OPTIMIZACIÓN DE RESUMEN AUTOMÁTICO MEDIANTE COMPRESIÓN DE FRASES

IRIA DA CUNHA[1,2]
ALEJANDRO MOLINA[1]

[1]*Laboratoire Informatique d'Avignon (UAPV), France*
[2]*Institut Universitari de Lingüística Aplicada (UPF), España*

RESUMEN

El objetivo de este estudio es confirmar si es adecuado emplear la compresión de frases como recurso para la optimización de sistemas de resumen automático. Para ello, en primer lugar, conformamos un corpus de resúmenes médicos producidos por diversos sistemas de resumen automático. En segundo lugar, realizamos una compresion manual de los mismos, siguiendo dos estrategias diferentes. Finalmente, comparamos los resúmenes originales con los resúmenes comprimidos, mediante el sistema ROUGE.

Palabras clave: compresión de frases, resumen automático, optimización, corpus textual,

ABSTRACT

This study aims to confirm if sentence compression could be used to improve automatic summarization systems. To perform this study, in the first place, we carry out a corpus that includes medical summaries of several automatic summarization systems. In the second place, we perform a manual compression of them, following two strategies. Finally, we compare the original summaries with the compressed summaries, using ROUGE metrics.

Keywords: sentence compression, automatic summarization, optimization, textual corpus

BUENO ALONSO, Jorge L., Dolores GONZÁLEZ ÁLVAREZ, Úrsula KIRSTEN TORRADO, Ana E. MARTÍNEZ INSUA, Javier PÉREZ GUERRA, Esperanza RAMA MARTÍNEZ & Rosalía RODRÍGUEZ VÁZQUEZ (Eds). 2010: *Analizar datos > Describir variación / Analysing data > Describing variation*. Vigo: Universidade de Vigo (Servizo de Publicacións), 808-819

1. INTRODUCCIÓN

La compresión de frases es un tema de investigación relativamente reciente, iniciado por Knight y Marcu (2000, 2002), quienes emplearon métodos estadísticos (*Noisy Channel*) y simbólicos (estructura sintáctica). Posteriormente, diversos autores trabajaron sobre este tema, como, por ejemplo, Lin (2003), Nguyen et al. (2004), Waszak y Torres-Moreno (2008), Yousfi-Monod y Prince (2008), y Fernández y Torres-Moreno (2009). El objetivo de la compresión de frases es, dada una frase, eliminar la información no esencial que incluye, manteniendo a la vez su gramaticalidad. También podría considerarse la compresión de frases como una manera de incrementar la cantidad de información relevante en un espacio limitado de palabras. La compresión de frases puede aplicarse a diversas tareas, como mejorar los sistemas de resumen automático o generar títulos automáticamente.

Nuestro trabajo está orientado al resumen automático. Partimos de la hipótesis de que una compresión adecuada de las frases de un resumen puede incluir más información relevante en el mismo espacio, conservando la gramaticalidad. De confirmarse esta hipótesis, podría emplearse la compresión de frases como recurso para la optimización de sistemas de resumen automático. Con este objetivo realizamos este trabajo. Para llevarlo a cabo, conformamos un corpus de resúmenes médicos producidos por diversos sistemas de resumen automático. A continuación realizamos una compresión manual de los mismos, siguiendo dos estrategias diferentes. Finalmente, comparamos los resúmenes originales con los resúmenes comprimidos, mediante el sistema ROUGE (Lin 2004), para evaluar si efectivamente los resúmenes comprimidos obtienen mejores resultados.

En la sección 2 detallamos la metodología empleada en nuestro estudio. En la sección 3 presentamos los experimentos realizados y los resultados obtenidos. En la sección 4 exponemos las conclusiones y el trabajo futuro.

2. METODOLOGÍA

La metodología empleada en nuestro trabajo incluye tres fases principales, que se detallan a continuación.

2.1 Conformación del corpus original

En primer lugar, conformamos un corpus de 40 textos médicos extraídos de la revista de investigación española *Medicina Clínica* (http://www.doyma.es/revistas/ctl_servlet?_f=7032&revistaid=2). Cada uno incluye un apartado de un artículo médico (de aproximadamente 400 palabras): *Fundamento*, *Pacientes y métodos*, *Resultados* o *Discusión*.

