El sistema de traducción automática castellano↔catalán interNOSTRUM

R. Canals-Marote, A. Esteve-Guillén, A. Garrido-Alenda, M.I. Guardiola-Savall, A. Iturraspe-Bellver, S. Montserrat-Buendia, S. Ortiz-Rojas, H. Pastor-Pina, P.M. Pérez-Antón y M.L. Forcada,

Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, E-03071 Alacant

Resumen Este artículo describe interNOSTRUM, un sistema de traducción automática (TA) castellano↔catalán actualmente en desarrollo que alcanza una gran velocidad mediante el uso de tecnología de estados finitos (lo que permite su integracion en navegadores de internet) y obtiene una precisión razonable utilizando una estrategia de transferencia morfológica avanzada (lo que permite producir rápidamente borradores de traducciones listos para una postedición ligera).

1 Introducción

Este artículo describe el sistema de traducción automática (TA) castellano↔catalán interNOSTRUM. Una de las razones principales razones de la demanda de traducciones del castellano al catalán es el impulso de la normalización o recuperación lingüística del catalán en las comunidades que tienen esta lengua como oficial (unos seis millones de catalanohablantes de un total de unos diez millones de habitantes); por otro lado, la traducción del catalán al castellano (principalmente para la asimilación de información) permite el acceso a documentos en catalán a las personas no catalanohablantes de España o de América Latina. El sistema interNOSTRUM todavía se esta desarrollando y tanto la Universitat d'Alacant, una universidad de tamaño medio, como la Caja de Ahorros del Mediterráneo (CAM), una de las cajas de ahorros más importantes de España, utilizan un prototipo del programa desde hace dos años; existe una versión de prueba disponible para el público en `http://internostrum.com`. Estas dos instituciones promovieron este proyecto de tres años (1999–2001) que actualmente se encuentra en su tercer año y que cuenta en su equipo con dos lingüistas y cuatro ingenieros informáticos. Si bien la precisión de las traducciones y la cobertura del vocabulario pueden mejorar (especialmente en la versión catalán→castellano), la velocidad del sistema —miles de palabras por segundo o decenas de millones de palabras por día en un PC típico de 1999 que actúa como servidor de internet— permite que sea usado ya como un sistema para obtener, de forma instantánea, borradores de traducciones que son relativamente fáciles de convertir en documentos publicables y para comprender información en otra lengua durante la navegación por internet (los accesos a nuestro servidor se incrementan todos los meses de forma regular; en marzo de 2001 se registraron 120.000 visitas). Estas velocidades se consiguen utilizando tecnología de estados finitos (Roche y Schabes 1997) en la mayoría de los módulos del sistema.

2 Prototipo actual y versiones futuras

Como se ha comentado anteriormente, interNOSTRUM no es un producto acabado; sin embargo, está disponible en internet. Esto se debe a dos de los objetivos básicos de nuestro proyecto: el primero, generar una versión operativa de interNOSTRUM tan pronto como fuera posible (la primera versión castellano→catalán fue publicada en noviembre de 1999) y el segundo, publicar la última versión estable lo más rápido posible. Estas son las principales razones para la configuración actual como servidor de internet en lugar de como programa para su distribución.

Actualmente, interNOSTRUM traduce textos ANSI, RTF (*rich text format* de Microsoft) y HTML (*hypertext markup language* o lenguaje de marcas para hipertextos) del castellano al catalán central (la variedad estándar usada oficialmente en la comunidad autónoma catalana) y viceversa. Una versión del sistema que aceptará y generará las variedades estándar valenciana y balear estará preparada en septiembre de 2001. La versión catalán→castellano es muy reciente (marzo de 2001) y todavía no es muy satisfactoria en cuanto a precisión y cobertura de vocabulario, aunque ya puede ser usada para la asimilación de información. Las tasas de error (número de palabras corregidas por cada 100 palabras en el texto meta para hacerlo aceptable) se situan en torno al 5% en sentido castellano→catalán cuando se traducen textos periodísticos (por ejemplo, tomados de `http://www.elpais.es`, y son algo peores en sentido inverso.

