowed states of each word form w in the corpus that belongs to F contains solely the state(s) associated with the corresponding paradigm according to the dictionary.[25] If the word is homograph, i.e. more than one paradigm may generate w, the set of allowed states contains the states associated with as many paradigms as found in the monolingual dictionary. With respect to the observable output, its value is the longest suffix of w that can be found in Σ. The whole word form cannot be used as observable output because the words to be inserted

[24] In order to simplify the computation of the probabilities, the estimation of the initial probability of each state (π) can be conveniently avoided by assuming that each sentence begins and ends with the end-of-sentence mark; thus, $\pi(\gamma)$ is 1 when γ is the end-of-sentence mark, and 0 otherwise.

[25] Note that a paradigm may correspond to multiple states in the HMM because of the following phenomenon. Consider again the four paradigms presented in Section 4.2: p_1 = {-ϵ, -*s*}; p_2 = {-*y*, -*ies*}; p_3 = {-*y*, -*ies*, -*ied*, -*ying*}; and p_4 = {-*a*, -*um*}; and the surface form to be inserted: *w=policies* (plural form of the noun *policy*). The compatible stem/paradigm pairs that need to be scored are: c_1=*policies*/p_1; c_2=*policie*/p_1; c_3=*polic*/p_2; and c_4=*polic*/p_3. If states were directly mapped to paradigms, the candidate stem/paradigm pairs c_1 and c_2 would have the same score. In order to avoid this situation and assign different scores to candidates with the same paradigm but different stem (and thus different suffix of the paradigm subtracted from the initial surface form in order to create the stem) the mapping between paradigms and states of the HMM is performed as follows. For each suffix $s_i \in p_a$ of the paradigm p_a for which exists another suffix of the paradigm $s_j \in p_a : s_i = ts_j$, being t any non-empty sequence of letters, a new *specialised* state $\gamma_{p_a \# s_i}$ is created in addition to the general state γ_{p_a} that corresponds to the paradigm. From the aforementioned set of paradigms, the states γ_{p_1}, $\gamma_{p_1 \# s}$, γ_{p_2}, γ_{p_3} and γ_{p_4} would be created. When assigning the set of allowed states to a word in the training corpus (or in the sentence in which the surface form to be inserted into the dictionary is contained), for each compatible paradigm, the specialised state that corresponds to the paradigm and the observable output is assigned. If it does not exist, the general state is assigned instead.

Table 4.6: Training data extracted from an example sentence, assuming that the dictionary only contains the paradigms p_1 = {*-ϵ*, *-s*}; p_2 = {*-y*, *-ies*}; p_3 = {*-y*, *-ies*, *-ied*, *-ying*}; p_4 = {*-a*, *-um*}; and p_5 = {*-ϵ*}. The word *today* is the only word in the sentence that cannot be found in the dictionary.

Sentence (word forms):	the	baby	is	crying	today
Observable output:	*-ϵ*	*-y*	*-ϵ*	*-ying*	*-y*
Allowed states:	$\{p_5\}$	$\{p_2\}$	$\{p_5\}$	$\{p_2\}$	$\{p_1, p_2, p_3, p_5\}$

are not known in advance and thus Σ could not be defined before training the HMM.

- The set of allowed states of each word form w' that cannot be found in F is the set of states obtained from the paradigms of its compatible candidates, as described in Section 4.2. The observable output is again the longest suffix of w' that can be found in Σ.

Once the HMM is trained, the feasibility score of the different candidate stem/paradigm pairs is computed with the help of the input sentence in which the word to be inserted into the monolingual dictionary is present. That input sentence is analysed as it was done with the sentences in the training corpus. Assuming that the word to be inserted is in position t_{unk} of the input sentence, the feasibility score $Score(c_n)$ for each candidate $c_n \in L$ whose associated state is γ_{c_n} is computed as the probability of γ_{c_n} being the state at position t_{unk}, given the previously trained model and the sequence of observables. It is computed applying the following equation, which corresponds to Eq. (27) in the tutorial by Rabiner (1989):

$$Score(c_n) = \frac{\alpha_{t_{\text{unk}}}(\gamma_{c_n})\beta_{t_{\text{unk}}}(\gamma_{c_n})}{\sum_{m=1}^{|L|} \alpha_{t_{\text{unk}}}(\gamma_{c_m})\beta_{t_{\text{unk}}}(\gamma_{c_m})}$$

This equation accounts for the probability mass of all the sequences of states that include γ_{c_n} at position t_{unk} and are compatible with the sequence of observable outputs (numerator) normalised by the probability mass of all possible sequences of states that contain a state associated with a candidate stem/paradigm pair at position t_{unk} and are compatible with the sequence of observable outputs (denominator). Given state j at position t, $\alpha_t(j)$ accounts for the (forward) probability of the sub-sentence from the beginning of the sentence to position t, whereas $\beta_t(j)$ corresponds to the (backward) probability of the sub-sentence from position $t+1$ to the end of the sentence (Rabiner, 1989). Let $Y = (y_1, y_2, ..., y_T)$ be the sequence of observable outputs of the input sentence in which the unknown word is present, the forward and backward probabilities

are defined as follows:

$$\alpha_1(j) = \pi_j B_{j,y_1}$$
$$\alpha_{t+1}(j) = B_{j,y_{t+1}} \sum_{i \in \Gamma} \alpha_t(i) A_{i,j}$$

$$\beta_T(j) = 1$$
$$\beta_t(j) = \sum_{i \in \Gamma} \beta_{t+1}(i) A_{j,i} B_{i,y_{t+1}}$$

4.5.2 Selecting the word forms to be asked with binary decision trees

As it has been explained in Section 4.3.1, the heuristic querying algorithm assumes that the user will accept any word form generated from the candidate stem/paradigm pair with the highest score, and selects the word form whose acceptance causes the discarding of the highest number of candidates. However, that algorithm presents a number of limitations that are addressed by the new querying algorithm presented in this section:

- It is not able to detect when it is worth asking about a word form from a candidate stem/paradigm with a low feasibility score that would discard many candidates when it is rejected.
- If the feasibility score is not reliable enough, i.e., the user does not accept word forms from the candidate paradigm with the highest feasibility score, the number of questions asked may be incremented drastically.

In other words, a better querying algorithm should be more robust to unreliable feasibility scores and balance better the number of candidate stem/paradigm pairs discarded when a word form is accepted, the number of candidate stem/paradigm pairs discarded when a word form is rejected, and the likelihood of a word form being accepted or rejected according to the feasibility score. This behaviour is achieved thanks to the use of binary decision trees.

Decision trees are a tool commonly used to implement classifiers. Given a set of input features and data points (each point is assigned a class and has a value for each of the features), the classifier predicts the class of new data points. The internal nodes (decision nodes) of a decision tree are labelled with an input feature, an arc coming from an internal node exists for each possible feature value, and leaves are labelled with classes. In order to classify a new data point, the tree is traversed from the root to the leaves according the values of its features. The ID3 algorithm (Quinlan,

1986) has been proposed in order to build these trees. This algorithm follows a greedy approach; the resulting trees are therefore sub-optimal. In each iteration, it selects the most appropriate attribute to split the data set. The algorithm starts from the root of the tree with the whole data set S. In each iteration, an attribute a is picked for splitting S, being a the attribute that provides the highest information gain. A child node is then created for each possible value of a, with a new data set that contains only the elements that match that value. The information gain measures the difference in entropy before and after S is split. The entropy of a data set S is computed as:

$$H(S) = -\sum_{x \in X} p(x) log_2 p(x),$$

where X is the set of classes and the probability $p(x)$ of class x is usually computed as the proportion of elements from S that belong to the class x. The information gain $\mathrm{IG}(a, S)$ obtained when the dataset is split by an attribute a is obtained as:

$$\mathrm{IG}(a, S) = H(S) - \sum_{u \in U} p(u) H(u),$$

where U is the set of subsets obtained as a result of splitting S using the attribute a, and $p(u)$ is usually calculated as the proportion of the number of data points in u to the number of data points in S.

A decision tree can be used to implement a more robust querying algorithm. For each word form w to be inserted into the bilingual dictionary, the corresponding decision tree is built by means of the ID3 algorithm as follows:

- The data set S is made of all the stem/paradigm pairs compatible with w.
- The class of each data point is the corresponding stem/paradigm pair.
- The feature set is made up of the set of different word forms, that is $\bigcup_{c_i \in L} I(c_i)$.
- There are only two possible feature values: *yes* and *no*. Hence, the resulting decision tree is binary.

The tree will be traversed once: the values of the features of the data point to be classified are the answers provided by the user. Note that if the proportion of data points that belong to class x were used to compute $p(x)$ when calculating the entropy $H(S)$, all the candidate stem/paradigm pairs would obtain the same probability because the data set from which the decision tree is built contains a single instance of each stem/paradigm pair. However, we can take advantage of the feasibility score computed by means of HMMs described previously and assign that value to $p(x)$. Similarly, during the computation of the information gain IG, the value of $p(u)$ is calculated as the sum of the feasibility scores of the candidate stem/paradigm pairs in t divided by the sum of the probabilities of the candidates in S. The positive effect on the resulting decision

Table 4.7: Example of the candidate stem/paradigm pairs and feasibility scores obtained when trying to insert the word form *copies* into an English monolingual dictionary and following the heuristic approach presented in Section 4.3.1.1 for computing the feasibility scores. The remainder of the process is described in Section 4.5.2.

c_n	$I(c_n)$	feasibility score
c_1 =cop/p_2	{copy, copies}	0.25
c_2 =cop/p_3	{copy, copies, copied, copying}	0.21
c_3 =copie/p_1	{copie, copies}	0.31
c_4 =copies/p_1	{copies, copiess}	0.23

tree of these new definitions of $p(x)$ and $p(u)$ that take into account the feasibility score when compared with their usual definitions has been empirically observed in the experiments presented in Section 4.5.3. The new definitions reduce the depth of the leaf nodes that represent the candidate stem/paradigm pairs with highest feasibility scores (at the expense of increasing the depth of the leaf nodes that represent the candidate stem/paradigm pairs with lower feasibility score) only when the difference between feasibility scores is relatively high.

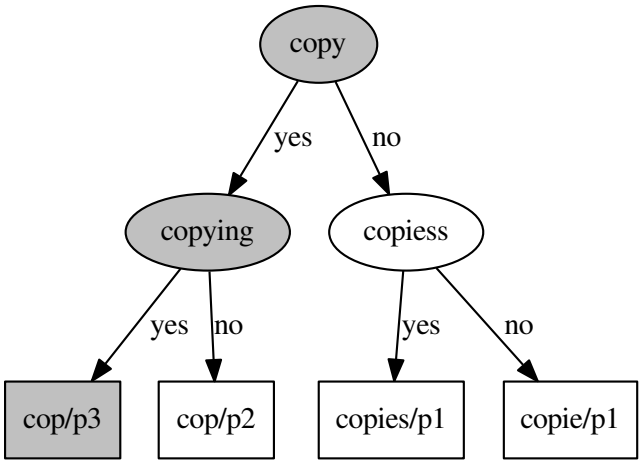
Let us illustrate the process with an example. Consider again the paradigms p_1 = {-ϵ, -*s*}; p_2 = {-*y*, -*ies*}; and p_3 = {-*y*, -*ies*, -*ied*, -*ying*}. If the word form *copies* (from the verb *to copy*) was to be inserted into the monolingual dictionary, the candidate stem/paradigm pairs depicted in Table 4.7 would be obtained. Suppose also that the monolingual corpus used is not reliable and the feasibility scores depicted in the table are obtained. The heuristic querying algorithm would need 3 questions to obtain the final stem/paradigm. First, it would ask the user to validate *copie*, the word form from the highest scored paradigm that causes the discarding of the highest amount of candidates when it is accepted. However, it would be rejected by the user, and only the candidate c_1 would be discarded. Then, the heuristic querying algorithm would choose *copies* by following the same criterion. It would also be rejected and the two remaining candidates would be c_2 and c_4. Finally, the algorithm, still in confirmation mode, would choose *copiess*, which would be rejected.

If the querying algorithm based on decision trees was chosen, the binary decision tree depicted in Figure 4.5 would be obtained and only 2 questions would be needed in order to reach the correct paradigm (the path is shadowed). Given the fact that the feasibility score of the different candidates is similar, all the leave nodes have the same depth.[26]

Now let us assume that the feasibility scores are more accurate, and the correct candidate, c_2, receives a very high feasibility score. The resulting decision tree, together with the new feasibility scores, is depicted in Figure 4.6. In this case, a single query

[26]This property depends on the particular membership relation between the different word forms and the candidate stem/paradigm pairs. The tree becomes less balanced as the difference between the feasibility scores of the different candidates grow.

Figure 4.5: Binary decision tree generated by applying the ID3 algorithm to the candidate stem/paradigm pairs listed in Table 4.7.



is needed in order to reach the correct stem/paradigm pair. The heuristic querying algorithm would need a single query too. This example shows that the querying algorithm based on binary decision trees is more efficient than the heuristic one when the feasibility score is not accurate, but at the same time it is also able to take advantage of an accurate feasibility score as the heuristic algorithm does. This fact is confirmed by the experiments presented in the next section.

