Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 103-110 recibido 22-04-2005; aceptado 01-06-2005

El tratamiento de la polisemia en la extracción de léxicos bilingües a partir de corpora paralelos *

Pablo Gamallo Otero
Dept. de Língua Espanhola
Univ. de Santiago de Compostela
pablogam@usc.es

Susana Sotelo Docío
Dept. de Língua Espanhola
Univ. de Santiago de Compostela
fesdocio@usc.es

Resumen: Este artículo propone un método de extracción de equivalentes léxicos de traducción a partir de corpora paralelos alineados a niveles superiores a la oración y etiquetados morfosintácticamente. La estrategia que seguiremos toma en cuenta el efecto discriminante y desambiguador que el contexto local ejerce sobre un elemento léxico. Dada una palabra de la lengua fuente y un tipo de contexto particular, el método propuesto aprende la traducción de la palabra en ese tipo de contexto.
Palabras clave: corpora paralelos, extracción de léxico multilingüe, traducción

Abstract: The paper proposes a method to extract translation equivalents from parallel corpora which were POS tagged and aligned using very large segments. The strategy we use takes into account the ability of contexts to select senses of words. Given a word of the source language and a particular type of context, we learn its word translation within an equivalent context.
Keywords: parallel corpora, multilingual lexical extraction, translation

1. Introducción

Los corpora paralelos pueden ser vistos como enormes depósitos de información bilingüe, mucha de ella de tipo léxico. En la actualidad, existe una gran cantidad de métodos de extracción de léxicos bilingües a partir de corpora paralelos, en su mayoría con alineamento a nivel de la oración. (Gale y Church, 1991; Melamed, 1997; Ahrenberg, Andersson, y Merkel, 1998; Tiedemann, 1998; Vintar, 2001; Kwong, Tsou, y Lai, 2004; Guinovart y Fontenla, 2004). Estos métodos presentan, al menos, dos dificultades. En primer lugar, a pesar de existir una gran variedad de sistemas automáticos de alineamento a nivel de la oración, los resultados son mediocres cuando los textos ofrecen fronteras poco claras en lo que se refiere a la identificación del fin de la oración. Las dificultades de alineamento se multiplican cuando el texto fuente o la traducción se presentan con repetidas elipsis, imprecisiones o reordenamientos (Fung y McKeown, 1996). En segundo lugar, muchos de los métodos de extracción citados se centran exclusivamente en generar correlaciones monosémicas, es decir cada expresión de la lengua fuente sólo tiene una traducción, y cada expresión de la lengua término es usada para traducir una única expresión-fuente. Tales correlaciones son generadas por un algorimo que va eliminando en ciclos sucesivos todas las expresiones que compiten por servir de traducción, quedándose al final con la más probable para cada palabra de la lengua fuente (Melamed, 1997). Dada la naturaleza excluyente de este algoritmo, se hace imposible el tratamiento de la polisemia léxica.

Con el objetivo de evitar la primera dificultad, este artículo propone un alineamento basado, no en la identificación de fin de oraciones, sino en la detección de fronteras naturales que delimitan las partes internas del corpus. La mayoría de los corpora contienen fronteras naturales que, de manera explícita, separan partes básicas del texto, tales como capítulos, artículos periodísticos, resúmenes, recetas, documentos legales, intervenciones de personajes o personas jurídicas, cartas, etc. Nuestro propósito es el de utilizar estos segmentos explícitos para alinear los corpora paralelos. Con respecto a la segunda dificultad (i.e., la polisemia), proponemos un método que asigna traducciones contextuales a las palabras de la lengua fuente. Para llevar a cabo esta estrategia, nuestro método trata de identificar primero correspondencias bilingües entre contextos capaces de discriminar significados. Una vez identificadas estas correspondencias, nuestro objetivo es usarlas para extraer traducciones contextual-

* Este trabajo ha sido subvencionado por el Ministerio de Educación y Ciencia a cargo del proyecto Gari-Coter, ref: HUM2004-05658-D02-02

izadas. En concreto, una palabra de la lengua fuente será asociada a otra palabra de la lengua término en un contexto discriminante específico. Esto nos permitirá vincular una palabra a varias traducciones, siendo cada traducción adecuada a un tipo de contexto. La precisión del sistema alcanza el 89 %. Aunque todavía estamos lejos del 99 % al que se puede llegar usando una extracción monosémica (Melamed, 1997), el resultado, dada las dificultades de la tarea, es ciertamente esperanzador.

