Resumen en español

Introducción

La traducción automática (TA) puede definirse como la utilización de un sistema informático para traducir un texto en un lenguaje natural, la lengua origen (LO), a otro lenguaje natural, la lengua meta (LM) o destino de la traducción. Aunque la TA también se emplea para traducir directamente el habla, esta tesis se centra únicamente en la traducción de textos escritos y gramaticalmente correctos.

Desde el punto de vista de la utilización de los sistemas de TA se distinguen principalmente dos usos bien diferenciados: asimilación y diseminación. El principal objetivo de la *asimilación* es la generación de traducciones en LM comprensibles, esto es, que permitan al usuario hacerse una idea del contenido del texto que ha sido traducido automáticamente, sin importar si la traducción es gramaticalmente correcta o contiene palabras sin traducir.

En el caso del uso de sistemas de TA para la *diseminación*, el objetivo es producir de manera automática un borrador para su posterior corrección por traductores profesionales; éste es el caso, por ejemplo, de instituciones públicas plurilingües, como la Generalitat de Catalunya o la Unión Europea, que promulgan leyes en más de un idioma.

Enfoques de la traducción automática

Los sistemas de TA pueden clasificarse atendiendo al tipo de conocimiento que se emplea en su construcción. Desde este punto de vista pueden distinguirse principalmente dos enfoques, uno basado en corpus y otro basado en reglas, aunque existen enfoques híbridos que conjugan ambos.

Los sistemas de TA basados en corpus normalmente requieren de grandes colecciones de textos paralelos a partir de los cuales el sistema aprende a realizar nuevas traducciones. Un texto *paralelo* se define como un texto en el cual se tiene el texto en un idioma junto a su traducción en otro idioma. Para aprender a traducir a partir de textos paralelos, estos deben encontrarse alineados a nivel oracional, es decir, para

cada frase en un idioma se debe identificar claramente su correspondiente traducción en la otra lengua. La obtención de alineamientos a nivel oracional no es trivial dado que durante el proceso de traducción algunas oraciones pueden haber sido borradas, insertadas o fusionadas con otras.

Los sistemas de TA basados en reglas emplean recursos y reglas de traducción, explícitamente codificados por lingüistas expertos, que tratan de describir el proceso de traducción (Hutchins and Somers, 1992). Este tipo de sistemas de TA requieren de recursos lingüísticos tales como diccionarios bilingües y morfológicos (con información léxica, sintáctica e incluso semántica), reglas de desambiguación o corpus desambiguados a mano para la desambiguación léxica y un extenso conjunto de reglas de transferencia estructural. El proceso de desarrollo de un sistema de TA basado en reglas requiere de un gran esfuerzo humano para la construcción de todos los recursos lingüísticos necesarios (Arnold, 2003).

Los enfoques híbridos que conjugan más de un paradigma de traducción automática están recibiendo en la actualidad un creciente interés. En la bibliografía podemos encontrar distintos enfoques híbridos (Dirix et al., 2005; Groves and Way, 2005; Oepen et al., 2007; Koehn and Hoang, 2007; Simard et al., 2007; Dugast et al., 2007); todos ellos tienen en común que su principal objetivo es intentar aliviar la necesidad de recursos tales como diccionarios, reglas de transferencia estructural o grandes colecciones de textos paralelos en la construcción de sistemas de TA.

Sistemas de traducción automática basados en reglas

Pese a que durante los últimos años los enfoques basados en corpus han visto incrementada su atención como resultado de la mayor disponibilidad de textos paralelos, los sistemas de TA basados en reglas siguen siendo activamente desarrollados principalmente por los siguientes motivos:

1. porque no todos los pares de lenguas para los cuales existe demanda tienen a su disposición la gran cantidad de textos paralelos necesarios para entrenar sistemas de TA de propósito general basados en corpus (Forcada, 2006), este es el caso de pares de lenguas tales como catalán–inglés, occitano–catalán o inglés–afrikaans, entre otros; y

2. porque los sistemas basados en reglas son más fácilmente diagnosticables durante su desarrollo y porque, además, los errores de traducción que producen suelen tener una naturaleza más repetitiva y previsible, lo cual ayuda a los profesionales que tienen que corregir su salida para su diseminación.

