Un método de extracción de equivalentes de traducción a partir de un corpus comparable castellano-gallego *

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

José Ramom Pichel Campos
Dept. de Tecnologia Linguística da
Imaxin|Software
Santiago de Compostela, Galiza
jramompichel@imaxin.com

Resumen: Los trabajos sobre extracción de equivalentes de traducción a partir de corpus comparables no-paralelos no han sido muy numerosos hasta ahora. La razón principal radica en los pobres resultados obtenidos si los comparamos con los enfoques que utilizan corpus paralelos y alineados. El método propuesto en este artículo, basado en el uso de *contextos semilla* generados a partir de diccionarios bilingües externos, obtiene tasas de precisión próximas a los métodos con corpus paralelos. Estos resultados apoyan la idea de que la ingente cantidad de corpus comparables disponibles via Web puede llegar a ser una fuente importante de conocimiento lexicográfico. En este artículo, se describen los experimentos realizados sobre un corpus comparable castellano-gallego.
Palabras clave: extracción de léxico multilingüe, corpus comparables, traducción automática

Abstract: So far, research on extraction of word translations from comparable, non-parallel corpora has not been very popular. The main reason was the poor results when compared to those obtained from aligned parallel corpora. The method proposed in this paper, relying on *seed contexts* generated from external bilingual dictionaries, allows us to achieve results similar to those from parallel corpus. In this way, the huge amount of comparable corpora available via Web can be viewed as a never-ending source of lexicographic information. In this paper, we desbribe the experiments performed on a comparable, Spanish-Galician corpus.
Keywords: multilingual lexical extraction, comparable corpora, automatic translation

1. Introducción

En las dos últimas décadas, han aparecido numerosos trabajos centrados en la extracción automática de léxicos bilingües a partir de corpus paralelos (Melamed, 1997; Ahrenberg, Andersson, y Merkel, 1998; Tiedemann, 1998; Kwong, Tsou, y Lai, 2004). Estos trabajos comparten una estrategia común: organizan primero los textos en pares de segmentos alineados para luego, en base a este alineamento, calcular las coocurrencias de palabras en cada par de segmentos. En algunos de estos experimentos, la precisión alcanzada al nivel de la palabra es muy alta: alrededor del 90 % para un *recall* del 90 %. Desgraciadamente, no hay todavía disponible una gran cantidad de texto paralelo, especialmente en lo que se refiere a lenguas minorizadas. Para evitar este problema, en los últimos años se han desarrollado técnicas de extracción de léxicos bilingues a partir de corpus comparables no-paralelos. Estas técnicas parten de la idea de que la Web es un enorme recurso de textos multilingües fácilmente organizados en corpus comparables no-paralelos. Un corpus comparable no-paralelo (de aquí en adelante "corpus comparable") está formado por textos en dos lenguas que, sin ser traducciones unos de otros, versan sobre temáticas parecidas. Sin embargo, la tasa de precisión de tales métodos es todavía bastante inferior a la de los algoritmos de extracción de corpus paralelos. Los mejores registros hasta ahora apenas alcanzan el 72 % (Rapp, 1999), y ello, sin dar cuenta de la cobertura alcanzada.

En este artículo, proponemos un nuevo método de extracción de léxicos bilingües a partir de corpus comparables. Este método se basa en el uso de diccionarios bilingües

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

con el propósito de identificar correspondencias bilingües entre pares de contextos léxico-sintácticos. A parte de los diccionarios, se utilizará para el mismo propósito la identificación de cognados en los textos comparables. La extracción del léxico bilingüe se realizará tomando en cuenta las coocurrencias de lemas mono y multi-léxicos en los contextos bilingües previamente identificados. Los resultados obtenidos mejoran el 72 % de precisión para una cobertura del 80 %, lo que supone un avance en el área de la extracción en corpus comparables. Estos resultados apoyan la idea de que la ingente cantidad de corpus comparables disponibles via Web puede llegar a ser una fuente casi inagotable de conocimiento lexicográfico.

