Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 295-302 recibido 1-05-2009; aceptado 5-06-2009

Evaluación de métodos semi-automáticos para la conexión entre FrameNet y SenSem *

Evaluation of semiautomatic methods for the connection between FrameNet and SenSem

Laura Alonso
FaMAF-UNC
Ciudad Universitaria
Córdoba, Argentina
alemany@famaf.unc.edu.ar

Irene Castellón
Facultat de Filologia
UB
Gran Via, 585
Barcelona
icastellon@ub.edu

Egoitz Laparra
Departamento de LSI
UPV-EHU
Manuel de Lardizábal, 1
San Sebastián
ego.laparra@gmail.com

German Rigau
Departamento de LSI
UPV-EHU
Manuel de Lardizábal, 1
San Sebastián
german.rigau@ehu.es

Resumen: En este artículo se presenta una aproximación a la conexión automática de modelos predicativos en castellano e inglés. El objetivo es establecer la dificultad de la tarea y medir el rendimiento de diferentes técnicas y métodos semi-automáticos. Por un lado perseguimos la reducción del esfuerzo para el enriquecimiento de dichos recursos y por otro al aumento de su cobertura y de su consistencia. Para ello combinamos la anotación manual con dos métodos automáticos para establecer correspondencias entre las distintas unidades semánticas.
Palabras clave: Semántica Léxica, Desambiguación semántica, Lexicón Verbal

Abstract: This paper presents an approach for the automatic connection between predicate model resources in Spanish and English. The objective is to assess the difficulty of the task and to evaluate the performance of different techniques and semi-automated methods. On the one hand we are pursuing a reduction in the effort for the enrichment of these resources, on the other, to increase their coverage and consistency. Thus, we combine manual annotation and two automatic methods in order to establish correspondences between the various semantic units.
Keywords: Lexical Semantics, Word Sense Disambiguation, Verbal Lexicon

1. Introducción

La construcción de recursos léxicos a nivel semántico de amplia cobertura resulta muy costosa, debido a la complejidad de la información que contienen y a su necesario volumen. Además, en muchos casos, los desarrolladores siguen criterios y objetivos distintos, cuando no inconsistentes.

Los modelos de predicados tales como FrameNet[1] (Baker, Fillmore, y Lowe, 1997), VerbNet (Kipper, 2005), PropBank (Palmer, Gildea, y Kingsbury, 2005) para el inglés o SenSem[2] (Alonso et al., 2007) para el castellano son recursos básicos en la mayoría de tareas avanzadas de PLN, como Sistemas de Pregunta Respuesta, Recuperación de la Información o Reconocimiento de Implicación Textual. La mayoría de los sistemas que disponen de cierta capacidad de comprensión del lenguaje natural requieren de conocimiento semántico a nivel de predicado. Este tipo de conocimiento permite identificar a los participantes de un evento en particular, independientemente de su realización en el texto. Así, usando estos modelos, pueden armonizarse en una representación semántica común diferentes fenómenos lingüísticos que expresan el mismo evento, como las transformaciones de activa/pasiva, alternancias verbales o nominalizaciones. En los últimos años se han desarrollado varios sistemas de Análisis Semántico Superficial y Etiquetado de Roles Semánticos utilizando estos recursos (Gildea y Jurafsky, 2002; Carreras y Màrquez, 2004; Carreras y Màrquez, 2005; Pradhan et al., 2007; Ruppenhofer et al., 2010).

* Esta investigación se ha llevado a cabo gracias a los proyectos TIN2006-1549-C03-02 del Ministerio de Educación y Ciencia y FFI2008-02579-E/FILO del Ministerio de Ciencia e Innovación

[1] http://framenet.icsi.berkeley.edu/

[2] http://grial.uab.es/recursos.php

Sin embargo, la construcción de modelos lo suficientemente grandes y suficientemmente ricos para el procesamiento semántico de amplia cobertura requiere del esfuerzo de grandes grupos de investigación durante largos periodos de desarrollo. Es por esto que la cobertura de los recursos de los que se disponen actualmente sigue siendo limitada. Además, si se desea disponer de estos recursos en diferentes idiomas se debe invertir el mismo esfuerzo para cada uno de ellos (Subirats y Petruck, 2003).

