Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos

Xavier Gómez Guinovart
Seminario de Lingüística Informática
Universidade de Vigo
sli@uvigo.es

Elena Sacau Fontenla
Seminario de Lingüística Informática
Universidade de Vigo
elesacau@uvigo.es

Resumen: En esta comunicación, presentamos las estrategias adoptadas en la investigación sobre el Corpus CLUVI para optimizar los resultados de la extracción automática de un léxico bilingüe inglés-gallego.
Palabras clave: corpus paralelos, extracción de léxico bilingüe, traducción

Abstract: In this paper, we present the strategies adopted in the research on the CLUVI Corpus in order to optimize the results of the automatic extraction of an English-Galician bilingual lexicon.
Keywords: parallel corpora, bilingual lexical extraction, translation

1 Introducción[1]

Los corpus paralelos, constituidos por textos en su versión original y traducida, ofrecen diversas posibilidades de explotación en el ámbito del procesamiento del lenguaje natural. Los resultados más importantes de la explotación de estos corpus se obtienen en las aplicaciones del PLN relacionadas con la traducción y la lexicología, en los campos de la traducción automática estadística (Knight, 1997), de las memorias de traducción y de la traducción automática basada en ejemplos (Turcato y Popowich, 2001), de la extracción léxica para recuperación de información multilingüe (Brown et al., 2000), de la extracción de terminología bilingüe (Vintar, 2001) y de la extracción de léxico bilingüe (Tiedemann, 2003).

En este último campo de trabajo, en el que se enmarca esta comunicación, el objetivo es la generación de diccionarios bilingües basados en las equivalencias léxicas de traducción identificadas en un corpus paralelo, en el que previamente se han establecido a nivel de frase u oración las correspondencias de traducción. En cierto sentido, el problema central de la extracción de léxico bilingüe consiste en convertir un corpus paralelo etiquetado con los alineamientos (es decir, con las equivalencias de traducción) a nivel de oración, en un corpus etiquetado paralelo con los alineamientos a nivel de palabra. Para lograr esta tarea, se han desarrollado diversos algoritmos, gran parte de ellos basados en medidas estadísticas relacionadas con la asociación mutua o con la coaparición de los elementos léxicos en las frases u oraciones bilingües alineadas (Och y Ney, 2003), y mostrando todos un margen de error nada desdeñable en los resultados (Tiedemann, 2003) debido a la naturaleza intrínsecamente "no literal" de la traducción y a otras dificultades relacionadas con las características del corpus, como la distancia lingüística entre las lenguas implicadas, el tipo de textos o el estilo de la traducción.

[1] Este trabajo fue financiado por la Xunta de Galicia, dentro del proyecto "Estudio e adquisición de recursos básicos de lingüística computacional do galego para a elaboración e mellora de aplicacións informáticas de tecnoloxía lingüística" (ref. PGIDT01TICC06E); y por el Ministerio de Ciencia y Tecnología (MCYT) y el Fondo Europeo de Desarrollo Regional (FEDER), dentro del proyecto "Procesamiento lingüístico-computacional del Corpus Lingüístico de la Universidad de Vigo (CLUVI)" (ref. BFF2002-01385), proyecto cofinanciado por la Dirección Xeral de I+D de la Xunta de Galicia y por la Universidade de Vigo. Más información en http://webs.uvigo.es/sli.

En esta comunicación, presentamos las estrategias adoptadas en la investigación sobre el Corpus CLUVI para optimizar los resultados de la extracción automática de un léxico bilingüe inglés-gallego. El Corpus CLUVI (Corpus Lingüístico da Universidade de Vigo[2]) es un corpus textual abierto de registros especializados de lengua gallega contemporánea. En su estado actual de desarrollo, los textos de la sección principal del CLUVI pertenecen a cuatro registros especializados (de los ámbitos jurídico-administrativo, periodístico, informático y literario) y a cuatro combinaciones lingüísticas en relación al gallego (monolingüe gallego, traducción gallego-español, traducción francés-gallego y traducción inglés-gallego), y poseen una extensión total aproximada de 6 millones de palabras. Este conjunto de textos del CLUVI se encuentra repartido en cinco subcorpus, cada uno de alrededor de un millón de palabras: el corpus paralelo TECTRA de textos literarios inglés-gallego, el corpus paralelo FEGA de textos literarios francés-gallego, el corpus paralelo LEGA de textos jurídico-administrativos gallego-español, el corpus monolingüe XIGA de textos sobre informática en gallego y el corpus monolingüe MEGA de lenguaje de los medios de comunicación social. La ampliación del CLUVI con los textos paralelos tetralingües inglés-gallego-francés-español de *El Correo de la Unesco*, y con textos paralelos cinematográficos inglés-gallego está en fase de elaboración.

