

Un Enfoque Integrado para la Desambiguación*

Jordi Atserias i Batalla

TALP Research Center
Jordi Girona Salgado, 1-3.
08034 Barcelona
{batala}@lsi.upc.edu

Resumen: En este artículo presentamos una extensión de una arquitectura integrada diseñada originalmente para *Semantic Parsing a WSD*. El marco propuesto permitirá que ambas tareas puedan colaborar y llevarse a cabo simultáneamente. Se ha probado la validez y robustez de esta arquitectura contra una tarea de *WSD* bien definida (el *SENSEVAL-II English Lexical Sample*) aplicando modelos sintáctico-semánticos adquiridos automáticamente.

Palabras clave: Arquitecturas integradas para el PLN, WSD, Semantic Parsing

Abstract: This paper presents an extension for WSD of an integrated architecture designed for Semantic Parsing. In the proposed framework, both tasks could be addressed simultaneously, collaborating between them. The feasibility and robustness of the proposed architecture have been proved against a well-defined task on WSD (the *SENSEVAL-II English Lexical Sample*) using automatically acquired models.

Keywords: Integrated Architectures for NLP, WSD, Semantic Parsing

1. Introducción

La mayoría de las diferentes tareas incluidas en el PLN, como la desambiguación de sentidos (*WSD*), la extracción de información (IE), la búsqueda de respuestas (QA) o la traducción automática aplican diferentes niveles de comprensión del lenguaje (NLU).

La comunidad del PLN se ha centrado en los últimos años en la evaluación de diferentes tareas, tales como *WSD* (*SENSEVAL*¹), parsing (*PARSEVAL*), IE (*MUCs*) y IR (*TRECS*). Para mejorar no sólo los resultados en estas tareas sino en la comprensión del lenguaje en general, es necesario abordar los interrogantes que han aparecido en estas competiciones, tales como: la integración de diferente tipos de conocimiento, la integración de los diferentes procesos dentro del PLN, el uso de ontologías, el razonamiento o el multilingüismo.

El objetivo de este trabajo es explorar el uso de nuevas arquitecturas para la comprensión del lenguaje, que sean lo más robustas y flexibles posibles y que intenten abordar con nuevos enfoques estas problemáticas. El trabajo que se presenta se enmarca dentro de

uno de los pasos fundamentales en NLU, la Interpretación Semántica. Nuestro objetivo a medio plazo es la integración de dos de las etapas de la interpretación semántica: La *Desambiguación de Sentidos (WSD)* y el *Análisis Semántico (Semantic Parsing)*.

El *WSD* consiste en determinar el sentido de una palabra dado el contexto en que aparece. Normalmente los sentidos posibles de una palabra se definen en un repositorio de sentidos, siendo WordNet (Fellbaum, 1998) el *de facto* estándar. Por otro lado, el *Semantic Parsing* conlleva la identificación de los roles semánticos de las entidades, (p.e. Agente) (Brill y Mooney, 1997). Este proceso es conocido como *Semantic Role Labeling* y a sido objetivo de las *shared tasks* de los recientes *SENSEVAL* y *CONLL*².

En este artículo aplicaremos una arquitectura integrada a la tarea de *WSD* que ya ha sido aplicada con anterioridad al *Semantic Parsing* (Atserias, Padró, y Rigau, 2001).

2. Semantic Parsing y WSD

Aunque el *WSD* y el *Semantic Parsing* están fuertemente ligados, tradicionalmente, la mayoría de sistemas los tratan por separado. Paradójicamente, el *WSD* puede mejorar

* El autor quiere agradecer la ayuda de Eli Comelles así como los comentarios de los revisores.