El artículo se organiza como sigue. En la sección 2, situaremos nuestro enfoque con respecto a otros trabajos relacionados. La sección 3 describirá con detalle las diferentes etapas del método propuesto. Seguidamente, en 4, analizaremos los experimentos realizados para un corpus castellano-gallego, y describiremos un protocolo de evaluación de los resultados. Acabaremos con una sección de conclusiones.

2. Trabajo relacionado

No existen muchos trabajos cuyo enfoque sea la extracción de léxicos bilingües en corpus comparables, en relación a los que usan textos paralelos y alineados. El método más eficiente, y en el que se basan la mayoría de los pocos trabajos en el área (Fung y McKeown, 1997; Fung y Yee, 1998; Rapp, 1999; Chiao y Zweigenbaum, 2002), se puede describir como sigue: la palabra o multipalabra w_1 es una traducción candidata de w_2 si las palabras que coocurren con w_1 dentro de una ventana de tamaño N son traducciones de las palabras que coocurren con w_2 dentro de la misma ventana. Esta estrategia se fundamenta, por tanto, en una lista de pares de palabras bilingües (llamadas *palabras semilla*), previamente identificadas en un diccionario bilingüe externo. En resumen, w_1 puede ser una traducción candidata de w_2 si ambas tienden a coocurrir con las mismas palabras semilla. El principal problema de este método es que, según la hipótesis de Harris (Harris, 1985), las ventanas de tamaño N son semánticamente menos precisas que los contextos locales de naturaleza léxico-sintáctica. Las técnicas más eficientes para la generación automática de relaciones semánticas (Grefenstette, 1994; Lin, 1998) no utilizan contextos definidos en forma de ventanas de palabras sino en forma de dependencias sintácticas. En este artículo, presentaremos un método de extracción de léxicos bilingües basado en la previa identificación de *contextos léxico-sintácticos bilingües*, y no en el uso de ventanas de *palabras semilla*, habitual en los trabajos más representativos del estado del arte.

Existen otros enfoques relacionados con la extracción de léxicos bilingües en corpus comparables que no requieren el uso de diccionarios externos (Fung, 1995; Rapp, 1995; Diab y Finch, 2001). Sin embargo, (Fung, 1995) obtiene resultados muy pobres lo que restringe enormemente sus potenciales aplicaciones, (Rapp, 1995) tiene graves limitaciones computacionales, y (Diab y Finch, 2001) sólo ha sido aplicado a corpus monolingües. Por último, cabe mencionar el enfoque descrito en (Gamallo y Pichel, 2005; Gamallo, 2007), que utiliza pequeños fragmentos de corpus paralelos como base para la extracción de contextos semilla.

3. Descripción de la estrategia

Nuestra estrategia se divide en tres etapas secuenciales: (1) procesamiento textual, (2) creación de una lista de contextos semilla por medio de la explotación de diccionarios bilingües y de la identificación de cognados, y (3) extracción de los equivalentes de traducción a partir de textos comparables usando como anclas la lista de contextos semilla.

3.1. Procesamiento del corpus comparable

En primer lugar, lematizamos, etiquetamos y desambiguamos morfosintácticamente el corpus comparable usando una herramienta de código abierto: Freeling (Carreras et al., 2004). En el proceso de etiquetación, se activa la identificación de nombres propios, que pueden ser mono y pluriléxicos. Una vez realizada esta tarea, se seleccionan potenciales dependencias sintácticas entre lemas con una estrategia básica de reconocimiento de patrones. Los determinantes son eliminados. Cada dependencia sintáctica identificada se descompone en dos contextos léxico-sintácticos complementarios. En el cuadro 1 se muestran algunos ejemplos. Dada una dependencia sintáctica identificada en el corpus, por

