Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 187-194 recibido 30-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Una Nueva Técnica de Construcción de Grafos Semánticos para la Desambiguación Bilingüe del Sentido de las Palabras *

A New Technique for Cross Lingual Word Sense Disambiguation based on Building Semantic Graphs

Andres Duque Fernandez[1], **Lourdes Araujo**[2], **Juan Martinez-Romo**[3]
NLP & IR Group
Universidad Nacional de Educación a Distancia (UNED)
28040 Madrid, Spain
{[1]aduque, [2]lurdes, [3]juaner}@lsi.uned.es

Resumen: En este trabajo presentamos unos resultados preliminares obtenidos mediante la aplicación de una nueva técnica de construcción de grafos semánticos a la tarea de desambiguación del sentido de las palabras en un entorno multilingüe. Gracias al uso de esta técnica no supervisada, inducimos los sentidos asociados a las traducciones de la palabra ambigua considerada en la lengua destino. Utilizamos las traducciones de las palabras del contexto de la palabra ambigua en la lengua origen para seleccionar el sentido más probable de la traducción. El sistema ha sido evaluado sobre la colección de datos de una tarea de desambiguación multilingüe que se propuso en la competición SemEval-2010, consiguiendo superar los resultados de todos los sistemas no supervisados que participaron en aquella tarea.
Palabras clave: Desambiguación lingüística, bilingüismo, métodos basados en grafos

Abstract: In this paper we present preliminary results obtained by the application of a new technique for building semantic graphs to the task of cross-lingual word sense disambiguation. Through the use of this unsupervised technique, we induce the senses associated with the translations of the ambiguous word in the target language. For this purpose, we use the translation of the words in the context of the ambiguous word in the source language to select the most likely sense. The system has been evaluated on a dataset from a cross-lingual word sense disambiguation task proposed in the SemEval-2010 competition, outperforming all unsupervised systems participating in that task.
Keywords: Cross-lingual, word-sense disambiguation, graph-based approaches

1. Introducción

Una gran parte de las palabras de un idioma son polisemas, es decir, tienen más de un significado y se interpretan de distintas formas según el uso que se hace de ellas. La Desambiguación del Sentido de las Palabras o Word Sense Disambiguation (WSD) se ha convertido por ello en uno de los problemas del Procesamiento del Lenguaje Natural (PLN) que ha atraído más atención, ya que también es un paso necesario para numerosos procesos de PLN (Ide y Veronis, 1998): traducción automática, recuperación de información, clasificación de textos, etc.

El problema de la desambiguación del sentido de las palabras se ha tratado frecuentemente como un problema de aprendizaje supervisado (Màrquez et al., 2006; Mihalcea, 2006). Sin embargo, este tipo de métodos requiere disponer de textos etiquetados semánticamente. Estos recursos son muy costosos y de hecho muy escasos. Por ello ha surgido un interés creciente en abordar el problema de forma no supervisada. Estos trabajos, que no requieren textos anotados semánticamente, se centran en la denominada *Inducción del sentido de las palabras* (ISP). El objetivo es distinguir los distintos usos o sentidos de una palabra determinada en un texto dado, pero sin clasificar dichos sentidos en base a un inventario de sentidos preexistente. La distinción se hace en base a grupos de palabras que presentan alguna relación acentuada con un sentido particular. Generalmente esta relación es la coaparición con la palabra considerada en los contextos observados en los textos. Esta es una característica muy interesante, como apunta Pedersen (2006), ya que por una parte no existe un inventario de sentidos para todas las palabras, y por otra parte, incluso cuando existe este inventario, la naturaleza y el grado de distinción de los sentidos que

* Trabajo financiado parcialmente por los proyectos Holopedia (TIN2010-21128-C02-01) y MA2VICMR (S2009/TIC-1542)

ISSN 1135-5948

nos interesa varía con las aplicaciones.

De forma muy genérica los métodos aplicados en ISP se pueden clasificar en dos categorías principales, los basados en vectores y los basados en grafos. Estos últimos (Veronis, 2004; Agirre et al., 2006; Agirre y Soroa, 2007; Klapaftis y Manandhar, 2008) son los más directamente relacionados con el trabajo que presentamos. Generalmente estos enfoques representan como un vértice a cada palabra que coaparece con la palabra objetivo (palabra ambigua que se quiere traducir) dentro de una ventana predefinida. Dos vértices están conectados por una arista si coaparecen en uno o más contextos de la palabra objetivo. Una vez que se ha construido el grafo para la palabra objetivo se aplican distintos algoritmos de detección de comunidades para inducir los sentidos. Cada comunidad de palabras se toma como uno de los sentidos inducidos.

