Procesamiento del Lenguaje Natural, Revista nº 40, marzo de 2008, pp. 35-42 recibido 29-01-08, aceptado 03-03-08

Bases de Conocimiento Multilíngües para el Procesamiento Semántico a Gran Escala*

Multilingual Knowledge Resources for wide–coverage Semantic Processing

Montse Cuadros
cuadros@lsi.upc.edu
TALP Research Center, UPC
Barcelona, Spain

German Rigau
german.rigau@ehu.es
IXA Group, UPV/EHU
Donostia-San Sebastian, Spain

Resumen: Este artículo presenta el resultado del estudio de un amplio conjunto de bases de conocimiento multilíngües actualmente disponibles que pueden ser de interés para un gran número de tareas de procesamiento semántico a gran escala. El estudio incluye una amplia gama de recursos derivados de forma manual y automática para el inglés y castellano. Con ello pretendemos mostrar una imagen clara de su estado actual. Para establecer una comparación justa y neutral, la calidad de cada recurso se ha evaluado indirectamente usando el mismo método en dos tareas de resolución de la ambigüedad semántica de las palabras (WSD, del inglés Word Sense Disambiguation). En concreto, las tareas de muestra léxica del inglés del Senseval-3.
Palabras clave: Adquisición y Representación del Conocimiento Léxico, WSD

Abstract: This report presents a wide survey of publicly available multilingual Knowledge Resources that could be of interest for wide–coverage semantic processing tasks. We also include an empirical evaluation in a multilingual scenario of the relative quality of some of these large-scale knowledge resources. The study includes a wide range of manually and automatically derived large-scale knowledge resources for English and Spanish. In order to establish a fair and neutral comparison, the quality of each knowledge resource is indirectly evaluated using the same method on a Word Sense Disambiguation task (Senseval-3 English Lexical Sample Task).
Keywords: Adquisition and Representation of Lexical Knowledge, WSD

1. *Introdución*

El uso de bases de conocimiento de amplia cobertura, tales como WordNet (Fellbaum, 1998), se ha convertido en una práctica frecuente, y a menudo necesaria, de los sistemas actuales de Procesamiento del Lenguaje Natural (NLP, del inglés Natural Language Processing). Incluso ahora, la construcción de bases de conocimiento suficientemente grandes y ricas para un procesamiento semántico de amplia cobertura, requiere de un gran y costoso esfuerzo manual que involucra a grandes grupos de investigación durante largos períodos de desarrollo. De hecho, centenares de años/persona se han invertido en el desarrollo de wordnets para varios idiomas (Vossen, 1998). Por ejemplo, en más de diez años de construcción manual (desde 1995 hasta 2006, esto es desde la versión 1.5 hasta la 3.0), WordNet ha pasado de 103.445 a 235.402 relaciones semánticas[1]. Es decir, alrededor de unas mil nuevas relaciones por mes. Sin embargo, estas bases de conocimiento no parecen ser suficientemente ricas como para ser usadas directamente por aplicaciones avanzadas basadas en conceptos. Parece que estas aplicaciones no se mostrarán eficaces en dominios abiertos (y también en dominios específicos) sin un conocimiento semántico de amplia cobertura más detallado y más rico construido mediante procedimientos automáticos. Obviamente, este hecho ha sido un

* Este trabajo ha sido parcialmente financiado por grupo IXA de la UPV/EHU y los proyectos KNOW (TIN2006-15049-C03-01) y ADIMEN (EHU06/113)

[1] Las relaciones simétricas se han contado una sola vez.

obstáculo al progreso del estado del arte en NLP.

Afortunadamente, en los últimos años, la comunidad investigadora ha desarrollado un amplio conjunto de métodos y herramientas innovadoras para la adquisición automática de conocimiento léxico a gran escala a partir de fuentes estructuradas y no estructuradas. Entre otros podemos mencionar eXtended WordNet (Mihalcea y Moldovan, 2001), grandes colecciones de preferencias semánticas adquiridas de SemCor (Agirre y Martinez, 2001)o adquiridas de British National Corpus (BNC) (McCarthy, 2001), Topic Signatures[2] para cada synset adquiridas de la web (Agirre y de la Calle, 2004) o adquiridas del BNC (Cuadros, Padró, y Rigau, 2005). Evidentemente, todos estos recursos semánticos han sido adquiridos mediante un conjunto muy diferente de procesos, herramientas y corpus, dando lugar a un conjunto muy amplio y variado de nuevas relaciones semánticas entre synsets. De hecho, cada uno estos recursos semánticos presentan volúmenes y exactitudes muy distintas cuando se evalúan en un marco común y controlado (Cuadros y Rigau, 2006). De hecho, que sepamos, ningún estudio empírico se ha llevado a cabo tratando de ver la forma en que estos grandes recursos semánticos se complementan entre sí.

