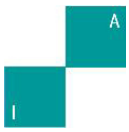
Inteligencia Artificial 16(51) (2013), 41-62



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

Compresión automática de frases: un estudio hacia la generación de resúmenes en español

Alejandro Molina
Laboratoire Informatique d'Avignon (LIA)
Université d'Avignon et des Pays de Vaucluse
alejandro.molina-villegas@univ-avignon.fr

Abstract The automatic generation of summaries is a challenging task that has generated two methodological families called extraction and abstraction. Sentence compression establishes a bridge between both families since it realizes a fine-grained extraction, creating a primary form of paraphrase. Here we present a study about sentence compression and proposes a linear model that predicts the removal of intra-sentence segments with application in summarization. The model uses a set of features measured directly from the text and was trained over 60 thousand decisions of remove or preserve a discourse segment considering the whole context and the produced summary. Using statistical analysis, we found the most significance features to predict with 75 % of accuracy the probability of deletion of segments. Then, we use the best model to generate summaries with compressed sentences. After a Turing-like test, we found that generated summaries are similar in quality than human-made summaries. We anticipate our assay to be a starting point of more sophisticated summarization methods and sentence compression algorithms. Furthermore, we put at the disposal of the community the data generated in this investigation, convinced that they will be of valuable utility for future research.

Keywords: sentence compression, discourse segments elimination, Turing test for NLP.
Keywords: compresión de frases, eliminación de segmentos discursivos, test de Turing aplicado al PLN.

1. Introducción

La compresión de frases consiste en eliminar las partes menos importantes de una oración o de una frase de manera automática. Aunque es un tema muy reciente, ya se han propuesto diversas aplicaciones para su uso, por ejemplo en los dispositivos móviles que cuentan con pantallas reducidas en tamaño y donde el número de caracteres mostrados es limitado. La compresión de frases permitiría reducir la extensión del texto mostrado y, de esta manera, incluir más información en un espacio reducido. Otro ejemplo es la generación automática de títulos de cualquier tamaño [52]. Las agencias de noticias reciben diariamente una gran cantidad de información proveniente de fuentes heterogéneas y sería una herramienta para los especialistas encargados de asignar un título a cada una de las informaciones que les llegan y que serán posteriormente convertidas en noticias. Otra aplicación es la de traducción de subtítulos. Un modulo de traducción automática puede estar acoplado con uno de compresión de frases de manera que se garantice una longitud específica del texto traducido. También pueden servir para ayudar a las personas con problemas visuales a recibir la información. [17] presenta un método de reducción de textos que tiene por objetivo disminuir el tiempo de lectura de un sintetizador para ciegos. El enfoque que presentaremos aquí es el de la compresión de frases como método de resumen automático de documentos. Medio siglo de investigación da muestra de lo desafiante de esta área que ha dado lugar a dos grandes familias:

ISSN: 1988-3064(on-line)

los métodos por extracción y los métodos por abstracción [48]. En los primeros se seleccionan alguna oraciones del documento de origen. No se realiza ninguna modificación de las oraciones seleccionadas. Los segundos intentan elaborar el resumen a partir del contenido original, pero no necesariamente usando las mismas oraciones del documento con escrupulosa exactitud. Debido probablemente a la dificultad de estos últimos, los métodos de resumen por extracción han tenido una mayor evolución. Así pues, la compresión de frases establece un puente entre ambas familias dado que realiza una extracción, pero de granularidad fina sobre el texto, creando una forma primaria de paráfrasis. Saca provecho de una de la mayores desventajas de los métodos por extracción: algunas de las oraciones seleccionadas para formar el resumen presentan información irrelevante. Un procesamiento más fino es necesario, un método de resumen al nivel de la oración.

En este artículo presentamos un método automático de compresión de frases, con aplicaciones en la generación de resumen, en el cual se eliminan algunos segmentos intra-oracionales de las frases. La sec. 2 está dedicada al estado del arte. En la sec. 3 explicamos la segmentación discursiva que utilizamos para delimitar los segmentos a eliminar. En la sec. 4 se define formalmente la compresión de frases y se establecen los criterios a considerar. Después, en la sec. 5 presentaremos un modelo para calcular la gramaticalidad de las frases comprimidas y en la sec. 6 presentaremos un modelo para estimar su informatividad. Luego, en la sec. 7 se describe la anotación multitudinaria del primer corpus de compresión de frases en español. En la sec. 8 presentamos un modelo de predicción de segmentos intra-frase basado en las anotaciones de dicho corpus y en la sec. 9 proponemos algoritmos para generar resúmenes con frases comprimidas. La sec. 10 trata de la problemática de la evaluación de frases comprimidas y describe una propuesta de evaluación por medio del test de Turing. La sección final está consagrada a las conclusiones y perspectivas del presente estudio.

2. Estado del Arte

Una de las primeras propuestas de compresión textual automática se atribuye a [17], donde se presenta un método de reducción de textos a su "versión telegráfica" con la finalidad de reducir el tiempo de lectura de un sintetizador para ciegos. Mediante el uso de reglas, el método intenta guardar lo más importantes de un documento, aunque el resultado sea agramatical. Estas reglas, originalmente propuestas por [19], indican: los nombres propios son más importantes que los nombres comunes, los sustantivos son más importantes que adjetivos, los adjetivos son más importantes que los artículos y la negación siempre es importante. Aunque la propuesta es original, las reglas aplicadas son discutibles por ser intuitivas y no se presenta evaluación alguna. Tiempo después, [52] presentan un método de resumen no extractivo, capaz de generar títulos de cualquier tamaño, el cual requiere una etapa de aprendizaje supervisado que intenta captar tanto las reglas de selección del contenido más importante como la realización del resumen. El modelo de selección está basado en la probabilidad de que una palabra ocurra en el resumen dada su aparición en el texto original y el modelo de realización consiste en un cadena de Markov de primer orden.

El trabajo de [18] es muy importante porque involucra estudios a partir de resúmenes realizados manualmente y muestra evidencia de que los humanos seleccionamos las partes relevantes de un texto para luego, mediante un trabajo de edición, hacerlas embonar en nuevas formas para construir un resumen. Los autores denominan a este proceso corta-y-pega (*cut-and-paste*) y se proponen descifrarlo para luego reconstruirlo. Tras analizar 120 oraciones provenientes de 15 resúmenes, se encontró que una proporción significativa (78 %) de oraciones provenientes de un resumen hecho por un humano fue elaborado mediante copia-y-pega. También se identificaron 6 operaciones básicas del proceso de corta-y-pega: reducción de oraciones, combinación de oraciones, transformación sintáctica, paráfrasis léxica, generalización/especificación y reordenamiento. De entre éstas, las dos operaciones principales fueron la reducción de frases y la combinación de frases lo que llevó al desarrollo de un "sistema de generación de texto" que intenta automatizar el proceso corta-y-pega.

Los trabajos de [23] marcaron el rumbo de la investigación en compresión de frases durante algunos años. Por primera vez, se utilizó el término *sentence compression*; se compiló un corpus de referencia con 1 067 oraciones y se propuso un método de evaluación manual. La compresión de frases se definió de la siguiente manera: sea la frase φ una secuencia de n palabras, $\varphi = (w_1, \ldots, w_n)$. Un algoritmo debe eliminar cualesquiera de ellas de manera que la secuencia restante es una compresión (no se permite cambiar el orden de las palabras). Para ello se proponen dos métodos: el primero basado en el modelo

del canal ruidoso y el segundo basado en árboles de decisión. Cabe mencionar que, en ambos casos, el procesamiento se realiza a partir de los árboles sintácticos de las frases y no directamente de éstas, lo que obliga en principio a usar un analizador sintáctico. En el método del canal ruidoso, se supone que existe una frase comprimida $\widetilde{\varphi}$ la cual fue alterada, añadiendo información complementaria ("ruido"), hasta llegar a su versión original φ. Así, el objetivo es optimizar $\mathbf{P}(\varphi, \widetilde{\varphi}) = \mathbf{P}(\widetilde{\varphi})\mathbf{P}(\varphi|\widetilde{\varphi})$, donde $\mathbf{P}(\widetilde{\varphi})$ corresponde a la probabilidad de que la frase $\widetilde{\varphi}$ exista y $\mathbf{P}(\varphi|\widetilde{\varphi})$ es la estimación de transformar el árbol sintáctico de $\widetilde{\varphi}$ en el de φ. En la práctica las estimaciones se obtienen a partir de las frecuencias del corpus *Penn Treebank*[1] y las transformaciones sintácticas se realizan usando un analizador probabilístico de gramática libre de contexto (*Standard Probabilistic Context-Free Grammar, SPCFG*) [6]. En el método de árboles de decisión, el árbol sintáctico de φ es transformado en un árbol reducido, mediante operaciones sucesivas y con la ayuda de una pila. La operaciones propuestas son: transferir la primera palabra a la pila (*shift*), obtener los dos subárboles en el tope de la pila y combinarlos en uno nuevo (*reduce*), remover algunas palabras de la pila (*drop*), cambiar la etiqueta de un subárbol en la pila (*assign type*) y regresar algún elemento de la pila a la secuencia original (*restore*). Para la evaluación, 4 jueces valoraron 32 frases mediante un score entre 1 y 5 indicando qué tan gramatical es el resultado y qué tanto se preservó la información importante. El mejor resultado fue para el canal ruidoso que obtiene un score global del 4.34 ± 1.02.

En los años subsecuentes, se han encontrado limitaciones a la propuesta de [23] muchas de ellas explicadas en detalle por [37]. La más importante es que no es posible generar los árboles de algunas frases comprimidas usando gramáticas libres de contexto. También, se ha observado que las oraciones son ambiguas si no se observa el contexto completo en el que está inscrita, por lo que [5] proponen que esto se tome en cuenta. Aunado a esto, el corpus tiene pocos árboles de ejemplo y algunas reglas no se aprenden porque éstas son vistas tan solo una vez. En [31] se propone procesar directamente la frase en lugar de su árbol sintáctico. Un algoritmo de programación dinámica decide, para cada palabra, si la frase obtiene un mejor score al conservarla o eliminarla. El score es una función lineal basada en características de las frases y en sus compresiones, cuyos pesos son calculados a partir de un corpus de entrenamiento.

Los estudios más recientes agregan dos aspectos a la compresión de frases: 1) se debe considerar el contexto de la frase en lugar de procesarla de manera aislada y 2) es más natural eliminar fragmentos al interior de la frase que palabras aisladas. El algoritmo propuesto por [44] fragmenta las frases en proposiciones y luego, con la ayuda de un analizador de dependencias, elimina las hojas del árbol. Se propone un score de informatividad basado en semántica latente [24] pero no se propone ningún componente para evaluar la gramaticalidad. Los autores consiguen mejorar este aspecto mediante un método de aprendizaje supervisado [45]. En [42] se explora la eliminación de estructuras discursivas en lugar de proposiciones. Los autores argumentan que, aunque el análisis discursivo automático a nivel del documento representa aún un desafío, la segmentación discursiva a nivel de la oración es una alternativa realista a la compresión de frases, pues algunos modelos de análisis discursivo a este nivel han mostrado un desempeño comparable al de los humanos [41].

