Extracción automática de léxico bilingüe: experimentos en español y catalán*

Automatic Bilingual Lexicon Extraction: Experiments in Spanish and Catalan

Raphaël Rubino **Iria da Cunha** **Georges Linarès**

Laboratoire Informatique d'Avignon
339, chemin des Meinajaries
84911 Avignon Cedex 9, Francia
raphael.rubino@univ-avignon.fr
georges.linares@univ-avignon.fr

Institut Universitari de
Lingüística Aplicada
Roc Boronat 138
08018 Barcelona, España
iria.dacunha@upf.edu

Resumen: En este artículo presentamos un sistema de extracción automática de léxico bilingüe catalán-español. Evitamos el empleo de corpus paralelos y usamos la información ofrecida por la Wikipedia como un corpus comparable entre el español y el catalán. Empleamos la similitud contextual para traducir unidades léxicas que no pueden traducirse por la distancia de edición. Los resultados obtenidos son positivos y confirman que este método podría aplicarse a las lenguas ibéricas.
Palabras clave: extracción automática, léxico bilingüe, traducción automática, español, catalán

Abstract: In this paper, we propose an automatic bilingual lexicon extraction system for Catalan and Spanish languages. Parallel corpora are not employed and Wikipedia is used as Catalan-Spanish comparable corpora. A contextual similarity approach is used to translate lexical units that are not translated by an edition distance. The obtained results are positive and confirm that this method could be applied to Iberian languages.
Keywords: Automatic Extraction, Bilingual Lexicon, Machine Translation, Spanish, Catalan

1. Introduction

En la Península Ibérica coexisten cinco lenguas oficiales: español, catalán, gallego, euskera y portugués. Para establecer vínculos entre estas lenguas y favorecer el multilingüismo, es necesario desarrollar recursos para todas ellas. Además, es indispensable crear recursos que permitan relacionarlas. Actualmente, hay una carencia de recursos de Procesamiento del Lenguaje Natural (NLP) para algunas de ellas, especialmente el gallego, el catalán y el euskera. Uno de los recursos necesarios para interrelacionar estas lenguas y diseñar herramientas de PLN (como sistemas de traducción automática) son los léxicos multilingües. Sin embargo, su desarrollo y actualización es costoso y lento, ya que normalmente supone la intervención humana.

El diseño de herramientas automáticas que ayuden en la construcción de léxicos bilingües (o multilingües) supone un reto en el ámbito del PLN. Existen trabajos que tratan este tema empleando diferentes estrategias. La mayor parte utilizan corpus paralelos (Brown et al., 1990; Wu y Xia, 1994; Koehn, 2005). No obstante, la creación de este tipo de corpus es costosa, lo cual encarece la investigación y no permite trabajar sobre todas las combinaciones de lenguas. Otra línea de investigación se basa en la utilización de un recurso más accesible, los corpus bilingües comparables, es decir, conjuntos de textos no paralelos con temáticas comunes pero escritos en cada lengua de manera independiente. Diversos autores han estudiado la

* Esta investigación ha sido parcialmente financiada por la Agence Nationale de la Recherche (ANR, Francia), proyecto AVISON (ANR-007-014); y los proyectos RICOTERM (FFI2010-21365-C03-01) y APLE (FFI2009-12188-C05-01) en España.

posibilidad de extraer unidades léxicas a partir de estos corpus, basándose en la hipótesis de que una unidad léxica y su traducción comparten similitudes en cuanto a su contexto (Fung, 1995; Rapp, 1995). Además de corpus comparables, esta aproximación emplea un léxico bilingüe preliminar de las lenguas analizadas.

La mayoría de las investigaciones sobre este tema se han realizado para relacionar el inglés con otras lenguas. Para las lenguas ibéricas, encontramos algunos trabajos, que utilizan principalmente métodos basados en corpus paralelos: para inglés-gallego (Guinovart y Fontenla, 2004), para portugués, español e inglés (Caseli y Nunes, 2007), y para inglés-gallego e inglés-portugués (Guinovart y Simoes, 2009).

