Comparación de dos métodos que determinan automáticamente el extracto de un texto

Hilario Salazar-Martínez, David Pinto, Héctor Jiménez-Salazar

Facultad de Ciencias de la Computación
Benemérita Universidad Autónoma de Puebla
C.U. 72570, Puebla, México
hilario_cam@yahoo.com.mx, dpinto@cs.buap.mx, hjimenez@fcfm.buap.mx

Abstract. In this paper a comparison between two automatic text extraction corpus-based methods is presented. Each method determines the most representative sentences by assigning a score to each sentence on a plain text. The second method uses the text body to calculate the score, and the first method uses only the text head. Despites the first method uses less information, in our test, it got similar results than the second one. Therefore, in such cases that the text head is available, it is preferable to use the first method, due to, its complexity time. In this case, the first method is in $O(n)$ and the second method is in $O(n^2)$; n is the number of sentences in a plain text.
Key Words: Automatic Text Extraction, Mutual Information, Sense Relationship

Resumen. Este trabajo compara dos métodos apoyados en un corpus para obtener el extracto de un texto. Ambos métodos asignan una puntuación a cada oración para determinar las oraciones más representativas. El primer método se apoya en el encabezado del texto para calcular la puntuación. El segundo, utiliza el cuerpo del texto para calcular su puntuación. Aunque el segundo método utiliza más información, en nuestras pruebas, obtuvo un desempeño semejante al primero. Así, resulta conveniente utilizar el primero, ya que su comportamiento, con un texto de n oraciones, está en la clase $O(n)$, mientras que el segundo en $O(n^2)$.
Palabras Clave: Generación Automática de Extractos, Relaciones de sentido.

1 Introducción

Los volumenes de información cada vez son mayores y con el surgimiento de Internet el manejo de éstos cobró mayor importancia, por lo que se hace necesaria la búsqueda de nuevas formas que ayuden a comprender eficazmente el contenido de un documento, sin tener que leer completamente el mismo. Por generación automática de resumen de un texto se entiende el proceso por el cual se identifica la información sustancial, proveniente de una fuente (o varias) para conocer de qué trata un documento sin necesidad de leerlo completamente, y producir una

versión abreviada destinada a un usuario particular (o grupo de usuarios) y a una tarea (o tareas) específica.

Existen muchos enfoques y desarrollos para resolver el problema de la generación del resumen automático. En algunos enfoques se han generado diversas herramientas, algunas de las cuales hacen uso de recursos lingüisticos grandes. Otros más emplean recursos semánticos y otros métodos estadísticos[4, 12]. En general, la mayoría de los métodos se basan en recursos copiosos.

Daniel Marcu [3] propuso un algoritmo que construye automáticamente un corpus para entrenar sistemas que determinan el extracto de un texto. Este algoritmo toma un conjunto de tuplas como entrada (Resumen, Texto) y genera el extracto correspondiente, es decir, el conjunto de cláusulas (oraciones) en el texto utilizado para escribir el correspondiente resumen. Para el desarrollo del algoritmo se realizó un experimento que fue evaluado por jueces. El experimento también sugiere estrategias de extracción para mejorar el desarrollo del sistema de resumen automático.

Debe reconocerse que la generación automática del resumen de un texto requiere mayor cantidad de recursos. Por ejemplo, el trabajo de Gustavo Crispino et al [7] propone una plataforma para resolver problemas de extracción de información y resumen automático. Este trabajo propone la definición de un modelo conceptual general de representación para conocimientos lingüisticos y el desarrollo de tareas especializadas que cooperen entre sí. En el modelo se identifican los conocimientos lingüisticos ubicándolos en sus contextos y organizándolos en tareas especializadas. Presenta, por un lado, la ventaja de permitir que el trabajo del lingüista se realice de manera independiente de su implementación informática, y, por otro, la de articular efectivamente en un mismo sistema informático los dos tipos de tareas.

Existen muchos trabajos que se apoyan en modelos de representación bien conocidos como el vectorial [5]. Así, pueden generarse resúmenes personalizados [8], cuyo fin es proveer al usuario de información contenida en textos de acuerdo con sus intereses.