Dep. binarias	Contextos
de (venta, azúcar)	< venta de [NOUN] > < [NOUN] de azúcar >
robj (ratificar, ley)	< ratificar [NOUN] > < [VERB] ley >
lobj (ratificar, gobierno)	< gobierno [VERB] > < [NOUN] ratificar >
iobj_contra(luchar, pobreza)	< luchar contra [NOUN] > < [VERB] contra pobreza >
modAdj(entrenador, adecuado)	< [NOUN] adecuado > < entrenador [ADJ] >

Cuadro 1: Dependencias binarias y sus contextos léxico-sintácticos asociados.

ejemplo:
de (venta, azúcar) ,
extraemos dos contextos léxico-sintácticos: < venta de [NOUN] >, donde NOUN representa al conjunto de nombres que pueden aparecer después de "venta de", es decir, "azúcar", "producto", "aceite", etc., y por otro lado, < [NOUN] de azúcar >, donde NOUN representa el conjunto de nombres que pueden aparecer antes del complemento "de azúcar": "venta", "importación", "transporte", etc. La caracterización de los contextos se basa en la noción de *co-requerimiento* descrita en (Gamallo, Agustini, y Lopes, 2005). Además de las dependencias preposicionales entre nombres, también utilizamos la dependencia *lobj*, que representa la probable relación entre el verbo y el nombre que aparece inmediatamente a su izquierda (*left* object); *robj* es la relación entre el verbo y el nombre que aparece a su derecha (*right* object); *iobj_prp* representa la relación entre el verbo y un nombre precedido de preposición. Por último, *modAdj* es la relación entre un nombre y el adjetivo que lo modifica.

Los léxicos bilingües que nos proponemos extraer no sólo se componen de lemas monoléxicos y nombres propios, sino también de lemas multi-léxicos, es decir, de expresiones con varios lexemas y un cierto grado de cohesión: "accidente de tráfico", "cadena de televisión", "dar a conocer", etc. Para poder extraer este tipo de expresiones, realizamos una segunda fase del procesamiento que consiste en identificar lemas multi-léxicos (que no son nombres propios) y sus contextos. En esta tarea, utilizamos un extractor automático básico, basado en la instanciación de patrones morfo-sintácticos (e.g, NOUN-PRP-NOUN, NOUN-ADJ, VERB-NOUN, etc.) que nos permite identificar un gran número de candidatos. Este extractor se ejecuta en el corpus comparable, por tanto, obtenemos lemas multi-léxicos en las dos lenguas. Posteriormente, reducimos la lista de candidatos con un filtro estadístico elemental que sólo retiene aquellos candidatos con un grado de cohesión elevado (medida *SCP*). Seguimos una estrategia parecida a la descrita en (Silva et al., 1999). Una vez constituida la lista de lemas multi-léxicos, extraemos sus contextos léxico-sintácticos de forma análoga a la empleada arriba para los lemas mono-léxicos y los nombres propios.

3.2. Generación de contextos bilingües

La principal estrategia que utilizamos para la generación de contextos léxico-sintácticos bilingües se fundamenta en la explotación de diccionarios bilingües externos. Supongamos que en un diccionario castellano-gallego la entrada castellana "venta" se traduce en gallego por "venda", ambos nombres. La generación léxico-sintáctica a partir de cada uno de estos nombres se lleva a cabo siguiendo reglas básicas como por ejemplo: un nombre puede ir precedido de una preposición que a su vez es precedida de otro nombre o un verbo, puede ir después de un nombre o verbo seguidos de una preposición, o puede ir antes o después de un adjetivo. Hemos centrado la generación en tres categorías: nombres, verbos y adjetivos. Para cada categoría sintáctica, hemos generado únicamente un subconjunto representativo de todos los contextos generables. El cuadro 2 muestra los contextos generados a partir de la correspondencia bilingüe entre "venta" y "venda" y un conjunto limitado de reglas.