Los experimentos de optimización se realizaron a partir de los resultados del programa de alineamiento léxico NATools (Simões y Almeida, 2003), que aplica una versión mejorada del algoritmo de Twente (Hiemstra, 1998) para calcular un índice de la correlación entre las coapariciones de los elementos léxicos en las oraciones bilingües alineadas. En lo que resta del artículo, desarrollamos en cuatro apartados (asimetrías de traducción en el corpus paralelo, anotación morfosintáctica paralela, preedición del corpus paralelo, y evaluación y filtrado selectivo de los resultados) las estrategias de optimización aplicadas en la generación de diccionarios de traducción inglés-gallego a partir del corpus paralelo TECTRA de textos literarios inglés-gallego, para rematar con algunas conclusiones y líneas futuras de investigación.

[2] Disponible para consulta pública en http://sli.uvigo.es/CLUVI/.

2 Asimetrías en el corpus paralelo

Las asimetrías de traducción en el corpus paralelo, es decir, los alineamientos no biunívocos y las alteraciones de orden que se dan en la traducción, son una de las principales causas de error en los resultados de la extracción léxica bilingüe. En el corpus paralelo TECTRA de textos literarios inglés-gallego alineado a nivel de oración que forma parte del Corpus CLUVI lo más frecuente es que a una frase del original en inglés le corresponda una frase de la traducción (alineamiento biunívoco o alineamiento 1:1). Sin embargo, no son pocos los casos en los que una frase del original no se traduce (alineamiento 1:0), o en los que a una frase del original le corresponde en la traducción media frase (1:1/2) o dos frases (1:2), o incluso en los que una frase de la traducción no se corresponde con ninguna frase del original (0:1). Además, la traducción también puede implicar alteraciones del orden original, mediante desplazamientos de frases enteras, o movimientos de fragmentos de frases del original a otras frases en la traducción. Todas estas asimetrías dificultan enormemente la extracción de equivalencias léxicas bilingües adecuadas, ya que los alineadores de palabras en general utilizan como base de sus cálculos la coaparición simultánea de los elementos léxicos bilingües candidatos en oraciones entre las que se haya determinado una equivalencia de traducción.

En el Corpus CLUVI, el sistema de codificación de los alineamientos oracionales de los textos paralelos está basado en el formato TMX, estándar para la codificación en XML de memorias de traducción, y utiliza como unidad básica de segmentación la frase ortográfica del texto original. El formato de anotación empleado en el CLUVI utiliza una versión adaptada de algunas de las etiquetas que forman parte de la especificación TMX 1.4 (Savourel, 2004) para representar tanto las correspondencias de traducción que no son biunívocas, como los reordenamientos. La especificación TMX no tiene en cuenta la codificación de estas asimetrías de las traducciones, ya que fue diseñada para el almacenamiento e intercambio de memorias de traducción, y no para la representación de segmentos equivalentes en corpus paralelos.

Las asimetrías traductológicas codificadas en el Corpus CLUVI se agrupan en tres categorías -omisiones, adiciones y reordenamientos-, y son etiquetadas mediante una versión adaptada de los elementos <hi> y <ph>, parte del TMX 1.4.