La transferencia entre recursos semánticos presenta diversos problemas, como por ejemplo distintas granularidades en la segmentación del significado, de forma que muchas veces resulta difícil encontrar correspondencias entre los inventarios de sentidos de los diferentes recursos. No obstante, ésta es una vía de gran potencial para enriquecer y mejorar los recursos semánticos de una lengua con los recursos de otra.

Siguiendo la línea de trabajos anteriores (ver Sección 2), presentamos una primera aproximación a la integración parcial de FrameNet y SenSem. El objetivo es evaluar la dificultad de la tarea y el rendimiento de diferentes técnicas y métodos semi-automáticos que pueden contribuir a la reducción del esfuerzo y al aumento de su consistencia.

En la primera sección presentamos una descripción de los recursos semánticos utilizados, junto con algunas menciones a trabajos previos en la integración de los mismos. En segundo lugar describimos la metodología general seguida en esta primera aproximación, basada en la combinación de anotación manual con dos métodos automáticos: desambiguación de sentidos y modelos de clasificación supervisada para establecer correspondencias entre sentidos y frames semánticos. En la Sección 4 explicaremos los dos métodos automáticos aplicados y presentamos sus resultados. Detallamos los aspectos relevantes de la anotación manual en la Sección 5. Finalmente presentamos las conclusiones a las que hemos llegado en esta investigación y las líneas de trabajo futuro.

2. Recursos semánticos

Los recursos semánticos que se quieren interconectar son de dos lenguas diferentes, SenSem para el castellano y FrameNet para el inglés. Para ello utilizamos un tercer recurso, el Repositorio Central Multilingüe (del inglés Multilingual Central Repository o MCR) (Atserias et al., 2004), que entre otros muchos recursos semánticos incorpora wordnets para estas dos lenguas.

SenSem (Alonso et al., 2007) es un banco de datos verbal compuesto por una base de datos léxica verbal[3] y un corpus[4] de 700.000 palabras, correspondientes a 25.000 frases y sus contextos. La base de datos se ha construido de forma inductiva a partir del corpus anotado a nivel sintáctico -categoría y función sintáctica- y a nivel semántico -sentido verbal, clase eventiva verbal, papeles semánticos y semántica oracional. La base de datos SenSem tiene como unidad principal el sentido verbal con información sintáctico-semántica. Concretamente, el sentido verbal incluye la definición del sentido, los esquemas de subcategorización, las estructuras oracionales en las que participa (agentiva, antiagentiva, pasiva, etc), su asociación a un synset de WordNet, la clase eventiva léxica, los papeles semánticos, en algunos casos sinónimos, ejemplos y la frecuencia de aparición del sentido en el corpus. La base de datos contempla un total de 998 sentidos.

FrameNet (Baker, Fillmore, y Lowe, 1997) es un recurso semántico que contiene descripciones y anotaciones de corpus de palabras del inglés siguiendo el paradigma de *Frame Semantics* (Fillmore, 1976). En este paradigma, un marco o *frame* corresponde a un escenario que implica la interacción de un conjunto de participantes típicos que desempeñan un papel especial en dicho escenario. FrameNet agrupa palabras (*lexical units*, en lo sucesivo LUs) en *frames* semánticos coherentes, y cada *frame* se caracteriza por su lista de participantes (*frame elements*, en lo sucesivo FEs). En FrameNet, los diferentes sentidos de una misma palabra están asignados a *frames* diferentes. Actualmente, FrameNet contiene más de 10.000 LUs agrupadas en 825 *frames*. Más de 6.100 de estos LUs proporcionan, además, ejemplos de corpus anotados con información lingüística. Las LUs de un *frame* pueden ser nombres, verbos, adjetivos y adverbios que representan un conjunto coherente de sentidos estrechamente relacionados que puede considerarse como un pequeño campo semántico. Por ejemplo, el *frame Education_Teaching* contiene LUs que representan la actividad docente y sus par-

[3] http://grial.uab.es/adquisicion/

[4] http://grial.uab.es/search/

ticipantes. Este *frame* es evocado por LUs como *student.n*, *teacher.n*, *learn.v*, *instruct.v*, *study.v*, etc. Este *frame* también define los participantes semánticos de esa actividad (o FEs) como *student*, *subject* o *professor*, que son participantes semánticos del *frame* y sus correspondientes LUs.