Otro enfoque que también se ha investigado para tratar la desambiguación es la utilización de textos paralelos. Como ha señalado Resnik (2004), no sólo para este problema, sino en general para el procesamiento del lenguaje, el significado oculto que comparten las traducciones paralelas permite inferir conocimiento de una lengua a partir de otra en la que se disponga de más recursos.

En este trabajo proponemos inducir los sentidos de las palabras ambiguas en distintos idiomas mediante un nuevo algoritmo de construcción de grafos y aplicarlos a la desambiguación semántica de textos.

La hipótesis de partida del algoritmo de construcción de grafos que utilizamos (Martinez-Romo et al., 2011) es que un documento tiene un contenido coherente, por lo que tiene sentido hacer el supuesto básico de que todas las palabras que aparecen en el mismo documento tienden a compartir un sentido común. El objetivo es enlazar cada par de palabras que compartan un sentido común, condición que aproximamos por la coaparición de ambas palabras en un mismo documento. Sin embargo, esto no es siempre cierto. Algunas palabras pueden aparecer en un documento sin tener realmente relación con el sentido general de dicho documento. Por lo tanto, se considera que dos palabras realmente comparten un sentido común si coaparecen *frecuentemente* en los mismos documentos.

Concretamente, generaremos un grafo en el idioma de destino (español) del que extraemos comunidades de palabras que representarán los sentidos inducidos. Un diccionario de traducciones entre lenguas nos permite identificar las comunidades asociadas a las traducciones que están más relacionadas en las lenguas consideradas. En concreto hemos utilizado un diccionario (López-Ostenero, 2002) bilingüe español-inglés creado en el Grupo de Procesamiento de Lenguaje Natural de la UNED (http://nlp.uned.es). Considerando textos alineados, esta identificación nos permitirá seleccionar el sentido más acertado de las palabras ambiguas de los textos buscando un sentido común entre las traducciones de los idiomas considerados.

El resto del artículo se organiza de la siguiente forma: en la sección 2 se citan los principales trabajos dentro del área de la desambiguación multilingüe del sentido de las palabras. En la sección 3 se describirán cada una de las etapas que componen la metodología seguida por el algoritmo propuesto. En la sección 4 se muestran los principales resultados. Finalmente, en la sección 5 se extraen una serie de conclusiones y se expone la línea de trabajo futuro.

2. *Estado del Arte*

Ha habido algunas propuestas de explotar los corpora paralelos para tratar la desambiguación del sentido de las palabras. Resnik y Yarowsky (1999) presentaron uno de los primeros análisis de la potencialidad de los recursos multilingües para la desambiguación, proponiendo un marco de evaluación y una medida de distancia multilingue entre sentidos. Diab y Resnik (2002) propusieron un método para anotar automáticamente el sentido de las palabras en grandes corpora paralelos. Este método requiere disponer de un inventario de sentidos para una de las lenguas y se apoya en el alineamiento a nivel de palabra para identificar las traducciones de palabras entre las lenguas consideradas. Ng, Wang, y Chan (2003) propusieron un método para adquirir datos de entrenamiento de desambiguación de sentidos basado en la selección manual de traducciones entre corpora paralelos que se utilizaban para entrenar un clasificador. Banea y Mihalcea (2011) también proponen un método supervisado, que en este caso utiliza rasgos multilingües para entrenar un clasificador. Los rasgos se obtienen traduciendo el contexto de las palabras ambiguas a varias lenguas. Fernandez-Ordonez, Mihalcea, y Hassan (2012) hacen una propuesta no supervisada suponiendo que la única información disponible es un diccionario con las definiciones de los distintos significados de las palabras ambiguas. En esta propuesta se utiliza una variante del algoritmo Lesk, que dada una secuencia de palabras intenta identificar las combinaciones de sentidos

que maximizan el solapamiento entre las definiciones correspondientes.