¹<http://www.senseval.org/>

²<http://www.cnts.ua.ac.be/conll/>

modelo <i>impersonal</i> de "hablar"					
Sint.	Prep.	Rol	Semántica	Concordancia	Opcional
SE	x	se	Top	no	no
PP	de, sobre	entity	Top	no	yes
PP	con	destination	Top	no	yes

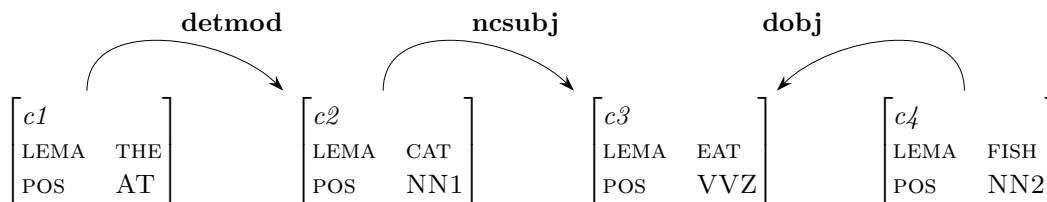
Cuadro 1: Ejemplo de modelo Sintáctico-Semántico para *Semantic Parsing*

Figura 1: Dependencias para "The cat eats fish"

los resultados del *Semantic Parsing*, ya que los diferentes sentidos de una palabra pueden presentar diferentes estructuras sintácticas (especialmente los verbos) y viceversa (p.e. usando preferencias de selección para determinar el sentido correcto de un verbo (Carroll y McCarthy, 2000)). En este trabajo presentaremos un enfoque integrado que podría utilizarse para combinar simultáneamente estas dos tareas.

Nuestra aproximación al *WSD* seguirá la misma formulación utilizada en (Atserias, Padró, y Rigau, 2001), donde se presenta una arquitectura para *Semantic Parsing* basada en la aplicación de modelos verbales lexicalizados, como el mostrado en el cuadro 1, donde se combina información sintáctica (preposición, concordancia, etc) y semántica (roles, preferencias de selección). En este sistema el *Semantic Parsing* se lleva a cabo encontrando el modelo más similar a la frase de entrada. Siguiendo el mismo enfoque, si asociamos el sentido de WordNet a estos modelos lexicalizados, a la vez que identificamos el modelo más similar estaremos identificando el sentido correcto de esa palabra. Además estaremos construyendo un marco de trabajo donde podremos aplicar el *Semantic Parsing* y el *WSD* simultáneamente.

Durante el preproceso la frase de entrada que contiene la palabra a desambiguar será analizada sintácticamente utilizando RASP (Carroll, Minnen, y Briscoe, 1998). La figura 1 muestra el análisis de dependencias resultante para la frase *The cat eats fish*.

Cada palabra será etiquetada con todos

sus posibles sentidos de WordNet. El etiquetado de texto libre con los sentidos de WordNet no es una tarea fácil, ya que los criterios de lematización o la identificación de términos multipalabra son diferentes en una herramienta general (RASP) que en WordNet. Por todo ello utilizamos un reconocedor de términos multipalabra específico para WordNet (Arranz, Atserias, y Castillo, 2005), en vez de la lematización/tokenización provista por RASP.

Una vez determinados los posibles sentidos en WordNet, podremos añadir la información semántica asociada a cada sentido de WordNet en el *Multilingual Central Repository* (MCR) (Atserias et al., 2004): etiquetas de la Top Concept Ontology de EuroWordNet (Vossen, 1998) expandida (Atserias, Climent, y Rigau, 2004), la Suggested Upper Merged Ontology (SUMO) (Niles y Pease, 2001) y los Dominios de MultiWordNet (Magnini y Cavaglia, 2000).

El resultado obtenido (dependencias gramaticales e información semántica) para cada palabra se convierte en una estructura de rasgos que será nuestra entrada al desambiguador. La Figura 2 muestra la estructura de rasgos obtenida para *fish* en sus dos sentidos: el de alimento (*fish#n#1*) y el de animal (*fish#n#2*). En adelante utilizaremos el término *objeto* para referirnos a estas estructuras de rasgos.

