Identificación de Antónimos en Textos Planos

Cupertino Lucero, David Pinto y Héctor Jimenez-Salazar

Facultad de Ciencias de la Computación
Benemérita Universidad Autónoma de Puebla
14 sur y Av. San Claudio. Edif. 135. Ciudad Universitaria
Puebla, Pue. 72570. México
Tel. (+52222) 229 55 00 ext. 7212 Fax (+52222) 229 56 72
QPrr@hotmail.com, dpinto@cs.buap.mx, hjimenez@fcfm.buap.mx

Abstract. WordNet has been used in several applications of Natural Language Processing field, however, the fact that it is not a specialized lexical database (LDB) led us to construct LDB's for specific domains. Thereby it is important to develop automatic methods for the recognition of lexical relationships. In this paper we present a new method for identifying word pairs (extracted from a raw corpus) in oppositeness relationship. The method was applied to a set of related word pairs obtained from a corpus by using Grefenstette's method. Also, we used features extracted from word contexts. These features were evaluated by a word-distance between pairs that appears in the same context, lexical-syntactical patterns that matches regular expressions with word contexts, and by lexical co-ocurrence networks built for each word of a pair. The method was tested with a set of word pairs taken from an economy corpus obtaining a 80-percent accuracy.
Key Words: Oppositeness relationship, lexical co-ocurrence network.

Resumen. WordNet se ha usado en varias aplicaciones y en muchos campos del procesamiento del lenguaje natural, sin embargo, el hecho de que no sea una base de datos léxica (BDL) especializada ha conducido a la construcción de BDLs para dominios específicos. Por ello, es muy importante desarrollar métodos automáticos para el reconocimiento de relaciones léxicas. En este trabajo se presenta un nuevo método para identificar pares de palabras en relación de oposición provenientes de un corpus sin formato. El método fue aplicado a pares de palabras relacionadas que fueron obtenidas de un thesaurus creado por el método de Grefenstette. Para lograr esto, también se usan algunos rasgos extraídos de los contextos de las palabras. Estos rasgos fueron evaluados a partir de la distancia entre las palabras que aparecen en el mismo contexto, por patrones léxico-sintácticos usados para empatar expresiones regulares en los contextos de las palabras, y por una red de co-ocurrencia léxica construida para cada palabra relacionada. El método fue probado en un conjunto de pares de palabras tomadas de un thesaurus de economía, y se obtuvo un 80 por ciento de precisión, lo cual es muy alentador.
Palabras Clave: Relación de oposición, redes de co-ocurrencia léxica.

1. Introducción

WordNet, es una Base de Datos Léxica (BDL) que tiene una amplia cantidad de relaciones léxicas entre palabras para el idioma inglés. A través del tiempo, WordNet ha tenido varias aplicaciones en muchos campos del procesamiento del lenguaje natural, como es el caso de la recuperación de información, desambiguación del sentido de una palabra, y refinamiento de otras bases de datos léxicas, entre otras [4,8,9,11]. Sin embargo, la poca especialización de esta base de datos conduce a la necesidad de construir BDL para dominios específicos, tarea que necesita métodos automáticos para reducir el tiempo y el esfuerzo, y permitir así su aplicabilidad a otros dominios.