La generación se completa con la instanciación de *prp*. Para ello, empleamos una lista cerrada de preposiciones específicas y sus correspondientes traducciones. De esta ma-

Castellano	Gallego
<venta *prp* [NOUN]>	<venda *prp* [NOUN]>
<[NOUN] *prp* venta>	<[NOUN] *prp* venda>
<[VERB] venta>	<[VERB] venda>
<[VERB] *prp* venta>	<[VERB] *prp* venda>
<venta [VERB]>	<venda [VERB]>
<venta [ADJ]>	<venda [ADJ]>
<[ADJ] venta>	<[ADJ] venda>

Cuadro 2: Contextos bilingües generados a partir de la correlación "venta-venda".

nera, obtenemos pares de contextos bilingües como: <venta de [NOUN]> y <venda de [NOUN]>, <venta en [NOUN]> y <venda en [NOUN]>, etc.

Por otro lado, usamos otra estrategia complementaria, basada en la identificación de cognados en los textos comparables. Llamamos aquí cognados a 2 palabras en lenguas diferentes que se escriben de la misma manera. Sólo nos interesamos en aquellos que no se encuentran en el diccionario bilingüe, y que son, en su mayoría, nombres propios. Generamos los contextos léxico-sintácticos correspondientes y los juntamos a la lista de pares de contextos bilingües.

Los pares bilingües generados por medio de estas dos estrategias servirán de anclas o referencias para marcar el corpus comparable en el que se va a realizar la última etapa del proceso de extracción.

3.3. Identificación de equivalentes de traducción en el corpus comparable

La etapa final consiste en la extracción de equivalentes de traducción con ayuda de los pares de contextos bilingües previamente generados. Esta etapa se divide en dos procesos secuenciales: filtrado de contextos y extracción de los equivalentes de traducción.

3.3.1. Filtrado

Dada la lista de pares de contextos bilingües generados en la etapa anterior, procedemos a la eliminación de aquellos pares con un grado elevado de *dispersión* y *asimetría* en el corpus comparable. Un par bilingüe de contextos se considera disperso si el número de lemas diferentes que aparecen en los dos contextos dividido por el número total de lemas de la categoría requerida es superior a un determinado umbral. Por otro lado, un par bilingüe se considera asimétrico si uno de los contextos del par tiene una frecuencia alta en el corpus mientras que el otro tiene una frecuencia baja. Los umbrales de dispersión y asimetría se establecen empíricamente y pueden variar en función del tipo y tamaño del corpus. Una vez filtrados los pares de contextos dispersos y asimétricos, nos queda una lista reducida que llamamos *contextos semilla*. Esta lista será utilizada en el siguiente proceso de extracción.

3.3.2. Algoritmo de extracción

Con el objetivo de extraer pares de lemas bilingües, proponemos el siguiente algoritmo.

> Dada una lista de pares de contextos semilla:
>
> (a) para cada lema w_i de la lengua fuente, se cuenta el número de veces que éste instancia cada contexto semilla y se construye un vector de contextos con esa información;
>
> (b) para cada lema w_j de la lengua meta, se cuenta el número de veces que éste instancia cada contexto semilla y se construye un vector de contextos con esa información;
>
> (c) Calculamos la similitud DICE entre pares de vectores: $DICE(w_i, w_j)$; si w_j está entre los N más similares a w_i, entonces seleccionamos w_j como el candidato a ser la traducción de w_i.

Veamos un ejemplo. El cuadro 3 ilustra algunas posiciones del vector de contextos asociado al nombre castellano "Bachillerato". El valor de cada posición (tercera columna en el cuadro) representa el número de veces que el nombre coocurre con el contexto en el corpus comparable. Cada contexto del vector de la entrada castellana tiene que tener su correlato gallego, pues forma parte de la lista de pares de contextos semilla. La primera columna del cuadro representa el índice o posición del contexto en el vector.

