Bases de Conocimiento Multilíngües para el Procesamiento Semántico a Gran Escala

Montse Cuadros
cuadros@lsi.upc.edu
TALP Research Center
Universitat Politécnica de Catalunya
Barcelona, Spain

German Rigau
german.rigau@ehu.es
IXA Group
Euskal Herriko Unibersitatea
Donostia-San Sebastian, Spain

Resumen

Este informe presenta el resultado de un amplio estudio de las bases de conocimiento multilíngües actualmente disponibles que pueden ser de interés para un gran número de tareas de procesamiento semántico a gran escala. También incluimos una evaluación empírica en un escenario multilíngüe de la calidad relativa de algunas de estas bases de conocimiento a gran escala. El estudio incluye una amplia gama de recursos derivados de forma manual y automática. Por lo que sabemos, esta es la primera vez que se ha realizado un estudio de la calidad relativa de dichos recursos. Con ello pretendemos mostrar una imágen 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 ambiedad semántica de las palabras (WSD, del inglés Word Sense Disambiguation). En concreto, las tareas de muestra léxica del inglés y castellano de Senseval-3. Primero, el estudio empírico demuestra que los recursos de conocimiento adquiridos automáticamente obtienen mejores resultados que los recursos del conocimiento derivados manualmente, y que la combinación del conocimiento contenido en estos recursos sobrepasa al clasificador que usa el sentido más frecuente para el inglés. En segundo lugar, también demostramos que este conocimiento a gran escala adquirido a partir de una lengua se puede transportar con éxito a otros idiomas (en este caso, castellano). Finalmente, concluimos con algunas perspectivas interesantes de trabajo futuro.

1. Introdución

El uso de bases de conocimiento de amplia cobertura, tales como WordNet [14], 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 [27]. 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 relaciones semánticas a 235.402 relaciones[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 obstaculizado seriamente el estado del arte de las aplicaciones avanzadas de 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 [21], grandes colecciones de preferencias semánticas adquiridas de SemCor [2, 3] o adquiridas de British National Corpus (BNC) [19], Topic Signatures[2] para cada synset adquiridos de la web [1] o adquiridos del BNC [8]. 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 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 [10]. Sin embargo, en la medida de lo que sabemos, 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 lenguaje (conocimiento representado en el plano semántico, como relaciones entre conceptos), hasta la fecha ninguna evaluación empírica se ha llevado a cabo mostrando hasta qué punto estos recursos semánticos adquiridos de un idioma (en este caso inglés) podrían ser de utilidad para otro (en este caso castellano).

Este artículo está organizado de la siguiente manera. Tras esta breve introducción, mostramos los recursos semánticos multilíngües que analizaremos. En

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

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

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. La sección 6 presenta una propuesta para la construcción automática de bases de conocimiento semántico muy denso a partir de corpus. Por último, la sección 7 se presentan algunas observaciones finales y el trabajo futuro.

2. Recursos Semánticos Multilíngües

La evaluación que aquí presentamos abarca una amplia variedad de recursos semánticos de gran tamaño: WordNet (WN) [14], eXtended WordNet [21], grandes colecciones de preferencias semánticas adquiridas de SemCor [2, 3] o adquiridos del BNC [19], y Topic Signatures para cada synset adquirido de la web [1] o SemCor [8].

A pesar de que estos recursos se han obtenido utilizando diferentes versiones de WN, utilizando la tecnología para alinear automáticamente wordnets [12], la mayoría de estos recursos se han integrado en un recurso común llamado Multilingual Central Repository (MCR) [5]. 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 las versiones de WN.

2.1. Multilingual Central Repository

El Multilingual Central Repository (MCR)[3] sigue el modelo propuesto por el proyecto EuroWordNet. EuroWordNet [27] es una base de datos léxica multiline 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. Por ejemplo, <*party*, *political_party*> forma un synset, ya que ambas palabras pueden ser utilizadas para referirse al mismo concepto. Un synset se describe a menudo mediante una glosa o definición, en este caso: "an organization to gain political power." Por último, los synsets pueden estar relacionados entre sí por relaciones semánticas, como hiponimia, meronimia, causa, etc.

La versión actual del MCR [5] 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 palabras en una lengua a las palabras equivalentes en cualquiera de las otras lenguas integradas en

[3] http://adimen.si.ehu.es/cgi-bin/wei5/public/wei.consult.perl
[4] http://nipadio.lsi.upc.es/~nlp/meaning

Fuente	**#relationes**
Princeton WN1.6	138.091
Preferencias de Selección de SemCor	203.546
Nuevas relaciones de WN2.0	42.212
Relationes Gold de eXtended WN	17.185
Relationes Silver de eXtended WN	239.249
Relationes Normal de eXtended WN	294.488
Total inglés	934.771
Total castellano	517.279

Cuadro 1: Relaciones semánticas incorporadas al MCR

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 [18], nuevas versiones de los Base Concepts y la Top Concept Ontology [4], y la ontología SUMO [25]. La versión actual del MCR contiene 934.771 relaciones semánticas entre synsets, la mayoría de ellos adquiridos automáticamente [5]. Esto representa un volúmen casi cuatro veces más grande que el de Princeton WordNet (235.402 relaciones semánticas únicas en WordNet 3.0).