Como se afirma en (Gamallo Otero y Pichel Campos, 2007), "desgraciadamente, no hay todavía una gran cantidad de texto paralelo, especialmente en lo que se refiere a lenguas minorizadas". Por esto, trabajar con lenguas como el gallego, catalán o euskera se hace más complicado. En (Gamallo Otero y Pichel Campos, 2007) se propone un método basado en corpus comparables de la Web, usando la idea de la similitud contextual. Lo aplican al español y el gallego, y, aunque sus resultados no superan los obtenidos usando corpus paralelos, son elevados. Esto refuerza la idea de que la gran cantidad de datos incluidos en la Web es una fuente de información importante y explotable para la construcción automática de léxicos bilingües. En esta línea, en (Gamallo y González, 2010) se propone un método automático para construir corpus comparables empleando la Wikipedia. En (Tomás et al., 2008) se construye un corpus que incluye dos tipos de artículos de la Wikipedia (paralelos y comparables) en español y catalán. En (Vivaldi y Rodríguez, 2010) se presenta un método de extracción de terminología bilingüe que emplea las categorías y estructura de la Wikipedia. La extracción de frases paralelas de la Wikipedia es también una tarea interesante que ha sido explorada por (Smith, Quirk, y Toutanova, 2010), por ejemplo, realizando diferentes experimentos a partir de la estructura de la Wikipedia.

El objetivo de este trabajo es desarrollar un sistema de extracción automática de léxico bilingüe para las lenguas de la Península Ibérica. Concretamente, trabajamos el par de lenguas español-catalán. Para ello, evitamos el empleo de corpus paralelos y aplicamos la idea de la similitud contextual entre una unidad léxica y su traducción (Fung, 1995; Rapp, 1995), empleando textos de la Wikipedia como corpus comparable. La metodología descrita en este trabajo está basada en el empleo de recursos y heurísticas existentes, pero aplicadas concretamente a la extracción de léxico bilingüe en estas dos lenguas.

2. Metodología

La metodología de nuestro trabajo incluye dos fases principales: Preprocesamiento y creación de recursos léxicos (FASE 0) y Aplicación del algoritmo (FASE 1).

2.1. FASE 0: Preprocesamiento y creación de recursos léxicos

Ya que nuestro trabajo se basa en un corpus comparable y un léxico bilingüe, en esta fase se construyen estos recursos. Concretamente, necesitamos dos léxicos bilingües: I) un léxico con candidatos a la traducción (con sus correspondientes traducciones) y II) un léxico "pivote" utilizado como elemento de relación entre las dos lenguas.

2.1.1. Preprocesamiento del corpus comparable

El preprocesamiento del corpus comparable incluye:

- Descarga de un fichero con todos los artículos de la Wikipedia (Wikipedia Dump) en las dos lenguas de trabajo (español y catalán).
- Eliminación de "páginas redirigidas" en Wikipedia, es decir, artículos que tienen un título pero no contienen texto en su interior. Por ejemplo, en la Wikipedia en español, la unidad "Proyección Azimutal" está vacía y redirigida a "Proyección azimutal" (simplemente cambia una "a" en mayúscula o minúscula); el año "4450" está redirigido al artículo sobre el "V milenio", etc.
- Eliminación de las stopwords en las dos lenguas. La lista de stopwords en catalán se ha obtenido del área de Ingeniería Lingüística del Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra (UPF)[1]. La lista

[1] http://latel.upf.edu/morgana/altres/pub/ca_stop.htm

de stopwords en español se ha obtenido del Laboratoire Informatique d'Avignon (LIA-UAPV)[2].

