Integrando una Base de Datos Léxica y una Colección de Entrenamiento para la Desambiguación del Sentido de las Palabras

L. Alfonso Ureña López
Departamento de Informática.
Universidad de Jaén
Avda. Madrid 35, 23071 Jaén. Spain
e-mail: laurena@ujaen.es

Manuel García Vega
Departamento de Informática.
Universidad de Jaén
Avda. Madrid 35, 23071 Jaén. Spain
e-mail: mgarcia@ujaen.es

José Mª Gómez Hidalgo
Departamento de Inteligencia Artificial
Universidad Europea de Madrid
e-mail: jmgomez@dinar.uem.es

Alberto Díaz Esteban
Departamento de Ingeniería del Software
Universidad Europea de Madrid
e-mail: alberto@drpis.esi.uem.es

Resumen

La resolución de la ambigüedad es una tarea compleja y útil para muchas aplicaciones del procesamiento del lenguaje natural. En concreto, la ambigüedad causa problemas en aplicaciones como: la *Recuperación de Información (IR)*, donde los problemas pueden ser substanciales y ser superados si se utilizan grandes consultas, y *la traducción automática*, donde es un gran problema inherente. Recientemente han sido varios los enfoques y algoritmos propuestos para realizar esta tarea.

Presentamos un nuevo enfoque basado en la integración de varios recursos lingüísticos de dominio público, como una base de datos léxica y una colección de entrenamiento. Nuestro enfoque integra la información de sinonimia de WordNet y la colección de entrenamiento SemCor para incrementar la efectividad de la desambiguación, a través del Modelo del Espacio Vectorial. Hemos probado nuestro enfoque sobre un gran conjunto de documentos con una fina granularidad de sentidos, como son los de WordNet, consiguiendo una alta precisión en la resolución de la ambigüedad léxica.

Palabras clave: Word Sense Disambiguation (WSD), Text Categorization (TC), WordNet, SemCor, Information Retrieval (IR), Machine Translation (MT), Ventana Contextual (VC).

1. Introducción

Un problema importante del Procesamiento del Lenguaje Natural (PLN) es determinar el significado de una palabra en un contexto particular. Las diferentes acepciones de una palabra son recogidas como varios sentidos en un diccionario. La tarea de desambiguación del sentido de las palabras (WSD) es identificar el sentido correcto de una palabra en un contexto. Esta tarea es compleja, pero muy útil en variadas aplicaciones del procesamiento en lenguaje natural [Kilgarriff 97a], como Categorización de Texto (TC) [Buenaga 97]; traducción automática [Brown 91]; restauración de acentos [Yarowsky 94]; encaminamiento y filtrado de textos; agrupamiento y segmentación de textos y, en general, en la recuperación de información [Kilgarriff 97a, 97b, Sanderson 96].

En este artículo presentamos un nuevo enfoque automático para WSD basado en el uso de varios recursos lingüísticos. Actualmente, muchos recursos, como colecciones de entrenamiento [Yarowsky 92, Yoshiki 94, Ureña 97] y bases de datos léxicas [Resnik 95, Agirre 96, Xiaobin 95] o thesaurus [Yarowsky 92], han sido satisfactoriamente empleadas para la

resolución de la ambigüedad léxica de manera aislada. Creemos que la idea clave para la mejora de la desambiguación es incrementar la cantidad de información de la que hace uso el sistema, a

Hemos elegido para nuestro enfoque el Modelo del Espacio Vectorial para la Recuperación de Información [Lesk 86], utilizando Ventana Contextual de tamaño variable [Ureña 98]. Los vectores de pesos son calculados para cada ventana contextual, empleando la base de datos léxica WordNet y el subconjunto de entrenamiento SemCor. Hemos calculado el vector de pesos para:

- un enfoque basado solamente en WordNet,
- un enfoque basado sólo en la Colección de entrenamiento,
- y, un enfoque combinado, basado en la integración de los recursos WordNet y SemCor.

