La compresión de frases: un recurso para la optimización de resumen automático de documentos

Alejandro Molina
LIA-Université d'Avignon y
GIL-Instituto de Ingeniería UNAM
alejandro.molina@etd.univ-avignon.fr

Iria da Cunha
IULA-Universitat Pompeu Fabra,
LIA-Université d'Avignon y
GIL-Instituto de Ingeniería UNAM
iria.dacunha@upf.edu

Juan-Manuel Torres-Moreno
LIA-Université d'Avignon,
École Polytechnique de Montréal y
GIL-Instituto de Ingeniería UNAM
juan-manuel.torres@univ-avignon.fr

Patricia Velázquez-Morales
VM Labs
patricia_velazquez@yahoo.com

Resumen

El objetivo de este trabajo de investigación es confirmar si es adecuado emplear la compresión de frases como recurso para la optimización de sistemas de resumen automático de documentos. Para ello, en primer lugar, creamos un corpus de resúmenes de documentos especializados (artículos médicos) producidos por diversos sistemas de resumen automático. Posteriormente realizamos dos tipos de compresiones de estos resúmenes. Por un lado, llevamos a cabo una compresión manual, siguiendo dos estrategias: la compresión mediante la eliminación intuitiva de algunos elementos de la oración y la compresión mediante la eliminación de ciertos elementos discursivos en el marco de la *Rhetorical Structure Theory* (RST). Por otro lado, realizamos una compresión automática por medio de varias estrategias, basadas en la eliminación de palabras de ciertas categorías gramaticales (adjetivos y adverbios) y una *baseline* de eliminación aleatoria de palabras. Finalmente, comparamos los resúmenes originales con los resúmenes comprimidos, mediante el sistema de evaluación ROUGE. Los resultados muestran que, en ciertas condiciones, utilizar la compresión de frases puede ser beneficioso para mejorar el resumen automático de documentos.

1. Introducción

La compresión de frases es un tema de investigación relativamente reciente. Los métodos sobre compresión de frases están orientados a la eliminación de la información no esencial de las frases de un documento, manteniendo al mismo tiempo su gramaticalidad. Las aplicaciones de la compresión de frases pueden ser muy diversas.

Un ejemplo de ello es la generación automática de títulos. Las agencias de noticias reciben diariamente una gran cantidad de información proveniente de fuentes heterogéneas. Estas agencias cuentan con especialistas encargados de asignar un título a cada una de las informaciones que les llegan y que serán posteriormente convertidas en noticias. (Mittal and Witbrock, 1999) presentan un sistema capaz de generar encabezados de tamaño arbitrario.

Otra aplicación es la generación de subtítulos para medios audiovisuales. Hoy en día, la mayor parte de las películas cuentan con subtítulos, pero la mayoría de las cadenas de televisión todavía ofrecen el subtitulado de manera limitada. Sin embargo, en los últimos años, este tema ha suscitado un gran interés, recibiendo una atención especial. Por un lado, los subtítulos pueden traducir una narración o diálogo que se realiza en un idioma extranjero y, por otro, pueden servir para ayudar a las personas con problemas visuales a recibir la información. (Grefenstette, 1998) presenta un método de reducción de textos que tiene por objetivo disminuir el tiempo de lectura de un sintetizador para ciegos.

Otra de las aplicaciones de la compresión de frases tiene que ver con la telefonía móvil. Actualmente, los dispositivos móviles cuentan con pantallas reducidas donde el número de caracteres mostrados es limitado. La compresión de frases es un método que permitiría reducir la extensión del texto mostrado y, de esta manera, incluir más información en un espacio determinado.

En otra línea de investigación, la compresión de frases podría servir como método para la op-

timización de los sistemas de resumen automático de documentos. El resumen automático es un tema de investigación muy relevante desde hace años y se han realizado estudios para diversos idiomas como el inglés (Marcu, 2000a; Teufel and Moens, 2002), el francés (Torres-Moreno, Velázquez-Morales, and Meunier, 2002; Boudin and Torres-Moreno, 2009), el español (da Cunha and Wanner, 2005; Mateo et al., 2003), el portugués (Salgueiro Pardo and Rino Machado, 2001) y el catalán (Fuentes, González, and Rodríguez, 2004); así como estudios multilingues (Lenci et al., 2002). Recientemente existen estudios sobre resumen de textos especializados en medicina (Afantenos, Karkaletsis, and Stamatopoulos, 2005; da Cunha, Wanner, and Cabré, 2007; Vivaldi et al., 2010), química (Pollock and Zamora, 1975; Boudin, Torres-Moreno, and Velazquez-Morales, 2008; Boudin, Torres-Moreno, and El-Bèze, 2008) y derecho (Farzindar, Lapalme, and Desclés, 2004), e incluso sistemas de resumen de sitios Web (Berger and Mittal, 2000).

Los sistemas de resumen automático, por lo general, siguen el paradigma de la extracción (Edmundson, 1969; Lal and Ruger, 2002), incluyendo las oraciones más relevantes del texto de manera literal. Regenerar automáticamente el texto extraído para crear un resumen por abstracción es sumamente complicado pues se deben incluir los contenidos más relevantes del texto original, pero redactados de manera diferente (Ono, Sumita, and Miike, 1994; Paice, 1990). La compresión de frases puede ser un vínculo en el camino de la extracción a la abstracción, es decir, una forma primaria de paráfrasis. Si partimos de la hipótesis de que, para determinadas tareas, un resumen posee una extensión limitada (como es el caso de los resúmenes de noticias), la compresión de frases conservando su gramaticalidad podría permitir una mayor cantidad de información en el mismo espacio. De confirmarse esta hipótesis, podría emplearse la compresión de frases como recurso para la optimización de sistemas de resumen automático de documentos. El objetivo de este trabajo es precisamente confirmar esta hipótesis.

Como antecedente directo podemos considerar el trabajo de (Lin, 2003), en el cual se comprimen las frases de un sistema de resumen extractivo multi-documento. Las diferencias entre nuestro trabajo y el de Lin son varias: en nuestro caso evaluamos varios sistemas mono-documento, utilizamos diversas estrategias de compresión, utilizamos ROUGE como métrica de evaluación y no empleamos componentes semánticos. Los resultados obtenidos confirman algunas observaciones de Lin, pero también enriquecen las conclusiones con un panorama experimental más amplio.

Nuestra metodología tiene varias etapas. En primer lugar, conformamos un corpus de textos especializados (en concreto, artículos médicos de investigación) acompañados de los resúmenes redactados por los mismos autores de los documentos. En segundo lugar, generamos resúmenes automáticos de los textos del corpus con diversos sistemas de resumen extractivo. En tercer lugar, realizamos una compresión de estos resúmenes, siguiendo tres estrategias diferentes: eliminación manual intuitiva de algunos elementos de la oración, eliminación manual de ciertos elementos discursivos con base en la *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1988) y compresión automática por medio de sistemas elementales. Finalmente evaluamos los resultados mediante los sistemas ROUGE (Lin, 2004) y BLEU (Papineni et al., 2002), a fin de verificar si efectivamente los resúmenes comprimidos obtienen mejores resultados al compararlos con los resúmenes del autor.