- Formateo de este fichero en Trec-text[3]. En el siguiente ejemplo se muestra un ejemplo de este tipo de formato, en donde la etiqueta <DOCNO> indica el número de documento, <TITLE> el título y <TEXT> el contenido:

```
<DOC>
 <DOCNO> 22 </DOCNO>
 <TITLE> Astronomía galáctica </TITLE>
 <TEXT>
   se denomina 'astronomía galáctica' a
   la investigación astronómica de nuestra
   galaxia, la vía láctea [...] seguros
   posee un agujero negro, etc.
 </TEXT>
</DOC>
```

- Indexación de los artículos con Lemur Indexation Toolkit[4]. Usamos esta herramienta para facilitar el cálculo de coocurrencias entre la unidad léxica que se quiere traducir y su contexto (es decir, las palabras del léxico II).

Actualmente, la Wikipedia en español contiene 761.727 artículos y en catalán 341.142. Depués de este preprocesamiento, nuestro corpus incluye 701.423 artículos en español y 296.465 en catalán. Esta reducción se debe a la eliminación de artículos redirigidos. No se realizó una selección temática de los artículos incluidos en el corpus, sino que se emplearon todos los temas de la Wikipedia. Tampoco se usó la estructura de la Wikipedia.

2.1.2. Recopilación del léxico I

En esta fase, creamos nuestro propio léxico bilingüe, que contiene los candidatos a la traducción en la lengua de partida (catalán), acompañados de su traducción en la lengua de llegada (español). Construimos estos recursos dada la carencia de léxicos bilingües extensos y actualizados gratuitos disponibles para el par de lenguas empleadas. Así, nuestro léxico podrá contener neologismos de reciente creación (como, por ejemplo, "mileurista")[5]. Esta fase incluye dos subfases:

1. Extracción de relaciones de correspondencia entre los títulos de los artículos de la Wikipedia en español y catalán, para obtener una lista preliminar de léxico bilingüe. Las relaciones entre los artículos en estas dos lenguas se establecen mediante enlaces interlengua (en el menú "En otros idiomas" de la Wikipedia en español). Establecemos las correspondencias en los dos sentidos (español-catalán y catalán-español) porque, en ocasiones, la estructura de la Wikipedia no correlaciona de la misma forma las entradas en los dos sentidos. Por ejemplo, en la Wikipedia en catalán encontramos la entrada "Prestige", que está correlacionada en la Wikipedia en español con "Desastre del Prestige". Sin embargo, la Wikipedia en español también ofrece la entrada "Prestige" (que se refiere al mismo petrolero), que solo muestra su correspondencia al inglés y al ruso, pero no al catalán. Vemos así que la estructura de la Wikipedia en español es más compleja que la de otras lenguas con menos entradas.
2. Filtrado de la lista preliminar de los dos léxicos bilingües mediante la eliminación automática de:

   - Pares de unidades léxicas que no mantienen la misma correlación en la estructura de la Wikipedia en los dos sentidos.
   - Pares de unidades léxicas que coinciden en las dos lenguas. Este criterio se aplica por dos motivos. Primero, porque consideramos que no es interesante evaluar los pares de unidades que son idénticas. Segundo, porque una gran cantidad de las unidades de este léxico bilingüe extraído de la Wikipedia serán entidades nombradas iguales en ambas lenguas, como por ejemplo "Harry Potter".
   - Pares de elementos numéricos, ya que no nos interesa traducir cifras, años, fechas, etc., aunque somos conscientes de que estás entidades podrían servir para poder paralelizar de forma eficiente frases en corpus comparables.
   - Pares de elementos en que solo uno tiene un signo de puntuación: generalmente

[2] http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/torres/logiciels/fonctionnels_esp.txt

[3] http://trec.nist.gov

[4] http://www.lemurproject.org

[5] Para más información sobre neología véase (Cabré y Estopà, 2009)

indican un error en la traducción (excepto el punto de la geminada del catalán). - Pares de elementos que pueden traducirse por la distancia de edición (Levenshtein, 1966). Por ejemplo, las siguientes unidades léxicas del catalán (a la izquierda) fueron traducidas correctamente al español por la distancia de edición (a la derecha), ya que las similitudes ortográficas son evidentes:

CATALÁN	ESPAÑOL
palau de westminster	*palacio de westminster*
lateralitat	*lateralidad*
fagocitosi	*fagocitosis*
província de bilecik	*provincia de bilecik*

En cambio, las siguientes unidades del catalán no se tradujeron adecuadamente:

CATALÁN	ESPAÑOL
surquillo	*bordillo*
floquet neu	*alquino*
tupaia	*tucana*
eratostenià	*río eno*

Comenzamos con un léxico de 140.137 unidades. Después del filtrado, antes de aplicar la distancia de edición, obtenemos 57.859 unidades y, después de la distancia de edición, 8.045 unidades, con las que trabajamos finalmente. Este léxico final contiene las unidades léxicas más difíciles de traducir, porque no pueden ser traducidas por una distancia de edición tradicional. Por este motivo, consideramos que la traducción automática de estas 8.045 unidades es el principal reto. Partimos de la idea de que el léxico bilingüe creado en esta fase es correcto. Sin embargo, no hemos realizado una revisión manual, dada su gran extensión. Esta revisión sería óptima para eliminar errores, pero intentamos evitar al máximo la intervención humana.

2.1.3. Recopilación del léxico II

Como ya hemos comentado, este léxico "pivote" se utiliza como elemento de relación entre las dos lenguas del trabajo. Por este motivo, este léxico debe ser correcto necesariamente, ya que gracias a él se realizan las corrrespondencias entre lenguas. Por esto, hemos decidido utilizar un léxico bilingüe español-catalán existente en la colección AULEX[6], que contiene vocabularios breves en línea de lenguas con recursos limitados, dirigida por Manuel Rodríguez Villegas, especialista compilador de diccionarios en línea.

2.2. FASE 1: Aplicación del algoritmo

El proceso de identificación de traducciones puede ser visto como un alineamiento palabra por palabra. Esta tarea se aborda normalmente mediante algoritmos basados en corpus paralelos, como el modelo IBM (Brown et al., 1993; González-Rubio et al., 2008). Sin embargo, como nosotros basamos nuestro proceso de extracción en corpus comparables (no paralelos), necesitamos otro método. Esta es la razón por la que nos centramos en la información contextual de la palabra que se quiere traducir y candidatos a traducciones. Nuestra aproximación se basa en las palabras adyacentes, asumiendo que podemos traducir parte del contexto del vocabulario. De hecho, como no se pueden traducir todas las unidades léxicas existentes alrededor de los candidatos en la lengua fuente y la lengua de llegada, necesitamos capturar la información más importante en las coocurrencias detectadas. Usamos medidas de normalización para resaltar las particularidades de las coocurrencias entre una palabra (léxico I) y las palabras del léxico "pivote" (léxico II).

En resumen, el método para identificar traducciones basado en la información contextual incluye cuatro pasos:

- cálculo de las coocurrencias entre una palabra (léxico I) y las palabras del léxico "pivote" (léxico II),
- normalización de las coocurrencias con una medida de asociación,
- construcción de un vector de contexto,
- comparación de los vectores de la lengua de partida y la lengua de llegada con una medida de similitud.

La Figura 1 resume el proceso general de extracción de traducción que presentamos en este trabajo.