El cuadro 1 indica el número de relaciones semánticas en el MCR entre pares de synsets. Como la versión actual del wordnet en castellano no tiene traducción equivalente para todas los synsets en inglés[6], el número total de las relaciones transportadas es alrededor de la mitad de las relaciones existentes para el inglés.

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

WN [14]: 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 (por ejemplo, tree#n#1–related–>teak#n#1 dado que tree#n#1–hyponym–>teak#n#2 y teak#n#2–has_substance–>teak#n#1), WN^3 utilizando las relaciones a distancias 1 a 3 (por ejemplo, tree#n#1–related3–>wood#n#1 dado que tree#n#1–hyponym–>teak#n#2, teak#n#2–has_substance–>teak#n#1 y teak#n#1–hypernym–>wood#n#1) y WN^4 utilizando las relaciones a distancias 1 a 4 (por ejemplo, tree#n#1–related4–>lumber#n#1 ya que tree#n#1–hyponym–>teak#n#2, teak#n#2–has_substance–>teak#n#1, teak#n#1–hypernym–>wood#n#1 y wood#n#1–substance_of–>lumber#n#1).

XWN [21]: Este recurso contiene las relaciones directas codificadas en eX-

[5] No consideramos las preferencias de selección adquiridos del BNC

[6] Actualmente, el wordnet en castellano tiene equivalentes de traducción para 62.720 synsets del WN inglés

tended WN (por ejemplo, teak#n#2–gloss–>wood#n#1).

WN+XWN: Este recurso contiene las relaciones directas incluidas en el WN y XWN. También hemos estudiado $(WN+XWN)^2$ (ya sea utilizando relaciones de WN o XWN a distancias 1 y 2, por ejemplo, tree#n#1–related–>wood#n#1).

spBNC [19]: Este recurso contiene 707.618 preferencias de selección con los sujetos y objetos típicos adquiridos del BNC.

spSemCor [3]: 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 [5]: 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 (como tree#n#1–hyponym–>teak#n#2), XWN (como teak#n#2–gloss–>wood#n#1), y spSemCor (como read#v#1–tobj–>book#n#1). Obsérvese que el MCR no incluye las relaciones indirectas de $(WN+XWN)^2$ (tree#n#1–related–>wood#n#1). No obstante, también hemos evaluado $(MCR)^2$ (utilizando las relaciones a distancia 1 y 2), que sí integra las relaciones de $(WN+XWN)^2$.

2.2. Topic Signatures

Las Topic Signatures (TS) son vectores de palabras relacionadas con un tema (o tópico) [17]. Las Topic Signatures 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. Básicamente, la adquisición de una TS consiste en:

- la adquisición de los mejores ejemplos posibles del corpus de un determinado sentido de la palabra (por lo general, caracterizando cada sentido como una consulta de palabras relacionadas y realizando una búsqueda en el corpus de los ejemplos que mejor se ajusten a las consultas) y a continuación,
- la construcción de las TS a partir de las palabras del contexto que mejor representan el sentido de la palabra en el corpus seleccionado.

Para este estudio hemos usado dos conjuntos de Topic Signatures 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 [1]. Las segundas TS se han obtenido directamente de SemCor.

TSWEB[7]: Inspirado en el trabajo de [16], 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

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

(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.

La figura 2 presenta un ejemplo de TSWEB para el primer sentido de la palabra *party*.

democratic	0.0126	socialist	0.0062
tammany	0.0124	organization	0.0060
alinement	0.0122	conservative	0.0059
federalist	0.0115	populist	0.0053
missionary	0.0103	dixiecrats	0.0051
whig	0.0099	know-nothing	0.0049
greenback	0.0089	constitutional	0.0045
anti-masonic	0.0083	pecking	0.0043
nazi	0.0081	democratic-republican	0.0040
republican	0.0074	republicans	0.0039
alcoholics	0.0073	labor	0.0039
bull	0.0070	salvation	0.0038

Cuadro 2: Topic Signature de parte#n#1 obtenida de la web (24 de las 15.881 palabras totales)

TSSEM: Estas Topic Signatures se han construido utilizando SemCor, un corpus en inglés donde todas sus palabras han sido anotadas semánticamente. Este corpus que 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 3, hay un ejemplo de los primeros sentidos que hemos obtenido para party#n#1.

Aunque hemos provado otras medidas, los mejores resultados se han obtenido utilizando la fórmula TFIDF [1].

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

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.

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
identification#n#1	0.3414
government#n#1	0.3414
designation#n#3	0.3414
authorities#n#1	0.3414

Cuadro 3: Topic Signature para party#n#1 obtenida de SemCor (11 de los 719 sentidos totales)

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

El presente estudio no incluye otras TS adquiridas del corpus BNC usando otras estrategias de construcción y de otras herramientas como ExRetriever o Infomap [8], [9] y [10].