3. El PLN como Problema de Satisfacción de Restricciones

Para encontrar el modelo más similar a la frase de entrada utilizaremos el marco

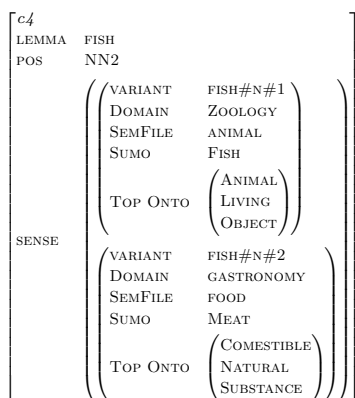


Figura 2: Objeto *Fish*

de los problemas de satisfacción de restricciones (CSP). Los CSPs han sido aplicados anteriormente en diferentes áreas del PLN, Part of Speech tagging (Padró, 1998), análisis sintáctico *Weighted Constraint Dependency Grammars* (Foth, Menzel, y Schröder, 2003). Incluso se han aplicado a tareas más complejas como la traducción automática (Mikrokosmos (Beale, 1996)).

En la mayoría de aplicaciones del PLN, y especialmente en el *WSD*, es necesario por un lado expresar vaguedad, posibilidades, preferencias, y por otro, integrar diferentes fuentes de conocimiento. Estos dos factores aumentan la complejidad de encontrar la mejor solución, ya que en muchos casos, o bien no existe una solución que satisfaga todas las preferencias o bien las fuentes de conocimiento pueden llegar a ser contradictorias. A pesar de los recientes esfuerzos en el área (p.e. (Rudova, 2001)), encontrar la mejor solución a este tipo de problemas a nivel global y de manera eficiente todavía es una cuestión abierta.

Los *Consistent Labeling Problem* (CLP) (Messeguer y Larossa, 1995), son una manera natural de modelar este tipo de CSPs. Una vez formulada nuestra tarea de NLU como un CLP, este puede ser resuelto eficientemente (máximo local que satisfaga en el mayor grado posible el conjunto de restricciones) usando el algoritmo iterativo de relajación de etiquetas (Relaxation Labeling).

Usando como marco teórico los CLPs, es posible integrar no sólo diferentes tipos de conocimiento sino también diferentes procesos del PLN. Simultáneamente, nos permitirá aplicar cualquier tipo de conocimiento (sintáctico, semántico, lingüístico, estadísti-

co) tan pronto como sea posible (Integrated Processing Hypothesis (Birnbaum, 1989)), pero manteniendo al mismo tiempo una representación independiente para cada uno de estos tipos de conocimiento.

4. CLP

Un *Consistent Labeling Problem* básicamente consiste en encontrar una asignación consistente de un conjunto de variables dado un conjunto de restricciones. Formalmente, un CLP se define mediante un conjunto de variables V_i , un conjunto de valores (dominio) para cada variable D_i , y una relación de compatibilidad entre tuplas. Las compatibilidades son funciones reales $r_{ij} : D_i D_j \rightarrow \mathbb{R}$ donde $r_{i,j}(a, b)$ indica la compatibilidad de asignar simultáneamente la etiqueta a a V_i y la etiqueta b a V_j .

De manera similar a los CSPs, que intentarían encontrar un conjunto de asignaciones que no violara las restricciones, un CLP busca un etiquetado tal que las asignaciones sean lo más compatibles posibles respecto a la función de compatibilidad.

Las estructuras de rasgos (objetos) que forman la entrada a nuestro sistema serán representadas en el CLP mediante un conjunto de asignaciones. Cada rasgo equivaldrá a una variable cuyo dominio (posibles valores) serán los valores del rasgo. Así la variable $c1.att$ denotará el rasgo *att* del objeto *c1*. Estos objetos se representan de manera amalgamada en el CLP, de manera que se combinan/superponen las representaciones asociadas a cada sentido. De una manera similar a las *Polaroid Words* (Hirst, 1987) a la vez que se escogerá el mejor modelo, también se escogerá la mejor representación del objeto y viceversa.

Muchos de los problemas modelados como CLP no tienen una estructura implícita. Para representar esta estructura entre objetos, es decir la aplicación de un modelo, utilizaremos dos variables especiales asociadas a cada objeto: la variable *model* que representa cual de los modelos del objeto se aplica y la variable *role* representa que rol desempeña el objeto dentro de un modelo.

Ya que un CLP asigna siempre un valor a una variable, necesitaremos una etiqueta especial NONE, para representar la posibilidad de no utilizar ningún modelo y otra variable TOP, para representar que el objeto no forma parte de ningún modelo.

