Una Metodología para la Creación de Thesauri

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

Resumen. La necesidad de incrementar la potencialidad de los recursos léxico semánticos, conlleva al empleo de técnicas que identifiquen de manera automática, relaciones semánticas entre palabras. En este trabajo se presenta una metodología para la construcción de *thesauri*, que se contempla utilizar en el proceso de enriquecimiento de otros recursos léxico semánticos como WordNet, y que, como se sabe, son indispensables en el proceso de WSD. El *thesaurus* que se presenta en este trabajo se constituye por pares de palabras relacionadas con tres tipos de relaciones semánticas, identificadas de manera automática, a saber: sinonimia, oposición y amplitud. En una evaluación manual, se obtuvo arriba del 80% de precisión en cuanto al tipo de relaciones semánticas identificadas.
Palabras Clave: Thesaurus, Recursos léxico semánticos.

1 Introducción

Con el fin de desarrollar sistemas robustos para el PLN, es necesario proveerse de ricos recursos, los cuales solamente podrán obtenerse con metodologías apoyadas en Internet [1] [2]. En particular, es necesario incrementar la potencialidad de los WordNets existentes para lograr eficacia en las aplicaciones de la tecnología del lenguaje humano. En el enriquecimiento de WordNet subyace la necesidad de la desambiguación del sentido de las palabras, además de técnicas que permitan aumentar su léxico o hacerlo específico para cierto dominio.

Es necesario, entonces, el empleo de técnicas que incidan en la identificación automática de relaciones semánticas entre palabras. El presente trabajo expone una metodología para construir un *thesaurus* a partir de texto plano.

Identificar automáticamente el tipo de relación semántica entre palabras relacionadas es una tarea difícil en la construcción de un *thesaurus*. En este trabajo se han considerado las relaciones semánticas de sinonimia, oposición y de amplitud. Como relaciones de oposición se consideran a los antónimos complementarios, de dirección, antipodales, reversivos, entre otros [8]. En las relaciones de amplitud se engloban a las jerárquicas: hiponimia-hiperonimia, holonimia-meronimia y cohiponimia.

En la siguiente sección se expone de manera breve la metodología propuesta para identificar relaciones semánticas entre pares de palabras relacionadas, extraídas de un texto plano y de un dominio particular. En la tercera sección se

presenta una evaluación de la aplicación de esta metodología en el dominio de Economía. Al final se analiza la manera en que esta metodología puede ser útil en el marco del enriquecimiento de diccionarios léxico semánticos.

2 Metodología Propuesta

Fundamentalmente, la presente metodología hace uso del método propuesto por G. Grefenstette [3] para identificar pares de términos relacionados. La identificación de relaciones semánticas se inspira en los patrones léxico sintácticos propuestos por P. Hearst [7] y se refina con las ideas de redes de co-ocurrencia léxica que utilizó P. Edmonds [5] con base en las medidas estadísticas empleadas por K. Church [4]. Así, el proceso de construcción del *thesaurus* consta de tres pasos que a continuación se describen.

2.1 Identificación automática de palabras relacionadas

La identificación de las palabras relacionadas del recurso inicia con un análisis estadístico basado en el corpus. Para cada sustantivo del corpus, con una frecuencia mayor que 10, se extraen sus contextos (oraciones donde se encuentra dicho vocablo); las palabras relacionadas del *thesaurus* son vecinos mutuos, si uno es altamente frecuente en los contextos del otro, y viceversa. Este grado de vecindad es estimado utilizando una variante de la medida de Jaccard.

El tamaño del corpus utilizado para construir este léxico es de 4.67MB de texto plano preprocesado (se eliminan las palabras cerradas y se lematiza el resto del texto), compuesto por 26,297 oraciones y 11,575 términos, incluyendo nombres propios y términos multipalabra.

2.2 Rasgos y umbrales empleados

Con el propósito de identificar los tres tipos de relaciones léxicas mencionadas, se hace uso de tres rasgos: patrones léxico-sintácticos que empatan con expresiones regulares en los contextos (PLS), redes de co-ocurrencia léxica (RCL) y una distancia promedio de separación (DPS). Dado que los rasgos son utilizados de manera diferente en la identificación de cada relación semántica, se presenta en esta misma sección un apartado para cada tipo de relación semántica.