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. En el cuadro 4 se muestra un ejemplo del corpus de prueba para la palabra *party* cuyo sentido correcto es el primero. Para la tarea en inglés se usó para la anotación los sentidos de WN1.7. 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 [22], en lugar de los que aparecen en WordNet.

Por otra parte, tratando de ser lo más neutral posible con respecto de los recursos estudiados, hemos aplicado sistemáticamente el mismo método de resolución de la ambigüedad a todos ellos. Recordemos que nuestro objetivo principal es establecer una comparación relativa de los recursos de conocimiento en vez de proporcionar la mejor técnica de WSD para una determinada base de conocimiento.

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[8]. El synset que tiene el recuento mayor es seleccionado. De hecho, se trata de un método 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 4 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. Como se puede ver, hay varias palabras importantes que aparecen en el texto, que también aparecen en la TS.

<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 4: Ejemplo de prueba número 00008131 para party#n cuyo sentido correcto es el primero

[8]También consideramos los términos multipalabra

Referencias	P	R	F1
TRAIN	65.1	65.1	65.1
TRAIN-MFS	54.5	54.5	54.5
WN-MFS	53.0	53.0	53.0
SEMCOR-MFS	49.0	49.1	49.0
RANDOM	19.1	19.1	19.1

Cuadro 5: Resultados de P, R, y F1 para los sistemas de referencia básicos de la muestra léxica en inglés

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 (SEMCOR-MFS): Este método selecciona el sentido más frecuente de la palabra según SemCor.

WordNet MFS (WN-MFS): Este método 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: Este método selecciona el sentido más frecuente de la palabra objetivo en el corpus de entrenamiento.

Train Topic Signatures (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.

El cuadro 5 presenta la precisión (P), recall (R) y F1 (media armónica del recall y precisión) de las diferentes referencias básicas. En este cuadro, TRAIN se ha calculado con un vector de como máximo 450 palabras. Como era de esperar, RANDOM obtiene el resultado más pobre. También como era de esperar, los sistemas que usan los sentidos más frecuentes de SemCor (SEMCOR-MFS) y WN (WN-MFS) están ambos por debajo del sistema que usa el sentido más frecuente del corpus de entrenamiento (TRAIN-MFS). Sin embargo, todos ellos están muy por debajo de las TS adquiridas utilizando el corpus de entrenamiento (TRAIN).

Recurso	P	R	F1	Tamaño medio
TSSEM	**52.5**	**52.4**	**52.4**	103
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

Cuadro 6: Resultados de P, R y F1 de los recursos evaluados individualmente en inglés

4.2. Evaluación de cada recurso en inglés

El cuadro 6 presenta ordenada por la medida F1, 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 en 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 aquellos recursos derivados que usan relaciones indirectas. Sorprendentemente, 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 interesante es 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 las TS construidas utilizando el corpus de entrenamiento (TRAIN-MFS).

En cuanto a las expansiones y otras combinaciones, el rendimiento de WN

KB	PM	DV	Rank
MCR+TSSEM	52.3	45.4	**52.7**
MCR+(WN+XWN)^2	47.8	37.8	51.5
(WN+XWN)^2+TSSEM	51.0	41.7	50.5
TSSEM+TSWEB	51.0	42.2	49.4
MCR+TSWEB	48.9	37.6	48.6
(WN+XWN)^2+TSWEB	41.5	34.3	45.4

Cuadro 7: Resultados de la combinación de dos recursos usando la medida F1

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 funciona mejor que WN+XWN y (MCR)^2 ligeramente mejor que (MCR)[9]No se han probado extensiones superiores).

4.3. Combinación de Recursos

Con el objetivo de evaluar de forma más detallada la contribución que tiene cada recurso, proporcionamos un análisis de su aportación combinada. Las combinaciones se han evaluado usando tres estrategias básicas diferentes [7].

Voto Directo (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.

Combinación de Probabilidad (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.

Combinación basada en el orden (Rank): Cada recurso semántico proporciona un orden de sentidos de la palabra que se quiere interpretar. Para cada sentido, se agregadan 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.

4.3.1. Combinando dos recursos

El cuadro 7 presenta las medidas de F1 corrrespondientes a la combinación de dos recursos usando los tres métodos de combinación. Las combinaciones están ordenadas según el resultado del metódo de combinación *Rank*. En negrita, destacamos el mejor resultado. En este caso, es el correspondiente a la combinación por orden de los recursos MCR y TSSEM[10].

[9](

[10]Obsérvese que los recursos al no ser totalmente disjuntos pueden presentar duplicaciones. Así, en este caso, algunas relaciones de SemCor puedan aparecer a la vez en spSemCor y

KB	PM	DV	Rank
MCR+TSSEM+(WN+XWN)^2	52.6	37.9	**54.6**
MCR+TSWEB+TSSEM	54.1	37.2	53.3
MCR+TSWEB+(WN+XWN)^2	49.8	33.3	52.1
(WN+XWN)^2+TSSEM+TSWEB	51.5	36.1	51.5

Cuadro 8: resultados de la combinación de tres recursos usando la medida F1

Observando el método de combinación aplicado, los métodos de la Combinación de Probabilidad (PM) y la combinación basada en el orden (Rank), se comportan parecido (cada método predomina en 3 de las 6 combinaciones) y ambos obtienen mejores resultados que el método de Combinación Directa (DV). De ahora en adelante, se va a usar el método de combinación basada en el orden (Rank) para comparar los resultados.