Este artículo se organiza como sigue. Primero, la sección 2 describe lo que llamamos *alineamento natural*. Después, la sección 3 introduce la medida estadística que utilizamos para calcular las correlaciones de traducción entre pares de expresiones así como las correlaciones entre pares de contextos. La sección 4 define la noción de *contexto discriminante*. En la sección 5, describimos el algoritmo de extracción usado. Y finalmente, en la sección 6, describimos un protocolo de evaluación de los resultados obtenidos.

2. Pre-procesamiento y alineamiento natural

Nuestro método requiere que los textos paralelos sean pre-procesados hasta el nivel morfosintáctico, es decir, segmentados, lematizados y etiquetados morfosintácticamente. Las experiencias se realizaron sobre textos en inglés, español y francés. Para el preprocesamiento de los textos utilizamos *software* libre disponible en la Red. Los textos ingleses y franceses fueron pre-procesados con el sistema TreeTagger (Schimd, 2002). Por su parte, los textos españoles fueron tratados con Freeling (Carreras et al., 2004). Como no efectuamos ninguna corrección al resultado del pre-procesamiento, el sistema de extracción del léxico bilingüe hereda forzosamente los errores provenientes del preprocesamiento. Por último, usamos un escaso número de reglas de tipo *"pattern-matching"*, a fin de identificar y extraer posibles contextos discriminantes en forma de *plantillas* léxico-sintácticas. En la sección 4, explicaremos la noción de contexto discriminante.

Una vez hecho esto, alineamos los textos fuente y término sirviéndonos de un proceso de detección de fronteras naturales. Llamamos fronteras naturales a aquellas marcas que organizan los textos en partes visibles y que tienen que ser conservadas en toda traducción si se quiere preservar la estructura interna del original. Fronteras naturales serían capítulos de novelas, artículos de periódicos, réplicas de obras teatrales, etc. Dejamos fuera de esta clasificación a los párrafos porque en todos los corpus que observamos existen numerosísimos casos de traducciones que no conservan la misma organización en párrafos que el texto fuente. El alineamento en párrafos exige una corrección manual tan costosa y penosa como el alineamento a nivel de la oración. Las experiencias evaluadas en nuestro trabajo se realizaron sobre varios corpora paralelos. Uno de ellos es un corpus inglés-español extraído de las actas del Parlamento Europeo (corpus *PE*). Estos textos se dividen en intervenciones de representantes políticos. Cada intervención, vista como una frontera natural de las actas, representa un segmento de alineamento. También hemos realizado experiencias con una novela, el Quijote, segmentada en capítulos, y con un corpus inglés-francés formado por textos de la Comisión Europea (corpus *CE*), que fue dividido en artículos jurídicos y directivas (Gamallo, 2005). En el corpus PE, detectamos más de 3.000 intervenciones que fueron usadas como segmentos para alinear el corpus. El principal problema de este tipo de alineamento es que los segmentos seleccionados pueden ser excesivamente largos. Sin embargo, el alineamento natural tiene una ventaja incuestionable. A pesar de que aparezcan asimetrías (omisiones o añadidos) en los textos paralelos, esto no influye para el correcto alineamento de todo el corpus. Es decir, con el alineamento natural, no necesitamos recurrir al penoso trabajo de la corrección manual.