2.1 Plataforma

interNOSTRUM actualmente se ejecuta sobre Linux utilizando Apache y PHP3 y está públicamente accesible en `http://internostrum.com`; en la CAM se está utilizando una versión interna sobre un servidor en Windows 2000–Internet Information Services. El sistema interNOSTRUM consiste en ocho módulos que se ejecutan en paralelo y se comunican a través de canales de texto.[1] Seis de los ocho módulos se generan automáticamente a partir de los correspondientes datos lingüísticos[2] utilizando compiladores escritos con ayuda de `yacc` y `lex`, herramientas del sistema operativo Unix (ver tabla 1). La velocidad actual del sistema está alrededor de las 1 000 palabras por segundo) en un PC típico de 1999 (un Pentium de 400 MHz).

2.2 Estrategia de traducción automática

interNOSTRUM es un sistema clásico de traducción automática indirecta que utiliza una estrategia de transferencia morfológica avanzada (parecida a la *arquitectura transformadora* o *trasformer architecture*, Arnold (1993) o a los *sistemas directos*, Hutchins y Somers (1992)) similar a algunas utilizadas en sistemas comerciales de TA para PC. [3] Como se dijo anteriormente, interNOSTRUM consiste en ocho módulos (ver figura 1): un módulo desformateador (que separa el texto de la información de formato), dos módulos de análisis (analizador morfológico y desambiguador léxico categorial), dos módulos de transferencia (módulo de diccionario bilingüe y módulo de procesamiento de patrones), dos módulos de generación (generador morfológico y postgenerador) y el módulo reformateador (que reintegra la información de formato original al texto traducido).

2.2.1 Módulos basados en tecnología de estados finitos

Cuatro de los módulos de interNOSTRUM, a saber, el analizador morfológico, el módulo del diccionario bilingüe, el generador morfológico y el postgenerador, están basados en *transductores de estados finitos* (TEF) (Roche y Schabes 1997). Esto permite que el procesamiento alcance velocidades del orden de 10 000 palabras por segundo, velocidad que es prácticamente independiente al tamaño de los diccionarios. Los TEF leen su entrada símbolo a símbolo; cada vez que leen un símbolo, se mueven a un nuevo estado y escriben en su salida uno o más símbolos.

El analizador morfológico se genera automáticamente (Garrido et al. 1999) a partir del *diccionario morfológico* (DM) para la lengua origen (LO). El DM contiene los lemas (formas canónicas o de base de las palabras con flexión), los paradigmas de flexión y las relaciones entre ellos. El subprograma lee las formas *superficiales* (FS) y escribe, para cada una, una o más *formas léxicas* (FL) que consisten en un lema, una categoría léxica e información de flexión.

El módulo del diccionario bilingüe , el cual es invocado por el módulo de procesamiento de patrones (véase más abajo), se genera automáticamente a partir de un fichero que contiene las correspondencias bilingües entre lemas. El programa lee una FL en LO y escribe su FL equivalente en la lengua meta (LM).

El generador morfológico básicamente lleva a cabo la tarea inversa del analizador morfológico, pero referida a la LM. El generador morfológico se genera a partir del DM de la LM.

El postgenerador : Aquellas FS involucradas en el guionado y la apostrofación

[1] El uso de canales de texto legible permite diágnosticar fácilmente algunos problemas y es una alternativa muy eficiente para las implementaciones en Linux (Unix).

[2] Esta característica convierte a interNOSTRUM en un sistema fácilmente adaptable a otras lenguas.

[3] Tales como Transcend RT de Transparent Technologies, las primeras versiones de Power Translator de Globalink (Mira i Gimènez y Forcada 1998; Forcada 2000), y Reverso de Softissimo.

(tales como los pronombres proclíticos, los artículos, algunas preposiciones, etc.) activan este módulo que normalmente está *dormido*. El postgenerador se genera a partir de un fichero que contiene las reglas correspondientes para la LM.