En la omisión, hay un fragmento del texto de partida que no tiene correspondencia en el texto de llegada, es decir, una frase o parte de una frase no es traducida. La omisión se codifica en los corpus paralelos del CLUVI con el elemento <hi>. Según la especificación TMX 1.4, el elemento <hi> (de nombre derivado del inglés "highlight") "delimits a section of text that has special meaning, such as a terminological unit, a proper name, an item that should not be modified, etc." (Savourel, 2004). En la especificación del CLUVI, basada en la TMX, el elemento <hi> marca en el texto de partida el elemento que se omite en el texto de llegada. Indicamos este uso de la etiqueta <hi> mediante un atributo type caracterizado con el valor de "supr". Por ejemplo, las frases alineadas inglés-gallego de (1) serían anotadas en el CLUVI como (2):

(1) *'Hello', I said.*
-Ola.

(2)
```
<tu><tuv xml:lang="en">
<seg>'Hello',<hi type="supr">
I said.</hi></seg></tuv>
<tuv xml:lang="gl">
<seg>-Ola.</seg></tuv></tu>
```

Por otra parte, la adición en la traducción implica una inserción de fragmentos en el texto de llegada que no tienen correspondencias en el texto de partida. La adición también se codifica en el CLUVI con el elemento <hi>, haciendo que este indique el fragmento insertado en la traducción. Este uso de la etiqueta <hi> se distingue mediante un atributo type caracterizado con el valor de "incl". El fragmento añadido se incorpora a la unidad de traducción en la que está inserto. Cuando el nuevo fragmento es una oración (o una secuencia de oraciones), se incorpora bien a la unidad de traducción anterior, bien a la siguiente, según su contexto. Véase un ejemplo simple de uso de esta etiqueta:

(3) *'Hello.'*
-Ola - dixen.

(4)
```
<tu> <tuv xml:lang="en">
<seg>'Hello.'</seg></tuv>
<tuv xml:lang="gl">
<seg>-Ola <hi type="incl">
- dixen.</hi></tuv></tu>
```

Finalmente, el reordenamiento implica desplazamientos de frases enteras, o movimientos de fragmentos de frases del original a otras frases en la traducción. Estos movimientos se reordenan en la sección de textos traducidos de los corpus paralelos del CLUVI para lograr que los segmentos equivalentes se localicen en la misma unidad de traducción y permitir de este modo una extracción léxica bilingüe menos problemática. El reordenamiento se codifica en el CLUVI mediante una combinación de los elementos <hi> y <ph>. Anotamos el fragmento o la oración movida mediante un elemento <hi> que incluye un atributo type con valor de "reord" y un atributo x con un valor numérico que actúa de índice. Por otro lado, indicamos con un elemento <ph> el lugar en el texto que ocupaba originalmente el elemento desplazado. De acuerdo con la especificación TMX 1.4, el elemento <ph> (o "placeholder") se utiliza "to delimit a sequence of native standalone codes in the segment. Standalone codes are codes that are not opening or closing of a pair, for example empty elements in XML" (Savourel, 2004). En la especificación del CLUVI, basada en la TMX, el elemento adaptado <ph> indica el punto de partida del movimiento, mientras que la relación entre el elemento desplazado y el lugar de partida es codificada en el elemento <ph> mediante un atributo x que comparte valor con el índice codificado en el elemento <hi> del segmento movido. Obviamente, la etiqueta que indica el lugar de origen siempre es una etiqueta vacía. Como criterio de etiquetado en la codificación del CLUVI, y con la finalidad de evitar incoherencias entre las distintas personas que participan en la codificación del corpus, los segmentos reordenados siempre son desplazados en dirección al inicio del texto. En consecuencia, en el CLUVI no hay ninguna secuencia semejante a <ph x="n"/> [...] <hi type="reord" x="n"> Elemento reordenado</hi>; en su lugar, las secuencias son siempre así: <hi type="reord" x="n"> Elemento reordenado</hi> [...] <ph x="n"/>. He aquí un ejemplo sencillo de codificación de un reordenamiento:

(5) *The front door!' she said in this loud whisper. 'It's them!'*

-A porta de fóra. ¡Son eles! - murmurou bastante alto.

