Spanish JavaSimLib: una herramienta para el cálculo de la similitud semántica entre palabras en castellano

Spanish JavaSimLib: a tool to compute the semantic similarity between words in Spanish

Isaac Lozano, Alexandre Trilla, Francesc Alías
GTM – Grup de Recerca en Tecnologies Mèdia
LA SALLE – UNIVERSITAT RAMON LLULL
Quatre Camins 2, 08022 Barcelona (Spain)
st18187@salle.url.edu, atrilla@salle.url.edu, falias@salle.url.edu

Resumen: En este artículo se adapta JavaSimLib al castellano, una herramienta capaz de calcular la similitud semántica entre palabras basándose en el contenido de información de cada uno de los términos. Para ello se adecua el WordNet en castellano al motor de búsqueda Lucene, permitiendo una distribución efectiva de toda la información necesaria. El proceso planteado logra replicar al castellano las prestaciones originales de JavaSimLib en inglés.
Palabras clave: Similitud semántica, WordNet, contenido de información, castellano

Abstract: This article adapts JavaSimLib to Spanish, which is a tool to compute the semantic similarity between words according to their information content. To this end, the Spanish WordNet is rearranged into the Lucene search engine, allowing an effective distribution of all the necessary information. The Spanish adaptation procedure that is shown attains the same performance as the original implementation of JavaSimLib in English.
Keywords: Semantic similarity, WordNet, information content, Spanish

1. Introducción

En la actualidad, el inglés es uno de los idiomas que más domina los ámbitos económicos, científicos y tecnológicos, dejando a otros idiomas, como el castellano, en un segundo plano. En este contexto, las herramientas lingüísticas de análisis y procesamiento de los idiomas se suelen desarrollar primero para el inglés y luego para los otros idiomas.

En la literatura se pueden encontrar distintos trabajos de investigación orientados al desarrollo y adaptación de herramientas lingüísticas en inglés para conseguir prestaciones similares para el castellano. Por ejemplo, se pueden destacar herramientas del campo del Procesamiento del Lenguaje Natural como Freeling (Padró et al., 2010), del campo del Análisis del Sentimiento como EmoTag (Francisco y Hervás, 2007) y EmoLib (García y Alías, 2008; Trilla y Alías, 2009), y del campo de la Traducción Automática como el Apertium (Vié et al., 2011), para nombrar algunos. Estas herramientas están típicamente constituidas por distintos módulos elementales (tokenizador, POS taggers, stemmers, desambiguación semántica, etc.). Este trabajo se centra en adaptar al castellano el trabajo de (Seco, Veale, y Hayes, 2004) que permite determinar la similitud semántica entre palabras en inglés. El resultado de esta adaptación se puede incorporar al módulo de desambiguación semántica de cualquier herramienta de las anteriormente enumeradas.

En (Seco, Veale, y Hayes, 2004) se describe una herramienta, denominada JavaSimLib[1], capaz de valorar la similitud semántica entre pares de palabras con resultados satisfactorios usando WordNet en inglés y el motor de búsqueda Lucene[2]. Para adaptar dicha herramienta al castellano, aquí nombrada Spanish JavaSimLib, es necesario contemplar ciertos detalles específicos de este idioma así como considerar el WordNet en castellano.

La organización de este trabajo se describe a continuación. En la Sección 2 se muestra el proceso seguido en el desarrollo de la herramienta para el castellano, detallando todos los pasos realizados para incorporar WordNet

[1] eden.dei.uc.pt/∼nseco/javasimlib.tar.gz
[2] http://lucene.apache.org/

en castellano al proceso de cálculo de la similitud semántica. En la Sección 3 se describe la evaluación de la herramienta y se valora si su efectividad es equivalente a la herramienta original desarrollada para el inglés. En las Secciones 4 y 5 se discuten los resultados obtenidos y se detallan las conclusiones obtenidas y el trabajo futuro.