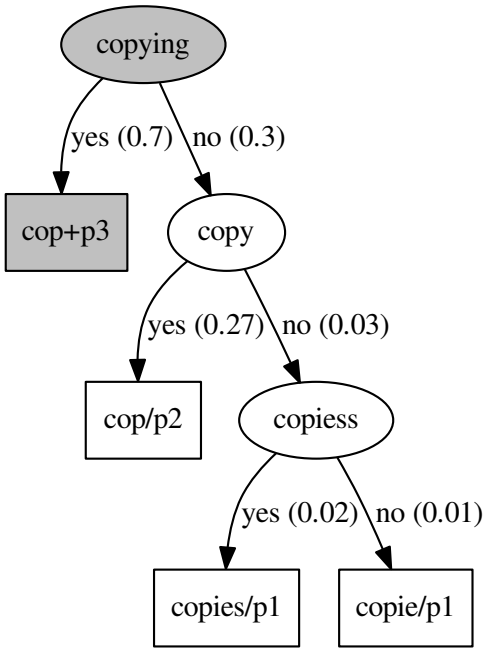
In summary, the use of a binary decision tree built with the ID3 algorithm permits overcoming the limitations of the heuristic feasibility score. The binary decision tree considers both affirmative and negative answers of the user and takes into account the feasibility scores thanks to the proposed modification to $p(x)$ and $p(u)$. Besides, it is more robust to unreliable feasibility scores, as shown in the previous example.

4.5.3 Experimental setup: automatic evaluation

In order to properly assess whether the probabilistic approaches presented in this section help to reduce the amount of questions asked to the users when a new word is inserted into a dictionary, an automatic evaluation has been performed. In this experimental setup, non-expert users, to which this method is eventually addressed, are replaced with an oracle that always chooses the option that leads to the right paradigm. Interferences caused by potential human errors are thus avoided.

The evaluation consisted of simulating the addition of a set of words to the Spanish monolingual dictionary of the English–Spanish Apertium MT system (Forcada et al., 2011). Six different test sets were built: each of them containing a set of word forms to be inserted into the dictionary and a context sentence for each word form. The average number of questions needed in order to obtain the correct paradigm was computed for the following three systems:

Figure 4.6: Binary decision tree generated by applying the ID3 algorithm to the candidate stem/paradigm pairs listed in Table 4.7, assuming that the feasibility scores of each candidate stem/paradigm pair are those depicted in the tree itself.



- the heuristic approaches for paradigm scoring and querying algorithm described in Section 4.3.1.

- the decision-tree-based querying algorithm described above, in which all the candidate stem/paradigm pairs have the same probability, that is, no feasibility score is used. The values initially defined by the ID3 algorithm for the probabilities $p(x)$ in the computation of the entropy $H(S)$ and $p(u)$ in the computation of the information gain $\mathrm{IG}(a, S)$ (see Section 4.5.2) are used (that is, those based on the proportion of elements that belong to each class).

- the decision-tree based querying algorithm described above, in which each candidate stem/paradigm pair is scored with the feasibility score based on hidden Markov models, as described in Section 4.5.2.

In addition to the average number of questions to be asked, the HMM probabilities and heuristic scores were compared by evaluating the success in detecting the correct paradigm, that is, in assigning the highest score or probability to the correct stem/ paradigm candidate.

4.5.3.1 Data set building

In order to build test sets that are as much realistic as possible, the words to be inserted were chosen from different stages of the development of the Apertium Spanish monolingual dictionary of the English–Spanish language pair. The revision history of the dictionary in the Subversion repository of the Apertium project[27] made this approach possible. Given a pair of dictionary revisions (R_1, R_2), being R_1 an earlier revision of R_2, the evaluation task consisted of adding to R_1 some of the entries[28] in R_2 that were not already in R_1. In order to ensure that all the paradigms assigned to these entries were also available in R_1, all the revisions of the dictionary were sequentially checked and grouped according to their paradigm definitions, thus obtaining ranges of *compatible revisions*. Then, the number of new entries added between the oldest and newest revisions of each range was computed, and six revision pairs among those with the greatest number of different entries were manually picked for the experiments.

Recall that a context sentence is needed for each word form to be inserted. They were extracted from the Spanish side of the News Commentary parallel corpus (Bojar et al., 2013) as follows.[29] The corpus was randomly split into two parts: the first one, that contained 90% of the sentences, was used for training the HMM; the second one, that contained the remaining 10%, was used for extracting the context sentences and the word forms of the test sets. Then, for each revision pair (R_1, R_2), the entries added between R_1 and R_2 were expanded in order to obtain the set of new word forms in R_2. For each word form in this set, the sentences in which it occurs were collected and the corresponding pairs of word form and context sentence (from now on, *evaluation pairs*) were added to the test set associated with (R_1, R_2). Note that, if a word form is found in multiple sentences, multiple evaluation pairs with the same word form are added to the test set. Concerning the training of the HMM from the 90% of the corpus, a different HMM was trained for each revision pair as described in Section 4.5.1. In all cases, the Baum–Welch algorithm was stopped after 9 iterations.[30] Table 4.8 shows the list of revision pairs, the number of entries differing between them, the number of evaluation pairs included in the evaluation, and the number of states and observable outputs in the HMM.

[27] `https://svn.code.sf.net/p/apertium/svn/trunk/apertium-en-es/apertium-en-es.es.dix`

[28] Recall that an entry is made by a stem and a paradigm. An entry can generate multiple word forms when it is expanded. Each word form has also morphological inflection information attached, although the morphological information is not included in most of the examples presented in this section for the sake of simplicity.

[29] This corpus was chosen because it belongs to an heterogeneous domain and it is already segmented into sentences. It contains 174 441 sentences and 5 100 982 words.

[30] Before the first iteration, the transition probabilities were uniformly initialised.Similarly, for each state, all the emission probabilities of observable outputs that contain the state in any of its sets of allowed states were also initialised to the same value. Emission probabilities of observable outputs that do not contain the state in any of their sets of allowed states were initialised to 0.

Table 4.8: Revision pairs of the Spanish monolingual dictionary of the Apertium English–Spanish MT system used in the experiments described in Section 4.5.3, number of entries (added between R_1 and R_2), number of evaluation pairs (made of a word form to be inserted and a context sentence), and number of states ($|\Gamma|$) and observable outputs ($|\Sigma|$) in the HMM used for paradigm scoring.

Revision pair			**# evaluation**		
R_1	R_2	**# entries**	**pairs**	$\|\mathbf{\Gamma}\|$	$\|\mathbf{\Sigma}\|$
7217	7287	109	485	4 180	27 487
11762	12415	1802	550	4 391	28 877
17582	20212	700	362	4 403	28 879
27241	27627	1048	297	5 021	28 880
34649	35985	1194	79	5 111	28 881
36838	44118	1039	650	5 111	28 881

Finally, the Spanish Wikipedia dump[31] was used as the monolingual corpus to compute the feasibility scores in the heuristic-based approach in Section 4.3.1. The value of the threshold Θ used to compute the set $\text{Unusual}_{\text{C}}(c_n)$ described in Section 4.3.1 was 0.1, as in the experiments presented in Section 4.3.2.

4.5.4 Results

Table 4.9 shows, for each of the three systems being evaluated, the average number of questions needed to determine the correct paradigm for the word forms in the test set. Since the objective of the probabilistic approaches for paradigm scoring and querying is reducing the amount of questions asked to the users, lower values represent better results. A cell in bold means that the corresponding system either outperforms or underperforms the other two systems by a statistically significant margin ($p \leq 0.05$).[32] If it outperforms them, the value in the cell is marked with the symbol $\uparrow$, whereas if it underperforms them, the value is marked with $\downarrow$.

It can be clearly observed that the system that contains the new probabilistic approaches for the feasibility score and querying algorithm needs fewer questions to reach the solution than the heuristic-based system in all the test sets but one. In addition, the difference in the number of questions asked is statistically significant in all the test sets. Although in the revision pair $(7217, 7287)$ the heuristic system needs fewer questions, the difference between both systems is relatively small.

[31] `http://dumps.wikimedia.org/eswiki/20110114/eswiki-20110114-pages-articles.xml.bz2`

[32] Statistical significance tests were performed with the randomization version of the paired sample t-test described by Yeh (2000), available at `http://www.nlpado.de/~sebastian/software/sigf.shtml`

Table 4.9: Average number of polar questions needed by the three approaches under evaluation (ID3-trained decision tree using HMM probabilities, ID3-trained decision tree in which all the candidates have the same probability, and heuristic-based approach) for each of the test sets. A cell in bold means that the corresponding system either outperforms or underperforms the other two systems by a statistically significant margin ($p \leq 0.05$). If it outperforms them, the value in the cell is marked with the symbol ↑, whereas if it underperforms them, the value is marked with ↓.

Revision pair		**average number of questions**		
R_1	R_2	**ID3+HMM**	**ID3**	**Heuristic**
7217	7287	3.26	**5.50**$^{\downarrow}$	**3.08**$^{\uparrow}$
11762	12415	5.22	5.26	**10.71**$^{\downarrow}$
17582	20212	**4.74**$^{\uparrow}$	**5.65**$^{\downarrow}$	5.18
27241	27627	**4.35**$^{\uparrow}$	5.72	5.85
34649	35985	6.22	6.32	**8.67**$^{\downarrow}$
36838	44118	**5.83**$^{\uparrow}$	6.11	**7.48**$^{\downarrow}$

Table 4.10: Average position of the correct paradigm in the list of candidate stem/paradigm pairs sorted by feasibility score, percentage of evaluation pairs in the test set for which the correct candidate is the first one for the HMM-based and heuristic feasibility scores, and proportion of evaluation pairs in the test set for which none of the word forms generated by expanding the correct stem/paradigm combination can be found in the monolingual corpus used for computing the heuristic feasibility score. A cell in bold means that the corresponding system outperforms the other system by a statistically significant margin ($p \leq 0.05$).

Revision pair		**average position**		**% correct is first**		**% test words with no word form**
R_1	R_2	HMM	Heuristic	HMM	Heuristic	**found in corpus**
7217	7287	1.47	**0.51**	70.31	**72.99**	0.20
11762	12415	**5.66**	10.45	**28.00**	8.36	54.36
17582	20212	1.87	1.72	**52.49**	40.88	0.00
27241	27627	7.11	**4.67**	39.73	**42.76**	9.66
34649	35985	6.66	**5.18**	45.57	45.57	37.79
36838	44118	**1.08**	3.51	**81.08**	70.52	2.10

It is also worth remarking how the decision tree without feasibility scores behaves: it is able to outperform the heuristic system in three out of the six test sets, even though it does not use any kind of feasibility score. It confirms its robustness and proves that a decision tree is more efficient (in terms of the number of questions asked) than the previous heuristic querying algorithm. In addition, results also confirm the successful integration of the feasibility scores computed using an HMM into the decision tree: when feasibility scores are used to compute the probabilities $p(x)$ and $p(u)$ defined in Section 4.5.2, the number of questions asked is reduced in all the test sets (the difference is statistically significant in four of them).

However, the fact that adding an HMM-based feasibility score to the decision tree leads to a reduction in the number of questions asked does not necessarily mean that the HMM-based feasibility score is more accurate than the heuristic one. In order to clarify this issue, Table 4.10 shows the average position of the correct paradigm in the sorted candidate list, as well as the percentage of evaluation pairs in the test set in which the correct paradigm was ranked as the first one for both types of feasibility scores. It also shows the proportion of evaluation pairs in the test set for which none of the word forms generated by expanding the correct stem/paradigm combination can be found in the monolingual corpus used for computing the heuristic feasibility score. The results vary across the different test sets: for some of them the HMM-based feasibility score is more accurate than the heuristic one, but for others it is the other way round.

The test set extracted from the revision pair $(11762, 12415)$ confirms that the HMM-based feasibility score is helpful when the monolingual corpus does not contain enough evidence to compute the heuristic feasibility score. For more than a half of the evaluation pairs in that test set, none of the word forms resulting from the inflection of the correct paradigm and the corresponding stem can be found in the monolingual corpus. As a consequence, the average position of the right stem/paradigm pair in the list L sorted by the heuristic feasibility score is very high. The HMM-based feasibility score does not suffer from this issue, and the position of the right stem/paradigm pair in the list sorted by HMM-based feasibility score is much closer to the first positions. In summary, the HMM-based feasibility score is able to find evidence in situations in which the heuristic one is not able to find it, but that does not mean that it is generally more reliable than the heuristic. The high number of states and observable outputs (see Table 4.8) makes difficult the proper estimation of the HMM parameters and this prevents the HMM approach from achieving better results. Nevertheless, since they extract a different type of information from the monolingual corpus, these results suggest that the two feasibility scores are complementary and could be combined in the future in order to make the most of the available monolingual corpora.

Another interesting conclusion that can be drawn from tables 4.9 and 4.10 is the confirmation of the robustness of the decision tree querying algorithm. In the test sets $(27241, 27627)$ and $(34649, 35985)$, the fully probabilistic system is able to outperform the heuristic-based despite the fact that the heuristic feasibility score is more accurate. When the heuristic feasibility score is less accurate than the HMM-based one, however,

the number of questions asked by the heuristic system grows, as can be observed in the test sets $(11762, 12415)$ and $(36838, 44118)$.

In conclusion, it has been proved that the probabilistic alternatives presented in this section reduce the number of questions that need to be asked to users in order to insert entries into a monolingual dictionary. Using a decision tree instead of the heuristic querying algorithm reduces the number of questions in almost all scenarios, but especially when the feasibility score is not accurate, while the HMM-based feasibility score seems to be complementary with the heuristic one and they could be combined in the future.