El artículo está organizado de la siguiente manera: en la sección 2 hacemos una breve presentación del estado del arte de la compresión automática de frases. En la sección 3 detallamos la metodología empleada en nuestro estudio. Los diversos experimentos realizados y los resultados obtenidos son presentados en la sección 4. Para finalizar, en la sección 5 exponemos las conclusiones y algunas perspectivas de trabajo futuro.

2. Estado del arte

La compresión automática de frases ha sido recientemente abordada utilizando tanto métodos simbólicos como estadísticos. A continuación mostramos un breve panorama sobre este tema.

Con respecto a las aproximaciones simbólicas para el idioma inglés, destaca el trabajo de (Cordeiro, Dias, and Brazdil, 2009), donde se propone un sistema completo, no supervisado, que comienza por identificar oraciones similares con alta probabilidad de ser paráfrasis a partir de notas periodísticas de la Web. Posteriormente, estas son alineadas y procesadas por un sistema de programación lógica inductiva (ILP) para deducir una serie de predicados de lógica de primer orden que constituyen las reglas de compresión. Igualmente, (Jing, 2000) describe un complejo sistema que contempla tanto la verificación de la coherencia mediante el análisis sintáctico como la información contextual utilizando WordNet[1]. En (Yousfi-Monod and Prince, 2006; Yousfi-Monod and Prince, 2008) se muestra un método basado

[1] `http://wordnet.princeton.edu`

en reglas de transformación aplicadas a árboles sintácticos de frases en francés.

En la línea de las aproximaciones estadísticas, los trabajos de (Knight and Marcu, 2000; Marcu, 2000b) constituyen quizás los pilares en el estudio de la compresión estadística. Los autores adoptan el modelo de canal ruidoso (*Noisy Channel*) utilizado comúnmente en el área de traducción automática estadística. Aunque este estudio fue realizado para el inglés, la metodología parece resultar lo suficientemente general para ser aplicada a otras lenguas u otros modelos de lengua. (Lin, 2003) confirma que este último método puede resultar interesante en la tarea de resumen automático y posteriormente (Hori and Furui, 2004) muestran que también resulta útil para el resumen del discurso oral (*Speech Summarization*). (Turner and Charniak, 2005) muestran algunos problemas ligados al modelo de *Noisy Channel*, como por ejemplo que este tiende a comprimir muy poco las frases. De manera similar, (Clarke and Lapata, 2006b) indican que, en dicho modelo, la compresión es dependiente del dominio de los corpus de aprendizaje.

En otras direcciones, (Clarke and Lapata, 2006a) presentan un método no supervisado, en el cual se aborda la tarea como un problema de programación lineal. Recientemente, (Fernández and Torres-Moreno, 2009) y (Waszak and Torres-Moreno, 2008) muestran resultados interesantes con métodos diversos basados en la física estadística aplicada a documentos en francés y en inglés.

Por último, cabe mencionar que hasta donde sabemos no existen trabajos sobre la compresión de frases en español, ni tampoco un corpus paralelo (frase/frase comprimida) en esta lengua que pueda utilizarse como referencia para evaluar o entrenar sistemas.

3. Metodología

La metodología empleada en nuestro trabajo incluye las fases principales que se detallan a continuación: 1) conformación del corpus original de documentos especializados, 2) selección de herramientas de resumen automático, 3) compresión manual y automática del corpus y 4) evaluación de resultados.

3.1. Conformación del corpus especializado

En primer lugar, conformamos un corpus especializado del dominio médico. Seleccionamos 40 artículos médicos extraídos de la revista de investigación en español *Medicina Clínica*[2], fundada en 1943[3]. La versión digital de la revista permite acceder a las ediciones electrónicas de años anteriores gratuitamente, posibilitando así la constitución del corpus de estudio.

Cada documento del corpus incluye un apartado de un artículo médico (de aproximadamente 400 palabras): Fundamento, Pacientes y métodos, Resultados y Discusión.

En segundo lugar, obtenemos los resúmenes de los 40 documentos del corpus mediante los diversos sistemas de resumen automático que se detallarán en la sección 3.2.

Además, creamos resúmenes *Baseline* (BL1 o BL-aleatorio) de cada resumen con oraciones seleccionadas aleatoriamente del texto original y otro resúmenes *Baseline* (BL2 o BL-1era frase) a partir de las primeras oraciones del texto original. Todos los resúmenes contienen el mismo número de oraciones, dependiendo del apartado del texto:

- Fundamento (2 oraciones),
- Pacientes y métodos (3 oraciones),
- Resultados (4 oraciones) y
- Discusión (2 oraciones).

Para determinar este número de oraciones se calculó el promedio de las oraciones incluidas en cada apartado de los resúmenes de los autores, ya que estos resúmenes se dividen en cuatro apartados, siguiendo la estructura del artículo original. Posteriormente, se tomó la decisión de incluir una oración adicional, debido a que percibimos que, en gran cantidad de ocasiones, en estos *abstracts* se fusionaron en una sola oración las informaciones de dos o más oraciones de los artículos. Podría decirse que ha sido una decisión empírica con el objetivo de evitar una pérdida de información (da Cunha, 2008).

3.2. Selección de herramientas de resumen autómatico

Los sistemas de resumen automático que hemos empleado en nuestro trabajo se describen a continuación.

1. Cortex (Boudin and Torres-Moreno, 2007; Torres-Moreno, Velázquez-Morales, and Meunier, 2001; Torres-Moreno, Velázquez-Morales, and Meunier, 2002) es un sistema

[2] `http://www.doyma.es/revistas/ctl_servlet?_f=7032&revistaid=2`

[3] *Science Citation Index, Current Contents, Index Medicus* y *Excerpta Medica*

de resumen automático basado en el Modelo de Espacio Vectorial (VSM) (Salton and McGill, 1983). Se trata de un sistema de resumen por extracción mono-documento que combina varias métricas sin aprendizaje. Estas métricas resultan de algoritmos de procesamiento estadísticos y de información sobre la representación vectorial del documento. La idea principal es representar un texto en un espacio vectorial adecuado y aplicar procesamiento estadístico.

2. ENERTEX (Fernández, 2009; Fernández, SanJuan, and Torres-Moreno, 2007; Fernández, SanJuan, and Torres-Moreno, 2008) también es un sistema de resumen automático basado en VSM, pero en este caso se trata de un enfoque de redes de neuronas inspirado en la física estadística. El algoritmo modela los documentos como una red de neuronas de la que se estudia su energía textual. La idea principal es que un documento puede ser procesado como un conjunto de unidades interactivas (las palabras), donde cada unidad se ve afectada por el campo creado por las demás.
3. DISICOSUM (da Cunha, 2008; da Cunha and Wanner, 2005; da Cunha, Wanner, and Cabré, 2007) es un modelo de resumen automático de textos médicos que parte de la idea de que los profesionales de un dominio especializado emplean técnicas concretas para resumir los textos de su ámbito. El algoritmo de DISICOSUM integra criterios basados en la estructura textual, en las unidades léxicas y en la estructura discursiva y sintáctico-comunicativa del texto. El modelo está formado por reglas que se relacionan con estos criterios lingüísticos.
4. RESUMIDOR HÍBRIDO (da Cunha et al., 2007a; da Cunha et al., 2009) consta de varios resumidores autónomos que se combinan de manera equilibrada para formar un único resumidor híbrido. Algunos de los resumidores utilizan métodos numéricos (CORTEX y ENERTEX), otro resumidor tiene un carácter estrictamente lingüístico (DISICOSUM) y en los dos sistemas restantes las métricas estadísticas (de CORTEX y ENERTEX) se combinan con la información lingüística procedente de un extractor de términos (YATE (Vivaldi, 2001; Vivaldi and Rodríguez, 2001; Vivaldi and Rodríguez, 2002)). Las características más relevantes de YATE son: el uso intensivo de información semántica junto con el uso de técnicas de combinación de los resultados obtenidos a partir de diferentes técnicas de extracción. Ha sido desarrollado para el ámbito médico en español, aunque está siendo adaptado con éxito a otros dominios (genómica, derecho, economía, informática y medio ambiente) y otras lenguas (catalán).
5. Dos sistemas de resumen automático relevantes a nivel del estado del arte de esta temática:
   - SWESUM: `http://swesum.nada.kth.se/index-eng.html`
   - OPEN TEXT SUMMARIZER (OTS): `http://libots.sourceforge.net`
6. Dos sistemas de resumen automático comerciales:
   - PERTINENCE SUMMARIZER: `http://www.pertinence.net/index.html`
   - WORD SUMMARIZER