Hemos abordado el problema de la generación automática de extractos mediante un algoritmo que ordena las oraciones contenidas en un texto según su similitud con el texto mismo [9]. El algoritmo está en $O(n^2)$ donde n es el número de oraciones del texto. Las extractos obtenidos fueron evaluados por cuatro jueces humanos y arrojaron resultados prometedores. En el presente trabajo exponemos una variante inspirada en el trabajo de Daniel Marcu [3], que permite reducir el tiempo de cálculo de $O(n^2)$ a $O(n)$ sin detrimento de su eficacia.

La hipótesis de que una oración O_i de un texto T es más representativa que otra O_j, si la similitud entre O_i y T es mayor que la similitud entre O_j y T, conlleva a la experimentación con diversas técnicas a fin de obtener un conjunto de oraciones que puedan considerarse como el extracto de un texto. Cabe mencionar que el algoritmo determina el extracto de textos planos (sin estructura, etiquetas, u otra información).

En la sección 2 de este trabajo se introducen dos conceptos importantes para el entendimiento del trabajo desarrollado: el uso de la relación de sentidos y el

uso de la medida de información mutua. El método mejorado se expone en la sección 3. En la sección 4 se presenta una comparación entre los dos métodos mencionados usando 40 textos del dominio "política". En la parte final de este artículo se presentan las conclusiones del trabajo.

2 Conceptos Básicos

El recurso en que se apoya el método que se presenta es el conjunto de relaciones de sentido de un término. Las relaciones del sentido de una palabra x, es el conjunto de todas aquellas palabras relacionadas con x [2]. En este caso, aproximamos las relaciones de sentido por los términos que ocurran en los contextos de x. Así, se están usando los términos de primer orden para representar x [1, 6]. Se han realizado otras aplicaciones que utilizan esta representación [13].

Para obtener el extracto de un texto, identificamos las oraciones más representativas de él. Para este fin, usamos las relaciones de sentido y "expandimos" los términos de cada oración del texto; esto último permite aplicar una función de similitud que califique cada oración con un nivel de representatividad.

Por otro lado, consideramos importante refinar la aproximación a las relaciones de sentido que provee el corpus, con la finalidad de incrementar la precisión de los resultados. Así, hemos hecho uso del concepto de información mutua. La información mutua (IM) [11] proviene de conceptos teóricos sobre la información de sistemas y finalmente se traduce a una medida. La IM es entonces un cociente de asociación, para medir la norma de asociación de palabras.

Si dos palabras, x e y, tienen probabilidades $\Pr(x)$ y $\Pr(y)$, entonces su información mutua, $I(x, y)$, se define como:

$$I(x, y) \equiv \log_2 \frac{\Pr(x, y)}{\Pr(x) \Pr(y)} \tag{1}$$

En este caso, las probabilidades $\Pr(x)$ y $\Pr(y)$ son estimadas calculando la frecuencia de ocurrencia relativa de x e y en un texto. La probabilidad conjunta, $\Pr(x, y)$, es estimada calculando el número de veces que ocurren x e y en la misma oración y dividiendo por el número oraciones del corpus (N). Considerando que la frecuencia conjunta de términos sea cero, la fórmula 1 queda reexpresada como:

$$IM(x, y) = \log_2 \left(\frac{N \cdot fr(x, y)}{fr(x) \cdot fr(y)} + 1 \right), \tag{2}$$

donde $fr(x)$ y $fr(y)$ son la frecuencia de ocurrencia del término x e y, respectivamente, y $fr(x, y)$ es la frecuencia de ocurrencia en una misma oración de los términos x e y.

3 Métodos de Extracción

Nuestro enfoque parte del supuesto de que el encabezado de un texto ofrece una buena aproximación al lector acerca del contenido del mismo y, por lo tanto,

éste podría representar al texto [10]. Valiéndose de una función de similitud, se buscan entonces las oraciones O_i que tengan mayor similitud con el encabezado. Como puede verse, para determinar la puntuación de todas las oraciones, se realizan n cálculos de similitud para un texto de n oraciones. En tanto, si para calcular la puntuación de una oración O_i se realizan n cálculos de similitud, usando el cuerpo del texto, el algoritmo tendría un comportamiento cuadrático para asignar la puntuación a las n oraciones del texto.