2. Adaptación de JavaSimLib al castellano

En este apartado se describe el proceso realizado para disponer de la Spanish JavaSimLib, que permite evaluar la similitud semántica entre dos palabras en castellano. En primer lugar, se describe el formato de los datos originales a partir de los cuales se realiza un proceso de adaptación para utilizar el motor de búsqueda Lucene. A continuación se realiza la adaptación y la ampliación de los datos con el objetivo de encontrar todos los parámetros necesarios en el desarrollo de la herramienta. Después de este proceso, se lleva a cabo la conversión del formato de datos para poder trabajar en un entorno basado en el motor de búsqueda Lucene. Finalmente, se muestra el formato de los datos generados con el objetivo de comprobar su estructura.

2.1. Datos originales de WordNet en castellano

Los datos iniciales proceden de la versión de noviembre de 2006 de WordNet en castellano, creado por el *LSI group* de la Universitat Politècnica de Catalunya (UPC), el *CL group* de la Universitat de Barcelona (UB) y el *NLP group* de la Universidad Nacional de Educación a Distancia (UNED). Este WordNet está enmarcado dentro del proyecto EuroWordNet (Vossen, 2002), el cual auna WordNets de diferentes idiomas tomando como referencia el WordNet original en inglés (Miller, 1995). Concretamente, toma como referencia la versión 1.6 de este WordNet.

Los datos iniciales y su formato están contenidos en un script SQL y, por lo tanto, son fácilmente importables a una base de datos. En este trabajo, se ha elegido MySQL por su facilidad de acceso.

Una vez dispuestos los datos, se procede a analizar su formato con más detenimiento. El WordNet en castellano contiene tres tablas principales: *synset*, *variant* y *relation*. En la tabla *synset* se encuentran todos los *synsets* pertenecientes al WordNet juntamente con la definición del concepto que describen cada uno de ellos. Un *synset* es un grupo de palabras que tienen el mismo significado. En la tabla *variant* aparecen todas las palabras del WordNet, cada una de ellas asociada con el *synset* al que pertenece y la acepción que define. Por último, en la tabla *relation* se muestran diferentes relaciones semánticas entre *synsets*, tales como hiponimia, meronimia, antonimia, etc. En el Cuadro 1 se muestra un ejemplo sobre la estructura y los campos más importantes de estas tablas, considerando el *synset* que contiene la palabra "chico" y sus sinónimos referentes a la acepción que define una persona joven de genero masculino.

En el Cuadro 1(a) se puede observar como están estructurados los *synsets*. El campo *pos* muestra la función de la palabra definida por el *synset*, el valor *offset* es el identificador del concepto dentro del WordNet, el número de hipónimos del *synset* viene determinado por el campo *sons* y por último se muestra la definición del concepto descrito en el campo *gloss*. El Cuadro 1(b) muestra la tabla *variant*, la cual comparte los campos *pos* y *offset*, que informan de los mismos parámetros que en la tabla de *synsets*. Los otros parámetros de interés son el campo *word*, el cual muestra la palabra definida por el *synset* y el campo *sense* el cual muestra la acepción de dicha palabra. Por último, se encuentra la tabla *relation* en el Cuadro 1(c). En esta tabla se muestra la relación semántica entre dos *synsets* según las especificaciones de (Vossen, 2002). Las funciones de palabra y los identificadores de *synset* que satisfacen esta relación se muestran en los campos *sourcePos*, *targetPos*, *sourceSynset* y *targetSynset*.

2.2. Adaptación y ampliación de los datos en función del contenido de información

Una vez analizado el formato inicial de los datos, es necesaria una ampliación y una adaptación de estos para poder desarrollar la herramienta planteada en el artículo. Para ello, es necesario dotar a los datos de los campos básicos para poder calcular la similitud entre dos conceptos.