Además, dado que este conocimiento es independiente de idioma (conocimiento representado en el plano semántico, es decir, como relaciones entre conceptos), hasta la fecha ninguna evaluación empírica se ha llevado a cabo mostrando: a) hasta qué punto estos recursos semánticos adquiridos de un idioma (en este caso inglés) podrían ser de utilidad para otro (en este caso castellano), y b) cómo estos recursos se complementan entre sí.

Este artículo está organizado de la siguiente manera. Tras esta breve introducción, mostramos los recursos semánticos multilíngües que analizaremos. En la sección 3 presentamos el marco de evaluación multilíngüe utilizado en este estudio. La sección 4 describe los resultados cuando evaluamos para el inglés estos recursos semánticos a gran escala y en la sección 5 para el castellano. Por último, la sección 6 se presentan algunas observaciones finales y el trabajo futuro.

[2]Topic Signatures es el término en inglés para referirse a las palabras relacionadas con un tópico o tema.

2. Recursos Semánticos Multilíngües

La evaluación que aquí presentamos abarca una amplia variedad de recursos semánticos de gran tamaño: WordNet (WN) (Fellbaum, 1998), eXtended WordNet (Mihalcea y Moldovan, 2001), grandes colecciones de preferencias semánticas adquiridas de SemCor (Agirre y Martinez, 2001)o adquiridos del BNC (McCarthy, 2001), y Topic Signatures para cada synset adquiridas de la web (Agirre y de la Calle, 2004).

A pesar de que estos recursos se han obtenido utilizando diferentes versiones de WN, utilizando la tecnología para alinear automáticamente wordnets (Daudé, Padró, y Rigau, 2003), la mayoría de estos recursos se han integrado en un recurso común llamado Multilingual Central Repository (MCR) (Atserias et al., 2004). De esta forma, mantenemos la compatibilidad entre todas las bases de conocimiento que utilizan una versión concreta de WN como repositorio de sentidos. Además, estos enlaces permiten transportar los conocimientos asociados a un WN particular, al resto de versiones de WN.

2.1. MCR

El Multilingual Central Repository[3] (MCR) sigue el modelo propuesto por el proyecto EuroWordNet. EuroWordNet (Vossen, 1998) es una base de datos léxica multilingüe con wordnets de varias lenguas europeas, que están estructuradas como el WordNet de Princeton. El WordNet de Princeton contiene información sobre los nombres, verbos, adjetivos y adverbios en inglés y está organizado en torno a la noción de un synset. Un synset es un conjunto de palabras con la misma categoría morfosintáctica que se pueden intercambiar en un determinado contexto.

La versión actual del MCR (Atserias et al., 2004) es el resultado del proyecto europeo MEANING del quinto programa marco[4]. El MCR integra siguiendo el modelo de EuroWordNet, wordnets de cinco idiomas diferentes, incluido el castellano (junto con seis versiones del WN inglés). Los wordnets están vinculados entre sí a través del Inter-Lingual-Index (ILI) permitiendo la conexión de las

[3]http://adimen.si.ehu.es/cgi-bin/wei5/public/wei.consult.perl

[4]http://nipadio.lsi.upc.es/~nlp/meaning

palabras en una lengua a las palabras equivalentes en cualquiera de las otras lenguas integradas en el MCR. De esta manera, el MCR constituye un recurso lingüístico multilíngüe de gran tamaño útil para un gran número de procesos semánticos que necesitan de una gran cantidad de conocimiento multilíngüe para ser instrumentos eficaces. Por ejemplo, el synset en inglés <*party*, *political_party*> está vinculado a través del ILI al synset en castellano <*partido*, *partido_político*>.