En relación a la evaluación, en la edición de 2010 de la campaña de SemEval en la que se proponen competiciones de sistemas sobre tareas relacionadas con la desambiguación del sentido de las palabras (WSD), se incluyó una dedicada a la desambiguación en un contexto multilingüe (Lefever y Hoste, 2010). En dicha tarea los participantes debían determinar automáticamente la traducción apropiada para el contexto de un nombre inglés dados cinco idiomas: Holandés, Alemán, Italiano, Español y Francés. Para la compilación de la colección, fueron empleados dos tipos de datos: un corpus paralelo extraído de Europarl (http://www.statmt.org/europarl/) sobre el que se construyó el "gold standard" y una colección de frases en inglés que contienen las palabras de la muestra léxica anotadas con sus correspondientes traducciones en cinco idiomas. En el idioma *español*, que es en el que nos centraremos en este trabajo, participaron cuatro sistemas. En cuanto a los sistemas supervisados, los sistemas UvT-WSD (van Gompel, 2010) y FCC (Vilariño et al., 2010) emplearon un clasificador basado en el algoritmo K-Nearest y Naive Bayes respectivamente. Los sistemas no supervisados utilizaron algoritmos basados en grafos con ciertas diferencias respecto al algoritmo presentado en este trabajo. El sistema T3-COLEUR (Guo y Diab, 2010) hizo uso de tablas de probabilidad de traducción bilingües que se derivan a partir del corpus Europarl. Por su parte el sistema UHD (Silberer y Ponzetto, 2010) construye para cada palabra objetivo un grafo de coapariciones multilingüe basado en los contextos alineados de la palabra objetivo, disponibles en los corpora paralelos. Los términos con una traducción cruzada por cada idioma se unen por un enlace específico de "traducción" entre los diferentes grafos correspondientes a cada idioma. Finalmente el grafo se transforma en un árbol de expansión mínimo y es utilizado para seleccionar las palabras más relevantes en el contexto y desambiguar cada instancia de evaluación.

En este trabajo también utilizaremos estos datos para su evaluación y aplicaremos una nueva técnica de construcción de grafos semánticos que se describe a continuación.

3. Descripción del Algoritmo

En esta sección se describirán cada una de las etapas que componen la metodología seguida para la aplicación del algoritmo basado en grafos de coaparición a una tarea de desambiguación semántica multilingüe.

3.1. Preprocesado del Corpus

El algoritmo basado en grafos de coaparición que se expone en el presente artículo es un algoritmo no supervisado, que utiliza el conocimiento extraído de los documentos completos que contienen alguna instancia de la palabra a desambiguar. En este caso, se ha utilizado el corpus paralelo multilingüe Europarl (Koehn, 2005), el cuál se extrae de las actas del Parlamento Europeo correspondientes a los años comprendidos entre 1996 y 2011, y se encuentra alineado a nivel de frase. Los idiomas utilizados para realizar la desambiguación han sido el inglés como idioma origen, y el español como idioma destino.

El corpus inicial está separado en documentos que representan las actas del Parlamento. Cada uno de los documentos contiene diferentes etiquetas que indican las distintas partes del mismo. Tras analizar los documentos que aparecen en el corpus utilizado, se decidió separar los documentos mediante la etiqueta "<SPEAKER>" que separan cada una de las intervenciones acontecidas a lo largo de la sesión. De esta forma, cada uno de los documentos que el algoritmo tiene en cuenta para construir un grafo de coaparición, representa la intervención de un único miembro del Parlamento Europeo en una sesión concreta y por lo tanto respeta nuestra hipótesis de que todas las palabras que aparecen en el mismo documento tienden a compartir un sentido común.

Una vez que se ha realizado la separación de los documentos, es necesario realizar un etiquetado de los mismos para clasificar las palabras que en ellos aparecen, en función de su categoría gramatical. Esta tarea se realizó de forma automática mediante la herramienta TreeTagger (Schmid, 1994), para los dos idiomas utilizados. Una vez realizado este etiquetado, disponemos de un conjunto de documentos etiquetados gramaticalmente, alineados a nivel de frase, en los dos idiomas que se van a utilizar para esta tarea.

3.2. Construcción del grafo semántico

El siguiente paso consiste en la construcción del grafo de coaparición de palabras en la lengua destino (español), a partir de los documentos etiquetados. Para ello se han utilizado grafos dedicados, es decir, para construir el grafo de palabras en español, sólo se han tenido en cuenta aquellos documentos que contienen al menos una de las posibles traducciones de la palabra ambigua. Estas posibles traducciones se obtienen del diccionario de referencia (López-Ostenero, 2002).