Aunque hasta ahora hemos mencionado únicamente trabajos para el inglés, la compresión de frases ha sido estudiada para otros idiomas. Para el francés, [53] presentan un estudio acerca de eliminación de artículos, adverbios, elementos parentéticos, aposiciones y locuciones; por su parte, [51] exploran un método basado en un perceptrón y [16] exploran un método basado en modelos termodinámicos. Para el holandés, [9] proponen un sistema de generación de subtítulos por compresión de frases. En portugués, [2] estudian la simplificación de frases por medio de reglas. Y en lo que concierne al español, [33] y [35] han estudiado este tema.

3. La segmentación discursiva

La primera parte de nuestro método se basa en la segmentación discursiva, que es la primera etapa del análisis discursivo automático descrito en [28]. En este tipo de análisis, se busca representar un documento a través de un árbol jerárquico que contiene información de tipo retórico/discursivo: (1) dónde comienzan y terminan los segmentos discursivos elementales del documento, (2) qué relación retórica/discursiva existe entre estos segmentos, (3) cómo es la estructura retórica/discursiva global del documento. En aras

[1] `http://www.cis.upenn.edu/~treebank/`

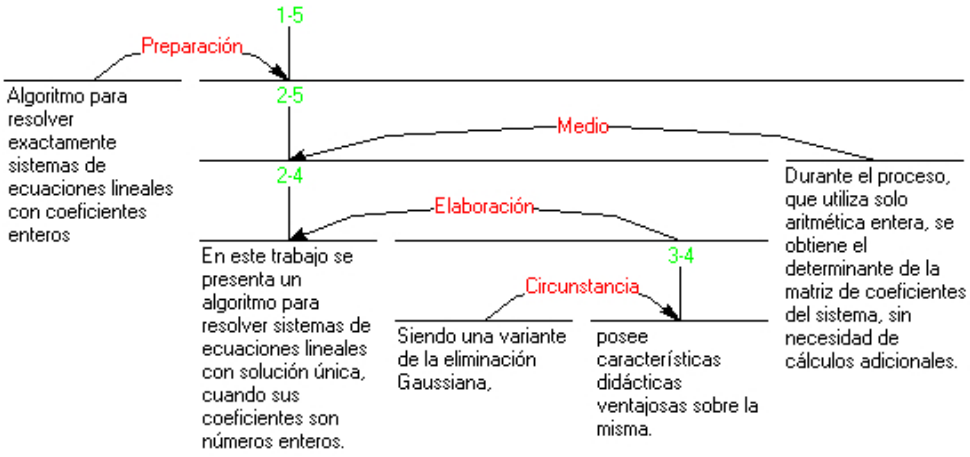


Figura 1: Ejemplo del análisis discursivo de un texto à través de su árbol RST.

de proveer un marco teórico que haga posible este tipo de análisis, se han propuesto teorías como la *Rhetorical Structure Theory* (RST) [26], ampliamente utilizada por la comunidad de Procesamiento del Lenguaje Natural (PLN).

La figura 1 muestra un ejemplo del análisis discursivo completo de un texto a través de su árbol RST. En la primera etapa, la segmentación discursiva, se identifican las fronteras de los segmentos discursivos:

> [Algoritmo para resolver exactamente sistemas de ecuaciones lineales con coeficientes enteros]$\_s_1$ [En este trabajo se presenta un algoritmo para resolver sistemas de ecuaciones linelaes con solución única, cuando sus coeficientes son números enteros.]$\_s_2$ [Siendo una variante de la eliminación Gaussiana,]$\_s_3$ [posee características didácticas ventajosas sobre la misma.]$\_s_4$ [Durante el proceso, que utiliza solo aritmética entera, se obtiene el determinante de la matriz de coeficientes del sistema, sin necesidad de cálculos adicionales.]$\_s_5$

En la segunda etapa, se identifican tanto el tipo de relación que enlaza a los segmentos como su nuclearidad; es decir, si un segmento (o un grupo de ellos) contiene la información medular de la relación, en cuyo caso se dice que es el núcleo; o en su defecto, contiene información complementaría acerca del núcleo, en cuyo caso se denomina satélite. En la figura 1, el sentido de las flechas determina la nuclearidad: las unidades desde las cuales parte la flecha corresponden a satélites, mientras que la puntas son los núcleos. El tipo de relación está indicado en medio de la flecha y las unidades involucradas en la relación se indican con los números en la cima de la linea vertical sobre los núcleos. Las líneas horizontales representan grupos de segmentos denominados *spans*, los cuales se aglutinan formando parte en relaciones de nivel superior, de donde surge la jerarquía que es conformada en la última etapa del análisis discursivo del documento.

Se ha propuesto la generación de resúmenes de manera automática construyendo el árbol discursivo completo y eliminando luego sus hojas [29]. Sin embargo, hasta donde sabemos, la construcción del árbol discursivo no ha sido aún automatizada para ningún idioma más que para el inglés. En lo que concierne al español, no se cuenta más que con programas para efectuar la segmentación discursiva.

Nuestra propuesta para la compresión de frases consiste en eliminar segmentos discursivos intra-frase, para lo cual nos valemos únicamente de la primera etapa del análisis discursivo. Así, nuestra unidad mínima eliminable es el segmento discursivo. Partimos de la idea de que si una frase ya es suficientemente simple, ésta será considerada como un solo segmento discursivo y por tanto no tendrá necesidad de ser comprimida. Si por el contrario, dicha frase es larga y compleja, estará formada por muchos segmentos discursivos, probablemente algunos de ellos conteniendo información complementaria que puede ser omitida. Cabe mencionar que estudios recientes han mostrado mejores resultados al eliminar segmentos discursivos en lugar de palabras aisladas y que existen segmentadores discursivos en español [7], francés [1], portugués [30], tailandés [21] e inglés [47], lo que permite proyectar nuestro estudio de compresión de frases a otros idiomas.

En nuestra investigación, hemos utilizado DiSeg [8], un segmentador discursivo para el español basado en la RST, el cual fue creado mediante la modificación de la gramática del analizador sintáctico FreeLing [3]. La modificación consiste en categorizar ciertas expresiones consideradas los "candidatos a marcadores discursivos". Éstos pueden ser simples (*e.g.* "como" y "entonces"); compuestos (*e.g.* "por ejemplo" y "al contrario") o complejos (*e.g.* "primero ... luego" y "por un lado ... por otro lado"). El proceso de segmentación se realiza en dos etapas [8]. Primero, se detectan las fronteras considerando la aparición de: 1) formas verbales, 2) conjunciones, 3) proposiciones subordinadas y 4) candidatos a marcadores discursivos. Luego, un programa lee de derecha a izquierda la salida de la primera etapa y aplica una serie de reglas para determinar si se debe insertar una marca de frontera. Una de las reglas aplicadas es, por ejemplo, que una forma verbal conjugada aparezca a la derecha y a la izquierda de un marcador discursivo. Los autores de DiSeg reportan un valor de F-Score de 80 % sobre un corpus de textos médicos [7].

Para verificar la utilidad de DiSeg en la compresión de frases, en [35] se evaluó hasta qué punto los humanos eliminan fragmentos textuales que corresponden a segmentos discursivos identificados por DiSeg. Para ello, se presenta el siguiente experimento: después de reunir un corpus con textos cortos de cuatro géneros (Wikipedia, resúmenes de artículos científicos, periodístico y cuento), se solicitó a cinco lingüistas eliminar palabras o grupos de palabras de las frases del corpus bajo las condiciones de no reescribir las frases, no modificar el orden de las palabras, no sustituir las palabras, asegurarse que las frases comprimidas sean gramaticales y asegurarse que las frases y el texto resultante conserve el significado de origen. Se comprobó que aproximadamente la mitad de los fragmentos eliminados corresponden con segmentos identificados por DiSeg. De la otra mitad, la mayoría de los fragmentos corresponde a alguno de estos tres casos:

1. fragmentos que comienzan por un un participio, *e.g.* "valorado en 40 mil dólares";

2. proposiciones relativas, *e.g.* "que agrupaba los videos más vendidos";

3. fragmentos que presentan un candidato a marcador discursivo pero que no incluyen un verbo, *e.g.* "a causa de la malnutrición durante la ocupación alemana".

Estos resultados llevaron a la creación de un segmentador discursivo orientado a la compresión de frases, el *Compression Segmenter* o CoSeg, mediante tres modificaciones realizadas a DiSeg que son:

1. la revocación de restricciones verbales;

2. la adición de nuevos candidatos a marcadores discursivos;

3. la adición de signos de puntuación como marcadores discursivos.

El cuadro 1 muestra un ejemplo de los segmentos resultantes usando DiSeg y CoSeg en el mismo documento. La figura 2 presenta la cobertura del segmentador CoSeg sobre un volumen de 675 fragmentos eliminados del corpus pero no reconocidos por DiSeg como segmentos discursivos. La parte oscura en la primera barra presenta la proporción de fragmentos cubiertos por CoSeg al eliminar las restricción verbal. La segunda barra muestra la proporción acumulada de los fragmentos eliminando la restricción verbal y agregando los nuevos marcadores. La tercera barra es la cobertura acumulada incluyendo también los signos de puntuación como marcadores. Estos resultados también dieron la pauta para explorar la posibilidad de un segmentador discursivo multilingüe de pocos recursos lingüísticos externos. En [38], la idea principal es poder agregar idiomas a dicho segmentador por medio de una lista de marcadores en texto plano. Se comprobó que usando un etiquetador de categorías gramaticales[2] y estrategias similares a las de CoSeg se puede obtener una precisión por encima del 60 % para el francés. En las secciones siguientes explicaremos en detalle de qué manera se pueden eliminar automáticamente segmentos discursivos identificados por DiSeg o por Coseg.

[2]Se utilizó TreeTagger disponible públicamente (`http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`) y con soporte en alemán, inglés, francés, italiano, holandés, español, búlgaro, ruso, griego, portugués, gallego, chino, swahili, latín y estonio.