Es interesante observar que solamente en dos casos la combinación de recursos se comportan por debajo de los recursos individuales. Ambos casos incluyen TSSEM (F1 of 52.4) cuando se combina con TSWEB (F1 of 49.4) y (WN+XWN)^2 (F1 of 50.5). De todos modos, en los casos restantes, parece que cada recurso añade algun tipo de conocimiento que los otros recursos no disponen. Por ejemplo, el conocimiento contenido en (WN+XWN)^2 parece no estar representado en el MCR. Más aún, a pesar de que (WN+XWN)^2+TSWEB obtiene los peores resultados cuando se combinan dos recursos (F1 of 45.4), la contribución individual a la combinación és impresionante (5.4 puntos con respeto a (WN+XWN)^2 y 9.4 puntos con respeto a TSWEB). De todas formas, el incremento más grande corresponde a la combinación de MCR+(WN+XWN)^2 (F1 de 51.5, 6.0 puntos por encima de MCR y 13.25 por encima de (WN+XWN)^2), indicando que ambos recursos contienen conocimiento complementario. De hecho, hay algún tipo de conocimiento contenido en el MCR no presente en TSSEM (debido al pequeño incremento de 0.3 puntos con respeto a TSSEM individual).

Observando los sistemas de referencia básicos, ninguna combinación alcanza el sentido más frecuente de WN (WN-MFS con F1 de 53.0). De todas formas, algunos de ellas sobrepasan el sentido más frecuente de SemCor (SEMCOR-MFS con F1 de 49.1). En particular, las combinaciones que incluyen información de SemCor (TSSEM or MCR).

4.3.2. Combinando tres recursos

El cuadro 8 presenta los resultados de la medida F1 cuando se combinan tres recursos semánticos distintos. Las combinaciones están ordenadas por el resultado de la combinación basada en *Rank*. El mejor resultado, en negrita, corresponde a la combinación basada en el orden de MCR que corresponde a WN+XWN+spSemCor, TSSEM y (WN+XWN)^2.

TSSEM, y por tanto, ser tomadas en cuenta dos veces

KB	PM	DV	Rank
MCR+(WN+XWN)^2+TSWEB+TSSEM	53.1	32.7	**55.5**

Cuadro 9: Resultados de la combinación de cuatro recursos usando la medida F1

Observando el método de combinación aplicado, parece ser que el comportamiento del método basado en el orden (Rank) es similar al del método basado en Combinación de Probabilidad (PM) (obteniendo mejores resultados en dos de las cuatro combinaciones, peores resultados en una y el mismo resultado en otra ocasion). Otra vez, ambas estrategias de combinación son mejores que la combinación por voto directo (DV).

Considerando sólo la combinación basada en orden (Rank), en general, la combinación de tres recursos semánticos obtiene ligeramente mejores resultados que combinando sólo dos o tres recursos. En este caso, sólo una combinación de recursos tiene unos resultados peores que los recursos tomados de forma individual. Esta combinación incluye otra vez TSSEM (F1 de 52.4), cuando es combinado con (WN+XWN)^2+TSWEB (F1 de 45.4). De todas formas, para el resto de casos, otra vez parece que la combinación de recursos integra algun tipo de conocimiento que no tienen los recursos de forma individual. En este caso, el mayor incremento corresponde a MCR+TSWEB+(WN+XWN)^2 (F1 de 52.1, 16.1 puntos por encima de TSWEB, 12.1 puntos por encima de (WN+XWN)^2, y 7.6 puntos por encima de MCR).

También en esta ocasión podemos observar que muchos de los recursos presentados contienen conocimiento complementario. Así por ejemplo, algún tipo de conocimiento contenido en MCR+(WN+XWN)^2 no se encuentra en TSSEM, debido al incremento de 2.2 puntos con respeto a TSSEM individualmente. De hecho, esta combinación de recursos obtiene 16.3 puntos por encima de (WN+XWN)^2 y 10.1 puntos por encima de MCR.

Si volvemos a considerar las referencias básicas, todas estas combinaciones superan al sentido más frecuente de SemCor (SEMCOR-MFS con F1 de 49.1), y dos combinaciones de tres recursos sobrepasan al sentido más frecuente de WN (WN-MFS con F1 de 53.0): MCR+TSWEB+TSSEM (F1 de 53.3) y MCR+TSSEM+(WN+XWN)^2 (F1 de 54.6). Ese último en concreto está por encima del sentido más frecuente del conjunto de entrenamiento (TRAIN con F1 of 54.5). Obviamente, debemos resaltar este resultado ya que en las tareas de WSD, los sistemas supervisados sobrepasan muchas veces con dificultad al sistema que escoge el sentido más frecuente del conjunto de datos de entrenamiento, a pesar de tener en cuenta el contexto local[15].