La división de un texto en palabras presenta aspectos no son triviales. Por una parte, existen grupos de palabras que no pueden traducirse palabra por palabra, y deben ser tratados como *unidades multipalabra* (UMP); estas unidades —aquellas que son de longitud fija— se están incorporando continuamente a los diccionarios morfológicos y al bilingüe; para ello, los módulos correspondientes son capaces de tratar tanto las UMP invariables como las que tienen flexión; su uso permite, además, evitar algunos de los problemas de traducción causados por la homografía, la polisemia o por estructuras no composicionales tales como algunas locuciones y colocaciones. Ejemplos: Cast. *con cargo a* → Cat. *a càrrec de*; Cast. *por adelantado* → Cat. *per endavant*; Cast. **echar** *de menos* → Cat. **trobar** *a faltar*; Cast. **tener** *que* + infinitivo → Cat. **haver** *de* + infinitivo. En los dos últimos ejemplos (una construcción modal y una locución), la UMP tiene un elemento susceptible de flexión (indicado en negrita). Por otra parte, existen combinaciones de ciertas formas verbales y pronombres enclíticos que se escriben como una sola palabra en castellano; estas combinaciones presentan transformaciones ortográficas tales como cambios en la acentuación o pérdida de consonantes: Cast. *dámelo* = *dá* + *me* + *lo* → Cat. *dóna* + *me* + *lo* = *dóna-me'l*; Cast. *presentémonos* = *presentemos* + *nos* → Cat. *presentem* + *nos* = *presentem-nos*. El analizador morfológico se ocupa de resolver todos estos casos si los paradigmas correspondientes se introducen convenientemente en los diccionarios morfológicos.

2.2.2 El desambiguador léxico categorial

La mayoría de las ambigüedades léxicas están dentro de dos grandes grupos: la *homografía* (cuando una FS tiene más de una FL o análisis) y la *polisemia* (cuando la FS tiene una sola FL pero el lema puede tener más de una interpretación). El módulo de desambiguación léxica o *desambiguador léxico categorial* (en inglés *part-of-speech tagger*) utiliza un modelo oculto de Markov basado en bigramas y trigramas (secuencias de dos o tres categorías léxicas) para resolver aquellos homógrafos que presentan ambigüedad categorial. Los parámetros del modelo reflejan las estadísticas de aparición conjunta de categorías observadas en un corpus de referencia; el desambiguador calcula en una pasada la desambiguación más probable de cada frase.

Actualmente estamos ajustando el conjunto de etiquetas utilizado[4] y construyendo un corpus de entrenamiento más grande para mejorar el funcionamiento de este módulo.[5] La polisemia no se trata (únicamente en algunos casos mediante el uso de unidades multipalabra de longitud fija que representan colocaciones): el diccionario bilingüe siempre proporciona el mismo equivalente para cada lema; se da el caso de que los errores debidos a la polisemia son mucho menos frecuentes que aquellos que se deben a la homografía. El problema de la polisemia se evitará en las aplicaciones bancarias y administrativas mediante el uso de un *castellano controlado* adaptado a estos tipos de texto (ver sección 3).

2.2.3 El módulo de procesamiento de patrones

A pesar del gran parecido entre el castellano y el catalán, existen bastantes divergencias gramaticales: divergencias de género y número que afectan a la concordancia — Cast. *la deuda contraída* (fem.) → Cat. *el deute contret* (masc.)—; construcciones de relativo con *cuyo*, ausente en catalán — Cast. *la cuenta cuyo titular es el asegurado* → Cat. *el compte el titular del qual és l'assegurat* —, o cambios preposicionales — Cast. *en Londres* → Cat. *a Londres* —.

Estas divergencias se tratan utilizando las reglas gramaticales oportunas. Como se ha dicho más arriba, interNOSTRUM utiliza una solución similar a la de algunos sistemas de

[4]El conjunto de etiquetas (unas 60 para ambos idiomas) se diferencia de los de propósito general en que en él se establecen distinciones que son relevantes para la traducción: por ejemplo, en castellano se distinguen el subjuntivo del indicativo para distinguir formas como *salen* que puede ser una forma de *salir* (catalán *surten*) o de *salar* (catalán *salin*).

[5]Una de las principales contribuciones a la actual tasa de error de interNOSTRUM son los pocos errores que aparecen en algunos homógrafos *difíciles* pero *frecuentes*, como puede ser *una* (artículo/verbo), *para* (verbo/preposición), y *como* (conjunción/verbo). Afortunadamente otros homógrafos son más fáciles de desambiguar. Cuando se analiza el catalán se encuentran problemas similares.