(6)
```
<tu> <tuv xml:lang=”en”>
<seg>’The front door!’ she
said in this loud
whisper.</seg></tuv>
<tuv xml:lang=”gl”>
<seg>-A porta de fóra.<hi
type=”reord” x=”1”>- murmurou
bastante alto.</hi></seg>
</tuv></tu><tu>
<tuv xml:lang=”en”>
<seg>It’s them.</seg>
</tuv><tuv xml:lang=”gl”>
<seg>¡Son eles!
<ph x=”1”/></seg></tuv></tu>
```

Si hubiera reordenamientos adicionales, estos se codificarían consecutivamente con los atributos <x=”2”>, <x=”3”>, ..., <x=”n”>, como se muestra en el siguiente ejemplo:

(7) *‘Leave him alone, hey’ Sunny said. ‘C’mon, hey. We got the dough he owes us. Let’s go.’*
-Déixao. Imos logo. Xa témo-lo que nos debe - dicía Sunny.

(8)
```
<tu><tuv xml:lang=”en”>
<seg>’Leave him alone, hey’
Sunny said.</seg></tuv>
<tuv xml:lang=”gl”>
<seg>-Déixao. <hi
type=”reord” x=”1”>- dicía
Sunny.</hi></seg></tuv></tu>
<tu><tuv xml:lang=”en”>
<seg><hi type=”supr”>’C’mon,
hey.</hi></seg></tuv>
<tu xml:lang=”gl”>
<seg></seg></tuv></tu>
<tu><tuv xml:lang=”en”>
<seg>We got the dough he owes
us.</seg></tuv>
<tuv xml:lang=”gl”><seg>
<hi type=”reord” x=”2”>Xa
témo-lo que nos debe
<ph x=”1”/></hi></seg>
</tuv></tu><tu>
<tuv xml:lang=”en”>
<seg>Let’s go.’</seg></tuv>
<tuv xml:lang=”gl”><seg>
Imos logo.<ph x=”2”/></seg>
</tuv></tu>
```

3 Anotación morfosintáctica paralela

Suele considerarse que la información morfosintáctica puede ser de utilidad para mejorar la precisión de la extracción léxica (Tiedemann, 2003). Para el etiquetado morfosintáctico de los corpus paralelos del CLUVI empleamos el etiquetario morfosintáctico elaborado por el SLI (Aguirre et al., 2003) de acuerdo con las directrices de EAGLES. El sistema probabilístico de etiquetado y desambiguación utilizado en el CLUVI, desarrollado conjuntamente por el SLI e Imaxin Software, utiliza un léxico computacional del gallego que contiene las especificaciones morfosintácticas definidas en el etiquetario del SLI.

Los textos en inglés del CLUVI son etiquetados con el programa Trigram’s Tags (TnT) (Brants, 2000), como se hace también para el corpus paralelo IJS-ELAN (Erjavec, 2002). Véase un ejemplo simplificado del corpus paralelo etiquetado con información categorial:

(9)
```
<tu><tuv xml:lang="en">
<seg>In_IN the_DT town_NN
they_PRP tell_VBP the_DT
story_NN of_IN the_DT
great_JJ pearl_NN ._.</seg>
</tuv> <tuv xml:lang="gl">
<seg>Na_PREP_ARDFS
cidade_NCFS
cóntase_VIPRS3_PPS3AR
a_ARDFS historia_NCFS
da_PREP_ARDFS gran_AXAPFS
perla_NCFS_perla
._PUNTO</seg></tuv></tu>
```

4 Preedición del corpus paralelo

Junto a la anotación morfosintáctica paralela, se explora la posibilidad de emplear como fuente para la generación de diccionarios una versión “limpia” de los corpus paralelos, preeditada para facilitar la extracción bilingüe mediante la eliminación de ciertas palabras gramaticales de alto índice de frecuencia y de los segmentos de texto que están marcados en el corpus paralelo como omisiones o como adiciones. Tanto en un caso como en el otro, se trata de elementos que generan una gran cantidad de errores y de ruido en la salida de la extracción léxica.

Dependiendo de la lengua del texto, se eliminarán unas unidades u otras, aunque en general para las dos lenguas se eliminan los signos de puntuación, excluyendo los guiones de unión de palabras compuestas; los dígitos; y los segmentos etiquetados como omisiones o inserciones, ya que indican unidades sin correspondencia de traducción.