Los procedimientos automáticos que se han propuesto para identificar las relaciones léxicas subyacentes en el lenguaje natural siguen de cierta manera el procedimiento manual. En esencia, se trata de observar las concordancias de dos términos para proponer una relación entre ellos. Por ejemplo, P. Hindle [14] propuso un método basado en la asignación de rasgos sintácticos frecuentes de un término, para representar y determinar la similitud entre sustantivos. Un método semejante es el que propuso G. Grefenstette [1] para construir thesauri por "conocimiento pobre", partiendo de un texto grande en un dominio específico. Este investigador observó que dos términos son mutuamente vecinos, si uno es altamente frecuente en los contextos del otro, y viceversa. En esta misma dirección, un trabajo para el portugués de C. Veraschin [5] refinó el método, incluyendo mayor información sintáctica en la representación de los términos a relacionar usando frases preposicionales. Tal vez el trabajo más significativo sea el de M. Hearst [7], quien obtuvo relaciones de hiponimia a partir de patrones léxico sintácticos. Por otra parte, Sanderson y Croft [15] identifican relaciones de hiponimia con base en la idea de subsunción, sustentada en la contención del conjunto de documentos que contienen ambos términos en el conjunto de documentos que contiene a uno de ellos, es decir: x subsume a y si los textos que contienen a y tienen una alta probabilidad de contener a x. Mas recientemente, H. Jiménez [6] senãló que el término puede usarse con diferentes sentidos y, por tanto, se pueden identificar algunas relaciones de sinonimia e hiponimia para algunos sentidos de las palabras en cuestión, refiriendo a la subsunción mediante el agrupamiento de los rasgos que representan los sentidos de los términos. Un enfoque que decide si un par de palabras son antónimos usa vectores conceptuales [12]: descompone cada término del par, por medio de un MRD (Machine Readable Dictionary) y un thesaurus. Este método determina si el par está en relación de antonimia, realizando operaciones en el espacio vectorial, como la medida del coseno.

En este trabajo se presenta un método para identificar antónimos (o en relación complementaria, como "verdadero"y "falso", Cruse [13]). Este método está basado en la representación de pares de términos -que se han encontrado como relacionados pero sin conocer el tipo de relación- mediante tres rasgos provenientes de sus contextos: la puntuación inversamente proporcional aportada por la distancia que separa el par de palabras, algunos patrones léxico sintácticos y el grado de similitud determinado por una red de co-ocurrencia léxica

construida para cada palabra de los pares. El corpus del cual se extraen los pares relacionados también provee los contextos para determinar los rasgos. Se tomaron un conjunto de pares que fueron usados como ejemplos positivos. Los umbrales útiles en la clasificación fueron definidos con base en los valores obtenidos con los ejemplos positivos.

El determinar la relación de antonimia es un problema difícil de resolver debido a que los antónimos y los sinónimos presentan rasgos similares. Así, es necesario establecer un mecanismo que permita distinguir entre estos dos tipos de relaciones. De manera específica, D. Cruse [13] comenta sobre los opositivos lo siguiente:

> "... in respect to all other features, they are identical, hence their semantic closeness; along the dimension of difference, they occupy opposing poles, hence the feeling of difference."

esto también fue expresado por L. Wanner [10]: los antónimos x y y cumplen que x tenga como rasgos ABC y y tenga $AB\neg C$.

En este trabajo se recurre al uso de las redes de co-ocurrencia léxica, las cuales fueron usadas por Philip Edmonds [3] para seleccionar el sinónimo más adecuado en un contexto, y que sirven también como un mecanismo de filtrado en el proceso de determinación de antónimos.

Cada uno de los rasgos usados en la clasificación de relaciones de oposición se encuentran descritos en la sección 2 de este trabajo. La sección 3 indica cómo calcular los pesos asociados a los rasgos, y muestra una prueba del procedimiento de clasificación sobre un ejemplo de pares relacionados. Al final, se presenta una discusión de los resultados obtenidos.

2. Rasgos Utilizados en la Identificación de Antónimos

El método de clasificación de rasgos propuesto, está basado en una función de puntaje, compuesta por los pesos de los mismos rasgos. El método usa los umbrales determinados por ejemplos positivos. A continuación se describe cada uno de estos rasgos necesarios en el cálculo del puntaje, el cual permite clasificar pares de palabras.

Distancia Inversamente Proporcional (DIP). Este rasgo tiene su base en la observación de contextos que tienen palabras relacionadas. DIP representa qué tan cercanas están dos palabras. En contextos de palabras relacionadas, las palabras antónimas muy frecuentemente co-ocurren con una distancia pequeña. Esta observación está basada por el uso de antónimos con propósitos de contraste. Se define DIP de la siguiente manera: la distancia entre las palabras relacionadas (número de palabras que las separan) es complementada con respecto a la distancia máxima aportada por ejemplos positivos. Se toma el valor máximo de la distancia complementaria en los ejemplos positivos como:

$$\Delta_M = \max_{(x_1,y_1)\in Pos} \{ \max_{(x_2,y_2)\in Pos} \bar{\Delta}(x_2,y_2) - \bar{\Delta}(x_1,y_1)\} \quad (1)$$

donde $\bar{\Delta}(x,y)$ es la distancia promedio entre las palabras x y y en sus contextos.