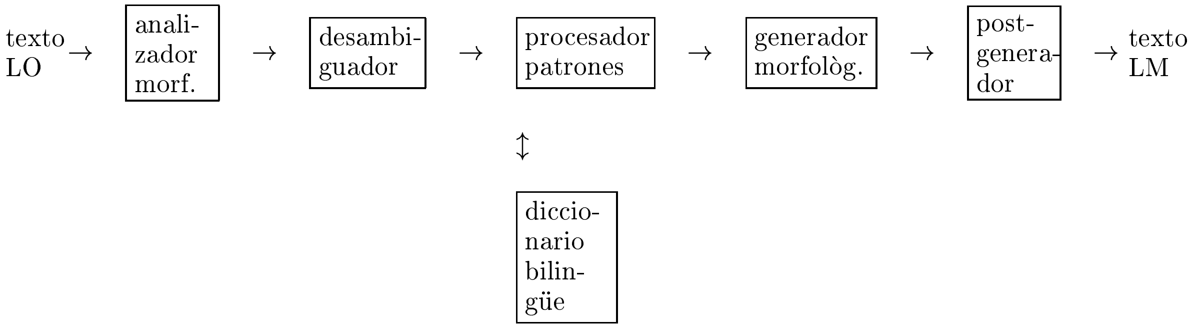

Figura 1: Los módulos lingüísticos de interNOSTRUM (los módulos desformateador y reformateador no aparecen en la figura).

Tabla 1: Generación automática de los módulos de interNOSTRUM a partir de datos lingüísticos

Lengua	datos lingüísticos	programa generador	módulo interNOSTRUM
LO	diccionario morfológico	compilador analizadores morfológicos	analizador morfológico
LO	corpus analizado morfológ.	entrenamiento desambiguador	desambiguador
LO, LM	diccionario bilingüe	compilador diccionarios bilingües	módulo diccionario bilingüe
LO, LM	reglas procesamiento patrones	compilador reglas procesamiento patrones	módulo procesamiento patrones
LM	diccionario morfológico	compilador generadores morfológicos	generador morfológico
LM	reglas guionado y apostrofación	compilador postgeneradores	postgenerador

TA comerciales: se basa en la detección y tratamiento de secuencias predefinidas de categorías léxicas (*patrones* o *fragmentos* bien definidos) que pueden considerarse como sintagmas rudimentarios: por ejemplo, **art.–subst.** o **art.–subst.–adj.** son dos sintagmas nominales válidos. Las secuencias conocidas por el programa constituyen su *catálogo de patrones*. Este módulo (Garrido-Alenda y Forcada 2001) funciona como sigue:

- El texto (analizado morfológicamente y desambiguado) se lee de izquierda a derecha, FL a FL.
- El módulo busca, a partir de la posición actual en el texto, la secuencia de FL más larga que coincide con algún patrón de su catálogo (por ejemplo, si la posición actual en el texto está en "una señal inequívoca ...", el módulo escogerá el patrón **art.–subst.–adj.** en lugar de **art.–subst.**.
- El módulo opera sobre este patrón (propaga el género y el número para asegurar la concordancia en la lengua meta, lo reordena o realiza otros cambios gramaticales) siguiendo las reglas asociadas al patrón.
- El módulo de procesamiento de patrones continua inmediatamente a continuación del patrón que se acaba de procesar (no vuelve a procesar una FL sobre la que ya ha realizado alguna operación).

Cuando no se detecta ningún patrón en la posición actual, el programa traduce una FL aisladamente y reanuda el proceso en la siguiente FL. Los fenómenos de "largo alcance", como la concordancia entre sujeto y verbo, requieren la propagación de información de un patrón a otro, que se realiza manteniendo información de estado que puede ser propagada de izquierda a derecha de un patrón a otro.

El módulo de procesamiento de patrones se genera automáticamente a partir de un fichero que contiene las reglas que especifican los patrones y sus acciones asociadas. Este módulo es el más lento (velocidades del orden de mil palabras por segundo), comparado con las decenas de miles de palabras por segundo del resto de los módulos. El catálogo actual de interNOSTRUM contiene unas dos docenas de patrones.