Concretamente, en la versión preeditada del corpus paralelo, eliminamos de la lengua

inglesa determinantes (*the, a, an*), pronombres personales (*I, you, he, she, it, we, you, they, me, you, him, her, you*), posesivos (*my, his, her*), demostrativos (*this, that*), conjunciones (*and, but, if, or*), preposiciones (*to, of, in, at, on, with, out, around, about*), partículas negativas (*no, not*), pronombres indefinidos (*all*), verbos auxiliares (*do, does, did, is, are, was, were, has, had*) y la marca de genitivo sajón.

Para el gallego eliminamos artículos (*o, a, os, as*), indefinidos (*un, uns, unha, unhas*), pronombres personales tónicos (*eu, ti, el, ela, nós, vós, eles, eles, me, se, nos*), posesivos (*meu, meus, seu, seus*), preposiciones (*a, con, de, en, para, por*), contracciones de preposición con artículo (*ó, ao, á, ós, aos, ás, co, coa, cos, coas, do, da, dos, das, no, na, nos, nas, polo, pola, polos, polas*), conjunciones (*que, e, se, nin, ou, pero*), verbos de tipo auxiliar (*é, era*) y partículas negativas (*non*).

5 Evaluación de la extracción léxica

A continuación, mostramos las conclusiones derivadas de la evaluación de los resultados de la extracción léxica bilingüe automática, utilizando el alineador léxico NATools (Simões y Almeida, 2003) con el corpus paralelo TECTRA de textos literarios inglés-gallego de 1.335.090 palabras alineado a nivel de oración que forma parte del Corpus CLUVI. La evaluación se realiza primero sobre los resultados de aplicar NATools al corpus preeditado (o versión "limpia" del corpus, descrita en el apartado anterior) sin etiquetar morfosintácticamente, y después al corpus preeditado etiquetado morfosintácticamente.

El objetivo de la evaluación es establecer un sistema automático de filtrado selectivo de los resultados de la extracción léxica bilingüe automática, para poder seleccionar los candidatos léxicos más fiables de acuerdo con criterios como la frecuencia de aparición de la palabra en la lengua fuente o como el índice de probabilidad del alineamiento léxico asociado con las palabras de la lengua término (Vintar, 2001).

El léxico inglés-gallego generado por NATools consiste en una lista bilingüe de todas las palabras distintas que aparecen en los textos en inglés del corpus. Cada palabra inglesa se acompaña de su frecuencia absoluta en el corpus y de las palabras en gallego (hasta un máximo de ocho) que el programa considera las traducciones más probables. A cada palabra gallega del léxico bilingüe generado se le adjunta un índice estimativo de la correlación entre su presencia en una frase y la presencia de la palabra inglesa original en la frase alineada correspondientes, es decir, un estimativo de la probabilidad de coaparición de los dos elementos léxicos (el inglés y el galego) en una misma unidad oracional de traducción. He aquí algunos ejemplos de las entradas bilingües generadas por NATools:

(10)
```
windows_15 ->
fiestras_0.84175086,
cristais_0.07912458,
garaxe_0.07912458

longing_3 ->
morriña_0.36656892,
señardade_0.15835778,
francia_0.15835778,
formara_0.15835778,
período_0.15835778

bed_96 ->
cama_0.83687806,
(null)_0.04077505,
leito_0.03240258,
deitar_0.02170320,
entre_0.01576635,
dei_0.01096033,
durmir_0.01087335,
sentárase_0.01054715
```

Para comprobar la precisión de los resultados de la extracción en función de la estrategia de optimización utilizada, en un primer experimento analizamos sólo la corrección de las traducciones que NATools ofrece como primera opción (T1), marcando como *incorrectas* todas aquellas que aparecían como `(null)` o vacías (alineamientos léxicos 1:0), y marcando como *posibles* los casos en que NATools ofrece como traducción una palabra que formaría parte de la expresión pluriléxica (locución, perífrasis, etc.), como en las opciones de traducción primera (*novo* es parte de la traducción correcta *de novo*) y segunda (*outra* es parte de la traducción correcta *outra vez*) de la siguiente entrada:

(11)
```
again_182 ->
novo_0.47650307,
outra_0.35298964,
volvín_0.04795518,
volveu_0.04691147,
repetiu_0.01373287,
volve_0.00785520,
estábame_0.00679777,
virou_0.00560301
```