Comparamos la similitud término-sentido, eligiendo el significado de mayor similitud, estudiando el ángulo que forman los vectores. Hemos realizado una serie de experimentos sobre un conjunto de prueba de la colección de evaluación SemCor, mostrando que un enfoque combinado puede mejorar la efectividad de la desambiguación.

Este trabajo está organizado como sigue. En primer lugar, introducimos la tarea de desambiguación y los recursos utilizados. Seguidamente, describimos el modelo en el que estos elementos son integrados, y examinamos ambos enfoques y la integración de los dos recursos. Después de esto, presentamos nuestra evaluación y los resultados obtenidos, y finalmente, describimos nuestras conclusiones y líneas futuras.

2. *Descripción de la Tarea*

Dado un conjunto de documentos, el objetivo de un sistema desambiguador es decidir el sentido correcto de los nombres y verbos que componen los documentos. El sistema hace uso de la información contenida en los textos para computar el grado de pertenencia del término a cada sentido.

través de varios recursos lingüísticos. Planteamos un modelo integrador para WSD, empleando de manera combinada corpus de entrenamiento y bases de datos léxicas.

El recurso más ampliamente utilizado para WSD es la colección de entrenamiento. Una colección de entrenamiento es un conjunto de documentos con los sentidos etiquetados manualmente, que permite al sistema asignar los sentidos a nuevos documentos, de acuerdo con su similitud a otros documentos de la colección de entrenamiento. Actualmente son varios los corpus de los que pueden ser obtenidos un conjunto de entrenamiento y otro de prueba. Hemos seleccionado SemCor por su amplia utilización y disponibilidad, lo que facilita la comparación de resultados.

Una base de datos léxica es un sistema con información léxica de uno o varios lenguajes. Desde este punto de vista, los diccionarios electrónicos pueden ser considerados como bases de datos léxicas. Actualmente, las bases de datos léxicas incluyen WordNet, EDR y Roget's Thesaurus. Hemos elegido WordNet dada su libre distribución, amplia cobertura y frecuencia de uso. Proponemos la integración de bases de datos léxicas y colecciones de entrenamiento para mejorar la efectividad del proceso WSD.

3. *Integración de recursos en el Modelo del Espacio Vectorial*

El Modelo del Espacio Vectorial (MEV) [Salton 83] fue originalmente desarrollado para la Recuperación de Información, pero provee un soporte muy adecuado para realizar otras tareas como WSD o TC. . También, el modelo está avalado por muchas experiencias en recuperación de texto [Lewis 92,Salton 89]. De hecho, el MEV es un entorno muy adecuado para expresar nuestro enfoque de WSD, pues permite la integración de múltiples fuentes de conocimiento para la desambiguación, y hace más fácil identificar el papel de cada fuente de conocimiento involucrada en la operación de desambiguación.

3.1 El Modelo del Espacio Vectorial para la Desambiguación de Términos

La clave del MEV para la Recuperación de Información (IR) es representar las expresiones del lenguaje natural mediante vectores de pesos. Cada peso representa la importancia de un término, en relación con un determinado sentido en la expresión del lenguaje natural. Una hipótesis fundamental en WSD es que cada palabra se utiliza con un único significado en un contexto concreto [Yarowsky 93]. Cada término s_{ji} queda representado o indexado por un vector de dimensión m, con los pesos asignados a cada uno de los términos de indexación. El término i con sentido j, queda representado con el peso del término, así como con los pesos de los términos circundantes.

$$s_{ji}=\langle ws_{j1}, ws_{k1}, ..., ws_{kn}\rangle$$

ws_{kc} peso de la palabra circundante c al término s_{ji}

Para el procesamiento de los textos a desambiguar, se obtienen los términos de indexación aparecidos en ellos, de una forma análoga al de los textos de la colección de entrenamiento. La representación de una consulta de un término c_k, se realiza mediante un vector de pesos asociados a los términos.

$$c_k=\langle wc_1, wck_1, ..., wc_{kn}\rangle$$

wc_{kc} peso de la palabra circundante c al término c_k

La similitud semántica entre el término i con sentido j y el término viene dada por el coseno del ángulo que forman sus vectores, con arreglo a la formula:

$$sim(s_{ji}, c_l) = \frac{\sum_{i=1}^{m} ws_{ji} \cdot wc_l}{\sqrt{\sum_{i=1}^{m} ws_{ji}^2 \cdot \sum_{i=1}^{m} wc_l^2}}$$

Los pesos de los términos pueden ser computados haciendo uso de la bien conocida fórmula basada en términos de frecuencias. Calculamos los pesos para los distintos términos de manera análoga a [Salton83]:

$$ws_{ji} = t_{ji} * \log_2(n/f_i)$$

Donde t_{ji} es la frecuencia del término j con sentido i en la ventana contextual, n es el número de sentidos de término i y f_i es el número de ventanas contextuales donde aparece el término i.

3.2 Enfoque Basado en la Colección de Entrenamiento SemCor

Un conjunto de documentos manualmente etiquetados con el sentido correcto de cada uno de los términos puede utilizarse para predecir el sentido de los términos que componen los nuevos documentos. Hemos utilizado SemCor por ser uno de los pocos corpus etiquetados de dominio público. SemCor es un subconjunto del Brown Corpus etiquetado con los sentidos de WordNet. Sin embargo, SemCor no es banco de pruebas óptimo para la resolución de la ambigüedad léxica, debido fundamentalmente a la fina granularidad de sentidos que utiliza [Padró 97].

La hipótesis clave cuando utilizamos una colección de entrenamiento para la resolución de la ambigüedad léxica es que un término aparece con un particular sentido en un determinado contexto. Los términos que constituyen ese contexto pueden ser buenos para predecir el sentido con que aparece el término. El conjunto de predictores del significado de un término, y su importancia, son computados estadísticamente por las ventanas contextuales [Ureña 98]), como un paso inicial del proceso de entrenamiento. Para ello, se representa cada término del corpus de entrenamiento con un vector, cuyas componentes son: el peso del término en el párrafo y los pesos de los términos que constituyen la ventana contextual. Así, para cada uno de los nombres de la colección de entrenamiento, calcularemos su ventana contextual. El programa desplaza la ventana desde el principio de todos los documentos que contiene el corpus de entrenamiento, hasta el final de cada uno, considerando un término en cada desplazamiento, y como palabras de contexto, cada una de las palabras circundantes. De esta manera, se construyen tantas ventanas

contextuales como términos con diferentes sentidos existan en la colección.

3.3 Enfoque Basado en WordNet

Las bases de datos léxicas contienen muchos tipos de información (conceptos, sinónimos y otras relaciones léxicas, hiponimia y otras correspondencias conceptuales, etc.). WordNet [Miller 95] es un lexicón que representa conceptos como conjuntos de sinónimos, o synsets (elementos básicos de WordNet).

El sentido no es un concepto bien definido, ofreciendo frecuentemente finas distinciones dependiendo de la colocación, contexto, etc. Para nuestro propósito consideramos los sentidos de las palabras presentes en WordNet. Y, seleccionamos la información de sinonimia como "categorías de sentidos", de esta manera, un término es consultado en WordNet, donde se obtiene la información de sus synsets o conceptos asociados.

Cada synset se trata como un sentido distinto para cada término, del que se obtiene un conjunto de palabras sinónimas para cada término. Se construyen ventanas contextuales con cada synset para cada uno de los términos. De esta manera, un término, en un determinado contexto, puede desambiguarse calculando la similitud entre dicho término y las ventanas contextuales asociadas a cada uno de sus synsets en WordNet.

3.4 Integración de SemCor y WordNet

Al incorporar además información proveniente de WordNet a la colección de entrenamiento, la efectividad de la desambiguación debe mejorar.

La integración se ha realizado, como sigue. En la fase de entrenamiento se construyen primeramente los vectores conforme a lo relatado en el enfoque basado en el entrenamiento, obteniendo un conjunto de vectores, uno por cada uno de los términos que constituyen la colección de entrenamiento. A continuación, cada uno de los términos, que representan a los vectores, se consulta en WordNet, si dicho término tiene un synset asociado para el sentido consultado se "une" dicho synset al vector, recalculándolo con mayor peso, en caso contrario se elimina dicho synset. La fase de prueba se realiza confrontando la consulta, por medio del cálculo de la similitud, con los vectores creados en el entrenamiento.

Esta técnica de integración identifica claramente el papel de cada recurso en este enfoque de desambiguación. Por un lado WordNet proporciona información relativa a la relación de sinonimia, ampliando el número de términos en relación con un determinado sentido, cuando los datos de entrenamiento no son grandes o no son seguros. Esto directamente contribuye con los términos usados en la representación del vector. Por otro, la colección de entrenamiento proporciona mayor información contextual para aquellos términos mejor entrenados.

4. Evaluación

La evaluación de la tarea WSD es muy heterogénea. Se han utilizado varias métricas y colecciones de prueba en distintos trabajos con variados enfoques. Esto ha producido un problema en lo que se refiere a la comparación de resultados entre distintos enfoques. Para minimizar este problema, hemos seleccionado para nuestro trabajo, un conjunto de métricas muy extendidas y frecuentemente utilizadas en el campo de la evaluación de los sistemas de recuperación de información [Salton 83] [Frakes 92].

4.1 Métricas de evaluación

Hemos utilizado la precisión como métrica básica para computar la efectividad de nuestros experimentos. El cálculo puede ser realizado utilizando macroaveraging y microaveraging [Lewis92].

La Precisión puede ser definida como el cociente entre el número de términos desambiguados satisfactoriamente y el número de términos desambiguados.

La precisión macroaveraging consiste en calcular la precisión para cada uno de los términos, y luego calcular la media para cada uno de ellos, como sigue:

$$P_{macroavg} = \frac{\sum_{i=1}^{n} P_i}{n} \qquad P_i = \text{Precisión del término } i$$

$$P_i = \frac{dc_i}{dc_i + di_i}$$

Donde dc_i es el número de desambiguaciones correctas del término i, di_i el número de desambiguaciones incorrectas del término i y n el número de términos desambiguados. Por otro lado, la precisión microaveraging consiste en calcular un sólo valor de precisión medio para todos los términos, según:

$$P_{microavg} = \frac{tdc}{tdc + tdi}$$

Siendo tdc el número de términos desambiguados correctamente y tdi el número de términos desambiguados incorrectamente.

```
<contextfile concordance=brown>
<context filename=br-a01 paras=yes>
<p pnum=1>
<s snum=1>
<wf cmd=ignore pos=DT>The</wf>
<wf cmd=done rdf=group pos=NNP lemma=group wnsn=1
lexsn=1:03:00:: pn=group>Fulton_County_Grand_Jury</wf>
<wf cmd=done pos=VB lemma=say wnsn=1
lexsn=2:32:00::>said</wf>
<wf cmd=done pos=NN lemma=friday wnsn=1
lexsn=1:28:00::>Friday</wf>
<wf cmd=ignore pos=DT>an</wf>
<wf cmd=done pos=NN lemma=investigation wnsn=1
lexsn=1:09:00::>investigation</wf>
....
....
<wf cmd=ignore pos=DT>any</wf>
<wf cmd=done pos=NN lemma=irregularity wnsn=1
lexsn=1:04:00::>irregularities</wf>
<wf cmd=done pos=VB lemma=take_place wnsn=1
lexsn=2:30:00::>took_place</wf>
<punc>.</punc>
</s>
</p>
```

Ilustración 1 Fragmento de un documento de SemCor

4.2 Colección de prueba

SemCor [Miller93] consta de un total de 103 ficheros de texto en SemCor, constituyendo un total de 11.628 frases y un total de palabras de 229.370. Además de un corpus de texto, SemCor es un lexicón, donde cada palabra en el texto hace referencia a su correcto significado en él. Puede definirse, bien como un corpus, en el que las palabras han sido etiquetadas sintáctica y semánticamente, o como un lexicón, en el que las frases de ejemplo pueden ser encontradas por varias definiciones. SemCor abarca el Brown Corpus donde sólo los nombres, verbos, adjetivos y adverbios son etiquetados semánticamente con los sentidos de WordNet. Las palabras (tales como preposiciones, determinantes, pronombres, verbos auxiliares, etc.) y caracteres no alfanuméricos, interjecciones y términos coloquiales no son etiquetados.

Hemos tomado como colección de prueba, un subconjunto de ficheros de SemCor seleccionados aleatoriamente. Un ejemplo de un fragmento de un fichero de SemCor se muestra en la Ilustración 1.

4.3 Resultados e interpretación

Para nuestros experimentos hemos seleccionado aleatoriamente cuatro documentos de SemCor considerados individualmente: br-a14, br-j09, br-k11 y br-k14. Estos textos (sin etiquetas) han representado el papel de ficheros de entrada.

Precisión	*Recursos Lingüísticos*		
	WordNet	*SemCor*	*WordNet+SemCor*
Microaveraging	57.7%	78.6%	81.3%
Macroaveraging	55.4%	83%	85.1%

Tabla 1 Resultados de nuestros experimentos

Los resultados para nuestra primera serie de experimentos son resumidos en la tabla 1. Esta tabla muestra las medias micro y macroaveraging para la precisión. Los valores obtenidos por el enfoque integrado muestran una apreciable ventaja sobre el enfoque basado en WordNet y el basado en SemCor.

5. *Conclusiones y Futuros Trabajos*

En este trabajo hemos presentado un nuevo enfoque para la desambiguación del significado de las palabras basado en la integración de varios recursos léxicos de libre distribución para mejorar la efectividad, como son los diccionarios electrónicos y los córporas de entrenamiento, empleados hasta ahora de manera aislada. Este enfoque integra la información proporcionada por la base de datos léxica WordNet, utilizando un algoritmo de entrenamiento fundamentado en el modelo del espacio vectorial. La técnica se basa en la mejora de la representación de los sentidos a través del uso de la base de datos léxica.

Hemos probado nuestro enfoque con la colección de prueba SemCor, obteniendo medidas que avalan, que utilizando el enfoque combinado, la precisión es mayor que la obtenida, tanto por el enfoque basado en el entrenamiento de SemCor, como por el basado en WordNet. A pesar de la complejidad de la tarea, se ha obtenido una buena precisión teniendo en cuenta la fina granularidad de sentidos utilizada, como consecuencia de la colección de entrenamiento SemCor y la base de datos léxica WordNet.

Actualmente, estamos realizando nuevos experimentos para la integración de mayor información proporcionada por WordNet (como son las relaciones de hiperonimia, hiponimia y meronimia) con enfoques basados en entrenamiento, todo ello unido al desarrollo de métodos más sofisticados para la desambiguación. Asimismo, extenderemos la idea utilizando múltiples recursos léxicos a otras tareas del procesamiento del lenguaje natural, donde la desambiguación puede ser muy útil.

Referencias

[Agirre 96] Agirre E., Rigau G. *Word sense disambiguation using conceptual density.* In Proceedings of COLING 1996.

[Buenaga 97] Buenaga Rodríguez M., Gómez Hidalgo J .M., Díaz Agudo B. *Using WordNet to Complement Training Information in Text Categorization.* Second International Conference on Recent Advances in Natural Language Processing, 1997

[Broown 91] Brown P. B., Pietra S. A., Pietra V. J.*Word Sense Disambiguation Using Statiscal Methods.* In Proc. Of ACL, pp. 264-270, 1991.

[Frakes 92] Frakes, W., Baeza, R., *Information retrieval: data structures and algorithms,* Prentice Hall, London. 1992.

[Kilgarriff 97a] Kilgarriff A *What is word sense disambiguation good for?* Proc. Natural Language Processing Pacific Rim Symposium. Phuket, Thailand. December 1997. pp 209-214.

[Kilgarriff 97b] Kilgarriff A. *Foreground and Background Lexicons and Word Sense Disambiguation for Information Extraction* Proc. International Workshop on Lexically Driven Information Extraction. Frascati, Italy. July 1997. pp 51-62.

[Lesk 86] Lesk, M. *Automatic sense disambiguation: how to tell a pine cone from an ice cream cone.* Proc. of the SIGDOC Conference 1986.

[Lewis 92] Lewis, D., *Representation and learning in information retrieval.* Ph.D. Thesis, Department of Computer and Information Science, University of Massachusetts. 1992.

[Miller 93] Miller G. Leacock C., Randee T. and Bunker R. *A Semantic concordance. In Proceedings of the 3rd DARPA Workshop on Human Language Technology,* New Jersey 1993.

[Miller 95] Miller G. *WordNet: lexical database.* Communications of the ACM Vol 38, No. 11.

[Padró 97] *A Hybrid Environment for Syntax-Semantic Tagging* PhD Thesis, Departament de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya. Barcelona, 1997.

[Resnik 95] Resnik P. *Disambiguating Noun Groupings with Respect to WordNet Senses* Proceedings of the 3rd Workshop on Very Large Corpora, MIT, 30 June 1995

[Salton 83] Salton G., McGill, M.J.*Introduction to modern information retrieval.* McGraw-Hill.1983.

[Salton 89] Salton, G., *Automatic Text Processing: the transformation, analysis and retrieval of information by computer.* Addison Wesley. 1989..

[Sanderson 96] Sanderson, M., *Word sense disambiguation and information retrieval.* Ph.D. Thesis, Department of Computing Science, University of University of Glasgow. 1996.

[Ureña 97] Ureña López, L. A., García Vega, M., Buenaga Rodríguez, M., Gómez Hidalgo, J. M. *Resolución de la ambigüedad léxica mediante información contextual y el modelo del espacio vectorial.* Séptima Conferencia de la Asociación Española para la Inteligencia Artificial. CAEPIA. 1997.

[Ureña 98] Ureña López, L. A., García Vega, M., Buenaga Rodríguez, M., Gómez Hidalgo, J. M. *Resolución automática de la ambigüedad léxica fundamentada en el modelo del espacio vectorial usando ventana contextual variable.* Asociación Española de Lingüística Aplaicada. AESLA. 1998.

[Xiaobin 95] Xiaobin Li; Szpakowicz S.; Matwin S. *A WordNet-based algorithm for word sense disambiguation.* Proceedings of the Fourteenth International Joint conference on Artificial Intelligence. pp. 1368-74, vol 2. 1995

[Yarowsky 92] Yarowsky D. *Word-sense disambiguation using statistical models of Roget's categories trained on large corpora.* In Proceedings of the 15th International Conference onr Computational Linguistics. 1992.

[Yarowsky 93] Yarowsky, D. *One Sense Per Collocation.* In *Proceedings, ARPA Human Language Technology Workshop.* Princeton, pp. 266-271, 1993.

[Yarowsky 94] Yarowsky D. *Decision lists for lexical ambiguity resolution: Aplication to accent restoration in Spanish and French.* In Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics, (ACL'94).1994.

[Yarowsky 95] Yarowsky D. *Unsupervised word sense disambiguation rivaling supervised methods.* In Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics, (ACL'95).1995.

[Yoshiki 94] Yoshiky N., Yoshihiko Nitta *Co-ocurrence vectors from corpora vs. distance vectors from.* In Proceedings of COLING94.