4.6 Conclusions

In this chapter, a novel method for allowing non-expert users to insert entries into the morphological dictionaries used in RBMT has been presented. It has been proved that non-expert users are able to successfully validate whether certain word forms are valid forms of the word to be inserted. The presented method creates the corresponding entry in the monolingual dictionary from the answers provided by the users with the help of existing inflection paradigms. For most of the words that the users were asked to add to the dictionary in the evaluation process, the right entry (stem and inflection paradigm) was inserted. Moreover, when the inserted entry was not the right one, it often shared most inflected word forms with the correct one, thus still increasing the coverage of the system. The use of a binary decision tree and an HMM for deciding which word forms need to be validated by the users ensures that the task is performed in a efficient way: only 5–6 questions on average were needed in order to insert a set of words selected from the revision history of a real Apertium monolingual dictionary.

Given that it has been proved that the strategy achieves good results with real users and only a few questions are needed on average to insert an entry thanks to the robustness provided by the decision tree, the strategy presented in this chapter opens the way to the cheap enlargement of dictionaries for RBMT when collaborators who master the particular encoding of the dictionary are not available. Even if they are available, this approach allows them to focus on the development of more complex parts of the system and let users with less experience carry out the task. This approach could be integrated in an online MT system and users could be asked to help inserting the words from the sentences to be translated that cannot be found in the dictionaries of the system. Moreover, users could also be contacted through a crowdsourcing platform (Wang et al., 2013).

It is worth noting that, as it has been stated in Section 4.2.2, the approach presented in this chapter is not able to choose among paradigms that generate the same set of word forms but with different associated morphological information. In the experiments described in Section 4.5.3, for around 87% of words in the test sets, the final

solution contained more than one paradigm with the same word forms. On average, the final solution contained 8.35 paradigms, while the average number of candidate stem/paradigm pairs was 29.77. Therefore, in order to use this approach without the intervention of expert users, either the missing morphological information is elicited by asking different questions to the user, or the paradigm with the most appropriate morphological information is selected in a fully automatic way. Some preliminary work (Sánchez-Cartagena et al., 2012a), to be described in more detail in Section 5.2, on an automatic process based on an n-gram language model of lexical categories and morphological inflection information has already been carried out. That preliminary work also shows that the only difference between the candidate stem/paradigm pairs that share the set of surface forms generated is often lexical category (it is usually restricted to adjective or noun) and gender.[33] This finding opens the door to multiple options for allowing non-expert users to complete the process and select the paradigm with the most appropriate linguistic information. On the one hand, they could be directly asked about the lexical category (adjective or noun) and gender of the word to be inserted, since this concepts are usually known by people with basic education. On the other hand, synthetic sentences in which the word to be inserted acts with different lexical categories or genders could be automatically built and presented to the user for validation. This options is described in more detail in Section 5.2. Nevertheless, the system described in this chapter can also be used out-of-the-box by more experienced developers of linguistic resources. If they choose the most appropriate morphological information at the end of the process, the tool helps to save time by reducing the number of options from which to choose: choosing among 8.35 candidate stem/paradigm pairs is easier and faster than choosing among 29.77.

[33]This happens for more than 63% of the entries in the Spanish monolingual dictionary of the Spanish–Catalan language pair in the Apertium project when trying to insert an inflected word form of each entry already present in the dictionary (the entry was temporarily removed) with the method described in this chapter (Sánchez-Cartagena et al., 2012a).

Chapter 5

Concluding remarks and future work

This chapter summarises the main contributions of this dissertation to the state of the art in machine translation and presents future research lines that may be followed in order to improve the approaches described.

5.1 Summary

In this dissertation, three novel approaches that ease the building of MT systems for language pairs with scarce resources have been presented. Each of them is addressed to the creation of a different type of resource used by MT systems:

- Shallow-transfer rules used by RBMT systems can be automatically inferred from very small parallel corpora by means of the novel approach described in Chapter 2.
- The statistical translation model of SMT systems can be enriched with rules and dictionaries from RBMT by following the novel hybridisation strategy presented in Chapter 3.
- New entries can be added by non-expert users to morphological dictionaries such as those used in RBMT with the new method described in Chapter 4.

The new rule inference approach described in Chapter 2 produces shallow-transfer rules from a small parallel corpus and RBMT dictionaries. They are encoded with a new rule formalism which is an extension of the alignment template formalism (Och and Ney, 2004) and can be converted to the format of a particular RBMT system, as it has been done for evaluating them with the Apertium RBMT platform (Forcada et al., 2011) (see Section A.3.2.1). The method has been evaluated with five

different language pairs and with parallel corpora of different sizes. The evaluation performed shows that the new method outperforms the previous alignment-template-based approach by Sánchez-Martínez and Forcada (2009), which uses a less-expressive formalism and a simpler learning algorithm. In addition, when the languages involved in the translation are closely related (e.g. Spanish↔Catalan), a few hundred parallel sentences (less than 10 000 words) have proved to be sufficient to obtain a set of competitive transfer rules, since the addition of more parallel sentences does not result in great improvements to the translation quality. Experiments also show that, for slightly bigger corpora (hundreds of thousands of words), the new approach reaches, and in some cases surpasses, the translation quality achieved by hand-crafted rules. For instance, the value of the TER score obtained by the automatically inferred rules in the Spanish–English evaluation described in Section 2.6 when the training corpus contains 25 000 sentences is 0.7256, while the hand-crafted rules in Apertium achieve a score of 0.7438.[1]

The algorithm described in Chapter 2 is the first rule inference approach that formalises the rule learning problem as a global minimisation problem that treats conflicts between rules at a global level. This way of approaching the problem allows it to achieve a high degree of generalisation over the linguistic phenomena observed in the training corpus. In addition, unlike previous approaches, the algorithm described in Chapter 2 addresses the problem of input segmentation in shallow-transfer RBMT. In this dissertation, it has been empirically proved that, thanks to these improvements, the rule learning algorithm is able to solve the main limitations of the previous approach by Sánchez-Martínez and Forcada (2009), namely, its inability to find the appropriate generalisation level for the alignment templates and to select the proper subset of alignment templates which ensures an adequate chunking of the input sentences.

The adoption of the rule inference approach presented in Chapter 2 will hopefully contribute towards making the development of transfer rules for new language pairs in MT systems like Apertium a much more cost-effective and technically feasible process. The new rule inference approach can ease the development of working systems in two ways. On the one hand, transfer rules constitute the RBMT linguistic resource that requires the deepest linguistic knowledge. It may therefore be difficult to find bilingual experts who are able to create them for a given language pair specially in the case of less-resourced languages. The new rule inference approach will permit creating the MT system without such experts, given the fact that it has shown to be effective for (five) language pairs in different families. On the other hand, if experts are available, the rule inference algorithm can be used to obtain a set of rules that may be afterwards refined by them. The amount of rules inferred by the new approach is generally one order of magnitude lower than the amount of rules obtained with the approach by Sánchez-Martínez and Forcada (2009), which eases their revision.

[1]The value of TER is inversely proportional to translation quality.

The high generalisation capacity of the new rule inference approach can also contribute to the improvement of MT systems that do not follow the rule-based approach. In particular, in Chapter 3 a new hybrid approach specifically designed for integrating shallow-transfer RBMT rules and dictionaries into a phrase-based SMT system has been presented. In that chapter, it has been proved that the new hybrid approach can be successfully applied to integrate rules inferred from the training corpus with the algorithm described in Chapter 2 and existing dictionaries into an SMT system. Thus, the resulting hybrid system is able to generalise the translation knowledge contained in the parallel corpus to sequences of words that have not been observed in the corpus but share lexical category or morphological inflection information with the words observed.

The experiments performed show that the new hybrid approach outperforms the general-purpose strategy aimed at improving phrase-based SMT models with data from other MT systems developed by Eisele et al. (2008), which is the only hybridisation strategy that can be found in the literature with which shallow-transfer RBMT linguistic resources can be integrated in an SMT architecture. Moreover, the experiments described in Chapter 3 also show that, when the hybrid system is built with automatically inferred rules, it is able to reach the translation quality that would be achieved by a hybrid system built with hand-crafted rules and that the automatically inferred rules often bring an improvement over a hybrid system that only uses dictionaries to enrich the SMT models. For instance, the value of the BLEU score obtained by the hybrid system enriched with automatically inferred rules in the English–Spanish out-of-domain evaluation described in Section 3.4.2 when the training corpus contains 600 000 sentences is 0.2508, while the hybrid system enriched with hand-crafted rules achieves a score of 0.2500 and that enriched only with dictionaries achieves a score of 0.2446. A system built with the hybridisation approach described in Chapter 3 (Sánchez-Cartagena et al., 2011c) and using hand-crafted rules from the Apertium project was one of the winners[2] in the pairwise manual evaluation of the WMT 2011 shared translation task (Callison-Burch et al., 2011) for the Spanish–English language pair.

The main novelty of the hybridisation strategy described in Chapter 3 lies in the fact that it takes advantage of the way in which the RBMT system uses shallow-transfer rules to split the input sentences in order to generate synthetic phrase pairs whose SL and TL phrases are mutual translations. This approach thus avoids the problems with wrong segment alignment suffered by the approach by Eisele et al. (2008). The new approach is also able to find an adequate balance between the probabilities of the phrase pairs extracted from the training corpus and from the synthetic ones.

The combination of the rule learning algorithm described in Chapter 2 and the hybridisation strategy presented in Chapter 3 constitutes a novel way of improving an SMT system with the sole use of morphological dictionaries. In the light of the results obtained in the experiments reported in Chapter 3, it can be concluded that the

[2]No other system was found statistically significantly better using the sign test at $p \leq 0.10$.

combination of both approaches will contribute to alleviate the data sparseness problem suffered by SMT systems when highly inflectional languages are involved, reduce the corpora size requirements for building SMT systems and also enable the creation of general-purpose SMT systems even when the only parallel corpus available belongs to a specialised domain, since the rules inferred are mostly domain-independent.

Regarding morphological dictionaries, which are the third type of MT resource addressed in this dissertation, it is worth pointing out that their availability is a requirement for the approaches described in chapters 2 and 3 to succeed. A parallel corpus is analysed to obtain its RBMT intermediate representations in the SL and in the TL prior to rule inference. In this process, the higher the coverage of the dictionaries, the more bilingual phrases can be used for rule inference. Similarly, in order to apply the inferred rules, either directly in an RBMT system or by means of the generation of synthetic phrase pairs to be integrated in an SMT system, SL words must be analysed previously. In order to facilitate that both approaches can be applied with high-coverage dictionaries, a novel method for allowing non-expert users to insert new entries into monolingual morphological dictionaries has been presented in Chapter 4. The experiments performed for the Spanish language with real users show that they are able to insert entries into the dictionary with a high success rate and that the approach is very efficient: only 5–6 questions on average were needed in order to insert a set of words selected from the revision history of a real Apertium monolingual dictionary. The main contribution of the approach presented in Chapter 4 has been the use of a principled method to make the choice of the words the user has to validate in order to find the most suitable inflection paradigm; more precisely, a binary decision tree built according to probabilities estimated with an HMM has been used.

The main drawback of the method presented in Chapter 4 is that it is not able choose among paradigms that generate the same set of word forms but with different associated morphological information. This situation happens for many words in the experiments carried out and limits the immediate application of the method for the creation of entries in morphological dictionaries only from non-expert users. Multiple solutions to this limitation will be described in Section 5.2. Nevertheless, the approach presented in Chapter 4 can still save costs in the development of dictionaries without further modifications if the most experienced users decide among the paradigms that generate the same set of word forms, while non-expert users carry out the remainder of the work.

Finally, the implementation of all the methods described in this dissertation has been released under the GNU GPL license; the tools released and the instructions for downloading and using them are described in Appendix B. The release of the tools has two main advantages. On the one hand, it ensures the reproducibility of the results presented in this dissertation and makes easier for the scientific community to continue the research, either by following the future research lines described next, or by starting new ones. On the other hand, it permits the effective achievement of the main objective

of the research carried out: easing the construction of MT system for language pairs with scarce resources.

5.2 Future research lines

In the light of the results obtained in the evaluation of the novel approaches that have been presented in this dissertation, some new research lines can be identified. They are listed below.

1. Concerning the new rule inference approach described in Chapter 2, the rule formalism could be enhanced in order to further improve the generalisation power and the translation quality achieved between languages that are not closely related. A new type of GAT that operates on sequences of chunks instead of sequences of words could be automatically inferred too. A chunk is a sequence of lexical forms that have been grouped together. A chunk can have a category and morphological inflection attributes, that are linked to the attributes of the lexical forms that are part of chunk. These new GATs could be converted to the format of the Apertium *interchunk* rules, described in Section A.2. An RBMT system that uses this new type of GAT would operate as follows. In the transfer step, the GATs inferred with the algorithm from Chapter 2 would be applied first, a TL chunk would be created from the sequence of lexical forms generated by each GAT applied, and afterwards the new GATs would be applied to sequences of chunks in order to change the value of their morphological inflection attributes or reorder them. Thus, this new type of rule would improve reordering and agreement between distant words in the sentence. In order to infer these GATs, chunks in the SL and the TL would be identified and aligned in the training parallel corpus before applying an algorithm similar to that described in Section 2.4. Most of the principles of the algorithm from Section 2.4 would be present: generation of multiple rules with different degrees of generalisation, a global minimisation problem, etc. The automatic inference of these new GATs poses multiple research challenges. One of them is the identification of chunks in the TL and their alignment with SL chunks. The information provided by the statistical word alignments in the training parallel corpus could be useful for this purpose: a sequence of TL words that constitutes a TL chunk will usually be aligned only with SL words in the same SL chunk. More sophisticated methods for chunk alignment could be also used, such as the edit-distance-style dynamic programming alignment algorithm by Tinsley et al. (2008), that uses word translation models and the lexical information in the chunks being aligned as sources of knowledge. The definition of chunk categories is another interesting challenge. For instance, a chunk in English that contains a determiner followed by an adjective and a noun and another chunk that contains a determiner followed by a

noun should probably by processed in the same way by the rules because both represent a noun phrase.

