Construcción rápida de un sistema de traducción automática español↔portugués partiendo de un sistema español↔catalán*

Patrícia Gilabert-Zarco, Javier Herrero-Vicente, Sergio Ortiz-Rojas, Antonio Pertusa-Ibáñez, Gema Ramírez-Sanchez, Felipe Sánchez-Martínez, Marcial Samper-Asensio, Miriam A. Scalco y **Mikel L. Forcada**
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071 Alacant

Resumen: Se describe el proceso seguido para construir rápidamente un sistema de traducción automática español–portugués y portugués–español, partiendo de un sistema existente que traduce entre el castellano y el catalán. Un equipo de cuatro desarrolladores ha producido en seis meses un sistema ya utilizable, con una cobertura de texto superior al 95 % y con una tasa de texto erróneo en torno al 10 %, y que traduce miles de palabras por segundo. Este proceso ha sido facilitado en gran parte por la existencia de un ingenio ("motor") de traducción independiente de los datos lingüísticos (un sistema clásico de transferencia) y de compiladores para convertir los datos lingüísticos en los formatos usados por este ingenio, y, por otro lado, por la existencia de datos lingüísticos de naturaleza morfológica para ambas lenguas.
Palabras clave: Traducción automática, desarrollo rápido, español, portugués

Abstract: This paper describes the rapid construction of a Spanish–Portuguese, Portuguese–Spanish machine translation system, starting from an existing system for the Spanish–Catalan pair, developed by the same research group. A team of four developers has produced in six months a useful system having a text coverage above 95 % and a word error rate around 10 % running at thousand of words a second. The process has partly been made easier, on the one hand, by the existence of a translation engine independent from the linguistic data and the avaliability of compilers to turn the linguistic data into the formats used by the engine, and, on the other hand, by the availability of morphological data for both languages.
Keywords: machine translation, rapid development, Spanish, Portuguese.

1. Introducción

Este artículo describe el proceso seguido para construir rápidamente un sistema de traducción automática (TA) español–portugués (`es`–`pt`) y portugués–español (`pt`–`es`), con énfasis en el portugués del Brasil, partiendo de un sistema existente que traduce entre el español y el catalán (`ca`). Un equipo de 4 desarrolladores ha producido en 6 meses (es decir, 2 personas·año) un sistema ya utilizable, con una cobertura de texto superior al 95 % y con una tasa de texto erróneo en torno al 10 % y que traduce miles de palabras por segundo. Este proceso ha sido facilitado en gran parte, por un lado, por la existencia de un ingenio ("motor") de traducción independiente de los datos lingüísticos (un sistema clásico de transferencia) y de compiladores para convertir los datos lingüísticos en los formatos usados por este ingenio, y, por otro lado, por la existencia de datos lingüísticos de naturaleza morfológica para ambas lenguas. Además, se ha podido aprovechar sin modificaciones el sistema de filtros y programas auxiliares (servidor de Internet) que permitían usar el sistema para (a) traducir documentos ASCII, RTF y HTML, (b) traducir documentos de Internet durante la navegación, con seguimiento de enlaces, (c) traducir correo electrónico y (d) permitir tertulias electrónicas (*chat*) en las que cada participante decide en qué lengua de las dos verá todas las intervenciones en la sala. El sistema resultante, que se desarrolla desde noviembre de 2002, está disponible en Internet en `http://copacabana.dlsi.ua.es`. El sistema español–catalán del que se partía está públicamente accesible en `http://www.`

* Trabajo financiado por Portal Universia, S.A., con apoyo de la Comisión Interministerial de Ciencia y Tecnología a través del proyecto TIC2000-1599-C02-01.

interNOSTRUM.com.

Debe mencionarse que existen ya dos sistemas (comerciales) de traducción entre el español y el portugués: ATS[1] es un sistema comercial basado en un servidor y que se tarifica por palabra (la versión gratuita permite probar textos de hasta 50 palabras) y Delta Translator[2] es un paquete comercial para instalar en ordenadores de sobremesa. El prototipo descrito en este artículo está, como ATS, disponible por internet, pero permite traducir gratuitamente, además de frases sueltas, documentos y páginas de internet completas;[3] no se ha realizado una evaluación comparativa con los productos comerciales, por un lado, porque nuestro sistema se encuentra aún en una fase relativamente temprana de su desarrollo, y, por otro, porque las modalidades de uso previstas para nuestro sistema difieren notablemente de las de los productos citados.