En segundo lugar obtenemos los resúmenes de los 40 textos del corpus mediante siete sistemas de resumen automático: Cortex (Torres-Moreno et al. 2002), Enertex (Fernández et al. 2008), Disicosum (da Cunha 2008), Pertinence (http://www.pertinence.net/index.html), Swesum (http://swesum.nada.kth.se/index-eng.html), OTS (http://libots.sourceforge.net/) y Word. Además, creamos una *baseline* (BL) de resúmenes con oraciones seleccionadas aleatoriamente del texto original. Todos los resúmenes contienen el mismo número de oraciones, dependiendo del apartado del texto: *Fundamento* (2), *Pacientes y métodos* (3), *Resultados* (4) y *Discusión* (2).

2.2. Compresión del corpus

Una vez obtenidos los resúmenes de los sistemas de resumen automático mencionados y la *baseline*, se procedió a su compresión manual siguiendo dos estrategias por separado:

1) Eliminación intuitiva de elementos no esenciales de la frase, como ciertos artículos, adverbios, elementos parentéticos, aposiciones, locuciones, etc., siguiendo la línea de los trabajos de Yousfi-Monod y Prince (2008). Esta estrategia implica cierta subjetividad, ya que pueden existir elementos que un anotador considere prescindibles, mientras que otro anotador considere necesarios para

el resumen. El ejemplo 1a muestra una oración original de uno de los resúmenes (resumen del apartado de *Pacientes y métodos* del resumidor Cortex); el ejemplo 1b indica en cursiva los elementos eliminados intuitivamente por parte de uno de los anotadores, y el ejemplo 1c muestra la oración final comprimida.

1a. "El Servicio de Epidemiología del Instituto Municipal de Salud Pública recoge de manera sistemática los casos de sida notificados por los médicos y, además, los casos procedentes de las altas hospitalarias y del registro de mortalidad."
1b. "El Servicio de Epidemiología del Instituto Municipal de Salud Pública recoge *de manera sistemática los* casos de sida notificados por *los* médicos y*, además, los* casos procedentes de *las* altas hospitalarias y del registro de mortalidad."
1c. "El Servicio de Epidemiología del Instituto Municipal de Salud Pública recoge casos de sida notificados por médicos y casos procedentes de altas hospitalarias y del registro de mortalidad."

2) Eliminación de "satélites" de la *Rhetorical Structure Theory* (RST) (Mann y Thompson 1988) del interior de la frase, en la línea de los trabajos de Marcu (1998, 2000). Esta estrategia implica el empleo de una base teórica más marcada. La RST es una teoría descriptiva de organización del texto muy útil para describirlo caracterizando su estructura a partir de las relaciones que mantienen entre sí los elementos discursivos del mismo (Circunstancia, Elaboración, Motivación, Evidencia, Justificación, Causa, Propósito, Antítesis, Condición, entre otras). Estas relaciones pueden ser asimétricas (núcleo-satélite) o simétricas (multinucleares): en las primeras el elemento principal se denomina "núcleo" y el secundario "satélite", mientras que en las segundas todos los elementos son núcleos. Por lo general, los satélites aportan una información adicional a sus núcleos. En la Figura 1 mostramos un árbol discursivo con relaciones de la RST, que incluye una relación multinuclear de Lista y dos relaciones núcleo-satélite, de Concesión y de Elaboración. El ejemplo ha sido extraído de uno de los textos médicos del corpus.

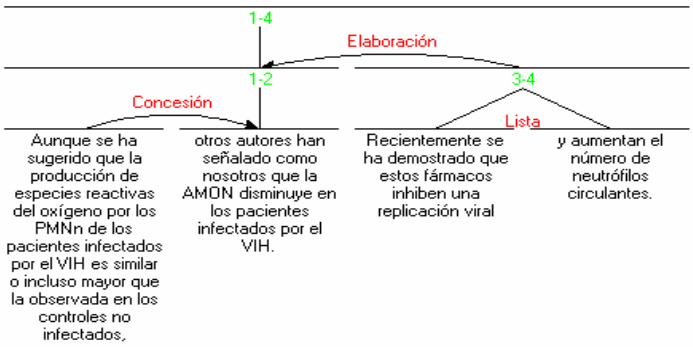

Figura 1. Árbol discursivo de la RST

El ejemplo 2a muestra una oración original de uno de los resúmenes (resumen del apartado de *Discusión* del resumidor Enertex); el ejemplo 2b indica en cursiva el satélite eliminado por parte de uno de los anotadores, y el ejemplo 2c muestra la oración final comprimida.

2a. "No existieron diferencias en las resistencias primarias o secundarias según la presencia o no de infección por el VIH como en otros estudios, aunque algunos autores comunicaron mayor frecuencia de resistencias primarias y secundarias en pacientes positivos para el VIH."