Patrones Léxico Sintácticos (PLS). Los PLSs han sido introducidos en el trabajo de P. Hearst [7] con buenos resultados en la identificación de hipónimos. Se identificaron diversos patrones en los contextos con relaciones de oposición guiados por palabras clave, signos de puntuación, y distancia entre palabras relacionadas. Los PLSs son representados como expresiones regulares. Ejemplos de estos patrones aparecen en la tabla 1. Los patrones contienen palabras clave como: *pero*, *desde*, *hasta*, *sino*, *y*, *o*.

Nr	Expresión Regular	Peso
1	Ant1 word*, **pero** word* Ant2	5
2	**desde** word* Ant1 **hasta** word* Ant2	4
3	Ant1 word* [,|;] **sino** word* Ant2	5
4	Ant1 word{0,4}[**y**|**o**] word{0,4} Ant2	1

Tabla 1. Expresiones regulares y sus pesos.

Redes de Co-ocurrencia Léxica (RCL). Debido a que los antónimos se comportan de manera similar a los sinónimos, se tuvo que discernir entre estas dos relaciones. Por lo tanto, se decidió manejar una representación para palabras que pudieran contener información usada en problemas con sinonimia. Las RCL fueron usadas en esta propuesta como un mecanismo de selección del sinónimo mas adecuado en un contexto [3]. El procedimiento usado para calcular una RCL para un término, llamado *raíz*, se describe a continuación:

1. Para formar el contexto de la raíz, x, se consideran las oraciones del corpus (usado en la construcción de nuestro thesaurus), $\mathcal{C}$, que la contienen:
$$A_1(x) = \{y | x, y \text{ co-occurren en una oración de } \mathcal{C}\} \quad (2)$$
2. El contexto de la raíz es filtrado, descartando todas las palabras cuya información mutua [2] es menor que 5; el resto de las palabras se les llama palabras asociadas de primer-orden:
$$A'_1(x) = \{y | y \in A_1(x) \wedge MI(x,y) > 5\} \quad (3)$$
3. El proceso se repite para las palabras $y \in A'_1(x)$ (palabras asociadas de segundo-orden). Esto depende del nivel deseado de la RCL. En general, las palabras asociadas de $n-$orden para x son determinadas de acuerdo con:
$$A'_n(x) = \bigcup_{y\in A_n^0 \Gamma_1(x)} A'_1(y) \quad (4)$$

La necesidad de discernir entre relaciones de sinonimia y antonimia de un par (a_1 y a_2), requirió del uso de RCL para decidir si pueden ser consideradas sinónimos. Una manera de hacer esto es calcular la similitud relativa entre las palabras. La similitud, en términos de RCL, está definida en [3] con el propósito de calcular la similitud de una palabra w en un contexto, la cual está basada en la similitud entre las palabras w y x, donde x debería estar en la RCL de w. Esto permite suponer que cada arco de la RCL tenga un peso (este tópico será explicado en la sección 3), y considerando $P = (w_0, w_1, \ldots, w_n)$ como el camino de costo mínimo de $w_0 = w$ a $w_n = x$, entonces, la similitud de w y x queda definida como sigue:

$$sig(w, x) = \frac{1}{d^3} \sum_{w_i \in P} \frac{t(w_{i-1}, w_i)}{i}, \tag{5}$$

donde $t(w_{i-1}, w_i)$ es el $t-$score definido en[2], y se calcula de la siguiente manera:

$$t(w_{i-1}, w_i) = \frac{P(w_{i-1}, w_i) - P(w_{i-1}) \cdot P(w_i)}{\sqrt{(\sigma^2(P(w_{i-1}, w_i)) + \sigma^2(P(w_{i-1}) \cdot P(w_i)))}} \tag{6}$$

con $P(w_{i-1}, w_i) = \frac{fr(w_i \Gamma_1, w_i)}{N}$ y $\sigma^2 P(w_{i-1}, w_i) \cong N \cdot P(w_{i-1}, w_i)$.