3. Una medida de semejanza adaptada al alineamento natural

Para estimar la probabilidad que tiene una expresión de la lengua término de ser una traducción de una expresión de la lengua fuente necesitamos contar la frecuencia de aparición de las dos expresiones en todo el corpus por separado, así como la frecuencia de coaparición en cada uno de los segmentos alineados. Utilizamos una versión del coeficiente Dice para medir la correlación entre dos tipos de expresiones. Dada una expresión fuente e_1 y su traducción candidata e_2, nuestra versión adaptada del coeficiente Dice se define como sigue:

$$Dice(e_1, e_2) = \frac{2F(e_1, e_2)}{F(e_1) + F(e_2)}$$

donde

$$F(e_1, e_2) = \sum_i min(f(e_1, s_i), f(e_2, s_i))$$

y

$$F(e_n) = \sum_i f(e_n, s_i)$$

$f(e_1, s_i)$ representa la frecuencia de la expresión e_1 en el segmento s_i. A diferencia de la mayoría de los trabajos sobre extracción de léxicos bilingües, consideramos que la frecuencia de aparición de una expresión en un segmento particular es una información muy significativa. Dado que los segumentos que utilizamos para alinear el corpus son más largos que los usados habitualmente, suele ocurrir que la misma expresión tenga varias ocurrencias en uno uno solo. De esta manera, una expresión de la lengua término, e_2, está fuertemente correlacionada con una expresión de la lengua fuente, e_1, si ambas tienden a tener las mismas frecuencias en cada segmento s_i. La toma en cuenta de las frecuencias por segmento es una diferencia fundamental con respecto a los enfoques booleanos estándar. En la mayoría de esos trabajos, la correlación entre dos expresiones se establece si ambas tienden a aparecer en los mismos segmentos. Lo que interesa es si una expresión aparece o no aparece en un segmento dado, y no el número de veces (Smadja, McKeown, y Hatzivassiloglou, 1996). Sin embargo, como en nuestros segmentos naturales muchas expresiones diferentes pueden aparecer en todos los segmentos, necesitamos un rasgo más informativo que la simple presencia o no, es decir, necesitamos contar el número de coapariciones por segmento.

4. Contextos discriminantes

La contribución más importante de este artículo se encuentra en la elaboración de un método de extracción de léxico bilingüe que utiliza contextos discriminadores de sentidos. Tomemos como ejemplo las dos expresiones inglesas siguientes:

1. vehicle registration
2. registration of the notification

La expressión (1) se traduce en español como "matrícula del vehículo", mientras que la (2) se traduce "registro de la notificación". En (1), el modificador "vehicle" se comporta como un contexto discriminante que selecciona un sentido específico de "registration": la matrícula del vehículo. Este sentido se diferencia claramente del seleccionado por "of the notification" en la expresión (2), donde "registration" denota una acción particular. Los dos contextos discriminantes de "registration" en las expresiones (1) y (2) pueden representarse por medio de estas dos plantillas léxico-sintácticas:

< vehicle [NOUN] >
< [NOUN] of the notification >

Estos contextos contienen la información suficiente para poder discriminar el sentido de las palabras que aparecen en ellos. Nótese que usamos plantillas léxico-sintácticas para representar los contextos discriminantes. Cada plantilla de este tipo puede ser vista como el resultado de abstraer una de las dos expresiones participantes en una relación sintáctica binaria de tipo Núcleo-Modificador. Dada la siguiente dependencia binaria:

of (registration, notification)

donde "registration" es el núcleo y "notification" el modificador, generamos dos plantillas léxico-sintácticas:

< registration of [NOUN] >
< [NOUN] of the notification >

Cada una de estas plantillas representa un contexto potencialmente discriminante. La formación de contextos discriminantes sigue el principio de la co-restricción (Gamallo, Agustini, y Lopes, 2005). Este principio estipula que no sólo el modificador puede discriminar el sentido del núcleo sino que también éste puede seleccionar un sentido particular del modificador. El algoritmo de extracción que describiremos en la siguiente sección parte de la idenficación automática de correlaciones bilingües entre contextos discriminantes.

5. Descripción del método de extracción

El método de extracción se divide en dos etapas. Primero, se extraen correlaciones bilingües entre contextos discriminantes. Y segundo, se utilizan los pares de contextos aprendidos en la primera etapa para extraer correlaciones entre palabras potencialmente polisémicas.