El primer paso está basado en la premisa de que una palabra y su traducción comparten similitudes contextuales en corpus comparables. Las palabras del léxico "pivote"

[6] http://aulex.org/aulex.php

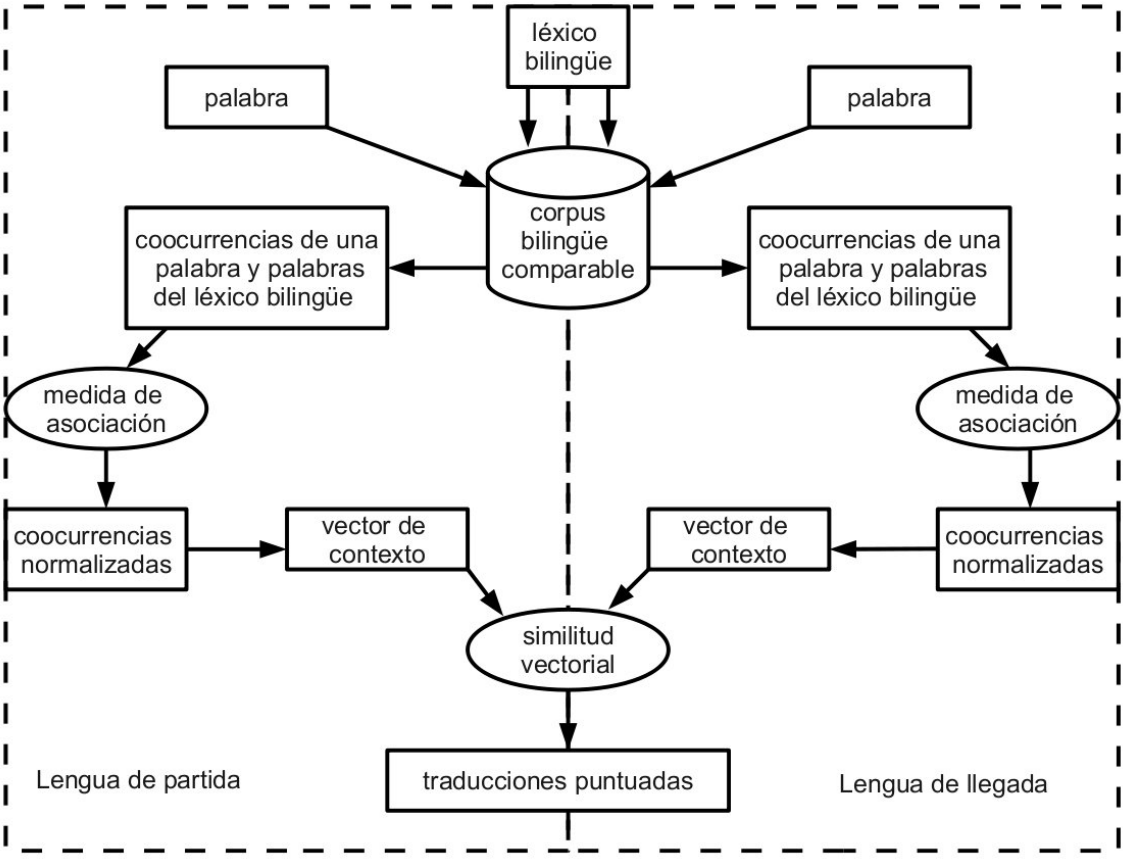

Figura 1: Esquema general del proceso de extracción de traducciones.

(léxico II) son los elementos de relación en ambas lenguas para modelizar el espacio contextual de donde vamos a extraer las traducciones. Las coocurrencias entre una palabra (léxico I) y las palabras del léxico "pivote" (léxico II) se contabilizan dentro de una ventana deslizante de un tamaño fijo (de 10 a 30 palabras en cada ejemplo) o dinámico (oraciones, párrafos, etc.).

