Exploración sobre la construcción automática de un tesauro a partir de un documento

Aarón Ramírez-De-la-Cruz, Héctor Jiménez-Salazar, Esaú Villatoro-Tello, Gabriela Ramírez-De-la-Rosa

Universidad Autónoma Metropolitana Unidad Cuajimalpa, Departamento de Tecnologías de la Información, Ciudad de México, México

{aaron.rc24, hgimenezs, villatoroe, a.gaby.rr}@gmail.com

Resumen Es común el problema de requerir el acceso a la información contenida en un texto con un vocabulario más alla del léxico general (por ejemplo, reglamentos o leyes, contratos, e instrucciones de cuidado, entre otros) cuando no se cuenta con diccionarios, glosarios, terminología, o tesauros, todos ellos de un dominio especial. En este trabajo se explora la generación automática de un tesauro a partir del texto de un documento semiestructurado, para establecer un acercamiento a los componentes de este proceso e iniciar un análisis sobre las variables que influyen en la extracción de una parte de la semántica de textos de dominio particular. Se aplicó una adaptación del método SEXTANT a varios textos de dominio especial para generar un tesauro. La revisión de las parejas de términos relacionadas y el texto original nos llevan a concluir una formulación que relaciona las características del texto y productividad del método.

Palabras clave: Tesauro, extracción semántica, estilo de texto.

Explorations on Automatic Thesaurus Construction from a Single Document

Abstract. It is common the need to access to the information at documents which use vocabulary far from the usual lexicon (for example laws, agreements, instructions, etc.) when there is no dictionaries, glossaries or thesauri, all of this fitting for a special domain. In this work it is explored the automatic thesauri generation from the text coming of a semi-structured document. In order to establish a view of the components of this process and begin an analysis about the variables that affect on the extraction of a portion of semantic in texts belonging to a specific domain, it was applied an adaptation of the SEXTANT method to several texts from different domain in order to automatically build, for each one, an specific thesaurus. The review of the related pairs of terms and the source text makes us to conclude the existence of a relation between text characteristics and the productivity of the method.

Keywords: Thesauri, semantic extraction, text style.

1. Introducción

En la actualidad se tiene acceso a grandes cantidades de información textual, mas no contamos con herramientas capaces de aprovechar esta información; en parte, por carecer de recursos lingüísticos suficientes. En efecto, podemos acceder, mediante una búsqueda, a una copiosa masa de documentos, pero en muchos casos es difícil con ellos dar respuesta a una pregunta específica. La navegación en documentos, la búsqueda de respuestas y diferencias sutiles entre documentos, son algunos de los problemas que enfrenta cotidianamente la sociedad del conocimiento y también las motivaciones del presente trabajo.

Una desventaja que tienen los recursos lingüísticos de propósito general es que pueden presentar ambigüedad cuando son aplicados a tareas que tratan con lenguaje o términos de un área específica. Por ejemplo, el término *celular* tiene una acepción si se encuentra en un texto de biología y una diferente cuando se encuentra en un texto de telecomunicaciones. Una forma para desambiguar el significado de los términos es tomar en cuenta su contexto, ya que puede aportar pistas sobre la función de las palabras. Sin embargo ello requiere, a su vez, de recursos lingüísticos para su procesamiento.

Un tesauro es una base de datos léxica que organiza términos de uso común o de un dominio particular. Cada entrada en un tesauro está acompañada por una serie de términos con los que mantiene una o más relaciones; algunas de ellas pueden ser de sinonimia, hiperonimia, hiponimia, entre otras. WordNet [1] es la conocida base de datos léxica en la que las palabras están agrupadas por su significado equivalente en estructuras llamadas *synsets*. Si bien los grupos que constituyen WordNet establecen relaciones semánticas, éstas son generales y no necesariamente coinciden con las usadas en el lenguaje de diferentes áreas de conocimiento.

Normalmente, para construir un tesauro se procede partiendo de un corpus grande de especialidad, y con él se genera el tesauro para después procesar textos de un dominio. Por esta razón hay diversos y eficientes algoritmos que automáticamente generan un tesauro [2], los cuales utilizan grandes cantidades de texto para ser aplicados. Por otro lado, es común el problema de requerir el acceso a la información contenida en un texto con un vocabulario más alla del léxico general (por ejemplo, reglamentos o leyes, contratos, e instrucciones de cuidado, entre otros) cuando no se cuenta con diccionarios, glosarios, terminología o tesauros, todos ellos de un dominio especial. Nuestro objetivo es conocer el alcance de la extracción semántica sobre un documento mediante el análisis de los parámetros que intervienen en él, con la finalidad de asegurar un nivel de precisión sobre el recurso generado.

En este trabajo se explora la generación automática de un tesauro a partir del texto de un documento semiestructurado, para establecer un acercamiento a los componentes de este proceso e iniciar un análisis sobre las variables que influyen en la extracción de una parte de la semántica de textos de dominio particular. Específicamente, se programó un método de extracción de pares de términos relacionados. El método está basado en conocimiento sintáctico [3], lo cual permite mejorar la selección de los contextos de los términos cuando no se

cuenta con un texto grande. Tanto la selección de los textos fuente para construir el tesauro como las evaluaciones de las parejas fueron actividades apoyadas por expertos en dominios específicos. Una vez que se obtuvieron los resultados se llevó a cabo un análisis para conocer la utilidad de este enfoque.

El resto del documento está organizado de la siguiente forma. En la sec. 2 se describen las características principales de trabajos orientados a la construcción automática de tesauros. La sec. 3 se dedica a exponer la adaptación hecha para el español del método SEXTANT. La descripción de las condiciones creadas para la aplicación del método adaptado y los resultados obtenidos se presentan en la sec. 4. Finalmente se lleva a cabo un análisis de los factores que influyen en la extracción de los términos relacionados a partir de un texto.

2. Trabajo relacionado

Los métodos más usados para generar un tesauro reúnen grandes colecciones de texto; generalmente se trata de tesauros de uso, como el caso de [2] o [4] que se apoyan en el British National Corpus (100M de palabras), los cuales son métodos impracticables en nuestro caso debido a que los documentos considerados en este trabajo no pasan de 10 mil palabras. Brevemente, esta sección se refiere a métodos que pueden verse complementarios al que aquí se aplicó.

En [5] proponen la construcción automática de tesauros basada en conceptos formales. La motivación de utilizar este enfoque es debido a que las representaciones de texto tradicionales basadas en un modelo vectorial ignoran las relaciones conceptuales entre términos, como pueden ser hiperonimia e hiponimia. Las pruebas de este trabajo fueron realizadas con el texto de un documento [6] con temática de expansión y mejora de consultas en sistemas de recuperación de información.

El análisis de conceptos formales es utilizado para la extracción de relaciones entre términos dentro de un contexto. Un concepto formal es definido como una tupla de tres elementos, (G, M, I), donde G son objetos, M son atributos y I es la relación binaria entre G y M. Para establecer relaciones jerárquicas entre conceptos utilizan una retícula, en la que cada nodo representa un concepto y las aristas son las relaciones de superconcepto y subconcepto. De esta forma, la relación superconcepto y subconcepto entre conceptos formales juegan roles de hiperónimo-hipónimo, y los que se encuentran un nivel abajo del mismo superconcepto pueden considerarse cohipónimos. La generación del tesauro consiste en una retícula de conceptos relacionados entre sí.

El trabajo de [7] está enfocado en la construcción de una ontología de términos legales. El corpora utilizado son 57 códigos pertenecientes a leyes francesas. Mediante un analizador sintáctico se identificaron 500 000 términos, que incluyen sustantivos, verbos, adverbios, adjetivos y las dependencias sintácticas entre éstas categorías gramaticales (e.g. sujeto de un verbo, objeto de un verbo, adjetivo de un sustantivo). Para identificar las relaciones entre los términos se

utiliza la información mutua (IM) entre palabras, usando la siguiente fórmula:

$$IM_{cw} = log\left(\frac{f_{cw}}{f_c f_w} + 1\right), \tag{1}$$

donde c son las palabras del contexto del término; w son los términos base; f_{cw} es la probabilidad conjunta de c y w en el corpus; f_c y f_w las frecuencias individuales de c y w, respectivamente. El resultado final es una lista de 103 994 términos, cada uno relacionado al menos a otro término.

Como se ha dicho, en este trabajo se eligió el método SEXTANT [3], el cual consideramos viable por no exigir un corpus grande; en cambio, el método se apoya en el análisis sintáctico para extraer contextos de los términos y encontrar parejas relacionadas. Dicho método será presentado en la siguiente sección a través de una adaptación al español.

3. Descripción del método

El método SEXTANT (Semantic Extraction from Text Via Analyzed Network of Terms, Extracción semántica de texto mediante análisis de términos relacionados) [3] emplea el contexto de las palabras para descubrir similitudes entre ellas. Se basa en la hipótesis de que las palabras que son usadas en un contexto similar a lo largo de un corpus de texto están relacionadas semánticamente. Dicho método está presentado en el libro de Grefenstette [3] apoyándose en reglas de la sintaxis inglesa.

El texto utilizado como corpus es dividido por oraciones que terminan con un punto (.). Dado que este signo no es exclusivo del final de una oración, se utilizó una lista de abreviaturas que no son consideradas como fin de línea (e.g., i.e., etc., Art., Vol., Dr., etc.). Las cifras numéricas con decimales (e.g. 3.14159) son ignoradas. Este conjunto de abreviaturas puede ser complementado con más elementos. Posteriormente cada oración es etiquetada con las partes del discurso (Part-Of-Speech) para conocer la categoría gramatical de las palabras (sustantivos, verbos, adjetivos, adverbios, etc.), utilizando el etiquetador Tree-Tagger [8]. En la siguiente descripción se presenta la adaptación hecha con reglas gramaticales del español.

El procedimiento de SEXTANT se realiza en cinco etapas secuenciales. La unidad que utiliza son oraciones simples, formadas por un sujeto, un verbo y el complemento.

Etapa 1: verbo principal. Se parte del hecho que una oración simple posee sólo un verbo conjugado [9]. Tomando en cuenta que en el proceso de etiquetado de texto puede haber casos de asignación incorrecta de categorías, se previeron cuatro casos para detectar dichos errores:

1. una conjunción o palabra clítica antes del verbo: *que le presenten, tuvieron que caminar*;
2. el verbo está en infinitivo, gerundio o pasado participio y está después de un sustantivo: *decisión tomada*;

3. antes del verbo aparece un artículo: *la solicitud, el cantar de las aves*;
4. una negación antes del verbo: *no indica, ningún proceder.*

Si una palabra fue etiquetada como verbo y cae en uno de los casos anteriores, se descarta como verbo principal. También mediante este procedimiento las oraciones compuestas (coordinadas o subordinadas) son segmentadas y conservadas como oraciones simples.

Etapa 2: sintagmas nominales. La estructura utilizada para capturar los sintagmas nominales es *Determinante + Núcleo + Complemento de la oración.* En la Tabla 1 se muestran las categorías gramaticales para cada componente de la estructura anterior. El núcleo puede ser secuencias con más de un sustantivo, como en el caso de las entidades nombradas.

Tabla 1. Categorías gramaticales permitidas por componente

Componente	Tipo	Ejemplo
Determinante	Artículo, demostrativo, posesivo, numeral, interrogativo, exclamativo.	los, esa, mi, tres, cada, cuántos, ¡qué!
Núcleo	Sustantivo propio, sustantivo común.	entidad, Universidad
Complemento	Complemento del sustantivo	fue velozmente

Etapa 3: sujeto principal. Utilizando los sintagmas detectados en la Etapa 2, el sujeto será el núcleo (un sustantivo o más). Si existen adjetivos después del núcleo, éstos se agregan al sujeto debido a que pueden formar parte de él.

Etapa 4: contexto del sujeto principal. Si del sintagma original de la Etapa 2 se omite el verbo principal (Etapa 1) y el sujeto principal (Etapa 3), el contenido restante es el contexto del sujeto principal. Este conjunto puede incluir sustantivos (propios y comunes), verbos y secuencias de sustantivos con adjetivos.

Etapa 5: cálculo de similitud contextual. La similitud entre dos sujetos S_i y S_j es calculada utilizando el contexto c_i y c_j, respectivamente. Se utiliza el coeficiente Jaccard, definido como:

$$sim(S_i, S_j) = sim(c_i, c_j) = \frac{c_i \cap c_j}{c_i \cup c_j}. \tag{2}$$

Una vez que se calcularon las similitudes de todos los sujetos, se ordenan de manera descendente y se presentan en el formato $[sim(S_i, S_j)]$ $[S_i]$ $[S_j]$
Este procedimiento es del orden $O(n^2)$, donde n es el número de sujetos en el corpus utilizado.

Finalmente para construir el archivo de tesauro, se toman los primeros pares de cada sujeto. Por defecto se eligieron 10, aunque éste puede cambiarse.

4. Experimento

El experimento descrito en esta sección está orientado a identificar las variables que influyen en la cantidad de parejas de términos relacionados. Para ello se tomó de cada uno de los textos utilizados tres porcentajes para generar términos relacionados. Una vez validadas las parejas de términos se analizaron las posibles variables que influyen en la producción del método.

Se describen a continuación los datos utilizados, la aplicación del método presentado en la sección anterior, asimismo, los resultados obtenidos.

4.1. Características de los documentos

Los documentos pertenecen a tres dominios diferentes: pedagogía, procesamiento del lenguaje natural, y psicología computacional. Dos de los cuatro autores proporcionaron su documento en formato de texto plano; los otros dos autores entregaron su documento como archivo PDF de los cuales se extrajo el texto y se almacenó como texto plano. La Tabla 2 muestra el tamaño de cada texto empleado para realizar el experimento.

Tabla 2. Características de los documentos

Texto	Número de oraciones	Número de palabras	Promedio de palabras por oración
1	159	5069	32.68
2	205	6124	30.68
3	197	5189	27.32
4	266	10014	38.64

4.2. Descripción del experimento

Previo a aplicar el método SEXTANT para construir el tesauro, de cada archivo de texto se eligieron aleatoriamente el número equivalente al 50 % y 75 % de las oraciones totales. Se generaron tres tesauros, uno para cada porcentaje de texto utilizado de los documentos (incluyendo el 100 % de las oraciones).

Para efectos de esta sección, se establecen las siguientes definiciones:

- **pareja:** relación semántica establecida entre los sujetos (S_i, S_k), en donde S_i y S_k $(k \neq i)$ son entradas del tesauro;
- **pareja útil:** la relación semántica (S_i, S_k) es válida;
- **grupo:** conformado por 10 parejas, de la forma *[(S_i,S_1), (S_i,S_2), (S_i,S_3), ..., (S_i,S_{10})]*

Para este experimento, los elementos del tesauro se ordenaron de forma decreciente con base en la suma de la similitud entre las parejas de cada grupo, por lo que las primeras entradas del tesauro son aquellas cuya similitud acumulada es mayor. De cada tesauro ordenado se extrajeron nueve grupos de 10 parejas cada uno y se presentaron a los respectivos autores para que indicaran la utilidad de las parejas.

Se entregó a los autores tres archivos de texto (uno para cada porcentaje de texto utilizado), cada uno con nueve grupos, correspondientes al tesauro generado a partir del documento que proporcionaron. Se muestra en la Fig. 1 el formato utilizado para que los autores marcaran las parejas que constituyen una relación semántica válida. En dicho formato de ejemplo, `término 1` y `término 3` son relaciones válidas con `entrada`.

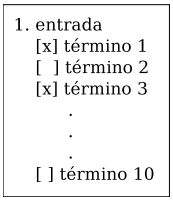

Fig. 1. Formato de un grupo de parejas

4.3. Resultados obtenidos

En la gráfica de la Fig. 2 aparece en el eje horizontal el porcentaje de texto utilizado y en el eje vertical, el número de parejas válidas identificadas por los autores.

5. Análisis de resultados

Los resultados resumidos en la Fig. 2, presentan dos textos que tienen líneas crecientes, uno línea decreciente, y otro no monótona. En el último caso, se detectaron problemas con el preprocesamiento, aún así se consideró útil su análisis. El texto que se comporta decreciente consideramos que pudo influir su sesgo multitématico. De esta manera, no en todos los casos las parejas aumentan al incrementar el tamaño de texto, pero sí hay una tendencia de aumento; lo cual, como veremos se relaciona con otros factores.

Por otro lado, la revisión de las parejas que produjo el método fue hecha tomando muestras de tres niveles de similitud: alta, media y baja. Dichas muestras se realizaron en cada uno de los tamaños de los textos de prueba: 50 %, 75 % y 100 %. En la Tabla 3 se muestra, para cada texto, el número de parejas válidas en cada nivel de similitud, además, la mínima similitud de las parejas que fueron

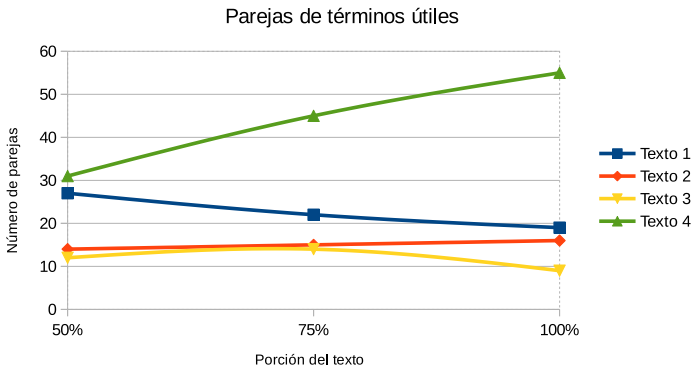

Fig. 2. Parejas útiles

elegidas por los revisores. Lo que se puede observar en esta tabla es que al variar el tamaño de texto (50 %, 75 % y 100 %) el mayor número de parejas válidas se concentra en el nivel alto de similitud. Esto permitiría establecer un umbral para elegir parejas válidas; en el caso de los textos usados en el experimento la similitud mínima promedio de las parejas elegidas en el nivel alto es 0.72.

Tabla 3. Similitud mínima por nivel

Nivel	Texto 1	Texto 2	Texto 3	Texto 4
Alto	28 (0.55)	21 (0.84)	18 (0.53)	51 (0.98)
Medio	22 (0.15)	14 (0.53)	11 (0.42)	40 (0.13)
Bajo	18 (0.03)	10 (0.08)	6 (0.11)	40 (0.05)

En lo que sigue hacemos un análisis sobre el léxico de los textos con el fin de extraer elementos que permitan caracterizar la productividad del método. Como se ha explicado, el método construye las parejas sacando provecho de la sintaxis del texto. En particular, el uso de la puntuación tiene influencia en la definición de contextos puesto que delimita sintagmas. En los textos utilizados se pudieron constatar diferencias sobre el uso de la puntuación. Bien que este elemento sintáctico sea parte del estilo del autor, el resultado conduce a que efectivamente deben adecuarse las reglas sintácticas utilizadas para identificar sintagmas en las cuales la puntuación no sea decisiva.

Asimismo, puesto que el método se basa en la similitud léxica de contextos, la cohesión del texto influye en los valores de similitud entre pares de palabras y en la selección de éstas para ser parte del tesauro. Por ejemplo, como el uso de pronombres constituye un elemento de cohesión, se espera que los textos que utilicen más pronombres tengan menos parejas y viceversa. En un conteo del