Variable	Values
C1.POS	{ NN1 }
C1.HEAD	{ cat }
C1.SENSE	{ cat#n#1, cat#n#2 ... }
C1.MODEL	{ NONE }
C1.ROLE	{ subj.m1.c3, subj.m2.c2 }
C2.POS	{ VVZ }
C2.HEAD	{ eat }
C2.SENSE	{ eat#v#1, eat#v#2 ... }
C2.MODEL	{ transitive }
C2.ROLE	{ TOP }
C3.POS	{ NN1 }
C3.HEAD	{ fish }
C3.SENSE	{ fish#n#1, fish#n#2 }
C3.MODEL	{ NONE }
C3.ROLE	{ dobj.m1.c2 }

Figura 3: CLP para *The cat eats fish*

La Figura 3 muestra la representación como CLP de la frase *The cat eats fish*. En esta figura se puede observar que para identificar el valor de una variable *role*, utilizamos una tupla (*rol*, *objeto*, *modelo*). Así el role *dobj* del modelo *m1* del objeto *eat* (*c3*) se representaría como la etiqueta (*dobj*, *c3*, *m1*).

4.1. Comparando Roles y Objetos

Para determinar qué roles de un modelo pueden ser desempeñados por los objetos de la frase, es necesario establecer una función de similitud entre un *objeto* y un *rol*: $sim(obj, rol)$. Una vez determinadas qué parejas rol-objeto se pueden instanciar, tendremos que decidir cuales se pueden utilizar conjuntamente para instanciar un modelo.

Sin embargo, parte de los atributos/asignaciones que determinan lo bien que un objeto puede desempeñar un role no cambian de valor (son estáticos). De manera que la función sim puede ser dividida en dos partes: una parte dinámica sim_{dyn} y otra estática sim_{static} que puede ser calculada una sola vez (p.e. cuando se construye el CLP) y utilizarse para determinar qué objetos pueden desempeñar un rol en el CLP. Por otra parte, la función de similitud que depende de atributos dinámicos sim_{dyn} tiene que ser convertida a un conjunto de restricciones que dependan del estado del CLP (los pesos asociados a las asignaciones en una iteración).

En los experimentos llevados a cabo, el único atributo dinámico considerado es el sentido. Así mismo y por simplicidad, en los experimentos se ha utilizado una función de similitud entre un rol y un objeto, que combina de manera independiente las funciones de similitud para cada rasgo:

$$sim(obj, rol) = \frac{\sum_{a \in Atts} sim_{att}(rol.a, obj.a)}{\#Atts}$$

A continuación, se describen el conjunto de restricciones que aseguran la buena formación de los modelos (estructurales) y la bondad de la aplicación del modelo y los roles (cotejo). Estas restricciones tienen asociadas un peso y pueden ser tanto de compatibilidad (\sim) como de incompatibilidad (\approx).

4.2. Restricciones Estructurales

- **Unicidad de Objeto:** El primer axioma establece que un objeto no puede ocupar más de un rol:

$$[c_x.role = a] \approx [c_x.role = b] \\ \forall x \in Obj \forall a, b \in Roles(c_x) \mid a \neq b$$

- **Unicidad de Rol:** Dado un role, este sólo puede ser instanciado por un objeto:

$$[c_x.role = a] \approx [c_y.role = a] \\ \forall x, y \in Obj \forall a \in Roles \mid x \neq y$$

Esta restricción evitaría por ejemplo que los objetos *cat* y *fish* ocuparan simultáneamente el mismo role.

- **Unicidad de Modelo:** Restringiremos los modelos asociados a un mismo objeto a ser incompatibles entre ellos:

$$[c_x.model = a] \approx [c_x.model = b] \\ \forall x \in Obj \forall a, b \in Models, x \neq y$$

- **Inconsistencia de Modelo:** Un role no puede ser instanciado por un objeto si el modelo al que pertenece el role no está siendo instanciado:

$$[c_x.model = m_b] \approx [c_y.role = (r, m_a, x)] \\ \forall x, y \in Obj (r, x, m_a) \in Roles(y) \\ m_b, m_a \in Modelos(x) \mid m_a \neq m_b$$