El MCR también integra WordNet Domains (Magnini y Cavaglià, 2000), nuevas versiones de los Base Concepts y la Top Concept Ontology (Álvez et al., 2008), y la ontología SUMO (Niles y Pease, 2001). La versión actual del MCR contiene 934.771 relaciones semánticas entre synsets, la mayoría de ellos adquiridos automáticamente[5]. Esto representa un volumen casi cuatro veces más grande que el de Princeton WordNet (235.402 relaciones semánticas únicas en WordNet 3.0).

En lo sucesivo, nos referiremos a cada recurso semántico de la siguiente forma:

WN (Fellbaum, 1998): Este recurso contiene las relaciones directas y no repetidas codificadas en WN1.6 y WN2.0 (por ejemplo, tree#n#1–hyponym–>teak#n#2). También hemos estudiado WN^2 utilizando las relaciones a distancia 1 y 2, WN^3 utilizando las relaciones a distancias 1 a 3 y WN^4 utilizando las relaciones a distancias 1 a 4.

XWN (Mihalcea y Moldovan, 2001): Este recurso contiene las relaciones directas codificadas en eXtended WN (por ejemplo, teak#n#2–gloss–>wood#n#1).

WN+XWN: Este recurso contiene las relaciones directas incluidas en WN y XWN. También hemos estudiado $(WN+XWN)^2$ (utilizando relaciones de WN o XWN a distancias 1 y 2).

spBNC (McCarthy, 2001): Este recurso contiene 707.618 preferencias de selección con los sujetos y objetos típicos adquiridos del BNC.

spSemCor (Agirre y Martinez, 2001): Este recurso contiene las preferencias de selección con los sujetos y los objetos típicos adquiridos de SemCor (por ejemplo, read#v#1–tobj–>book#n#1).

MCR (Atserias et al., 2004): Este recurso contiene las relaciones directas incluidas en el MCR. Sin embargo, en los experimentos descritos a continuación se excluyó el recurso spBNC debido a su pobre rendimiento. Así, el MCR contiene las relaciones directas de WN , XWN, y spSemCor. Obsérvese que el MCR no incluye las relaciones indirectas de $(WN+XWN)^2$. No obstante, también hemos evaluado $(MCR)^2$ (utilizando las relaciones a distancia 1 y 2), que sí integra las relaciones de $(WN+XWN)^2$.

[5]No consideramos las preferencias de selección adquiridos del BNC (McCarthy, 2001).

2.2. Topic Signatures

Las Topic Signatures (TS) son vectores de palabras relacionadas con un tema (o tópico) (Lin y Hovy, 2000). Las TS pueden ser construidas mediante la búsqueda en un corpus de gran tamaño del contexto de un tema (o tópico) objetivo. En nuestro caso, consideramos como un tema (o tópico) el sentido de una palabra.

Para este estudio hemos usado dos conjuntos de TS distintos. Las primeras TS constituyen uno de los mayores recursos semánticos disponibles actualmente con alrededor de 100 millones de relaciones semánticas (entre synsets y palabras) que ha sido adquirido automáticamente de la web (Agirre y de la Calle, 2004). Las segundas TS se han obtenido directamente de SemCor.

TSWEB[6]: Inspirado en el trabajo de (Leacock, Chodorow, y Miller, 1998), estas Topic Signatures se adquirieron utilizando para la construcción de la consulta del tópico (o sentido de WN en nuestro caso), los sentidos monosémicos próximos al tópico en WordNet (esto es, sinónimos, hiperónimos, hipónimos directos e indirectos, y hermanos), consultando en Google y recuperando hasta un millar de fragmentos de texto por consulta (es decir, por sentido o tópico), y extrayendo de los fragmentos las palabras con frecuencias distintivas usando TFIDF. Para estos experimentos, se ha utilizado como máximo las primeras 700 palabras distintivas de cada TS resultante.

Debido a que éste es un recurso semántico entre sentidos y palabras, no es posible transportar sus relaciones al wordnet castellano sin introducir gran cantidad de errores.

El cuadro 1 presenta un ejemplo de TSWEB para el primer sentido de la palabra *party*.