Dadas las palabras a_1 y a_2 con nodos en RCL $L(a_1)$ y $L(a_2)$, respectivamente, se espera que ellas tengan una alta similitud si la suma de las similitudes de las palabras en $L(a_1) \cap L(a_2)$ para ambas RCLs es alta. Por lo tanto, es necesario saber cuándo la similitud es alta. En este trabajo se hace referencia a la similitud total para calcular la similitud relativa. La similitud total, s_t, es calculada sumando todos los pesos de ambas $L(a_1)$ y $L(a_2)$. En resumen, la similitud relativa entre las palabras a_1 y a_2 está definida como:

$$s_r(a_1, a_2) = \frac{1}{s_t} \sum_{w \in \{a_1, a_2\}, x \in L(a_1) \cap L(a_2)} sig(w, x), \tag{7}$$

Se puede entonces decir que los rasgos ayudan a determinar el puntaje total de un par de términos relacionados, para determinar si estos son sinónimos.

3. Determinación del Puntaje Total

El puntaje total considera cada rasgo descrito anteriormente. De acuerdo con los valores observados en ejemplos positivos, un peso es asignado a cada rasgo. Los pesos más altos expresan que los valores de dichos rasgos fueron observados en ejemplos positivos. Por ejemplo, un peso bajo dado a una expresión regular indica que algunos ejemplos negativos empatan con tal patrón.

El puntaje total $S_g(a_1, a_2)$ es la suma de todos los valores de los rasgos. Este valor se calcula como:

$$S_g(a_1, a_2) = W_{er}(a_1, a_2) + W_d(a_1, a_2) + W_{net}(a_1, a_2), \tag{8}$$

donde $W_{er}(a_1, a_2)$ es el peso obtenido por las expresiones regulares que empatan los contextos que contienen tanto a a_1 como a a_2, $(W_d(a_1, a_2))$ es el peso aportado por la distancia inversamente proporcional y $s_r(a_1, a_2)$ $(W_{net}(a_1, a_2))$ un valor inversamente proporcional al valor de similitud entre a_1 y a_2, obtenido a través de las RCLs. Cada peso está normalizado en el rango $[0, 1]$. Así, S_g es menor que 3. A continuación se describen cada uno de estos valores:

Dado E, el conjunto de todas las expresiones regulares que empatan los contextos de a_1 y a_2; $weight(e)$, el peso de e (algunos ejemplos se muestran en la tabla 1); y $fr(e)$ la frecuencia relativa e que empata con los contextos, se define $W_{er}(a_1, a_2) = \sum_{e \in E} weight(e) \cdot fr(e)$.

$W_d(a_1, a_2)$ es considerado el puntaje máximo para ejemplos positivos. A partir de Δ_M (ec. 1) se determina $W_d(a_1, a_2)$ para lograr un valor normalizado entre $[0, 1]$:

$$W_d(a_1, a_2) = \frac{\Delta_M - \bar{\Delta}(a_1, a_2)}{\Delta_M} \tag{9}$$

Finalmente, $W_{net}(a_1, a_2)$ sigue un cálculo similar a la ecuación 9:

$$W_{net}(a_1, a_2) = \frac{\max\limits_{(x,y) \in Pos} \{s_r(x, y)\} - s_r(a_1, a_2)}{\max\limits_{(x,y) \in Pos} \{s_r(x, y)\}} \tag{10}$$

En la prueba se usa un corpus compuesto por 26297 oraciones, 11575 términos (incluyendo nombre propios, después de eliminar las palabras cerradas y lematizar el resto). En la tabla 2 se muestra el conjunto de entrenamiento, el cual está compuesto por 15 pares, diez positivos y cinco negativos, que fueron usados para afinar los umbrales. El conjunto de prueba estuvo compuesto por 8 pares de antónimos y 10 pares que tienen algún otro tipo de relación semántica. La tabla 3 muestra algunos ejemplos de la aplicación del método, el veredicto indica si los pares de palabras, positivos y negativos, han sido identificados como antónimos o no antónimos respectivamente, de manera correcta o incorrecta. Es importante notar que solo cuatro pares fueron clasificados incorrectamente, y que el método también descarta hipónimos.