4.3.3. Combinando cuatro recursos

El cuadro 9 presenta los resultados segun la medida F1 con respeto a los tres métodos (DV, PM, Rank) cuando combinamos cuatro recursos semanticos distintos. En negrita se presenta el mejor resultado. En este caso, en la combinación

basada en el orden con MCR, TSSEM, TSWEB y (WN+XWN)^2.

Otra vez, el método basado en en el orden (Rank) tiene un mejor comportamiento que el método basado en el voto directo (DV) o en la combinación de probabilidades (PM).

Considerando sólo la combinación basada en el orden (Rank), como se esperaba, 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 [20].

5. Evaluación en castellano

5.1. Referencias básicas para el 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 (Minidir-MFS): Este método 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 con la frecuencia de los sentidos de palabras del conjunto de entrenamiento. Por eso, para el castellano, el sentido más frecuente de Minidir (Minidir-MFS) es el mismo que el sentido más frecuente del conjunto de entrenamiento (TRAIN-MFS).

Train Topic Signatures (TRAIN): Este método 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 ita-

Referencias básicas	P	R	F1
TRAIN	81.8	68.0	74.3
MiniDir-MFS	67.1	52.7	59.2
RANDOM	21.3	21.3	21.3

Cuadro 10: Resultados de las referencias básicas de la evaluación usando la muestra léxica en castellano, a nivel de Precisión, Recall y F1

Bases de conocimiento	P	R	F1	Tamaño medio
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

Cuadro 11: Resultados de las medidas de P, R y F1 para los recursos evaluados individualmente en castellano.

liano[11].

Además, solamente pueden ser transportadas de un idioma a otro sin introducir demasiados errores las relaciones que existan en un recurso entre sentidos[12]. Como TSWEB relaciona palabras en inglés a un synset, no ha sido transportado ni evaluado al castellano.

El cuadro 10 presenta las medidas de precisión (P), recall (R) y F1 de las diferentes referencias básicas. Para el castellano, el recurso TRAIN ha sido evaluado con un tamaño de vector máximo de 450 palabras. Como se esperaba, RANDOM obtiene el resultado más pobre, y el sentido más frecuente obtenido de Minidir (Minidir-MFS, que es igual a TRAIN-MFS) es bastante más bajo que las TS obtenidas usando el conjunto de entrenamiento (TRAIN).

5.2. Evaluando cada recurso del castellano por separado

El cuadro 11 presenta ordenado por la medida F1, el comportamiento de los distintos recursos semánticos y su tamaño medio. En negrita, aparecen los mejores resultados por Precision (P), Recall (R) y F1. 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 of 22.5). Es interesante notar que en terminos de Precision, 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

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

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

idioma, no puede ser transportado directamente a otro idioma. Otros posibles motivos de los bajos resultados podrian ser el pequeño tamaño de los recursos castellano (comparándolos con los existentes en inglés), 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 RANDOM, pero ninguno de ellos llega a Minidir-MFS (igual que al 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 en español original (F1 de 22.5).

6. Una propuesta para la construcción de bases de conocimiento ultra conectadas y densas

En nuestra opinión, un procesamiento semántico preciso (como el WSD) no se basa sólo en algoritmos más o menos sofisticados, sino en aproximaciones basadas en el uso intensivo de conocimiento semántico. Los resultados presentados en este trabajo, sugieren que es necesaria mucha más investigación en la adquisición y uso de recursos semánticos a gran escala. De hecho, la arquitectura basada en ciclos del proyecto MEANING demostró que la adquisición de mejor conocimiento permitía una mejor resolución de la ambigüedad semántica de las palabras (WSD) y que si se mejoraban los sistemas de WSD, podiamos ser capaces de obtener más y mejor conocimiento [26].

Por lo tanto, nuesto próximo objetivo va a ser la adquisición por medios automáticos de bases de conocimiento ultra conectadas y densas a partir de grandes colecciones de corpus y usando el conocimiento existente y codificado que tenemos en el MCR. De esta forma, queremos incrementar el número de relaciones semánticas actualmente disponibles en el MCR (alrededor de un millón) hasta decenas de millones.

La propuesta actual consiste en:

- Seguir [8] y [10] para la adquisición de Topic Signatures (TS) muy precisas para todas las palabras monosémicas de WN (por ejemplo, usando InfoMap [13]). Es decir, adquirir vectores de palabras muy relacionadas con una palabra monosémica en concreto (por ejemplo, airport#n#1) a partir de corpus como el BNC u otras colecciones de textos como GigaWord o la web.

- Aplicar un algoritmo de WSD basado en conocimiento a cada TS. Es decir, obtener vectores de sentidos en vez de vectores de palabras. Por ejemplo, usando una version del algorimo "Structural Semantic Interconnections"(SSI) [24]).

palabra+pos	peso	num. sentidos
airport#n	1.000000	1
heathrow#n	0.843162	0
gatwick#n	0.768215	0
flight#n	0.765804	9
airfield#n	0.740861	1
train#n	0.739805	6
travelling#n	0.732794	1
passenger#n	0.722912	1
station#n	0.722364	4
ferry#n	0.717653	2

Cuadro 12: Primeras diez palabras de la TS de airport#n#1 obtenida a partir del BNC usando InfoMap.