5.1. Etapa 1: extracción de correlaciones bilingües entre contextos discriminantes

El primer paso de nuestro método es identificar y extraer pares bilingües de contextos discriminantes. Para ello, identificamos dependencias binarias e inducimos todas las plantillas léxico-sintácticas posibles. Recuérdese que las plantillas léxico-sintácticas representan contextos discriminantes. El cuadro 1 muestra algunos pocos ejemplos representativos de los diferentes tipos de dependencias que manejamos, así como de sus correspondientes plantillas. En (Gamallo, Agustini, y Lopes, 2005), explicamos en pormenor el método usado para identificar las dependencias binarias.

Nótese que *lobj* y *robj* representan dependencias verbales: el objeto a la izquierda del verbo (left_object) y el objeto a la derecha (right_object). Por último *modAdj* y *modN* simbolizan respectivamente la modificación adjectival y nominal.

Una vez identificadas las plantillas de los textos fuente y término, calculamos el coeficiente Dice de todos los pares y escogemos, para cada plantilla de la lengua fuente, su más probable traducción siempre que sobrepase un determinado valor (>0.4). En los cuadros 2, 3 y 4 aparecen algunos ejemplos de correlaciones bilingües entre plantillas que fueron extraídas del corpus PE.

Es evidente que estas correlaciones de plantillas léxico-sintácticas se asemejan mucho a las llamadas "plantillas de traducción" (*translation templates*) dentro del enfoque *Example-Based Machine Translation* (EBMT) (Carl, 1999). Sin embargo, a diferencia de la estrategia inductiva y supervisada en EBMT, nuestro método no induce las plantillas a partir de ejemplos de traducción. Lo que proponemos aquí es un método de aprendizaje no supervisado. Las correlaciones entre plantillas se adquieren automáticamente por medio de la selección de los pares con mejor coeficiente Dice.

En la segunda etapa del algoritmo de extracción, utilizaremos las plantillas como contextos lingüísticos más precisos y apropiados para aprender correlaciones entre palabras.

5.2. Etapa 2: extracción de correlaciones entre palabras

Dividimos esta segunda etapa en dos procesos: la extracción de correlaciones entre palabras monosémicas y la extracción de correlaciones entre palabras potencialmente polisémicas. Los pares de plantillas aprendidos en la etapa anterior serán utilizados en la segunda extracción. Sólo serán consideradas las palabras con frecuencia absoluta > 2.

Extracción de monosémicas: Cuando una palabra y su posible traducción superan el 0,8 de correlación (medida con el coeficiente Dice), entonces inferimos que las dos palabras comparadas tienen un comportamiento monosémico a lo largo del corpus que sirvió de base a la comparación. De aquí se deduce que los índices de correlación más altos dan lugar a equivalentes de traducción exentos de ambigüedad. Al igual que en el "enfoque de correlaciones competitivas" (*competitive linking approach*) propuesto en (Melamed, 1997), las correlaciones que superen el nivel de semejanza arriba indicado van a ser eliminadas del espacio de búsqueda antes de comenzar el segundo proceso. Nuestro proceso de extracción de pares monosémicos no introduce, por tanto, ningún tipo de método innovador. La contribución principal de nuestro trabajo está en la manera cómo extraemos palabras polisémicas.

Extracción de polisémicas: El espacio de búsqueda del segundo proceso contiene ahora las palabras que no han sido consideradas monosémicas en el proceso anterior y que pueden instanciar, al menos, uno de los pares de plantillas extraídos en la etapa 1. El proceso de extracción opera de la siguiente manera: cada palabra no monosémica que instancia una plantilla de la lengua fuente es comparada con todas las palabras que instancian la plantilla correspondiente de la lengua término. Seleccionamos de todas estas, aquella que tiene el coeficiente Dice más alto siempre que sobrepase un determinado valor (e.g., >0.3). Dado que las palabras candidatas a ser traducciones se escogen entre aquellas que aparecen en el mismo contexto discriminante, nos es posible reducir