La evaluación se realiza primero sobre los resultados obtenidos a partir del corpus preeditado sin etiquetas morfosintácticas y, como ya se dijo, teniendo en cuenta únicamente la corrección de las traducciones que NATools ofrece como primera opción (T1). Los resultados de esta evaluación de la precisión de la extracción léxica, realizada de manera manual sobre un conjunto de entradas aleatoriamente seleccionadas que representan alrededor del 5% del total de entradas bilingües inglés-gallego extraídas del corpus (28.384), se muestran en la Tabla 1, donde *prob* es el índice de probabilidad del alineamiento léxico entre la palabra del inglés (L) y la primera opción (T1) de traducción del galego, *frec* es la frecuencia absoluta en el corpus de la palabra en inglés (L), *cobrt* es la cobertura o porcentaje de palabras analizadas que abarcan las palabras en inglés en cada banda de probabilidad y frecuencia, *correc* es el porcentaje de T1 correctas, *posib* es el porcentaje de T1 posibles (en el sentido que se explica más arriba), y *prec* la precisión o porcentaje acumulado de T1 correctas y posibles.

prob	*frec*	*cobrt*	*correc*	*posib*	*prec*
1	>4	4%	94,1%	5,9%	100%
	4	1,9%	87,5%	0%	87,5%
	<4	0,9%	50%	0%	50%
<1 - >0,5	>4	29,0%	93,5%	1,6%	95,1%
	4	1,7%	57,1%	0%	57,1%
	<4	2,1%	33,3%	0%	33,3%
<=0,5	>4	45,0%	71,2%	5,2%	76,4%
	4	2,1%	0%	11,1%	11,1%
	<4	13,2%	28,6%	0%	28,6%

Tabla 1: Evaluación de T1 en el corpus preeditado sin etiquetas morfosintácticas

De estos resultados, parece derivarse que la *frontera de calidad* de la alineación automática, a nivel de L+T1, se encuentra en la intersección entre el valor de frecuencia de L mayor de 4 y el valor de probabilidad de T1 mayor de 0,5. Si seleccionamos para el diccionario sólo las parejas L+T1 que cumplan estos dos criterios, la precisión o fiabilidad de la extracción automática sería realmente alta (un 93,6 % de acierto, que se eleva al 95,7% con la incorporación de las traducciones marcadas como posibles). El problema de emplear como filtro automático de la extracción léxica este criterio combinado es que la cobertura o cantidad de léxico adquirido no es muy grande (en total abarca un 33% del total del léxico analizado).

A partir de esta constatación, consideramos preciso desglosar los resultados obtenidos para las palabras con una probabilidad menor o igual que 0,5, con el fin de encontrar un corte que nos permitiera aumentar significativamente el porcentaje de palabras adquiridas sin disminuir de manera inaceptable el grado de acierto. De este modo, hallamos que la relación entre la fiabilidad de los resultados y la cobertura léxica mejoraba asumiendo como valores de corte una frecuencia de L mayor de 4 y una probabilidad de T1 mayor de 0,3 pero diferente de 0,5 (Tabla 2). Con este criterio, la fiabilidad de la extracción automática sigue siendo bastante alta (un 87,5% de acierto, que aumenta al 91,4% con las traducciones posibles), mientras que la cantidad de léxico adquirido crece hasta llegar al 54,7% del total del léxico analizado). La razón por la que eliminamos los casos de probabilidad 0,5 para T1 es que, en todos ellos, aparece una segunda opción de traducción (T2) con una probabilidad del mismo valor, que dificulta enormemente el acierto automático.

prob	*frec*	*cobrt*	*correc*	*posib*	*prec*
<0,5 - >=0,4	>4	9,4%	77,5%	7,5%	85%
<0,4 - >=0,3	>4	12,3%	78,8%	5,8%	84,6%
<0,3	>4	23,1%	65,3%	4,1%	69,4%