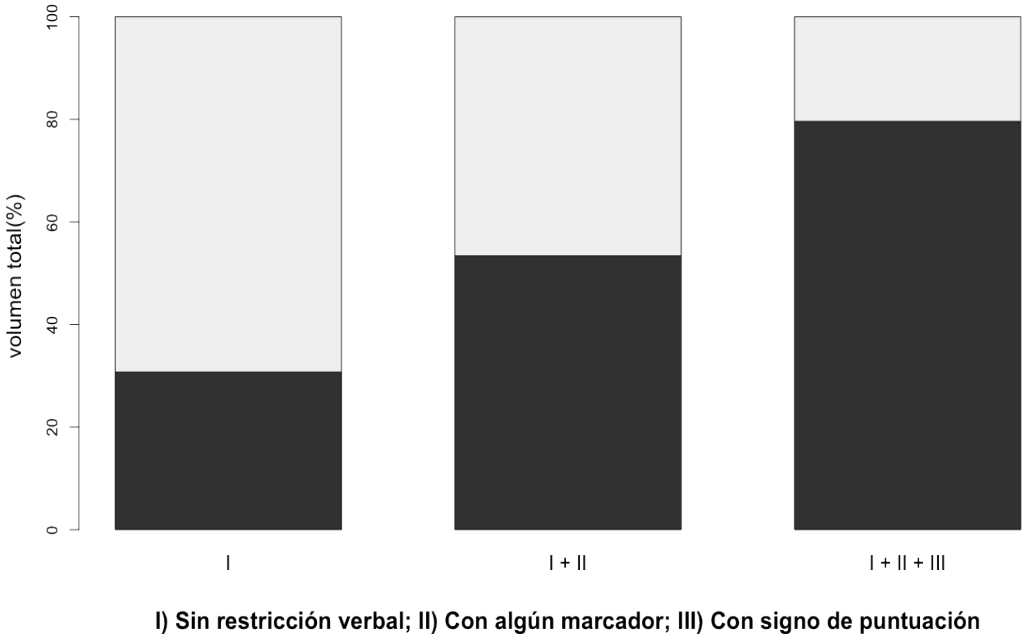


Figura 2: Cobertura del segmentador discursivo para la compresión de frases CoSeg sobre 657 fragmentos (2 651 palabras) no reconocidos por DiSeg.

Cuadro 1: Ejemplo de un documento, titulado "El termómetro", segmentado por DiSeg y por CoSeg.

El termómetro (segmentado por DiSeg)
[Para saber qué tan caliente o frío está algo, es decir,]$\_s_1$ [si se desea conocer la temperatura, debe utilizarse un instrumento que ofrezca un dato confiable, el termómetro.]$\_s_2$ [Este instrumento tiene muchos usos en los hogares, en las industrias y en las unidades de salud.]$\_s_3$ [En casa es útil tener un termómetro]$\_s_4$ [para saber con precisión]$\_s_5$ [si alguien de la familia tiene fiebre.]$\_s_6$ [En la industria los termómetros miden la temperatura de hornos y calderas, así como de diversos materiales]$\_s_7$ [y sustancias que cambian a través de un proceso productivo.]$\_s_8$ [Como ves,]$\_s_9$ [con frecuencia es necesario medir la temperatura de distintas cosas, del aire, del cuerpo humano, de un horno o del agua de una alberca,]$\_s_{10}$ [por lo que existen distintos tipos de termómetros.]$\_s_{11}$ [No importa el tipo de termómetro, en todos ellos la temperatura se mide en unidades llamadas grados.]$\_s_{12}$ [Cada marca del instrumento es un grado]$\_s_{13}$ [y cada tipo de termómetro incluye una escala de medición que, por lo general, se da en grados centígrados.]$\_s_{14}$
El termómetro (segmentado por CoSeg)
[Para saber qué tan caliente o frío está algo,]$\_s_1$ [es decir,]$\_s_2$ [si se desea conocer la temperatura,]$\_s_3$ [debe utilizarse un instrumento que ofrezca un dato confiable,]$\_s_4$ [el termómetro.]$\_s_5$ [Este instrumento tiene muchos usos en los hogares,]$\_s_6$ [en las industrias y en las unidades de salud.]$\_s_7$ [En casa es útil tener un termómetro]$\_s_8$ [para saber con precisión]$\_s_9$ [si alguien de la familia tiene fiebre.]$\_s_{10}$ [En la industria los termómetros miden la temperatura de hornos y calderas,]$\_s_{11}$ [así como de diversos materiales]$\_s_{12}$ [y sustancias que cambian a través de un proceso productivo.]$\_s_{13}$ [Como ves,]$\_s_{14}$ [con frecuencia es necesario medir la temperatura de distintas cosas,]$\_s_{15}$ [del aire,]$\_s_{16}$ [del cuerpo humano,]$\_s_{17}$ [de un horno]$\_s_{18}$ [o del agua de una alberca,]$\_s_{19}$ [por lo que existen distintos tipos de termómetros.]$\_s_{20}$ [No importa el tipo de termómetro,]$\_s_{21}$ [en todos ellos la temperatura se mide en unidades llamadas grados.]$\_s_{22}$ [Cada marca del instrumento es un grado]$\_s_{23}$ [y cada tipo de termómetro incluye una escala de medición que,]$\_s_{24}$ [por lo general,]$\_s_{25}$ [se da en grados centígrados.]$\_s_{26}$

4. Compresión de frases por eliminación de segmentos discursivos

Hasta ahora hemos visto que la segmentación discursiva puede ser ventajosa para la compresión e incluso hemos propuesto un segmentador especializado. Ahora hablaremos de los criterios para decidir si un segmento debe eliminarse o conservarse. Estos se basan en la gramaticalidad de la frase resultante; en su informatividad (entendida como la cantidad de información importante retenida) y en la tasa de compresión. En la ecuación (1), la tasa de compresión τ expresa el volumen conservado después que la frase φ fuera resumida a la frase $\widetilde{\varphi^\star}$ y donde $\mathbf{Lon}(\bullet)$ es la longitud.

$$\tau = \frac{\mathbf{Lon}(\widetilde{\varphi^\star})}{\mathbf{Lon}(\varphi)} \tag{1}$$

Así, dada una frase φ, un algoritmo de compresión de frases debe encontrar una versión comprimida (que denotaremos con $\widetilde{\varphi^\star}$) que considere a la vez los tres criterios siguientes:

- ser más corta que la original o de igual longitud (si no se puede comprimir): $\mathbf{Lon}(\widetilde{\varphi^\star}) \leq \mathbf{Lon}(\varphi)$;
- preservar la información original (equivalencia semántica): $\mathbf{Inf}(\widetilde{\varphi^\star}) \approx \mathbf{Inf}(\varphi)$;
- ser gramaticalmente correcta: $\mathbf{Gram}(\widetilde{\varphi^\star}) \approx \mathbf{Gram}(\varphi)$.

Por ejemplo, el Cuadro 2 presenta las 2^3 subsecuencias posibles de la frase $\varphi =$ [En casa es útil tener un termómetro]$\_s_1$ [para saber con precisión]$\_s_2$ [si alguien de la familia tiene fiebre.]$\_s_3$. Según el criterio de longitud, $\widetilde{\varphi}_8$ sería la solución óptima pero sería también la peor solución considerando el criterio de informatividad, dado que no preserva ninguna información. Note también que los candidatos $\widetilde{\varphi}_3$, $\widetilde{\varphi}_4$, $\widetilde{\varphi}_6$ y $\widetilde{\varphi}_7$ no pueden ser soluciones porque no son gramaticales o porque cambian el sentido original de la frase. No obstante, sería discutible la elección entre $\widetilde{\varphi}_1$, $\widetilde{\varphi}_2$ y $\widetilde{\varphi}_5$ como solución dado que el criterio definitivo sería la informatividad y mientras alguien se puede inclinar por no eliminar nada, argumentando que todo es importante, alguien más podría determinar que lo más importante se encuentra en el primer segmento y que el resto puede eliminarse sin perdidas importantes en la semántica. Esto nos lleva a algunas observaciones interesantes que debemos hacer notar. La primera es que según nuestra definición los tres criterios se deben poder medir y comparar mediante sendas funciones $\mathbf{Lon}(\bullet)$, $\mathbf{Inf}(\bullet)$ y $\mathbf{Gram}(\bullet)$. El criterio de longitud es de hecho el único que se puede calcular trivialmente, basta con contar el número de palabras o segmentos. Con respecto a la gramaticalidad, en la sección 5 discutiremos este criterio y veremos cómo podemos asignar un valor de probabilidad a una frase de acuerdo con las frecuencias obtenidas a partir de un corpus grande. En todo caso, siempre podemos leer directamente la frase resultante y decidir si es gramatical o no. La informatividad es sin duda el criterio más difícil de tratar. En la sección 6 explicaremos un modelo para cuantificarla artificialmente, el cual ha sido utilizado en varias aplicaciones de procesamiento automático del lenguaje. Sin embargo, al analizar los resultados de la anotación manual de más de 60 mil segmentos anotados manualmente, comprobamos que existe un cierto grado de subjetividad inherente a la compresión de frases ligado a la tarea de determinar qué es importante. Esto nos lleva a que la solución $\widetilde{\varphi^\star}$ puede no ser única y esto debe ser considerado en el diseño de métricas artificiales para determinar la informatividad de un texto.

Cuadro 2: Ejemplo de los candidatos a la compresión para una frase con tres segmentos

$\widetilde{\varphi}_1$	(s_1, s_2, s_3)	En casa es útil tener un termómetro para saber con precisión si alguien de la familia tiene fiebre.
$\widetilde{\varphi}_2$	(s_1, s_3)	En casa es útil tener un termómetro si alguien de la familia tiene fiebre.
$\widetilde{\varphi}_3$	(s_1, s_2)	En casa es útil tener un termómetro para saber con precisión.
$\widetilde{\varphi}_4$	(s_2, s_3)	Para saber con precisión si alguien de la familia tiene fiebre.
$\widetilde{\varphi}_5$	(s_1)	En casa es útil tener un termómetro.
$\widetilde{\varphi}_6$	(s_2)	Para saber con precisión.
$\widetilde{\varphi}_7$	(s_3)	Si alguien de la familia tiene fiebre.
$\widetilde{\varphi}_8$	$()$	

5. Gramaticalidad

Dado que uno de los criterios para la compresión de frases es la gramaticalidad, necesitamos un método para poder determinar de manera automática si una frase es correcta. Más que eso, según nuestra definición, necesitamos una función con la cual poder comparar el "grado de gramaticalidad" de una frase. Esto es justamente uno de los temas más desafiantes del PLN, basta con utilizar algún traductor automático para darse cuenta que, hoy por hoy, no hay un método eficaz para saber si una frase es gramaticalmente correcta.

Aunque contamos con analizadores sintácticos muy sofisticados, estos están diseñados para construir árboles sintácticos y no precisamente para determinar si una secuencia de palabras es una frase gramaticalmente correcta. En consecuencia, es posible obtener el árbol sintáctico a partir de cualquier frase (incluso una agramatical). El árbol de la Figura 3 se obtuvo mediante el analizador FreeLing 3.0[3] usando como entrada la frase "útil un tener casa En termómetro es". Un experimento similar es descrito en [20] para la frase "*He is an amazing*" Utilizando un analizador sintáctico estadístico [22] y un analizador sintáctico relacional (*link parser*) [40].