Dependencias binarias	Plantillas
of (import, sugar)	< import of [NOUN] > < [NOUN] of sugar >
robj (approve, law)	< approve [NOUN] > < [VERB] law >
lobj (approve, president)	< president, [VERB] > < [NOUN] approve >
modAdj(legal, document)	< legal [NOUN] > < [ADJ] document >
modN(area, protection)	< protection [NOUN] > < [NOUN] area >

Cuadro 1: Dependencias binarias y plantillas

Inglés	**Español**	**Dice**
< active [NOUN] >	< [NOUN] activo >	0,65
< african[NOUN] >	< [NOUN] africano >	0,65
< agricultural [NOUN] >	< [NOUN] agrícola >	0,63
< alarming [NOUN] >	< [NOUN] alarmante >	0,45
< albanian [NOUN] >	< [NOUN] albanés >	0,61

Cuadro 2: Correlaciones bilingües entre plantillas provenientes de dependencias nombre-adjetivo (extracto de las 5 primeras entradas de la lista con este tipo de plantillas)

Inglés	**Español**	**Dice**
< [NOUN] after year>	< [NOUN] tras año>	0,41
< crime against [NOUN] >	< crimen contra [NOUN] >	0,66
< fight against [NOUN] >	< lucha contra [NOUN] >	0,43
< violence against [NOUN] >	< violencia contra [NOUN] >	0,81
< [NOUN] against discrimination >	< [NOUN] contra discriminación >	0,50

Cuadro 3: Correlaciones bilingües entre plantillas provenientes de dependencias nominales con preposición (extracto de las 5 primeras entradas de la lista con este tipo de plantillas)

Inglés	**Español**	**Dice**
< africa [VERB] >	< África [VERB] >	0,66
< agreement [VERB] >	< acuerdo [VERB] >	0,45
< agriculture [VERB] >	< agricultura [VERB] >	0,50
< aid [VERB] >	< ayuda [VERB] >	0,57
< alcohol [VERB] >	< alcohol [VERB] >	0,76

Cuadro 4: Correlaciones bilingües entre plantillas provenientes de dependencias verbales de tipo "left-object" (extracto de las 5 primeras entradas de la lista con este tipo de plantillas)

el nivel exigido de correlación. Con respecto a un contexto discriminante, basta con sobrepasar el 30 % de semejanza para que podamos considerar que existe correlación entre la palabra fuente y su posible traducción. Por consiguiente, consideramos que una palabra de la lengua término es susceptible de ser una traducción contextualmente motivada de una palabra fuente si ambas instancian un par de plantillas correlacionadas y si su índice de semejanza supera el 0,3. Por otro lado, como las comparaciones se realizan en diversas plantillas léxico-sintácticas (es decir, en diferentes contextos), una misma palabra puede correlacionarse con varias de la otra lengua. Es de esta manera que podemos proponer traducciones contextualizadas de una palabra polisémica. A diferencia de los métodos estándar, el nuestro nos permite generar correlaciones correctas de una-a-varias palabras. En algunos casos conseguimos hasta tres traducciones diferentes de una palabra.

Veamos un ejemplo. Nuestro sistema aprende que la palabra inglesa "area" no tiene ninguna posible traducción con un índice de correlación superior a 0,8. Por tanto, el primer proceso no la selecciona como una palabra monosémica. Pasamos entonces al segundo proceso. Aquí descubrimos que "area" instancia plantillas como: < forest [NOUN] > y < adopt in [NOUN] > (ya que en el corpus aparecen las expresiones "forest area" y "adopt in the area"). Además, descubrimos las siguientes correlaciones entre

plantillas:

<forest [NOUN]> <[NOUN] de bosque> 0.64
<adopt in [NOUN]> <aprobar en [NOUN]> 0.57

Por último, comparamos "area" con todas las palabras españolas que pueden instanciar las plantillas correspondientes y seleccionamos los índices más altos siempre que sobrepasen el valor mínimo exigido. Así, aprendemos que "area" se traduce como "zona" en el contexto <[NOUN] de bosque>, puesto que el índice de semejanza entre "area" / "zona" es el más alto en ese contexto: $0,39$. Pero también aprendemos que "area" se traduce como "ámbito" en el contexto <aprobar en [NOUN]>, puesto que ambas palabras obtienen el mayor índice de correlación, 0,36, de ese contexto. Por consiguiente, nuestro sistema consigue aprender que "area" tiene, al menos, dos traducciones posibles, "zona" y "ámbito", y aprende también en qué tipo de contexto discriminante aparece la traducción.