Tabla 2: Evaluación de T1 para *L* >4 y *prob* <= 0.5

También consideramos interesante la adquisición de T2 mediante un sistema de filtrado de la extracción automática. En la evaluación de los resultados para T2 (Tabla 3), previa al establecimiento de su frontera de calidad, no se tiene en cuenta las entradas para las que el programa no ofrece más que una traducción, es decir, los casos en que T1 tiene probabilidad 1.

prob	*frec*	*cobrt*	*correc*	*posib*	*prec*
<0,5 - >=0,4	>4	1,0%	50%	0%	50%
	=<4	0,5%	50%	0%	50%
<0,4 - >=0,3	>4	4,9%	73,7%	10,5%	84,2%
	=<4	3,6%	7,1%	0%	7,1%

<0,3	>4	74%	50,3%	2,4%	52,8%
	=<4	13,6%	28,3%	0%	28,3%

Tabla 3: Evaluación de T2 en el corpus preeditado sin etiquetas morfosintácticas

A partir de los resultados de la Tabla 3, decidimos aplicar a la selección de T2 el mismo criterio que aplicamos para T1, es decir, escoger los pares L+T2 con un valor de *frec* superior a 4 y un valor de *prob* mayor o igual que 0,3 pero diferente de 0,5. Con este filtro, la fiabilidad de la selección es del 69,6% si contamos sólo las traducciones claramente correctas, y del 78,3% si contamos también las marcadas como posibles, si bien el porcentaje del léxico abarcado por este conjunto de palabras es realmente pequeño (5,9%).

Finalmente, para comprobar la rentabilidad de la incorporación de información morfosintáctica al corpus paralelo, en lo que respecta a los resultados de la extracción léxica, evaluamos la precisión de T1 en la versión etiquetada del corpus (Tabla 4).

En comparación con los resultados obtenidos en la evaluación del corpus paralelo preeditado sin etiquetas morfosintácticas, los resultados son algo mejores con el corpus etiquetado. Asumiendo una frontera de calidad semejante a la adoptada para T1 y T2 en la versión no etiquetada, es decir, una frecuencia de L superior a 4 y una probabilidad de T1 mayor o igual que 0,3 pero diferente de 0,5, la extracción léxica automática obtiene un 89,7% de resultados correctos (casi 2 puntos por encima de los resultados obtenidos con el corpus no etiquetado) y un 93,9% de fiabilidad (2,5 puntos por encima) contando las traducciones anotadas como posibles, si bien la cobertura del léxico fiable (50,2%) se ve sensiblemente disminuida.

prob	*frec*	*cobrt*	*correc*	*posib*	*prec*
1	>4	5,9%	92,0%	8,0%	100,0%
	4	0,5%	100,0%	0,0%	100,0%
	<4	2,8%	33,3%	0,0%	33,3%
<1 - >0,5	>4	28,1%	93,3%	3,4%	96,6%
	4	0,9%	75,0%	0,0%	75,0%
	<4	2,6%	27,3%	0,0%	27,3%
0,5	>4	0,0%	0,0%	0,0%	0,0%
	4	0,0%	0,0%	0,0%	0,0%
	<4	1,7%	28,6%	0,0%	28,6%
<0,5- >=0,4	>4	6,8%	93,1%	3,4%	96,6%
	4	0,9%	75,0%	0,0%	75,0%
	<4	0,0%	0,0%	0,0%	0,0%
<0,4 - >=0,3	>4	9,4%	75,0%	5,0%	80,0%
	4	0,2%	0,0%	0,0%	0,0%
	<4	3,5%	26,7%	0,0%	26,7%
<0,3	>4	27,6%	64,1%	5,1%	69,2%
	4	1,9%	25,0%	0,0%	25,0%
	<4	7,1%	13,3%	0,0%	13,3%