3.3. Herramientas de compresión de frases

Una vez obtenidos los extractos de los sistemas de resumen automático mencionados y las *baselines*, se procedió a su compresión. No se verificó el efecto en el orden inverso, es decir, no se realizó en mi primer lugar la compresión de las frases del texto original para posteriormete realizar un extracto, ya que el objetivo de este trabajo es confirmar si es adecuado emplear la compresión de frases como recurso para la optimización de sistemas de resumen automático. De tal manera que, bajo este enfoque, concebimos la extracción como la primera etapa y la compresión como la segunda etapa.

Para la compresión usamos las siguientes estrategias manuales y automáticas de eliminación de información:

Dos estrategias manuales:

1. Eliminación manual intuitiva
2. Eliminación manual basada en la RST

Cuatro estrategias automáticas:

1. Eliminación adjetival
2. Eliminación adverbial
3. Eliminación adjetival y adverbial
4. Eliminación aleatoria *baseline*

Estos sistemas serán descritos a continuación.

3.3.1. Compresión manual

Con respecto a la compresión manual empleamos dos estrategias:

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 and Prince, 2008). Esta estrategia implica cierta subjetividad, ya que pueden existir elementos que un anotador considere presindibles, mientras que otro anotador considere necesarios para el resumen. Para realizar esta tarea, utilizamos el mismo protocolo usado en la construcción del corpus de frases comprimidas en francés[4] del proyecto ANR-RPM2[5] (de Loupy et al., 2010).

   El ejemplo a) del Cuadro 1 muestra una oración original procedente de uno de los resúmenes (resumen del apartado de PACIENTES Y MÉTODOS del resumidor CORTEX) y el ejemplo b) la misma oración final comprimida.

a) "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."
b) "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."

Cuadro 1: Ejemplo de compresión manual por eliminación intuitiva.

2. Eliminación de satélites de la *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1988) del interior de la frase, en la línea de los trabajos de Marcu (Marcu, 1998; Marcu, 2000b). 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 información adicional a sus núcleos. Estos elementos pueden ser oraciones completas, pero también pueden encontrarse a nivel intraoracional, es decir, estar formados por fragmentos del interior de las oraciones. Es en estos casos en los que nos centraremos, ya que, en este trabajo, la compresión de frases se realiza dentro de las oraciones, independientemente de su contexto discursivo en el texto. Aunque existen trabajos sobre análisis discursivo automático para el portugués basados en la RST (Leal, Quaresma, and Chishman, 2006), la compresión de frases mediante esta estrategia se realizó de manera manual, debido a que no existe en la actualidad ningún analizador discursivo completo para el español que pueda detectar núcleos y satélites. Sin embargo, hay un proyecto vigente sobre el tema (da Cunha et al., 2007b; da Cunha et al., 2010), por lo que, en cuanto este analizador discursivo esté operativo, podremos llevar a cabo este tipo de compresión de manera automática.

   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.

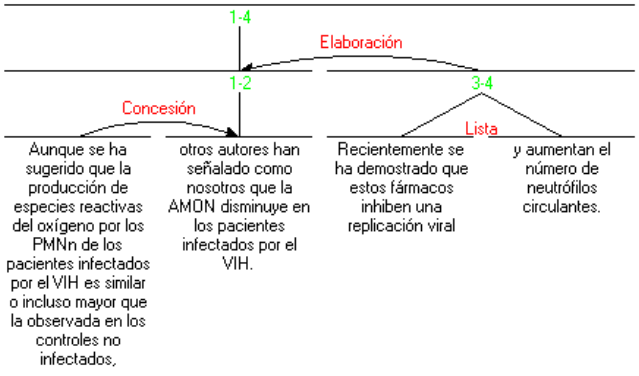

Figura 1: Ejemplo de árbol discursivo con relaciones de la RST.

   El ejemplo a) del Cuadro 2 muestra una oración original de uno de los resúmenes (resumen del apartado de DISCUSIÓN del resumi-

[4] El corpus puede ser recuperado en el sitio web: `http://lia.univ-avignon.fr/rpm2`

[5] `http://labs.sinequa.com/rpm2/`

dor Enertex) y el ejemplo b) muestra la oración final comprimida.

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

Cuadro 2: Ejemplo de compresión manual por eliminación de satélites.

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

3.3.2. Compresión automática

Con respecto a la compresión automática, hemos desarrollado cuatro sistemas:

1. Sistema de eliminación adjetival (elimADJ). Elimina todas las apariciones de adjetivos dejando los elementos restantes intactos.
2. Sistema de eliminación adverbial (elimADV). Análogo al anterior, pero eliminando adverbios.
3. Sistema de eliminación mixto (elimADJ-ADV). Elimina ambas categorías, adjetivos y adverbios.
4. Sistema de referencia de base (elimALE). Elimina un porcentaje fijo de palabras aleatoriamente (16 % en este caso –de acuerdo con la tasa de compresión promedio de los anotadores humanos–).

El Anexo 1 muestra algunos ejemplos. El ejemplo a) refleja una oración original de uno de los resúmenes (resumen del apartado de Discusión del resumidor Enertex). El ejemplo b) corresponde a la versión comprimida automática obtenida por el sistema elimADJ. El c) corresponde a la versión comprimida obtenida por el sistema elimADV y el d) corresponde a la versión comprimida del sistema elimADJ-ADV. Finalmente el ejemplo e) corresponde a la salida del sistema de base elimALE. En todos los casos se eliminó el texto entre paréntesis. Estos sistemas se explican en detalle a continuación.