uso de pronombres en cada texto se observó que aquél con mayor proporción de pronombres tiene mayor cantidad de parejas. Es decir, las diferencias observadas en la obtención de parejas útiles se debe a otros factores (por ejemplo, la repetición de palabras). Por último se calcularon algunas medidas para precisar la influencia del estilo sobre los resultados del método. Se determinó para cada texto T la cantidad de términos con una sola ocurrencia, $H(T) = \{x | frec(x, T) = 1\}$; la ponderación de términos de ocurrencia unitaria por tamaño de vocabulario, $PH(T) = |H(T)|/|V(T)|$, donde $V(T)$ es el vocabulario de T; la suma del número de oraciones que utiliza cada uno de los términos, $R(T)$; y la proporción de $R(T)$ con respecto al número de oraciones, excluyendo palabras cerradas, $Rec(T)$. Estas medidas se presentan en la siguiente tabla, al igual que el número de parejas producidas por el método para cada texto.

Tabla 4. Medidas de los textos utilizados

T	$\|V(T)\|$	$\|H(T)\|$	$PH(T)$	$R(T)$	$Rec(T)$	#Parejas
1	1486	969	0.65	2182	13.7	20
2	1542	937	0.60	2611	12.7	17
3	1337	1144	0.68	2179	11.0	10
4	2359	1012	0.72	3954	14.8	55

Puede notarse que hay una proporción directa entre la relación de repetición por oración, y el número de términos de ocurrencia unitaria normalizado con el número de parejas válidas: hay mayor productividad del método para los textos con estas características estilísticas.

También es cierto que el aumento de parejas útiles podría ser mejorado con la resolución de anáfora. En suma, con los textos de prueba utilizados hay variación de los resultados que proporciona el método frente a diversos estilos de escritura.

6. Conclusiones

Se ha aplicado una adaptación del método SEXTANT a varios textos para generar un tesauro. Asimismo, se llevó a cabo un análisis para conocer los factores que influyen en la cantidad de las parejas obtenidas. Las pruebas se realizaron con textos de diversos dominios, y de tamaño limitado por 10 000 palabras (en promedio de 6 500 palabras); lo cual es una restricción realista para aplicaciones donde se requiere información semántica del texto para, por ejemplo, hacer consultas sobre su contenido. El análisis de los resultados lleva a concluir que el método tiene sensibilidad al estilo de escritura y que es posible extraer una parte de la semántica del texto representada por parejas de términos relacionados.

Agradecimiento. Deseamos agradecer el apoyo brindado por la Coordinación de la Licenciatura en Tecnologías y Sistemas de Información de la UAM-C.

Asimismo, reconocemos la gentil colaboración de los jueces que participaron en la revisión de las parejas para poder realizar los experimentos aquí presentados, en particular al Dr. Tiburcio Moreno Olivos.

Referencias

1. Miller, G.: WordNet: A Lexical Database for English. Communications of the ACM 38, pp. 39–41 (1995)
2. Curran, J., Moens, M.: Improvements in Automatic Thesaurus Extraction. In: Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition, vol. 9, pp. 59–66 (2002)
3. Grefenstette, G.: Explorations on Automatic Thesaurus Discovery. Kluwer Academic Publishers (1994)
4. Rychlý, P., Kilgarriff, A.: An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, pp. 41–44 (2007)
5. Jehng, J., Shihchieh, C., Cheng, C.: A Formal Concept Analysis-Based Domain-Specific Thesaurus and Its Application in Document Representation. In: Taniar, D., Gervasi, O., Murgante, B., Pardede, E., Apduhan, B. (Eds.): Computational Science and Its Applications, vol. 6018, Springer, pp. 431–442 (2010)
6. Xu, J., Croft, W.: Improving the Effectiveness of Information Retrieval with Local Context Analysis. In: ACM Transactions on Information Systems, vol. 18, ACM, pp. 79–112 (2000)
7. Lame, G.: Using NLP Techniques to Identify Legal Ontology Components: Concepts and Relations. In: Benjamins, V., Casanovas, P., Breuker, J., Gangemi, A. (Eds.): Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications, Springer, pp. 169–184 (2005)
8. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of the International Conference on New Methods in Language Processing, pp. 44–49 (1994)
9. Cohen, S.: Redacción sin dolor. Planeta (2010)