Para comprobar si la coaparición de dos palabras en un documento es significativa, se define un modelo nulo en el que las palabras se distribuyen aleatoria e independientemente entre un conjunto de documentos de un corpus. Concretamente, se calcula la probabilidad de que dos palabras coincidan por puro azar. Este valor nos permite determinar un p-valor p para la coaparición de dos palabras. Si $p \ll 1$ se puede considerar que la aparición de las dos palabras en el mismo documento es significativa, y por lo tanto, es probable que su significado esté relacionado.

Concretamente, si dos palabras que se encuentran respectivamente en dos documentos n_1 y n_2 de entre los N que componen el corpus, para contar cuantos casos existen en los que dos palabras coincidan en exactamente k documentos, debemos tener en cuenta que hay cuatro tipos de documentos: k documentos que contienen ambas palabras, $n_1 - k$ documentos que contienen sólo la primera palabra, $n_2 - k$ documentos que contiene sólo la segunda palabra, y $N - n_1 - n_2 + k$ documentos que no contienen ninguna de las dos palabras. Por lo tanto, el número de disposiciones que buscamos viene dado por el coeficiente multinomial:

$$\binom{N}{k}\binom{N-k}{n_1-k}\binom{N-n_1}{n_2-k} \tag{1}$$

Así, la probabilidad de que dos palabras que aparecen en los documentos n_1 y n_2 respectivamente y que están distribuidas de forma aleatoria e independiente entre N documentos, coincidan en exactamente k de ellos viene dada por:

$$p(k) = \frac{\binom{N}{k}\binom{N-k}{n_1-k}\binom{N-n_1}{n_2-k}}{\binom{N}{n_1}\binom{N}{n_2}} \tag{2}$$

si $\max\{0, n_1 + n_2 - N\} \leq k \leq \min\{n_1, n_2\}$ y cero en otro caso.

Podemos escribir la ecuación (2) de una forma más fácil de tratar computacionalmente. Para ello introducimos la notación $(a)_b \equiv a(a-1)\cdots(a-b+1)$, para cualquier $a \geq b$, y sin pérdida de generalidad suponemos que la primera palabra es la más frecuente, es decir $n_1 \geq n_2 \geq k$. Entonces:

$$\begin{aligned} p(k) &= \frac{(n_1)_k (n_2)_k (N-n_1)_{n_2-k}}{(N)_{n_2}(k)_k} \\ &= \frac{(n_1)_k (n_2)_k (N-n_1)_{n_2-k}}{(N)_{n_2-k}(N-n_2+k)_k (k)_k}, \end{aligned} \tag{3}$$

donde en la segunda forma se ha usado la identidad $(a)_b = (a)_c(a-c)_{b-c}$ valida para $a \geq b \geq c$. La ecuación (3) se puede reescribir como

$$\begin{aligned} p(k) = &\prod_{j=0}^{n_2-k-1}\left(1 - \frac{n_1}{N-j}\right) \\ &\times \prod_{j=0}^{k-1} \frac{(n_1-j)(n_2-j)}{(N-n_2+k-j)(k-j)}. \end{aligned} \tag{4}$$

Esto nos permite determinar un p-valor para la coaparición de dos palabras como

$$p = \sum_{k \geq r} p(k), \tag{5}$$

donde r es el número de documentos en el corpus en el que se han encontrado realmente las dos palabras juntas. Si $p \ll 1$ podemos considerar que la aparición de las dos palabras en el mismo documento es significativa, y por lo tanto es probable que su significado esté relacionado. Podemos cuantificar aún más esta significancia tomando la mediana (correspondiente a $p = 1/2$) como una referencia y calculando el peso de un enlace como $\ell = -\log(2p)$, es decir una medida de cuanto se desvía de la mediana el valor real de r.

Con esto podemos construir un grafo que tiene las palabras como nodos, y conectar con un enlace los pares de palabras que coinciden en al menos un documento y con un peso de coaparición por debajo de un valor umbral. El peso ℓ asignado a los enlaces mide la desviación de la coaparición de las dos palabras con respecto al caso nulo. Al grafo resultante le denominamos grafo de coaparición.