Un análisis estadístico de los elementos eliminados por los anotadores, mediante el protocolo de compresión intuitiva del corpus RPM2 (de Loupy et al., 2010), arrojó resultados interesantes. El Cuadro 3 muestra las cinco secuencias más comúnmente eliminadas mediante este protocolo. Para llevar a cabo este análisis, se extrajeron por separado las secuencias de palabras eliminadas y sus equivalentes en términos de categorías gramaticales. Las categorías gramaticales fueron obtenidas mediante TreeTagger[6]. Elegimos esta herramienta por ser independiente del idioma, además de ser flexible, en el sentido de que es inmediato cambiar de un idioma a otro, lo que nos permitirá emplear, sin complicaciones, la misma metodología en trabajos futuros. Las etiquetas utilizadas en el análisis (que pueden ser consultadas en el sitio Web)[7]) fueron las siguientes: LP (paréntesis izquierdo), RP (paréntesis derecho), CARD (cifras), PERCT (símbolo %), ART (artículo), NP (nombre propio), ADJ (adjetivo) y ADV (adverbio). Observando el Cuadro 3, se puede inferir que la simple extracción de un adjetivo o un adverbio constituye una práctica común en la tarea de compresión. Del total de secuencias eliminadas, el 19.86 % incluyó al menos un adjetivo y el 9.15 % al menos un adverbio. También puede comprobarse que la eliminación del contenido entre paréntesis resulta ineluctable en la tarea de compresión. Del total de secuencias eliminadas, el 36.16 % contiene un texto entre paréntesis y el 31.91 % constituye toda la secuencia eliminada en sí. Otros resultados menos evidentes nos dieron la pauta para nuevas investigaciones al respecto. Por ejemplo, se observó que el 27.45 % de las secuencias contienen al menos una coma y, de estas, aproximadamente en la mitad es el primer símbolo de la secuencia. En sistemas posteriores consideraremos la segmentación de oraciones a partir de delimitadores ortográficos.

El análisis de las secuencias comprimidas nos llevó a construir tres sistemas de compresión elementales: el sistema de eliminación adjetival (elimADJ), el sistema de eliminación adverbial (elimADV) y el sistema de eliminación mixto (elimADJ-ADV). Además se construyó un sistema de referencia (elimALE) que extrae el 16 % de las palabras aleatoriamente –de acuerdo con la tasa de compresión promedio de los anotadores–.

[6] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`

[7] `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/spanish-tagset.txt`

Secuencia	**Ejemplos**	**Porcentaje de eliminación**
LP CARD PERCT RP	(33,7 %), (5 %)	7,14 %
ART	el, la, las, los	6,03 %
LP NP RP	(VIH), (ELISA)	4,46 %
ADV	generalmente, probablemente	4,02 %
ADJ	principales, importante	3,79 %

Cuadro 3: Lista de las secuencias más frecuentemente eliminadas en el corpus de resúmenes comprimidos intuitivamente.

En todos los casos se eliminó el contenido entre paréntesis.

4. Evaluación

Todos los resúmenes (comprimidos y sin comprimir) fueron evaluados con el sistema automático ROUGE (Lin, 2004), comparándolos utilizando como referencia los *abstracts* de los autores de los artículos. El protocolo utilizado involucra el uso de resúmenes modelo o de referencia (escritos por personas) y el paquete ROUGE, un sistema de evaluación de resúmenes que se basa en la co-ocurrencia de n-gramas entre resúmenes candidatos (los que se quiere evaluar) y resúmenes modelo. ROUGE mide los máximos, los mínimos y el valor medio (reportado en este artículo) de la intersección de los n-gramas en los resúmenes candidatos y de referencia (por ejemplo, ROUGE-1 compara unigramas, ROUGE-2 compara bigramas, ROUGE-SU4 compara bigramas con huecos, etc.). Las campañas de evaluación del NIST[8] han adoptado este test para medir la relevancia de los resúmenes. Para ser consistentes con la metodología del NIST, adoptamos el mismo protocolo en la evaluación de los resúmenes producidos por nuestro sistema. Los resúmenes fueron previamente truncados a 10, 20, 30 y así consecutivamente hasta 100 palabras automáticamente. Esta proceso garantiza una evaluación en condiciones iguales de tamaño en número de palabras.

Además de la evaluación con ROUGE, decidimos verificar la calidad de las oraciones comprimidas generadas por los sistemas automáticos. Para ello, hemos utilizado BLEU, un método de evaluación semiautomático desarrollado por IBM para la tarea de traducción automática (*Machine Translation* o MT) (Papineni et al., 2002). La idea central en MT es que, a medida que una traducción (hecha por un sistema) se acerca más (comparando la co-ocurrencia de n-gramas) a una referencia hecha por un experto, la traducción es mejor. Hemos optado por utilizar esta herramienta dado que, hasta nuestro conocimiento, no existe aún un método automático de evaluación de oraciones comprimidas. Sin embargo, reconocemos que con este método es posible que aún una frase agramatical obtenga un buen *score* BLEU. La evaluación consistió en tomar como referencia las oraciones comprimidas por los humanos mediante la estrategia intuitiva y la RST, y comparar con las oraciones comprimidas por los sistemas automáticos (elimADJ, elimADV, elimADJ-ADV y elimALE).

La figura 6 del Anexo 2 ilustra la metodología completa empleada en nuestro estudio, detallada en los apartados anteriores.

4.1. Experimentos con compresión manual

Se calculó una media normalizada (en porcentaje) de las compresiones manuales, de la siguiente manera:

$$C = \frac{\langle A \rangle - \langle B \rangle}{\langle A \rangle} \times 100 \qquad (1)$$

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 2 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 del porcentaje de compresión de 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.

Para comprobar si los resúmenes comprimidos son mejores que los resúmenes originales de los sistemas de resumen automático y los resúmenes *Baselines*, los evaluamos por separado con ROUGE. En concreto, empleamos ROUGE-2. Como ya hemos comentado, esta medida evalúa la co-ocurrencia de bigramas entre los resúmenes candidatos (es decir, los resúmenes que se desea evaluar) y los resúmenes de referencia o modelos realizados por humanos (es decir, *abstracts* de los autores de los artículos médicos).

Una vez realizada la evaluación de ambos ti-

[8] `http://www.nist.gov/index.html`

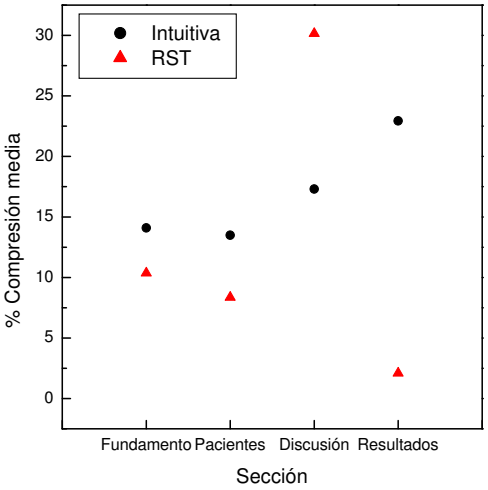

Figura 2: Porcentajes medios de compresión intuitiva y RST por sección.

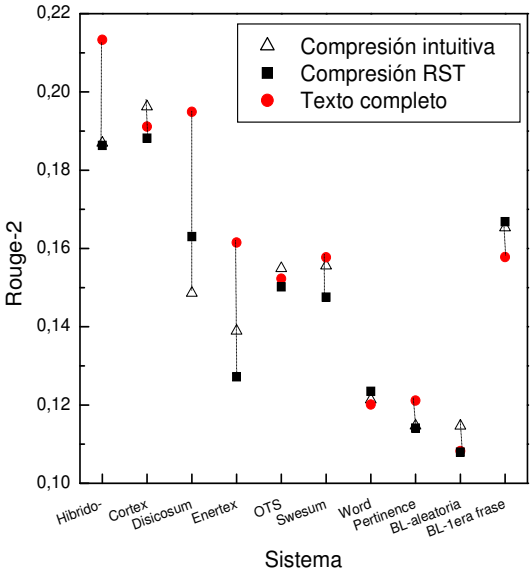

Figura 3: ROUGE-2 para cada sistema: en función del tipo de compresión realizada (truncamiento a 50 palabras) intuitiva, RST y texto completo.

pos de resúmenes (comprimidos y sin comprimir, ambos truncados de 10 a 100 palabras), comparamos el *score* obtenido con ROUGE-2.