El primer dato necesario que se requiere de cada uno de los *synsets* que componen el WordNet es su contenido de información. Este valor da una idea de la concreción de cada concepto. Originalmente, en (Resnik, 1995) se propone la siguiente expresión para

synset			
pos	**offset**	**sons**	**gloss**
n	07389783	12	Persona joven de sexo masculino

(a) Tabla synset

variant			
pos	**offset**	**word**	**sense**
n	07389783	garzón	1
n	07389783	muchacho	1
n	07389783	mozo	1
n	07389783	chaval	2
n	07389783	chico	1
n	07389783	niño	2

(b) Tabla variant

relation				
relation	**sourcePos**	**sourceSynset**	**targetPos**	**targetSynset**
near-antonym	n	07389783	n	07260273
has-hyponym	n	07389783	n	07071609

(c) Tabla relation

Cuadro 1: Estructura original del WordNet en castellano para la palabra "chico" usada como ejemplo.

calcularlo:

$$ic_{res} = -logp(c) \quad (1)$$

En la expresión 1 se valora la probabilidad que tiene un concepto 'c' del WordNet en un texto. Si una palabra tiene pocas probabilidades de aparecer, probablemente se deba a que describa un concepto muy concreto. Por el contrario, las palabras que aparecen frecuentemente suelen ser conceptos muy globales y muy poco específicos de una temática concreta. En el caso propuesto en este artículo se trabaja con un WordNet. Por lo tanto, es interesante valorar el contenido de información de una palabra en base a un criterio distinto al cálculo de probabilidades. En (Seco, Veale, y Hayes, 2004) se propone la siguiente expresión para calcular el contenido de información en un entorno basado en WordNets:

$$ic_{wn}(c) = 1 - \frac{log(hypo(c) + 1)}{log(max_{wn})} \quad (2)$$

En la expresión 2, $hypo(c)$ es el número de hipónimos de un *synset* concreto del WordNet y max_{wn} es el número total de conceptos que hay en el WordNet. Esta expresión indica la dependencia del contenido de información con el número de hipónimos de un *synset* concreto. Si un *synset* tiene un gran número de hipónimos, indica que se está ante un concepto muy global y por ello su contenido de información es bajo. En el caso contrario, un *synset* con pocos hipónimos señala un concepto muy concreto, y por ello su contenido de información es alto. En el caso que nos ocupa, al trabajar con palabras con distintas funciones (nombres, verbos y adjetivos), se calcula el contenido de información de cada *synset* en función del número total de *synsets* que existen para cada una de estas funciones. Consecuentemente, el valor de max_{wn} varia de la forma siguiente:

- Nombres: $max_{wn} = 71410$
- Verbos: $max_{wn} = 12342$
- Adjetivos: $max_{wn} = 18189$

De esta forma, se realizarán tres estructuras diferentes según la función de palabra ya que las jerarquías en el WordNet son independientes. El cálculo del contenido de información de cada *synset* se realiza con la ayuda del campo *sons* de la tabla *synset* existente en los datos originales. Como se ha comentado anteriormente, este campo especifica el número de hipónimos de cada *synset*. Por lo tanto se puede aplicar la expresión 2, utilizando $hypo(c) = sons$ para cada uno de los *synsets*.

Para acabar de calcular la similitud entre dos conceptos, es necesario conocer el *Most Specific Common Abstraction* (MSCA)

de los conceptos a valorar. El MSCA indica la información compartida entre dos conceptos cualquiera. Por norma general, es interesante calcular el concepto perteneciente al MSCA con el contenido de información más alto. Originalmente, este valor se define según la similitud de Resnik (1995):

$$sim_{res}(c_1, c_2) = max_{c \in S(c_1,c_2)} ic_{res}(c) \quad (3)$$

En la expresión 3, $S(c_1, c_2)$ representa el conjunto de información que comparten los conceptos c_1 y c_2. La expresión 3 propone el cálculo del contenido de información más alto de todos los conceptos compartidos por c_1 y c_2. A partir de esta base, en la literatura se muestran formas más complejas de calcular la similitud entre dos conceptos. Una de ellas es la siguiente expresión propuesta en (Jiang y Conrath, 1998), donde se calcula la similitud entre dos conceptos a partir de su información y de la similitud de Resnik, y por lo tanto, a partir del MSCA entre los dos conceptos:

$$\begin{aligned} dist_{jcn}(c_1, c_2) &= (ic_{res}(c_1) + ic_{res}(c_2)) - \\ &\quad -2 \cdot sim_{res}(c_1, c_2) \end{aligned} \quad (4)$$

En el entorno en el que se trabaja para el desarrollo de la herramienta planteada, se utiliza la siguiente adaptación de la fórmula de Jiang y Conrath utilizada en (Seco, Veale, y Hayes, 2004), donde se calcula la similitud de Resnik a partir del MSCA en un WordNet:

$$\begin{aligned} sim_{jcn}(c_1, c_2) &= 1 - \frac{(ic_{res}(c_1) + ic_{res}(c_2))}{2} - \\ &\quad - \frac{2 \cdot sim_{res'}(c_1, c_2)}{2} \end{aligned} \quad (5)$$

En un WordNet, el *synset* perteneciente al MSCA de dos conceptos es el *synset* hiperónimo común entre ellos que posee el mayor contenido de información. El *synset* que cumple esta condición siempre es el *synset* intersección entre los dos conceptos de los cuales se estudia la similitud. Para poder calcular el MSCA de dos conceptos, y por lo tanto su similitud, se ha implementado un algoritmo que retorna todos los identificadores de *synsets* hiperónimos de cada concepto existente en el WordNet con la ayuda de la tabla *relation*, véase Cuadro 1(c). Es decir, se calcula el hiperónimo del concepto original, el hiperónimo del hiperónimo, hasta llegar al *synset* con menor contenido de información. Comparando las cadenas de hiperónimos de dos conceptos, se puede encontrar el *synset* perteneciente al MSCA de los dos conceptos. Para ello ha sido necesario corregir dos incongruencias de la tabla *relation*:

- El *synset* 1702479 aparece como hipónimo de él mismo.
- El *synset* 1022027 aparece como hipónimo del *synset* 1021384 al mismo tiempo que el *synset* 1021384 aparece como hipónimo del *synset* 1022027.

Si no se eliminan estas relaciones se generan bucles infinitos en la ejecución del algoritmo necesario para obtener las diferentes cadenas de hiperónimos.

2.3. Conversión de datos SQL-Lucene

La herramienta desarrollada se apoya en el motor de búsqueda Lucene. En este entorno, los datos están estructurados en documentos.

Para realizar el proceso de conversión se ha utilizado el conversor LuSQL[3] (Newton, 2008). Esta herramienta permite transformar un conjunto de datos en un formato de índices Lucene, los cuales son aptos para el desarrollo de la herramienta adaptada en este artículo. La utilidad del conversor LuSQL está en su capacidad de extraer información de una base de datos con comandos propios de un entorno SQL. De esta forma se pueden organizar los campos de los índices Lucene con sentencias idénticas a las utilizadas en la selección de información en una base de datos MySQL. Los índices Lucene resultantes deben contener los siguientes campos en la herramienta Spanish JavaSimLib:

- hypernym: Árbol de hiperónimos de un *synset* concreto. Necesario para poder calcular el MSCA entre dos *synsets*, y por consiguiente, la similitud entre dos conceptos.
- ic: Contenido de información de un *synset* concreto. Interviene de forma activa en el cálculo de la similitud entre dos conceptos.
- synset: Identificador de *synset*.

[3] http://lab.cisti-icist.nrc-cnrc.gc.ca/cistilabswiki/index.php/LuSql

- word: Cadena de caracteres formada por todas las palabras pertenecientes a un *synset* concreto, juntamente con el identificador de acepción. Aunque en la base de datos original no se encuentra esta estructura, la capacidad del conversor LuSQL permite el uso de sentencias de concatenación para generar este campo.

A pesar de que algunos campos están formados por cadenas de caracteres, Lucene puede extraer información individualizada de estas estructuras debido a la opción de organizar los datos en *tokens*. De esta forma, se puede extraer información individual de los campos *hypernym* y *word*. Con la generación de los índices Lucene, ya se tienen los datos en un formato adecuado para la adaptación de la herramienta capaz de valorar la similitud semántica entre palabras en castellano.

2.4. Visualización de los índices

Para revisar el formato de los índices desarrollados, se puede optar por visualizarlos con el visor de índices Lucene Luke[4]. Luke es una herramienta que permite observar datos basados en el motor de búsqueda Lucene. También permite la edición de los índices en el caso que sea necesaria alguna modificación. Pero aunque se disponga de esta funcionalidad, en el caso de trabajar con un gran volumen de datos como es el caso que se describe en este artículo, es preferible realizar la elaboración de los índices con una herramienta como el conversor LuSQL ya que permite obtener un grado de automatización que el visor Luke no es capaz de conseguir. En la Figura 1 se pueden ver los diferentes campos que componen el *synset* correspondiente a la palabra "chico" y los sinónimos correspondientes a la acepción referida a una persona joven de género masculino.

3. Evaluación

Con el fin de evaluar la herramienta desarrollada utilizando el mismo procedimiento descrito en (Seco, Veale, y Hayes, 2004), resulta necesario traducir los 30 pares de palabras en inglés seleccionados en (Miller y Charles, 1991). En este estudio, la similitud de estos pares de palabras fue evaluada por 38 estudiantes universitarios, puntuando cada par con una nota comprendida entre 0 (mínima similitud) y 4 (máxima similitud o sinonimia total). En (Seco, Veale, y Hayes, 2004) se muestra el coeficiente de correlación de Pearson de estas puntuaciones con la similitud calculada usando la ecuación 5, obteniendo una correlación de 0,84. Para evaluar las prestaciones de la adaptación de JavaSimLib al castellano, se calcula la correlación de los pares de palabras traducidos al castellano usando la similitud semántica calculada a partir de los índices de similitud que se obtienen mediante la herramienta Spanish JavaSimLib.

En la traducción y el cálculo de similitud semántica ha sido necesario descartar 6 pares de palabras del experimento original por distintas circunstancias. En los pares de palabras "caldera-estufa" y "gema-joya" se han encontrado inconsistencias en el MSCA y en la estructura del WordNet en castellano. En el caso inglés de "caldera-estufa" (*furnace-stove*), también aparecen inconsistencias, tal y como se comenta en (Jiang y Conrath, 1998). En la Figura 2 se pueden ver las diferencias entre el WordNet inglés y el castellano para los pares de palabras "*gem-jewel*" y "gema-joya". En ella, sólo aparecen las acepciones más características de cada *synset* con el objetivo de facilitar la comprensión de las estructuras. En el caso de los pares de palabras en inglés "*midday-noon*" y "*noon-string*" no se ha encontrado una traducción razonable para el término *noon*, ya que este describe una hora concreta del día (las 12 del mediodía). En este caso el problema detectado es debido a la mayor riqueza léxica del inglés, el cual contiene términos que describen conceptos de los cuales no existen traducciones directas (una sola palabra) al castellano. Por último, también se han descartado los pares de palabras en inglés "*bird-crane*" y "*crane-implement*" ya que el término *crane* puede tener dos traducciones distintas (grulla y grúa), las cuales modifican el escenario del experimento considerablemente. Además, la palabra "grulla" no aparece en el WordNet en castellano, haciendo inviable el uso del par de palabras antes mencionado en la evaluación de la herramienta.

Con los 24 pares de palabras seleccionados, se obtiene un valor de correlación de Pearson de 0,774. Este resultado permite validar el funcionamiento de Spanish JavaSimLib, si se compara con las valoraciones individuales entre los diferentes pares de palabras vistos en (Seco, Veale, y

[4]http://code.google.com/p/luke/

Field	ITSVopfOLBC	Norm	Value
<hypernym>	ITS--------	0.4375	7389783 6951621 4123 3731 1740
<ic>	--S----O---	---	0.770499
<synset>	I-S--------	1.0	7389783
<word>	ITS--------	0.375	garzón.1 muchacho.1 mozo.1 chaval.2 chico.1 niño.2

Figura 1: Documento y correspondientes campos del synset correspondiente a la palabra "chico".

Hayes, 2004) para el inglés. En el Cuadro 2 se muestra en detalle los pares de palabras utilizados y descartados en el proceso de evaluación.

4. Discusión de los resultados

Como se ha podido ver anteriormente, en el experimento realizado para comprobar el funcionamiento de Spanish JavaSimLib se ha obtenido un valor de correlación de Pearson de 0,774, mientras que para el inglés este valor era de 0,84 (Seco, Veale, y Hayes, 2004). Se ha obtenido un resultado inferior al inglés debido a diferentes circunstancias. El primer aspecto reseñable es el hecho de que no se ha podido realizar exactamente el mismo experimento debido a que ha sido necesario descartar algunos pares de palabras por los problemas de transferencia de idioma, impidiendo traducir ciertas palabras. También se han localizado algunas limitaciones del WordNet en castellano, como la omisión de algún término así como la existencia de algunas inconsistencias. Es lógico que aparezcan este tipo de dificultades al tratarse del único WordNet en castellano estable disponible hasta el momento, el cual además está basado en una versión desactualizada del WordNet en inglés. Estas circunstancias provocan que el proceso de evaluación se haya realizado con menos datos, lo que conlleva diferencias con el experimento original realizado en inglés. Sin embargo, se puede afirmar que el resultado obtenido en los pares de palabras traducidos demuestran un buen funcionamiento de la herramienta en castellano.

5. Conclusiones

En este artículo se ha mostrado el proceso y las herramientas necesarias para poder adaptar al castellano la herramienta JavaSimLib (Seco, Veale, y Hayes, 2004), desde ahora nombrada Spanish JavaSimLib, para calcular la similitud semántica entre pares de palabras en castellano. Para ello, ha sido necesario realizar distintos procesos de revisión manual y de tratamiento de los

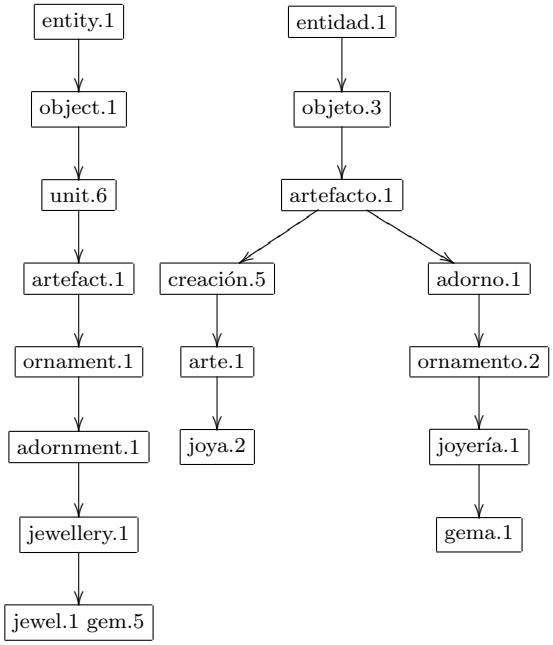

Figura 2: Estructura del WordNet para los pares de palabras "gem-jewel" (izquierda) y "gema-joya" (derecha).

datos para poder replicar esta herramienta en castellano. La evaluación ha demostrado que la adaptación ha logrado unas prestaciones equiparables a la versión original en inglés. Asimismo, el proceso descrito en este trabajo permite la integración de futuras versiones de WordNet en castellano (o en otros idiomas).

En un futuro, se pretende completar este trabajo mediante la evaluación de la similitud semántica entre palabras según nativos españoles para que los resultados no se vean afectados por contextos sociolingüísticos distintos.

Bibliografía

Francisco, V. y R. Hervás. 2007. EmoTag: Automated Mark Up of Affective Information in Texts. *EUROLAN 2007 Summer School Doctoral Consortium*, páginas 5–12.

García, D. y F. Alías. 2008. Emotion identification from text using semantic disambiguation. En *Procesamiento del Lengua-*

je Natural, numero 40, páginas 75–82 *(in Spanish)*, Mar.

Jiang, J. y D. Conrath. 1998. Semantic similarity based on corpus statistics and lexical taxonomy. En *Proceedings of the International Conference on Research in Computational Linguistics*.

Miller, G. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Miller, G. y W.G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6:1–28.

Newton, G. 2008. Lusql v0.9 user manual. Informe técnico, Canada Institute for Scientific and Technical Information (CISTI). National Research Council Canada.

Padró, L., M. Collado, S. Reese, M. Lloberes, y I. Castellón. 2010. FreeLing 2.1: Five Years of Open-Source Language Processing Tools. En *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*, La Valetta, Malta.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. En *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, páginas 448–453.

Seco, N., T. Veale, y J. Hayes. 2004. An Intrinsic Information Content Metric for Semantic Similarity in WordNet. En *Proceedings of ECAI-04*, páginas 1089–1090.

Trilla, A. y F. Alías. 2009. Sentiment classification in English from sentence-level annotations of emotions regarding models of affect. En *Proceedings of InterSpeech2009*, páginas 516–519, Brighton (UK).

Vié, A., L. Villarejo, M. Farrús, y J. O'Regan. 2011. Apertium advanced web interface: a first step towards interactivity and language tools convergence. En F. Sánchez-Martínez y J.A. Pérez-Ortiz, editores, *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, páginas 45–51, Barcelona, Spain.

Vossen, P. 2002. EuroWordNet General Document. University of Amsterdam.

Pares en inglés		Pares en castellano		EH	SS
car	automobile	coche	automóvil	3,92	1
journey	voyage	trayecto	viaje	3,84	0,88
boy	lad	chico	chaval	3,76	1
coast	shore	costa	orilla	3,70	0,96
asylum	madhouse	asilo	manicomio	3,61	0,42
magician	wizard	mago	hechicero	3,50	1
food	fruit	comida	fruta	3,08	0,76
bird	cock	pájaro	gallo	3,05	0,72
tool	implement	herramienta	instrumento	2,95	1
brother	monk	hermano	monje	2,82	0,38
lad	brother	chaval	hermano	1,66	0,44
journey	car	trayecto	coche	1,16	0
monk	oracle	monje	oráculo	1,10	0,13
cemetery	woodland	cementerio	arboleda	0,95	0,55
food	rooster	comida	gallo	0,89	0,44
coast	hill	costa	colina	0,87	0,68
forest	graveyard	bisque	cementerio	0,84	0,25
shore	woodland	orilla	arboleda	0,63	0,28
monk	slave	monje	esclavo	0,55	0,36
coast	forest	costa	bosque	0,42	0,25
lad	wizard	chaval	hechicero	0,42	0,47
chord	smile	acorde	sonrisa	0,13	0,37
glass	magician	cristal	mayo	0,11	0,37
rooster	voyage	gallo	viaje	0,08	0
CORRELACIÓN				1	0,77

(a) Pares de palabras seleccionados

Pares en inglés		Pares en castellano		EH	SS	Motivo del descarte
gem	jewel	gema	joya	3,84	0,21	Inconsistencia del WordNet
midday	noon	mediodía	–	3,42	–	Traducción inexistente para *noon*
furnace	stove	caldera	estufa	3,11	0,27	Inconsistencia del WordNet
bird	crane	pájaro	grulla/grúa	2,97	–	Doble traducción de la palabra *crane*
crane	implement	grulla/grúa	instrumento	1,68	–	Doble traducción de la palabra *crane*
noon	string	–	cuerda	0,08	–	Traducción inexistente para *noon*

(b) Pares de palabras descartados

Cuadro 2: Palabras consideradas en la experimentación. Correlación obtenida entre la similitud semántica de referencia (Seco, Veale, y Hayes, 2004) y la obtenida con Spanish JavaSimLib. EH muestra la evaluación supervisada mostrada en (Miller y Charles, 1991) y SS muestra la similitud semántica más alta posible de cada par de palabras utilizando la expresión 5.