2.3 El desformateador y el reformateador

Ambos módulos están escritos en `lex` y `C++`; el reformateador es mucho más sencillo. Hay tres versiones de cada módulo: la versión ANSI, la versión RTF y la versión HTML. El desformateador se utiliza para identificar y separar los comandos de formato del texto que se quiere traducir. La información sobre formato, las imágenes incluídas en el documento, etc. se encapsulan (entre dobles corchetes "`[[...]]`") para formar los denominados *superblancos*, que los restantes módulos ven como simples espacios en blanco entre palabras (los segmentos de material de formato muy grandes (> 8 kB) se escriben en ficheros temporales cuyo nombre unívoco se envia entre corchetes en lugar de los datos); cuando los módulos lingüísticos producen traducciones que tienen más o menos palabras que el original, una heurística asegura una colocación razonable de los superblancos y asegura que no se pierdan. Una versión especial del desformateador convierte los URL de las etiquetas HTML "`<A HREF=...>`" en URL redirigidos a través de interNOSTRUM con el fin de permitir traducciones en tiempo real durante la navegación.[6]

3 Herramientas de apoyo proyectadas para interNOSTRUM

Actualmente se trabaja sobre tres herramientas de apoyo: (a) un *asistente de estilo* para ayudar a los autores de textos en castellano a evitar algunas ambigüedades difíciles de resolver utilizando las reglas sintácticas, léxicas y de estilo especificadas en un *castellano controlado*; (b) un *asistente de preedición*, para la desambiguación manual de palabras y estructuras problemáticas (se hace clic sobre ellas y se obtiene un menú de opciones, lo que será útil cuando la estrategia estadística utilizada por el programa sea incapaz de escoger la opción correcta); y (c) un *asistente de postedición*, en el cual los autores podrán hacer clic sobre aquellas palabras del texto meta que sospeche que son traducciones incorrectas, de forma que le permitirá sustituirlas por una de las opciones compatibles con el texto original, ofrecidas en un menú.

[6] También permite el tratamiento de muchos tipos de páginas con *frames*.

4 Conclusiones

Hemos presentado interNOSTRUM, un sistema de traducción automática castellano↔catalán actualmente en desarrollo que alcanza una gran velocidad con el uso de tecnología de estados finitos y una precisión razonable utilizando una estrategia de transferencia morfológica avanzada. El sistema está disponible como servidor de internet y se está utilizando para obtener borradores de traducciones del castellano al catalán y para navegar a través de servidores de internet catalanes en castellano.

Agradecimentos: Este trabajo ha sido financiado por la Caja de Ahorros del Mediterráneo y por el Vicerrectorado de Nuevas Tecnologías de la Universitat d'Alacant, y más recientemente por la Comisión Interministerial de Ciencia y Tecnología a través del proyecto TIC2000-1599-C02-02.

Referencias

Arnold, D. (1993). Sur la conception du transfert. En Bouillon, P. y Clas, A., editores, *La traductique*, pages 64–76. Presses Univ. Montréal, Montréal.

Forcada, M. L. (2000). Learning machine translation strategies using commercial systems: discovering word-reordering rules. En Lewis, D. y Mitkov, R., editores, *MT 2000: machine translation and multilingual applications in the new millennium*, pages 7–1–7–8. British Computer Society.

Garrido, A., Iturraspe, A., Montserrat, S., Pastor, H., y Forcada, M. (1999). A compiler for morphological analysers and generators based on finite-state transducers. *Procesamiento del Lenguaje Natural*, (25):93–98.

Garrido-Alenda, A. y Forcada, M. L. (2001). Morphtrans: un lenguaje y un compilador para especificar y generar módulos de transferencia morfológica para sistemas de traducción automática. En *Actas del XVII Congreso de la Sociedad Española de Procesamiento del Lenguaje Natural, Jaén, septiembre de 2001.*

Hutchins, W. y Somers, H. (1992). *An Introduction to Machine Translation.* Academic Press.

Mira i Gimènez, M. y Forcada, M. L. (1998). Understanding PC-based machine translation systems for evaluation, teaching and reverse engineering: the treatment of noun phrases in Power Translator. *Machine Translation Review (British Computer Society)*, 7:20–27. (disponible en `http://www.dlsi.ua.es/~mlf/mtr98.ps.Z`).

Roche, E. y Schabes, Y. (1997). Introduction. En Roche, E. y Schabes, Y., editores, *Finite-State Language Processing*, pages 1–65. MIT Press, Cambridge, Mass.