El cuadro 4, por su parte, muestra los valores asociados a las mismas posiciones en el vector del nombre gallego "Bacharelato". Los contextos de la segunda columna son las traducciones de los castellanos que aparecen en el cuadro 3. Por ejemplo, en la posición 00198 de los dos vectores, aparecen los contextos: <estudio de [NOUN]> y <estudo de

índice	contexto	freq.
00198	<estudio de [NOUN]>	123
00234	<estudiante de [NOUN]>	218
00456	<curso de [NOUN]>	69
01223	<asignatura de [NOUN]>	35
02336	<[NOUN] en Lugo>	6
07789	<estudiar [NOUN]>	98
08121	<cursar [NOUN]>	56

Cuadro 3: Extracto del vector asociado al sustantivo español **Bachillerato**.

índice	contexto	freq.
00198	<estudo de [NOUN]>	78
00234	<estudante de [NOUN]>	145
00456	<curso de [NOUN]>	45
01223	<materia de [NOUN]>	41
02336	<[NOUN] en Lugo>	35
07789	<estudar [NOUN]>	23
08121	<cursar [NOUN]>	13

Cuadro 4: Extracto del vector asociado a la nombre gallego **Bacharelato**.

[NOUN]>. Como forman un par de contextos semilla, tienen que aparecer en la misma posición vectorial.

Tal y como muestran los cuadros 3 y 4, el nombre gallego "Bacharelato" coocurre con numerosos contextos que son traducciones de los contextos con los que también coocurre el nombre castellano "Bachillerato". Para calcular el grado de similitud entre dos lemas, w_1 y w_2, utilizamos una versión del coeficiente Dice:

$$\text{Dice}(w_1, w_2) = \frac{2\sum_i \text{mín}(f(w_1, c_i), f(w_2, c_i))}{f(w_1) + f(w_2)}$$

donde $f(w_1, c_i)$ representa el número de coocurrencias entre el lema w_1 y el contexto c_i. Como ya se ha dicho anteriormente, los lemas pueden ser mono o multi-léxicos. Para cada lema de la lengua fuente (castellano), seleccionamos los lemas de la lengua meta (gallego) con el valor de similitud Dice más alto, lo que los sitúa como sus posibles traducciones. En nuestros experimentos "Bacharelato" es el lema gallego con el valor de similitud más alto con respecto a "Bachillerato".

4. Experimentos y evaluación

4.1. El corpus comparable

El corpus comparable se compone de noticias de diarios y semanarios *on line*, publicados desde finales de 2005 hasta finales de 2006. El corpus castellano contiene 13 millones de palabras de artículos de *La Voz de Galicia* y *El Correo Gallego*. Por su parte, el corpus gallego contiene 10 millones de palabras de artículos extraídos de *Galicia-Hoxe, Vieiros* y *A Nosa Terra*. La mayoría de los textos gallegos están escritos respetando la normativa del 2003 de la Real Academia Galega, dejando para otros proyectos corpus con ortografías convergentes con el portugués. Los artículos recuperados cubren un amplio espectro temático: política regional, nacional e internacional, cultura, deporte y comunicación.

4.2. El diccionario bilingue

El diccionario bilingue que hemos utilizado para generar los contextos semilla es el empleado por el sistema de traducción automática de código abierto Opentrad, con el motor de traducción Apertium (Armentano-Oller et al., 2006) para los pares castellano-gallego. Nuestros experimentos tienen como objetivo actualizar el diccionario, que contiene actualmente cerca de 30.000 entradas, para mejorar los resultados del traductor castellano-gallego, implantado en *La Voz de Galicia*, sexto periódico en número de lectores de España. Este proyecto se realizó en colaboración con el área de ingeniería lingüística de imaxin|software.