- **Unicidad del TOP** Sólo hay un TOP:

$$[c_x.model = TOP] \approx [c_y.model = TOP] \\ \forall x, y \in Obj, x \neq y$$

- **Existencia del TOP** Tiene que existir como mínimo un TOP:
 $[c_x.model = TOP] \sim \# [c_y.model = TOP] \forall x, y \in Obj \mid x \neq y$
- **Soporte al NONE** El modelo NONE es compatible con la no existencia de ninguna asignación de role:

$$[c_y.model = NONE] \sim \# [c_y.role = a] \forall y \in Obj$$

4.2.1. Restricciones de Cotejo

- **Aplicabilidad del Modelo** Para no penalizar excesivamente los modelos pequeños el soporte que recibe cada modelo se normaliza por el número de roles que tiene.

$$[c_x.model = m] \sim [c_y.role = (r, m, x)] \forall (r, m, x) \in Roles$$

Por ejemplo, si el modelo *eat-V4* tiene tres roles (subj, dobj, dobj2), la restricción que daría soporte al modelo dependiendo de la asignación del rol *dobj2* sería $[c3_{model} = eat-V4] \sim \frac{1}{3} [c3_{role} = (dobj2, eat-V4, c2)]$. El modelo tendría además una restricción similar para sus otros dos roles (subj y dobj).

- **Aplicabilidad del Rol** El soporte dado a los roles debe tener en cuenta el sentido del objeto que lo instancia, por lo que tendremos que comparar cada sentido del objeto y del role:

$$[c_{role} = (r, m, x)] \sim^w [c_{sense} = s] \forall c, x \in Obj \forall s \in c.sense \text{ y donde } w \text{ es } sim_{dyn} \text{ entre } s, \text{ el sentido del objeto, y el sentido del role.}$$

Por ejemplo, la restricción $[c3_{role} = (dobj2, eat-V4, c2)] \sim^{2,45} [c3_{sense} = fish\#n\#2]$ dará soporte a la asignación del rol $(dobj2, eat-V, c2)$ teniendo en cuenta el peso de la asignación del sentido *fish\#v\#2* y su similitud en WordNet³ con el/los sentido/s del role $(dobj2, eat-V4, c2)$.

³En los experimentos hemos utilizado el nivel del primer ancestro común

4.3. Restricciones de Sentido

A la vez que aplicamos un modelo, queremos que se seleccione el sentido asociado a este. Tanto para los roles como para el núcleo del modelo. Para ello se definen dos conjuntos de restricciones.

- **Head Sense Disambiguation** Este grupo de restricciones asociarán el modelo con el sentido de su núcleo:

$$[c_{sense} = s] \sim^{100} Or_{i=1}^n [c_{model} = m_i] \forall s \in c.sense \text{ y donde } m_1 \dots m_n \text{ es el conjunto de modelos de } c \text{ cuyo sentido es } s$$

Por ejemplo, la restricción $[c2_{sense} = eat\#v\#3] \sim^{100} [c2_{model} = eat-V17] \text{ or } [c2_{model} = eat-V52] \text{ or } [c2_{model} = eat-V50]$ dará más peso a la asignación del tercer sentido de *eat* si alguno de los modelos asociados a este sentido (*eat-V17*, *eat-V52*, *eat-V50*) es seleccionado.

- **Role Sense Disambiguation** Este grupo de restricciones asociará el sentido del rol con el objeto que lo instancia:

$$[c_{sense} = r.sense] \sim^w [c_{role} = (r, m, x)] \forall c \in Obj \text{ donde } w \text{ es } sim_{static}(obj, role)$$

Por ejemplo, $[c3_{sense} = fish\#n\#2] \sim^2 [c3_{role} = (dob2, eat-V4, c2)]$ seleccionará el segundo sentido de *fish* si el objeto *c3* instancia el role *dobj2* del modelo *eat-V4*.

4.4. Etiquetado inicial

Dado que el método de *relaxation labeling* es un algoritmo iterativo de convergencia local, uno de los parámetros más significativos es el estado inicial a partir del cual empezamos las iteraciones. Heurísticamente hemos inicializado las asignaciones de roles y modelos de acuerdo con la función estática de similitud, mientras que para las relativas a los sentido hemos utilizado la frecuencia en Sencor más uno.