El segundo paso ha sido ya ampliamente estudiado en la literatura. Se han probado diversas medidas de asociación, basadas en tablas de contingencia 2*2 como la mostrada en el Cuadro 1, y se observa que las más efectivas son información mutua (Church y Hanks, 1990), *log-likelihood* (Dunning, 1993) y *odds-ratio* (Evert, 2004). En la Sección 3 presentamos los resultados obtenidos con las medidas de información mutua y *odds-ratio*, cuyas fórmulas ofrecemos en la Ecuación 1 y 2, respectivamente.

$$mi(w,s) = log\frac{a}{(a+b)(a+c)} \qquad (1)$$

$$odds(w,s) = log\frac{(a+\frac{1}{2})(d+\frac{1}{2})}{(b+\frac{1}{2})(c+\frac{1}{2})} \qquad (2)$$

	s	$\overline{s}$
w	$a = occ(\ w\ ,\ s\ )$	$b = occ(\ w\ ,\ \overline{s}\ )$
$\overline{w}$	$c = occ(\ \overline{w}\ ,\ s\ )$	$d = occ(\ \overline{w}\ ,\ \overline{s}\ )$

Cuadro 1: Tabla de contingencias entre dos palabras

El Cuadro 1 contiene las coocurrencias comunes en una ventana de una palabra del léxico I (reflejada como w) y las palabras del léxico "pivote" o II (reflejadas como s), pero también los casos en los que w aparece sin s, s aparece sin w, y finalmente en los que no aparecen juntas. Este paso de normalización es particularmente útil para tratar diferencias entre lenguas en corpus comparables. Por ejemplo, el corpus extraído de la Wikipedia

	Documentos	Unidades léxicas
Candidatos	-	300
Léxico "pivote"	-	1.944
Wikipedia CA	296.465	1.461.325
Wikipedia ES	701.423	3.931.243

Cuadro 2: Recursos empleados para los experimentos

en español contiene una mayor cantidad de unidades léxicas, por eso el número de ocurrencias de palabras es mayor que el número de ocurrencias de su traducción en una lengua con menos recursos (como el catalán).

El tercer paso se refiere básicamente a la modelización de una palabra (léxico I) en un espacio contextual. Para cada palabra (léxico I) en la lengua de partida y de llegada, el contexto se modeliza como un vector de contexto. Cada componente de este vector contiene un cálculo de coocurrencias normalizado. Los componentes tienen que ser fijos porque queremos que las dimensiones sean comparables entre los vectores de la lengua de partida y de llegada.

El cuarto paso se basa en medidas de vectores de similitud para comparar los vectores de contexto en la lengua de partida y de llegada. El objetivo es detectar similitudes entre las asociaciones contextuales de las palabras. Los vectores más similares son traducciones posibles. Estas medidas son otro parámetro bien estudiado en la literatura, y las más populares son el coseno, la distancia euclidiana y la métrica *City Block* (Morin et al., 2007). La fórmula de la distancia del coseno entre los vectores de la lengua de partida y de llegada, con la medida de asociación *odds-ratio*, se detalla en la Ecuación 3 (donde V es un vector, s es la lengua de partida, t es la lengua de llegada, y n es una unidad del léxico "pivote").

$$cosine_{V_s}^{V_t} = \frac{\sum_n odds_n^s odds_n^t}{\sqrt{(\sum_n odds_n^s)^2}\sqrt{(\sum_n odds_n^t)^2}} \quad (3)$$

3. Experimentos y resultados

Para evaluar nuestro método, hemos empleado los recursos incluidos en el Cuadro 2. Hemos extraído aleatoriamente 300 candidatos a traducir del léxico I.

Hemos realizado diversos experimentos empleando las medidas de asociación y las medidas de similitud vectorial, presentadas en 2.2. Observamos que los mejores resultados se obtienen con la utilización de la medida de asociación *odds-ratio* y la similitud de *cosenos*. Los resultados se presentan en el Cuadro 3 (P = Precisión, C = Cobertura, F = F-measure). Consideramos que es interesante presentar también los resultados obtenidos con las otras medidas de asociación, como las coocurrencias y la información mutua.