Una de las diferencias esenciales entre estos dos recursos, además de la lengua, es que FrameNet está basado en clases (frames) mientras que los sentidos de SenSem no estan asociados a clases verbales. SenSem se ha construido unidad por unidad sin tener en cuenta ninguna generalización semántica a la que pudiera pertenecer un sentido. Además, FrameNet incorpora un conocimiento mucho más específico, SenSem utiliza papeles semánticos que generalizan más la función de los elementos que acompañan al predicado. Por otro lado, la cobertura de ambos recursos es muy diferente, SenSem contiene 998 sentidos correspondientes a los 250 verbos más frecuentes del español (Davies, 2006) y FrameNet, para el inglés, tiene una cobertura de unas 10.000 unidades léxicas.

El recurso que utilizamos para conectar ambos recursos es **WordNet** [5] (Fellbaum, 1998). WordNet es de lejos la base de conocimiento más utilizada en tareas de Procesamiento de Lenguaje Natural (PLN). Contiene información codificada manualmente sobre nombres, verbos, adjetivos y adverbios del inglés y está organizada en base a la noción de *synset*. Un synset es un conjunto de palabras con la misma categoría morfosintáctica que se pueden intercambiar en un determinado contexto. Un synset suele describirse más detalladamente mediante una glosa y mediante las relaciones semánticas explícitas con otros synsets.

Por los beneficios que se obtienen de integrar varios recursos léxicos verbales, se ha trabajado en la integración entre WordNet y FrameNet (Burchardt, Erk, y Frank, 2005; Johansson y Nugues, 2007; Pennacchiotti et al., 2008; Tonelli y Pianta, 2009; Tonelli y Pighin, 2009; Laparra y Rigau, 2009), WordNet, FrameNet y VerbNet (Shi y Mihalcea, 2005), FrameNet, VerbNet y PropBank (Giuglea y Moschitti, 2006) o WordNet, VerbNet y PropBank (Pazienza, Pennacchiotti, y Zanzotto, 2006). En algunos casos se menciona la idea de integrar recursos de diferentes lenguas, gracias al ILI de WordNet (Johansson y Nugues, 2007).

[5] `http://wordnet.princeton.edu/`

3. Metodología

El objetivo final de nuestro experimento es interconectar FrameNet y SenSem para transferir información semántica del primero al segundo. Para ello, aplicaremos métodos y técnicas automáticos para reducir el esfuerzo manual de conexión. Dado que la tarea de conectar estos recursos es compleja, es probable que los métodos automáticos no proporcionen buenos resultados. Para evaluar la dificultad de la tarea, se llevó a cabo la anotación manual de parte de las correspondencias entre sentidos de SenSem y *frames* de FrameNet, tal como se explica en la Sección 5. Se obtuvo el grado de acuerdo entre jueces y se estudiaron los problemas a resolver.

Por otro lado, experimentamos con dos métodos automáticos para ayudar a la conexión: desambiguación automática de sentidos y aprendizaje automático de clasificadores.