3.3. Detección de comunidades

Una vez que se ha definido el grafo de coaparición se puede pasar a agrupar los nodos, por ejemplo, aplicando una descomposición en comunidades. Las comunidades son subgrafos cuyos nodos presentan algún tipo de afinidad estructural o dinámica, y por lo tanto es plausible suponer que cada comunidad comparte un sentido común, diferente del de las restantes comunidades. Existen diversas propuestas de algoritmos de extracción de comunidades. En los experimentos realizados se ha utilizado *Walktrap*, presentado por Pons y Latapy (Pons y Latapy, 2005). Este algoritmo es de los que se apoyan en la idea de que un *camino aleatorio* o *random walk* queda atrapado más fácilmente en las partes del grafo densamente conectadas, que corresponden a

las comunidades. También utiliza heurísticas para fusionar las comunidades iterativamente, hasta conseguir un conjunto óptimo de comunidades. Dado que la optimización de este algoritmo de comunidades se escapa de los objetivos principales de nuestro trabajo, hemos decidido adoptar los valores por defecto de su implementación (http://www-rp.lip6.fr/ latapy/PP/walktrap.html) para obtener la descomposición óptima.

3.3.1. Grafo de Comunidades

Las comunidades que se han obtenido son posteriormente representadas en un nuevo grafo para poder calcular las distancias entre ellas. Para la construcción de este nuevo grafo GC, se recorrerá el grafo de coaparición de palabras GP, y se generará un enlace entre dos comunidades C_1 y C_2 siempre que una palabra $x \in C_1$ esté enlazada en el grafo de coaparición GP con una palabra $y \in C_2$.

3.4. Desambiguación de la Palabra Objetivo

A través de la construcción de los grafos de coaparición y la detección de comunidades, se consigue una estructura de representación del conocimiento que nos permitirá realizar la tarea de desambiguación de las palabras objetivo.

3.4.1. Extracción del contexto

La competición del SemEval-2010 que estamos utilizando para evaluar nuestro sistema, proporciona tan solo una frase (en la que aparecía la palabra ambigua) cómo única información de contexto. Dicho contexto es el que se va a utilizar para buscar la traducción más probable como conocimiento auxiliar y seleccionar las posibles traducciones de la palabra ambigua. En la Figura 1 se muestra un esquema del funcionamiento de la desambiguación de la palabra objetivo.

En primer lugar, se realiza un análisis de la frase en la que aparece la palabra objetivo, y se extraen las palabras más representativas. En los experimentos realizados hemos tenido en cuenta los nombres, adjetivos y verbos, aunque el uso de verbos ha sido descartado debido a la reducción significativa del rendimiento del algoritmo. Una vez hecho esto, utilizando el diccionario, se obtienen todas las posibles traducciones de cada una de las palabras del contexto. Dado que los grafos con los que trabajamos están formados únicamente por nombres o nombres y adjetivos, se seleccionan las traducciones de las palabras del contexto que corresponden a estas categorías gramaticales.

Posteriormente se identifican aquellas comunidades que contienen al menos una traducción ya sea de las palabras del contexto o de la palabra objetivo. Como resultado, tenemos un conjunto de comunidades en las que aparece al menos una traducción de la palabra objetivo M_T, y otro conjunto de comunidades en las que aparece al menos una traducción de las palabras del contexto M_C. Utilizando el grafo de comunidades, se calculan las distancias entre cada comunidad dentro de M_T y cada comunidad de M_C. Teniendo en cuenta que una traducción de la palabra objetivo puede pertenecer a la misma comunidad que otras traducciones de las palabras del contexto, la distancia en este caso sería 1, y por tanto se normalizan el resto de las distancias sumando 1 a su valor original.

Nuestra hipótesis en este punto, se basa en que aquella traducción de la palabra objetivo que se encuentre más cerca de las traducciones de las palabras del contexto, tiene una mayor probabilidad de ser la traducción correcta. De esta forma hemos establecido una ponderación de cada una de las traducciones de la palabra objetivo en base a dos factores. El primero de ellos es la distancia entre las comunidades que contienen las traducciones de la palabra objetivo y las comunidades que contienen las traducciones de las palabras del contexto. El segundo se basa en la cantidad de traducciones de las palabras del contexto que contiene la comunidad considerada. El peso asignado a cada traducción, w_{t_j}, viene dado por la fórmula

$$w_t = \max_{M_C^i \in M_C} \frac{A_C^i}{(d_{M_C^i M_T^t} + 1)} \qquad (6)$$

donde A_C^i es el número de traducciones del contexto que contiene M_C^i, y $d_{M_C^i M_T^t}$ es el número de pasos (distancia) existente entre la comunidad M_C^i y la comunidad M_T^t, es decir, aquella en la que se encuentra la traducción analizada.

Esta ponderación obtenida por cada traducción de la palabra ambigua, se utiliza posteriormente para ordenar dichas traducciones.

4. Experimentación y resultados

En este apartado se expondrá el método que se ha seguido para realizar la evaluación del sistema propuesto, así como los resultados obtenidos y su comparación con otros sistemas.

4.1. Método de evaluación

La metodología seguida para realizar la evaluación del sistema es la misma utilizada en la ta-

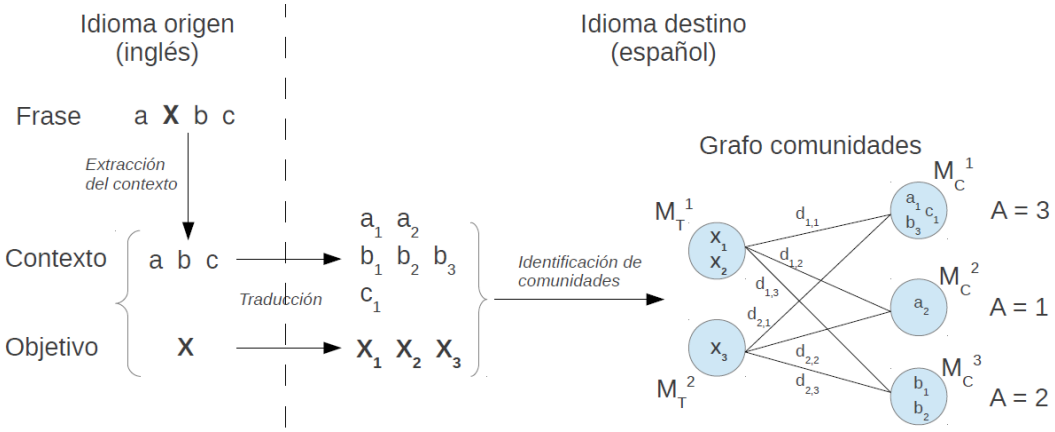

Figura 1: Diagrama del funcionamiento del algoritmo de desambiguación multilingüe de palabras.

rea 3 de la competición SemEval 2010, denominada *Cross-Lingual Word Sense Disambiguation* (Lefever y Hoste, 2010). Los conjuntos de datos proporcionados para la tarea consisten en un subconjunto *Trial*, compuesto por cinco palabras y veinte frases etiquetadas con la palabra a desambiguar para cada una de ellas, y un subconjunto *Test*, compuesto por veinte palabras, y cincuenta frases etiquetadas con la palabra ambigua. En (Lefever y Hoste, 2010) se puede encontrar una descripción más detallada del proceso de construcción de estos conjuntos de datos.

La evaluación se realiza sobre las palabras de test, a partir de un *Gold-Standard* que ofrece las traducciones más probables, ordenadas según su peso, para cada una de las palabras en cada uno de los contextos proporcionados. En la competición SemEval 2010 se realizaron dos tipos de evaluaciones. La primera denominada *Best*, permitía al sistema evaluado proponer tantas traducciones como decidiese que eran adecuadas, pero la puntuación se dividía por el número de palabras propuestas. En la segunda evaluación, denominada *Out-Of-Five*, se permitía al sistema proponer hasta un máximo de cinco traducciones para cada palabra, y la puntuación se obtiene únicamente en función de si las palabras propuestas se encuentran en el *Gold-Standard* y en qué posición. El sistema propuesto en el presente trabajo está enfocado hacia la evaluación *Out-Of-Five*, por lo que se proponen las cinco traducciones que el sistema considera más probables.

La puntuación de los sistemas se ofrece en términos de precisión y cobertura (*precision* y *recall*), y el ranking de los sistemas se obtiene a partir de la Medida-F (*F-Measure*).

4.2. Resultados

Como ya hemos indicado, en este trabajo nos centramos únicamente en la traducción del inglés al español. El sistema ha sido evaluado utilizando, por una parte, únicamente nombres para construir el grafo de coaparición de palabras, y por otra parte, utilizando nombres y adjetivos. También se ha variado el umbral a partir del cuál se considera que una coaparición de palabras es estadísticamente significativa, es decir, el valor máximo que puede tomar el *p-valor* explicado en la sección 3.2 para generar un enlace entre dos palabras en el grafo.