En la figura 3 pueden observarse los resultados de ROUGE-2 obtenidos con un truncamiento promedio a 50 palabras, mediante la compresión intuitiva y mediante la compresión RST. El Cuadro 4 incluye los datos numéricos de esta evaluación. Como puede observarse, con este truncamiento, los resúmenes del sistema HÍBRIDO mejoran notablemente después de realizar la compresión mediante la estrategia intuitiva (de 0.18696 a 0.21331), mientras que mantienen una puntuación similar al ser comprimidos mediante la estrategia RST (de 0.18696 a 0.18632). El sistema CORTEX no mejora con la compresión, aunque mediante la compresión con la estrategia intuitiva no pierde excesiva información (disminuye de 0.19624 a 0.19116). DISICOSUM, por su parte, mejora sus resultados con la compresión llevada a cabo mediante ambas estrategias, pasando de 0.14862 a 0.19492 con la estrategia intuitiva y a 0.16303 con la estrategia RST. ENERTEX obtiene valores más elevados después de la compresión intuitiva de sus resúmenes (de 0.13893 a 0.16151).

El sistema OTS no mejora sus resúmenes con ningún tipo de compresión. SWESUM, WORD y PERTINENCE mejoran ligeramente sus resultados con alguno de los tipos de compresión: el primero mediante la compresión intuitiva (de 0.15558 a 0.15773) y el segundo y el tercero mediante la compresión RST (de 0.12136 a 0.12350. y de 0.11471 a 0.12115, respectivamente). Los resúmenes BL-1era frase mejoran ligeramente con la compresión RST. Finalmente, los resúmenes BL-aleatoria no mejoran sus resultados con la compresión, como era de esperarse.

Sistema	ROUGE-2 Texto completo	ROUGE-2 Comp. intuitiva	ROUGE-2 Comp. RST
HÍBRIDO	0.18696	**0.21331**	0.18632
CORTEX	0.19624	0.19116	0.18817
DISICOSUM	0.14862	0.19492	0.16303
ENERTEX	0.13893	0.16151	0.12719
OTS	0.15492	0.15227	0.1502
SWESUM	0.15558	0.15773	0.14756
WORD	0.12136	0.12012	0.1235
PERTINENCE	0.11471	0.12115	0.11408
BL-ALEAT.	0.11466	0.10821	0.10794
BL-1ERA.	0.16533	0.15782	0.16683

Cuadro 4: Resultados de la evaluación ROUGE-2 para cada sistema con truncamiento a 50 palabras.

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 al mismo número de palabras (50), algunos de ellos pueden incluir menos palabras una vez realizada la compresión. Este hecho puede haber provocado que estos resúmenes obtengan un valor más bajo de ROUGE-2, ya que al contener frases comprimidas ROUGE-2 castigará la falta de co-ocurrencias de bigramas entre resúmenes con frases compri-

midas y los 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 los textos médicos son breves, porque normalmente reflejan datos o informaciones concretas (sobre todo los resúmenes de los apartados de PACIENTES Y MÉTODOS y RESULTADOS), que generalmente no incluyen satélites.

La figura 4 reporta los resultados de ROUGE-2 obtenidos por cada sistema para resúmenes completos truncados de 10 a 100 palabras. La figura 5 muestra los resultados de ROUGE-2 de todos los sistemas, con resúmenes comprimidos mediante la estrategia intuitiva con un truncamiento de 10 a 100 palabras, además de sin truncamiento. Como puede observarse, el comportamiento de los resúmenes comprimidos intuitivamente con los diferentes niveles de truncamiento (de 10 a 100 palabras) es bastante similar al descrito para los resúmenes truncados a 50 palabras. Los resultados más destacables son la mejora evidente de los resúmenes del Resumidor HÍBRIDO mediante la compresión con un truncamiento de 30, 40, 50 y 60 palabras, la ligera mejora del sistema CORTEX con un truncamiento de 40 palabras, la clara mejora de DISICOSUM con un truncamiento de 30 y 40 palabras y la mejora, también evidente, de los resúmenes de ENERTEX con un truncamiento de 30, 40 y 50 palabras.

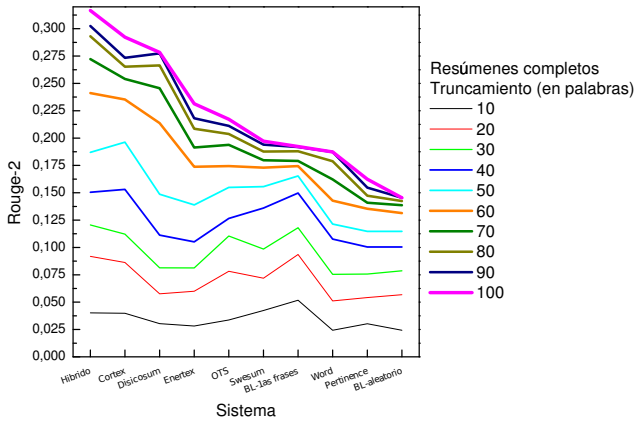

Figura 4: Resultados ROUGE-2 para resúmenes sin compresión con truncamiento de 10 a 100 palabras.

Con respecto al *ranking* de los sistemas, por un lado, al realizar la evaluación de los resúmenes completos, por lo general CORTEX se posiciona en primer lugar, seguido muy de cerca por el Resumidor HÍBRIDO, y posteriormente de la BL-1era frase, OTS, DISICOSUM, ENERTEX, WORD, PERTINENCE y BL-aleatoria.

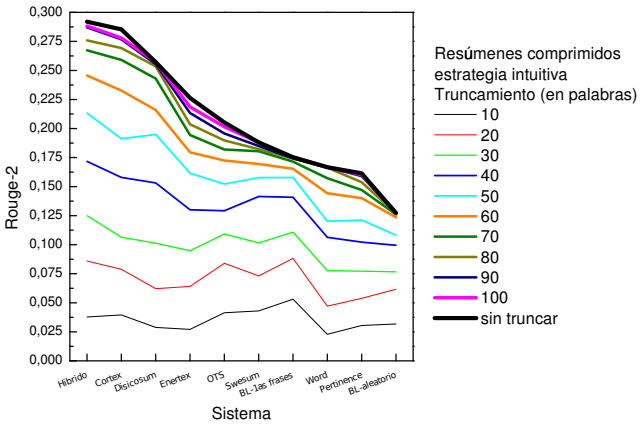

Figura 5: Resultados ROUGE-2 para resúmenes comprimidos mediante la estrategia intuitiva con truncamiento de 10 a 100 palabras y sin truncamiento.