TSSEM: Estas TS se han construido utilizando SemCor, un corpus en inglés donde todas sus palabras han sido anotadas

[6]`http://ixa.si.ehu.es/Ixa/resources/~sensecorpus`

democratic	0.0126
tammany	0.0124
alinement	0.0122
federalist	0.0115
missionary	0.0103
anti-masonic	0.0083
nazi	0.0081
republican	0.0074
alcoholics	0.0073

Cuadro 1: Topic Signature de party#n#1 obtenida de la web (9 de las 15.881 palabras totales)

political_party#n#1	2.3219
party#n#1	2.3219
election#n#1	1.0926
nominee#n#1	0.4780
candidate#n#1	0.4780
campaigner#n#1	0.4780
regime#n#1	0.3414
government#n#1	0.3414
authorities#n#1	0.3414

Cuadro 2: Topic Signature para party#n#1 obtenida de SemCor (9 de los 719 sentidos totales)

semánticamente. Este corpus tiene un total de 192.639 palabras lematizadas y etiquetadas con su categoría y sentido según WN1.6. Para cada sentido objetivo (o tópico), obtuvimos todas las frases donde aparecía ese sentido. De esta forma derivamos un subcorpus de frases relativas al sentido objetivo. A continuación, para cada subcorpus se obtuvo su TS de sentidos utilizando TFIDF.

En el cuadro 2, mostramos los primeros sentidos obtenidos para party#n#1.

Aunque hemos probado con otras medidas, los mejores resultados se han obtenido utilizando la fórmula TFIDF (Agirre y de la Calle, 2004).

$$TFIDF(w,C) = \frac{wf_w}{max_w wf_w} \times log\frac{N}{Cf_w} \quad (1)$$

Donde w es la palabra del contexto, wf la frecuencia de la palabra, C la colección (todo el corpus reunido para un determinado sentido), y Cf es la frecuencia en la colección.

El número total de las relaciones entre synsets de WN adquiridos de SemCor es 932.008. En este caso, debido al menor tamaño del wordnet castellano, el número total de las relaciones transportadas es de sólo 586.881.

3. Marco de evaluación

Con el fin de comparar los distintos recursos semánticos descritos en la sección anterior, hemos evaluado todos estos recursos como Topic Signatures (TS). Esto es, para cada synset (o tópico), tendremos un simple vector de palabras con pesos asociados. Este vector de palabras se construye reuniendo todas las palabras que aparecen directamente relacionados con un synset. Esta simple representación intenta ser lo más neutral posible respecto a los recursos utilizados.

Todos los recursos se han evaluado en una misma tarea de WSD. En particular, en la sección 4 hemos utilizado el conjunto de nombres de la tarea de muestra léxica en inglés de Senseval-3 (Senseval-3 English Lexical Sample task) que consta de 20 nombres, y en la sección 5 hemos utilizado el conjunto de nombres de la tarea de muestra léxica en castellano de Senseval-3 (Senseval-3 Spanish Lexical Sample task) que consta de 21 nombres. Ambas tareas consisten en determinar el sentido correcto de una palabra en un contexto. Para la tarea en inglés se usó para la anotación los sentidos de WN1.7.1. Sin embargo, para el castellano se desarrolló especialmente para la tarea el diccionario MiniDir. La mayoría de los sentidos de MiniDir tienen vínculos a WN1.5 (que a su vez está integrado en el MCR, y por tanto enlazado al wordnet castellano). Todos los resultados se han evaluado en los datos de prueba usando el sistema de puntuación de grano fino proporcionado por los organizadores. Para la evaluación hemos usado sólo el conjunto de nombres etiquetados porque TSWEB se contruyó sólo para los nombres, y porque la tarea de muestra léxica para el inglés usa como conjunto de sentidos verbales aquellos que aparecen en el diccionario WordSmyth (Mihalcea, T., y A., 2004), en lugar de los que aparecen en WordNet.

Así, el mismo método de WSD se ha aplicado a todos los recursos semánticos. Se realiza un simple recuento de las palabras coincidentes entre aquellas que aparecen en la Topic Signature de cada sentido de la palabra objetivo y el fragmento del texto de test[7]. El synset que tiene el recuento mayor es seleccionado. De hecho, se trata de un méto-

[7]También consideramos los términos multipalabra que aparecen en WN.

do muy simple de WSD que sólo considera la información de contexto en torno a la palabra que se desea interpretar. Por último, debemos señalar que los resultados no están sesgados (por ejemplo, para resolver empates entre sentidos), mediante el uso del sentido más frecuente en WN o cualquier otro conocimiento estadístico.