Ejemplos de PLS para relaciones de oposición y de amplitud aparecen en la tabla 1. En este caso, los pesos para cada PLS fueron asignados con base en la precisión de cada expresión regular. Así, se asignó un punto al PLS por cada 5 puntos porcentuales de su precisión. Este peso es usado posteriormente, en el cálculo del puntuación global, básicamente, sumando los pesos de todas los PLS que son satisfechos por un par de palabras relacionadas, (a_1, a_2), y normalizando el resultado obtenido (este valor es representado posteriormente como $W_{PLS}(a_1, a_2)$).

El procedimiento para construir una RCL es como se presenta en [9], excepto por las siguientes consideraciones (que tienen su base en la experimentación):

Nr	Expresión Regular	Peso
1	Ant word*, **pero** word* Ant	20
2	**desde** word* Ant **hasta** word* Ant	16
3	Ant word* [,|;] **sino** word* Ant	20
1	Hip, **incluyendo** word,* [o | y] Hip	20
2	Hip word{0,1} **a semejanza de** word{0,5} Hip	12
3	**tal** Hip **como** {word,}* [o | y]{0,1} Hip	20

Tabla 1. PLS para opuestos y amplios con sus pesos.

1. Para el filtrado de los contextos, la información mutua (IM) se toma dentro del rango [4.5,6]. Los pares cuya IM está fuera de este rango se descartan.
2. Los cálculos de similitud y contención entre dos redes, se realizan sólo considerando las palabras comunes que se encuentran en los mismos niveles.
3. En caso de que en sus asociaciones de primer orden, las RCL se contengan mutuamente (RCL reflexivas), dichas subredes son ignoradas en los cálculos; en caso contrario (RCL disjuntas), se usan de manera completa.

El puntuación aportado por el rasgo RCL es normalizado y se representa mediante $W_{RCL}(a_1, a_2)$.

El rasgo de Distancia Promedio de Separación (DPS), fue usado también en [9] para identificar antónimos en textos planos. DPS indica que tan cercanas están dos palabras en los contextos comunes. Se ha observado que las palabras antónimas frecuentemente co-ocurren más cercanas que las palabras en relación de amplitud. Específicamente, los experimentos revelan que la DPS entre palabras opuestas es de 2.5, y que la mayoría de estas DPS caen en el rango [0,4.5]; por otro lado, la DPS entre palabras en relación de amplitud es de 5.1 y el rango es: [3.5,7]. Estos resultados se normalizan, y son contemplados en la metodología como umbrales en algunos casos. Su representación en el puntuación global es a través de $W_{DPS}(a_1, a_2)$.

2.3 Identificación automática del tipo de relación semántica

Relaciones de oposición. En esta parte, se hace uso de los tres rasgos previamente descritos, en una función de puntuación global y de umbrales establecidos con base en ejemplos positivos. PLS mostró ser el rasgo más importante, por lo cual se utiliza, primero, para seleccionar el grupo de pares de términos relacionados del *thesaurus* con mayor peso, y después como una puntuación parcial que contribuye a la puntuación total. En este trabajo se descubrieron 22 PLS para detectar relaciones de oposición, mediante los cuales se identificó un grupo de 65 pares de términos relacionados del *thesaurus*, que sobrepasaron el umbral establecido de 1.4.

RCL es el segundo rasgo prioritario en esta identificación, pero su efectividad está en relación de sus características, es decir, si las redes son reflexivas o disjuntas. Observaciones sobre estos dos tipos de RCL, hacen suponer que las RCL reflexivas representan palabras más fuertemente relacionadas y comunes

en el contexto que las RCL disjuntas. Bajo esta hipótesis, las RCL reflexivas aportarían una puntuación más estable a la función de puntuación que las RCL disjuntas, por lo que se decidió dividir el grupo de los 65 pares identificados del *thesaurus* en dos: grupo de RCL reflexivas y grupo de RCL disjuntas.