El número de contextos bilingües generados a partir de las entradas del diccionario es de 539.561. A este número hay que sumarle aquellos contextos generados usando la estrategia de identificación de cognados en el corpus que no se encuentran en el diccionario. Estos son 754.469. En total, consiguimos 1.294.030 contextos bilingües. Este número se reduce drásticamente cuando pasamos el filtro que elimina los que tienen un comportamiento disperso y asimétrico en el corpus comparable. La lista final de contextos semilla es de: 127.604.

4.3. Evaluación

El protocolo de evaluación que elaboramos sigue, en algunos aspectos, el de (Melamed, 1997), que fue definido para evaluar un método de extracción de léxicos a partir de corpus paralelos. La precisión del léxico extraído se calcula con respecto a diferentes niveles de cobertura. En nuestro trabajo, la cobertura se define poniendo en relación las entradas del léxico y su presencia en el corpus compa-

rable. En particular, la cobertura se calcula sumando las frecuencias en el corpus de las ocurrencias de los lemas que forman el léxico extraído, y dividiendo el resultado por la suma de las frecuencias de todos los lemas en el corpus. El cálculo de la cobertura se hace separadamente para cada una de las categorías gramaticales en estudio: nombres, verbos y adjetivos. Y basta con calcularlo usando los lemas y el corpus de la lengua fuente. De esta manera, decimos que el léxico extraído alcanza un nivel de cobertura del 90 % para los nombres si, y sólo si, los nombres del léxico castellano (lengua fuente) tienen una frecuencia en el corpus que alcanza el 90 % de la frecuencia de todos los nombres en el mismo corpus.

Para calcular la precisión, fijamos una categoría gramatical y un nivel de cobertura del léxico, y extraemos aleatoriamente 150 lemas-test de esa categoría. Calculamos, en realidad, dos tipos de precisión: *precisión-1* se define como el número de veces que la traducción candidata seleccionada en primer lugar es la correcta, dividido por el número de lemas-test. *Precisión-10* es el número de candidatos correctos que aparecen en la lista de los 10 más similares de cada lema, dividido por el número de lemas-test.

Hasta ahora, en los protocolos de evaluación de otros métodos de extracción de léxicos bilingües a partir de corpus comparables no se había definido ningún tipo de cobertura. La única información sobre las palabras o lemas testados es su frecuencia absoluta. Es decir, se testan palabras o lemas con una frecuencia mayor a N, donde N suele ser ≥ 100. (Chiao y Zweigenbaum, 2002). El problema reside en que las frecuencias absolutas, al ser totalmente dependientes del tamaño del corpus de entrenamiento, no son útiles para comparar las tasas de precisión alcanzadas por diferentes métodos. En nuestro trabajo, sin embargo, la noción de nivel de cobertura intenta subsanar dicha limitación.

4.4. Resultados

El cuadro 5 muestra los resultados de la evaluación. Para cada una de las categorías gramaticales, incluidos los nombres multiléxicos, y para cada nivel de cobertura (90 %, 80 %, y 50 %), calculamos los dos tipos de precisión.

Con respecto a los nombres, los tres niveles de cobertura del 90, 80 y 50 por ciento corresponden a léxicos compuestos por 9.798, 3.534 y 597 nombres, respectivamente. En la categoría "Nombres" se incluyen nombres propios mono y multi-léxicos. La precisión al nivel del 90 % es relativamente baja (entre 50 y 60 por ciento) debido al elevado número de nombres propios incluidos en el léxico y a la dificultad de encontrar la correcta traducción de un nombre propio usando el método propuesto.[1] En la figura 1 ilustramos la evolución de la precisión (1 y 10) en función de los tres niveles de cobertura. Con una cobertura del 80 %, la precisión es bastante aceptable: entre el 80 y el 90 por ciento. A este nivel de cobertura, la frecuencia de las entradas evaluadas es ≥ 129. Se trata, por tanto, de un nivel próximo al empleado en la evaluación de otros trabajos relacionados, donde se calculaba la precisión de palabras con frecuencia ≥ 100. Sin embargo, en estos trabajos relacionados, las tasas de precisión son sensiblemente inferiores: alrededor del 72 % en los mejores casos (Rapp, 1999). Conviene precisar aquí que el hecho de tener resultados aceptables sólo con palabras o lemas frecuentes no es un problema insalvable ya que, al trabajar con corpus comparables, podemos fácilmente incrementar el tamaño del corpus y, con ello, el número de lemas que sobrepasen el umbral de la frecuencia 100. Por ejemplo, al incrementar nuestro corpus el doble del tamaño inicial, conseguimos obtener 1/3 más de lemas con una frecuencia superior a 100.