5.3. Polisemia y asociaciones indirectas

Según Melamed, el modelo monosémico es el más eficaz contra los problemas creados por las asociaciones indirectas. Son asociaciones indirectas correlaciones como "christiam" y "demócrata", o "amsterdam" y "tratado". Estas asociaciones tienen un índice elevado de semejanza debido a que estas palabras aparecen casi siempre en colocaciones como "demócrata cristiano" o "tratado de amsterdam". La solución propuesta por el modelo monosémico es escoger sólo la mejor correlación (Melamed, 1997) y eliminar todas las demás. Obviamente, esta solución impide el tratamiento de la polisemia. Nuestro modelo polisémico, en cambio, al efectuar comparaciones en contextos discriminantes, elimina de manera natural las asociaciones indirectas. Así por ejemplo, la palabra inglesa "amsterdam" no será comparada con la palabra española "tratado" puesto que aparecen en contextos discriminantes diferentes. Al nunca ser comparadas una con otra, se impide que pueda escogerse "tratado" como posible traducción de "amsterdam".

6. Experimentos y evaluación

Realizamos tres experimentos de extracción sobre tres corpus diferentes: un corpus inglés-francés extraído de la legislación de la Comisión Europea (CE) que contiene sobre 2 millones de ocurrencias de palabras y 1.050 segmentos, y donde cada segmento es una directiva o artículo de ley; un corpus inglés-español extraído de las actas del parlamento europeo (PE) con 1 millón de ocurrencias y 3.000 segmentos, y donde cada segmento se corresponde con una intervención de un agente político; un corpus formado por las versiones inglesa y española del Quijote, consituído por unas 460 mil ocurrencias y 350 segmentos, y donde cada segmento es un capítulo de la novela. El tamaño medio de los segmentos generados es el siguiente: las directivas de CE forman segmentos alineados de unos 20 Kbytes, las intervenciones de PE son en media de 2 Kbytes, y los capítulos del Quijote oscilan entre 10 y 20 Kbytes. Esto quiere decir que la alineación natural puede dar lugar a segmentos que sobrepasan las 2.000 palabras.

El algoritmo de extracción generó tres diccionarios bilingües, uno para cada corpus: a partir de CE, generamos un diccionario de 2.232 lemas (nombres y adjetivos) y 1.797 plantillas léxico-sintácticas; a partir de PE, se generó un diccionario de 2.368 lemas (nombres, verbos y adjetivos) y 2.551 plantillas; y a partir del Quijote, extraímos 1.161 lemas (nombres, verbos y adjetivos) y 1.270 plantillas[1].

A fin de evaluar nuestro método de extacción, utilizamos en cada experimento un test de 600 ocurrencias de palabras seleccionadas aleatoriamente. Cada test se divide en 4 partes: 150 ocurrencias de todas las categorías, 150 sólo con nombres, 150 sólo con adjetivos y 150 sólo con verbos. Para facilitar la tarea a los evaluadores, cada palabra seleccionada se acompaña de su contexto inmediato. El cuadro 5 muestra los resultados de la evaluación. Llamamos *precisión* al número de traducciones correctas que fueron propuestas por el sistema dividido por el número total de traducciones sugeridas. *Recall* es el número de traducciones correctas propuestas por el sistema dividido por el número de todas las ocurrencias del test. La primera línea del cuadro refleja los valores obtenidos en los tests donde se utilizaron palabras de las tres

[1]Los diccionarios generados pueden ser cargados libremente en: http://gramatica.usc.es/~gamallo/dicos.htm.