Esta tesis se centra en el desarrollo de sistemas de TA basados en reglas y más concretamente en sistemas de TA por transferencia estructural superficial (Hutchins

and Somers, 1992) para la traducción entre lenguas emparentadas. El desarrollo de este tipo de sistemas de TA implica, por lo general, el desarrollo de:

- diccionarios morfológicos que para cada palabra en LO proporcionan su posibles *formas léxicas*, consistentes en lema, categoría léxica e información de flexión;

- métodos para resolver la ambigüedad léxica de aquellas palabras en LO que pueden tener más de una interpretación. Ésto implica el desarrollo de desambiguadores léxicos categoriales (Manning and Schütze, 1999, cap. 10) y, dependiendo del sistema de TA, de métodos para resolver la polisemia (Ide and Véronis, 1998; Agirre and Edmonds, 2007);

- diccionarios bilingües que para una determinada palabra en LO (y tal vez alguna información acerca del sentido) ofrecen su traducción en LM; y

- reglas de transferencia estructural que detectan patrones que requieren un procesamiento especial para asegurar el correcto orden de las palabras en LM o la concordancia de género y número, entre otros fenómenos.

El apéndice A ofrece una descripción detalla del sistema de TA de código abierto Apertium usado a lo largo de esta tesis. Este sistema de TA sólo necesita de desambiguadores léxicos categoriales para la resolución de la ambigüedad léxica de los textos en LO dado que el diccionario bilingüe ofrece un único equivalente en LM para cada palabra en LO; este enfoque ha demostrado ser adecuado para la TA entre lenguas emparentadas como español–catalán u occitano–catalán.

De todo los recursos que son necesarios para construir un sistema de TA por transferencia (estructural) superficial esta tesis se centra en la obtención, a partir de corpus, de:

- los desambiguadores léxicos categoriales empleados para resolver la ambigüedad léxica categorial de los textos a traducir, y

- el conjunto de reglas de transferencia que se emplean para adecuar la traducción a la reglas gramaticales de la LM.

El objetivo final es reducir en la medida de lo posible el esfuerzo humano necesario para la construcción de este tipo de sistemas de TA. El sistema resultante puede considerarse híbrido pues integra métodos basados en corpus dentro de un sistema de TA basado en reglas.

Desambiguadores léxicos categoriales para la traducción automática

La desambiguación léxica categorial es un paso habitual en aplicaciones de procesamiento del lenguaje natural; consiste en determinar la categorial léxica de todas y cada una de las palabras de un texto dado. Generalmente los desambiguadores léxicos categoriales se basan en la hipótesis de que a una palabra se le puede asignar una única categoría léxica atendiendo a las categorías léxicas de las palabras que aparecen en su vecindad.

En TA, la correcta elección de la categoría léxica de las palabras a traducir es crucial dado que la traducción de una palabra en LO a la LM puede diferir de una categoría léxica a otra. Por ejemplo, la palabra en español *ahorro* puede traducirse al inglés como *saving* cuando es *etiquetada* (desambiguada) como nombre, mientras que su traducción en inglés sería *save* cuando la palabra es etiquetada como una forma conjugada del verbo *ahorrar*.

De entre los diferente enfoques existentes para la obtención de desambiguadores léxicos categoriales de propósito general, esta tesis se centra en el desarrollo de desambiguadores léxicos categoriales basados en modelos ocultos de Markov (MOM, véase el apéndice B). Los MOM puede ser entrenados de forma *supervisada* mediante el empleo de textos desambiguados (o etiquetados) a mano, o bien de forma *no supervisada* mediante el uso del algoritmo de Baum y Welch con texto no etiquetado. Obviamente, el enfoque supervisado proporciona mejores resultados, pero lleva parejo el coste humano que requiere el etiquetar a mano un texto para el entrenamiento, lo que lo convierte en un recurso caro que no siempre está disponible.

