Pruebas con Algoritmos de Agrupamiento para generar una Base de Datos Léxica

Sofía Paniagua Rivera, Héctor Jiménez-Salazar & David Pinto

Facultad de Ciencias de la Computación
B. Universidad Autónoma de Puebla,
14 sur y Av. San Claudio. Edif. 135. Ciudad Universitaria,
Puebla, Pue. 72570. México,
Tel. (01222) 229 55 00 ext. 7212 Fax (01222) 229 56 72,
sofiadp@terra.com.ve, hjimenez@fcfm.buap.mx, dpinto@cs.buap.mx

Abstract. Advances on constructing a Lexical Database (LDB) are presented in this paper. The goal is to use that LDB for Word Sense Disambiguation (WSD), thus, we present an evaluation of eight ambiguous sentences over a set of clusters (generated by the MOD-SLC algorithm). Each cluster is composed by features obtained by two methods: 1) using a efficacy formula and 2) using intersection by context-pairs. Although the corpus is small, the results encourage to proceed in this direction. The first method obtained a 75-percent accuracy and the second method obtained 62.5-percent accuracy.
Keywords: Clustering, Lexical Databases, Disambiguation

Resumen. En este trabajo se presentan avances sobre la construcción de una base de datos léxica (BDL) que estará conformada por grupos de sustantivos con su respectivos rasgos. Se espera que la BDL apoye en el proceso de desambiguación del sentido de una palabra (WSD). Por tanto, en este artculo se reporta la evaluación de ocho oraciones (con un término a desambiguar) sobre un conjunto de grupos generados a través del algoritmo MOD-SLC, con solución de rasgos por medio de dos métodos: uso de fórmula de eficacia e intersección a pares entre contextos. A pesar de usar un corpus sumamente reducido, los resultados son interesantes, ya que para el primer método se obtiene un 75% de precisión mientras que para el segundo método se obtiene un 62.5% de precisión.
Keywords: Agrupamiento, Bases de Datos Léxicas, Desambiguación.

1 Introducción

Una base de datos léxica es una base de conocimientos a gran escala, normalmente hecha a mano [7], y la construcción automática de ella es todavía una línea de estudio. Una base de datos léxica destinada, por lo general, a desambiguar el sentido de las palabras, también se considera como un diccionario que posee términos relacionados, cada uno con sus respectivas características (rasgos). La desambiguación del sentido de las palabras (WSD) ha sido un tema de investigación en los últimos años y una tarea más en el procesamiento del

lenguaje natural (PLN). El trabajo de desambiguación involucra, esencialmente, empatamiento del contexto de la instancia de una palabra a ser desambiguada con información de una fuente de conocimiento externa, o información acerca de contextos de instancias de la palabra anteriormente desambiguada. Se plantea la necesidad de crear un recurso que permita desambiguar palabras de textos en español y en dominios específicos. En este trabajo se presentan avances sobre la construcción de una base de datos léxica que estará conformada por grupos de sustantivos con su respectivos rasgos. Las bases de datos léxicas tienen fundamentalmente información de términos y relaciones de significado.

El desarrollo de este recurso ha pasado por diversas fases. En la primera fase hubo necesidad de su uso en la traducción automática (MT), en donde se enfocó en la traducción de textos técnicos y de un mismo dominio. Después la MT fue orientada al desarrollo de diccionarios especializados. El sistema de Masterman(1962) utilizó una red semántica para derivar la representación de oraciones de conceptos de un lenguaje fuente [1]. Wilks(1990) usó las primitivas de Masterman para el lenguaje natural y fue uno de los primeros diseños para el tratamiento de la desambiguación de sentidos [2–4]. Otras aplicaciones se centran en la creación de diccionarios, por ejemplo, Lesk en 1986 [5] creó una base de datos donde asoció una firma a cada sentido en un diccionario. En algunos casos, se tiene un conjunto de palabras y su definición (diccionario de términos), en otros casos se tiene, además, la relación entre cada uno de los términos (por ejemplo Wordnet). El concepto de bases de datos léxicas no es reciente, sin embargo sigue en estudio; por ejemplo la base de datos léxica de EuroWordNet se está ampliando, utilizando nuevos conceptos basados en los algoritmos de Lesk, Wilks y otros precursores [6].
Los seres humanos tenemos un recurso similar a las bases de datos léxicas inmerso en nuestro conocimiento, y a partir de éste, es fácil identificar a que se refiere una palabra ambigua, siempre tomando como referencia el contexto en el cual se emite dicha palabra. Esto ha sido la base de muchos trabajos para desambiguar el sentido de las palabras y es el que seguimos en nuestro trabajo. Nuestras pruebas harán uso de corpora (homogéneo y heterogéneo) para proporcionar los contextos de los sustantivos a desambiguar. El paso más importante es agrupar automáticamente estos contextos con el fin de extraer características de cada uno de los sentidos de una palabra; pretendemos tomar cada grupo formado como un sentido. Así varios procedimientos ocupan un papel importante, como la función de similitud y el criterio de selección de características.

2 Planteamiento y solución

Tres técnicas están siendo implantadas con la finalidad de reportar su rendimiento en el ámbito de agrupamiento de contextos para sustantivos ambiguos. La primera técnica está fundamentada en el uso de un algoritmo voraz que utiliza una función de similitud basada en el valor del coseno del ángulo entre vectores representativos de cada contexto. La segunda técnica se le denomina MOD-SLC (en la que se concentra el presente trabajo) y toma sus bases en la propuesta

de Hassan et al [10] para fabricación de partes. La última técnica es una propuesta original de este trabajo y propone refinar la técnica MOD-SLC mediante iteraciones continuas, usando la técnica de temple simulado. Cada técnica hace uso de un mecanismo de representación distinto. Para representar los contextos de cada sustantivo, primero se extraen de cada contexto el mínimo número de palabras que describan al mismo (palabras representativas o rasgos), y se realiza de dos maneras. Una, calculando la eficacia y asumiendo que cada grupo está compuesto de solamente una palabra (la palabra a elegir). Se revisa para cada grupo cuál obtiene la mayor eficacia y a ese grupo se asigna la palabra como representativa de él. La otra forma de elegir las palabras representativas del grupo es obteniendo la unión de todas las intersecciones por pares de contextos (solución SR). A continuación se describe en detalle cada una de las técnicas implantadas en este trabajo.

2.1 Técnica MOD-SLC

El método SLC (Single Linkage Cluster; usado en biología) propuesto por Sneath [11, 12] utiliza el coeficiente de similitud de Jaccard para encontrar la similitud entre bacterias. Por su parte la versión modificada del método SLC (MOD-SLC), utiliza una variante en el coeficiente de similitud de Jaccard conocido como coeficiente de similitud de Jaccardian o non-Jaccardian [10], el cual es una medida del nivel de empatamiento en donde el número de empatamientos es dividido por una cantidad normalizada. Este coeficiente tiene un término adicional en el numerador y es, básicamente, el número de parejas perdidas, el cual es dividido por la normalización de términos. En nuestro caso, hemos hecho uso de la medida de Jaccardian con la finalidad de medir el grado de similitud entre contextos [10]. El coeficiente de similitud queda definido como sigue:

$$SB_{ij} = (X_{ij} + sqrt(X_{ij} * Y_{ij}))/(X_i + X_j + X_{ij} + sqrt(X_{ij} * Y_{ij})),$$

donde X_{ij} es el número de unos en común entre los contextos i y j en la matriz de incidencia (palabras en común). Y_{ij} es número de ceros en común (palabras que ambos contextos no tienen), X_i es el número de unos que están en el contexto i y no están en el contexto j (palabras diferentes). X_j se define de manera similar a X_i pero para el contexto j.

El procedimiento para la generación de los grupos se describe en la figura 1.

Paso 0 Generar la matriz de incidencia contexto-palabras.
Paso 1 Usando la matriz de incidencia crear la matriz de similitud de contextos.
Paso 2 Localizar el máximo valor de similitud (SB_{ij}) en la matriz de similitud.
Paso 3 Asignar el contexto i y el contexto j al mismo grupo.
Paso 4 Eliminar la similitud (SB_{ij}) de la matriz de similitud.
Paso 5 Si no están todos los contextos asignados regresar al paso 2.

Figura 1. Algoritmo MOD-SLC para agrupamiento de contextos.

Posteriormente se procede a asignar palabras a cada grupo usando una función de eficacia. Para cada palabra, se verifican los contextos en los que ocurre y se calcula la eficacia entre estos contextos y para cada uno de los grupos generados. La medida de la eficacia global para los grupos generados, y que se usa también para la asignación de palabras a contextos se obtiene mediante la siguiente fórmula:

$$Eficacia = (e - e_0)/(e + e_1),$$

donde, e es la cantidad de unos en la matriz, e_0 es el número de elementos excepcionales (fuera de los grupos) y e_1 es el número de ceros dentro de los grupos. La eficacia puede ser un buen mecanismo para medir la calidad de las soluciones obtenidas, ya que en el mejor de los casos, debería haber cero elementos excepcionales y ningún valor cero dentro de los grupos; lo que arrojaría un valor 1 y por tanto serían grupos bien definidos.

El experimento realizado en este trabajo se basó en el algoritmo MOD-SLC. Este algoritmo, como se mencionó anteriormente, permite agrupar los contextos obtenidos, para cada sustantivo del corpus, con base en una función de similitud. Para obtener estos contextos se utilizó un corpus heterogéneo. Se programaron las etapas de preprocesamiento, extracción de sustantivos, extracción de contextos, agrupamiento y desambiguación. En particular, se eligió el sustantivo "banco" como prueba y el proceso de prueba llevado a cabo con este sustantivo se describe a continuación.

Una vez generado el archivo de los contextos de la palabra ambigua "banco", se procedió a obtener el vocabulario de sus contextos; esto debido a que el vocabulario es necesario para generar la matriz de incidencia contextos-palabras. Se procedió a aplicar el algoritmo MOD-SLC, obteniendo los grupos, y para cada grupo se extrajeron sus rasgos. Se utilizaron 8 oraciones de prueba en las cuales aparece la palabra "banco". Las oraciones de prueba fueron extraídas del corpus del CIC[1]. Las palabras contenidas en las oraciones de prueba se presentan en la tabla 1. El proceso de desambiguación consistió en aplicar el cociente de Jaccard entre cada uno de los rasgos de los contextos obtenidos y la oración de consulta. Si el valor de similitud es igual a cero significa que la oración de consulta no pertenece a ningún grupo, en otro caso, simplemente se verifica el valor máximo obtenido (por grupo) y se concluye que el sentido de la oración de consulta corresponde a dicho grupo. Los resultados obtenidos se muestran en la tabla 2 y 3. Dichas tablas están estructuradas de la siguiente manera: en la primera columna se encuentra el número de oración a ser evaluada, en la segunda columna aparece el grupo al que debería pertenecer la oración (evaluación manual), en la tercera columna el grupo al que fue asignada (a través del cociente de Jaccard), y en la última columna se encuentra el valor de similitud entre la oración y el grupo a la que fue asignada.

[1] Recurso amablemente proporcionado por el Laboratorio de Procesamiento de Lenguaje Natural del CIC-IPN.

3 Resultados

En la tabla 2 se presenta el resultado de evaluar 8 oraciones sobre un conjunto de grupos generados a través del algoritmo MOD-SLC, con selección de rasgos por la fórmula de eficacia. Puede observarse que el número de oraciones asignadas correctamente a su grupo es 5, eso deja 3 oraciones que aparentemente no están siendo bien asignadas. Sin embargo, cabe aclarar que la oración 6 no debía ser asignada a grupo alguno, pues no existe un grupo que la represente (de acuerdo con los contextos que se obtuvieron del corpus). Por lo anterior puede decirse que, en general, el algoritmo asignó 6 respuestas correctas de 8 que se utilizaron para la desambiguación. En la tabla 3 se presenta el resultado de evaluar las mismas oraciones sobre un conjunto de grupos generados a través del algoritmo MOD-SLC, con selección de rasgos por intersección a pares entre contextos. Se puede observar que de las 8 pruebas realizadas, 5 de ellas son correctas y las 3 restantes están asignadas de manera incorrecta. Aunque nuevamente debemos tomar en cuenta que la oración 6 no debía ser asignada a grupo alguno, y sin embargo el sistema la coloca en el grupo que se espera más adecuado para dicha oración.

Tabla 1. Oraciones para realizar la prueba de desambiguación

No.	Consulta
1	El dolar puede cambiarse en varias instituciones financieras , tales como las casas de cambio o el banco
2	demanda a banco mundial ayuda monetaria para abatir pobreza de mujeres
3	Se acentúa la baja de las tasa financieras para depósitos en los bancos
4	Esta chica siempre toma asiento en el mismo banco
5	El banco no pudo estar cerrado sino el acreedor habría ido a otro lugar a cambiar el cheque
6	Esta evaluación cuenta con un banco de docencia curricular donde se brinda un seguimiento académico al profesor
7	Puesto que el banco rocoso se encuentra con pronunciada inclinación desde el río Troja
8	El banco condone parte de la deuda agrícola

4 Discusión

El trabajo presentado es el inicio de una herramienta que apoye el proceso de WSD. Para obtener el objetivo esperado y probar su efectividad se contempla realizar los siquientes pasos. En primer lugar, hacer pruebas con dos corpora, uno homogéneo y otro heterogéneo. Se usará enseguida un etiquetador de partes de discurso, para proveer mayor información a los algoritmos de agrupamiento. Asimismo, se hará uso de un lematizador para normalizar las palabras

Tabla 2. Resultados con la selección de rasgos utilizando eficacia de las palabras

Número de oración	Grupo correcto	Grupo asignado	Valor de la similitud
1	5	5	0.04
2	6	6	0.04
3	3	3	0.01
4	2	-	0.00
5	4	4	0.01
6	-	-	0.00
7	1	5	0.02
8	4	4	0.01

Tabla 3. Resultados con la selección de rasgos utilizando la solución SR

Número de oración	Grupo correcto	Grupo asignado	Valor de la similitud
1	3	3	0.43
2	3	3	0.25
3	3	3	0.5
4	2	3	0.25
5	3	3	0.17
6	-	3	0.33
7	1	3	0.2
8	3	3	0.2

que pertenecen a una misma familia morfológica, con ello se espera conseguir que el algoritmo de agrupamiento obtenga mayor eficacia al formar los grupos. Por ahora, están siendo probados los tres algoritmos de agrupamiento mencionados. Es importante destacar que el proceso de agrupamiento es complicado, debido a que existen realmente pocas palabras en común para aquellos contextos que se refieren a un mismo sentido, sin embargo, el algoritmo MOD-SLC ha mostrado resultados alentadores en el proceso de agrupamiento de contextos y se pretende mejorar esta técnica mediante la adición de heurísticas de refinamiento, como es el caso del recocido simulado. Finalmente, se tiene planeado usar la base de datos léxica en un sistema de recuperación de información para conocer la efectividad del recurso generado.

Referencias

1. Masterman, Margaret., "Semantic message detection for machine translation using interlingua", Stationery Office, London, pp. 437-475, 1962.
2. Wilks, Y. and Fass D., "Preference semantics: A family history", *Report MCCS*, pp. 90-194, 1990.
3. Wilks, Y. and Stevenson M., "The grammar of sense: Is word sense tagging much more than part-of-speech tagging?", *Technical report CS-96-05*, University of Sheffield, Sheffield, UK, 1996.

4. Wilks, Y. and Stevenson M., "Sense Tagging: Semantic tagging with a lexicon", 1996.
5. Lesk, Michael., "Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone", *SIGDOC Conference*, pp. 24-26, 1986.
6. Pedersen, T. and Banerjee, S., "An adapted Lesk Algorithm for Word Sense Disambiguation using WordNet", *CICLing 2002*, pp. 136-145, 2002.
7. Ide, N., and Veroni J., "Introduction to the Special Issue on Word Sense Disambiguation", *Computacional Linguistics*, Vol. 24, Number 1, pp. 1-29, 1998.
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, 2003.
10. Hassan M. S., Reda M. S. A. A., Araby I. M., "Formation of Machine Groups and Part Families: A modified SLC Method and Comparative Study", *Integrated Manufacturing Systems*, pp. 123-137, 2003.
11. Sneath, P.H., "The application of Computers to Taxonomy", *Journal of General Microbiology*, Vol. 17, pp. 201-226, 1957.
12. Sneath, P.H., "Some Thoughts of Bacterial Classification", *Journal of General Microbiology*, Vol. 17, pp. 184-200, 1957.