Cat.	**Precisión (%)**			**Recall (%)**		
	PE	**EC**	**Quij**	**PE**	**EC**	**Quij**
todos	89	82	77	76	68	53
nombres	95	83	79	80	72	55
adjs	90	79	58	58	61	26
verbos	86	-	77	64	-	59

Cuadro 5: Evaluación de los resultados en tres corpora: PE, CE y Quijote.

categorías[2]. La precisión alcanzada oscila entre el 89 %, para el corpus PE, que fue dividido en 3.000 segmentos, y 77 % para el Quijote, dividido en apenas 300 segmentos. Entre ambos, se sitúa CE, con 82 % y dividido en 1,050 segmentos. La diferencia de *recall* entre los tres corpora mantiene proporciones análogas. Parece existir, por tanto, una relación fuerte entre número de segmentos alineados y calidad de extracción. Estos experimentos tienden a mostrar que, para nuestro método de extracción, el número de segmentos alineados es un factor más importante que el tamaño de esos segmentos. Si esto es así, el alineamento por segmentos naturales (capítulos de novela, cartas, artículos, etc.) es una opción tan válida como el alineamento por oraciones. No es el tamaño de los segmentos (20 o 2.000 palabras) lo que determina la calidad de nuestra extracción, sino la cantidad de segmentos alineados. Cuando el número de segmentos comienza a ser significativo, como en el caso del corpus PE, la precisión que se obtiene, 89 %, ya es comparable con la de muchos de los trabajos en este área basados en alineamento a nivel de oraciones. Además, la ventaja que tiene trabajar con segmentos grandes es que permite manipular corpus más voluminosos y así tener mayor cobertura.

Sin embargo, esta evaluación no nos permite aun medir de forma exacta la influencia de los contextos discriminantes en el proceso de extracción. Para efectuar esta medición, observamos por un lado los resultados obtenidos en contextos discriminantes y por otro, qué traducciones se obtendrían si no se tienen en cuenta estos contextos (experiencia que nos sirve de *baseline*). El cuadro 6 muestra que el grado de precisión de las traducciones propuestas es mucho más alto cuando se utilizan contextos discriminantes que cuando se proponen independientemente del contexto. En particular, si tomamos como referencia los pares de palabras con un índice de correlación no muy alto, entre $0,3$ y $0,4$, el número de pares correctos sin contexto es muy bajo: apenas llega al 25 %. En cambio, cuando los pares con este índice se extraen a partir de contextos discriminantes, el grado de precisión alcanza el 78 %. Como la palabra fuente y su posible traducción están obligadas a co-aparecer en, al menos, un par de contextos discriminantes equivalentes, nuestro sistema es capaz de aprender equivalentes de traducción utilizando valores de semejanza relativamente bajos. Al poder bajar los índices de correlación admisibles, el sistema consigue un *recall* más alto, es decir, mayor cobertura de traducción. Además, dado que podemos aceptar traducciones con índices de correlación bajos, estamos en condiciones de proponer, sin perder demasiada precisión, que cada palabra polisémica tenga varias traducciones posibles.

Dice	**Precisión (%)**	
	+cntx	−cntx
>0,3 - <0,4	78	25
>0,5 - <0,6	97	78
>0,7 - <0,8	99	95

Cuadro 6: Precisión con contextos discriminantes y sin contextos.

El cuadro 6 muestra cómo a medida que subimos el índice Dice de correlación, la extracción en contexto no aporta casi ninguna mejora con respecto a la extracción sin contexto. De ello se deduce que las palabras que participan en correlaciones con más de 0,8 de semejanza tienen un comportamiento tendente a la monosemia. Esto justifica que se pueda y deba realizar el primer proceso de la etapa 2 del algoritmo (es decir, la extracción de palabras monosémicas) sin tomar en cuenta los contextos discriminantes.

Además de estos dos tipos de experiencias, contamos realizar en brebe experimentos donde se utilicen un número muy superior de plantillas provenientes de diferentes corpus. Esto nos permitirá saber si es posible mantener la precisión y mejorar la cobertura con el aumento de plantillas independientes del dominio en cuestión.

[2]En el caso del corpus CE no fueron extraídos los verbos.