Ante esto, existen avances en torno a la evaluación automática de la gramaticalidad y uno de los modelos más flexibles es el modelo probabilístico del lenguaje [4, 27]. Éste permite estimar la probabilidad de una secuencia de palabras a partir de frecuencias de n-gramas obtenidas de un corpus. Para una frase $\varphi = (w_1, w_2, \ldots, w_n)$, el modelo estima la probabilidad $\mathbf{P}(\varphi) = \mathbf{P}(w_1, w_2, \ldots, w_n)$ mediante la ecuación (2), donde $w_i^j = (w_i, \ldots, w_j)$ representa la subsecuencia que va de la palabra w_i a la palabra w_j (un n-grama). En la práctica, se debe prever la situación de que un n-grama no aparezca en el corpus (en la ecuación (2), basta con que un término sea cero para que la estimación de la secuencia completa sea cero). Para evitar esta situación, se debe estimar la probabilidad de los n-gramas inexistentes mediante el suavizado de la distribución de frecuencias [4].

$$\mathbf{P}(w_1^n) = \mathbf{P}(w_1) \times \mathbf{P}(w_2|w_1) \times \mathbf{P}(w_3|w_1^2) \times \ldots \times \mathbf{P}(w_n|w_1^{n-1}). \quad (2)$$

Con el fin de verificar la utilidad de un modelo de lengua probabilístico para la compresión de frases, se han realizado experimentos que comparan dos sistemas de resumen automático por compresión de frases contra un humano y contra un sistema aleatorio [34]. Todos comprimieron cada frase de una serie de textos siguiendo estrategias diferentes. Un anotador lingüísta eliminó palabras o grupos de palabras de las frases asegurándose que las frases comprimidas fueran gramaticales y que el texto resultante conservara el significado de origen. Un primer sistema, S_{todo}, elegía como solución aquella compresión con el score de gramaticalidad más alto de entre todas las subsecuencias para cada frase dada. El segundo sistema, S_{prim}, elegía como solución la compresión con el score de gramaticalidad más alto, pero eligiendo únicamente de entre las compresiones que incluían el primer segmento. El tercer sistema, S_{ale}, eliminó palabras de manera aleatoria con una tasa de compresión fija al 30 %. El score de gramaticalidad fue calculado mediante el programa SRLIM [46] junto con el corpus *Google Web 1T*[4] siguiendo la configuración recomendada para dicho corpus. Los resultados de dicho estudio son que la proporción de compresiones agramaticales producidas por el anotador es de 0 %, la de S_{todo} es 8.1 %, la de S_{prim} es apenas 6.9 % y la de S_{ale} es de 76.6 %. Así, tomando como referencias al humano y S_{ale}, se observó que un modelo de lengua probabilístico puede ser útil como criterio de gramaticalidad. Otro resultado importante es que una heurística simple como conservar sistemáticamente el primer segmento puede mejorar los resultados de gramaticalidad de frases comprimidas porque es más probable que no se pierda el sujeto principal, lo que nos llevó a considerar las posiciones de los segmentos en nuestro modelo.

La conclusión principal es que aunque un modelo probabilístico del lenguaje no puede garantizar si una frase es gramatical o no lo es, éste sí es sensible a perturbaciones en la gramática de frases comprimidas y es muy flexible para ser utilizado en varios idiomas.

[3] `http://nlp.lsi.upc.edu/freeling/demo/demo.php`

[4] Google Web 1T 5-grams, 10 European Languages, Version 1. LDC Catalog No.: LDC2009T25

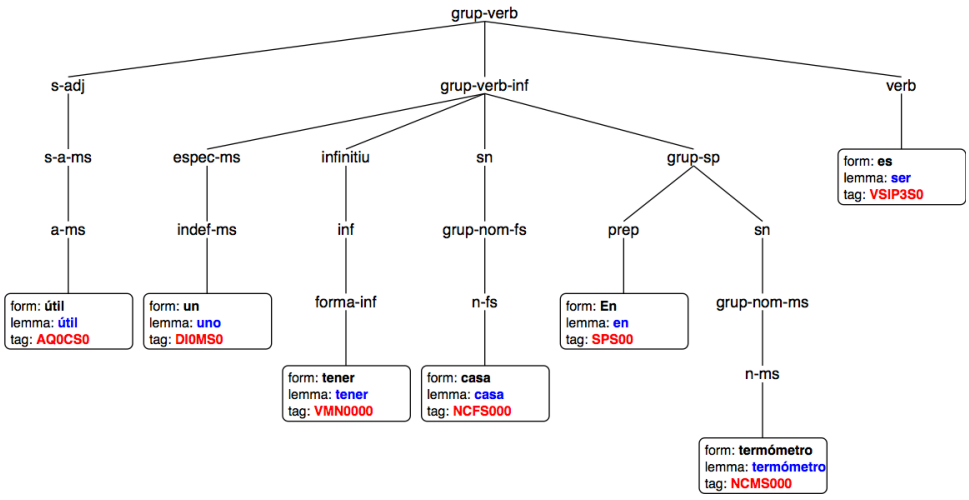


Figura 3: Árbol sintáctico de una frase agramatical.

6. Informatividad

Cuantificar la informatividad de un texto no es una cuestión trivial; lo que ha dado la pauta a diversas propuestas, basadas en su mayoría en frecuencias de palabras. Desde el modelo del espacio vectorial [39], es posible representar textos a través de vectores. Cada elemento en el vector tiene un valor de "peso" que representa la importancia de una palabra en el texto y éste se utiliza para ponderar las frases. La forma más extendida de asignar dicho valor de peso es calculando el TF-IDF (*term frecuency-inverse document frecuency*) [43]. Otros modelos muy utilizados son *Latent Semantic Indexing* [10], LexRank [12] y Enertex, también conocido como "energía textual" [13, 14, 15].Todos ellos comparten la ventaja principal del procesamiento numérico: independencia del idioma. Nuestro criterio de informatividad está basado en la energía textual porque ha demostrado buenos resultados en resumen automático y, a diferencia de los otros dos modelos, funciona incluso cuando se trata con frecuencias unitarias. Pero, el aspecto más importante es que este modelo toma en cuenta el contexto de las frases pues para calcularla se toman en cuenta las relaciones léxicas entre ellas. A continuación sintetizamos la manera de calcular el valor de energía textual de las frases y/o los segmentos de un documento.

Sea T el número de palabras distintas de un texto, el tamaño del vocabulario. Representaremos la i-ésima frase del texto como una secuencia de T elementos $\varphi_i = (a_{i,1}, \cdots, a_{i,\mathrm{T}})$ en la cual $a_{i,j}$ es la frecuencia del término w_j en φ_i. Así, para un texto con Φ frases, la matriz $\mathbb{A} := (a_{ij})$ que lo representa tiene dimensiones $[\Phi \times \mathrm{T}]$, donde $1 \geq i \leq \Phi$,$1 \geq j \leq \mathrm{T}$. Según lo descrito en [13, 14, 15], los pesos de la interacción entre las palabras se calculan mediante la regla de Hebb en su forma matricial usando el producto de la ecuación (3). Donde $j_{i,j} \in \mathbb{J}$ considera las frecuencias de las palabras que ocurren en la misma frase (relación de primer orden), así como las palabras que co-ocurren con otras previamente relacionadas (relación de segundo orden.)

$$\mathbb{J} = \mathbb{A}^{\mathrm{t}} \times \mathbb{A} \tag{3}$$

La energía textual de un documento es finalmente calculada mediante el producto de la ecuación (4). El valor de energía, $e_{i,j} \in \mathbb{E}$, representa el grado de relación léxica entre φ_i y φ_j y la energía de una frase φ_i se calcula sumando los valores de i-ésimo renglón de la matriz de energía.

$$\mathbb{E} = \mathbb{A} \times \mathbb{J} \times \mathbb{A}^{\mathrm{t}} = (\mathbb{A} \times \mathbb{A}^{\mathrm{t}})^2 \tag{4}$$

Nuestra propuesta es utilizar los valores de energía textual como medida de informatividad de los segmentos discursivos. El Cuadro 3 muestra los valores de energía textual de un texto de nuestro corpus titulado "El termómetro" segmentado con CoSeg. La primer columna corresponde a la energía de la frase y la segunda columna a la del segmento. En los experimentos se detectó que la distribución de los valores de

energía textual, en su versión original, es asimétrica y con una cola muy pesada hacia los valores cercanos a cero, lo cual fue analizado y corregido mediante la transformación **Log**(**Ener**(•)) de los valores.

Comparando los valores del Cuadro 3, podemos decidir si un segmento es importante y se debe conservar o si puede ser eliminado de la frase. Consideremos, por ejemplo, la frase más energética del texto (**Ener**$(\varphi) = 119$), $\varphi_5 = (s_{14}, s_{15}, s_{16}, s_{17}, s_{18}, s_{19}, s_{20})$. Si conservamos solamente los segmentos cuyo valor es mayor que cero, obtenemos la compresión:

$\widetilde{\varphi'}_5 = (s_{14}, s_{15}, s_{18}, s_{19}, s_{20})$ =Como ves, con frecuencia es necesario medir la temperatura de distintas cosas, de un horno o del agua de una alberca, por lo que existen distintos tipos de termómetros.

Por supuesto, se puede argumentar que los segmentos s_{14}, s_{18} y s_{19} también deberían ser eliminados, pues a juzgar por los valores de energía textual, son los segmentos s_{15} y s_{20} los que parecen portar lo más importante, lo que genera la compresión:

$\widetilde{\varphi''}_5 = (s_{15}, s_{20})$ =Con frecuencia es necesario medir la temperatura de distintas cosas, por lo que existen distintos tipos de termómetros.

De hecho, llevando esto al extremo, podríamos conservar únicamente el segmento con la energía textual más alta en la frase, lo cual podría resultar en una frase agramatical. Pero aún, dejando de lado los resultados agramaticales, cómo podríamos determinar que una solución es mejor que otra. Nuestra propuesta aborda esta cuestión basándose en el aprendizaje supervisado. En la sección siguiente, explicaremos el protocolo de anotación de más de 60 mil segmentos para los cuales se decidió si se debía eliminar o conservar tomando en cuenta los criterios de la compresión de frases.

Cuadro 3: Ejemplo de valores de energía textual de los segmentos CoSeg de un texto.