A continuación mostramos algunos ejemplos de traducciones correctas:

CATALÁN	ESPAÑOL
formatge blau	*queso azul*
floridura	*moho*
momificació	*embalsamamiento*
senglar calidó	*jabalí calidón*
vaga	*huelga*

Y también ejemplos de traducciones incorrectas:

CATALÁN	ESPAÑOL
creu nòrdica	*idioma islandés*
castellà mèxic	*alfabetización*
bombeta elèctrica	*cuenco*
astúries	*labor*
bitxo	*salsa pescado*

Los resultados obtenidos muestran la eficacia en cuanto a la precisión en el rango 1 de la medida *odds ratio* combinada con la similitud de *cosenos*. El aumento de la cobertura según el número de candidatos tenidos en cuenta (un rango entre 5 y 10) implica un descenso significativo de la precisión. El cálculo de la precisión tiene en cuenta el número de unidades léxicas de la lengua de llegada consideradas como una buena traducción. Para el rango 10, por ejemplo, una sola traducción es válida según la referencia (léxico I), pero el sistema ofrece 10. En este rango, la información mutua y *odds ratio* son equivalentes en cuanto a precisión y cobertura.

Estos resultados son difícilmente comprables con los de otros trabajos. Sin embargo, observamos que, para el dominio periodístico, los experimentos de (Rapp, 1999) muestran una precisión del rango 1 del 72 % sobre 100 candidatos evaluados. El autor utiliza un corpus en alemán que contiene 135 millones de palabras y un corpus en inglés que incluye 163 millones. Además, el léxico "pivote" que emplea en sus experimentos contiene 16.380 entradas, es decir, que es muy superior al léxico "pivote" que nosotros empleamos en este tra-

	TOP 1			TOP 5			TOP 10		
	P	*C*	*F*	*P*	*C*	*F*	*P*	*C*	*F*
Coocurrencias	45,00	45,00	45,00	15,33	76,67	25,56	8,17	81,67	14,85
Información mutua	57,67	57,67	57,67	16,60	83,00	27,67	9,07	90,67	16,48
Odds ratio	58,00	58,00	**58,00**	16,47	82,33	27,44	9,07	**90,67**	16,48

Cuadro 3: Resultados obtenidos a tres rangos (mejores 1, 5 y 10 traducciones) por similitud de cosenos entre los vectores de contexto

bajo. De hecho, creemos que la precisión del rango 1 del 58 %, que hemos obtenido, podría mejorarse con un léxico con un mayor número de entradas. Este aspecto está relacionado con la cantidad de recursos disponibles para el catalán, menos dotado que otras lenguas. La evaluación de los candidatos ubicados en el primer rango es el modo más apropiado de observar si el léxico bilingüe extraído podría ser incluido en un sistema de traducción automática. Sin embargo, es necesario mejorar la precisión de los resultados con el objetivo de aportar recursos robustos.

En nuestro trabajo no abordamos la construcción de modelos estadísticos de traducción, sino que nos centramos en la tarea de la extracción de léxico bilingüe. Sin embargo, existen diversos trabajos que se están realizando actualmente por otros autores en relación con el entrenamiento de sistemas de traducción automática con datos no paralelos, obteniendo resultados prometedores (Ravi y Knight, 2011).

4. Conclusiones y trabajo futuro

En este trabajo presentamos un sistema de extracción automática de léxico bilingüe, que aplicamos a un par de lenguas de la Península Ibérica: español-catalán. Para los experimentos no empleamos corpus paralelos, sino corpus comparables usando como recurso la información ofrecida por la Wikipedia, aplicando la idea de las similitudes contextuales entre una unidad léxica y su traducción. Los resultados obtenidos son positivos, dado que se logró traducir correctamente más de la mitad de los candidatos. Además, consideramos que la precisión del rango 1 podrá mejorarse mediante un léxico "pivote" que incluya más unidades léxicas, lo cual planeamos hacer como trabajo futuro.