7. Conclusiones y trabajo futuro

Hemos descrito un método enfocado en la generación de equivalentes de traducción a partir de corpora bilingües, previamente alineados por medio de las fronteras naturales presentes en todo tipo de texto. Los segmentos alineados pueden llegar a ser documentos de más de 80 ou 90 Kbytes. Para poder llevar a cabo la extracción de equivalencias entre expresiones léxicas en segmentos de tales dimensiones, utilizamos una medida de semejanza que toma en cuenta la frecuencia de aparición de las expresiones en cada segmento. El objetivo es desarrollar métodos de extracción de léxico bilingüe sin recurrir al alineamento a nivel de la oración. Podemos pasar directamente del alineamento natural al alineamento a nivel de palabra sin pasar por la discutida oración.

Por otro lado, nuestro método da cuenta de la polisemia gracias al uso de contextos discriminantes. Cada palabra no se compara con el resto de palabras de su categoría, sino con aquéllas que aparecen en contextos discriminantes correlacionados. Estas comparaciones contextualizadas son las que nos permiten aprender traducciones correctas, a pesar de tener índices de semejanza bajos. Así mismo, nuestro sistema es capaz de aprender diferentes traducciones contextualizadas de una palabra polisémica. Es esta la principal contribución de nuestro trabajo. Sin embargo, la principal desventaja de este método es que no está bien adaptado para identificar sentidos o usos minoritarios de las palabras en contextos de colocaciones. Por ejemplo, no permite descubrir que "heavy" se puede traducir como "fuerte" en la colocación "heavy rain".

Finalmente, como trabajo futuro, nos proponemos tratar algo que todavía no hemos abordado: las correlaciones entre multipalabras y palabras simples. También nos proponemos utilizar los pares bilingües de plantillas léxico-sintácticas como punto de partida para el aprendizaje de léxicos bilingües a partir de corpora no-paralelos.

Bibliografía

Ahrenberg, Lars, Mikael Andersson, y Magnus Merkel. 1998. A simple hybrid aligner for generating lexical correspondences in parallel texts. En *COLING-ACL'98*, páginas 29–35, Montreal.

Carl, Michael. 1999. Inducing translation templates for example-based machine translation. En *MT-Summit VII*.

Carreras, X., I. Chao, L. Padró, y M. Padró. 2004. An open-source suite of language analyzers. En *LREC'04*, Portugal.

Fung, Pascale y Kathleen McKeown. 1996. A technical word and term translation aid using noisy parallel corpora across language groups. *Machine Translation Journal*, páginas 53–87.

Gale, Willian y Kenneth Church. 1991. Identifying word correspondences in parallel texts. En *Workshop DARPA SNL*.

Gamallo, Pablo. 2005. Extraction of translation equivalents from parallel corpora using sense-sensitive contexts. En *10th Conference of the European Association on Machine Translation (EAMT'05)*, páginas 97–102, Budapest, Hungary.

Gamallo, Pablo, Alexandre Agustini, y Gabriel Lopes. 2005. Clustering syntactic positions with similar syntactic requirements. *Computational Linguistics*, 31(1).

Guinovart, Xavier G. y Elena Sacau Fontenla. 2004. Métodos de optimización de la extracción de léxico bilingue a partir de corpus paralelos. *Procesamiento del Lenguaje Natural*, 33:133–144.

Kwong, Oi Yee, Benjamin K. Tsou, y Tom B. Lai. 2004. Alignment and extraction of bilingual legal terminology from context profiles. *Terminology*, 10(1):81–99.

Melamed, Dan. 1997. A word-to-word model of translational equivalence. En *ACL'97*, Madrid, Spain.

Schimd, Helmut. 2002. Treetagger. En *A language independent part-of-speech tagger*, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

Smadja, F., K. McKeown, y V. Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons. *Computational Linguistics*, 22(1).

Tiedemann, Jorg. 1998. Extraction of translation equivalents from parallel corpora. En *11th Nordic Conference of Computational Linguistics*, Copenhagen, Denmark.

Vintar, Ŝ. 2001. Using parallel corpora for translation-oriented term extraction. *Babel Journal*, 47(2):121–132.