4. Discusión

Se considera que la identifación de relaciones de antonimia con el uso combinado de patrones léxico sintácticos y las redes de co-ocurrencia léxica es una línea interesante de investigación. Se ha observado que es conveniente usar una profundidad de tamaño 3 para una RCL, ya que a mayor profundidad, la cantidad de términos involucrados tiende a debilitar la asociación de términos del nodo raíz. Se destaca también la importancia que posee el cálculo de la información mutua como mecanismo de filtrado, ya que decrementa la posibilidad de que dos cualesquiera términos posean una relación de primer orden, y el puntaje t que define los pesos entre los términos.

El umbral tomado para la determinación de la relación entre un par de palabras relacionadas en este trabajo, está basado en el cálculo obtenido a

Pares de Palabras	Distancia	PLS	RCL	Veredicto
Absoluto-Relativo	0.64814815	0.79333333	0.97665313	Correcto
Compra-Venta	0.59259259	0.16666667	0.47302784	Correcto
Natural-Artificial	0.68518519	0.58	0.9724478	Correcto
Consumidor-Productor	0.62962963	0.56	0.47911833	Correcto
Escasez-Abundancia	1	1	0.36180394	Correcto
General-Particular	0	0.50666667	1	Correcto
Máximo-Mínimo	0.51851852	0.26666667	0.64022622	Correcto
Positivo-Negativo	0.44444444	0.80666667	0	Correcto
Verdad-Mentira	0.72222222	0.66666667	0.9650522	Correcto
Vida-Muerte	0.88888889	0.86666667	0.77392691	Correcto
Corrección-Ajuste	0.75925926	0.2	0.08164153	Correcto
Hombre-Humano	-0.38888889	0.10666667	0.97056265	Correcto
Mercancía-Producto	-0.05555556	0.18666667	0.83236659	Correcto
Moneda-Dinero	0.05555556	0.02	0.87558005	Correcto
Obrero-Trabajador	-0.48148148	0	0.86934455	Correcto

Tabla 2. Conjunto de entrenamiento: los primeros 10 pares son antónimos.

Pares de Palabras	Distancia	PLS	RCL	Veredicto
Bajo-Alto	0.53703704	0.2	0.62601508	Correcto
Activo-Pasivo	0.67037037	0.76666667	0.28741299	Correcto
Grande-Pequeño	0.47407407	0.53333333	0.97317285	Correcto
Oferta-Demanda	0.79259259	0.76666667	0.04350348	Correcto
Pregunta-Respuesta	0.53703704	0.02666667	0.1712587	Incorrecto
Público-Privado	0.48148148	0.53333333	0.75710557	Correcto
Social-Individual	0.27407407	0.63333333	1.0549594	Correcto
Interior-Exterior	0.50740741	0.73333333	0.62427494	Correcto
Confianza-Fe	-2.05555556	0	0.90762761	Correcto
Crédito-Préstamo	0.33333333	0.33333333	0.36107889	Correcto
Cultura-Democracia	0.7037037	0.26666667	0.67444896	Incorrecto
Harina-Trigo	0.6	0.43333333	0.05191415	Correcto
Inversión-Gasto	0.53148148	0.1	0.33990719	Correcto
Miembro-Comunidad	0.90740741	0	0.70562645	Correcto
Pobreza-Problema	0.56851852	0.03333333	1.03045244	Incorrecto
Productor-Benefactor	0.32222222	0.16666667	0.34353248	Correcto
Rasgo-Característica	0.85185185	0	0.72186775	Incorrecto
Semana-Día	0.30185185	0	0.50101508	Correcto

Tabla 3. Conjunto de prueba: los primeros 8 pares son antónimos.

través del conjunto de entrenamiento. Sin embargo, sería deseable obtener dicho umbral, por ejemplo, mediante una relación entre el tamaño del corpus y el número de contextos del par de palabras relacionadas.