5. Experimentos

SENSEVAL proporciona un marco ideal para la evaluación del WSD. Así que para demostrar la flexibilidad y robustez de nuestra

arquitectura para *WSD*, decidimos aplicar nuestro sistema a la tarea *English Lexical Sample* del SENSEVAL-II. Esta tarea consiste en desambiguar las ocurrencias de 73 palabras⁴ diferentes (nombres, verbos y adjetivos) en un corpus de 4.328 párrafos. Escogimos esta tarea específica ya que pensamos que reduciría el impacto tanto de los errores en el preproceso, como la falta de modelos (ya que se proporciona un corpus de entrenamiento). Además, a diferencia de su equivalente en el SENSEVAL-III, SENSEVAL-II utiliza directamente los sentidos de WordNet para los verbos.

Para afrontar esta tarea con nuestro sistema, necesitaremos modelos sintácticos que a la vez contengan información de los sentidos en WordNet de las palabras a desambiguar. Desgraciadamente, en la actualidad, incluso para el inglés existen pocos recursos de amplia cobertura que proporcionen modelos sintáctico-semánticos que se puedan vincular a WordNet. Aunque en este sentido cabe destacar los recientes esfuerzos para vincular recursos como FrameNet y VerbNet con WordNet (Shi y Mihalcea, 2005).

Por lo tanto, a pesar de su inherente complejidad, decidimos obtener de manera automática este tipo de modelos a partir de corpus. Obtener esta clase de modelos automáticamente presenta múltiples dificultades. Primeramente, la falta de corpus desambiguado o su pequeño tamaño hace difícil poder obtener modelos que cubran todos los sentidos de WordNet o que contemplen todas las alteraciones dietéticas o patrones de subcategorización que estos puedan presentar. Además, tanto los métodos de desambiguación como de análisis sintáctico todavía presentan una alta tasa de error que difícilmente puedan superar los algoritmos de aprendizaje.

5.1. Adquisición de Modelos

Para crear nuestros modelos a partir de corpus etiquetados semánticamente utilizamos la misma cadena de proceso que para la entrada a nuestro sistema (ver sección 2), obteniendo para cada frase un conjunto de análisis de dependencias enriquecido con información semántica del MCR. En cada frase, extraeremos para cada palabra las estructuras de rasgos del conjunto de dependencias directas (p.e. subj / obj / dobj) que llegan a

⁴Que corresponden a 241 términos multipalabras

esta e interpretaremos estas como los roles de un nuevo modelo para esta palabra. Debido al enorme número de modelos, nuestra primera aproximación para este experimento ha sido restringir el uso de los modelos a utilizar a aquellos en los cuales el núcleo sintáctico es una de las palabras a desambiguar.

Los modelos utilizados se generaron a partir de dos corpus de muy diferentes características. Por una parte, SemCor (Miller et al., 1993), que en su mayor parte está totalmente desambiguado pero debido a su escaso tamaño (cerca de 250.000 palabras) tiene muy poca cobertura, y por otra parte, el propio corpus de entrenamiento del *English Lexical Sample (training)* cuyos 8.611 ejemplos sólo tienen etiquetado el sentido de la palabra a desambiguar. El cuadro 2 muestra el número de modelos para las palabra a desambiguar obtenidos de cada corpus. Cabe remarcar que aunque se obtengan más modelos de Semcor la distribución de estos modelos entre los diferentes sentidos es muy diferente a la del *Training*. Ya que, mientras que en el *Training* los modelos están distribuidos entre todos los sentidos que aparecen en el test, en el Semcor se concentran solamente en los sentidos más frecuentes y no llegan a cubrirlos todos.