2b. "No existieron diferencias en las resistencias primarias o secundarias según la presencia o no de infección por el VIH como en otros estudios, *aunque algunos autores comunicaron mayor frecuencia de resistencias primarias y secundarias en pacientes positivos para el VIH*."

2c. "No existieron diferencias en las resistencias primarias o secundarias según la presencia o no de infección por el VIH como en otros estudios."

El fragmento eliminado ("aunque […] para el VIH".) constituye un satélite de Concesión de la RST, puesto en evidencia mediante el conector discursivo "aunque".

Así, el corpus comprimido consta de 1760 oraciones (880 comprimidas mediante la estrategia intuitiva y 880 mediante la estrategia RST). Cada uno de los cuatro autores de este trabajo llevó a cabo la compresión intuitiva y mediante la RST de una sección: *Fundamento* (160 oraciones), *Pacientes y métodos* (240 oraciones), *Resultados* (320 oraciones) y *Discusión* (160 oraciones).

2.3. Evaluación

Todos los resúmenes (comprimidos y sin comprimir) fueron evaluados con el sistema automático ROUGE, comparándolos con los *abstracts* de los autores de los artículos. Los resúmenes fueron previamente truncados a 50 palabras automáticamente para evaluarlos en condiciones iguales de tamaño en número de palabras. La Figura 2 ilustra la metodología empleada en el estudio.

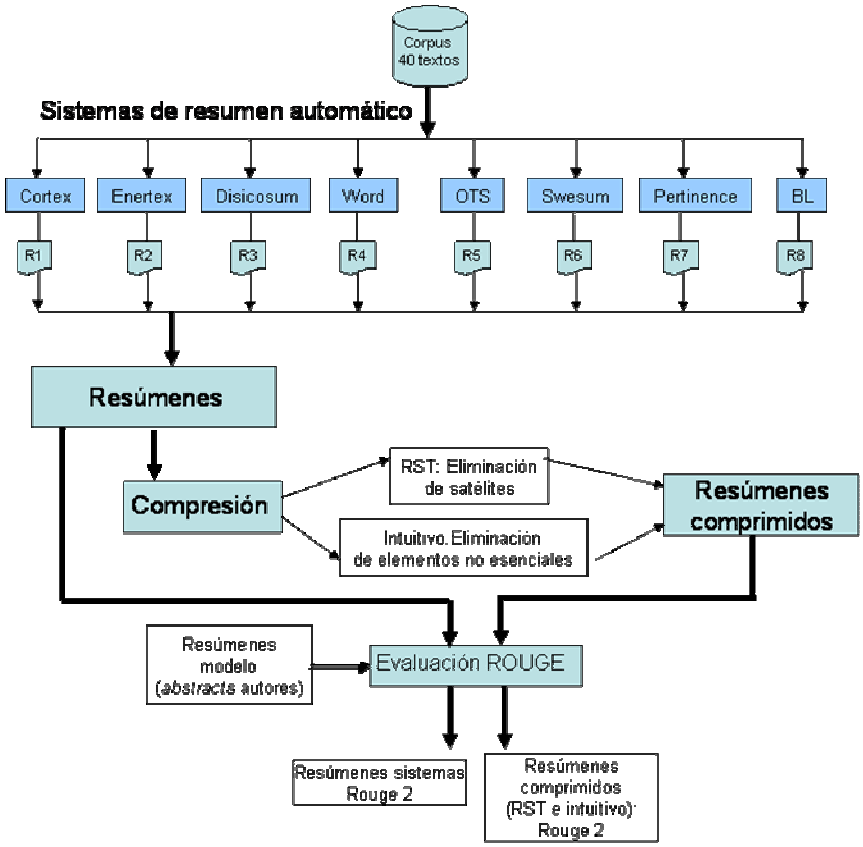

Figura 2. Metodología del estudio

3. EXPERIMENTOS Y RESULTADOS

Hemos realizado un estudio estadístico del corpus de resúmenes para medir los porcentajes de compresión *C* (intuitiva y RST) en cada sección. Una media normalizada fue calculada de la siguiente manera:

$$C = \frac{\langle A \rangle - \langle B \rangle}{\langle A \rangle} \; x \; 100$$

donde $\langle A \rangle$ es el número de palabras promedio antes de comprimir y $\langle B \rangle$ el número de palabras promedio después de la compresión.