Los resultados obtenidos son alentadores y motivan el estudio profundo del uso de la técnica propuesta como un mecanismo para la identificación de antónimos en un texto sin información adicional, como etiquetas o algunas estructuras.

Los resultados apuntan al uso de expresiones regulares como primer mecanismo de restricción para posteriormente filtrar a través del uso de redes de co-ocurrencia léxica. Por último, si existieran pares aún sin categorizar, entonces podría hacerse uso de la distancia promedio esperada para términos en relación de antonimia. Aún es necesario validar los resultados encontrados, en una muestra grande de pares de palabras y conocer si el método propuesto podría aplicarse a dominios diversos.

Agradecimientos

Agradecemos los comentarios de los árbitros de este trabajo, así también el apoyo parcial recibido por parte del proyecto VIEP III 9-04/ING/G.

Referencias

1. G. Grefenstette: "Explorations in Automatic Thesaurus Discovery", Kluwer Academic Publishers, Boston Hardbound, ISBN 0-7923-9468-2 July 1994.
2. Church, Kenneth Ward; Gale, William; Hanks, Patrick; Hindle, Donald; Moon, Rosamund: "Lexical Substitutability", In: Atkins, B. T. S.; Zampolli, Antonio (eds.): *Computational Approaches to the Lexicon*. Oxford University Press, pp. 153-180, 1994.
3. Edmonds P.: "Choosing the word most typical in context using a lexical co-ocurrence network", *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, pp. 507-509, 1997.
4. Morato J., Marzal M.A., Lloréns J., Moreiro J.: "WordNet Applications", Petr Sojka, Karel Pala, Pavel Smrc, Christiane Fellbaum, Piek Vossen (Eds.): *Proceedings GWC 2004*, pp. 270-278, 2004.
5. Caroline Varaschin Gasperin Vera Lúcia Strube de Lima, "Experiments on Extracting Semantic Relations from Syntactic Relations", *CiCLing 2003*, LNCS 2588, pp. 314-324, 2003.
6. Jiménez-Salazar, H., "A Method of Automatic Detection of Lexical Relationships Using a Raw Corpus", *CiCLing 2003*, LNCS 2588, pp. 325-328, 2003.
7. Hearst, M.: "Automatic acquisition of hyponyms from large text corpora". *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France, 1992.
8. Yorick Wilks, Roberta Catizone: "Lexical Tuning". *CICLing 2002*, pp. 106-125, 2002.
9. Paolo Rosso, Francesco Masulli, Davide Buscaldi, Ferran Pla, Antonio Molina: "Automatic Noun Sense Disambiguation", *CICLing 2003*: pp. 273-276

10. L. Wanner: "Lexical Functions in Lexicography and Natural Language Processing", John Benjamins Publishing Company, 1996.
11. Hearst, M.: "Automated Discovery of WordNet Relations", in *WordNet and Electronic Lexical Database*, C. Fellbaum (Ed.), The MIT Press, 1999, pp. 131-152.
12. Schwab, D.,Lafourcade, M., Prince, V.: "Antonymy and Conceptual Vectors", in *the Proceedings of the 19th Conference on Computational Linguistics*, 2002, pp. 904-910.
13. Cruse, D.: "Lexical Semantics", Cambridge , Cambridge University Press, 1986.
14. Hindle, D.: "Noun Classification from Predicate-Argument Structures", in *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 1990, pp. 268-275.
15. Sanderson M., Croft B.: "Deriving concept hierarchies from text", In *Proceedings of the 22 a Annual lnt. ACM S1G1R Conference on Research and Development in Information Retrieval*, pp. 206 - 213, Berkeley, CA, August 1999.