2. Moreover, alternative approaches could be considered for some of the steps of the algorithm described in Chapter 2 in order to further improve the results obtained. The word alignment quality (Section 2.4.1) could be improved by integrating symmetrisation in the training of the alignment models as shown by Liang et al. (2006), who have reported a reduction in the alignment error rate with small parallel corpora.

3. Similarly, as regards the optimisation performed to discard rules that cause a deficient chunking of the sentences to be translated, some alternatives to the evaluation metric used to compute the set of key text segments $\mathcal{I}$ (defined in Section 2.4.5) could be considered. In the experiments reported in Section 2.5, BLEU (Papineni et al., 2002) with the smoothing implemented by the National Institute of Standards and Technology (NIST)[3] was the metric employed. Nakov et al. (2012) suggest some improvements to the BLEU smoothing, which are well-suited to sentence-level optimisation. Their impact on the optimisation for chunking could be studied in the future. Another alternative that deserves to be studied consists of scoring the TL sentences generated during the optimisation process described in Section 2.4.5 with a TL model (if a big TL monolingual corpus is available).

4. The optimisation of the thresholds δ and θ used for discarding unreliable GATs, described in Section 2.4.3, is also subject to improvement by studying some alternatives. For instance, their optimum value could be obtained by means of a simplex algorithm (Spendley et al., 1962) rather than following the method described in Section 2.5, that consists of trying all the values in the interval $[0, 1]$ at increments of 0.05 for δ and using θ only to reduce the complexity of the minimisation problem. By using the simplex algorithm, the optimum value could be found after trying fewer values, which would reduce the overall time required by the rule inference algorithm.

5. In Chapter 2, a set of experiments aimed at evaluating the combination of hand-crafted and automatically inferred rules is presented. The results show a degradation of the translation quality achieved by the RBMT system when automatically inferred rules are added to the hand-crafted ones. One possible cause of the degradation observed is the fact that the existing hand-crafted rules have not been taken into account when optimising the automatically inferred rules for chunking. Thus, it is worth exploring the result of optimising both types of rules together. The optimisation process could be modified in order to give a higher priority to the hand-crafted rules. For each sentence, all the segments matching

[3]MTeval utility version 13; `ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13.pl`.

a hand-crafted rule could be added to the set of key segments $\mathcal{I}$. Thus, the automatically inferred rules that usually prevent the application of a hand-crafted rule would be discarded.

6. As it has been previously pointed out, the degree of success of the rule inference algorithm depends on the coverage of the RBMT dictionaries. If a word cannot be analysed, it cannot appear in any bilingual phrase used for rule inference. Moreover, if the unknown word is in a text to be translated with the learned rules, no rule can be applied to it (this problem also occurs when the rules are hand crafted). In order to alleviate the negative effect of a low-coverage dictionary when extracting rules from parallel corpora, some alternatives can be considered. On the one hand, unknown words could be considered as another lexical category. Thus, for instance, an English–Spanish rule that matches the `DT UNKNOWN N` sequence would perform the agreement between the determiner and the noun, even though the word between them (probably an adjective) is unknown. On the other hand, the approach for allowing non-expert users to insert entries into monolingual dictionaries described in Chapter 4 could also help to mitigate the negative impact of unknown words. When analysing the corpus from which to infer the rules, the paradigm with the highest feasibility score could be assigned to unknown words, so that they can be properly analysed. A morphological guesser (see Section 1.3.3.2) could also be used. In any case, the translation according to the bilingual dictionary of the unknown words would still be unavailable.

7. The rule inference algorithm described in Chapter 2 could also be applied when a parallel corpus is not available if a crowdsourcing (Wang et al., 2013) approach is followed. Given an SL monolingual corpus, segments could be extracted from it and provided to the users of the crowdsourcing platform in order to be translated. The rule learning algorithm could then be applied to the resulting bilingual segments. The process could be executed iteratively and the coverage of the existing rules could be used as part of the active learning strategy, that is, as the way to decide which segments will generate the best rules if they are translated by non-expert users and added to the set of bilingual segments from which the rules are inferred. A similar scheme has already been proposed in the context of SMT by Haffari et al. (2009). However, unlike in the approach by Haffari et al. (2009), the segments that the users of the crowdsourcing platform would translate could be sub-sentential units (an appropriate context would have to be displayed in order to help them to translate ambiguous sub-segments).

8. With regard to the hybrid approach described in Chapter 3, which involves the integration in an SMT system of shallow-transfer rules automatically inferred with the algorithm presented in Chapter 2, some alternatives could be studied for a better integration of these rules. In Section 3.4.2, the effect on the hybrid system of the optimisation of rules for chunking is discussed: on the one hand, it prevents incorrect bilingual phrases, which may not be assigned a probability low enough

by the language model, from being taken into account in the decoding process; on the other hand, it may also remove some rules that are able to generate correct and useful bilingual phrases, but that usually prevent other more important rules from being applied in an RBMT system. The optimisation of rules for chunking (described in Section 2.4.5) could be adapted so that it generates rules more suitable for being integrated in an SMT system. For instance, only the rules from sequences of lexical categories that are never (or hardly ever) identified as key segments could be discarded.

9. Another limitation that affects the integration of automatically inferred shallow-transfer rules into SMT is the low scalability of the rule inference algorithm. As it has been pointed out in Section 3.4.1, due to the high complexity of the minimisation process, only a subset of the training corpus that contains 160 000 sentences has been used for rule inference in the experiments carried out in Chapter 3. Even though the improvement in translation quality brought by the automatically inferred rules grows slowly with the size of the portion of the training corpus used for rule inference (see Section 3.4.2), it is worth exploring an alternative option that would permit processing the whole training corpus. It would involve not solving the minimisation problem, but using as a solution all the GATs that correctly reproduce any bilingual phrase used for training (this process is described in Section 2.4.2). When generating and scoring synthetic phrase pairs from the inferred rules (see Section 3.2), the frequency assigned to each synthetic phrase pair could be proportional to the ratio of phrase pairs from the training corpus that are correctly reproduced by the rule from which the synthetic phrase pair has been generated (this ratio is defined as $\frac{\mathcal{Q}(\mathcal{G}(z))}{\mathcal{Q}(\mathcal{M}(z))}$ in Section 2.4.3).

10. Concerning the new approach for allowing non-expert users to insert entries into morphological dictionaries presented in Chapter 4, the experiments reported in Section 4.5.4 show that the feasibility score based on an HMM and the heuristic one that is simply based on the amount of words that are present in a monolingual corpus are complementary. Thus, it would be worth exploring the creation of a new feasibility score that combines both sources of information, in line with the work by Šnajder (2013).

11. The approach presented in Chapter 4 cannot choose among paradigms that generate the same set of inflected word forms, but labelled with different morphological information.[4] One possible solution to this limitation would be automatically determining the most appropriate morphological information using information from a monolingual corpus. To that end, preliminary experiments have been carried out (Sánchez-Cartagena et al., 2012a) with an n-gram model (such as those used in SMT for TL modelling) that assigns probabilities to sequences of

[4]For example, in Spanish many adjectives such as *alto* and nouns such as *gato* are inflected identically. Therefore, two paradigms that produce the same collection of suffixes {*-o* (masculine, singular), *-a* (feminine, singular), *-os* (masculine, plural), *-as* (feminine, plural)} but with different morphological inflection information are defined in the monolingual dictionary.

analysed words (lexical forms without lemma, that is, lexical forms made only of lexical category and morphological inflection information). In order to determine the most appropriate paradigm, a set of sentences that contain any of the word forms resulting from the expansion of the candidate paradigm pairs are collected (recall that all the candidates generate the same set of word forms). Then, the paradigm that encodes the more likely morphological information given the context in which the word forms appear is selected. Experiments showed around 75% success rate on a test set built from the whole Apertium Spanish monolingual dictionary. Note also that at least one sentence with a word form generated by the candidate paradigms will always be available: the sentence that the user was trying to translate with the MT system. The effect on the results of the number of sentences containing the targeted word forms that can be found in the monolingual corpus remains to be studied, as does the use of the feasibility scores provided by the HMM.

12. Another alternative would be asking users more sophisticated questions to allow the system to infer the missing linguistic information. For instance, they could be asked to validate sentences that contain inflected word forms of the word to be inserted into the dictionary. As it has been pointed out in Section 4.6, the difference between the candidate paradigms in Spanish is often lexical category (it is usually restricted to adjective or noun) or gender. Thus, synthetic sentences in which the word to be inserted acts with different lexical categories or genders could be automatically built and presented to the user for validation. For instance, in order to obtain the gender in Spanish, the user would validate sentences that simply contain a determiner (with different genders) followed by the word to be inserted (e.g. if the user wants to insert the word *coche* in Spanish, she would have to validate *el coche* and *la coche*).

13. When the approach in Chapter 4 is used to allow non-expert users to insert entries into monolingual dictionaries of languages different for Spanish, the difference between the candidate paradigms that generate the same set of word forms may comprise linguistic information different from gender and lexical category. Consequently, a more general approach for choosing the sentences to be validated should be defined. It could be performed as follows. For each inflected word form obtained after the user answers the polar questions (e.g. *gato*, *gata*, *gatos*, *gatas*), and for each lexical category–morphological inflection information pair corresponding to each candidate paradigm (e.g. `N-gen:m.num:sg` and `ADJ-gen:m.num:sg` for *gato*[5]), all the sentences in a monolingual corpus that contain a lexical form with that lexical category and morphological inflection information would be collected. For instance, in the sentence *Me gusta este perro* (*I like this dog* in English), *perro* acts as a masculine singular noun (`N-gen:m.num:sg`) and in the sentence *Él está cansado* (*He feels tired* in English)

[5]For the sake of simplicity, in this example the only difference between the candidate paradigms is the lexical category.

cansado acts as a masculine singular adjective (`ADJ-gen:m.num:sg`). Then, for each sentence, the lemma originally found in the sentence would be replaced by the lemma of the word to be inserted, all the lexical forms would be inflected and the user would be asked to validate whether the resulting sentence is correct.[6] In the previous example, the user would have to validate *Me gusta este gato* (*I like this cat*) and *él está gato* (*he feels cat*), and she would probably accept only the first one. The paradigm with the highest rate of sentences accepted would be the chosen one.

14. Finally, the number of states and observable outputs of the HMM used for computing feasibility scores is quite large, which makes it difficult to estimate reliable values for the parameters of the HMM. The effect of disabling the creation of additional states when a paradigm contains suffixes that can lead to the creation of multiple candidate stem/paradigm pairs that share the same paradigm (described in Section 4.5.1) could also be studied. On the one hand, the amount of parameters of the HMM would be drastically reduced but, on the other hand, candidate stem/paradigm pairs that share the same paradigm would have the same feasibility score, which may also increase the number of questions asked.

In summary, in this dissertation three new methods that ease the building of MT systems for language pairs with scarce resources have been presented. The rule inference approach described in Chapter 2 will significantly reduce the effort and time needed to develop RBMT systems. Its impact will be especially remarkable for less-resourced language pairs, since the bilingual experts who usually create the rules of RBMT systems, who may be difficult to find for this kind of language pairs, will not be needed. The hybridisation strategy described in Chapter 3, in combination with the rule inference approach from Chapter 2, will contribute towards a better use of the linguistic resources available for a language pair. It will also alleviate the data sparseness suffered by SMT systems when dealing with highly inflected languages, while keeping other advantages of SMT, such as the good lexical selection and the fluency of the output. Thus, the size of the parallel corpus needed to obtain a competitive SMT system will be reduced. Finally, the approach for allowing non-expert users to insert entries into morphological dictionaries described in Chapter 4 will also speed up the creation of RBMT resources and facilitate the application of the approaches described in chapters 2 and 3.

[6] Actually, the users should state whether the sentence is grammatically correct or not. They could be instructed about what *grammatically correct* means with an example (e.g. *the wireless boy* is grammatically correct although it makes no sense) or they could be simply asked to validate whether the sentence is correct, without further explanation, and an n-gram language model (operating either on lemmas or on surface forms) could be used to discard sentences that, after inserting the lemma of the word to be inserted, are semantically incorrect.