La investigación se piensa desarrollar en los siguientes pasos:

1. **Conexión de FrameNet con WordNet**, mediante un algoritmo automático de interpretación semántica de las palabras (en ingles Word Sense Disambiguation) basado en el conocimiento. Este algoritmo toma las LUs asociadas a cada *frame* de FrameNet y las asocia a un synset de WordNet (ver Sección 4.1).
2. **Conexión de SenSem con FrameNet** a través de los synsets asociados a cada sentido de SenSem y los synsets asociados a las LUs de FrameNet por el procedimiento automático del punto anterior.
3. **Validación manual** de una parte (45 %) de las correspondencias entre *frames* de FrameNet y sentidos de SenSem (ver Sección 5).
4. **Aprendizaje de clasificadores** a partir de ejemplos positivos y negativos de correspondencias *frame* – sentido de SenSem generadas en el punto anterior (ver Sección 4.2).
5. **Uso de clasificadores para prevalidar correspondencias** que todavía no habían sido validadas manualmente (55 %).
6. **Validación manual** de las correspondencias pre-validadas por los clasificado-

res en el punto anterior, para evaluar el rendimiento de los clasificadores.

En total, mediante este procedimiento a través de WordNet se han encontrado 329 pares candidatos a correspondencia entre frame y sentido[6], cubriendo un 42 % de los *frames* de FrameNet y un 44 % de los sentidos de SenSem asociados a un synset[7]. De estas 329, 181 (un 55 %) fueron validadas como correspondencias efectivas.

4. Métodos automáticos de soporte

4.1. Desambiguación automática de sentidos: algoritmo SSI

Partiendo del trabajo de (Cuadros y Rigau, 2008), hemos implementado una versión del algoritmo Structural Semantic Interconnections (SSI) (Navigli y Velardi, 2005). SSI es una aproximación iterativa incremental de desambiguación semántica basada en el conocimiento. El algoritmo SSI es muy simple y consiste en una fase de inicialización y una serie de pasos iterativos.

Dado W, una lista ordenada de palabras, el algoritmo procede de la siguiente forma. Durante la fase de inicialización, todas las palabras monosémicas se incluyen en el conjunto I de palabras interpretadas, y las palabras polisémicas se introducen en el conjunto P (todas ellas pendientes de ser interpretadas). En cada paso, se interpreta una palabra del conjunto P utilizando las palabras del conjunto I, seleccionando el sentido de la palabra que sea más cercano a los sentidos de las palabras ya interpretadas del conjunto I. Una vez que el sentido de una palabra es seleccionado, la palabra se elimina del conjunto P y se incluye en I. El algoritmo termina cuando no hay más palabras pendientes en el conjunto P.

Para medir la distancia de un synset (de la palabra que se interpreta en cada paso) a un conjunto de synsets (los sentidos de las palabras ya interpretados en I), el algoritmo SSI original utiliza una base de conocimiento propia derivada semi-automáticamente que integra varios recursos en línea (Navigli, 2005). Con el fin de evitar una explosión exponencial de posibles caminos entre synsets, no se consideran todos estos caminos. El algoritmo SSI original utiliza una gramática de relaciones entrenada sobre SemCor para filtrar caminos inapropiados y procurar un peso a los caminos apropiados.

SSI-Dijkstra (Cuadros y Rigau, 2008) calcula varias veces el algoritmo de Dijkstra. El algoritmo de Dijkstra es un algoritmo voraz que calcula el camino más corto entre un nodo y el resto de nodos de un grafo. La librería BoostGraph[8] puede ser utilizada de manera muy eficiente para calcular la distancia más corta entre dos nodos en grafos de gran tamaño.

Además, en nuestro caso hemos utilizado parte del conocimiento disponible en recursos públicos para construir un gran grafo con 99.635 nodos (synsets) y 636.077 arcos (el conjunto de las relaciones directas entre synsets obtenida de WordNet[9] (Fellbaum, 1998) y eXtended WordNet[10] (Mihalcea y Moldovan, 2001)).

En la tabla 1 se puede ver el resultado del proceso de desambiguación semántica que hemos aplicado sobre algunos LUs del frame *Education_Teaching*. También se incluye el grado de polisemia de cada palabra (#sentidos), y la definición (glosa) del sentido (Synset) seleccionado por el algoritmo.

Evaluado sobre el mismo conjunto de evaluación que el utilizado por (Tonelli y Pianta, 2009) este algoritmo obtiene una Precisión del 78 %, un Recall del 63 % y un F1 del 69 % (Laparra y Rigau, 2009).