Por otro lado, al realizar la evaluación de los resúmenes truncados, el orden del *ranking* cambia ligeramente, posicionándose claramente el Resumidor HÍBRIDO en primer lugar, seguido de CORTEX, DISICOSUM, BL-1era frase, ENERTEX, SWESUM, OTS, WORD, PERTINENCE y BL-aleatoria.

Es destacable el hecho de que, inesperadamente, la BL-1era frase obtiene resultados muy elevados tanto en la evaluación de resúmenes completos como de resúmenes comprimidos, en comparación con algunos otros resumidores. Este hecho puede deberse a que, en el tipo de documentos utilizados (artículos médicos de investigación), las primeras oraciones de cada apartado generalmente contienen las informaciones más relevantes.

4.2. Experimentos con compresión automática

En el Cuadro 5 se comparan los résumenes con frases comprimidas sin truncamiento y utilizando ROUGE-SU4. Bajo estas condiciones se observa que el sistema elimADV da mejores resultados y resulta comparable a la eliminación RST e intuitiva. Sin embargo, una lectura directa de los resúmenes comprimidos muestra que en muchas ocasiones los resúmenes generados por el sistema elimADV perdieron la consistencia debido a la generación de frases agramaticales. En el caso de la compresión intuitiva y la compresión por RST esto no sucede, ya que estas se realizan de manera manual.

El Cuadro 5 muestra los resultados de la evaluación ROUGE para los resúmenes con fra-

Sistema	SU4 RST	SU4 intuitiva	SU4 elimADV	SU4-elimADJ	SU4-elimADJ-ADV	SU4-elimALE
Híbrido	**0.31756**	**0.31333**	**0.31548**	**0.29045**	**0.28276**	0.26315
Cortex	0.31299	0.30925	0.31136	0.28514	0.27765	**0.27532**
Disicosum	0.28297	0.28454	0.28545	0.28297	0.26269	0.23440
Enertex	0.25624	0.26229	0.26235	0.24207	0.23519	0.21496
OTS	0.24546	0.24737	0.24521	0.22589	0.22004	0.20397
Swesum	0.21797	0.22126	0.21993	0.21179	0.20650	0.18582
Word	0.22048	0.20522	0.20971	0.19629	0.19508	0.17530
Pertinence	0.21829	0.21029	0.21268	0.20443	0.20153	0.18305
BL-aleat.	0.16933	0.16412	0.17155	0.15777	0.15558	0.14366
BL-1era.	0.22766	0.21544	0.21950	0.20225	0.20106	0.18756

Cuadro 5: Evaluación Rouge-SU4 para resúmenes con frases comprimidas.

ses comprimidas por los sistemas elimADJ, elimADV, elimADJ-ADV y elimALE. El Cuadro 6 muestra el promedio de la evaluación Bleu obtenido por los sistemas de compresión. En tanto que Bleu devuelve valores entre 0 y 1 (1 es asumido como una buena compresión en relación a la referencia), se puede notar que, en general, las heurísticas utilizadas por los sistemas automáticos se correspondieron mejor a la compresión intuitiva que a la compresión RST.

De acuerdo con el Cuadro 6, podría entenderse que la estrategia de eliminación de adverbios se asemeja mucho más al comportamiento intuitivo. Esta última conclusión es engañosa si se considera que, en el corpus original, los adjetivos constituyen la cuarta categoría más frecuente (6,90 %), mientras que los adverbios ocupan el dieciseisavo lugar con apenas un 1,10 %. Es decir, que la heurística de eliminar adjetivos es, en cierto sentido, mucho más arriesgada que aquella de eliminar adverbios por el simple hecho de que estos últimos aparecerán con menos frecuencia.

El sistema elimADV tiende a dejar las frases intactas con más frecuencia y el *score* Bleu, en este caso, resulta óptimo por que se compara una frase consigo misma (todos los n-gramas son encontrados intactos).

Sistema	**Referencia**	
	Intuitiva (RPM2)	**Satélites (RST)**
elimALE	0.67408	0.70669
elimADJ-ADV	0.74427	0.67549
elimADJ	0.76857	0.70757
elimADV	**0.82538**	**0.77098**

Cuadro 6: Evaluación Bleu para los cuatro sistemas de compresión automática contra las dos referencias manuales.

5. Conclusiones

En este trabajo hemos explorado la posibilidad de emplear la compresión de frases para la optimización de sistemas de resumen automático de documentos. La metodología empleada consistió en extraer las frases que conformarían el resumen y posteriormente comprimirlas mediante diversas estrategias. Este método nos permitió analizar y evaluar diversas características de ambos procesos por separado. Sin embargo, nuestros trabajos futuros estarán orientados a concebir la selección y la compresión como una tarea conjunta, pues, como se menciona en (Daumé III and Marcu, 2002), este enfoque puede llevar a mejores resultados.

La principal conclusión de nuestros experimentos es que la compresión de frases puede beneficiar a algunos sistemas de resumen automático. Esta mejora parece no ser excesivamente elevada y creemos que se debe a que los resúmenes contienen un cierto número de palabras (de 10 a 100) que después de la compresión disminuye y esto les perjudica en la evaluación Rouge, pues ésta considera la co-ocurrencia de n-gramas como una buena práctica y es de suponer que algunas de estas co-ocurrencias se pierdan en la compresión. Tenemos razones para creer que esto penaliza injustamente los resúmenes con frases comprimidas. También hemos explorado la implementación de sistemas de compresión que simulen la eliminación humana intuitiva de elementos de la frase para optimizar sistemas de resumen automático. Esta tarea plantea interesantes retos e interrogantes que deben resolverse en el futuro, comenzando por los recursos necesarios para analizar el problema (corpus alineados de frases-frases comprimidas) pues estos son, hasta nuestro conocimiento, escasos y, en algunos casos, como el del español, aún inexistentes. Sin embargo, como parte de este trabajo hemos elaborado de manera semiautomática un corpus alineado experimental para el español. Este corpus está disponible en el sitio web `http://lia.univ-avignon.fr/fileadmin/axes/TALNE/index.html`. También será interesante comprobar, en trabajos futuros, cómo se comporta la compresión en otros géneros, como noticias periodísticas. Tenemos la intuición de que algunos dominios son más sensibles a la compresión que otros.

Los sistemas de compresión descritos aquí son aún prototipos elementales pero nos permitirán contrastar los resultados de sistemas más complejos en un futuro. Por ejemplo, ahora que contamos con un conjunto de secuencias comprimidas, podemos utilizar métodos de aprendizaje supervisado para generar reglas de compresión.

Además, queremos realizar más pruebas de cara a profundizar en los motivos que han hecho que la compresión siguiendo la estrategia de la RST no obtenga resultados demasiado positivos.

Creemos que este hecho ha sido provocado por haber eliminado todos los satélites, independientemente de su tipo. En este tipo de textos científicos, por ejemplo, puede ser que los satélites del apartado RESULTADOS sean relevantes para un resumen.

A su vez, al eliminar los satélites de Condición, se pierde una información necesaria para la compresión del texto.

Finalmente, nos restan por explorar otros experimentos interesantes de compresión contextual de frases: por ejemplo, dada una frase en la posición i, su compresión podría depender del contexto generado por las $i-1$ frases precedentes $j = 1, 2, ..., i-1$. Algoritmos que consideren esta contextualización son actualmente objeto de estudio en nuestro equipo.