A modo de ejemplo, el cuadro 3 muestra uno de los textos de prueba de Senseval-3 correspondiente al primer sentido de la palabra *party*. En negrita se muestran las palabras que aparecen en la TS correspondiente al sentido party#n#1 de la TSWEB.

4. Evaluación para el inglés

4.1. Referencias básicas para el English

Hemos diseñado una serie de referencias básicas con el fin de establecer un marco de evaluación que nos permita comparar el rendimiento de cada recurso semántico en la tarea WSD en inglés.

RANDOM: Para cada palabra este método selecciona un sentido al azar. Esta referencia puede considerarse como un límite inferior.

SEMCOR-MFS: Esta referencia selecciona el sentido más frecuente de la palabra según SemCor.

WN-MFS: Esta referencia selecciona el sentido más frecuente según WN (es decir, el primer sentido en WN1.6). Los sentidos de las palabras en WN se ordenaron utilizando las frecuencias de SemCor y otros corpus anotados con sentidos. Así, WN-MFS y SemCor-MFS son similares, pero no iguales.

TRAIN-MFS: Esta referencia selecciona el sentido más frecuente de la palabra objetivo en el corpus de entrenamiento.

TRAIN: Esta referencia utiliza el corpus de entrenamiento de cada sentido proporcionado por Senseval-3 construyendo directamente una TS con las palabras de su contexto y utilizando la medida TFIDF. Téngase en cuenta que en los marcos de evaluación de WSD, este es un sistema muy básico. Sin embargo, en nuestro marco de evaluación, este sistema "de referencia"podría ser considerado como un límite superior. No esperamos obtener mejores palabras relativas a un sentido que de su propio corpus.

4.2. Evaluación de cada recurso en inglés

El cuadro 4 presenta ordenadas por la medida F1, las referencias y el rendimiento de cada uno de los recursos presentados en la sección 2 y el tamaño medio de las TS por sentido de palabra. El tamaño medio de las TS de cada recurso es el número de palabras asociadas a un synset de promedio. Obviamente, los mejores recursos serán aquellos que obtengan los mejores resultados con un menor número de palabras asociadas al synset. Los mejores resultados de precisión, recall y medida F1 se muestran en negrita. También hemos marcado en cursiva los resultados de los sistemas de referencia. Los mejores resultados son obtenidos por TSSEM (con F1 de 52,4). El resultado más bajo se obtiene por el conocimiento obtenido directamente de WN debido principalmente a su escasa cobertura (R, de 18,4 y F1 de 26,1). También es interesante notar que el conocimiento integrado en el (MCR) aunque en parte derivado por medios automáticos obtiene mucho mejores resultados en términos de precisión, recall y medida F1 que utilizando cada uno de los recursos que lo integran por separado (F1 con 18,4 puntos más que WN, 9,1 más que XWN y 3,7 más que spSemCor).

A pesar de su pequeño tamaño, los recursos derivados de SemCor obtienen mejores resultados que sus homólogos usando corpus mucho mayores (TSSEM vs. TSWEB y spSemCor vs. spBNC).

En cuanto a los sistemas de referencia básicos, todos los recursos superan RANDOM, pero ninguno logra superar ni WN-MFS, ni TRAIN-MFS, ni TRAIN. Sólo TSSEM obtiene mejores resultados que SEMCOR-MFS y está muy cerca del sentido más frecuente de WN (WN-MFS) y el corpus de entrenamiento (TRAIN-MFS).

En cuanto a las expansiones y otras combinaciones, el rendimiento de WN se mejora utilizando palabras a distancias de hasta 2 (F1 de 30,0), y hasta 3 (F1 de 34,8), pero disminuye utilizando distancias de hasta 4 (F1 de 33,2). Curiosamente, ninguna de estas ampliaciones de WN logra los resultados de XWN (F1 de 35,4). Por último, $(WN+XWN)^2$ va mejor que WN+XWN y $(MCR)^2$ ligeramente mejor que MCR[8].