4.2. Clasificadores para conectar sentidos y frames

Como se explica en la Sección 5, en una fase anterior se validaron manualmente 150 de las 329 correspondencias entre sentidos y *frames* propuestos a través de los synsets asociados a cada recurso 4.1. Esta validación proporcionó 106 ejemplos positivos y 44 negativos de correspondencias *frame*–sentido. Estos ejemplos se usaron para entrenar diferentes clasificadores que, dado un par *frame*–sentido, puedan distinguir si se puede establecer una conexión entre ellos. Estos clasificadores se usaron para pre-validar los candida-

[6] Hay que tener en cuenta que sólo 9.325 LUs, son reconocidos por WordNet (un total del 92 %) correspondientes a sólo 672 *frames*.

[7] SenSem tiene diversos sentidos que no han sido asociados a ningún synset, para ellos el método de conexión a través de WordNet no se aplica.

[8] `http://www.boost.org/doc/libs/1_35_0/libs/graph/doc/index.html`

[9] `http://wordnet.princeton.edu`

[10] `http://xwn.hlt.utdallas.edu`

Lexical Unit	synset	#senses	Gloss
education.n	00567704-n	2	"activities that impart knowledge"
teacher.n	07632177-n	2	"a person whose occupation is teaching"
instruct.v	00562446-v	3	"impart skills or knowledge"
study.v	00410381-v	6	"be a student; follow a course of study; be enrolled at an institute of learning"
student.n	07617015-n	2	"a learner who is enrolled in an educational institution"
pupil.n	07617015-n	3	"a learner who is enrolled in an educational institution"

Cuadro 1: Resultado parcial del proceso de desambiguación de LUs del frame *Education_Teaching*.

tos a correspondencia restantes, para aliviar la tarea de anotación manual y para estudiar hasta qué punto el uso de clasificadores podía resultar útil.

Los pares de ejemplo fueron caracterizados mediante sus roles y *Frame Elements* (FEs) asociados al sentido. Para cada ejemplo se creó un vector cuyas dimensiones eran los posibles elementos de frame (FE) y sus roles. El valor en la n-ésima dimensión de este vector era 1 si el rol o FE estaba asociado al sentido o frame del par, y 0 si no lo estaba.

Para caracterizar los frames se asociaron al *frame* no solamente sus propios FEs, sino también los FEs de los frames heradados, según se desprende de la jerarquía de *frames* de FrameNet. Por ejemplo, a partir del FE *Speaker* asociado al frame *Communicate_categorization* se le asociaron también los elementos de frame *Expressor*, *Sentient_entity* y *Cognizer*, asociados a los frames de los cuales hereda sus propiedades *Communicate_categorization*.

Se han aplicado distintos sistemas de aprendizaje automático usando el entorno Weka (Witten y Frank, 2005): Support Vector Machines (SVM), clasificadores bayesianos (Naive Bayes), árboles de decisión (J48) y basados en reglas de decisión (JRip). Al tener pocos ejemplos anotados, se realizó una evaluación por validación cruzada (ten-fold cross-validation), y se tomó el factor kappa como indicador de la fiabilidad del clasificador automático para reproducir la clasificación humana.

Como se puede ver en el Cuadro 2, SVM obtienen los mejores resultados, alcanzando una buena coincidencia con la clasificación humana, con un factor kappa de $\kappa = 0{,}71$, lo cual indica buena fiabilidad del clasificador.

A partir de estos ejemplos iniciales, parece que los clasificadores obtenidos podrían ser útiles en un paso de pre-validación de las

clasificador	instancias bien clasificadas
SVM	87 %
Naive Bayes	83 %
J48	84 %
JRip	76 %

Cuadro 2: Resultados obtenidos por distintos clasificadores a partir de 150 ejemplos positivos y negativos de correspondencias frame – sentido, evaluados mediante ten-fold cross-validation.

correspondencias establecidas mediante los synsets de WordNet. En la siguiente sección detallamos el uso de los clasificadores en pre-validación.