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, France), en el marco del proyecto *Resumé Plurimédia Multidocument* (RPM2), concedido a Juan-Manuel Torres-Moreno.

Anexo 1

a) **Oración original**
"Todos presentaron concentraciones de cocaína detectables en la orina, status epiléptico e inestabilidad hemodinámica, falleciendo dos de ellos, el tercero se encuentra en estado de coma vegetativo y el cuarto paciente, una vez estabilizado, fue sometido a laparotomía y se extrajeron 10 paquetes intactos y uno roto, evolucionando favorablemente y siendo dado de alta (tres de estos casos han sido publicados previamente)."

b) **elimADJ**
"Todos presentaron concentraciones de cocaína en la orina, status e inestabilidad, falleciendo dos de ellos, el tercero se encuentra en estado de coma vegetativo y el cuarto paciente, una vez estabilizado, fue sometido a laparotomía y se extrajeron 10 paquetes y uno roto, evolucionando favorablemente y siendo dado de alta."

c) **elimADV**
"Todos presentaron concentraciones de cocaína detectables en la orina, status epiléptico e inestabilidad hemodinámica, falleciendo dos de ellos, el tercero se encuentra en estado de coma vegetativo y el cuarto paciente, una vez estabilizado, fue sometido a laparotomía y se extrajeron 10 paquetes intactos y uno roto, evolucionando y siendo dado de alta."

d) **elimADJ-ADV**
"Todos presentaron concentraciones de cocaína en la orina, status e inestabilidad, falleciendo dos de ellos, el tercero se encuentra en estado de coma vegetativo y el cuarto paciente, una vez estabilizado, fue sometido a laparotomía y se extrajeron 10 paquetes y uno roto, evolucionando y siendo dado de alta."

e) **elimALE**
"Todos presentaron concentraciones de cocaína detectables en la orina, status epiléptico e inestabilidad hemodinamica, falleciendo ellos, el se encuentra en estado coma vegetativo y el cuarto paciente, una vez, fue sometido a laparotomía y se extrajeron 10 paquetes y roto, favorablemente y siendo dado alta."

Anexo 2

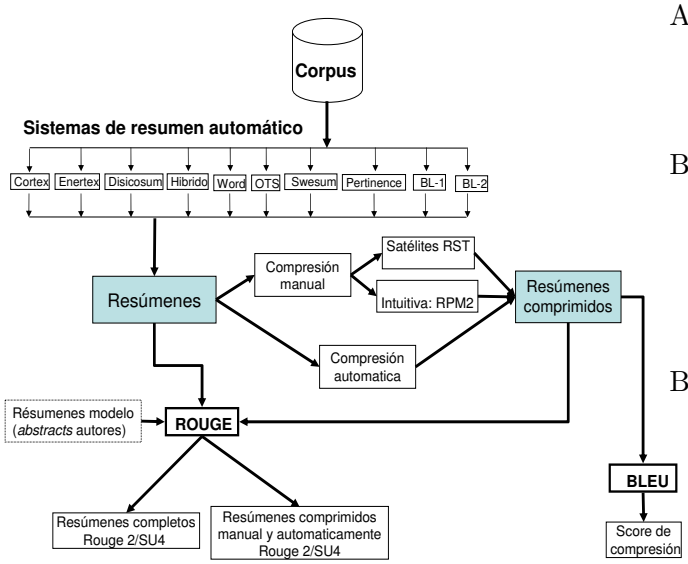

Figura 6: Metodología empleada para la generación de resúmenes, la compresión de frases y sus evaluaciones.

References

Afantenos, S., V. Karkaletsis, and P. Stamatopoulos. 2005. Summarization from medical documents: a survey. *Artificial Intelligence in Medicine*, 33(2):157–177.

Berger, A.L. and V.O. Mittal. 2000. OCELOT: a system for summarizing Web pages. In *Proceedings of the 23rd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 144–151. ACM.

Boudin, F. and J.-M. Torres-Moreno. 2007. NEO-CORTEX: A Performant User-Oriented Multi-Document Summarization System. In *Computational Linguistics and Intelligent Text Processing (CICLing'07)*, volume 4394 of *Lecture Notes in Computer Science*, pages 551–562. Springer.

Boudin, F. and J-M. Torres-Moreno. 2009. Résumé automatique multi-document et indépendance de la langue : une première évaluation en français. In *Proceedings of Traitement Automatique de la Langue Naturelle (TALN'09), Senlis.*

Boudin, F., J-M. Torres-Moreno, and M. El-Bèze. 2008. Mixing Statistical and Symbolic Approaches for Chemical Names Recognition. In *Proceedings of the conference CICLing'08, Haifa (Israel), 2008 17-23 February*, pages 334–349. The Springer LNCS 4919.

Boudin, F., J-M. Torres-Moreno, and P. Velazquez-Morales. 2008. An efficient Statistical Approach for Automatic Organic Chemistry Summarization. In *Proceedings of the International Conference on Natural Language Processing (GoTAL), Gothenburg (Sweden)*, pages 89–99. The Springer LNCS 5221.

Clarke, J. and M. Lapata. 2006a. Constraint-based sentence compression: An integer programming approach. In *COLING/ACL 2006 Main Conference Poster Sessions*, pages 144–151.

Clarke, J. and M. Lapata. 2006b. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, page 384. Association for Computational Linguistics.

Cordeiro, J., G. Dias, and P. Brazdil. 2009. Unsupervised induction of sentence compression

rules. In *UCNLG+Sum '09: Proceedings of the 2009 Workshop on Language Generation and Summarisation*, pages 15–22, Morristown, NJ, USA. Association for Computational Linguistics.

da Cunha, I. 2008. *Hacia un modelo lingüístico de resumen automático de artículos médicos en español*. Ph.D. thesis, IULA-UPF, Barcelona, España.

da Cunha, I., S. Fernández, P. Velázquez Morales, J. Vivaldi, E. SanJuan, and J.-M. Torres-Moreno. 2007a. A new hybrid summarizer based on Vector Space model, Statistical Physics and Linguistics. In *Lecture Notes in Computer Science, 4827*, pages 872–882. Springer.

da Cunha, I., S. Fernández, P. Velázquez, J. Vivaldi, E. SanJuan, and J.M. Torres-Moreno. 2007b. A new hybrid summarizer based on Vector Space Model, Statistical Physics and Linguistics. In *MICAI 2007: Advances in Artificial Intelligence. Lecture Notes in Computer Science*, pages 872–882. Gelbukh, A. and Kuri Morales, A. F. (eds.), Berlín: Springer.

da Cunha, I., E. SanJuan, J.-M. Torres-Moreno, M. Lloberes, and I. Castellon. 2010. DiSeg : Un segmentador discursivo automatico para el español. *Procesamiento de Lenguaje Natural, ISSN: 1989-7553*, 2010(45).

da Cunha, I., J.-M. Torres-Moreno, P. Velázquez-Morales, and J. Vivaldi. 2009. Un algoritmo lingüístico-estadístico para resumen automático de textos especializados. *Linguamática*, 2(2):67–79.

da Cunha, I. and L. Wanner. 2005. Towards the Automatic Summarization of Medical Articles in Spanish: Integration of textual, lexical, discursive and syntactic criteria. In *Crossing Barriers in Text Summarization Research (RANLP-2005)*, pages 46–51. Saggion, H. and Minel, J. (eds.), Borovets (Bulgaria): INCOMA Ltd.

da Cunha, I., L. Wanner, and T. Cabré. 2007. Summarization of specialized discourse: The case of medical articles in Spanish. *Terminology*, 13(2):249–286.