[8]No se han probado extensiones superiores.

<instance id="party.n.bnc.00008131"docsrc="BNC"> <context> Up to the **late** 1960s , catholic nationalists were **split** between two **main political** groupings . There was the Nationalist Party , a weak **organization** for which **local** priests had to **provide** some **kind** of legitimation . As a <head>party</head> , it really only exercised a modicum of **power** in relation to the Stormont administration . Then there were the republican **parties** who focused their attention on Westminster **elections** . The disorganized **nature** of catholic nationalist **politics** was only turned round with the emergence of the **civil rights movement** of 1968 and the subsequent forming of the SDLP in 1970 . </context> </instance>

Cuadro 3: Ejemplo de prueba número 00008131 para party#n cuyo sentido correcto es el primero.

KB	P	R	F1	Size
TRAIN	*65.1*	*65.1*	*65.1*	
TRAIN-MFS	*54.5*	*54.5*	*54.5*	
WN-MFS	*53.0*	*53.0*	*53.0*	
TSSEM	**52.5**	**52.4**	**52.4**	103
SEMCOR-MFS	*49.0*	*49.1*	*49.0*	
MCR^2	45.1	45.1	45.1	26,429
MCR	45.3	43.7	44.5	129
spSemCor	43.1	38.7	40.8	56
$(WN+XWN)^2$	38.5	38.0	38.3	5,730
WN+XWN	40.0	34.2	36.8	74
TSWEB	36.1	35.9	36.0	1,721
XWN	38.8	32.5	35.4	69
WN^3	35.0	34.7	34.8	503
WN^4	33.2	33.1	33.2	2,346
WN^2	33.1	27.5	30.0	105
spBNC	36.3	25.4	29.9	128
WN	44.9	18.4	26.1	14
RANDOM	*19.1*	*19.1*	*19.1*	

Cuadro 4: Resultados de los recursos evaluados individualmente para el Inglés según las medidas de P, R y F1.

4.3. Combinación de Recursos

Con el objetivo de evaluar de forma más detallada la contribución que tiene cada recurso, proporcionamos un pequeño análisis de su aportación combinada. Las combinaciones se han evaluado usando tres estrategias básicas diferentes (Brody, Navigli, y Lapata, 2006).

DV (del inglés Direct Voting): Cada recurso semántico tiene un voto para el sentido predominante de la palabra a interpretar. Se escoge el sentido con más votos.

PM (del inglés Probability Mixture): Cada recurso semántico proporciona una distribución de probabilidad sobre los sentidos de las palabras que serán interpretadas. Estas probabilidades (normalizadas), serán contabilizadas y se escogerá el sentido con mayor probabilidad.

Rank: Cada recurso semántico proporciona un orden de sentidos de la palabra que se quiere interpretar. Para cada sentido, se agregarán las posiciones de cada uno de los recursos evaluados. El sentido que tenga un orden menor (más cercano a la primera posición), será el escogido como el correcto.

El cuadro 5 presenta las medidas de F1 correspondientes a las mejores combinaciones de dos, tres y cuatro recursos usando los tres métodos de combinación.

Observando el método de combinación aplicado, los métodos de la Combinación de Probabilidad (PM) y la combinación basada en el orden (Rank) son los que dan mejores resultados, comparando con el de Combinación Directa (DV), sin embargo, el método basado en el orden da mejores resultados.

La combinación de los cuatro recursos semánticos obtiene mejores resultados que usando sólo tres, dos o un recurso. Parece ser que la combinación de los recursos aporta un conocimiento que no tienen los diferentes recursos individualmente. En este caso, 19.5 puntos por encima que TSWEB, 17.25 puntos por encima de $(WN+XWN)^2$, 11.0 puntos por encima de MCR y 3.1 puntos por encima de TSSEM.

Observando las referencias básicas, esta combinación supera el sentido más frecuente de SemCor (SEMCOR-MFS con F1 de 49.1), WN (WN-MFS con F1 de 53.0) y el conjunto de entrenamiento (TRAIN-MFS con F1 de 54.5). Este hecho, indica que la combinación resultante de recursos a gran escala codifica el conocimiento necesario para tener un etiquetador de sentidos para el inglés que se comporta como un etiquetador del sentido más frecuente. Es importante mencionar que el sentido más frecuente de una palabra, de acuerdo con el orden de sentidos de WN es un desafio difícil de superar en las tareas de WSD (McCarthy et al., 2004).