Con la idea anterior se propuso realizar dos experimentos, en los que se hace uso de un corpus C (en este caso del dominio "política") que permite representar los términos. Tanto C como el texto T deben ser preprocesados, C_1 y T_1 respectivamente; es decir, se eliminan las palabras cerradas, y se trunca cada término. Además, T_1 se descompone en su encabezado T_e y su cuerpo T_c. El corpus C_1 se emplea para representar cada una de las oraciones con los términos de asociación de primer orden. Así, se obtiene lo que llamamos "oración representada". La similitud en cada una de las pruebas se calcula con las oraciones representadas. A continuación se precisan las ideas antes expuestas.

3.1 Representación de Texto

Las relaciones de sentido con respecto al corpus C_1 es un conjunto de pares [9]:

$$V = \{(x, y) | x, y \in O, \text{ para una oración } O \in C_1\}. \tag{3}$$

La representación de una palabra x con base en V es:

$$\text{Expand}(x, V) = \{y | (x, y) \in V \wedge IM(x, y) > 5\}. \tag{4}$$

La representación de una palabras es vista como un conjunto de palabras (sin repeticiones). A partir de la representación de una palabra, es posible representar una oración formando una tupla. Corresponde a $O = (x_1, \ldots, x_k)$ la representación:

$$\text{ExpandO}(O, V) = (\text{Expand}(x_1, V), \ldots, \text{Expand}(x_k, V)). \tag{5}$$

La representación, T_2, del texto T_1 es, entonces:

$$T_2 = (\text{ExpandO}(O_1, V), \ldots, \text{ExpandO}(O_n, V)). \tag{6}$$

donde T_1 está formado por $(O_1, O_2, \ldots, O_n)$.

3.2 Extracto del Texto

Veamos ahora los métodos de extracto automático de un texto. El primero emplea el encabezado para determinar la puntuación de cada oración del cuerpo y, el segundo usa el mismo cuerpo para calcular esta puntuación.

A: Encabezado. Para calcular la puntuación de cada oración del texto, calculamos la similitud entre el encabezado del texto T, T_e, y su cuerpo T_c. En esta tarea utilizamos el coeficiente de Jaccard. Lo anterior considerando la representación de cada una de estas oraciones:

$$sim(T_{e2}, O_{ic2}) = \frac{\#(T_{e2} \cap O_{ic2})}{\#(T_{e2} \cup O_{ic2})}, \tag{7}$$

donde O_{ic2} es la i-ésima oración representada del cuerpo de T_2 y T_{e2} es el encabezado expandido de $T2$.

B: Complemento. Este método se basa en el complemento de una oración [9], con respecto al texto, para elegir las oraciones más representativas. La puntuación de una oración se calcula con la siguiente simplificación del coeficiente de Jaccard:

$$sim'(O_{ic2}, \bar{O}_{ic2}) = \#(O_{ic2} \cap \bar{O}_{ic2}). \tag{8}$$

Aquí, sim' cuenta el número de elementos comunes de la oración i y su complemento en T_2, $\bar{O}_{ic2}$; es decir $\bar{O}_{ic2}$ es T_2 sin O_{ic2}.

4 Comparación de los métodos

El corpus utilizado en las pruebas realizadas consta de 375 textos, todos del dominio "Política". Dicho corpus tiene un tamaño de 2.7 kb, con 21559 pseudolexemas (considerando nombres propios) y 29741 oraciones. Cabe mencionar que los 40 textos usados en la prueba, no son parte del corpus.

La comparación de los métodos antes descritos se llevó a cabo tomando como punto de referencia el texto original. En esencia, se comparan las posiciones de las oraciones de un extracto (Ext) con las posiciones de esas oraciones en el texto original. Así, cuando las posiciones de las oraciones extraídas por cada método son cercanas con respecto al texto original, entonces su representatividad es similar, y podemos concluir que los métodos obtienen resultados semejantes. Con esta idea se aplicaron funciones de cercanía (Cer_{ora}), a cada oración de Ext con cada $O_i \in Ext$. Esta medida se basa en la diferencia de la posición i de $O_i \in Ext$ con la posición j de O_i en T_c. En suma, la valoración de una oración del extracto con respecto del texto original es:

$$Cer_{ora}(T_c, O_i) = \frac{1}{|j - i| + 1}. \tag{9}$$

La evaluación del extracto del documento se obtiene promediando:

$$Cer_{doc}(T_c) = \frac{\Sigma_{i=1}^{5} Cer_{ora}}{5}, \tag{10}$$

donde 5 fue un número arbitrario de oraciones que se tomaron para definir Ext.