Creemos que este trabajo es relevante, dado que proponemos un sistema que casi no requiere esfuerzo humano, es rápido y, sobre todo, permite la actualización constante del léxico bilingüe, ya que la Wikipedia se amplía cada día con nuevas entradas. Tomando la Wikipedia como un corpus abierto y en constante evolución, podremos emplear este método para aumentar el léxico de cualquier lengua de la Península Ibérica de una manera dinámica y, así, favorecer el multilingüismo, las relaciones entre lenguas y el desarrollo de herramientas de PLN, como los sistemas de traducción automática. La principal ventaja de la metodología empleada en este trabajo es que es independiente de lengua. Para emplearla en diferentes lenguas solo se necesita un corpus comparable y un léxico "pivote" entre las dos lenguas que se quieren tratar.

Como trabajo futuro, nos gustaría aplicar el sistema sobre otros pares de lenguas. Especialmente, estamos interesados en el español-euskera, dada la gran diferencia ortográfica entre las unidades léxicas de estas dos lenguas. Además, nos gustaría incorporar nuestro sistema de extracción a un sistema de traducción automática, para:

1. realizar una evaluación extrínseca de nuestro sistema,
2. aumentar la cobertura de vocabulario de un traductor automático.

Bibliografía

Brown, P.F., S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, y P.S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

Brown, P.F., S.D. Pietra, V.J.D. Pietra, y R.L. Mercer. 1993. The Mathematic of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Cabré, M.T. y R. Estopà. 2009. *Les paraules noves criteris per detectar i mesurar els neologismes*. Eumo editorial.

Caseli, HM y MGV Nunes. 2007. Automatic Induction of Bilingual Lexicons for Machi-

ne Translation. *International Journal of Tranlation*, 19:29–43.

Church, K.W. y P. Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Dunning, T. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

Evert, S. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. tesis, Universität Stuttgart. 353 páginas.

Fung, P. 1995. Compiling Bilingual Lexicon Entries from a Non-parallel English-Chinese Corpus. En *Workshop on Very Large Corpora*, páginas 173–183.

Gamallo, P. y I. González. 2010. Wikipedia as a Multilingual Source of Comparable Corpora. En *LREC Workshop on Building and Using Comparable Corpora*, páginas 19–26.

Gamallo Otero, P. y J.R. Pichel Campos. 2007. Un método de extracción de equivalentes de traducción a partir de un corpus comparable castellano-gallego. *Lenguaje Natural*, páginas 241–248.

González-Rubio, J., G. Sanchis-Trilles, A. Juan, y F. Casacuberta. 2008. A Novel Alignment Model Inspired on IBM Model 1. En *EAMT*, páginas 47–56.

Guinovart, X.G. y E.S. Fontenla. 2004. Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos. *Lenguaje Natural*, 33:133–140.

Guinovart, X.G. y A. Simoes. 2009. Parallel Corpus-Based Bilingual Terminology Extraction. En *International Conference on Terminology and Artificial Intelligence*.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. En *MT Summit X*, páginas 79–86.

Levenshtein, V.I. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. En *Soviet Physics Doklady*, páginas 707–710.

Morin, E., B. Daille, K. Takeuchi, y K. Kageura. 2007. Bilingual Terminology Mining-Using Brain, not Brawn Comparable Corpora. En *ACL*, páginas 664–671.

Rapp, R. 1995. Identifying Word Translations in Non-parallel Texts. En *ACL*, páginas 320–322.

Rapp, R. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. En *ACL*, páginas 519–526.

Ravi, S. y K. Knight. 2011. Deciphering Foreign Language. En *ACL*, páginas 12–21.

Smith, J.R., C. Quirk, y K. Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. En *NAACL/HLT*, páginas 403–411.

Tomás, J., J. Bataller, F. Casacuberta, y J. Lloret. 2008. Mining wikipedia as a parallel and comparable corpus. En *Language Forum*.

Vivaldi, J. y H. Rodríguez. 2010. Finding domain terms using wikipedia. En *LREC*, páginas 386–393.

Wu, D. y X. Xia. 1994. Learning an English-Chinese lexicon from a Parallel Corpus. En *AMTA*, páginas 206–213.