Con respecto a los adjetivos y verbos, resalta la disparidad en los resultados. Mientras la precisión para los verbos roza el 100 % al

[1] Buscamos la traducción de todo tipo de nombres propios pues el diccionario bilingüe del traductor necesita esta información. El motor Apertium 1.0 no integra todavía un detector de entidades.

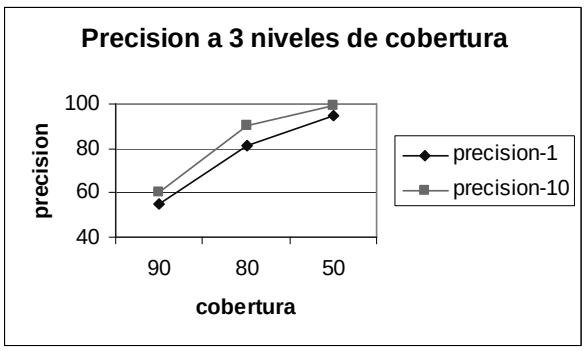

Figura 1: Precisión de los nombres a 3 niveles de cobertura

Categoría	Cobertura	Precisión-1	Precisión-10	Tamaño del léxico
Nombre	90 %	55 %	60 %	9798
Nombre	80 %	81 %	90 %	3534
Nombre	50 %	95 %	99 %	597
Adj	90 %	61 %	70 %	1468
Adj	80 %	81 %	87 %	639
Adj	50 %	94 %	98 %	124
Verbo	90 %	92 %	99 %	745
Verbo	80 %	97 %	100 %	401
Verbo	50 %	100 %	100 %	86
N multi-lex	50 %	59 %	62 %	2013

Cuadro 5: Resultados de la evaluación

80 % de cobertura, los adjetivos se sitúan entre el 81 y el 87 por ciento a ese mismo nivel. Los problemas para tratar los adjetivos radican sobre todo en la dificultad del desambiguador morfosintáctico para distinguir entre adjetivos y participios verbales. Un lema etiquetado como adjetivo por el desambiguador castellano puede tener su traducción en gallego etiquetada como verbo. Con respecto a la cobertura, en el 80 % el léxico de adjetivos consta de 639 lemas y el de verbos de 401. Los léxicos aprendidos para estas categorías son, por tanto, relativamente pequeños, pero el número puede y debe crecer con la explotación de más cantidad de corpus comparables.

Por último, evaluamos los lemas nominales multi-léxicos que no son nombres propios. La precisión se sitúa en torno al 60 % para una cobertura del 50 % del léxico. El principal problema relacionado con los lemas multiléxicos es su baja frecuencia en el corpus. Los 2.013 lemas evaluados a ese nivel de cobertura parten de frecuencias relativamente bajas, ≥ 40, lo que impide obtener resultados satisfactorios. Aún así, los resultados son sensiblemente mejores a los obtenidos por otros trabajos similares con términos multipalabra (Fung y McKeown, 1997), que no superan el 52 % de precisión para pequeños léxicos.[2]