Ener(φ)	**Ener**(s_i)	
78	3	[Para saber qué tan caliente o frío está algo,]$\_s_1$
78	0	[es decir,]$\_s_2$
78	2	[si se desea conocer la temperatura,]$\_s_3$
78	3	[debe utilizarse un instrumento que ofrezca un dato confiable ,]$\_s_4$
78	7	[el termómetro.]$\_s_5$
36	2	[Este instrumento tiene muchos usos en los hogares,]$\_s_6$
36	10	[en las industrias y en las unidades de salud.]$\_s_7$
60	11	[En casa es útil tener un termómetro]$\_s_8$
60	0	[para saber con precisión]$\_s_9$
60	7	[si alguien de la familia tiene fiebre.]$\_s_{10}$
65	18	[En la industria los termómetros miden la temperatura de hornos y calderas,]$\_s_{11}$
65	4	[así como de diversos materiales]$\_s_{12}$
65	2	[y sustancias que cambian a través de un proceso productivo.]$\_s_{13}$
119	2	[Como ves,]$\_s_{14}$
119	13	[con frecuencia es necesario medir la temperatura de distintas cosas,]$\_s_{15}$
119	0	[del aire,]$\_s_{16}$
119	0	[del cuerpo humano,]$\_s_{17}$
119	1	[de un horno]$\_s_{18}$
119	5	[o del agua de una alberca,]$\_s_{19}$
119	16	[por lo que existen distintos tipos de termómetros.]$\_s_{20}$
49	15	[No importa el tipo de termómetro,]$\_s_{21}$
49	3	[en todos ellos la temperatura se mide en unidades llamadas grados.]$\_s_{22}$
66	6	[Cada marca del instrumento es un grado]$\_s_{23}$
66	14	[y cada tipo de termómetro incluye una escala de medición que,]$\_s_{24}$
66	0	[por lo general,]$\_s_{25}$
66	0	[se da en grados centígrados.]$\_s_{26}$

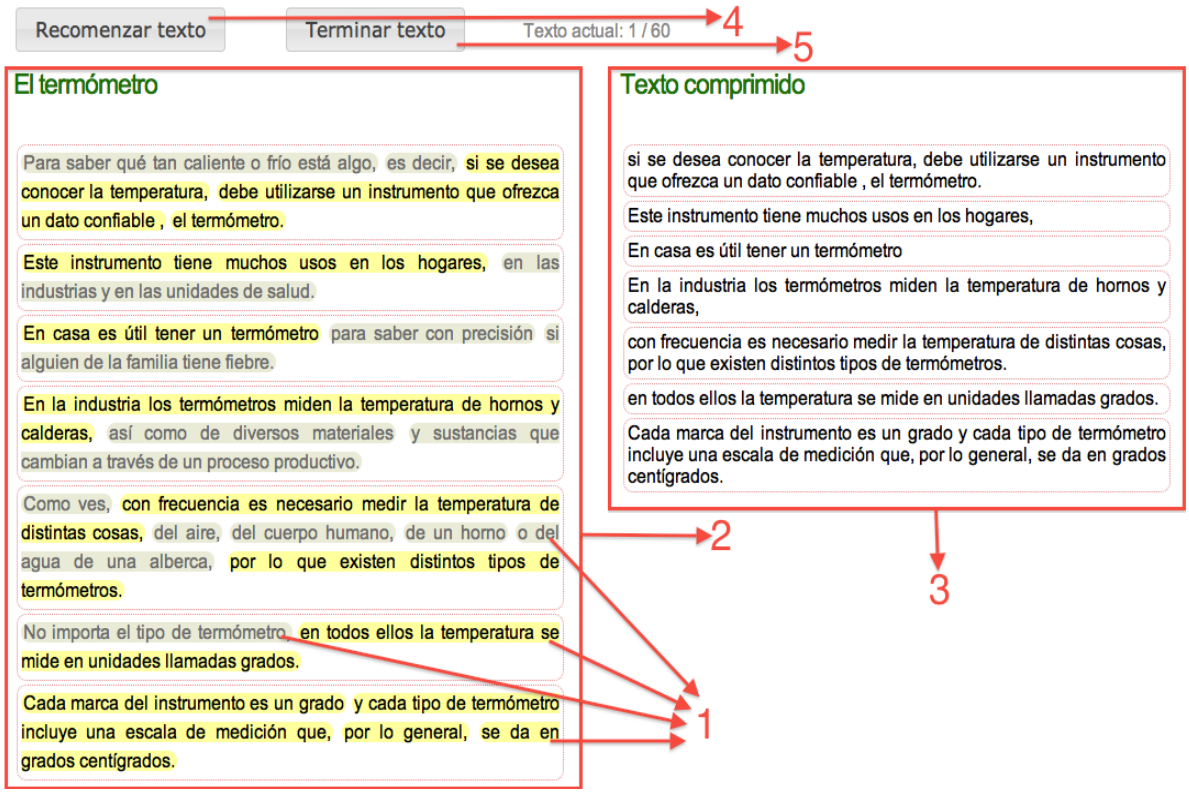


Figura 4: Componentes del sistema de generación de resúmenes por eliminación de segmentos discursivos.

7. Corpus

Debido a la falta de un corpus en compresión de frases en español fue necesario desarrollar una plataforma de anotación multitudinaria y lanzar una campaña de anotación masiva. Al final de ésta, se logró capturar más de 60 mil veces la decisión de eliminar o conservar un segmento discursivo y al mismo tiempo, se obtuvieron 2 877 resúmenes elaborados manualmente, cada uno elaborado por una persona diferente.

Primero se eligieron 30 documentos que fueron segmentados usando DiSeg y CoSeg. Cerca de 150 voluntarios fueron registrados para anotar el corpus, todos ellos hablantes nativos del español y con nivel de estudios superior al bachillerato. Siguiendo el paradigma de "ciencia ciudadana" aplicado al PLN [32], se elaboró un manual de anotación, muy sencillo y con normas precisas[5]. Cada voluntario anotaba tantos documentos como quería y en el momento que le parecía adecuado pero nunca anotaba el mismo texto dos veces por lo que fue necesario implementar módulos de administración de usuarios y *pooling*.

La figura 4 muestra la interfaz del sistema de anotación así como sus componentes que son: 1) los segmentos discursivos que pueden ser activados o desactivados haciendo clic sobre ellos; 2) el área donde se muestra el texto original; 3) el área donde se muestra el texto resultante al desactivar los segmentos; 4) el botón que restaura la interfaz a las condiciones iniciales y 5) el botón que envía el resultado de la anotación a la base de datos y realiza una petición para desplegar un texto nuevo.

Consideramos que tanto el sistema de anotación como los datos permitirá realizar experimentos en el futuro por lo cual hemos publicado el código de un sistema similar[6], así como los datos resultantes de esta investigación[7].

[5] http://molina.talne.eu/compress4/man/
[6] http://molina.talne.eu/sentaatool/info/systeme_description.html
[7] http://molina.talne.eu/sentence_compression/data/

8. Modelo de predicción de la eliminación de segmentos discursivos

En uno de los trabajos más representativos en resumen automático, H.P. Edmunson, estableció un paradigma basado en el método de regresión lineal [11]. Él propuso ponderar las oraciones de un documento con base en cuatro características que consideró importantes según las frecuencias de palabras de un corpus con doscientos artículos de química. La ponderación se realizó mediante una regresión lineal cuyos coeficientes fueron ajustados manualmente según los resultados obtenidos. En general, un modelo de regresión lineal simple es una herramienta estadística para identificar la relación entre una variable explicada y y un vector de variables explicativas $(x_1, x_2, \ldots, x_p)$. La ecuación (5) corresponde a la forma general de una regresión lineal simple en la cual y depende de $(x_1, x_2, \ldots, x_p)$.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{5}$$

Los datos de la campaña de anotación de la sección 7 nos sirvieron para determinar el vector de parámetros β que predice la eliminación de un segmento. Las variables explicativas de nuestro modelo corresponden esencialmente a una de cuatro categorías según su tipo: variables de informatividad (energía textual), variables de gramaticalidad (modelo probabilístico de lenguaje), variables de segmentación (segmentación discursiva) y variables de longitud. El cuadro 4 enlista las variables utilizadas. La idea es modelar la probabilidad de eliminación de los segmentos utilizando las variables del cuadro 4 y considerando las 60 844 decisiones de los anotadores en la variable de respuesta. Ésta variable corresponde a $\mathbf{P_{elim}}(s)$ de la ecuación (6) en la cual, un valor de $\mathbf{P_{elim}}(s) = 1$ indica que el segmento s fue eliminado sistemáticamente por todos mientras que $\mathbf{P_{elim}}(s) = 0$ indica que el s siempre fue retenido.

$$\mathbf{P_{elim}}(s) = \frac{\text{número de eliminaciones de } s}{\text{total de anotaciones de } s} \tag{6}$$

La ecuación (7) define un modelo lineal de predicción de la probabilidad de eliminar un segmento discursivo de una frase, la cual depende de las variables del cuadro 4. Un primer grupo de variables mide la informatividad ($\mathbf{Ener}(s,\varphi) = \beta_1 Es + \beta_2 E\varphi + \beta_3 \widetilde{E}$); el segundo grupo mide la gramaticalidad ($\mathbf{Gram}(s,\varphi) = \beta_4 Gs + \beta_5 G\varphi + \beta_6 \widetilde{G}$); el tercero mide el impacto de la segmentación ($\mathbf{Seg}(s,\varphi) = \beta_7 S + \beta_8 P + \beta_9 \widetilde{P}$); y el cuarto, el impacto de la longitud ($\mathbf{Lon}(s,\varphi) = \beta_{10} Ls + \beta_{11} L\varphi + \beta_{12} \widetilde{L}$).

$$\mathbf{P_{elim}}(s,\varphi) = \mathbf{Ener}(s,\varphi) + \mathbf{Gram}(s,\varphi) + \mathbf{Seg}(s,\varphi) + \mathbf{Lon}(s,\varphi) \tag{7}$$

Decimos que el modelo de la ecuación (7) es "completo" en el sentido de que todas las variables del cuadro 4 son utilizadas. La significación estadística de los modelos completos para DiSeg y CoSeg se

Cuadro 4: Lista de variables explicativas utilizadas para el ajuste de regresión lineal.

Var.	Tipo	Descripción
Es	Ener	La energía textual del segmento
$E\varphi$	Ener	La energía textual de la frase
$\widetilde{E}$	Ener	El cociente entre la energía textual del segmento y la de la frase: $Es/E\varphi$
Gs	Gram	El valor del segmento en el modelo de lenguaje de $n-$gramas
$G\varphi$	Gram	El valor de la frase en el modelo de lenguaje de $n-$gramas
$\widetilde{G}$	Gram	El cociente entre el valor del segmento en el modelo de lenguaje y el de la frase: $Gs/G\varphi$
S	Seg	El número total de segmentos en la frase
P	Seg	La posición del segmento en la frase
$\widetilde{P}$	Seg	La posición relativa del segmento en la frase: P/S
Ls	Lon	La longitud del segmento en número de palabras
$L\varphi$	Lon	La longitud de la frase en número de palabras
$\widetilde{L}$	Lon	El cociente entre la longitud del segmento y la de la frase: $Ls/L\varphi$

muestra en las tablas 5 y 6. Con el fin de facilitar la lectura, agregaremos los prefijos "D" o "C" según se trate de segmentos DiSeg o CoSeg.

Comparando los cuadros 5 y 6 se observa que el ajuste del modelo completo de CoSeg es superior al de DiSeg. También se observan diferencias de significación estadística de algunas variables. Por ejemplo, la energía textual de un segmento DiSeg no tiene impacto en la estimación de la probabilidad de eliminación, mientras que la energía de un segmento CoSeg es muy significativa. También, el número total de segmentos es apenas significativo para DiSeg, mientras que para CoSeg tiene más impacto. Por el contrario, las variables de longitud de segmento y de frase son más determinantes para DiSeg que para CoSeg. Estas diferencias enfatizan una de nuestras intuiciones iniciales: los segmentos a eliminar no deben ser muy largos. La probabilidad de eliminación de los segmentos DiSeg, que suelen ser largos, parece estar más influenciada por la longitud que por el contenido.