Appendix A

Apertium: an open-source shallow-transfer machine translation platform

In this appendix, the Apertium free/open-source rule-based machine translation platform is briefly described. It is the platform in which the new methods presented in this dissertation have been evaluated. Apertium is being widely used to build machine translation systems for a variety of language pairs, especially in those cases where shallow transfer suffices to produce good-quality translations (mainly with related-language pairs). This appendix describes the translation engine and the encoding of linguistic data, including the method followed for converting the generalised alignment templates generated by the rule inference approach described in Chapter 2 into Apertium shallow-transfer rules.[1]

A.1 Introduction

The Apertium free/open-source machine translation (MT) platform comprises an engine, a toolbox, and data to build rule-based MT systems. The platform was initially aimed at related-language pairs (such as Spanish–Portuguese) but it was expanded later to deal with more divergent pairs (such as English–Spanish). Apertium uses finite-state transducers (Roche and Schabes, 1997) for lexical processing, hidden Markov models for part-of-speech tagging (Cutting et al., 1992), and multi-stage finite-state chunking for structural transfer.

Apertium may be used to build MT systems for a variety of language pairs; to that end, the platform uses simple, standard formats to encode the linguistic data needed,

[1]This appendix is largely based on a paper by Forcada et al. (2011).

and documented procedures[2] to build those data and to train the necessary modules. Apertium is licensed under the GNU General Public License[3] (GNU GPL) and can be downloaded from the project's website: `http://www.apertium.org`.

The MT engine and tools in Apertium were not built from scratch, but are rather the result of a complete rewriting and extension of two previous MT systems, namely the Spanish–Catalan MT system `interNOSTRUM.com` (Canals-Marote et al., 2001) and the Spanish–Portuguese MT system `traductor.universia.net` (Garrido-Alenda et al., 2004), both developed by the Transducens group at Universitat d'Alacant. The first version of the whole system (Apertium level 1) was released on July 29, 2005, and closely followed the architecture of those two non-free systems. An enhanced version of the engine (Apertium level 2) was released on December 22, 2006, featuring an extended implementation of the structural transfer of Apertium level 1 to perform more complex transformations for the translation between less-related language pairs.

The remainder of this appendix is organised as follows: Section A.2 describes the different modules of the Apertium architecture, while in Section A.3 the encoding of linguistic data in Apertium is explained. Section A.3 also describes how the generalised alignment templates (GATs) generated by the rule inference approach presented in Chapter 2 are encoded as Apertium shallow-transfer rules.

A.2 The Apertium MT architecture

Apertium is a classical shallow-transfer or transformer system consisting of a 10-module Unix-style *pipeline* or *assembly line*. To ease diagnosis and independent testing, modules communicate between themselves using text streams. This allows for some of the modules to be used in isolation, independently from the rest of the MT system, for other natural-language processing tasks, or for research purposes. A description of each module in the pipeline in given below. Afterwards, two translation examples are provided.

A.2.1 Modules in the Apertium pipeline

The Apertium pipeline contains the following modules (see Fig. A.1):

- A *deformatter* which encapsulates the format information in the input as *superblanks* that will then be seen as blanks between words by the rest of the modules.

[2] `http://wiki.apertium.org/wiki/Documentation`
[3] `http://www.gnu.org/licenses/#GPL`

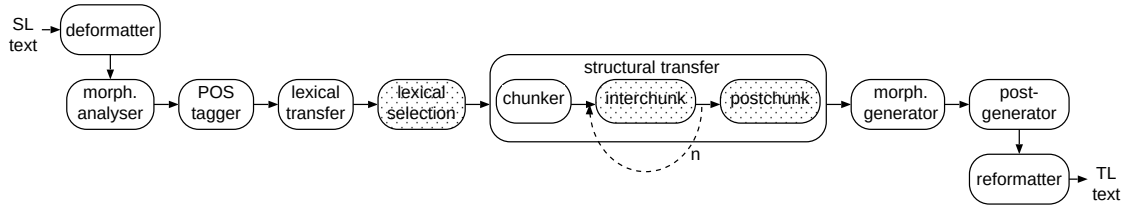


Figure A.1: The Apertium architecture. Shadowed modules are optional and intended for less-related pairs. Apertium level 2 allows for an arbitrary number of interchunk modules.

- A *morphological analyser* which segments the text in surface forms (*words*, or, where detected, multi-word lexical units) and delivers, for each of them, one or more *lexical forms* consisting of lemma, lexical category and morphological inflection information. It reads a finite-state transducer compiled from an SL morphological dictionary in XML.

- A *statistical part-of-speech tagger* which chooses, using a first-order hidden Markov model (Cutting et al., 1992), the most likely lexical form corresponding to an ambiguous surface form.

- A *lexical transfer* module which reads each SL lexical form and delivers the corresponding TL lexical form by looking it up in a bilingual dictionary encoded as a finite-state transducer compiled from the corresponding XML file. Multiple TL translations for a single SL lexical forms may be encoded in the bilingual dictionary. In that case, a *lexical selection* module selects the most appropriate translation (Tyers et al., 2012) given the context. The lexical selection is carried out by a set of rules also encoded as a finite-state transducer and compiled from an XML file.

- A *structural transfer* module which consists of three sub-modules:

  - A mandatory *chunker* which performs local syntactic operations and segments the sequence of lexical units into chunks. A *chunk* is defined as a fixed-length sequence of lexical categories that corresponds to some syntactic feature such as a noun phrase or a prepositional phrase.
  - An optional *interchunk* module which performs longer-range operations with the chunks and between them. More than one interchunk module can be used in sequence to perform increasingly higher-level transfer transformations.
  - An optional *postchunk* module which performs finishing operations on each chunk and removes chunk encapsulations so that a plain sequence of lexical forms is generated.

  Some language pairs use only the first sub-module (chunker), which is equivalent to Apertium level 1, while others use one or more interchunk submodules and

an additional postchunk submodule (Apertium level 2). Nevertheless, it is worth pointing out that the structural transfer module does not rely on a full parse tree of the whole sentence. Even if it is possible for a processed pattern to leave information for later patterns, which can be used for left-to-right long-range agreement processes, RBMT systems that perform a full syntactic analysis are more effective than Apertium when dealing with other phenomena such as long-range reorderings. All of these modules are compiled from XML files containing rules. Patterns are applied in a left-to-right, longest match way.

- A *morphological generator* which delivers a TL surface form for each TL lexical form, by suitably inflecting it. It reads an finite-state transducer compiled from a TL morphological dictionary in XML.

- A *post-generator* which performs orthographic operations, such as contractions (e.g. Spanish *a + el = al* or Portuguese *por + as = pelas*), apostrophations (e.g. Catalan *el + institut = l'institut*) or epenthesis (e.g. English *a + institute = an institute*), using a finite-state transducer generated from a rule file written in XML.

- A *reformatter* which de-encapsulates any format information.

Two examples of the outputs of each module for different input texts and language pairs are given below. As the modules interchunk and postchunk are optional, firstly a translation example for a language pair in which they are present (Apertium level 2) is discussed, and afterwards an example for a language pair that do not use them (Apertium level 1) is also presented. In the first translation example, the lexical selection module is also executed. The transfer rules generated by the approach described in Chapter 2 are designed to be used in an Apertium level 1 system (see Section A.3.2 for more details). The Spanish↔Catalan Apertium-based system used in the experiments reported in Chapter 2 follows the Apertium level 1 pattern, while the English↔Spanish and Breton–French systems used in chapters 2 and 3 follow the Apertium level 2 model.

A.2.2 Apertium level 2 example: English–Spanish

Table A.1 shows the output of each module in the Apertium pipeline (see Fig. A.1) when translating one sentence written in HTML from English to Spanish with Apertium level 2. First, the deformatter encapsulates format information (in this case, HTML tags) in square brackets, so that the rest of the modules treat it as simple blanks between words. Then, the morphological analyser delivers one lexical form for each of the unambiguous input surface forms, and two or more for the surface forms that, according to the English monolingual dictionary, may be assigned different lexical categories or morphological inflection information (*will* can be a noun or an auxiliary verb; *go* can be a verb in infinitive or in present tense; *park* can be a noun or a verb, in infinitive or in present tense); the rest of the words are tagged as subject pronoun (*we*), preposition

Table A.1: An example of step-by-step execution of Apertium when translating the HTML text "We will go to the old park" into Spanish. The output of each module becomes the input of the next one (see text for details).

Module	Output
Deformatter	`We will go to the[ <b>]old park[</b>]`

Morph. analyser	`^We/Prpers<prn><subj><p1><mf><pl>$`
	`^will/will<n><sg>/will<vaux><inf>$`
	`^go/go<vblex><inf>/go<vblex><pres>$ ^to/to<pr>$`
	`^the/the<det><def><sp>$[ <b>]^old/old<adj><sint>$`
	`^park/park<n><sg>/park<vblex><inf>/`
	`park<vblex><pres>$[</b>]`

PoS tagger	`^Prpers<prn><subj><p1><mf><pl>$ ^will<vaux><inf>$`
	`^go<vblex><inf>$ ^to<pr>$ ^the<det><def><sp>$`
	`[ <b>]^old<adj><sint>$ ^park<n><sg>$[</b>]`

Lex. transfer	`^Prpers<prn><subj><p1><mf><pl>/Prpers<prn><tn><p1><GD><pl>$`
	`^will<vaux><inf>/ser<vaux><inf>$`
	`^go<vblex><inf>/ir<vblex><inf>$ ^to<pr>/a<pr>$`
	`^the<det><def><sp>/el<det><def><GD><ND>$[ <b>]`
	`^old<adj><sint>/viejo<adj><GD><ND>/anciano<adj><GD><ND>$`
	`^park<n><sg>/parque<n><><sg>$[</b>]`

Lex. selection	`^Prpers<prn><subj><p1><mf><pl>/Prpers<prn><tn><p1><GD><pl>$`
	`^will<vaux><inf>/ser<vaux><inf>$`
	`^go<vblex><inf>/ir<vblex><inf>$ ^to<pr>/a<pr>$`
	`^the<det><def><sp>/el<det><def><GD><ND>$[ <b>]`
	`^old<adj><sint>/viejo<adj><GD><ND>$`
	`^park<n><sg>/parque<n><><sg>$[</b>]`

Chunker	`^Prnsubj<SN><tn><p1><GD><pl>{^prpers<prn><2><p1><4><pl>$}$`
	`^verbcj<SV><vblex><fti><PD><ND>{^ir<vblex><3><4><5>$}$`
	`^pr<PREP>{^a<pr>$}$`
	`^det_nom_adj<SN><DET><m><sg>{^el<det><def><3><4>$`
	`[ <b>]^parque<n><3><4>$ ^viejo<adj><3><4>$}$[</b>]`

Interchunk	`^Verbcj<SV><vblex><fti><p1><pl>{^ir<vblex><3><4><5>$}$`
	`^pr<PREP>{^a<pr>$}$`
	`^det_nom_adj<SN><DET><m><sg>{^el<det><def><3><4>$`
	`[ <b>]^parque<n><3><4>$ ^viejo<adj><3><4>$}$[</b>]`

Postchunk	`^Ir<vblex><fti><p1><pl>$ ^a<pr>$ ^el<det><def><m><sg>$` `[ <b>]^parque<n><m><sg>$ ^viejo<adj><m><sg>$[</b>]`
Morph. generator	`Iremos ~a el[ <b>]parque viejo[</b>]`
Postgenerator	`Iremos al[ <b>]parque viejo[</b>]`
Reformatter	`Iremos al <b>parque viejo</b>`

(*to*), definite determiner singular/plural (*the*), and synthetic adjective (*old*).[4] The characters "^" and "$" delimit the analysis for each surface form, and the different lexical forms for each surface form are separated by "/". The string after the "^" and before the first "/" is the surface form as it appears in the input text; the string before each group of lexical labels is the lemma. In the next step, the ambiguous words are correctly tagged by the part-of-speech tagger.

The lexical transfer module delivers one or more lexical forms in Spanish for each lexical form in English. The lexical form in English and its corresponding lexical forms in Spanish are separated by "/". Note that the word *old* has two possible translations into Spanish according to the bilingual dictionary used by the lexical transfer module: *viejo* and *anciano*. The lexical selection module decides that *viejo* is the most suitable one. The labels `GD` and `ND` in Spanish lexical forms (meaning *gender to be determined* and *number to be determined*) indicate that there was not enough information at word level to determine this grammatical information.

The chunker detects patterns of words, creating four chunks in this case. It executes the local actions programmed for each detected pattern, which can imply local reorderings, deletion or insertion of words. Here, the chunk labelled `verbcj` is generated for the detected sequence *auxiliary verb–verb* (*will go*), and it contains only one lexical form, the Spanish verb *ir*; the auxiliary is used to determine the value `fti` (future) of the verb chunk. The sequence *determiner–adjective–noun* (*the old park*) is labelled `det_nom_adj`, and the adjective is moved after the noun. Two other chunks are generated, one for the pronoun (labelled `Prnsubj`) and another for the preposition (labelled `pr`). The lexical forms belonging to each chunk are enclosed between curly brackets, and the labels outside correspond to the lexical information from the head of the TL chunks (for example, the noun in the noun phrase) or, in the absence of this information, from some of the other constituents in order of *importance*. Note that the labels with numbers link the grammatical information of elements inside the chunk to that of elements outside the chunk. This is how the postchunk module will be able to determine later that *el* and *viejo* must be assigned the tags `m` (masculine) and `sg` (singular) to match the gender and number of the noun *parque*, or that the verb *ir* must be assigned the future tense (`fti`). Note also that, as happened in the lexical transfer step, the labels `GD`, `PD` and `ND` in the first and second chunks (`PD` means *person to be determined*) indicate that there was not enough information at chunk level to determine this grammatical information, so that the task is passed on to the next module, where operations between chunks can be performed.