KB	PM	DV	Rank
2.system-comb: MCR+TSSEM	52.3	45.4	**52.7**
3.system-comb: MCR+TSSEM+(WN+XWN)^2	52.6	37.9	**54.6**
4.system-comb: MCR+(WN+XWN)^2+TSWEB+TSSEM	53.1	32.7	**55.5**

Cuadro 5: Combinaciones de 2, 3, y 4 sistemas según la medida de F1

5. Evaluación en castellano

Del mismo modo que en el caso del inglés, hemos definido unas referencias básicas para poder establecer un marco de evaluación completo y comparar el comportamiento relativo de cada recurso semántico cuando es evaluado en la tarea de WSD en castellano.

RANDOM: Para cada palabra este método selecciona un sentido al azar. Esta referencia puede considerarse como un límite inferior.

Minidir-MFS: Esta referencia selecciona el sentido más frecuente de la palabra según el diccionario Minidir. Minidir es un diccionario construido para la tarea de WSD. La ordenación de sentidos de palabras corresponde exactamente a la frecuencia de los sentidos de palabras del conjunto de entrenamiento. Por eso, Minidir-MFS és el mismo que TRAIN-MFS.

TRAIN: Esta referencia usa el conjunto de entrenamiento para directamente construir una Topic Signature para cada sentido de palabra usando la medida de TFIDF. Igual que para el inglés, en nuestro caso, esta referencia puede considerarse como un límite superior.

Debemos indicar que el WN castellano no codifica la frecuencia de los sentidos de las palabras y que para el castellano no hay disponible ningún corpus suficientemente grande que esté etiquetado a nivel de sentido del estilo del italiano[9].

Además, solamente pueden ser transportadas de un idioma a otro sin introducir demasiados errores las relaciones que existan en un recurso entre sentidos[10]. Como TSWEB relaciona palabras en inglés a un synset, no ha sido transportado ni evaluado al castellano.

5.1. Evaluando cada recurso del castellano por separado

El cuadro 6 presenta las medidas de precisión (P), recall (R) y F1 de las diferentes

Knowledge Bases	P	R	F1	Size
TRAIN	81.8	68.0	74.3	
MiniDir-MFS	67.1	52.7	59.2	
MCR	46.1	41.1	**43.5**	66
WN^2	56.0	29.0	42.5	51
(WN+XWN)^2	41.3	**41.2**	41.3	1,892
TSSEM	33.6	33.2	33.4	208
XWN	42.6	27.1	33.1	24
WN	**65.5**	13.6	22.5	8
RANDOM	21.3	21.3	21.3	

Cuadro 6: Resultados de los recursos evaluados individualmente pare el castellano según las mediadas de P, R y F1.

referencias básicas y recursos semánticos, ordenados por la medida de F1. En cursiva aparecen las referencias y en negrita los mejores resultados. Para el castellano, el recurso TRAIN ha sido evaluado con un tamaño de vector máximo de 450 palabras. Como se esperaba, RANDOM obtiene el menor resultado, y el sentido más frecuente obtenido de Minidir (Minidir-MFS, que es igual a TRAIN-MFS) es bastante más bajo que las TS obtenidas del corpus de entrenamiento (TRAIN).

WN obtiene la precision más alta (P de 65.5) pero dado su pequeña cobertura (R de 13.6), tiene la F1 más baja (F1 de 22.5). Es interesante notar que en terminos de precisión, recall y F1, el conocimiento integrado en el MCR supera a los resultados de TSSEM. Este hecho, posiblemente indica que el conocimiento actualmente contenido en el MCR es más robusto que TSSEM. Este hecho también parece indicar que el conocimiento de tópico obtenido de un corpus anotado a nivel de sentido de un idioma, no puede ser transportado directamente a otro idioma. Otros posibles motivos de los bajos resultados podrían ser el menor tamaño de los recursos en castellano (comparándolos con los existentes en inglés) o los diferentes marcos de evaluación, incluyendo el diccionario (diferenciación de sentidos y enlace a WN).