5. Anotación manual

Dos lingüistas han validado manualmente los 329 pares entre un sentido de SenSem y un *frame* de FrameNet para los que se encontró una equivalencia mediante los synsets de WordNet asociados a cada uno de ellos. En una primera fase, se validaron 150 pares frame–sentido y se encontró correspondencia para el 70 %. De esta primera fase de validación se generaron los 106 ejemplos positivos y 44 negativos con los que se entrenaron los clasificadores descritos en la Sección 4.2.

Con estos clasificadores se pre-validaron los 179 pares todavía no validados por humanos. Los anotadores humanos validaron estos 179 pares, y se encontró correspondencia para el 42 % (un total de 75 pares). Esta segunda validación sirvió para evaluar el funcionamiento de los clasificadores, comparando las conexiones propuestas automáticamente por los clasificadores con las de los humanos.

Como vemos en el Cuadro 3, el acuerdo entre jueces y clasificadores es bajo, probablemente por discrepancias entre los ejemplos de los cuales se aprendieron los modelos y los ejemplos a los cuales se aplicaron los modelos.

Para empezar, observamos que el porcentaje de correspondencias encontradas en la primera fase es mucho mayor que el encontrado en la segunda (70 % vs. 42 %). Para tratar de mejorar la consistencia de la anotación manual, realizamos un estudio del acuerdo entre los jueces humanos y una casuística detallada de los pares a conectar, que presentamos en las siguientes dos secciones.

5.1. Acuerdo entre anotadores humanos

Para evaluar el grado de acuerdo entre anotadores humanos, 35 casos fueron validados por ambos anotadores. Inicialmente, el porcentaje de acuerdo entre ambos anotadores fue del 71 %. El coeficiente kappa de acuerdo entre jueces (Carletta, 1996) obtenido fue de $\kappa = 0,42$, con un rango entre 0,13 y 0,71 en el intervalo de 95 % de confianza. Este valor indica acuerdo moderado entre los anotadores.

Para mejorar el grado de acuerdo entre jueces se establecieron algunas pautas para determinar la correspondencia entre un *frame* y un sentido. Se estudiaron los casos más conflictivos y se explicitó el proceso de toma de decisiones para los casos más frecuentes. Se muestran algunos ejemplos en la Sección 5.2.

Después de establecer estas pautas, el porcentaje de acuerdo llegó al 94 %, con $\kappa = 0,88$ (entre $\kappa = 0,72$ y $\kappa = 1,04$ para 95 % de confianza). Este valor indica buena fiabilidad de la anotación humana, es decir, que diferentes anotadores tienden a producir los mismos resultados para diferentes anotaciones. Por lo tanto, parece que la tarea puede resolverse con un importante grado de acuerdo, siempre que se cuente con pautas explícitas para establecer correspondencias, como suele suceder en las tareas de anotación semántica por humanos.

Como consecuencia de este análisis de resultados, hemos creado unas directrices para mejorar la consistencia de la anotación. Estas directrices se usarán para mejorar la consistencia del conjunto de ejemplos de aprendizaje, y así mejorar el funcionamiento de los clasificadores.

5.2. Análisis de casos

Los dos anotadores humanos estudiaron diferentes pares de correspondencia sentido – *frame* para establecer unas pautas que ayuden a la consistencia de la anotación humana.

En primer lugar, recordemos que la relación se establece siempre entre un *frame*, es decir una clase genérica del inglés, y una unidad léxica del castellano. Sin embargo, los métodos automáticos detectan candidatos a correspondencia mediante synsets asociados a la unidad léxica del castellano y a las unidades léxicas asociadas a la clase genérica del inglés. Muy a menudo, las unidades léxicas asociadas a un frame son hipónimos de ese *frame*, así que la correspondencia propuesta involucra en realidad a una unidad léxica del inglés muy especializada y al sentido asociado al mismo synset, también muy especializado. En estos casos, no se puede establecer una correspondencia entre el sentido y el *frame* que contiene a la unidad léxica, como se puede ver en los dos primeros ejemplos de la Figura 1.