La Figura 2 nos muestra gráficamente los valores que toma la Medida-F para cada una de las configuraciones de parámetros. Los nombres se representan como “NN”, mientras que los adjetivos se representan como “JJ”. Como se puede observar, los mejores resultados, tanto para el grafo que utiliza sólo los nombres, como para el grafo que utiliza nombres y adjetivos, se consiguen con los umbrales de $10^{(-13)}$ y $10^{(-11)}$ respectivamente. La evolución de estos valores confirma la utilidad del algoritmo empleado, ya que el aumento del umbral repercute en la permanencia de unas relaciones cada vez más significativas. De esta forma, se prueba el hecho de que el grafo establece unos vínculos entre términos tan representativos que el hecho de eliminar enlaces superfluos se traduce en una mejora de los resultados. Esta mejora se mantiene hasta que el grafo comienza a perder enlaces realmente relevantes como se puede apreciar en la figura. También se muestra que los mejores resultados se obtienen cuando se utilizan nombres y adjetivos, que enriquecen la representación frente a la utilización únicamente de nombres.

En el Cuadro 1 se puede observar la compara-

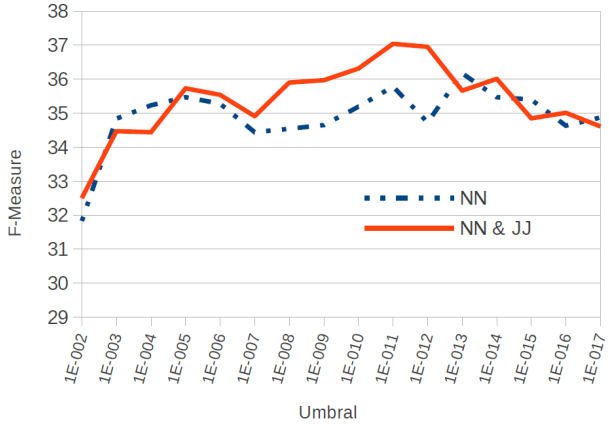

Figura 2: Evolución de la Medida-F según el umbral utilizado para los métodos basados en el uso de nombres (NN), y nombres y adjetivos (NN & JJ)

ción entre los resultados obtenidos por el sistema propuesto, y los obtenidos por los diversos sistemas no supervisados que compitieron en la tarea 3 del SemEval-2010. El sistema propuesto, en su configuración óptima (utilizando nombres y adjetivos para construir el grafo y con un umbral de 10^{-11}) supera a todos los sistemas no supervisados que presentaron resultados para el idioma español en dicha competición. Como se puede observar, nuestro algoritmo presenta el mejor valor de cobertura, en detrimento de algo de precisión. Esto es debido a que el método de evaluación elegido para desarrollar el sistema fue el *Out-Of-Five*, en donde se valora la proposición de un número de hasta cinco traducciones por encima de la selección de una única traducción.

Sistema	Medida-F	P	C
Propuesto	**37.04**	37.04	**37.04**
T3-COLEUR	35.67	35.84	35.46
UHD-1	34.95	**38.78**	31.81
UHD-2	34.22	37.74	31.30

Cuadro 1: Resultados en función de la Medida-F, Precisión (P) y Cobertura (C) obtenidos por nuestro algoritmo, en comparación con los sistemas no supervisados participantes en la tarea 3 del SemEval-2010.

5. *Conclusiones y Trabajo Futuro*

En este trabajo hemos abordado la tarea de la desambiguación del sentido de las palabras en un contexto multilingüe y con un enfoque no supervisado. La idea subyacente ha sido inducir los sentidos de las palabras en el idioma destino, en este caso el español, mediante una nueva técnica de construcción de grafos. Después hemos utilizado las traducciones de las palabras del contexto de la palabra origen para identificar la correspondencia más probable con las traducciones de la palabra considerada. Esta metodología, aplicada a los datos de la tarea 3 de la competición Semeval-2010, ha conseguido superar a todos los sistemas no supervisados que participaron en la tarea. Sin embargo, hay muchos aspectos de la propuesta cuya investigación nos proponemos abordar. Uno de ellos es refinar la medida de distancia entre comunidades utilizando los pesos de los enlaces del grafo de coaparición. Otro aspecto a estudiar, consiste en analizar distintas alternativas para tratar los casos de empates que se puedan producir al seleccionar la mejor traducción. También queremos investigar otros algoritmos de detección de comunidades así como aplicar nuestro algoritmo a otras lenguas.