.

Como ejemplo, podemos considerar las primeras diez palabras con peso y categoria gramatical, que aparecen en la Topic Signature (TS) del nombre monosemico airport#n#1 (ver cuadro 12). Esta TS ha sido obtenida a partir del corpus BNC usando Infomap. De las diez palabras que aparecen en la TS, dos de ellas no aparecen en WN (correspondientes a los nombres propios heathrow#n y gatwick#n), cuatro palabras són monosémicas (airport#n, airfield#n, travelling#n y passenger#n) y las otras cuatro són polisémicas (flight#n, train#n, station#n and ferry#n). En esta TS, no hay verbos, adjetivos ni adverbios.

Hemos implementado una version del algoritmo "Structural Semantic Interconnections"(SSI), un algoritmo iterativo de WSD basada en el conocimiento [24]. El algoritmo SSI, es muy simple y consta de dos etapas. Hay una primera etapa inicialización y una segunda etapa iterativa. Dada W, una lista de palabras ordenadas para ser interpretadas[13], el algoritmo de SSI hace lo siguiente. Durante la etapa de inicialización construye un conjunto llamado I (palabras interpretadas), que incluye todas las palabras monosémicas, y otro conjunto P (palabras pendientes de ser interpretadas) que contiene las palabras polisémicas. En cada paso, el conjunto I es usado para ayudar a resolver la ambigüedad de una palabra de P, seleccionando el sentido de la palabra que és más próximo al conjunto I de palabras ya interpretadas. Cada vez que una palabra es interpretada, la palabra es eliminada del conjunto P y incluida en el conjunto I. El algoritmo termina cuando no hay ninguna palabra pendiente en el conjunto P.

Inicialmente, la lista de las palabras interpretadas I, incluye los sentidos de la palabras monosémicas en W, o un conjunto de sentidos prefijados[14]. De todas formas, en nuestro caso, cuando procesamos una TS derivada de una palabra monosémica m, la lista I incluye desde el principio como mínimo el sentido de

[13] Obtener sus sentidos más plausibles dado el resto de palabras que aparecen en W.

[14] Si no hay palabras monosémicas, el algoritmo puede suponer un conjunto inicial basándose en el sentido mas probable, o el sentido menos ambiguo, etc.

Synsets	Distance
4	6
4530	5
64713	4
29767	3
597	2
20	1
1	0

Cuadro 13: Distancias mínimas desde airport#n#1 al resto del grafo

la palabra monosémica m (en nuestro ejemplo, airport#n#1).

Para medir la proximidad de un synset (relacionado con la palabra que se quiere interpretar en cada paso) a un conjunto de synsets (sentidos de las palabras que ya han sido interpretados en I), SSI usa una base de conocimiento propia derivada de forma semi-automatica de un conjunto de recursos semiestructurados [23]. Esta base de conocimiento se usa para calcular distancias entre synsets en un grafo. Con el interés de evitar la explosión exponencial de posibilidades, no todos los posibles caminos son considerados por el algoritmo. SSI usa una gramatica libre de contexto prefijada de relaciones con peso para filtrar los caminos inapropiados y para proporcionar pesos a los caminos apropiados. Los pesos de las reglas de la gramática han sido precalculados usando SemCor.

Nuestra versión usa parte del conocimiento contenido en el MCR para construir un gran grafo conectado que incluye 99.635 nodos y 636.077 arcos. Por ejemplo, el cuadro 13 refleja las distancias mínimas desde airport#n#1 al resto de synsets del grafo. Así, desde airport#n#1 todos los synsets del grafo són acesibles a distancias menores de 6 arcos. Hay sólo un synset a distancia cero (airport#n#1) y veinte synsets directamente conectados a airport#n#1. De hecho, el 95 % del grafo es accesible a distancia cuatro o menos.

Por ejemplo, el cuadro 13 refleja las distancias mínimas desde airport#n#1 al resto de synsets del grafo. Así, desde airport#n#1 todos los synsets del grafo són acesibles a distancias menores de 6 arcos. Hay sólo un synset a distancia cero (airport#n#1) y veinte synsets directamente conectados a airport#n#1. De hecho, el 95 % del grafo es accesible a distancia cuatro o menos.

Sobre este grafo, usamos una libreria muy eficiente para aplicar el algoritmo de Dijsktra [15]. Dijkstra es un algoritmo voraz para calcular el camino con menor distancia entre un nodo y el resto de ellos en un grafo. De esta forma, podemos calcular de forma muy eficiente, el camino más corto entre dos nodos cualquiera del grafo. Esta version del algoritmo SSI, la llamamos SSI-Dijkstra.