Pero la diferencia en granularidad de las unidades también se da en el sentido inverso, es decir, a menudo encontramos sentidos de SenSem que son más generales que el *frame* al cual han sido asociados, como se ve en el último ejemplo de la Figura 1. En el caso del *frame* "Trust", se trata siempre de una comunicación (trasmisión de información), mientras que en español, "creer_3" está definido tanto como una comunicación como por un comportamiento, caso que no está contemplado en el esquema *Trust*. En estos casos, la solución pasa por o bien doblar la unidad del castellano para poder realizar una relación unívoca o bien relacionar la entrada con dos esquemas.

6. Conclusiones y Trabajo Futuro

Hemos presentado un estudio de métodos y técnicas para establecer una metodología para integrar los modelos semánticos verbales de diferentes lenguas, FrameNet del inglés y SenSem del castellano. El objetivo final es conectar los predicados verbales de SenSem con FrameNet.

Hemos estudiado la dificultad intrínseca de establecer correspondencias entre sentidos de SenSem y frames de FrameNet. Hemos estudiado casos conflictivos y hemos obtenido el grado de acuerdo entre jueces. Hemos conseguido aumentar la fiabilidad del grado de acuerdo mediante algunos criterios que aumentan la consistencia de la anotación.

Con el objetivo de reducir el coste de producir recursos tan complejos y aumentar la consistencia de la conexión manual entre am-

	sí	no	auto
sí	67	8	
no	92	12	
manual			(**44 %**)

SVM

	sí	no	auto
sí	47	28	
no	77	27	
manual			(**41 %**)

Naive Bayes

	sí	no	auto
sí	49	26	
no	62	42	
manual			(**50 %**)

J48 (árboles de decisión)

Cuadro 3: Descripción del acuerdo entre jueces humanos (filas, *manual*) y clasificadores automáticos (columnas, *auto*) para validar correspondencias entre pares *frame*–sentido, para los clasificadores SVM (Support Vector Machines), Naive Bayes y J48 (árboles de decisión).

Frame	**synset**	**Sentido**
Arriving: *An object Theme moves in the direction of a Goal. The Goal may be expressed or it may be understood from context, but its is always implied by the verb itself.* unidad léxica: **come**	1262658	**acercar_5:** Arrimarse a alguien o a algo.
Trust: *A Cognizer thinks that the Information given by a particular Source is correct. The specific Content or Topic of the Information may also be described.* unidad léxica: **believe**	0461554	**creer_2:** Dicho de la religión, tener la esperanza de un mundo mejor o de la existencia de un dios.
Trust: (ídem anterior) unidad léxica: **trust**	0488417	**creer_3:** Estar seguro que algo o alguien obrará correctamente o que no nos engañará.

Figura 1: Ejemplos de diferentes granularidades entre los sentidos de SenSem y los frames de FrameNet. Se muestra también la LU de FrameNet a partir de la cual se ha realizado la correspondencia, y el synset asociado a la LU y al sentido.

bos dos recursos, hemos aplicado dos técnicas automáticas: desambiguación de sentidos y clasificación de los pares sentido–*frame* candidatos a conexión. Se han aplicado técnicas de desambiguación de sentidos a las unidades léxicas de FrameNet, para asociarles automáticamente un synset de WordNet. Se establecen como candidatos a conexión los pares de frames y sentidos asociados al mismo synset. El 55 % de estos candidatos fueron validados como correspondencias.

Además, se aprendieron clasificadores para determinar automáticamente si un candidato era efectivamente una correspondencia. Los clasificadores aprendidos ofrecieron resultados muy variables, probablemente debido a que los ejemplos de aprendizaje fueron generados por dos jueces sin consenso previo en los criterios de anotación. Después del análisis de resultados se crearon unas directrices para mejorar la consistencia de la anotación.