Bibliografía

Agirre, Eneko, David Martínez, Oier Lopez de Lacalle, y Aitor Soroa. 2006. Two graph-based algorithms for state-of-the-art wsd. En *EMNLP*, páginas 585–593.

Agirre, Eneko y Aitor Soroa. 2007. Ubc-as: A graph based unsupervised system for induction and classification. En *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, páginas 346–349, Prague, Czech Republic, June. Association for Computational Linguistics.

Banea, Carmen y Rada Mihalcea. 2011. Word sense disambiguation with multilingual features. En *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, páginas 25–34. Association for Computational Linguistics.

Diab, Mona T. y Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. En *ACL*, páginas 255–262.

Fernandez-Ordonez, Erwin, Rada Mihalcea, y Samer Hassan. 2012. Unsupervised word sense disambiguation with multilingual representations. En *LREC*, páginas 847–851.

Guo, Weiwei y Mona Diab. 2010. Coleur and colslm: A wsd approach to multilingual lexical substitution, tasks 2 and 3 semeval 2010. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, páginas 129–133, Strouds-

burg, PA, USA. Association for Computational Linguistics.

Ide, Nancy y Jean Veronis. 1998. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24:1–40.

Klapaftis, Ioannis P. y Suresh Manandhar. 2008. Word sense induction using graphs of collocations. En *Proceeding of the 2008 conference on ECAI 2008*, páginas 298–302, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. En *MT summit*, volumen 5.

Lefever, Els y Veronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, páginas 15–20, Stroudsburg, PA, USA. Association for Computational Linguistics.

López-Ostenero, Fernando. 2002. *Un sistema interactivo para la búsqueda de información en idiomas desconocidos por el usuario*. Ph.D. tesis, Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia.

Màrquez, Lluís, Gerard Exsudero, David Martínez, y German Rigau. 2006. Supervised corpus-based methods for wsd. En *Word Sense Disambiguation: Algorithms and Applications*, volumen 33 de *Text, Speech and Language Technology*. Springer, Dordrecht, The Netherlands, páginas 167–216.

Martinez-Romo, Juan, Lourdes Araujo, Javier Borge-Holthoefer, Alex Arenas, José A. Capitán, y José A. Cuesta. 2011. Disentangling categorical relationships through a graph of co-occurrences. *Phys. Rev. E*, 84:046108, Oct.

Mihalcea, Rada. 2006. Knowledge-based methods for wsd. En *Word Sense Disambiguation: Algorithms and Applications*, volumen 33 de *Text, Speech and Language Technology*. Springer, Dordrecht, The Netherlands, páginas 107–132.

Ng, Hwee Tou, Bin Wang, y Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. En *ACL*, páginas 455–462.

Pedersen, Ted. 2006. Unsupervised corpus-based methods for WSD. En *Word Sense Disambiguation: Algorithms and Applications*. Springer, páginas 133–166.

Pons, P. y M. Latapy. 2005. Computing communities in large networks using random walks. *Lect. Notes Comput. Sci.*, 3733:284.

Resnik, Philip. 2004. Exploiting hidden meanings: Using bilingual text for monolingual annotation. En *Int. Conf. Computational Linguistics and Intelligent Text Processing (CICLing)*, páginas 283–299.

Resnik, Philip y David Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. En *Proceedings of international conference on new methods in language processing*, volumen 12, páginas 44–49. Manchester, UK.

Silberer, Carina y Simone Paolo Ponzetto. 2010. Uhd: Cross-lingual word sense disambiguation using multilingual co-occurrence graphs. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, páginas 134–137, Stroudsburg, PA, USA. Association for Computational Linguistics.

van Gompel, Maarten. 2010. Uvt-wsd1: A cross-lingual word sense disambiguation system. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, páginas 238–241, Stroudsburg, PA, USA. Association for Computational Linguistics.

Veronis, Jean. 2004. Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252.

Vilariño, Darnes, Carlos Balderas, David Pinto, Miguel Rodríguez, y Saul León. 2010. Fcc: Modeling probabilities with giza++ for task #2 and #3 of semeval-2. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, páginas 112–116, Stroudsburg, PA, USA. Association for Computational Linguistics.