La DPS se utiliza en los dos agrupamientos como una puntuación determinada al complementarla con respecto a 4.5 (distancia complementada), que es el límite superior del rango de números donde caen los promedios de la mayoría de las palabras en relación de oposición. La distancia complementada arroja una puntuación menor que uno, de manera proporcional. Se toma el valor máximo de la distancia complementada en los ejemplos positivos como:

$$\Delta_M = \max_{(x,y)\in Pos} \{4.5 - \bar{\Delta}(x,y)\} \tag{1}$$

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

A continuación se describen los criterios y umbrales establecidos para la detección de relaciones de oposición, particulares a los dos agrupamientos ya descritos.

1. **RCL reflexivas.** La puntuación total (S_g) se calcula como:

$$S_g(a_1,a_2) = W_{PLS}(a_1,a_2) + W_{DPS}(a_1,a_2) + W_{RCL}(a_1,a_2), \tag{2}$$

   donde (a_1,a_2) es un par de palabras relacionadas. Así, dado que los valores de $W_{PLS}(a_1,a_2)$, $W_{DPS}(a_1,a_2)$ y $W_{RCL}(a_1,a_2)$ están normalizados a 1, entonces $S_g(a_1,a_2) \leq 3$. Este cálculo está descrito también en [9].
2. **RCL disjuntas.** El método considera los rasgos: DPS y PLS. Este criterio tuvo su base en la experimentación, y durante la etapa de entrenamiento se comprobó que el uso conjunto de estos dos rasgos arroja mejores resultados que la combinación de los tres rasgos. Así, la función de puntuación cumplirá $S_g(a_1,a_2) \leq 2$.

Identificación de relaciones de amplitud. Nuevamente los PLS son el rasgo más confiable, por lo que fueron utilizadas como mecanismo de selección de pares de palabras relacionadas a ser analizadas. Por medio de los PLS se obtuvo un grupo de 105 pares de los 635 del *thesaurus*.

Bajo la hipótesis de que el grado de contención y las diferencias de tamaño entre redes pueden señalar algún tipo de relación de amplitud [6], el grupo seleccionado por los PLS fue dividido en tres sub-grupos: diferencias grandes, diferencias medianas y diferencias pequeñas, y el grado de contención de las RCL funcionó como mecanismo de eliminación.

La DPS sólo se utilizó en el grupo donde las diferencias en los tamaños de las redes fue pequeña, debido a que las RCL en este caso son más estables; lo cual hace suponer que también las distancias están bien establecidas.

A continuación se describen las características particulares de cada uno de los tres sub-grupos de PLS generados.

– **Diferencias grandes.** Estuvieron en el rango de [67%, 100%) y 21 pares cayeron en este grupo. Mediante el grado de contención se eliminaron 5 correctamente, con un umbral de contención del 20%.
– **Diferencias medianas.** Estuvieron en el rango de [33.3%, 67%) y 38 pares cayeron en este grupo. Mediante el grado de contención se eliminaron 4 correctamente, con un umbral de contención del 13%.
– **Diferencias pequeñas.** Estuvieron en el rango de [0%, 33.3%) y 45 pares cayeron en este grupo. Mediante el grado de contención se eliminaron 12 pares, 11 correctamente, con un umbral de contención del 20%; en este grupo se utilizó DPS; mediante este rasgo se eliminaron 22 pares, 18 de manera correcta.

3 Evaluación

Términos opuestos De los 65 pares de palabras relacionadas, seleccionadas por los PLS, 20 cayeron en el grupo de RCL reflexivas y 45 en RCL disjuntas. En las RCL reflexivas se marcaron 15 pares en relación de oposición, 9 de manera correcta y 6 de manera incorrecta. En las RCL disjuntas se marcaron 33 de los cuales 18 fueron correctos. De manera general, de 65 pares, la metodología marcó 48, de los cuales 27 fueron marcados de manera correcta, obteniéndose así una precisión de 56.25%.

Términos amplios Una vez empleada la metodología anterior, de los 105 pares de palabras relacionadas, seleccionadas por los PLS, se eliminaron 43, de los cuales 38 fueron eliminados correctamente, quedando 62 pares de los cuales 40 están en relación de amplitud, lo que arroja una precisión del 64.5%.