5. Conclusiones

Hasta ahora no han sido muy numerosos los trabajos sobre extracción a partir de corpus comparables no-paralelos. La principal razón de esta escasez es, sin duda, la dificultad de conseguir resultados satisfactorios con los que se puedan crear recursos útiles. El método propuesto en este artículo presenta unos resultados que, sin llegar a las tasas de precisión de los métodos basados en corpus paralelos, dejan claro que los corpus comparables pueden ser una fuente muy interesante de conocimiento lexicográfico. Y existe todavía un amplio margen para mejorar los resultados. Dado que los corpus comparables crecen diariamente con el asombroso crecimiento de la Web, no resultaría complicado actualizar e incrementar los léxicos bilingües de forma incremental tomando en cuenta, en cada actualización, sólo aquellos lemas que juntos sumen una frecuencia, en los textos de la lengua fuente, del 80 % de la frecuencia total. Esta tarea de actualización incremental del léxico forma parte de nuestro trabajo en curso. De esta manera, pretendemos aumentar y mejorar el diccionario bilingüe del sistema de traducción Apertium.

[2] Si bien, el trabajo de (Fung y McKeown, 1997) tiene el mérito de extraer léxicos bilingües de dos lenguas muy dispares: inglés y japonés.

Bibliografía

Ahrenberg, Lars, Mikael Andersson, y Magnus Merkel. 1998. A simple hybrid aligner for generating lexical correspondences in parallel texts. En *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, páginas 29–35, Montreal.

Armentano-Oller, Carme, Rafael C. Carrasco, Antonio M. Corbí-Bellot, Mikel L. Forcada, Mireia Ginestí-Rosell, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, y Miriam A. Scalco. 2006. Open-source portuguese-spanish machine translation. En *Lecture Notes in Computer Science, 3960*, páginas 50–59.

Carreras, X., I. Chao, L. Padró, y M. Padró. 2004. An open-source suite of language

analyzers. En *4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Chiao, Y-C. y P. Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. En *19th COLING'02*.

Diab, Mona y Steve Finch. 2001. A statistical word-level translation model for comparable corpora. En *Proceedings of the Conference on Content-Based Multimedia Information Access (RIAO)*.

Fung, Pascale. 1995. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. En *14th Annual Meeting of Very Large Corpora*, páginas 173–183, Boston, Massachusettes.

Fung, Pascale y Kathleen McKeown. 1997. Finding terminology translation from non-parallel corpora. En *5th Annual Workshop on Very Large Corpora*, páginas 192–202, Hong Kong.

Fung, Pascale y Lo Yuen Yee. 1998. An ir approach for translating new words from nonparallel, comparable texts. En *Coling'98*, páginas 414–420, Montreal, Canada.

Gamallo, Pablo. 2007. Learning bilingual lexicons from comparable english and spanish corpora. En *Machine Translation SUMMIT XI*, Copenhagen, Denmark.

Gamallo, Pablo, Alexandre Agustini, y Gabriel Lopes. 2005. Clustering syntactic positions with similar semantic requirements. *Computational Linguistics*, 31(1):107–146.

Gamallo, Pablo y José Ramom Pichel. 2005. An approach to acquire word translations from non-parallel corpora. En *12th Portuguese Conference on Artificial Intelligence (EPIA'05)*, Evora, Portugal.

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, USA.

Harris, Z. 1985. Distributional structure. En J.J. Katz, editor, *The Philosophy of Linguistics*. New York: Oxford University Press, páginas 26–47.

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.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. En *COLING-ACL'98*, Montreal.

Melamed, Dan. 1997. A portable algorithm for mapping bitext correspondences. En *35th Conference of the Association of Computational Linguistics (ACL'97)*, páginas 305–312, Madrid, Spain.

Rapp, Reinhard. 1995. Identifying word translations in non-parallel texts. En *33rd Conference of the ACL'95*, páginas 320–322.

Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated english and german corpora. En *ACL'99*, páginas 519–526.

Silva, J. F., G. Dias, S. Guilloré, y G. P. Lopes. 1999. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. En *Progress in Artificial Intelligence*. LNAI, Springer-Verlag, páginas 113–132.

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