The interchunk module detects the sequence `Prnsubj`–`verbcj` and uses the grammatical information of the pronoun chunk to assign person and number to the verb chunk, so that `PD` is now *first person* (`p1`) and `ND` is now *plural* (`pl`). It also deletes the pronoun chunk.

[4] *Synthetic* adjectives, such as *old*, are inflected for comparison by adding a morpheme (i.e. *old*, *older*, *oldest*) in opposition to analytic adjectives, e.g. *expensive*, that are not inflected.

Table A.2: An example of step-by-step execution of Apertium when translating from Spanish to Catalan the HTML text "vi una señal" (that means "I saw a signal" in English). The output of each module becomes the input of the next one (see text for details).

Module	Output
Deformatter	`vi[ <b>]una señal[</b>]`

Morph. analyser	`^vi/ver<vblex><ifi><1><sg>$[ <b>]`
	`una/un<det><ind><f><sg>/unir<vblex><prs><1><sg>/`
	`unir<vblex><prs><3><sg>$ ^señal/señal<n><f><sg>$[</b>]`

PoS tagger	`^ver<vblex><ifi><1><sg>$[ <b>]^un<det><ind><f><sg>$` `^señal<n><f><sg>$[</b>]`

Lex. transfer	`^ver<vblex><ifi><1><sg>/veure<vblex><ifi><1><sg>$[ <b>]`
	`^un<det><ind><f><sg>/un<det><ind><f><sg>$`
	`^señal<n><f><sg>/senyal<n><m><sg>$[</b>]`

Chunker (transfer)	`^anar<vbaux><pres><1><pl>$ ^veure<vblex><inf>$[ <b>]` `^un<det><ind><m><sg>$ ^senyal<n><m><sg>$[</b>]`
Morph. generator	`vaig veure[ <b>]un senyal[</b>]`
Postgenerator	`vaig veure[ <b>]un senyal[</b>]`
Reformatter	`vaig veure <b>un senyal</b>`

In the generation phase, the morphological generator delivers a TL surface form for each TL lexical form by looking them up in the Spanish monolingual dictionary. After that, the postgenerator performs the contraction of *a+el* into *al*. Finally, the reformatter restores the format information (HTML tags) into the translated text.

A.2.3 Apertium level 1 example: Spanish–Catalan

Table A.2 shows the output of each module in the Apertium pipeline (see Fig. A.1) when translating one sentence written in HTML from Spanish to Catalan with Apertium level 1. The actions carried out by the deformatter, morphological analyser and part-of-speech taggers are similar to those described in the previous example. In this example, the first and third surface forms are unambiguous: *vi* can only be a verb and *señal* can only be a noun. The second one, however, can be a determiner or a verb (either first-person or third-person). The part-of-speech tagger decides that it is a determiner.

Regarding the transfer step, in Apertium level 1 the chunker detects chunks, executes local actions within the lexical forms of each chunk independently, and produces a sequence of lexical forms (the postchunk module is therefore not needed). In the running example, a *determiner-noun* rule is used to change the gender of the determiner so that it agrees with the noun, and another rule introduces the auxiliary Catalan verb *anar* and changes the tense of the verb Catalan *veure* to infinitive. As a result, a sequence of four lexical forms is obtained. Finally, the morphological generator delivers

one surface form for each lexical form and the postgenerator does not perform any modification.

A.3 Formats for linguistic data

As has been already pointed out, the formats used by this architecture are declarative and based on XML[5] for interoperability; in particular, for easier parsing, transformation, and maintenance. Moreover, the use of well-defined XML formats allow third-party tools to automatically generate data, such as bilingual dictionaries or transfer rules, to be used by the translation engine. The XML formats for each type of linguistic data are defined through conveniently-designed XML document-type definitions.

A.3.1 Dictionaries

Apertium uses monolingual morphological dictionaries, bilingual dictionaries and postgeneration dictionaries for lexical processing. They use a common XML format.

Morphological dictionaries establish the correspondences between surface forms and lexical forms and contain: (a) a definition of the alphabet (used by the tokenizer), (b) a section defining the grammatical symbols used in a particular application to specify lexical forms (symbols representing concepts such as *noun*, *verb*, *plural*, *present*, *feminine*, etc.), (c) a section defining paradigms, and (d) one or more labelled dictionary sections containing lists of surface form–lexical form correspondences for whole lexical units, usually defined by referencing paradigms.

A small example follows to show how a simple entry is encoded in a XML monolingual dictionary (the definitions of the alphabet and grammatical symbols are omitted for the sake of simplicity). A paradigm named `par123` to be used in English nouns with singular ending in *-um* which change it to *-a* to form the plural form will be defined as follows:

```
<pardef n="par123">
 <e><p>  <l>um</l>  <r>um<s n="n"/><s n="sg"/></r>  </p></e>
 <e><p>  <l>a</l>   <r>um<s n="n"/><s n="pl"/></r>  </p></e>
</pardef>
```

Now, the words *baterium*/*bacteria* and *datum*/*data* will be defined as follows:

```
<e lm="bacterium"><i>bacteri</i><par n="par123"/></e>
<e lm="datum"><i>dat</i><par n="par123"/></e>
```

[5] `http://www.w3.org/XML/`

The part inside the `i` element contains the prefix of the word that is common to all inflected forms (lemma), and the element `par` refers to the inflection paradigm of the word. In this case, *bacterium* will be analysed into `bacterium<n><sg>` and *bacteria* into `bacterium<n><pl>`.

It is also possible to create entries consisting of two or more words if these words are considered to build a single *translation unit*. Dictionaries may also contain *nested paradigms* used in other paradigms (for instance, paradigms for enclitic pronoun combinations are included in all Spanish verb paradigms).

Along with morphological analysers, Apertium also has a number of bilingual lexica. These are encoded in the same XML-based format used by the morphological analysers, but represent correspondences between lemmata, including multi-word lexical units. Each bilingual correspondence is an entry in the dictionary, where lemma and lexical category are specified and, in some cases, morphological inflection information is also included (e.g. to specify changes in the inflection information from SL to TL, and also to mark some ambiguities that should be solved by the structural transfer module). The example below shows how the bilingual correspondence between the English noun *datum* and the Spanish noun *dato* is encoded in a bilingual dictionary entry. Note that the tag `m` is included in order to indicate that the Spanish noun is masculine.

```
<e><p><l>datum<s n="n"/></l><r>dato<s n="n"/><s n="m"/></r></p></e>
```

Finally, post-generation dictionaries are used to establish correspondences between input and output strings corresponding to the orthographical transformations to be performed by the post-generator on the TL surface forms generated by the generator.

A.3.2 Structural transfer rules

Structural transfer rule files contain pattern–action rules which describe what has to be done for each pattern (much like in languages such as `perl` or `lex`). The `pattern` section of a chunker rule is used to specify the lexical category, lemma, and morphological inflection attributes of the lexical forms to be matched: lemma and morphological inflection attributes are optional. The instructions working with the matched lexical forms are placed in the `action` section of the rule. Apertium provides instructions that permit access to the SL lexical forms matched by the rule and the translation provided for them in the bilingual dictionary. There are also instructions that permit the TL lexical forms to be built by assembling the aforementioned elements. Some flow control structures (mainly, loops and conditionals) are also allowed. Interchunk rules work in a similar way, but the elements they process are whole chunks instead of lexical forms. As the rule inference algorithm described in Chapter 2 generates rules encoded as GATs, they need to be converted to the Apertium XML format in order to be integrated in an Apertium RBMT system. The conversion process is described below.

A.3.2.1 Encoding generalised alignment templates as Apertium structural transfer rules

The GATs produced by the rule inference algorithm described in Chapter 2 are converted into rules to be executed by the Apertium chunker module. Thus, an Apertium level 1 system can be created without manually writing a single transfer rule. The automatic inference of interchunk rules is a research line that may be tackled in the future, as explained in Section 5.2.

The set of GATs obtained are converted into rules by grouping those GATs that match the same sequence of lexical categories under the same rule. Each rule detects the corresponding sequence of lexical categories in its `pattern` section (regardless of the lemma and morphological inflection attributes). GATs are then included in the `action` section in decreasing order of specificity (see Section 2.4.6) signifying that the most specific GAT is always applied when more than one GAT can be applied to the sequence of lexical forms matched by the rule. For each GAT, the body of the rule checks whether the lemmas, morphological inflection attributes and restrictions of the sequence of SL lexical forms matched by the rule are compatible with the GAT, and if they are then the GAT is applied and the execution of the rule ends. If after checking all the GATs in a rule, none of them can be applied, the engine attempts to apply a shorter rule to the input text.[6]

The following example illustrates how GATs for the translation of a sequence of two verbs from Catalan to Spanish (like that shown in Figure 2.13 on page 63) are encoded as an Apertium rule. Figure A.2 shows the `pattern` section that matches a sequence of two verbs. The action section of the rule consists of several GATs; Figure A.3 shows the fragment of the `action` section that corresponds to the GAT in Figure 2.13. The XML tags `choose`, `when`, `test` and `otherwise` work as the `switch` instruction in many programming languages. The first `equal` instruction checks whether the lemma of the first verb is *anar*, the following two instructions ensure that the tense of the first SL verb is *past* and that the result obtained after looking it up in the bilingual dictionary is also in the past tense too (restriction). The two remaining `equal` instructions apply the same verification to the infinitive mood of the second SL verb. If the five tests are passed, one lexical form is generated (defined by the `lu` tag inside the `out` element). Its lemma is obtained by looking up in the bilingual dictionary the second lexical form matched by the rule (first `clip` tag). The lexical categories and the first morphological inflection attribute (verb tense) are explicitly defined with the tag `lit-tag` and the values of the other two morphological inflection attributes (person and number) are obtained using the `clip` tag with the `side` attribute set to "`tl`" (TL references). The `reject-current-rule` instruction discards the rule and attempts to apply other (shorter) rule to the input sequence; it is executed only when none of the GATs in the rule can be applied.

[6]This behaviour differs from the standard behaviour in Apertium. To implement it, the Apertium engine has been modified in order to add support for the cancellation of the execution of a rule.

```
<section-def-cats>
  <def-cat n="CAT_VERB">
    <cat-item tags="VERB.*"/>
  </def-cat>
  ...
</section-def-cats>
...
<section-rules>
<rule>
  <pattern>
    <pattern-item n="CAT_VERB"/>
    <pattern-item n="CAT_VERB"/>
  </pattern>
  <action>
    ...
  </action>
</rule>
...
</section-rules>
```

Figure A.2: Header (`pattern` section) of an Apertium shallow-transfer rule containing GATs for the translation of the Catalan verb *anar* in the past tense followed by a verb in infinitive mood into Spanish. The `section-def-cats` section is used to define identifiers for the patterns to be matched by the rules.

```
<action>
  <choose>
    ...
    <when>
      <test><and>
        <equal>
          <clip pos="1" side="sl" part="lemma" />
          <lit v="anar"/>
        </equal>
        <equal>
          <clip pos="1" side="sl" part="tense" />
          <lit-tag v="past"/>
        </equal>
        <equal>
          <clip pos="1" side="tl" part="tense" />
          <lit-tag v="past"/>
        </equal>
        <equal>
          <clip pos="2" side="sl" part="tense" />
          <lit-tag v="inf"/>
        </equal>
        <equal>
          <clip pos="2" side="tl" part="tense" />
          <lit-tag v="inf"/>
        </equal>
      </and></test>
      <out>
         <lu><clip pos="2" side="tl" part="lemma"/><lit-tag v="verb.
            past"/><clip pos="2" side="tl" part="person"/><clip pos="
            2" side="tl" part="number"/></lu>
      </out>
    </when>
        ...
    <otherwise>
      <reject-current-rule shifting="no" />
    </otherwise>
  </choose>
</action>
```

Figure A.3: Fragment of the `action` section of an Apertium shallow-transfer rule encoding the structural transformation provided by the GAT shown in Figure 2.13 (see page 63) for the translation of the Catalan verb *anar* in the past tense followed by a verb in infinitive mood into Spanish. The `pattern` section of the rule is shown in Figure A.2.

Appendix B

Open-source software released as part of this thesis

All the methods and techniques described in this thesis have been released under open-source licenses in order to ensure the reproducibility of all the experiments conducted, and to allow other researchers to use and improve them. This appendix briefly overviews the open-source software released and relates each software package with the experiments conducted in each chapter.

B.1 `apertium-transfer-tools v.2.0`

The algorithm for inferring shallow-transfer rules from a small parallel corpus and existing RBMT dictionaries presented in Chapter 2 is implemented in the software package `apertium-transfer-tools`, which can be downloaded from the Apertium subversion repository at `http://sourceforge.net/projects/apertium/`.[1] It has been released under the GNU GPL v3 free software license. It is meant to replace a previous version of the `apertium-transfer-tools` package initially released by Sánchez-Martínez and Forcada (2009). The current implementation relies on the `Giza++` package for computing word alignments and on the python library `PuLP` and the Cbc solver[2] in order to solve the minimisation problem defined in Section 2.4.4.1 with integer linear programming.

The `apertium-transfer-tools` tool produces shallow-transfer rules encoded in the Apertium XML format (described in Appendix A). A system-independent list of generalised alignment templates (see Section 2.3) is also generated by the tool. From this list, rules encoded in the particular format of any other shallow-transfer RBMT system could be easily generated.