El artículo se organiza como sigue: el apartado 2 describe brevemente la estructura del ingenio de traducción; en el apartado 3 se describe en detalle la construcción de los datos lingüísticos necesarios para los sistemas español–portugués y portugués–español; se muestran resultados preliminares en el apartado 4; finalmente, el apartado 5 establece las conclusiones.

2. El ingenio de traducción

Este apartado describe brevemente la estructura del ingenio de traducción, procedente del sistema español–catalán interNOSTRUM (Canals-Marote et al., 2001; Garrido-Alenda y Forcada, 2001; Garrido et al., 1999). Se trata de un sistema clásico de traducción automática indirecta que utiliza una estrategia de transferencia morfológica avanzada similar a la de algunos sistemas comerciales de TA para PC. Consiste en una *cadena de montaje* de ocho módulos, seis de los cuales se compilan automáticamente a partir de ficheros con los correspondientes datos lingüísticos. Los módulos, cuya organización se muestra en la figura 1, son:

- El *desformateador*, que separa el texto a traducir de la información de formato.
- El *analizador morfológico*, que segmenta el texto en formas superficiales (FS) (las unidades léxicas tal como se presentan en los textos) y entrega para cada FS una o más formas léxicas (FL) consistentes en un *lema* (la forma base comúnmente usada para las entradas de los diccionarios clásicos), la *categoría léxica* (nombre, verbo, preposición, etc.) y la información de flexión morfológica (número, género, persona, tiempo, etc.). La división de un texto en FS presenta aspectos complejos debida a la existencia, por un lado, de contracciones (*del*, *teniéndolo*, *vámonos*) y, por otro, de unidades léxicas de más de una palabra (*a pesar de*, *echó de menos*). Este módulo se genera a partir de un diccionario morfológico para la lengua origen (LO) usando un compilador de analizadores morfológicos (Garrido et al., 1999).
- El *desambiguador léxico categorial*, elige, usando un modelo estadístico (modelo oculto de Markov), una de las FL de una FS ambigua de acuerdo con su contexto; las FS ambiguas por tener más de un lema, más de una categoría o representar más de una flexión son muy comunes (en torno a una de cada tres FS) y son una fuente muy importante de errores de traducción en caso de elegir el equivalente incorrecto. El modelo estadístico se entrena sobre corpus suficientemente representativos de textos en LO.
- El *módulo de transferencia estructural*, que detecta y trata patrones de palabras (sintagmas) que exigen tratamiento especial por causa de las divergencias gramaticales entre las lenguas (cambios de género y número, reordenamientos, cambios preposicionales, etc.). Este módulo se genera a partir de un archivo de reglas escrito en un lenguaje de programación sencillo (Garrido-Alenda y Forcada, 2001).
- El *módulo de transferencia léxica*, que gestiona un diccionario bilingüe y es invocado por el módulo de transferencia estructural, lee cada FL en LO y entrega la FL correspondiente en lengua meta (LM). Este módulo se genera a partir de un diccionario bilingüe usando un compilador de módulos de transferencia léxi-

[1] http://www.automatictrans.es

[2] http://www.deltatranslator.com/dtspa.htm

[3] Con limitaciones razonables de tamaño para evitar la congestión del sistema

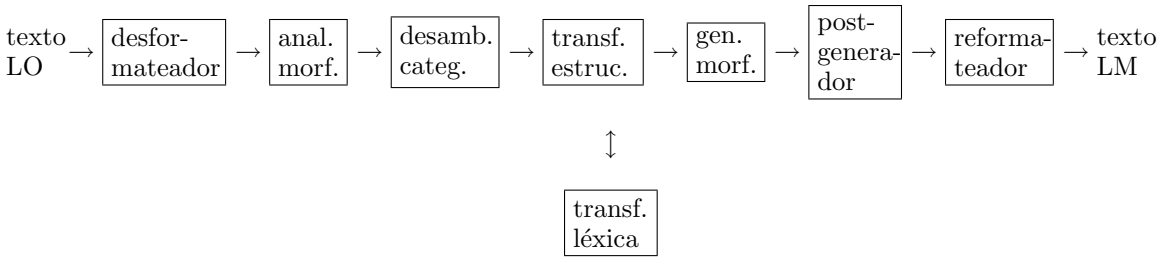

Figura 1: Los ocho módulos que forman la cadena de montaje del sistema de traducción automática.

ca.