Los dos métodos (supervisado y no supervisado) ya mencionados para el entrenamiento de desambiguadores léxicos categoriales basados en MOM sólo emplean información de la lengua que pretenden desambiguar, un enfoque obvio cuando el desambiguador resultante se va a aplicar en aplicaciones de procesamiento de lenguaje natural que solo implican a un idioma. Sin embargo, cuando se utilizan desambiguadores léxicos categoriales en TA, es decir, cuando la desambiguación léxica categorial no es más que un paso intermedio en el proceso de traducción entre dos lenguas, hay dos hechos a los que la comunidad científica en general no ha prestado suficiente atención:

- por un lado, que hay una fuente de conocimiento, aparte del uso de textos paralelos (Yarowsky and Ngai, 2001; Dien and Kiem, 2003), que se puede utilizar de forma no supervisada para la obtención de mejores desambiguadores léxicos categoriales: el uso de un modelo estadístico de la LM, y

- por otro lado, que en TA la desambiguación léxica categorial no es más que un paso necesario en el proceso de traducción a la LM; por tanto, lo que realmente importa es la calidad final de la traducción, no la precisión del desambiguador,

> o en otras palabras, no importa si una palabra es erróneamente desambiguada si su traducción es correcta.[1]

Esta tesis propone un nuevo método, inspirado en los dos hechos arriba mencionados, para el entrenamiento de desambiguadores léxicos categoriales de la LO, basados en MOM, mediante el empleo de información de la LM, así como del resto de módulos del sistema de TA en el que el desambiguador resultante se integra. El objetivo es obtener, de forma enteramente no supervisada, desambiguadores léxicos categoriales que maximicen la calidad de traducción del sistema en que se integran.

La aplicación de este nuevo método implica los siguientes pasos:

- en primer lugar, el texto en LO se divide en segmentos suficientemente pequeños y de los que se tiene la certeza que serán traducidos independientemente del resto de segmentos;
- después, cada segmento se traduce a la LM atendiendo a todas y cada una de sus posibles desambiguaciones (combinaciones de categorías léxicas), esto provoca que para un segmento dado tengamos distintas traducciones;
- una vez calculadas todas las posibles traducciones de un segmento, cada una de ellas es evaluada haciendo uso de un modelo estadístico de la LM;
- la verosimilitud en LM de cada una de las traducciones es utilizada para calcular la probabilidad de cada una de las desambiguaciones del segmento en cuestión; y
- finalmente, estas probabilidades se utilizan para ajustar los parámetros del MOM.

En el capítulo 2 se describen en detalle los experimentos que han sido realizados para la evaluación de este nuevo método. Los experimentos consisten en el entrenamiento de los desambiguadores léxicos categoriales de tres lenguas diferentes —español, francés y occitano[2]— para su uso dentro del sistema de TA de código abierto Apertium (véase el apéndice A) con la finalidad de traducir al catalán (LM). Tras el entrenamiento, los desambiguadores léxicos categoriales son evaluados de forma indirecta a partir de la evaluación de la calidad de traducción del sistema de TA que integra el desambiguador resultante. Los resultados demuestran que el sistema de TA usado en los experimentos ofrece mejores resultado cuando el desambiguador léxico categorial es entrenado usando este nuevo método en comparación con desambiguadores léxicos categoriales entrenados a través del algoritmo no supervisado clásico (Baum y Welch).

En el caso del español también se evaluó el desambiguador de forma aislada, esto es, se evaluó la calidad del desambiguador tras etiquetar un corpus y comparar el

[1] Nótese que en ocasiones una palabra puede ser traducida del mismo modo para una o más de sus posibles categorías léxicas.

[2] Los experimentos se han realizado con la variante aranesa del dialecto del occitano conocido como gascón y que se habla en el Valle de Arán.

resultado con un corpus de referencia desambiguado a mano; en esta evaluación el resultado ofrecido por el nuevo método de entrenamiento es, una vez más, mejor que el ofrecido por el algoritmo de Baum y Welch.

Además, para español también se ha realizado un comparación con un desambiguador léxico categorial entrenado de forma supervisada a partir de un corpus etiquetado de tamaño medio; sorprendentemente, el método propuesto en esta tesis produce desambiguadores léxicos categoriales que al usarse en TA ofrecen resultado similares a los ofrecidos cuando el desambiguador es entrenado de forma supervisada. Por el contrario, al evaluar los desambiguadores de forma aislada, el desambiguador entrenado de forma supervisada ofrece mejores resultados; ésto es debido a que no siempre la traducción de una palabra incorrectamente desambiguada es errónea. Este resultado viene a confirmar que un desambiguador léxico categorial que es apropiado para su uso en TA no necesariamente tiene por qué serlo para otras tareas de procesamiento del lenguaje natural.