[1] `svn://svn.code.sf.net/p/apertium/svn/trunk/apertium-transfer-tools`
[2] `https://projects.coin-or.org/Cbc`

B.2 rule2Phrase

The method described in Chapter 3 for enriching a phrase-based SMT system with linguistic resources from RBMT is implemented in the software tool `rule2Phrase` (Sánchez-Cartagena et al., 2012b), released under the GNU GPL v3 free software license. It can be freely downloaded from `http://www.dlsi.ua.es/~vmsanchez/Rule2Phrase.tar.gz`. The tool enriches phrase-based SMT models built with the Moses toolkit (Koehn et al., 2007) with linguistic data from the Apertium RBMT platform (Forcada et al., 2011).

As has been explained in Chapter 3, the hybridisation strategy involves the generation of a set of synthetic phrase pairs and their integration in the MST phrase table. In order to generate the synthetic phrase pairs, the SL text to be translated with the hybrid system must be processed previously with the `--extract-n-grams` option of the tool, and afterwards the generation of the synthetic phrase pairs can be performed by executing `rule2Phrase` with the `--gen-phrases` option.

Concerning the integration of the synthetic phrases in the SMT system, the strategies described in sections 3.2.2.3 and 3.2.2.4 are implemented in the `rule2Phrase` tool. For both hybridisation strategies, when the tool is executed with the `--buildSMT` option, it runs the whole phrase-based SMT training and tuning process and produces enriched SMT models that are ready to use with the Moses decoder.

B.3 apertium-dixtools

An implementation of the approach for allowing non-expert users to insert entries in monolingual morphological dictionaries described in Chapter 4 has been added to the package `apertium-dixtools`, which is an existing toolbox that allows Apertium developers to perform multiple operations with dictionaries, such as sorting their entries, formatting XML files or creating new bilingual dictionaries by combining existing ones. It has been released under the GNU GPL v2 free software license and can be downloaded from the Apertium subversion repository at `http://sourceforge.net/projects/apertium/`.[3]

In order to insert new entries in a monolingual dictionary by answering polar questions, `apertium-dixtools` must be executed with the `guessparadigm` option and the monolingual dictionary, a monolingual corpus and a file with the word forms to be inserted must be provided as command line arguments. The tool currently implements the heuristic approaches for the feasibility score and querying algorithm described in Section 4.3.1. After being invoked, the tool asks the polar questions through a command-line interface, and prints the chosen stem and paradigm for each word form to be inserted.

[3] `svn://svn.code.sf.net/p/apertium/svn/trunk/apertium-dixtools`

Index of abbreviations

Index of frequently used symbols

List of Figures

List of Tables

Bibliography

Alcázar, A. (2005). Consumer Corpus: Towards linguistically searchable text. In *Proceedings of BIDE (Bilbao-Deusto) Summer School of Linguistics 2005*, Bilbao, Spain.

Ambati, V., Vogel, S., and Carbonell, J. (2010). Active learning and crowd-sourcing for machine translation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 2169–2174, Valletta, Malta.

Arnold, D. (2003). Why translation is difficult for computers. In *Computers and Translation: A translator's guide*. Benjamins Translation Library.

Arnold, D., Balkan, L., Meijer, S., Humphreys, R., and Sadler, L. (1993). *Machine Translation: an Introductory Guide*. Blackwells-NCC, London.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, USA.

Bangalore, B., Bordel, G., and Riccardi, G. (2001). Computing consensus translation from multiple machine translation systems. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 351–354.

Bartůšková, D. and Sedláček, R. (2002). Tools for Semi-automatic Assignment of Czech Nouns to Declination Patterns. In Sojka, P., Kopeček, I., and Pala, K., editors, *Text, Speech and Dialogue*, volume 2448 of *Lecture Notes in Computer Science*, pages 159–162.

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process. *Inequalities*, 3:1–8.

Bisazza, A., Ruiz, N., and Federico, M. (2011). Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In *International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, USA.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on

Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.

Bojar, O. and Hajič, J. (2008). Phrase-based and deep syntactic English-to-Czech statistical machine translation. In *Proceedings of the third Workshop on Statistical Machine translation*, pages 143–146, Columbus, Ohio, USA.

Brandt, M. D., Loftsson, H., Sigurpórsson, H., and Tyers, F. M. (2011). Apertium-IceNLP: a rule-based Icelandic to English machine translation system. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 217–224, Leuven, Belgium.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Goldsmith, M. J., Hajic, J., Mercer, R. L., and Mohanty, S. (1993a). But dictionaries are data too. In *Proceedings of the Workshop on Human Language Technology*, pages 202–205, Princeton, New Jersey.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993b). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Brown, R. D. (1999). Adding linguistic knowledge to a lexical example-based translation system. In *Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pages 22–32, Chester, England.

Buck, C., Heafield, K., and Ooyen, B. V. (2014). N-gram counts and language models from the common crawl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 3579–3584, Reykjavik, Iceland.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. (2011). Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland.

Canals-Marote, R., Esteve-Guillen, A., Garrido-Alenda, A., Guardiola-Savall, M., Iturraspe-Bellver, A., Montserrat-Buendia, S., Ortiz-Rojas, S., Pastor-Pina, H., Perez-Antón, P., and Forcada, M. (2001). The Spanish–Catalan machine translation system interNOSTRUM. In *Proceedings of the Machine Translation Summit VIII*, pages 73–76, Santiago de Compostela, Spain.

Carl, M. (1999). Inducing translation templates for example-based machine translation. In *Proceedings of the Machine Translation Summit VII*, pages 250–258, Sigapore.

Carl, M. (2001). Inducing probabilistic invertible translation grammars from aligned texts. In *Proceedings of the 2001 workshop on Computational Natural Language Learning - ConLL '01*, volume 7, pages 1–7, Morristown, NJ, USA.

Carl, M. (2007). METIS-II: The German to English MT system. In *Proceedings of the Machine Translation Summit XI*, pages 65–73, Copenhagen, Denmark.

Carl, M. and Way, A., editors (2003). *Recent Advances in Example-Based Machine Translation*, volume 21. Springer.

Caseli, H. M., Nunes, M. G. V., and Forcada, M. L. (2006). Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. *Machine Translation*, 20(4):227–245. Published in 2008.

Chanod, J. P. and Tapanainen, P. (1999). Creating a tagset, lexicon and guesser for a French tagger. Technical Report cmp-lg/9503004.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Cholakov, K. and Van Noord, G. (2009). Combining Finite State and Corpus-based Techniques for Unknown Word Prediction. In *Recent Advances in Natural Language Processing*, pages 60–64, Borovets, Bulgaria.

Cicekli, I. and Güvenir, H. A. (2001). Learning translation templates from bilingual translation examples. *Applied Intelligence*, 15(1):57–76.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Costa-Jussà, M. R. and Farrús, M. (2014). Statistical machine translation enhancements through linguistic levels: A survey. *ACM Comput. Surv.*, 46(3).

Costa-Jussà, M. R. (2012). An overview of the phrase-based statistical machine translation techniques. *The Knowledge Engineering Review*, 27:413–431.

Costa-Jussà, M. R. and Fonollosa, J. A. (2015). Latest trends in hybrid machine translation and its applications. *Computer Speech & Language*, 32(1):3 – 10. Hybrid Machine Translation: integration of linguistics and statistics.

Crego, J. (2014). SYSTRAN RBMT engine: hybridization experiments. In *3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)*, Gothenburg, Sweden.

Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. (1992). A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 133–140, Trento, Italy.

Détrez, G. and Ranta, A. (2012). Smart paradigms and the predictability and complexity of inflectional morphology. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–653, Avignon, France.

Dugast, L., Senellart, J., and Koehn, P. (2008). Can we Relearn an RBMT System? In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 175–178, Columbus, Ohio, USA.

Détrez, G., Sánchez-Cartagena, V. M., and Ranta, A. (2014). Sharing Resources Between Free/Open-Source Rule-based Machine Translation Systems: Grammatical Framework and Apertium. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 4394–4400, Reykjavik, Iceland.

Eberle, K. (2014). Hybrid strategies for better products and shorter time-to-market. page 97.

Eisele, A., Federmann, C., Saint-Amand, H., Jellinghaus, M., Herrmann, T., and Chen, Y. (2008). Using Moses to integrate multiple rule-based machine translation engines into a hybrid system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 179–182, Columbus, Ohio, USA.

Enache, R., España Bonet, C., Ranta, A., Màrquez Villodre, L., et al. (2012). A hybrid system for patent translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 269–276, Trento, Italy.

Eskander, R., Habash, N., and Rambow, O. (2013). Automatic Extraction of Morphological Lexicons from Morphologically Annotated Corpora. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1032–1043, Seattle, Washington, USA.

Esplà-Gomis, M., Sánchez-Cartagena, V. M., and Pérez-Ortiz, J. A. (2011a). Enlarging monolingual dictionaries for machine translation with active learning and non-expert users. In *Proceedings of Recent Advances in Natural Language Processing*, pages 339–346, Hissar, Bulgaria.

Esplà-Gomis, M., Sánchez-Cartagena, V. M., and Pérez-Ortiz, J. A. (2011b). Multimodal Building of Monolingual Dictionaries for Machine Translation by Non-Expert Users. In *Proceedings of the Machine Translation Summit XIII*, pages 147–154, Xiamen, China.

Esplà-Gomis, M., Sánchez-Cartagena, V. M., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Forcada, M. L., and Carrasco, R. C. (2014). An efficient method to assist non-expert users in extending dictionaries by assigning stems and inflectional paradigms to unknown words. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 19–26, Dubrovnik, Croatia.

Federmann, C., Eisele, A., Uszkoreit, H., Chen, Y., Hunsicker, S., and Xu, J. (2010). Further experiments with shallow hybrid mt systems. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 77–81, Uppsala, Sweden.

Federmann, C. and Hunsicker, S. (2011). Stochastic parse tree selection for an existing rbmt system. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 351–357, Edinburgh, Scotland.

Fiscus, J. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 347–354.

Font-Llitjós, A. (2007). *Automatic Improvement of Machine Translation Systems*. PhD thesis, Carnegie Mellon University.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., F. Sánchez-Martínez, G. R.-S., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144. Special Issue: Free/Open-Source Machine Translation.

Forsberg, M., Hammarström, H., and Ranta, A. (2006). Morphological lexicon extraction from raw text data. In Salakoski, T., Ginter, F., Pyysalo, S., and Pahikkala, T., editors, *Advances in Natural Language Processing*, volume 4139 of *Lecture Notes in Computer Science*, pages 488–499.

Garey, R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Series of Books in the Mathematical Sciences. W. H. Freeman.

Garfinkel, R. S. and Nemhauser, G. L. (1972). *Integer programming*, volume 4. Wiley New York.

Garrido-Alenda, A., Gilabert Zarco, P., Pérez-Ortiz, J. A., Pertusa-Ibáñez, A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Scalco, M. A., and Forcada, M. L. (2004). Shallow parsing for Portuguese–Spanish machine translation. In *Language technology for Portuguese: shallow processing tools and resources*, pages 135–144. Edições Colibri, Lisboa.

Gaspari, F. and Hutchins, J. (2007). Online and free! Ten years of online machine translation: Origins, developments, current use and future prospects. In *Proceedings of the Machine Translation Summit XI*, pages 199–206, Copenhagen, Denmark.

Goodman, J. and Chen, S. F. (1998). An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University.

Graham, Y. and van Genabith, J. (2010). Factor templates for factored machine translation models. In *IWSLT 2010 : 7th International Workshop on Spoken Language Translation*, pages 275–283, Paris, France.

Green, S. and DeNero, J. (2012). A class-based agreement model for generating accurately inflected translations. In *Proceedings of the 50th Annual Meeting of the*

Association for Computational Linguistics: Long Papers - Volume 1, pages 146–155, Jeju Island, Korea.

Haffari, G., Roy, M., and Sarkar, A. (2009). Active learning for statistical phrase-based machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 415–423, Boulder, Colorado.

Hildebrand, A. S. and Vogel, S. (2008). Combination of Machine Translation Systems via Hypothesis Selection from Combined N-Best Lists. In *Proceedings of 8th AMTA conference*, pages 254–261, Hawaii, USA.

Hlaváčová, J. (2001). Morphological guesser of czech words. In Matoušek, V., Mautner, P., Mouček, R., and Tašer, K., editors, *Text, Speech and Dialogue*, volume 2166 of *Lecture Notes in Computer Science*, pages 70–75.

Hutchins, J. (2010). Machine translation: A concise history. *Journal of Translation Studies*, 13(1–2):29–70. Special Issue: The teaching of computer-aided translation.

Hutchins, W. J. and Somers, H. L. (1992). *An introduction to machine translation*, volume 362. Academic Press New York.

Ivars-Ribes, X. and Sánchez-Cartagena, V. M. (2011). A Widely Used Machine Translation Service and its Migration to a Free/Open-Source Solution: the Case of Softcatalà. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 61–68, Barcelona, Spain.

Kahn, A. B. (1962). Topological sorting of large networks. *Communications of the ACM*, 5(11):558–562.

Kaji, H., Kida, Y., and Morimoto, Y. (1992). Learning translation templates from bilingual text. In *Proceedings of the 14th Conference on Computational Linguistics*, pages 672–678, Nantes, France.