- El *generador morfológico*, que genera a partir de la FL en LM una forma superficial adecuada en LM, flexionándola adecuadamente. Este módulo se genera a partir de un diccionario morfológico para la LM mediante un compilador de generadores morfológicos.
- El *postgenerador*, que realiza algunas operaciones ortográficas en LM tales como contracciones y apostrofaciones, y que es generado por un compilador a partir de reglas de transformación en un formato similar al de los diccionarios anteriores.
- El *reformateador*, que reintegra la información de formato original al texto traducido.

Cuatro de estos módulos, a saber, el analizador morfológico, el módulo de transferencia léxica (diccionario bilingüe), el generador morfológico y el postgenerador, están basados en *transductores de estados finitos* (Garrido et al., 1999).

Este diseño permite obtener sistemas de TA *rápidos* (que traducen miles de palabras por segundo en ordenadores de sobremesa comunes) y cuyos resultados son, a pesar de los errores, razonablemente inteligibles y fáciles de corregir. En el caso de lenguas emparentadas como las que nos ocupan, una traducción mecánica palabra por palabra (con un equivalente fijo) presentaría errores la mayoría de los cuales se pueden resolver con un análisis bastante somero (un análisis morfológico seguido de un análisis sintáctico superficial, local y parcial) y con un tratamiento adecuado de las ambigüedades léxicas (principalmente de la homografía). El diseño presentado sigue estas directrices con resultados muy interesantes.

3. Construcción de los datos lingüísticos

3.1. Diccionarios morfológicos

El diccionario morfológico (DM) del español (`es`) fue sencillo de construir ya que consiste esencialmente en un subconjunto del desarrollado en el sistema español–catalán existente.

Para la construcción del diccionario morfológico del portugués (`pt`) fueron de gran utilidad las bases de datos de portugués europeo (`pt_PT`) para el programa `jspell` (Dias de Almeida, 1994), disponibles libremente en `http://natura.di.uminho.pt/~jj/pln/`; este programa es una modificación del corrector ortográfico `ispell` para Unix y Linux, con la particularidad de que las reglas de generación de las FS aceptadas a partir de los lemas incluye información gramatical sobre la flexión. La adaptación consistió básicamente en las siguientes tareas: (a) conversión de los indicadores de flexión y de las categorías gramaticales al formato requerido por nuestros compiladores; (b) adición de las variantes en portugués brasileño (`pt_BR`) de aquellas palabras cuya ortografía es diferencial (p.e., `pt_PT` *projecto* = `pt_BR` *projeto*; `pt_PT` *dezasseis* = `pt_BR` *dezesseis*; `pt_PT` *connosco* = `pt_BR` *conosco*, etc.); (c) adecuación de los lemas (por ejemplo, obtención de lemas propios para aquellas palabras que en `jspell` se obtienen mediante derivación), y (d) reconstrucción completa de los paradigmas de flexión verbal para permitir una gestión más sencilla de estos, en particular en lo referente a las variaciones ortográficas debidas a la contracción con pronombres enclíticos (*traduz* + *los* → *tradu-los*, etc.). Una buena parte del DM del `pt` del sistema de traducción automática se ha obtenido a partir de estos datos; también ha sido necesario añadir entradas nuevas.

3.2. Diccionario bilingüe

La creación del diccionario bilingüe ha sido una de las tareas más costosas del proyecto. Debe tenerse en cuenta que, dada la extensión actual de los DM español y portugués, la cobertura de traducción del sistema resultante viene básicamente determinada por la cobertura de las correspondencias léxicas del diccionario bilingüe. Inicialmente se usaron listas de correspondencias de palabras encontradas en Internet (muchas de ellas resultado del trabajo de personas no expertas), lo que permitió tener unas 3000 correspondencias que exigieron un gran esfuerzo de depuración tras la extracción automática. Este diccionario se completó con correspondencias para las palabras de clases léxicas cerradas (determinantes, pronombres, preposiciones) y con listas de nombres propios de persona y de lugar que no varían de una lengua a la otra y se construyeron los primeros prototipos completos de traductores automáticos.

Posteriormente, se usaron estos prototipos sobre corpus de textos reales (algunos obtenidos de la prensa electrónica accesible en Internet y otros de la base de datos *Portuguese Newswire Text* que distribuye el *Linguistic Data Consortium*) para obtener listas de palabras desconocidas por el sistema ordenadas por su orden de frecuencia. Este es el ciclo básico de crecimiento de los diccionarios del sistema: se obtienen listas de palabras frecuentes y desconocidas, se incorporan a los diccionarios (añadiendo palabras derivadas de la misma familia donde se puede hacer automáticamente), y se repite el proceso; de esta manera, se dispone periódicamente de un nuevo prototipo con mayor cobertura de traducción.