Cuadro 5: Modelo de regresión lineal completo para la eliminación de segmentos DiSeg.

	Parámetro	Desviación estándar	Estadístico t	$\mathbf{P}(>\|t\|)$	Significación
DEs	-0.027	0.016	-1.595	0.111	
$DE\varphi$	0.026	0.009	2.710	0.007	***
$D\widetilde{E}$	-0.008	0.092	-0.092	0.926	
DGs	0.005	0.002	1.897	0.058	*
$DG\varphi$	-0.003	0.001	-2.159	0.031	**
$D\widetilde{G}$	0.056	0.257	0.221	0.824	
DS	0.024	0.014	1.668	0.096	*
DP	-0.032	0.019	-1.642	0.101	
$D\widetilde{P}$	0.672	0.069	9.679	0.000	****
DLs	0.016	0.006	2.784	0.005	***
$DL\varphi$	-0.007	0.002	-2.699	0.007	***
$D\widetilde{L}$	-0.768	0.262	-2.929	0.003	***
Estadístico F= 107.1 con 12 y 433 grados de libertad, valor $p: < 2{,}2e-16$					
R^2= 0.695, R^2 ajustado=0.690					

Cuadro 6: Modelo de regresión lineal completo para la eliminación de segmentos CoSeg.

	Parámetro	desviación estándar	Estadístico t	$\mathbf{P}(>\|t\|)$	Significación
CEs	-0.057	0.017	-3.372	0.000	****
$CE\varphi$	0.057	0.006	8.488	0.000	****
$C\widetilde{E}$	0.033	0.100	0.337	0.736	
CGs	0.005	0.002	1.795	0.073	*
$CG\varphi$	-0.002	0.001	-2.425	0.015	**
$C\widetilde{G}$	0.099	0.199	0.502	0.615	
CS	-0.013	0.006	-2.192	0.028	**
CP	-0.004	0.007	-0.669	0.503	
$C\widetilde{P}$	0.395	0.051	7.720	0.000	****
CLs	0.015	0.006	2.297	0.021	**
$CL\varphi$	-0.004	0.002	-1.951	0.051	*
$C\widetilde{L}$	-0.584	0.220	-2.653	0.008	***
Estadístico F= 153.5 con 12 y 809 grados de libertad, valor $p: < 2{,}2e-16$					
R^2= 0.748, R^2 ajustado=0.741					

Surge ahora la cuestión de encontrar las regresiones lineales "óptimas", en el sentido que contengan exclusivamente variables explicativas significativas. Para ello hemos generado todas las regresiones posibles por medio de un contador binario. Una vez generadas, las regresiones fueron ordenadas, primero con respecto al coeficiente de determinación ajustado y luego por el estadístico F como segundo criterio.

Los resultados para los modelos lineales "óptimos" corresponden a los cuadros 7 y 8, de donde se

deduce la ecuación de predicción de la eliminación de un segmento DiSeg (ecuación (8)) y la ecuación de predicción de la eliminación de un segmento CoSeg (ecuación (9)). Se observa que los resultados de ambas segmentaciones mejoran ligeramente al eliminar las variables sin significación estadística, aunque CoSeg continua siendo superior. Esta vez, la diferencia más importante es que, en el modelo óptimo, la posición absoluta de un segmento CoSeg no afecta su probabilidad de eliminación, pero esto no ocurre para los segmentos DiSeg. Lo que sí es común es la influencia de la posición relativa de los segmentos.

A partir de estos resultados surgen dos conclusiones. La primera es que las variables elegidas para caracterizar la compresión de frases (segmentación discursiva, energía textual y modelos del lenguaje) son pertinentes para este propósito. No obstante, su uso podría ser mayormente explotado con modelos más complejos que la regresión lineal simple. La segunda es que, aún en los modelos óptimos, la varianza no está completamente explicada por las regresiones debido a la subjetividad inherente a la compresión de frases. La discrepancia en los resultados de la anotación del corpus tiene un efecto en la dispersión de las observaciones y, podemos suponer, que un efecto similar surgiría utilizando otras técnicas de aprendizaje supervisado sobre el mismo corpus.

Cuadro 7: Modelo de regresión lineal óptimo para la eliminación de segmentos DiSeg.

	Parámetro	desviación estándar	Estadístico t	$\mathbf{P}(>\|t\|)$	Significación
DEs	-0.028	0.006	-4.097	0.000	****
$DE\varphi$	0.027	0.008	3.163	0.001	***
DGs	0.004	0.001	2.584	0.010	**
$DG\varphi$	-0.002	0.001	-2.559	0.010	**
DS	0.024	0.013	1.838	0.066	*
DP	-0.031	0.018	-1.687	0.092	*
$D\widetilde{P}$	0.669	0.065	10.303	0.000	****
DLs	0.015	0.004	3.756	0.000	****
$DL\varphi$	-0.007	0.002	-3.142	0.001	***
$D\widetilde{L}$	-0.713	0.070	-10.070	0.000	****
Estadístico F= 129.1 con 10 y 435 grados de libertad, valor $p: < 2{,}2e-16$					
R^2= 0.696, R^2 ajustado=0.691					

Cuadro 8: Modelo de regresión lineal óptimo para la eliminación de segmentos CoSeg.

	Parámetro	desviación estándar	Estadístico t	$\mathbf{P}(>\|t\|)$	Significación
CEs	-0.053	0.005	-9.613	0.000	****
$CE\varphi$	0.059	0.005	10.508	0.000	****
CGs	0.003	0.001	2.024	0.043	**
$CG\varphi$	-0.002	0.001	-2.585	0.009	***
CS	-0.015	0.005	-2.961	0.003	***
$C\widetilde{P}$	0.369	0.029	12.451	0.000	****
CLs	0.013	0.005	2.589	0.009	***
$CL\varphi$	-0.004	0.001	-2.048	0.040	**
$C\widetilde{L}$	-0.446	0.068	-6.480	0.000	****
Estadístico F= 205.1 con 9 y 812 grados de libertad, valor $p: < 2{,}2e-16$					
R^2= 0.748, R^2 ajustado=0.753					

$$\hat{\mathbf{P}}_{\mathbf{elim\_DiSeg}}(s,\varphi) = -0{,}028DEs + 0{,}027DE\varphi + 0{,}004DGs - 0{,}002DG\varphi \\ +0{,}024DS - 0{,}031DP + 0{,}669D\widetilde{P} + 0{,}015DLs - 0{,}007DL\varphi - 0{,}713D\widetilde{L} \tag{8}$$

$$\hat{\mathbf{P}}_{\mathbf{elim\_CoSeg}}(s,\varphi) = -0{,}053CEs + 0{,}059CE\varphi + 0{,}003CGs + -0{,}002CG\varphi \\ -0{,}015CS + 0{,}369C\widetilde{P} + 0{,}013CLs - 0{,}004CL\varphi - 0{,}446C\widetilde{L} \tag{9}$$

9. Dos algoritmos de generación de resúmenes por eliminación de segmentos

Usando los resultados de la sección anterior se proponen dos algoritmos de resumen automático por compresión de frases. La parte medular de los algoritmos que aquí se presenta consiste en una expresión condicional del tipo **si** $(\widehat{\mathbf{P}_{\mathbf{elim}}}(s,\varphi) > \alpha)$ **entonces eliminar**(s), donde α es un valor entre cero y uno que representa el umbral de probabilidad con el que decidiremos eliminar o no un segmento en cada iteración y la estimación de la probabilidad de eliminación se calculará por medio de las ecuaciones (8) y 9 según el tipo de segmento.

El algoritmo 1 toma como argumentos el umbral de probabilidad α y el documento a resumir *Doc*. Primero *Doc* es segmentado en frases y luego en segmentos discursivos. Posteriormente, se decide para cada segmento si éste debe ser eliminado según el valor de $\widehat{\mathbf{P}_{\mathbf{elim}}}(s,\varphi)$. El algoritmo 1 termina cuando se han procesado todas las frases del documento y produce un resumen del mismo comprimiendo las frases.

Algoritmo 1 Generación de resúmenes por eliminación de segmentos.

Argumentos : (umbral de probabilidad $\alpha \in [0,1]$, Documento *Doc*)
$Segmentar_{\varphi}(Doc)$ //En frases
$Segmentar_{s}(Doc)$ //En segmentos discursivos
$CalcularVars(Doc)$ //Los valores del cuadro 4
para todo φ en *Doc* **hacer**
  para todo s en φ **hacer**
    si $(\widehat{\mathbf{P}_{\mathbf{elim}}}(s,\varphi) > \alpha)$ **entonces**
      $Eliminar(s)$ de φ
    fin si
  fin para
fin para
  devolver resumen

Una limitación del algoritmo 1 es su incapacidad para controlar la tasa de compresión τ. Frente a esto, se propone el algoritmo 2, en el cual se controla τ, de manera parcial, incrementando "lentamente" el valor de α en el bucle principal hasta que el resumen tenga el tamaño deseado. Note que debido a la naturaleza del método de compresión, no siempre es posible obtener de manera exacta la compresión deseada (en número de palabras) porque la unidad mínima eliminable es el segmento discursivo.

Algoritmo 2 Generación de resúmenes por eliminación de segmentos con tasa de compresión como argumento.

Argumentos : (τ deseada, Documento *Doc*)
$\alpha \leftarrow 0$
$Segmentar_{\varphi}(Doc)$ //En frases
$Segmentar_{s}(Doc)$ //En segmentos discursivos
$CalcularVars(Doc)$ //Los valores del cuadro 4
repetir
  para todo φ en *Doc* **hacer**
    para todo s en φ **hacer**
      si $(\widehat{\mathbf{P}_{\mathbf{elim}}}(s,\varphi) > \alpha)$ **entonces**
        $Eliminar(s)$ de φ
      fin si
    fin para
  fin para
  $\alpha \leftarrow \alpha + 0{,}01$
hasta que τ de resumen $\geq \tau$ deseada
  devolver resumen

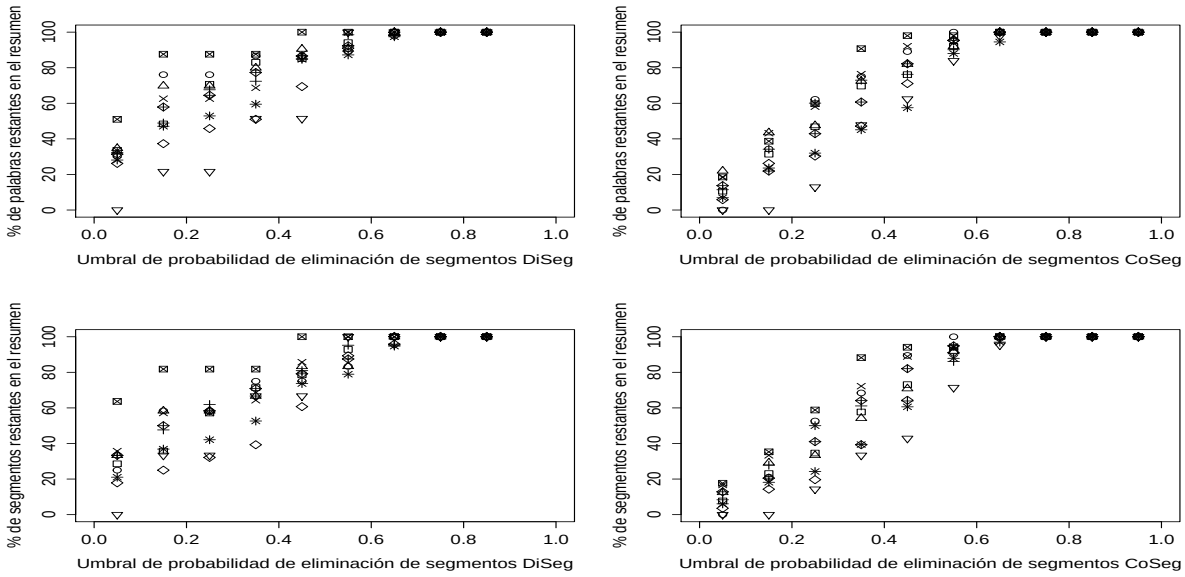


Figura 5: Tasa de compresión de nueve resúmenes automático en función del valor umbral de probabilidad de eliminación de segmentos discursivos.

Con estos dos algoritmos, hemos producido resúmenes con frases comprimidas para una serie de textos distintos a los utilizados en el ajuste de las regresiones. Para el algoritmo 1, produjimos nueve resúmenes por documento, variando el valor de α desde 0.05 hasta 0.95, con un incremento de 0.1. La figura 5 presenta la tasa de compresión en función de α para los nueve documentos. Los cuadros de arriba presentan la tasa de compresión en número de palabras y los de abajo en número de segmentos. A la izquierda se muestran los resultados para DiSeg y a la derecha para CoSeg. Observamos que en todos los casos, a partir del valor de cierto valor, los documentos ya no son comprimidos. En efecto, los modelos lineales fueron ajustados para emular el proceso de eliminación como lo harían los humanos, lo que refleja la tendencia a conservar la información a partir de un momento dado. Para el algoritmo 2 se produjeron nueve resúmenes por documento, variando τ desde el 10 % hasta el 90 % en incrementos del 10 %. Los resúmenes obtenidos son prácticamente los mismos que los del algoritmo 1; siendo la única diferencia el poder usar la tasa de compresión como argumento. El cuadro 9 presenta los resúmenes más cortos de un documento producidos por ambos algoritmos para los cuales el título forma parte del documento.

10. Evaluación de resumen automático mediante el test de Turing

Después de haber generado resúmenes, surge la problemática de la evaluación de los resultados. La evaluación en compresión de frases continua siendo de hecho un problema abierto puesto que las métricas clásicas de evaluación de resumen automático no pueden ser utilizadas para medir la calidad de frases comprimidas. Se ha observado en [33, 34] que ni ROUGE [25], ni FRESA [49] ni BLUEU [36] son sensibles a la gramaticalidad, uno de los criterios de la compresión de frases, por lo tanto no pueden ser utilizados para diferenciar una buena compresión de una mala. Frente a esta problemática y con la finalidad de exhibir la calidad de los resúmenes producidos, hemos propuesto una versión del test de Turing [50] en el cual una persona que funge como juez debe diferenciar entre un resumen elaborado por una máquina, usando nuestro método, y un resumen elaborado por otra persona durante la anotación de nuestro corpus. Cabe mencionar que al ser la primera vez que se utiliza este tipo de prueba para evaluar resúmenes automáticos, estamos interesados en realizar un cuidadoso análisis estadístico de los resultados de la apreciación de los jueces. Es por ello, que hemos diseñado un protocolo de evaluación en el cual es

más importante tener muchos jueces que tener muchos documentos.

Seis resúmenes humanos fueron elegidos aleatoriamente de los 2 877 resúmenes del corpus. Otros seis resúmenes producidos automáticamente y con alta gramaticalidad fueron elegidos: tres de ellos segmentados con DiSeg y tres segmentados con CoSeg. Para cada segmentador elegimos un resumen corto ($\tau < 50\,\%$), un resumen mediano ($\tau \approx 50\,\%$) y un resumen largo ($\tau > 50\,\%$) con respecto al tamaño del documento original. Nuestra única intervención fue la de asegurar que las frases en los resúmenes comenzaran por una mayúscula y terminaran por un punto. Los 54 jueces del test (todos hispanos con nivel superior de estudios) ignoraban toda esta información y no tenían acceso al documento de origen (para que no infirieran como funciona el programa); La única consigna fue la de determinar para cada resumen si éste había sido producido por un humano o por una máquina: el test de Turing aplicado al resumen automático.

Para saber si los resultados son significativos, se aplicó el test exacto de Fisher a las respuestas de cada uno de los jueces tomando como hipótesis nula (H_0) que la calidad de los resúmenes automáticos es similar a la de los resúmenes manuales. De entre los 54 jueces, solamente 1 obtuvo evidencia estadística suficiente para rechazar H_0 Para los 53 jueces restantes, no hay suficiente evidencia estadística para asociar sus respuestas con el origen verdadero de los resúmenes.

Verificamos mediante el mismo test si las respuestas de los 54 jueces estuvieron influenciadas por el tipo de segmentación. El cuadro 10 corresponde a la tabla de contingencia del número de veces que los jueces identificaron correctamente o incorrectamente el origen de los resúmenes según el segmentador. En este caso, H_0 corresponde a que la identificación es independiente del tipo de segmentador. El resultado da un valor $p = 0{,}496$ a 95 % de confianza. Por tanto, como $p > 0{,}05$, aceptamos H_0: el tipo de segmento, DiSeg o CoSeg, no hace más fácil la identificación del origen de los resúmenes.

Finalmente se evaluó la influencia de la tasa de compresión τ y del tamaño del documento en las respuestas de los jueces. Para ello utilizamos el test de χ^2, puesto que la tabla de contingencia asociada es de 3×2, el cuadro 11, y el test exacto de Fisher no puede utilizarse en este caso. Los resultados del test de χ^2 dan un valor $p = 0{,}055$, apenas superior al valor crítico para aceptar la hipótesis de que el tamaño del resumen final no tiene influencia en las respuestas. Para tener más elementos en el análisis observamos la varianza residual entre los valores observados y los esperados. El cuadro 12 muestra la desviación estándar de los residuos entre estos. Se puede observar que a medida que los resúmenes son más largos, la desviación de los residuos aumenta considerablemente. El origen de los resúmenes largos ($\tau > 50\,\%$) es el más difícil de identificar por los jueces, seguido del de los resúmenes medianos ($\tau \approx 50\,\%$) y al final el de los cortos ($\tau < 50\,\%$).

Cuadro 9: Ejemplos de resúmenes automáticamente generados por el método de compresión de frases.

Descubrimiento de mamut emociona a científicos (resumen) seg=CoSeg, $\alpha = 0{,}05$, $\tau = 24{,}8$
Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno.
Descubrimiento de mamut emociona a científicos (resumen) seg=DiSeg, $\alpha = 0{,}05$, $\tau = 33{,}3$
Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales.
Descubrimiento de mamut emociona a científicos (documento original)
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacer se con la enorme bestia. El hallazgo es raro porque al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Eso permite a los científicos recolectar polen y evidencias de otras plantas en el sitio de la excavación que podrían revelar detalles respecto del medio ambiente de Iowa hace más de 12000 años. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra para ver si pueden determinar qué tanto de los restos del mamut siguen enterrados. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales para proteger el sitio.

Cuadro 10: Identificación del origen de los resúmenes según el segmentador.

Segmentador	Origen correctamente identificado	Origen erróneamente identificado
DiSeg	45	63
CoSeg	19	35

Cuadro 11: Evaluación de la influencia de la tasa de compresión τ para la identificación del origen de los resúmenes.

Tasa de compresión relativa al documento	Correctamente identificado	Erróneamente identificado
$\tau < 50\,\%$	27	27
$\tau \approx 50\,\%$	30	24
$\tau > 50\,\%$	18	36

Cuadro 12: Residuos estandarizados de la desviación entre el valor esperado y observado de la identificación del origen de los resúmenes con diferente tasa de compresión.

Tasa de compresión relativa al documento	Correctamente identificado	Erróneamente identificado
$\tau < 50\,\%$	0.668	-0.668
$\tau \approx 50\,\%$	1.671	-1.671
$\tau > 50\,\%$	-2.339	2.339

11. Conclusiones y perspectivas a futuro

Hemos estudiado la compresión de frases por eliminación de segmentos discursivos bajo un enfoque de aplicación al resumen automático en español. Sin embargo, otras aplicaciones son posibles, ya que tanto la metodología como el marco teórico son flexibles para ser adaptados en otras direcciones y otros idiomas. Por ello, hemos puesto a disposición de la comunidad los datos generados en esta investigación, convencidos de que serán de valiosa utilidad para futuras investigaciones: [8].

Concluimos que la compresión de frases es ideal para establecer un puente desde el resumen por extracción hacia la generación de resúmenes abstractivos. También, que la compresión de frases tiene un cierto grado de subjetividad inherente que merece ser estudiado más a detalle. Creemos que se debe abrir el debate con respecto al tema de la subjetividad en el resumen automático.

Mostramos que la segmentación discursiva intra-frase sirve para generar frases comprimidas gramaticales e informativas. Contrario al enfoque clásico, observamos que no es necesario elaborar todo el análisis discursivo del texto para identificar segmentos eliminables. Así, ni la identificación de nuclearidad, ni la identificación del tipo de relación discursiva ni la creación del árbol discursivo son necesarios para este método. También hemos propuesto las bases de un segmentador discursivo orientado a la compresión de frases y hemos discutido la posibilidad de un segmentador multi-lingüe de pocos recursos lingüísticos.

Ante la falta de un método de evaluación, propusimos un test de Turing validado con un test combinatorio que abre una perspectiva prometedora, dado que este tipo de evaluación puede ser utilizado en otras tareas automáticas como la traducción, la simplificación, la reformulación y la paráfrasis.

Con respecto a los resultados de nuestro sistema de generación de resúmenes por compresión de frases, encontramos que los retos principales a analizar a futuro son la falta de gramaticalidad y la falta de cohesión de algunos resúmenes generados. El primero se refiere a garantizar que todas las frases del resumen como el resumen en su totalidad sean gramaticales. Basta con un pequeño error gramatical para que la calidad final del resumen sea puesta en duda. Pero garantizar la gramaticalidad no es posible usando

[8] `http://molina.talne.eu/sentence_compression/data/`

un modelo probabilístico como lo hemos hecho hasta ahora. En investigaciones futuras se deben incluir otros elementos, además del conteo de n-gramas. El segundo reto es debido a dos razones principales: 1) la supresión excesiva de marcadores discursivos contenidos en los segmentos eliminados y 2) la incapacidad del método de identificar sinónimos y paráfrasis. Para evitar la eliminación excesiva de marcadores tendríamos primero que conocer su función en la frase, es decir, a qué tipo de relación corresponde. Aunque la identificación del tipo de relación no es trivial, existen avances en esta dirección. Con respecto a la incapacidad de identificación de sinónimos, podemos decir que la energía textual simplemente relaciona los lemas de las palabras en un documento y por tanto es incapaz de reconocer que dos símbolos distintos correspondan al mismo concepto. Esta cuestión puede mejorar considerando modelos más robustos que relacionen las palabras de manera semántica.