Tabla 4: Evaluación de T1 en el corpus preeditado con etiquetas morfosintácticas

6 Conclusiones

Los resultados de la extracción léxica bilingüe automática a partir de corpus paralelos pueden optimizarse utilizando diversas técnicas. En esta comunicación hemos explorado algunas estrategias para tratar de superar las limitaciones impuestas por la extracción léxica cuando ésta se basa únicamente en los alineamientos oracionales. Las estrategias de optimización que presentamos están basadas, por una parte, en la codificación en el corpus paralelo de la información traductológica relativa a las asimetrías de traducción (alineamientos no biunívocos y alteraciones de orden en la traducción) y, por otra parte, en la preedición del corpus paralelo mediante la eliminación de palabras gramaticales muy frecuentes y de los segmentos de texto que están marcados en el corpus paralelo como omisiones o como adiciones y, por tanto, no tienen una correspondencia traductológica precisa. Estos elementos son eliminados por tratarse de una fuente de ruido importantísima y con una incidencia directa en los errores de la extracción. Adicionalmente, exploramos las posibilidades de optimización de la extracción léxica mediante la incorporación de información morfosintáctica en el corpus paralelo.

Asimismo, investigamos la posibilidad de crear filtros automáticos para cribar los resultados de la extracción léxica automática, eliminando del diccionario bilingüe generado los candidatos de traducción menos fiables. Estos filtros automáticos se elaboran mediante criterios inferidos de la evaluación de los resultados de la extracción automática utilizando las optimizaciones anteriormente desarrolladas. Como regla general, concluimos que el mejor criterio de selección para la

generación de diccionarios a partir de los resultados de la extracción léxica automática es un filtro que combina la frecuencia del lema (superior a 4) con la probabilidad de su traducción más probable (mayor o igual que 0,3 pero diferente de 0,5). Aplicando este criterio, que amalgama precisión y cobertura, la fiabilidad de las entradas generadas alcanza el 93,9% (con el 50,2% de cobertura) en el caso de los corpus paralelos con etiquetas morfosintácticas, y el 91,4% (con el 54,7% de cobertura) en el caso de los corpus paralelos no etiquetados morfosintácticamente.

El esfuerzo que supone etiquetar morfosintácticamente un corpus paralelo puede no verse compensado por un aumento en la precisión de la extracción léxica bilingüe proporcional a tal esfuerzo, si bien es cierto que los beneficios del etiquetado categorial acompañado de lematización permite efectuar procesamientos automáticos que facilitan ciertos aspectos de la creación de diccionarios bilingües basados en corpus. En este sentido, pretendemos seguir explorando las posibilidades de explotación de la incorporación de información lingüística en los corpus paralelos en diversas aplicaciones de PLN.

Bibliografía

Aguirre, J.L., A. Álvarez Lugrís y X. Gómez Guinovart. 2003. Aplicación do etiquetario morfosintáctico do SLI ó corpus de traduccións TECTRA. *Viceversa*, 7-8: 189-212.

Brants, T. 2000. TnT: A Statistical Part-of-Speech Tagger. *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*.

Brown, R.D., J.G. Carbonell y Y. Yang. 2000. Automatic dictionary extraction for cross-Language Information Retrieval. En J. Véronis (ed.), *Parallel Text Processing: Alignment and Use of Translation Corpora*. Kluwer, Dordrecht, pp. 275-298.

Erjavec, T. 2002. Compiling and Using the IJS-ELAN Parallel Corpus. *Informatica*, 26: 299-307.

Hiemstra, D. 1998. Multilingual Domain Modeling in Twenty-One. Automatic Creation of a Bi-directional Translation Lexicon from a Parallel Corpus. En *Proceedings of the 8th CLIN Meeting*, pp. 41-58.

Knight, K. 1997. Automating Knowledge Acquisition for Machine Translation. *AI Magazine*, 18(4): 81-96.

Och, F.J. y H. Ney 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1): 19-51.

Savourel, Y. *ed.* 2004. *TMX 1.4b Specification*. Localisation Industry Standards Association. [http://www.lisa.org/tmx/tmx.htm]

Simões, A.M. y J.J. Almeida. 2003. NATools: A Statistical Word Aligner Workbench. *Procesamiento del Lenguaje Natural*, 31: 217-224.

Tiedemann, J. 2003. *Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. Acta Universitatis Upsaliensis, Upssala.

Turcato, D. y F. Popowich. 2001. What is Example-Based Machine Translation?. En *Proceedings of the Workshop on Example-Based Machine Translation (MT Summit VIII).*

Vintar, Š. 2001. Using Parallel Corpora for Translation-oriented Term Extraction. *Babel Journal*, 47(2): 121-132.