3.3. Reglas de transferencia estructural

Una buena parte de las reglas de transferencia estructural existentes en el sistema español–catalán interNOSTRUM han podido ser utilizadas directamente sin modificaciones tanto para el sentido `es`–`pt` como para el sentido `pt`–`es`: por ejemplo, se han conservado intactas todas aquellas reglas que aseguran que al traducir sintagmas nominales (SN) sencillos (determinante–substantivo, determinante–substantivo–adjetivo, determinante–adjetivo–substantivo, determinante–adjetivo, etc.) en los que se debe revisar la concordancia, bien porque el substantivo ha cambiado de género o número (*una señal roja* (`es`) → *um sinal vermelho* (`pt`)), o bien porque es necesario determinar el género o el número cuando el original era ambiguo para alguna de las palabras (*una crisis* (`es`) → *uma crise* (`pt`), pero *dos crisis* (`es`) → *duas crises* (`pt`)). La reutilización de reglas es posible porque el orden de las palabras en la mayoría de los SN es el mismo en las tres lenguas (`es`, `pt`, `ca`). Además de estas reglas, se han añadido reglas para tratar los siguientes problemas frecuentes de transferencia `es`–`pt` y `pt`–`es`:

- Posición diferencial de adverbios y adjetivos comparativos como *más*/*mais* y *menos*/*menos* en SN (p.ej. *dos coches más* (`es`) → *mais dois carros* (`pt`))
- Reglas para la elección correcta de tiempos de verbo; por ejemplo, el portugués usa el futuro de subjuntivo tanto en expresiones temporales (*quando vieres*) como condicionales (*se vieres*) mientras que en español se usa el presente de subjuntivo si es una expresión temporal (*cuando vengas*) pero se usa el pretérito imperfecto de subjuntivo cuando es una expresión condicional (*si vinieras*).
- Reglas para los artículos delante de locativos (*da França* (`pt`) → *de Francia* (`es`)).
- Algunas reglas (aún insuficientes) para reposicionar los pronombres átonos o clíticos, cuya distribución en portugués es diferente que en castellano (proclítico en portugués y enclítico en castellano o viceversa): (*enviou-me* (`pt`) → *me envió* (`es`); *para te dizer* (`pt`) → *para decirte* (`es`), etc.).
- Reglas para la preposición *a* en algunas construcciones modales con *ir* y *venir* (*vai comprar* (`pt`) → *va a comprar* (`es`)).
- Reglas para traducir el español *que* por el portugués *do que* en oraciones comparativas y viceversa.
- Reglas léxicas, por ejemplo, para decidir la traducción correcta del adverbio *muito* (`pt`) → *muy*/*mucho* (`es`) o la del adjetivo *primeiro* (`pt`) → *primer*/*primero* (`es`).
- Reglas para los gerundios (p.ej., *em* + pronombre átono + gerundio (`pt`) → gerundio + pronombre átono (`es`)).

3.4. Desambiguación léxica categorial

La construcción de un desambiguador léxico categorial basado en modelos de Markov ocultos (Cutting et al., 1992) para la LO de un sistema de TA comporta: (a) el diseño de un etiquetario (conjunto de partes de la oración) reducido, es decir, que agrupa las etiquetas finas entregadas por el analizador morfológico en etiquetas más gruesas pero elegidas de manera que sean adecuadas para la tarea de traducción; (b) la construcción de un corpus de entrenamiento suficientemente representativo de la LO, una pequeña parte del cual deberá ser etiquetada manualmente al menos para la evaluación; (c) la obtención de las probabilidades del modelo de Markov oculto mediante el entrenamiento propiamente dicho sobre el corpus.

Para el traductor `es`–`pt`, se ha usado directamente el módulo de desambiguación léxica categorial del español que existía en el sistema español–catalán. Para el traductor `pt`–`es` era necesaria la construcción de un desambiguador completo siguiendo los pasos indicados. La necesidad —mostrada por la entidad privada financiadora del proyecto— de tener también disponible, tan pronto como fuese posible, un prototipo `pt`–`es`, nos hizo adoptar una solución provisional consistente en usar para el portugués el etiquetador del español (etiquetario y probabilidades). Este desambiguador está lejos de ser adecuado para la tarea, ya que asume, por ejemplo, que la sintaxis (la distribución de etiquetas) del portugués y la del español es la misma, pero ha permitido la construcción de un prototipo `pt`–`es` utilizable que desambigua correctamente en muchos casos. Se espera incorporar un desambiguador específico para el portugués (actualmente en construcción) en setiembre de 2003.