	Número de Modelos
Semcor	7.344
Training	4.438

Cuadro 2: Modelos obtenidos

6. Resultados

El cuadro 3 muestra los resultados (Precisión y Recall) obtenidos para el *English Lexical Sample* de SENSEVAL-II utilizando los modelos obtenidos a partir de Semcor y del corpus de entrenamiento de la propia tarea (*Training*). Usando los modelos obtenidos de cada corpus, se realizaron tres experimentos variando el nivel de semántica utilizado para determinar la bondad de la instanciación de los roles: sin semántica (*Syntax*), utilizando sólo la posición del sentido en WordNet (*Synset*) y utilizando toda la información asociada a un sentido en el MCR. Respecto a los atributos sintácticos, restringimos los objetos que pueden instanciar un rol a aquellos que tienen el mismo tipo de relación sintáctica y la misma preposición que el propio rol. Esta condición es probable-

mente demasiado restrictiva y reduce el impacto del uso de más información semántica en los resultados.

	Modelos			
	Training		Semcor	
	P	R	P	R
MCR	48.3	26.9	28.3	15.9
Synset	48.2	26.9	27.5	15.5
Syntax	47.9	26.8	27.0	15.2

Cuadro 3: Resultado en **P**recisión y **R**ecall

Utilizando la evaluación del SENSEVAL con mayor granularidad (*coarse*) nuestro sistema llega al 59 % de precisión (41 % usando Semcor), creemos que este salto tan inusual es debido a la falta de modelos aplicables del sentido correcto, sobretodo para Semcor (nuestra formulación fuerza a escoger el sentido más parecido de entre los que hay un modelo aplicable). Si para cada frase del test calculamos si existe un modelo para el sentido correcto con la misma dependencia-preposición que la palabra a desambiguar, podemos establecer que la cota superior a la que el sistema actual podría llegar es del 70 %.

Por eso, los resultados obtenidos demuestran la validez de la aproximación, aunque estén ligeramente por debajo del estado del arte en *WSD*. Más aun, podemos valorar nuestra aproximación como satisfactoria, teniendo en cuenta que estamos usando nuestra formulación sin ningún tipo de ajuste (atributos o función de similitud) y que los modelos se han obtenido automáticamente (sin ningún tipo de generalización).

Además el marco en el cual se formaliza nuestro sistema (CLP) permite la integración/combinación de otras fuentes de conocimiento o procesos (p.e. heurísticos utilizando dominios) complementarios a esta formulación con los que mejorar los resultados.

7. Limitaciones de los modelos

Los modelos que hemos obtenidos automáticamente son limitados, y no siempre permiten construir una representación semántica adecuada, por ejemplo en el caso de ... *clean dental surface* ..., la representación semántica correspondiente a *dental surface* queda asociada al sentido de *surface*. Como consecuencia desambiguaremos

erróneamente el verbo *clean*, ya que los modelos de *clean#v#3* (limpiar la casa) están más relacionado con limpiar una *superficie*.

Por otro lado, el prototipo actual sólo integra de una manera superficial el nivel sintáctico con el semántico y por tanto no es robusto a errores en el análisis sintáctico ni le es posible desambiguar una palabra, si no tenemos un análisis de dependencias. Añadiendo a nuestro sistemas heurísticos de *WSD* que no requieran sintaxis (p.e. basados en dominios) se ganaría robustez.

En el caso de los modelos obtenidos a partir del Semcor, aunque semánticamente más informativos también son mucho más esparsos con lo que es difícil superar las inconsecuencias o errores de etiquetado en el corpus.

La capacidad de desambiguación del sistema también depende enormemente de la información disponible para discriminar los sentidos, por lo tanto difícilmente desambiguará palabras cuyos sentidos tengan un comportamiento sintáctico parecido y una representación semántica parecida (p.e. los cinco sentidos del nombre *child*).

8. Conclusiones y Trabajo futuro

Hemos demostrado que es posible desarrollar arquitecturas más robustas y flexibles, utilizando problemas de satisfacción de restricciones y que estos pueden ser resueltos de manera razonable utilizando métodos de optimización bien conocidos. Además, el enfoque empleado puede ser extendido a otros modelos que combinen información sintáctico-semántica aunque sean diferentes en su naturaleza y objetivos (p.e. FrameNet).