12. Agradecimientos

Este trabajo fue realizado durante la tesis doctoral del autor, auspiciado por la beca doctoral de Conacyt (México) 211963 (y parcialmente financiado por el proyecto Conacyt 178248) bajo la dirección de Juan-Manuel Torres-Moreno (Laboratoire Informatique d'Avignon, Francia), y co-dirigido por Eric San-Juan (Laboratoire Informatique d'Avignon, Francia) y Gerardo Sierra (UNAM, Instituto de Ingénieria, México).

Referencias

[1] S.D. Afantenos, P. Denis, and L. Danlos. Learning recursive segments for discourse parsing. *Cornell University ArXiv:1003.5372,Computation and Language (cs.CL)*, 2010.

[2] Sandra M Aluísio, Lucia Specia, Thiago AS Pardo, Erick G Maziero, and Renata PM Fortes. Towards brazilian portuguese automatic text simplification systems. In *8th ACM symposium on Document engineering*, pages 240–248. ACM, 2008.

[3] Irene Castellón, Montse Civit, and Jordi Atserias. Syntactic parsing of unrestricted spanish text. In *Proceedings First International Conference on Language Resources and Evaluation (LREC'98), Granada, Spain*, 1998.

[4] S.F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393, 1999.

[5] J. Clarke and M. Lapata. Constraint-based sentence compression: An integer programming approach. In *COLING/ACL'06 on Main Conference Poster Sessions*, pages 144–151, Sydney, Australie, 2006.

[6] Michael John Collins. A new statistical parser based on bigram lexical dependencies. In *34th annual meeting on Association for Computational Linguistics*, pages 184–191. ACL, 1996.

[7] Iria da Cunha, Eric SanJuan, Juan-Manuel Torres-Moreno, Marina Lloberes, and Irene Castellón. Discourse segmentation for spanish based on shallow parsing. In Grigori Sidorov, Arturo Hernández Aguirre, and Carlos Reyes García, editors, *Advances in Artificial Intelligence*, volume 6437 of *Lecture Notes in Computer Science*, pages 13–23. Springer Berlin / Heidelberg, 2010.

[8] Iria da Cunha, Eric SanJuan, Juan-Manuel Torres-Moreno, Marina Lloberes, and Irene Castellón. Diseg 1.0: The first system for spanish discourse segmentation. *Expert Systems with Applications*, 39(2):1671–1678, 2012.

[9] Walter Daelemans, Anja Höthker, and Erik Tjong Kim Sang. Automatic sentence simplification for subtitling in dutch and english. In *4th International Conference on Language Resources and Evaluation*, pages 1045–1048, 2004.

[10] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[11] H. P. Edmundson. New Methods in Automatic Extraction. *Journal of the Association for Computing Machinery*, 16(2):264–285, 1969.

[12] G. Erkan and D. R. Radev. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479, 2004.

[13] Silvia Fernández. *Applications exploratoires des modèles de spins au Traitement Automatique de la Langue*. doctorat en physique statistique, Département de Physique de la Matière et des Matériaux, Université Henri Poincaré, Nancy, France, 2009.

[14] Silvia Fernández, Eric SanJuan, and Juan-Manuel Torres-Moreno. Energie textuelle des mémoires associatives. In *Proceedings de la conférénce Traitement Automatique de la Langue Naturelle (TALN'07)*, volume 1, pages 25–34, Toulouse, France, 5-8 Juin 2007.

[15] Silvia Fernández, Eric SanJuan, and Juan-Manuel Torres-Moreno. Textual Energy of Associative Memories: performants applications of Enertex algorithm in text summarization and topic segmentation. In *Mexican International Conference on Artificial Intelligence (MICAI'07)*, pages 861–871, Aguascalientes, Mexique, 2007. Springer-Verlag.

[16] Silvia Fernández and Juan-Manuel Torres-Moreno. Une approche exploratoire de compression automatique de phrases basée sur des critères thermodynamiques. In *Proceedings de la conférénce Traitement Automatique de la Langue Naturelle (TALN'09)*, Senlis, France, 24-26 Juin 2009.

[17] G. Grefenstette. Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind. In *AAAI Spring Symposium on Intelligent Text summarization (Working notes)*, pages 111–118, Stanford University, CA, Etats-Unis, 1998.

[18] Hongyan Jing and Kathleen R McKeown. The decomposition of human-written summary sentences. In *22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 129–136. ACM, 1999.

[19] Yufeng Jing and W Bruce Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO*, volume 94, pages 146–160, 1994.

[20] Siddhartha Jonnalagadda and Graciela Gonzalez. Sentence simplification aids protein-protein interaction extraction. *Cornell University ArXiv:1001.4273*, 2010.

[21] Nongnuch Ketui, Thanaruk Theeramunkong, and Chutamanee Onsuwan. A rule-based method for thai elementary discourse unit segmentation (ted-seg). In *Knowledge, Information and Creativity Support Systems (KICSS), 2012 Seventh International Conference on*, pages 195–202. IEEE, 2012.

[22] Dan Klein and Christopher D Manning. Accurate unlexicalized parsing. In *41st Annual Meeting on Association for Computational Linguistics*, volume 1, pages 423–430. ACL, 2003.

[23] Kevin Knight and Daniel Marcu. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107, Juillet 2002.

[24] Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25(2):259–284, 1998.

[25] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In Marie-Francine Moens and Stan Szpakowicz, editors, *Workshop Text Summarization Branches Out (ACL'04)*, pages 74–81, Barcelone, Espagne, Juillet 2004. ACL.

[26] W. C. Mann and S. A. Thompson. *Rhetorical Structure Theory: A Theory of Text Organization*. Information Sciences Institute, Marina del Rey, 1987.

[27] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, 1999.

[28] D. Marcu. *The rhetorical parsing, summarization, and generation of natural language texts.* PhD thesis, Computer Science, University of Toronto, Toronto, Canada, 1998.

[29] Daniel Marcu. *The Theory and Practice of Discourse Parsing Summarization.* MIT Press, Cambridge, 2000.

[30] E. Maziero, T. Pardo, and M. Nunes. Identificação automática de segmentos discursivos: o uso do parser palavras. Série de relatórios do núcleo interinstitucional de lingüística computacional, Universidade de Sao Paulo, São Carlos, Brésil, 2007.

[31] Ryan McDonald. Discriminative sentence compression with soft syntactic evidence. In *Proceedings of EACL*, volume 6, pages 297–304, 2006.

[32] Alejandro Molina. Sistemas web colaborativos para la recopilación de datos bajo el paradigma de ciencia ciudadana. *Komputer Sapiens*, 1:6–8, Janvier-Juin 2013.

[33] Alejandro Molina, Iria da Cunha, Juan-Manuel Torres-Moreno, and Patricia Velazquez-Morales. La compresión de frases: un recurso para la optimización de resumen automático de documentos. *Linguamática*, 2(3):13–27, 2010.

[34] Alejandro Molina, Juan-Manuel Torres-Moreno, Iria da Cunha, Eric SanJuan, and Gerardo Sierra. Sentence compression in spanish driven by discourse segmentation and language models. *Cornell University ArXiv:1212.3493, Computation and Language (cs.CL), Information Retrieval (cs.IR)*, 2012.

[35] Alejandro Molina, Juan-Manuel Torres-Moreno, Eric SanJuan, Iria da Cunha, Gerardo Sierra, and Patricia Velázquez-Morales. Discourse segmentation for sentence compression. In *Advances in Artificial Intelligence*, LNCS, pages 316–327. Springer-Verlag, Berlin, Heidelberg, 2011.

[36] K. Papineni, S. Roukos, T. Ward, , and W. J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 311–318, Philadelphie, PA, Etats-Unis, 6-12 Juillet 2002. ACL.

[37] Emily Pitler. Methods for sentence compression. Technical Report MS-CIS-10-20, University of Pennsylvania, 2010.

[38] Rémy Saksik, Alejandro Molina, Linhares Andréa, and Torres-Moreno. Segmentaçao discursiva automática: uma avaliaçao preliminar em francês. In *4th Meeting RST and Discourse Studies, STIL 2013 Symposium in Information and Human Language Technology*, 2013.

[39] G. Salton. *The SMART Retrieval System – Experiments un Automatic Document Processing.* Prentice-Hall, Englewood Cliffs, 1971.

[40] Daniel DK Sleator and Davy Temperley. Parsing english with a link grammar. *arXiv preprint cmp-lg/9508004*, 1995.

[41] Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *HLT-NAACL*, pages 149–156, Edmonton, Canada, 2003.

[42] Caroline Sporleder and Mirella Lapata. Discourse chunking and its application to sentence compression. In *conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 257–264, Stroudsburg, PA, USA, 2005. ACL.

[43] Karen Spärck-Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 1(28):11–21, 1972.

[44] Josef Steinberger and Karel Jezek. Sentence compression for the lsa-based summarizer. In *7th International conference on information systems implementation and modelling*, pages 141–148, 2006.

[45] Josef Steinberger and Roman Tesar. Knowledge-poor multilingual sentence compression. In *7th Conference on Language Engineering (SOLE'07)*, pages 369–379, Le Caire, Egypte, 2007.

[46] A. Stolcke. Srilm – an extensible language modeling toolkit. In *Intl. Conf. on Spoken Language Processing*, volume 2, pages 901–904, Denver, 2002.

[47] Milan Tofiloski, Julian Brooke, and Maite Taboada. A syntactic and lexical-based discourse segmenter. In *ACL-IJCNLP*, pages 77–80, 2009.

[48] Juan-Manuel Torres-Moreno. *Résumé automatique de documents : une approche statistique.* Hermès-Lavoisier, France, 2012.

[49] Juan-Manuel Torres-Moreno, Horacio Saggion, Iria da Cunha, and Eric SanJuan. Summary Evaluation With and Without References. *Polibits: Research journal on Computer science and computer engineering with applications*, 42:13–19, 2010.

[50] Alan M Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

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

[52] Michael J. Witbrock and Vibhu O. Mittal. Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries. In *22nd Conference SIGIR'99*, pages 315–316, Berkeley, CA, Etats-Unis, 15-19 Aout 1999. ACM.

[53] M. Yousfi-Monod and V. Prince. Compression de phrases par élagage de l'arbre morpho-syntaxique. *Technique et Science Informatiques*, 25(4):437–468, 2006.