La Figura 3 muestra los valores *C* promedios en cada sección (círculos), que oscilan entre el 13% y el 24%. Esta variación indica una cierta independencia del número de frases en la compresión e, inversamente, una fuerte dependencia de la longitud de las mismas. En cuanto a la RST, es importante señalar el comportamiento en las secciones *Discusión* y *Resultados*. En la primera, las frases contienen muchos satélites que, al ser eliminados, aumentan la compresión. En la segunda, las frases conservan una estructura mayoritariamente nuclear, que las hace poco candidatas a ser comprimidas.

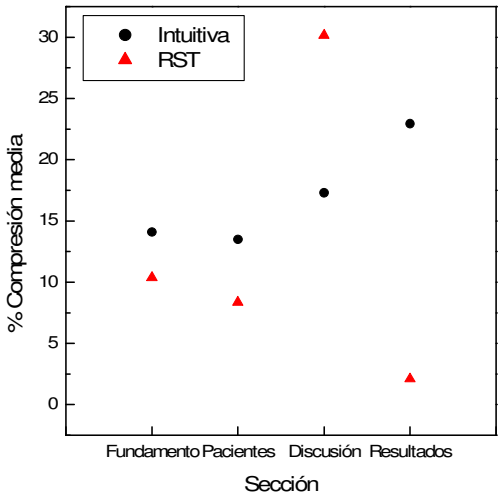

Figura 3. Porcentajes medios de compresión intuitiva y RST por sección

Para comprobar si los resúmenes comprimidos son mejores que los resúmenes originales de los sistemas de resumen automático y el resumen *baseline*, los evaluamos por separado con ROUGE. En concreto, empleamos ROUGE-2. Esta medida evalúa la coocurrencia de bigramas entre los resúmenes candidatos (es decir, los resúmenes que se desea evaluar) y los resúmenes de referencia (es decir, resúmenes modelo realizados por humanos; en nuestro trabajo los resúmenes de referencia son los *abstracts* de los autores de los artículos médicos). Una vez realizada la evaluación de ambos tipos de resúmenes (comprimidos y sin comprimir, ambos truncados a 50 palabras), comparamos la puntuación obtenida con ROUGE-2.

Los resultados reflejan que algunos de los resúmenes comprimidos de manera intuitiva obtienen mejores resultados que los resúmenes no comprimidos correspondientes, confirmando nuestra hipótesis inicial. Sin embargo, la mejora no es tan significativa como se pensó en un primer momento. Esto puede deberse a que, aunque todos los resúmenes están truncados a 50 palabras, algunos de ellos pueden incluir menos de 50 palabras una vez realizada la compresión. Este hecho puede haber provocado que estos resúmenes obtengan una puntuación más baja con ROUGE, ya que "pierden información" según este sistema (recordemos que ROUGE-2 evalúa la coocurrencia de bigramas entre resúmenes candidatos y resúmenes de referencia). Asimismo, se observa que, en general, los resúmenes comprimidos mediante la eliminación de satélites de la RST no mejoran demasiado con respecto a los resúmenes no comprimidos. Esta situación puede deberse a que las oraciones de los resúmenes de este tipo de textos son breves, porque normalmente reflejan datos o informaciones concretas (sobre todo los resúmenes de los apartados de *Pacientes y métodos* y *Resultados*), y por tanto no incluyen satélites. En la Figura 4 pueden observarse los resultados obtenidos.

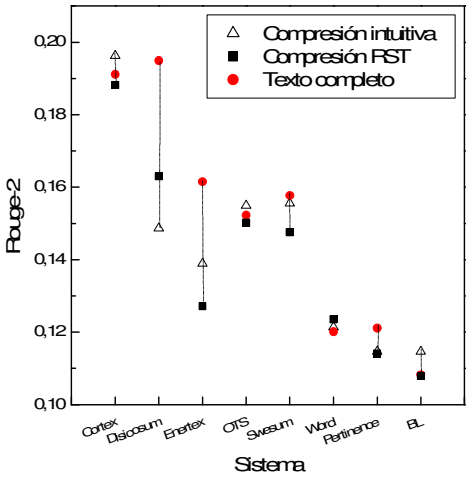

Figura 4. ROUGE-2 para cada sistema: *score* en función de las compresiones intuitiva, RST y en texto completo

4. CONCLUSIONES Y TRABAJO FUTURO

La principal conclusión de nuestro trabajo es que la compresión de frases puede beneficiar a algunos sistemas de resumen automático. Aunque la mejora no es excesivamente elevada, creemos que los resultados son esperanzadores. Además, hemos detectado los posibles motivos por los cuales algunos resúmenes comprimidos podrían haber obtenido una menor puntuación: en primer lugar, contienen menos de 50 palabras después de la compresión y esto les perjudica en la evaluación ROUGE-2. En segundo lugar, la evaluación ROUGE considera la coocurrencia de bigramas que se pierden en la compresión. Esto penaliza injustamente los resúmenes con frases comprimidas. Para evitar esto, pensamos desarrollar una nueva medida, bien adaptada a los resúmenes con compresión, que considere la pertinencia de las palabras incluidas. También nos gustaría realizar nuevos experimentos combinando ambos tipos de compresiones (en primer lugar, compresión RST y en segundo lugar compresión intuitiva).