Kaufmann, T. and Pfister, B. (2010). Semi-automatic extension of morphological lexica. In *Proceedings of the 2010 International Multiconference on Computer Science and Information Technology (IMCSIT)*, pages 403–409, Wisła, Poland.

Kirchhoff, K. and Yang, M. (2005). Improved language modeling for statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 125–128, Ann Arbor, USA.

Knight, K. (1999). Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4).

Koehn, P. (2004a). Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In Frederking, R. E. and Taylor, K. B., editors, *Machine Translation: From Real Users to Research*, volume 3265 of *Lecture Notes in Computer Science*, pages 115–124.

Koehn, P. (2004b). Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, volume 4, pages 388–395, Barcelona, Spain.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit X*, pages 12–16, Phuket, Thailand.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Koehn, P. and Hoang, H. (2007). Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 868–876, Prague, Czech Republic.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, Prague, Czech Republic.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 48–54, Edmonton, Canada.

Koehn, P. and Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic.

Korte, B. and Vygen, J. (2012). *Combinatorial Optimization: Theory and Algorithms*. Springer, 5th edition.

Labaka, G., España-Bonet, C., Màrquez, L., and Sarasola, K. (2014). A hybrid machine translation architecture guided by syntax. *Machine Translation*, 28(2):91–125.

Lavie, A. (2008). Stat-XFER: A General Search-Based Syntax-Driven Framework for Machine Translation. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 4919 of *Lecture Notes in Computer Science*, pages 362–375.

Liang, P., Taskar, B., and Klein, D. (2006). Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 104–111, New York, New York, USA.

Lindén, K., Tuovila, J., et al. (2009). Corpus-based paradigm selection for morphological entries. In *Proceedings of the 17th Nordic Conference of Computational Linguistics*, pages 96–102, Odense, Denmark.

Liu, Y. and Zong, C. (2004). The technical analysis on translation templates. In *Proceedings of the IEEE International Conference on Systems, Man & Cybernetics*, volume 5, pages 4799–4803.

Marino, J. B., Banchs, R. E., Crego, J. M., de Gispert, A., Lambert, P., Fonollosa, J. A., and Costa-Jussà, M. R. (2006). N-gram-based machine translation. *Computational Linguistics*, 32(4):527–549.

Matusov, E., Ueffing, N., and Ney, H. (2006). Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–40, Trento, Italy.

McCreight, E. M. (1976). A space-economical suffix tree construction algorithm. *J. ACM*, 23(2):262–272.

McShane, M., Nirenburg, S., Cowie, J., and Zacharski, R. (2002). Embedding knowledge elicitation and MT systems within a single architecture. *Machine Translation*, 17:271–305.

Menezes, A. and Richardson, S. (2003). A best-first alignment algorithm for automatic extraction of transfer mappings from bilingua corpora. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, volume 21 of *Text, Speech and Language Technology*, pages 421–442.

Monson, C. (2009). *ParaMor: From Paradigm Structure to Natural Language Morphology Induction*. PhD thesis, Carnegie Mellon University.

Nakov, P., Bonev, Y., Angelova, G., Cius, E., and Von Hahn, W. (2003). Guessing morphological classes of unknown German nouns. In *Recent Advances in Natural Language Processing*, pages 347–356, Borovets, Bulgaria.

Nakov, P., Guzman, F., and Vogel, S. (2012). Optimizing for sentence-level BLEU+1 yields short translations. In *Proceedings of COLING 2012*, pages 1979–1994, Mumbai, India.

Och, F. J. (2002). *Statistical machine translation: from single-word models to alignment templates*. PhD thesis, RWTH Aachen University.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Och, F. J. and Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA.

Peradin, H. and Tyers, F. (2012). A rule-based machine translation system from Serbo-Croatian to Macedonian. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 55–62, Gothenburg, Sweden.

Popovic, M. and Ney, H. (2006). Statistical machine translation with a small amount of bilingual training data. In *5th LREC SALTMIL Workshop on Minority Languages*, page 25–29, Genoa, Italy.

Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montreal, Canada.

Probst, K. (2005). *Automatically Induced Syntactic Transfer Rules for Machine Translation under a Very Limited Data Scenario*. PhD thesis, Carnegie Mellon University.

Probst, K., Levin, L., Peterson, E., Lavie, A., and Carbonell, J. (2002). MT for minority languages using elicitation-based learning of syntactic transfer rules. *Machine Translation*, 17(4):245–270.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Riezler, S., King, T. H., Kaplan, R. M., Crouch, R., Maxwell, III, J. T., and Johnson, M. (2002). Parsing the wall street journal using a lexical-functional grammar and discriminative estimation techniques. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 271–278, Philadelphia, USA.

Riezler, S. and Maxwell III, J. T. (2006). Grammatical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 248–255, New York City, USA.

Roche, E. and Schabes, Y. (1997). Introduction. In Roche, E. and Schabes, Y., editors, *Finite-State Language Processing*, pages 1–65. MIT Press, Cambridge, Mass.

Rosa, R., Mareček, D., and Dušek, O. (2012). Depfix: A system for automatic correction of czech mt outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 362–368, Montreal, Canada.

Rosti, A.-V., Matsoukas, S., and Schwartz, R. (2007). Improved word-level system combination for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 312–319, Prague, Czech Republic.

Rubino, R., Toral, A., Sánchez-Cartagena, V. M., Ferrández-Tordera, J., Ortiz-Rojas, S., Ramírez-Sánchez, G., Sánchez-Martínez, F., and Way, A. (2014). Abu-MaTran at WMT 2014 translation task: Two-step data selection and RBMT-style synthetic rules. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 171–177, Baltimore, MD, USA.

Sánchez-Cartagena, V. M., Esplá-Gomis, M., and Pérez-Ortiz, J. A. (2012). Source-Language Dictionaries Help Non-Expert Users to Enlarge Target-Language Dictionaries for Machine Translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 3422–3429, Istanbul, Turkey.

Sánchez-Cartagena, V. M., Esplá-Gomis, M., Sánchez-Martínez, F., and Pérez-Ortiz, J. A. (2012a). Choosing the correct paradigm for unknown words in rule-based machine translation systems. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 27–39, Gothenburg, Sweden.

Sánchez-Cartagena, V. M. and Pérez-Ortiz, J. A. (2009). An open-source highly scalable web service architecture for the Apertium machine translation engine. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 51–58, Alacant, Spain.

Sánchez-Cartagena, V. M. and Pérez-Ortiz, J. A. (2010a). ScaleMT: a Free/Open-Source Framework for Building Scalable Machine Translation Web Services. *The Prague Bulletin of Mathematical Linguistics*, 93:97–106.

Sánchez-Cartagena, V. M. and Pérez-Ortiz, J. A. (2010b). Tradubi: Open-source social translation for the apertium machine translation platform. *The Prague Bulletin of Mathematical Linguistics*, 93:47–56.

Sánchez-Cartagena, V. M., Pérez-Ortiz, J. A., and Sánchez-Martínez, F. (2014). The UA-Prompsit hybrid machine translation system for the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 178–185, Baltimore, MD, USA.

Sánchez-Cartagena, V. M., Pérez-Ortiz, J. A., and Sánchez-Martínez, F. (2015). A generalised alignment template formalism and its application to the inference of shallow-transfer machine translation rules from scarce bilingual corpora. *Computer Speech & Language*, 32(1):46–90. Hybrid Machine Translation: integration of linguistics and statistics.

Sánchez-Cartagena, V. M., Sánchez-Martínez, F., and Pérez-Ortiz, J. A. (2011a). Enriching a Statistical Machine Translation System Trained on Small Parallel Corpora with Rule-Based Bilingual Phrases. In *Proceedings of Recent Advances in Natural Language Processing*, pages 90–96, Hissar, Bulgaria.

Sánchez-Cartagena, V. M., Sánchez-Martínez, F., and Pérez-Ortiz, J. A. (2011b). Integrating shallow-transfer rules into phrase-based statistical machine translation. In *Proceedings of the Machine Translation Summit XIII*, pages 562–569, Xiamen, China.

Sánchez-Cartagena, V. M., Sánchez-Martínez, F., and Pérez-Ortiz, J. A. (2011c). The Universitat d'Alacant hybrid machine translation system for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 457–463, Edinburgh, Scotland.

Sánchez-Cartagena, V. M., Sánchez-Martínez, F., and Pérez-Ortiz, J. A. (2012b). An open-source toolkit for integrating shallow-transfer rules into phrase-based satistical machine translation. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 41–54, Gothenburg, Sweden.

Sánchez-Martínez, F. (2011). Choosing the best machine translation system to translate a sentence by using only source-language information. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, pages 97–104, Leuven, Belgium.

Sánchez-Martínez, F. and Forcada, M. L. (2009). Inferring shallow-transfer machine translation rules from small parallel corpora. *Journal of Artificial Intelligence Research*, 34(1):605–635.

Sánchez-Martínez, F., Pérez-Ortiz, J. A., and Forcada, M. (2008). Using target-language information to train part-of-speech taggers for machine translation. *Machine Translation*, 22(1-2):29–66.

Schwenk, H., Abdul-Rauf, S., Barrault, L., and Senellart, J. (2009). SMT and SPE machine translation systems for WMT'09. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 130–134, Athens, Greece.

Sennrich, R. (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549, Avignon, France.

Simard, M., Ueffing, N., Isabelle, P., and Kuhn, R. (2007). Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 203–206, Prague, Czech Republic.

Šnajder, J. (2013). Models for Predicting the Inflectional Paradigm of Croatian Words. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 1(2):1–34.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th biennial conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, USA.

Spendley, W., Hext, G. R., and Himsworth, F. R. (1962). Sequential application of simplex designs in optimisation and evolutionary operation. *Technometrics*, 4(4):441–461.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing*, pages 901–904, Denver, USA.

Suykens, J. A. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300.

Thomforde, E. and Steedman, M. (2011). Semi-supervised ccg lexicon extension. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1246–1256, Edinburgh, United Kingdom.

Thurmair, G. (2009). Comparing different architectures of hybrid Machine Translation systems. In *Proceedings of the Machine Translation Summit XII*, Ottawa, Canada.

Tiedemann, J. (2009). Character-based PSMT for Closely Related Languages. In *Proceedings of the 13th Annual Conference of the European Association of Machine Translation*, pages 12–19, Barcelona, Spain.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 2214–2218, Istanbul, Turkey.

Tinsley, J., Ma, Y., Ozdowska, S., and Way, A. (2008). Matrex: the dcu mt system for wmt 2008. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 171–174, Columbus, Ohio, USA.

Trosterud, T. and Unhammer, K. B. (2012). Evaluating North Sámi to Norwegian assimilation RBMT. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 13–26, Gothenburg, Sweden.

Tyers, F. M. (2009). Rule-based augmentation of training data in Breton–French statistical machine translation. In *Proceedings of the 13th Annual Conference of the European Association of Machine Translation*, pages 213–217, Barcelona, Spain.

Tyers, F. M. (2010). Rule-based Breton to French machine translation. In *Proceedings of the 14th Annual Conference of the European Association of Machine Translation*, pages 174–181, St Raphael, France.

Tyers, F. M. (2013). *Feasible lexical selection for rule-based machine translation*. PhD thesis, Universitat d'Alacant, Alacant, Spain.

Tyers, F. M., Sánchez-Martínez, F., and Forcada, M. L. (2012). Flexible finite-state lexical selection for rule-based machine translation. In *Proceedings of the 16th Annual Conference of the European Association of Machine Translation*, pages 213–220, Trento, Italy.

Varga, I. and Yokoyama, S. (2009). Transfer rule generation for a Japanese-Hungarian machine translation system. In *Proceedings of the Machine Translation Summit XII*, Ottawa, Canada.

Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in machine translation. In *IFIP Congress*, volume 68, pages 254–260.

Vilar, D., Peter, J.-T., and Ney, H. (2007). Can We Translate Letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 33–39, Prague, Czech Republic.

Vogel, S., Ney, H., and Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*, pages 836–841, Copenhagen, Denmark.

Wang, A., Hoang, C., and Kan, M. (2013). Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, 47(1):9–31.

Watanabe, T., Suzuki, J., Tsukada, H., and Isozaki, H. (2007). Online large-margin training for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 764–773, Prague, Czech Republic.

Way, A. (2010). Machine translation. In Clark, A., Fox, C., and Lappin, S., editors, *The Handbook of Computational Linguistics and Natural Language Processing*, Blackwell Handbooks in Linguistics. Wiley.

Weaver, W. (1955). Translation. *Machine translation of languages: fourteen essays*, pages 15–23.

Wilks, Y. (1978). Machine translation and artificial intelligence. In *Translating and the Computer: proceedings of a seminar*, pages 27–43, London, United Kingdom.

Xu, Y., Ralphs, T. K., Ladányi, L., and Saltzman, M. J. (2009). Computational experience with a software framework for parallel integer programming. *INFORMS Journal on Computing*, 21(3):383–397.

Yeh, A. (2000). More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, pages 947–953, Saarbrücken, Germany.

Zaidan, O. F. and Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, Portland, USA.

Zbib, R., Kayser, M., Matsoukas, S., Makhoul, J., Nader, H., Soliman, H., and Safadi, R. (2012). Methods for integrating rule-based and statistical systems for Arabic to English machine translation. *Machine Translation*, 26(1-2):67–83.

Zollmann, A. and Vogel, S. (2011). A Word-Class Approach to Labeling PSCFG Rules for Machine Translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1–11, Portland, USA.