Consideraciones adicionales Considerando que los pares clasificados como opuestos y a la vez como amplios, mantienen una relación de cohiponímia, es posible trasladarlos del grupo de opuestos al de términos con relación de amplitud. De esta manera, la precisión para el grupo de opuestos aumenta a 79.6%, mientras que para el grupo de relaciones de amplitud se obtiene un 84.6%.

En la tabla 2 se muestran algunos ejemplos de los 635 pares relacionados extraídos de un corpus de Economía.

4 Perspectivas

El método reconoce relaciones semánticas de cohiponimia y oposición de manera indistinta. Se ha observado que este tipo de comportamiento ocurre incluso en los seres humanos, sin embargo, pensamos que convendría experimentar con el método a fin de obtener algún rasgo adicional que permita dicriminar entre estos tipos de relaciones semánticas.

Palabra Uno	Palabra Dos	Relación	Evaluación
Países_ricos	**Países_pobres**	CHIP	Correcto
Metal	**Oro**	Hiper	Correcto
Oro	**Metales_preciosos**	HIP	Correcto
Oro	**Mina**	HOLO	Sin Clase
Plata	**Cobre**	CHIP	Correcto
Pescado	**Harina**	CHIP	Correcto
Aceite	**Harina**	CHIP	Correcto
Harina	**Trigo**	HOLO	Sin Clase
Positivo	**Negativo**	ANT	Correcto
Falso	**Verdadero**	ANT	Correcto
Justicia	**Equidad**	SIN	Incorrecto

Tabla 2. Fragmento del *thesaurus* de Economía.

Es posible mejorar la precisión de las relaciones semánticas identificadas afinando los rasgos que se han descrito y, sobre todo, apoyándose en una fuente que provea grandes volúmenes de contextos para los términos.

La presente metodología incide solamente en la identificación de relaciones semánticas, pero se está considerando el sentido predominante de las palabras relacionadas. Es necesario agrupar los contextos para obtener las diferentes relaciones semánticas entre parejas de sentidos [6]. Se requiere, así, hacer uso de la Web [1], que es el único medio en el cual es posible obtener la suficiente riqueza para este fin.

Algunos trabajos que hacen referencia a medidas de similitud de WordNet [10], permiten asociar sentidos de WordNet a grupos de contextos que representan el sentido de una palabra. Así, si es posible tener un mapeo entre los sentidos de la palabra y aquellos sentidos ya definidos en WordNet, será posible también enriquecer los recursos léxico semánticos existentes.

Referencias

1. Mihalcea, R. & Moldovan, I.: "Automatic Acquisition of Sense Tagged Corpora" *Proc. of FALIRS99*, Florida, 1999.
2. Leacock, C.; Chodorow, M. & Miller, G. A.: "Using Corpus Statistics and WordNet Relations for Sense Identification", *Computational Linguistics*, 24(1), 1998.
3. Grefenstette, G.: "Explorations in Automatic *thesaurus* Discovery", Kluwer Academic Publishers, Boston Hardbound, ISBN 0-7923-9468-2 July 1994.
4. Church, K. W.; Gale, W.; Hanks, P.; Hindle, D.; Moon, R.: "Lexical Substitutability", In: Atkins, B. T. S.; Zampolli, A. (eds.): *Computational Approaches to the Lexicon*. Oxford University Press, pp. 153-180, 1994.
5. Edmonds, P.: "Choosing the word most typical in context using a lexical cooccurrence network", *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, pp. 507-509, 1997.
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.: "Automated Discovery of WordNet Relations", in *WordNet and Electronic Lexical Database*, C. Fellbaum (Ed.), The MIT Press, 1999, pp. 131-152.
8. Cruse, D.: "Lexical Semantics", Cambridge , Cambridge University Press, 1986.
9. Lucero, C.; Pinto, D.; Jiménez-Salazar, H.: "Identificación de antónimos en textos planos", In *Cuarto encuentro de computación*, pp. 203 - 211, Colima-México, CA, Septiembre 2004.
10. Patwardhan, S.; Banerjee, S.; and Pedersen, T., 2003. Using measures of semantic relatedness for word sense disambiguation. In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics(CICLing 2003), Mexico City.