Observando los sistemas de referencia, todos los recursos de conocimiento superan

[9] http://multisemcor.itc.it/

[10] Es decir, relaciones semánticas synset a synset.

RANDOM, pero ninguno de ellos llega a Minidir-MFS (que es igual a TRAIN-MFS) ni a TRAIN.

De todas formas, podemos remarcar que el conocimiento contenido en el MCR (F1 de 43.5), parcialmente derivado con medios automáticos y transportado al WN castellano del ingles, casi dobla los resultados del WN castellano original (F1 de 22.5).

6. Conclusiones

Creemos, que un procesamiento semántico de amplia cobertura (como WSD) debe basarse no sólo en algoritmos sofisticados sino también en aproximaciones basadas en grandes bases de conocimiento. Los resultados presentados en este trabajo, sugieren que es necesaria mucha más investigación en la adquisición y uso de recursos semánticos a gran escala.

Además, el hecho que esos recursos presenten relaciones semánticas a nivel conceptual, nos permite trasladar estas relaciones para ser evaluadas en otros idiomas.

Por lo que sabemos, esta es la primera vez que un estudio empírico demuestra que las bases de conocimiento adquiridas automáticamente obtienen mejores resultados que los recursos derivados manualmente, y que la combinación del conocimiento contenido en estos recursos sobrepasa al clasificador que usa el sentido más frecuente para el inglés. Tenemos planificada la validación empírica de esta hipótesis en las tareas donde se interpretan todas las palabras de un texto *all-words*.

Bibliografía

Agirre, E. y O. Lopez de la Calle. 2004. Publicly available topic signatures for all wordnet nominal senses. En *Proceedings of LREC*, Lisbon, Portugal.

Agirre, E. y D. Martinez. 2001. Learning class-to-class selectional preferences. En *Proceedings of CoNLL*, Toulouse, France.

Álvez, J., J. Atserias, J. Carrera, S. Climent, A. Oliver, y G. Rigau. 2008. Consistent annotation of eurowordnet with the top concept ontology. En *Proceedings of Fourth International WordNet Conference (GWC'08)*.

Atserias, J., L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, y Piek Vossen. 2004. The meaning multilingual central repository. En *Proceedings of GWC*, Brno, Czech Republic.

Brody, S., R.Ñavigli, y M. Lapata. 2006. Ensemble methods for unsupervised wsd. En *Proceedings of COLING-ACL*, páginas 97–104.

Cuadros, M., L. Padró, y G. Rigau. 2005. Comparing methods for automatic acquisition of topic signatures. En *Proceedings of RANLP*, Borovets, Bulgaria.

Cuadros, M. y G. Rigau. 2006. Quality assessment of large scale knowledge resources. En *Proceedings of EMNLP*.

Daudé, J., L. Padró, y G. Rigau. 2003. Validation and Tuning of Wordnet Mapping Techniques. En *Proceedings of RANLP*, Borovets, Bulgaria.

Fellbaum, C., editor. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.

Leacock, C., M. Chodorow, y G. Miller. 1998. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24(1):147–166.

Lin, C. y E. Hovy. 2000. The automated acquisition of topic signatures for text summarization. En *Proceedings of COLING*. Strasbourg, France.

Magnini, B. y G. Cavaglià. 2000. Integrating subject field codes into wordnet. En *Proceedings of LREC*, Athens. Greece.

McCarthy, D. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Aternations, Subcategorization Frames and Selectional Preferences*. Ph.D. tesis, University of Sussex.

McCarthy, D., R. Koeling, J. Weeds, y J. Carroll. 2004. Finding predominant senses in untagged text. En *Proceedings of ACL*, páginas 280–297.

Mihalcea, R. y D. Moldovan. 2001. extended wordnet: Progress report. En *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA.

Mihalcea, R., Chlovski T., y Killgariff A. 2004. The senseval-3 english lexical sample task. En *Proceedings of ACL/SIGLEX Senseval-3*, Barcelona.

Niles, I. y A. Pease. 2001. Towards a standard upper ontology. En *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, páginas 17–19. Chris Welty and Barry Smith, eds.

Vossen, P., editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* . Kluwer Academic Publishers .