4. Resultados

Para dar una idea aproximada de los resultados que se pueden obtener con el prototipo del sistema en su versión de mayo de 2003, se presenta, a título ilustrativo, una evaluación de los resultados sobre dos textos de unas dos mil palabras construidos tomando porciones de noticias de secciones de temática diferente de diarios electrónicos (en portugués, *O Globo*, `http://oglobo.globo.com` y en español, *El Mundo*, `http://www.elmundo.es`). Para ello definimos la tasa de error como el número mínimo de operaciones de postedición (corrección humana) necesarias para conseguir que la traducción automática en bruto sea adecuada, dividido por el número de palabras total del texto, y expresada como tanto por ciento. Las operaciones de postedición son: (a) el *borrado* de una palabra, (b) la *inserción* de una palabra y (c) la substitución de una palabra por otra (otras operaciones como la transposición de palabras se reducen a combinaciones de estas tres). La tabla 1 muestra los resultados obtenidos en ambas direcciones (`es`–`pt` y `pt`–`es`); en particular, se anota cuántas substituciones han sido necesarias por culpa de palabras desconocidas por el sistema y cuántos errores se deben a la resolución incorrecta de una palabra homógrafa. Los resultados obtenidos en dirección `es`–`pt` son claramente mejores que en dirección contraria; aunque en principio se podría pensar que esto es debido al hecho de que el desambiguador léxico del portugués no es adecuado, los resultados indican que el impacto de la homografía mal resuelta sólo puede explicar parcialmente la diferencia.

	es–pt	pt–es
Palabras totales	**2107**	**2314**
Homógrafos mal resueltos	4	35
Palabras desconocidas	13	53
Otros errores	64	152
Palabras corregidas	**81**	**240**
Tasa de error	3,8 %	10,4 %

Cuadro 1: Datos sobre la calidad de traducción automática obtenida por el prototipo actual.

La velocidad de traducción medida con textos llanos (sin formato) de 6 millones de palabras en un PC de sobremesa con procesador AMD Athlon 2100 es de 5600 palabras por segundo en ambas direcciones.

5. Conclusiones

Los resultados presentados confirman que la estrategia seguida para construir un sistema de TA español–catalán es adecuada para el par español–portugués, como muestran los resultados preliminares presentados para un prototipo realizado. Por un lado, la separación de datos lingüísticos e ingenio de traducción, articulada mediante el uso de compiladores, ha permitido el uso del ingenio existente sin modificaciones; por otro lado, la estrategia seguida para la adquisición y la implementación de los datos lingüísticos pa-

ra el par español–portugués ha permitido la construcción de un prototipo útil (disponible en `http://copacabana.dlsi.ua.es`) con un esfuerzo de 2 años-persona; en la actualidad, el prototipo es claramente mejor en dirección `es`–`pt` que en la contraria.

Bibliografía

Canals-Marote, R., A. Esteve-Guillén, A. Garrido-Alenda, M. Guardiola-Savall, A. Iturraspe-Bellver, S. Monserrat-Buendia, S. Ortiz-Rojas, H. Pastor-Pina, P.M. Perez-Antón, y M.L. Forcada. 2001. El sistema de traducción automática castellano-catalán interNOSTRUM. *Procesamiento del Lenguaje Natural*, 27:151–156. XVII Congreso de la Sociedad Española de Procesamiento del Lenguaje Natural, Jaén, Spain, 12-14.09.2001.

Cutting, D., J. Kupiec, J. Pedersen, y P. Sibun. 1992. A practical part-of-speech tagger. En *Third Conference on Applied Natural Language Processing. Association for Computational Linguistics. Proceedings of the Conference.*, páginas 133–140, Trento, Italia, 31 marzo–3 abril.

Dias de Almeida, José João. 1994. Jspell — um módulo para análise léxica genérica de linguagem natural. En *Encontro da Associação Portuguesa de Lingüística*. disponible en `http://natura.di.uminho.pt/~jj/pln/`.

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

Garrido-Alenda, Alicia y Mikel L. Forcada. 2001. Morphtrans: un lenguaje y un compilador para especificar y generar módulos de transferencia morfológica para sistemas de traducción automática. *Procesamiento del Lenguaje Natural*, 27:157–164.