SSI-Dijkstra tiene unas propiedas muy interesantes. Por ejemplo, siempre devuelve la distancia mínima entre dos synsets. Esto es, el algoritmo de Dijkstra siempre proporciona una respuesta, siendo la distancia mínima más o menos

[15] Usamos una version modificada de la libreria BoostGraph.

palabra	offset	peso	Definición
flight#n	00195002n	0.017	a scheduled trip by plane between designated airports
travelling#n	00191846n	0	the act of going from one place to another
train#n	03528724n	0.012	a line of railway cars coupled together and drawn by a locomotive
passenger#n	07460409n	0	a person traveling in a vehicle (a boat or bus or car or plane or train etc) who is not operating it
station#n	03404271n	0.019	a building equipped with special equipment and personnel for a particular purpose
airport#n	02175180n	0	an airfield equipped with control tower and hangers as well as accommodations for passengers and cargo
ferry#n	02671945n	0.010	a boat that transports people or vehicles across a body of water and operates on a regular schedule
airfield#n	02171984n	0	a place where planes take off and land

Cuadro 14: TS interpretada para airport#n#1 obtenida automática del BNC usando InfoMap y SSI-Dijkstra

.

cercana[16]. De hecho, el algoritmo SSI-Dijkstra compara las distancias mínimas entre los synsets de una palabra y todos los synsets ya interpretados en I. En cada paso, SSI-Dijkstra seleciona el synset de la palabra polisémica que se esté tratando que es más próximo a I.

El cuadro 14 presenta los resultados del processo de interpretación aplicando el algoritmo SSI-Dijkstra a la TS presentada en el cuadro 12[17]. Ahora, parte de la TS obtenida a partir del BNC usando InfoMap ha sido interpretada a nivel de synset resultando en una TS de sentidos interpretados. Las palabras no presentes en WN1.6 han sido ignoradas (*heathrow* y *gatwick*). Algunas otras que eran monosémicas en WN1.6 han sido consideradas interpretadas (*travelling, passenger, airport* y *airfield*). El resto ha sido interpretado correctamente (*flight* con nueve sentidos, *train* con seis sentidos, *station* con cuatro y *ferry* con dos).

Esta TS interpretada representa siete nuevas relaciones semánticas entre airport#n#1 y las primeras palabras de la TS. Éstas nuevas relaciones podrian ser integradas en una nueva versión del MCR (por ejemplo, airport#n#1–related–>flight#n#9). Sin embargo, también podrían integrarse las relaciones indirectas de la TS interpretada (por ejemplo, flight#n#9–related–>travelling#n#1). De este modo, teniendo n sentidos de palabras interpretadas de la TS, se podría crear un total de $(n^2 - n)/2$ nuevas. Es decir, para las diez primeras palabras de la TS de airport#n#1, se podrían crear veintiocho nuevas relaciones entre synsets.

Así, podríamos repetir este proceso para todas las palabras monosémicas de WordNet que aparecieran en un corpus concreto. El número total de palabras monosémicas en WN1.6 es de 98.953. Obviamente, no todas las palabras

[16] A diferencia del algoritmo SSI original que no siempre da un camino ya que sólo se explora una parte del grafo.

[17] Este proceso tardó 31 segundos en procesar la TS en un ordenador personal de sobremesa

monosémicas aparecerán en el corpus, aunque esperamos obtener siguiendo este proceso varios millones de nuevas relaciones semánticas entre synsets. Este método va a permitirnos derivar por medios completamente automáticos una base de conocimientos con millones de nuevas relaciones que podrán ser incorporadas a una nueva version del MCR. Más aún, este proceso podría ser repetido varias veces, procesando en cada paso palabras aún no procesadas de las TS, incrementando en cada paso la versión previa del MCR.

Como prueba de la validez de la propuesta, hemos aplicado el método propuesto a un subconjunto de las TS de TSWEB correspondientes a los 20 nombres de la tarea de la muestra léxica del inglés de SensEval-3. Para esas TS, ya sabemos a priori el sentido de una palabra en que la TS deriva. El cuadro 2 nos muesta la TS de TSWEB para el primer sentido del nombre "party". En este caso, la tarea consiste en assignar el sentido correcto a la palabra "party"dándole al algoritmo SSI-Dijkstra las primeras diez palabras del contexto de la TS derivada para el sentido party#n#1. De las 114 TS de TSWEB correspondientes a los nombres seleccionados para la tarea de muestra léxica del inglés de SensEval–3, el algoritmo selecciona correctamente 105. Esto es un 95 % de respuestas correctas[18]. Obviamente, resultados menos interesantes se pueden preveer para el resto de palabras de la misma TS. De todas formas, esos resultados iniciales parecen ser muy prometedores.

Considerado que tenemos una base de conocimiento con 99.635 synsets, el número total de posibles conexiones directas entre systems deberia ser de alrededor de 10.000,000.000 (diez mil millones). Eso puede ser considerado como el límite superior del número potencial de relaciones binarias que pueden existir en una base de conocimiento como WordNet[19].

De esta forma, planeamos adquirir grandes porciones de nuevo conocimiento estableciendo sólo las conexiones apropiadas entre synsets. Además, este sistema és independiente del idioma. Podria ser repetido para cualquier idioma que tuviese palabras conectadas al MCR (por ejemplo, el castellano).