En este artículo hemos presentado una arquitectura capaz de integrar *Semantic Parsing* y *WSD*, y donde ambas tareas pueden colaborar. Los resultados obtenidos en una tarea real de *WSD* (SENSEVAL-II English Lexical Sample) utilizando modelos adquiridos automáticamente muestran la viabilidad de la propuesta, a pesar de estar lejos todavía de combinar totalmente estas dos etapas.

Las líneas de trabajo futuro se centrarán básicamente en: por un lado, ampliar la integración entre el *Semantic Parsing* y el *WSD*, utilizando modelos semánticos más ricos que permitan realmente llevar a cabo el *Semantic Parsing*, y por otro lado, mejorar el propio sistema, tanto los métodos de selección de los modelos a aplicar como incluyendo otros métodos de *WSD* o explorando como ajustar

las funciones de similitud.

Bibliografía

- Arranz, Victoria, Jordi Atserias, y Mauro Castillo. 2005. Multiword expressions and word sense disambiguation. En Alexander Gelbukh, editor, *CICLING'05*, volumen LNCS 3406.
- Atserias, J., S. Climent, y G. Rigau. 2004. Towards the meaning top ontology: Sources of ontological meaning. En *4rd International Conference on Language Resources and Evaluations (LREC)*.
- Atserias, Jordi, Lluís Padró, y German Rigau. 2001. Integrating multiple knowledge sources for robust semantic parsing. En *Proceedings of the International Conference, Recent Advances on Natural Language Processing RANLP'01*, Bulgaria.
- Atserias, Jordi, Luis Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, y Piek Vossen. 2004. The MEANING multilingual central repository. En *Proceedings of the Second International Global WordNet Conference (GWC'04)*, Brno, Czech Republic, January. ISBN 80-210-3302-9.
- Beale, Stephen. 1996. *Hunther-Gatheter: Applying Constraint Satisfaction, Branch-and-Bound and Solution Synthesis to computational Semantics*. Ph.D. thesis, Computer Research Laboratory, New Mexico State University.
- Birnbaum, L. 1989. A critical look at the foundations at autonomous syntactic analysis. En *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*, páginas 99–106.
- Brill, Eric y Raymond J. Mooney. 1997. An Overview of Empirical Natural Language Processing. *Artificial Intelligence Magazine*, 18(14):13–24. Special Issue on Empirical Natural Language Processing.
- Carroll, J. y D. McCarthy. 2000. Word sense disambiguation using automatically acquired verbal preferences. *Computers and the Humanities. Senseval Special Issue*, 34(1-2).
- Carroll, J., G. Minnen, y E. Briscoe. 1998. Can subcategorisation probabilities help a statistical parser? En *Proceedings of the Sixth ACL/SIGDAT Workshop on Very Large Corpora*, páginas 118–126.
- Fellbaum, C., editor. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.
- Foth, Kilian, Wolfgang Menzel, y Ingo Schröder. 2003. Robust parsing with weighted constraints. to appear in *Natural Language Engineering*.
- Hirst, Graeme. 1987. *Semantic Interpretation and the Resolution of the ambiguity*. Studies in Natural Language Processing. Cambridge University Press.
- Magnini, B. y G. Cavaglia. 2000. Integrating subject field codes into wordnet. En *Proceedings of the Second International Conference on Language Resources and Evaluation LREC'2000*, Athens. Greece.
- Messeguer, Pedro y Javier Larossa. 1995. Constraint satisfaction as global optimization. En *IJCAI'95*.
- Miller, G., C. Leacock, R. Teng, y R. Bunker. 1993. A Semantic Concordance. En *Proceedings of the ARPA Workshop on Human Language Technology*.
- Niles, I. y A. Pease. 2001. Towards a standard upper ontology. En *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*, páginas 17–19. Chris Welty and Barry Smith, eds.
- Padró, Lluís. 1998. *A Hybrid Environment for Syntax-Semantic Tagging*. Ph.D. thesis, Departament de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya. Barcelona.
- Rudova, Hana. 2001. *Constraint Satisfaction with Preferences*. Ph.D. thesis, Faculty of Informatics, Masaryk University,.
- Shi, Lei y Rada Mihalcea. 2005. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. En *CICLING'05*, Mexico.
- Vossen, P., editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers .