Desambiguación Léxica de Sustantivos usando la Web

Aarón Pancardo-Rodríguez[1], Manuel Montes-y-Gómez[1,2], Paolo Rosso[2],
Davide Bucaldi[3], Luis Villaseñor-Pineda[1]

[1] Laboratorio de Tecnologías del Lenguaje
Instituto Nacional de Astrofísica, Óptica y Electrónica, México
{aaron_cyberman, mmontesg, villasen}@inaoep.mx

[2] Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, España
{mmontes, prosso@dsic.upv.es}

[3] Dipartimento di Informatica e Scienze dell'Informazione,
Università di Genova, Italia
buscaldi@disi.unige.it

Resumen. La selección del sentido más apropiado de una palabra ambigua en una oración es uno de los problemas centrales del procesamiento del lenguaje natural. En este artículo se presenta un método que permite desambiguar el sentido de un sustantivo usando la información conceptual de su contexto. La selección del sentido más frecuente se realiza considerando exclusivamente las estadísticas de co-ocurrencia en la Web de los sustantivos del contexto con los sinónimos e hiperónimos del sustantivo en cuestión. Los resultados obtenidos en esta primera aproximación son aún muy modestos, 28% de precisión en la selección del mejor sentido, sin embargo confirman el potencial de la Web como fuente de información para la desambiguación del sentido de las palabras, y orientan su uso hacia la discriminación de los sentidos menos probables.

1. Introducción

El objetivo de la desambiguación del sentido de las palabras (WSD, por sus siglas en inglés) es identificar el sentido correcto de una palabra en un contexto particular. Los trabajos más recientes [6] muestran que el paradigma más eficiente para esta tarea es el de aprendizaje supervisado. Sin embargo, debido a la escasez de corpora etiquetados y la dificultad para crearlos manualmente, se considera que estos métodos ya han alcanzado sus más altos rendimientos.

Intentando evitar o reducir el problema de adquisición de conocimiento al que se enfrentan los métodos supervisados, los métodos no supervisados no necesitan ningún proceso de aprendizaje y se basan sólo en el conocimiento léxico proporcionado por algún recurso externo (p.e., Wordnet). Entre éstos destacan los métodos basados en densidad conceptual [1,8] y marcas de especificidad [7]. Asimismo se han propuesto algunos métodos para la adquisición automática de corpus etiquetado a partir de la Web [2,5].

En este artículo se presenta un método no supervisado y basado en la Web para la desambiguación léxica de sustantivos. A diferencia de otros métodos que usan exclusivamente la Web para extraer más ejemplos de entrenamiento, este método aprovecha la redundancia en la Web para deducir directamente las probabilidades

asociadas a cada uno de los sentidos de un sustantivo de acuerdo con su contexto. Es importante señalar que esta aproximación fue motivada por algunos trabajos recientes en la búsqueda de respuestas en la Web [3,4].

Por otra parte, el método propuesto se fundamenta en la hipótesis de que los documentos mantienen una *cohesión temática*, que se refleja en una fuerte relación semántica entre los conceptos que mencionan. Así pues, nuestro método considera que el sustantivo que se desea desambiguar y los sustantivos de su contexto deben tener una relación semántica clara, y que dicha relación debe manifestarse en la Web en una alta co-ocurrencia de los sustantivos del contexto con los sinónimos e hiperónimos del sustantivo en cuestión.

El resto del artículo se organiza de la siguiente manera. En la sección 2 se describe el método de desambiguación léxica de sustantivos basado en la Web y soportado en la hipótesis de la cohesión temática. En la sección 3 se muestran los resultados preliminares obtenidos al aplicar este método en la desambiguación de todos los sustantivos del corpus "Senseval-3 English all words". Finalmente en la sección 3 se discuten los resultados obtenidos y se presentan algunas propuestas de trabajo futuro.

2. Descripción del método

Dado un sustantivo w, con $|w|$ sentidos, inmerso en el contexto C formado por los sustantivos presentes en su contexto inmediato (en la misma oración), la función $\Gamma(w_k,C)$ indica la cohesión temática entre el sentido w_k del sustantivo y su contexto.

La estimación de $\Gamma(w_k,C)$ en la Web considera los n sinónimos de w_k definidos en Wordnet 2.0 $\{s_{ik},\ 0< i \leq n\}$, así como sus m hiperónimos $\{h_{jk},\ 0 < j \leq m\}$. La co-ocurrencia de estos elementos con el contexto se calcula mediante la función $f_s(x,y)$. Esta función retorna el número de páginas en la Web que contienen el patrón x AND y de acuerdo con el motor de búsqueda S. Asimismo, $f_s(x)$ es una función que regresa el número de páginas web que contienen la cadena x usando el buscador S.

Si se asume que $\Gamma(w_k,C) \approx P_{web}(w_k|C)$, es decir, que la cohesión temática entre el sentido w_k del sustantivo en cuestión y su contexto es aproximadamente proporcional a la probabilidad de, dada una página web que contiene los sustantivos pertenecientes a C, encontrar el sustantivo w con el sentido k, entonces la relación temática entre w_k y C puede calcularse de las siguientes dos maneras:

$$\Gamma(w_k,C) = \frac{1}{n+m}\left(\sum_{i=1}^{n} P(s_{ik}|C) + \sum_{j=1}^{m} P(h_{jk}|C)\right) \quad (1)$$

$$\Gamma(w_k,C) = \arg\max_{\substack{0<i\leq n \\ 0<j\leq m}} \left(P(s_{ik}|C), P(h_{jk}|C)\right) \quad (2)$$

donde, $P(s_{ik}|C) = f_s(C,s_{ik}) / f_s(C)$, y $P(h_{ik}|C) = f_s(C,h_{ik}) / f_s(\mathrm{C})$.

Asimismo, si suponemos que $\Gamma(w_k,C) \approx P_{web}(C|w_k)$, es decir, que la cohesión temática entre el sentido w_k del sustantivo en cuestión y su contexto es aproximadamente proporcional a la probabilidad de, dada una página web que contiene el sustantivo w con su sentido k, encontrar todos y cada uno de los conceptos de su contexto, entonces la relación temática entre w_k y C puede calcularse de las siguientes dos maneras:

$$\Gamma(w_k, C) = \frac{1}{n+m}\left(\sum_{i=1}^{n} P(C|s_{ik}) + \sum_{j=1}^{m} P(C|h_{jk})\right) \quad (3)$$

$$\Gamma(w_k, C) = \arg\max_{\substack{0<i\le n \\ 0<j\le m}} \left(P(C|s_{ik}), P(C|h_{jk})\right) \quad (4)$$

donde, $P(C|s_{ik}) = f_s(C,s_{ik}) / f_s(s_{ik})$, y $P(C|h_{ik}) = f_s(C,h_{ik}) / f_s(h_{jk})$.

Las fórmulas 1 y 3 se basan en el promedio de las probabilidades. Ellas suponen que todos los sinónimos e hiperónimos del sentido w_k deben tener una relación con el contexto C para lograr distinguir una cohesión temática entre ellos. Por el contrario, las fórmulas 2 y 4 se basan en el máximo de las probabilidades, indicando que es suficiente que alguno de los sinónimos o hiperónimos de w_k tenga relación con C para establecer su cohesión temática.

El algoritmo para desambiguar un sustantivo considerando la cohesión temática con su contexto se detalla a continuación:

1. Seleccionar el conjunto C de sustantivos alrededor de w usando la oración como limite del contexto.
2. Para cada sentido w_k de w, y para cada sinónimo s_{ik} e hiperónimo h_{jk} de w_k, calcular $f_s(C,s_{ik}), f_s(C,h_{ik}), f_s(s_{ik})$ y $f_s(h_{jk})$.
3. Asignar a cada sentido w_k un peso dependiendo de la función $\Gamma(w_k,C)$, la cual combina los resultados obtenidos en el paso anterior.
4. Seleccionar el sentido (o sentidos) con los pesos mayores.

3. Resultados experimentales

La tabla 1 muestra los resultados obtenidos en la desambiguación del sentido de 215 sustantivos[1] seleccionados del corpus "Senseval-3 English all words" [9]. En la evaluación se usaron tres motores de búsqueda diferentes: Google, Altavista y MSN. Para cada una de ellos la tabla 1 muestra la precisión obtenida en la selección de la mejor respuesta, las dos mejores respuestas, y la mitad de los sentidos más probables.

Algunas observaciones interesantes de estos resultados son:

Por una parte, el desempeño alcanzado por las distintas máquinas de búsqueda es similar. La diferencia promedio en la precisión fue de 0.03. Además, el análisis detallado de los resultados señaló que los aciertos de los distintos buscadores tienen un alto traslape, lo que implica que ninguna combinación de éstos presentará resultados considerablemente superiores.

Por otra parte, los resultados obtenidos bajo la consideración $\Gamma(w_k,C) \approx P_{web}(C|w_k)$, fórmulas 3 y 4, fueron los mejores en todos los casos. Este resultado es importante pues implica que la normalización respecto a $f_s(s_{ik})$ y $f_s(h_{ik})$ hace menos sensible el cálculo de las probabilidades a la presencia de sinónimos e hiperónimos muy generales o raros.

[1] Este conjunto representa todas las palabras del corpus "Senseval-3 English all words" que fueron marcados como sustantivos después de aplicar un proceso de etiquetado de partes de la oración.

Tabla 1. Desambiguación léxica usando la Web

Tipo de evaluación	Fórmula	Google	Altavista	MSN
1 sentido	1	0.227	0.186	0.181
	2	***0.279***	0.186	0.19
	3	0.237	0.215	0.209
	4	0.251	0.215	0.218
2 sentidos	1	0.348	0.353	0.362
	2	0.395	0.339	0.358
	3	0.432	0.386	0.4
	4	0.451	0.427	***0.469***
⌈n/2⌉ sentidos	1	0.595	0.534	0.548
	2	0.6	0.576	0.609
	3	0.641	0.613	0.637
	4	0.646	0.623	***0.679***

Finalmente, los resultados obtenidos a partir de la máxima probabilidad encontrada entre el contexto y uno de los sinónimos o hiperónimos del sustantivo en cuestión, formulas 2 y 4, fueron los mejores en la mayoría de los casos. Igual que en el caso anterior, creemos que este resultado se debe a que el máximo es menos sensible que el promedio a la presencia de sinónimos e hiperónimos muy generales o raros.

4. Conclusiones y trabajo futuro

En este artículo se propuso un método para la desambiguación léxica de sustantivos basado en la web y en la hipótesis de la cohesión temática de los documentos. Con este método la selección del sentido más frecuente se realiza a partir de las estadísticas de co-ocurrencia en la Web del contexto y los sinónimos e hiperónimos del sustantivo en cuestión. El enfoque es novedoso, pues la mayoría de los métodos de desambiguación léxica que emplean la Web lo hacen para recolectar corpora y no para el cálculo directo de las probabilidades de los sentidos.

Los resultados obtenidos en nuestro primer experimento son aún muy modestos, 28% de precisión en la selección del mejor sentido y 68% en la selección de la mitad de los sentidos más probables. Sin embargo, estos resultados confirman el potencial de la Web como fuente de información para la desambiguación del sentido de las palabras, y orientan su uso hacia la llamada "desambiguación léxica difusa".

Los experimentos permitieron observar lo siguiente. En primer lugar, el método falló principalmente en la desambiguación de las palabras más polisemicas, con más de 5 sentidos en promedio. En segundo lugar, en la mayoría de los aciertos, independientemente de la máquina de búsqueda usada, la probabilidad del sentido correcto fue considerablemente mayor que las probabilidades del resto de los sentidos. Ambas observaciones sugieren la integración de los métodos basados en la Web con otros métodos de desambiguación léxica.

A manera de trabajo futuro se propone lo siguiente: (i) considerar los hipónimos de cada sentido en lugar de los hiperónimos; (ii) medir la cohesión temática entre sentido y contexto considerando su co-ocurrencia en la misma oración y no en un mismo

documento. Para ello será necesario redefinir las consultas a la Web, o en su defecto usar alguna otra colección grande de documentos a manera de corpus (e.g. el corpus BNC). (iii) utilizar más información contextual, es decir, mayor cantidad de sustantivos del contexto así como palabras con otras partes de la oración. Consideramos que esta medida permitirá incrementar la precisión de la desambiguación de los sustantivos altamente polisemicos; (iv) definir un esquema para medir el grado de confianza de los resultados de la web. Con ello pretendemos facilitar la combinación de este enfoque con algún otro método tanto supervisado como no supervisado.

Agradecimientos

Los autores agradecen a CONACYT (43990A-1, U39957-Y), R2D2 CICYT (TIC2003-07158-C04-03), ICT EU-India (ALA/95/23/2003/077-054), así como a la Secretaría de Estado de Educación y Universidades de España por el soporte financiero otorgado.

Referencias

[1] E. Agirre and G. Rigau (1995). A proposal forWord Sense Disambiguation using Conceptual Distance. Proceedings of the International Conference on Recent Advances in NLP, RANLP´95. 1995.

[2] E. Agirre and D. Martinez (2000). Exploring automatic word sense disambiguation with decision lists and the web. In Procedings of the COLING 2000.

[3] E. Brill , J. Lin, M. Banko, S. Dumais, A. Ng (2001). *Data-intensive question answering*. In Proceedings of the Tenth Text REtrieval Conference (TREC 2001).

[4] A. Del-Castillo, M. Montes-y-Gómez, L. Villaseñor-Pineda (2004). QA on the Web: A Preliminary Study for Spanish Language. International Conference on Computer Science. ENC 2004. Colima, México, Septiembre, 2004.

[5] R. Mihalcea and I. Moldovan (1999). An Automatic Method for Generating Sense Tagged Corpora. Proceedings of the 16th National Conference on Artificial Intelligence. AAAI Press, 1999.

[6] R. Mihalcea and P. Edmonds (2004). Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. Barcelona, Spain, July, 2004.

[7] A. Montoyo (2000). Método basado en Marcas de Especificidad para WSD. Procesamiento del Lenguaje Natural, Revista nº 26, Septiembre, 2000.

[8] P. Rosso, F. Masulli, D. Buscaldi, F. Pla, A. Molina (2003). Automatic Noun Disambiguation. Lecture Notes in Computer Science, Vol. 2588, Springer Verlag, February 2003.

[9] B. Zinder and M. Palmer (2004). The English All Words Task. Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3:). Barcelona, Spain, July, 2004.