El primer paso del trabajo futuro será la creación de ejemplos más consistentes para mejorar el aprendizaje de clasificadores. Después aplicaremos estos clasificadores para conectar las unidades de FrameNet y SenSem que no se asociaron mediante correspondencias entre synsets. Se generarán candidatos con el producto cartesiano de todas las unidades no conectadas, esos candidatos serán pre-validados mediante clasificadores y validados manualmente. Una vez validadas, estas correspondencias pasarán a engrosar el conjunto de ejemplos a partir de los cuales se aprenden los clasificadores, lo cual debería mejorar el funcionamiento de los mismos.

Bibliografía

Alonso, L., J. A. Capilla, I. Castellón, A. Fernández, y G. Vázquez. 2007. The sensem project: Syntactico-semantic annotation of sentences in spanish. *Selected papers from RANLP 2005*.

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*.

Baker, C., C. Fillmore, y J. Lowe. 1997. The berkeley framenet project. En *COLING/ACL'98*, Montreal, Canada.

Burchardt, A., K. Erk, y A. Frank. 2005. A WordNet Detour to FrameNet. En *Proceedings of the GLDV 2005 GermaNet II Workshop*, páginas 408–421.

Carletta, J. 1996. Assessing agreement on classification tasks: the kap-

pa statistic. *Computational Linguistics*, 22(2):249–254.

Carreras, X. y L. Màrquez. 2004. Introduction to the conll-2004 shared task: Semantic role labeling. En *CoNLL-2004*.

Carreras, X. y L. Màrquez. 2005. Introduction to the conll-2005 shared task: Srl. En *CoNLL-2005*.

Cuadros, M. y G. Rigau. 2008. KnowNet: a proposal for building knowledge bases from the web. En *First Symposium on Semantics in Systems for Text Processing, STEP'08*, Venice, Italy.

Davies, Mark. 2006. *Diccionario Español de Frecuencia*. Routledge.

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

Fillmore, Charles J. 1976. Frame semantics and the nature of language. 280:20–32.

Gildea, D. y D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Giuglea, A. M. y A. Moschitti. 2006. Semantic role labeling via framenet, verbnet and propbank. En *ACL-44*, páginas 929–936.

Johansson, Richard y Pierre Nugues. 2007. Using WordNet to extend FrameNet coverage. En *Proceedings of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages*.

Kipper, Karen. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. tesis, Univ. of Pennsylvania.

Laparra, Egoitz y German Rigau. 2009. Integrating wordnet and framenet using a knowledge-based word sense di sambiguation algorithm. En *Proceedings of RANLP'09*, Borovets, Bulgaria.

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

Navigli, R. 2005. Semi-automatic extension of large-scale linguistic knowledge bases. En *FLAIRS*.

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

Palmer, Martha, Daniel Gildea, y Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Pazienza, Maria Teresa, Marco Pennacchiotti, y Fabio Massimo Zanzotto. 2006. Mixing WordNet, VerbNet and PropBank for studying verb relations. En *LREC'06*.

Pennacchiotti, Marco, Diego De Cao, Roberto Basili, Danilo Croce, y Michael Roth. 2008. Automatic induction of FrameNet lexical units. En *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*.

Pradhan, Sameer, Edward Loper, Dmitriy Dligach, y Martha Palmer. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. En *SemEval-2007*, páginas 87–92.

Ruppenhofer, Josef, Caroline Sporleder, Roser Morante, Collin Baker, y Martha Palmer. 2010. Semeval 2010 shared task: Filling the gaps in semantic role labelling. `http://www.coli.uni-saarland.de/projects/semeval2010_FG/`.

Shi, Lei y Rada Mihalcea. 2005. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. En *CICLing*, Mexico.

Subirats, Carlos y Miriam R.L. Petruck. 2003. Surprise: Spanish framenet! En *Proceedings of the International Congress of Linguists*, Praga.

Tonelli, Sara y Emanuele Pianta. 2009. A novel approach to mapping framenet lexical units to wordnet synsets. En *Proceedings of IWCS-8*.

Tonelli, Sara y Daniele Pighin. 2009. New features for framenet - wordnet mapping. En *Proceedings of CoNLL'09*, Boulder, CO, USA.

Witten, Ian H. y Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.