El tener que traducir a la LM cada segmento de la LO atendiendo a todas y cada una de sus posibles desambiguaciones conlleva un elevado coste computacional debido a que el número de traducciones a realizar crece exponencialmente con la longitud de los segmentos. Para paliar este problema en el capítulo 3 se propone un método que hace uso de información *a priori*, obtenida de forma enteramente no supervisada, para descartar el mayor número de desambiguaciones posible antes de traducirlas. Este método se basa en la hipótesis de que un modelo de categorías léxicas de la LO que no es suficientemente bueno para usarse en la desambiguación léxica categorial, sí lo es para, dado un segmento de texto en LO, seleccionar un reducido conjunto de entre sus posibles desambiguaciones, de modo que la correcta se encuentre en dicho conjunto.

El uso de este nuevo enfoque requiere de un modelo de categorías léxicas de la LO que se usa para calcular la probabilidad a priori de todas las posibles desambiguaciones de un segmento dado para posteriormente seleccionar las desambiguaciones con mayor verosimilitud y traducirlas a la LM. En los experimentos para validar este método se ha utilizado un MOM entrando mediante el método de Kupiec (Kupiec, 1992; Manning and Schütze, 1999, p.358), un método de inicialización clásico que también se emplea como modelo inicial cuando el entrenamiento del MOM se hace a través del algoritmo de Baum y Welch. Cabe añadir que el MOM que se emplea para decidir que desambiguaciones se traducen y cuales se descartan puede ser actualizado de forma dinámica durante el entrenamiento.

Los experimentos para evaluar la bondad de este método han sido realizados con los mismos pares de lenguas y datos lingüísticos ya mencionados más arriba. El método que usa información de la LM para el entrenamiento de los desambiguadores léxicos categoriales no se ve modificado en su funcionamiento, salvo por el número de desambiguaciones que son tenidas en cuenta. Los resultado obtenidos demuestran que la

hipótesis inicial es correcta y que se pueden evitar en torno a un 80 % de las traducciones a realizar sin que la calidad de la traducción alcanzada por el sistema de TA que integra el desambiguador léxico categorial resultante se vea afectada.

Inferencia automática del conjunto de etiquetas para la desambiguación léxica categorial

Hasta ahora hemos visto que en esta tesis se usan MOM para la desambiguación léxica categorial de los textos en LO a traducir. Si bien, aunque los MOM pueden entrenarse de forma no supervisada mediante el método ya descrito, ello no evita que tengamos que definir manualmente el conjunto de estados a usar por el MOM para la desambiguación léxica categorial.

En principio, podrían usarse como estados del MOM directamente las categorías léxica de *grano fino* que se obtienen directamente al obviar el lema de las formas léxica obtenidas tras el análisis morfológico del texto a traducir. Sin embargo, su empleo como estados del MOM hace que el número de parámetros a estimar sea excesivamente grande, piénsese que el número de probabilidades de transición entre estados del MOM crece de forma cuadrática con el número de estados cuando se considera un MOM de primer orden (basado en bigramas). Por otra parte, el tener un número elevado de estados hace que la cantidad de éstos que no se ve representado en el corpus de entrenamiento crezca, lo que puede redundar en una peor estimación de los parámetros del MOM.

Por todo ello, se suele definir manualmente el conjunto de estados (*etiquetario*) a usar por el MOM. Esta definición consiste en determinar como se agrupan las categorías léxicas de grano fino en categorías léxicas más generales. Normalmente, la definición del etiquetario se realiza atendiendo a criterios lingüísticos, esto es, evitando agrupar bajo una misma etiqueta general (estado) dos o más categorías de grano fino con distinta función sintáctica. Sin embargo, cuando el desambiguador léxico categorial va a usarse como módulo embebido en un sistema de TA, lo que realmente importa es distinguir entre categorías léxicas de grano fino que dan lugar a traducciones divergentes en la LM, o que ayudan a desambiguar otras palabras con traducciones divergentes que podrían aparecer en su vecindad.

Esta tesis aborda la obtención de forma totalmente automática y no supervisada del etiquetario (conjunto de estados) a usar por el MOM para la desambiguación léxica categorial. Se propone la obtención automática del etiquetario a emplear porque, en primer lugar, ello evitará la necesidad de que su definición corra a cargo de lingüistas expertos y, en segundo lugar, porque la asunción subyacente que se suele hacer cuando se define el etiquetario manualmente no necesariamente tiene por qué redundar en una mejora de la calidad de traducción alcanzada por el traductor automático que integre el desambiguador resultante. Dicha asunción viene a decir que aquellas categorías de

grano fino que tiene la misma categoría léxica presentan, por lo general, distribuciones de probabilidad similares.

En el capítulo 4 se propone la aplicación de un algoritmo de agrupamiento jerárquico aglomerativo para la obtención automática del etiquetario. El algoritmo se aplica sobre un MOM inicial que contiene tanto estados como categorías léxicas de grano fino diferentes proporciona el analizador morfológico del sistema de TA a su salida; este MOM inicial se entrena haciendo uso de información de la LM mediante el nuevo método ya descrito.

El algoritmo comienza con tantos agrupamientos como etiquetas de grano fino o estados se emplean en el MOM inicial. En cada paso, los dos agrupamientos que se encuentran más próximos, atendiendo a una medida de similitud entre ellos, se fusionan en un único agrupamiento. El algoritmo termina cuando no hay más agrupamientos que fusionar o la distancia (disimilitud) entre ellos es mayor que un umbral dado. La media de disimilitud o distancia entre dos agrupamientos se basa en las probabilidades de transición entre estados del MOM e integra, además, una restricción que hace que dos agrupamientos no puedan fusionarse si ello supone una pérdida de información. Esta restricción se hace necesaria porque tras la desambiguación léxica categorial se debe poder recuperar la categoría léxica de grano fino proporcionada por el analizador morfológico a su salida; esta información es requerida por el resto de módulos del sistema de TA tras el desambiguador léxico categorial para llevar a cabo la traducción.

La evaluación del método propuesto para la inferencia automática del conjunto de estados a emplear por el MOM se ha centrado en los mismos pares de lenguas mencionados anteriormente. Para ello se entrenaron desambiguadores léxicos categoriales para español, francés y occitano mediante el método que emplea información de la LM (catalán) pero en este caso utilizando directamente las etiquetas de grano fino que el analizador morfológico proporciona a su salida. Después el algoritmo de agrupamiento jerárquico aglomerativo se aplicó en cada caso sobre el MOM previamente entrenado para distintos valores del umbral de distancia entre agrupamientos. Los resultados obtenidos muestran que un gran número de etiquetas de grano fino no aparecen en los corpus de entrenamiento, lo cual provoca que todas ellas acaben formando parte del mismo agrupamiento.

En cuanto a la calidad de traducción alcanzada por el sistema de TA Apertium cuando integra estos desambiguadores léxicos categoriales, los resultados obtenidos varían de unos pares de lenguas a otros. Mientras que en el caso de español–catalán la calidad de traducción no se ve afectada por el hecho de usar un menor número de estados, en los otros dos casos ésta se deteriora. Si bien, el deterioro en la calidad de la traducción no es muy significativo y puede ser asumido en escenarios de uso de la TA en los que la velocidad o el consumo de memoria sean un factor determinante, pues un menor número de estados implica un menor número de parámetros que cargar en memoria antes de llevar a cabo la desambiguación léxica categorial.

Inferencia automática de reglas de transferencia estructural

Como se mencionó en la introducción, esta tesis también aborda el problema de la inferencia de reglas de transferencia (estructural) superficial a partir de textos paralelos de pequeño tamaño.[3] Las reglas de transferencia son necesarias para producir traducciones gramaticalmente correctas en LM.

En el capítulo 5 se propone un método no supervisado para la inferencia de reglas de transferencia morfológica. Esta reglas se basan en plantillas de alineamiento (Och and Ney, 2004) como las usadas en traducción automática estadística. Las plantillas de alineamiento ofrecen una generalización que se realiza a partir de pares bilingües extraídos de textos paralelos, una vez obtenidos los alineamientos a nivel de palabra, mediante el uso de clases de palabras que pueden ser definidas mediante métodos estadísticos o siguiendo criterios lingüísticos.

Para su empleo en sistemas de TA basados en reglas las plantillas de alineamiento han tenido que ser adaptadas y extendidas con un conjunto de restricciones que controlan su aplicación como reglas de transferencia. Con este fin:

- el diccionario bilingüe del sistema de TA en el que el conjunto de reglas inferidas se empleará se usa para cerciorarse de que los pares bilingües que se obtienen de los textos paralelos, y de los cuales se extraen las plantillas de alineamiento, pueden reproducirse con el sistema de TA;
- el diccionario bilingüe se usa también para obtener de forma automática el conjunto de restricciones que cada plantilla de alineamiento lleva asociado y que vela por su correcta aplicación; y
- las clases de palabras empleadas para generalizar los pares bilingües dando lugar a plantillas de alineamiento se definen siguiendo criterios lingüísticos.

Una vez obtenido un conjunto de plantillas de alineamiento, éstas son filtradas atendiendo a su frecuencia de aparición en la colección de textos paralelos. Dos criterios han sido probados para la selección del conjunto de plantillas de alineamiento, el primero de ellos tiene en cuenta únicamente la frecuencia de aparición de cada plantilla de alineamiento; el segundo considera además la longitud, número de palabras, de cada plantilla de alineamiento para evitar penalizar aquellas plantillas que contemplan un mayor contexto. Finalmente las plantillas de alineamiento seleccionadas se emplean para la generación de reglas de transferencia en el formato usado por el ingenio de TA que las integrará, en este caso Apertium.

[3]Textos paralelos de pequeño tamaño en relación al tamaño de los textos paralelos normalmente empleados para el entrenamiento de sistema de TA de propósito general basados en corpus (Och, 2005).

Cada regla de transferencia se compone de un conjunto de plantillas de alineamiento que comparten la misma secuencia de clases de palabras en LO. Dentro de la misma regla, y ya en tiempo de ejecución, se aplica aquella plantilla de alineamiento con mayor frecuencia de aparición siempre y cuando se cumplan las restricciones que velan por su correcta aplicación. Si se diera el caso de que ninguna de las plantillas de alineamiento puede aplicarse la regla acaba traduciendo el segmento de texto tratado palabra por palabra.

Para evaluar las reglas inferidas automáticamente mediante el método descrito se han realizado experimentos con los pares de lenguas de Apertium español–catalán, español–gallego y español–portugués; en todos los casos se han utilizado dos corpus paralelos, uno que contiene en torno a medio millón de palabras y otro que contiene en torno a dos millones. En cuanto a los corpus utilizados para evaluar la calidad de la traducción se han empleado corpus de dos tipos para cada par de lenguas y dirección de traducción, uno cuya naturaleza es análoga a la de los corpus utilizados para el entrenamiento, y otro obtenido a partir de la post-edición (corrección) del resultado de traducir un corpus con Apertium cuando se usan las reglas de transferencia codificadas a mano que vienen con los paquetes lingüísticos usados en los experimentos.

La calidad de traducción alcanzada por las reglas inferidas automáticamente se compara en el capítulo 5 con la calidad obtenida cuando no se emplean reglas de transferencia, esto es cuando la traducción se realiza palabra por palabra, y con la calidad de traducción obtenida cuando se emplean reglas de transferencia codificadas a mano. Los resultados obtenidos muestran que el uso de reglas de transferencia inferidas automáticamente mediante la adaptación de las plantillas de alineamiento al paradigma de la TA basada en reglas ofrece mejores resultados que la traducción palabra por palabra, y que la calidad de traducción alcanzada se aproxima a la obtenida cuando las reglas de transferencia son codificadas a mano por lingüistas; es más, cuando en la evaluación se emplea un corpus análogo al utilizado para el entrenamiento la calidad de traducción ofrecida por las reglas inferidas automáticamente es similar (en algunos casos incluso mejor) a la ofrecida por las reglas codificadas a mano.

En cuanto a la cantidad de corpus paralelo necesario para obtener un conjunto de reglas de transferencia que proporcionen una calidad de traducción aceptable, los experimentos realizados con distintos tamaños de corpus demuestran que con un corpus de medio millón de palabras la calidad de las reglas inferidas es satisfactoria, incluso para algunos pares de lenguas la calidad es similar a la obtenida cuando las reglas de transferencia se obtiene a partir de un corpus de entrenamiento de dos millones de palabras.

Finalmente, cabe mencionar que los dos criterios utilizados para la selección del conjunto de plantillas de alineamiento a usar en la generación de las reglas de transferencia ofrecen resultados similares. Este resultado sugiere que el criterio que trata de priorizar las plantillas de alineamiento que contemplan un mayor contexto ofrecería

mejores resultados si las reglas contemplaran algún método de "backoff" que permitiera aplicar otras plantillas de alineamiento dentro de la misma regla cuando ninguna de las otras es aplicable en lugar de realizar un traducción palabra por palabra.

Discusión

El principal objetivo de los métodos que se discuten en esta tesis ha sido el de facilitar el desarrollo de sistemas de TA por transferencia superficial evitando la intervención humana en algunas etapas del desarrollo de tales sistemas de TA. Más concretamente esta tesis se ha centrado en:

- un nuevo método no supervisado para el entrenamiento de desambiguadores léxicos categoriales para su uso como módulo embebido en sistemas de TA;
- la aplicación de un algoritmo de agrupamiento para obtener de forma automática el conjunto de estados a usar por desambiguadores léxicos categoriales basados en MOM; y
- la inferencia automática de reglas de transferencia superficial a partir de una colección de textos paralelos de pequeño tamaño.

En cuanto al método de entrenamiento de desambiguadores léxicos categoriales que emplea información de la LM y del resto de módulos del sistema de TA en el que se integra, cabe decir que este nuevo método supone un avance en el desarrollo de desambiguadores léxicos categoriales para su empleo en TA. Éste es el primer método, al menos del que yo tenga constancia, que para el entrenamiento de un desambiguador léxico categorial para una lengua dada usa información de otra lengua sin necesidad de textos paralelos. Este enfoque podría aplicarse para el entrenamiento de otros modelos de la LO, por ejemplo para resolver la polisemia, que podrían beneficiarse de la información que la LM nos brinda.

Para beneficiarse de este nuevo método, los desarrolladores de sistemas de TA basados en reglas solo necesitan construir el resto de módulos del sistema de TA antes de entrenar el desambiguador léxico categorial a ser embebido en dicho sistema. Es más, si las lenguas entre las que se va a llevar a cabo la traducción se encuentra íntimamente emparentadas, como es el caso de español–catalán, los desambiguadores léxicos categoriales podrían entrenarse antes incluso de tener un sistema de TA completo.

El método para la obtención del conjunto de estados a usar por los desambiguadores léxicos categoriales no ha ofrecido los resultados esperados, pues la calidad de la traducción ha empeorado levemente con algunos de los pares de lenguas usados en los experimentos. Esto puede ser debido en parte al método empleado para calcular la distancia o similitud entre dos agrupamientos; en el capítulo 6 se proponen otros métodos que podrían mejorar los resultados obtenidos.

Con respecto al método que emplea plantillas de alineamiento para la generación de reglas de transferencia superficial para su uso en TA, cabe mencionar que esta es la primera aproximación que adapta la plantillas de alineamiento usadas en TA estadística al paradigma de la TA basada en reglas. Las reglas generadas son totalmente legibles, lo que permite su edición por lingüistas expertos. Esto significa que durante el desarrollo de un nuevo sistema de TA los desarrolladores pueden usar este nuevo método para obtener un conjunto de reglas de transferencia superficial básicas y después introducir manualmente nuevas reglas; de este modo los desarrolladores podrían concentrar sus esfuerzo en la escritura de aquellas reglas que revisten especial dificultad. Desde mi punto de vista ésta es una gran ventaja con respecto a otros métodos basados en corpus porque permite la coexistencia de reglas inferidas automáticamente y reglas codificadas a mano.

Por último, hay que mencionar que todos los métodos descritos en esta tesis han sido implementados y liberados como código abierto bajo la licencia GNU GPL (véase el apéndice C). La disponibilidad del código fuente en su totalidad, sin restricción alguna, garantiza la reproducibilidad de todos los experimentos realizados y permite, además, a otros investigadores la mejora de los algoritmos discutidos sin necesidad de reimplementarlos.

Todos los métodos y algoritmos han sido implementados de tal forma que se integran perfectamente en el proceso de desarrollo de nuevos pares de lenguas para el sistema de TA de código abierto Apertium; esto beneficia, por un lado, a otros investigadores que utilizan Apertium como plataforma para la investigación y prueba de nuevas ideas y, por otro lado, a la gente interesada en desarrollar nuevos pares de lenguas para Apertium, como demuestra el hecho de que los desambiguadores léxicos categoriales de algunos pares de lenguas, como francés–catalán u occitano–catalán, hayan sido obtenidos mediante el método de entrenamiento de desambiguadores léxicos categoriales propuesto en esta tesis.