Como conclusión general, creemos que llevar a cabo la implementación de un sistema de compresión que simule la eliminación humana intuitiva de elementos de la frase podría servir para optimizar sistemas de resumen automático. Actualmente estamos desarrollando un sistema de compresión de frases en español, inglés y francés.

AGRADECIMIENTOS

Parte de este trabajo ha sido financiado mediante una ayuda de movilidad posdoctoral otorgada por el Ministerio de Ciencia e Innovación de España (Programa Nacional de Movilidad de Recursos Humanos de Investigación; Plan Nacional de Investigación Científica, Desarrollo e Innovación 2008-2011) a Iria da Cunha. Asimismo este trabajo fue financiado parcialmente mediante la beca 211963 del CONACYT (México) a Alejandro Molina. El proyecto ha sido además parcialmente financiado por la Agence Nationale pour la Recherche (ANR) de Francia, en el marco del proyecto Resumé Plurimédia Multidocument (RPM2), concedido a Juan-Manuel Torres-Moreno.

REFERENCIAS BIBLIOGRÁFICAS

da Cunha, I. 2008. *Hacia un modelo lingüístico de resumen automático de artículos médicos en español*. Barcelona: Institut Universitari de Lingüística Aplicada. [CD-ROM] (Sèrie Tesis; 23)

Fernández, S.; SanJuan, E.; Torres-Moreno, J. M. 2008. “Enertex : un système basé sur l'énergie textuelle”. En actas del congreso *Traitement Automatique des Langues Naturelles*. 99-108. Avignon.

Fernández, S.; Torres-Moreno, J-M. 2009. “Une approche exploratoire de compression automatique de phrases basée sur des critères thermodynamiques”. En actas de la *16ème conference sur le*

Traitement Automatique des Langues Naturelles (TALN). Senlis, Francia.

Knight, K.; Marcu, D. 2000. "Statistics-based summarization - step one: Sentence compression". En actas de la *Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*. 703-710. Sapporo, Japón.

Knight, K.; Marcu, D. 2002. "Summarization beyond sentence extraction: a probabilistic approach to sentence compression". *Artificial Intelligence* 139 (1). 91-107.

Lin, C. 2003. "Improving summarization performance by sentence compression - a pilot study". En actas del *Sixth International Workshop on Information Retrivial with Asian Language* (IRAL). 1-8. Sapporo, Japón.

Lin, C. 2004. "Rouge: A Package for Automatic Evaluation of Summaries". En actas del *Workshop on Text Summarization Branches Out* (WAS 2004). 25-26. Barcelona.

Mann, W. C.; Thompson, S. A. 1988. "Rhetorical structure theory: Toward a functional theory of text organization". *Text* 8 (3). 243-281.

Marcu, D. 1998. The rhetorical parsing, summarization, and generation of natural language texts. Toronto, University of Toronto. [Tesis doctoral]

Marcu, D. 2000. *The Theory and Practice of Discourse Parsing Summarization*. Massachusetts: Institute of Technology.

Nguyen, M. L.; Shimazu, A.; Horiguchi, S.; Ho, B. T.; Hukushi, M. 2004. "Probabilistic sentence reduction using support vector machine". En actas del *20th international conference computational linguistics* (COLING-2004). 743-749. Ginebra, Suiza.

Torres-Moreno, J.M.; Velázquez-Morales, P.; Meunier, J.G. 2002. "Condensés de textes par des méthodes numériques". En actas de las *Journées internationales d'Analyse statistique des Données Textuelles*. 723-734. St. Malo.

Waszak, T.; Torres-Moreno, J-M. 2008. "Compression entropique de phrases contrôlée par un perceptron". En actas de la *9th*

International Conference on the Statistical Analysis of Textual Data (JADT). 1163-1173. Lyon, Francia.

Yousfi-Monod, M.; Prince, V. 2008. "Sentence Compression as a Step in Summarization or an Alternative Path in Text Shortening". En actas de *Coling 2008: Companion volume – Posters and Demonstrations*. 139-142. Manchester, UK.