Daumé III, H. and D. Marcu. 2002. A noisy-channel model for document compression. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 449–456. Association for Computational Linguistics.

de Loupy, C., C. Ayache M. Guigan, S. Seng, and J.-M. Torres-Moreno. 2010. A French Human Reference Corpus for multi-documents summarization and sentence compression. In *International Conference on Language Resources and Evaluation (LREC'10)*.

Edmundson, H. P. 1969. New methods in automatic extracting. *Journal of ACM*, 16(2):264–285.

Farzindar, A., G. Lapalme, and J.P. Desclés. 2004. Résumé de textes juridiques par identification de leur structure thématique. *Traitement Automatique des Langues (TAL), Numéro spécial sur: Le résumé automatique de texte: solutions et perspectives*, 45(1):26.

Fernández, S. and J.-M. Torres-Moreno. 2009. Une approche exploratoire de compression automatique de phrases basée sur des critères thermodynamiques. In *Actes de la Conférence sur le Traitement Automatique du Langage Naturel*.

Fernández, S. 2009. *Applications exploratoires des modèles de spins au Traitement Automatique de la Langue*. Ph.D. thesis, Université Henri Poincaré Nancy 2, France.

Fernández, S., E. SanJuan, and J.-M. Torres-Moreno. 2007. Énergie textuelle de mémoires associatives. In *Traitement Automatique des Langues Naturelles*, pages 25–34. Toulouse, France.

Fernández, S., E. SanJuan, and J.-M. Torres-Moreno. 2008. Enertex : un système basé sur l'énergie textuelle. In *Traitement Automatique des Langues Naturelles*, pages 99–108. Avignon, France.

Fuentes, M., E. González, and H. Rodríguez. 2004. Resumidor de noticies en catala del projecte hermes. In *Proceedings of the II Congres d'Enginyeria en Llengua Catalana (CELC04), Andorra*.

Grefenstette, G. 1998. Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind. In *Working notes of the AAAI Spring Symposium on Intelligent Text summarization*, pages 111–118.

Hori, C. and S. Furui. 2004. Speech summarization: an approach through word extraction and a method for evaluation. *IEICE TRANSACTIONS on Information and Systems*, 87:15–25.

Jing, H. 2000. Sentence reduction for automatic text summarization. In *Proceedings of the sixth conference on Applied natural language processing*, pages 310–315. Association for Computational Linguistics.

Knight, K. and D. Marcu. 2000. Statistics-based summarization-step one: Sentence compression. In *National Conference on Artificial Intelligence*, pages 703–710. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Lal, P. and S. Ruger. 2002. Extract-based summarization with simplification. In *Document Understand Conference (DUC'02)*. NIST.

Leal, Ana, Paulo Quaresma, and Rove Chishman. 2006. From syntactical analysis to textual segmentation. In Renata Vieira, Paulo Quaresma, Maria Nunes, Nuno Mamede, ClÃ¡udia Oliveira, and Maria Dias, editors, *Computational Processing of the Portuguese Language*, volume 3960 of *Lecture Notes in Computer Science*, pages 252–255. Springer Berlin / Heidelberg.

Lenci, A., R. Bartolini, N. Calzolari, A. Agua, S. Busemann, E. Cartier, K. Chevreau, and J. Coch. 2002. Multilingual summarization by integrating linguistic resources in the MLIS-MUSI project. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*, pages 29–31.

Lin, Chin-Yew. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Marie-Francine Moens and Stan Szpakowicz, editors, *Text Summarization Branches Out: ACL-04 Workshop*, pages 74–81, Barcelona, July.

Lin, C.Y. 2003. Improving summarization performance by sentence compression-a pilot study. In *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages*, pages 1–8.

Mann, W. C. and S. A. Thompson. 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*. Ph.D. thesis, Dep. of Computer Science, University of Toronto.

Marcu, D. 2000a. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

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

Mateo, P.L., J.C. González, J. Villena, and J.L. Martínez. 2003. Un sistema para resumen automático de textos en castellano. *DAEDALUS SA, Madrid, España*.

Mittal, V. O. and M. J. Witbrock. 1999. Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries. In *SIGIR 9'9: proceedings of 22nd International Conference on Research and Development in Information Retrieval, August 1999*, page 315. University of California, Berkeley.

Ono, K., K. Sumita, and S. Miike. 1994. Abstract generation based on rhetorical structure extraction. In *Proceedings of the 15th conference on Computational linguistics - Volume 1*, pages 344–348. Association for Computational Linguistics (ACL).

Paice, C.D. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing & Management*, 26(1):171–186.

Papineni, K., S. Roukos, T. Ward, and W.-j. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Pollock, J.J. and A. Zamora. 1975. Automatic abstracting research at chemical abstracts service. *Journal of Chemical Information and Computer Sciences*, 15(4):226–232.

Salgueiro Pardo, T.A. and L.H. Rino Machado. 2001. A Summary Planner Based on a Three-Level Discourse Model. In *6th NLPRS - Natural Language Processing Pacific Rim Symposium*, pages 533–538.

Salton, G. and M. McGill. 1983. *Introduction to Modern Information Retrieval*. Computer Science Series, McGraw Hill Publishing, Company.

Teufel, S. and M. Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Torres-Moreno, J.-M., P. Velázquez-Morales, and J.G. Meunier. 2002. Condensés de textes par des méthodes numériques. In *JADT*, volume 2, pages 723–734.

Torres-Moreno, J.-M., P. Velázquez-Morales, and J.G. Meunier. 2001. Cortex : un algorithme pour la condensation automatique des textes. In *ARCo 2001*, pages 65–75. Lyon, France.

Turner, J. and E. Charniak. 2005. Supervised and unsupervised learning for sentence compression. In *Association for Computational Linguistics*, volume 43, pages 290–297.

Vivaldi, J. 2001. *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona.

Vivaldi, J., I. da Cunha, J.M. Torres-Moreno, and P. Velázquez-Morales. 2010. Automatic summarization using terminological and semantic resources. In *International Conference on Language Resources and Evaluation (LREC'10)*.

Vivaldi, J. and H. Rodríguez. 2001. Improving term extraction by combining different techniques. *Terminology*, 7(1):31–47.

Vivaldi, J. and H. Rodríguez. 2002. Medical term extraction using the EWN ontology. In *Terminology and Knowledge Engineering*, pages 137–142. Nancy.

Waszak, T. and J.-M. Torres-Moreno. 2008. Compression entropique de phrases contrôlée par un perceptron. In *Journées internationales d'Analyse statistique des Données Textuelles (JADT'08) Lyon*, pages 1163–1173.

Yousfi-Monod, M. and V. Prince. 2006. Compression de phrases par élagage de leur arbre morpho-syntaxique. *Technique et Science Informatiques*, 25:437–468.

Yousfi-Monod, M. and V. Prince. 2008. Sentence Compression as a Step in Summarization or an Alternative Path in Text Shortening. In *Coling'08*.