Resta estudiar e investigar más profundamente cómo convertir las relaciones creadas de esta forma en relaciones más específicas e informativas. Una posibilidad sería tratar de adquirirlas de textos o derivarlas automáticamente de las Ontologías integradas en el propio MCR.

7. Conclusiones

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. Obviamente, se pueden imaginar métodos más sofisticados para usar apropiadamente estos recursos [24]. Además, el hecho

[18] Inesperadamente, en TSWEB tres sentidos no tenian una TS associada

[19] Cada relación binaria puede estar establecida potencialmente entre dos synsets diferentes

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

Creemos, que un procesamiento semántico preciso (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.

Parece ser que la combinación de los recursos a gran escala disponibles actualmente codifican el conocimiento necesario para comportarse como un etiquetador del 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*.

Son necesarios más experientos en un marco multilingüe para clarificar el distinto comportamiento del MCR y TSSEM en los dos idiomas, quizás usando el WN del italiano (también integrado en el MCR) y MultiSemCor [6].

Además, tenemos planificado la adquisición a gran escala de Topic Signatures de sentidos muy precisos para todos los idiomas integrados en el MCR. Estas nuevas Topic Signatures podrian ser fácilmente integradas en versiones futuras del MCR.

8. Agradecimientos

Algunas partes de este trabajo han sido también publicadas previamente en [11]. 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).

Referencias

[1] E. Agirre and O. L. de la Calle. Publicly available topic signatures for all wordnet nominal senses. In *Proceedings of LREC*, Lisbon, Portugal, 2004.

[2] E. Agirre and D. Martinez. Learning class-to-class selectional preferences. In *Proceedings of CoNLL*, Toulouse, France, 2001.

[3] E. Agirre and D. Martinez. Integrating selectional preferences in wordnet. In *Proceedings of GWC*, Mysore, India, 2002.

[4] J. Ález, J. Atserias, J. Carrera, S. Climent, A. Oliver, and G. Rigau. Consistent annotation of eurowordnet with the top concept ontology. In *Proceedings of Fourth International WordNet Conference (GWC'08)*, to appear.

[5] J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, and P. Vossen. The meaning multilingual central repository. In *Proceedings of GWC*, Brno, Czech Republic, 2004.

[6] L. Bentivogli, E. Pianta, and M. Ranieri. Multisemcor: an english italian aligned corpus with a shared inventory of senses. In *Proceedings of the the Meaning Workshop 2005*, Trento, Italy, 2005.

[7] S. Brody, R.Ñavigli, and M. Lapata. Ensemble methods for unsupervised wsd. In *Proceedings of COLING-ACL*, pages 97–104, 2006.

[8] M. Cuadros, L. Padró, and G. Rigau. Comparing methods for automatic acquisition of topic signatures. In *Proceedings of RANLP*, Borovets, Bulgaria, 2005.

[9] M. Cuadros, L. Padró, and G. Rigau. An empirical study for automatic acquisition of topic signatures. In *Proceedings of GWC*, pages 51–59, 2006.

[10] M. Cuadros and G. Rigau. Quality assessment of large scale knowledge resources. In *Proceedings of EMNLP*, 2006.

[11] M. Cuadros, G. Rigau, and M. Castillo. Evaluating large-scale knowledge resources across languages. In *Proceedings of Recent Advanced of Natural Language Processing (RANLP'07)*, 2007.

[12] J. Daudé, L. Padró, and G. Rigau. Validation and Tuning of Wordnet Mapping Techniques. In *Proceedings of RANLP*, Borovets, Bulgaria, 2003.

[13] B. Dorow and D. Widdows. Discovering corpus-specific word senses. In *EACL*, Budapest, 2003.

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

[15] V. Hoste, W. Daelemans, I. Hendrickx, and A. van den Bosch. Evaluating the results of a memory-based word-expert approach to unrestricted word sense disambiguation. In *Proceedings of the Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 95–101, 2002.

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

[17] C. Lin and E. Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of COLING*, 2000. Strasbourg, France.

[18] B. Magnini and G. Cavaglià. Integrating subject field codes into wordnet. In *Proceedings of LREC*, Athens. Greece, 2000.

[19] D. McCarthy. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Aternations, Subcategorization Frames and Selectional Preferences*. PhD thesis, University of Sussex, 2001.

[20] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. Finding predominant senses in untagged text. In *Proceedings of ACL*, pages 280–297, 2004.

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

[22] R. Mihalcea, T.Chlovski, and A.Killgariff. The senseval-3 english lexical sample task. In *Proceedings of ACL/SIGLEX Senseval-3*, Barcelona, 2004.

[23] R.Ñavigli. Semi-automatic extension of large-scale linguistic knowledge bases. In *Proc. of 18th FLAIRS International Conference (FLAIRS)*, Clearwater Beach, Florida, 2005.

[24] R.Ñavigli and P. Velardi. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(7):1063–1074, 2005.

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

[26] G. Rigau, B. Magnini, E. Agirre, P. Vossen, and J. Carroll. Meaning: A roadmap to knowledge technologies. In *Proceedings of COLING'2002 Workshop on A Roadmap for Computational Linguistics*, Taipei, Taiwan, 2002.

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