En la siguiente tabla se muestran los resultados de promediar la evaluación de cada extracto de las 40 oraciones aplicando los métodos A y B.

Tabla 1. Semejanza relativa de los métodos

Método	Cer_{ora} de Ext					Cer_{doc}
	O_1	O_2	O_3	O_4	O_5	
A	**0.190**	**0.126**	**0.172**	**0.211**	**0.198**	**0.179**
B	**0.235**	**0.184**	**0.151**	**0.133**	**0.173**	**0.175**

En la tabla 1 observamos un valor promedio de cercanía semejante para A y B. De la misma forma se calculó la semejanza entre los extractos (ec. 9). El resultado de este cálculo fue 0.66, lo cual nos permite afirmar que la eficacia de ambos métodos es similar y, por tanto, puede usarse con mayor ventaja el método A.

5 Conclusiones

Se compararon dos métodos para obtener el extracto de un texto. Ambos métodos usan un corpus con el fin de representar las oraciones y, así, asignarles una puntuación sobre su representatividad acerca del texto. El método A requiere un tiempo en $O(n)$ y el método B, $O(n^2)$, dado un texto de n oraciones. Que el encabezado sea revelador del contenido del texto es una hipótesis del método A. Mientras que B no depende del encabezado. Al aplicar una medida de cercanía con respecto al texto original fue posible constatar, en una prueba que usó cuarenta textos, que ambos metódos obtienen extractos similares. Se concluye, entonces, que el método basado en el encabezado del texto es ventajoso.

Referencias

[1] Ruge, Gerda: Combining corpus linguistics and human memory models for automatic term association, *Text information retrieval*, T. Strzalkowski (Ed.), Kluwer, 1999.

[2] Lyons, J.: *Semantics*, Cambridge University Press, 1977.

[3] Daniel Marcu: "The Automatic construction of large-scale corpora for summarization research", ACM-SINGIR '99, pp. 137-144, 1999.

[4] F.C. Johnson,C.D. W.J. Paice,Black & A.P. Neal "The application of linguistic processing to automatic abstract generation", *SIGIR* '94

[5] Gerard Salton, James Allan & Amit Singhal: "Automatic text decomposition and structuring", "Information Processing and Management", V. 32, pp. 127-138, Elsevier, 1996.

[6] Grefentette, Gregory: "Automatic thesaurus generation from raw text using knowledge-poor techiniques", "Xerox", Grenoble Lab.,1995.

[7] Gustavo Crispino,Jean-Luc Minel & Javier Couto "Contexto: Una plataforma para la extracción de información y el resumen automático de textos", "Proceedings of the 2nd. Workshop on Spanish Processing and Language Technologies",pp. 153-157, Universidad de Jaén, España, 2001.

[8] I. Acero, M. Alcojor,A. Díaz, J.M. Gómez & M. Maña "Generación Automática de Resúmenes personalizados", "Procesamiento de Lenguaje Natural 27 ", pp. 281-298, SEPLN 2001.

[9] Salazar, H., Pinto, David & Jiménez, H. "Text extraction: a corpus-based approach", XXX Aniversario FCC-BUAP, noviembre 2003. ISBN: 968 863 711 4, pag. 92-94, 2003.

[10] Salazar, H. *Obtención del extracto de un texto*, Tesis de Licenciatura en Ciencias de la Computación, FCC-UAP, 2003.

[11] Kenneth Ward Church & Patrick Hanks "Word Association Norms, Mutual Information", Computational Linguistics Volume 16, Number 1, March 1990.

[12] Kevin Knight & Daniel Marcu "Summarizing beyond sentence extraction: A probabilistic approach to sentence comprensión", Elsevier 2002, pp. 91-107, 2002.

[13] Varaschin Gasperin, C. & Strube de Lima, V.L.: "Experiment on extracting semantic relations from syntactic relation", "Lecture Notes in Computer Science ", Vol. 2588, Springer, 2003.