Red Temática
en Tecnologías
del Lenguaje

“Creación de corpus en español mexicano para la generación de resúmenes”

Que Presenta

M.C.C. Griselda Areli Matias Mendoza

Tutor Académico:

Dra. Yulia Nikolaevna Ledeneva

TIANGUISTENCO, MÉXICO

Contenido

Introducción

En este documento se describe la construcción de un corpus de textos en español para resúmenes, construido con el fin de servir como apoyo al área de Procesamiento del Lenguaje Natural (PLN) en el idioma español, principalmente en el área de la generación de resúmenes automáticos. El corpus fue creado bajo el proyecto de la Red Temática en Tecnologías del Lenguaje (Red TTL).

El corpus está compuesto por noticias en español mexicano y dos resúmenes generados por dos humanos. El objetivo principal del corpus es servir como corpus para evaluar los métodos del estado del arte y las herramientas comerciales para la generación de resúmenes extractivos. Sin embrago, puede ser utilizado para diversos fines como, análisis lingüísticos de textos, ya sea para resúmenes o solo análisis de texto, sistemas de recuperación de información o detección de tópicos.

Actualmente existen métodos del estado del arte que trabajan con la generación de resúmenes extractivos, como (Ledeneva, 2008), (Ledeneva, 2008ª), (García, 2008), (Montiel, 2009), (García, 2013) (Mendoza, 2014), (Meena, 2015), (Bhargava, 2016), entre otros. Sin embargo, todos ellos trabajan sólo para el lenguaje inglés. Existen otros que son independientes del lenguaje y prueban más de una colección de documentos, como, (Mihalcea, 2005), (Patel, 2007), (Last, 2010), (Saggion, 2011). Pero, a pesar de probar con más de una colección de documentos en lenguajes como inglés, portugués, chino, entre otros, dejan de lado uno de los más importantes lenguajes, el español.

Según (Cervantes, 2013), los expertos predicen que para el año 2050 habrá más de 530 millones de hispanohablantes, de los cuales 100 millones estarán viviendo en los Estados Unidos. Esto nos muestra un amplio campo de trabajo para el PLN en español. Por eso la importancia de contar con un corpus en el lenguaje español, pero que además sea en español mexicano para poder conocer mejor el comportamiento de los diferentes métodos y herramientas para la generación de resúmenes extractivos en nuestro lenguaje.

Corpus de textos en español para resúmenes

1 Características generales

El corpus creado es un corpus en español mexicano, exclusivo para la generación de resúmenes extractivos para un sólo documento. El corpus se presenta en formato digital sobre noticias periodísticas.

1.1. Protocolo de compilación

1.1.1. Búsqueda y acceso a la información

El corpus fue creado a partir de noticias electrónicas disponibles en la red. Las noticias fueron obtenidas de la página oficial del periódico "Crónica" (Crónica, s.f.). Se seleccionaron 20 noticias de las siguientes categorías, Academia, Bienestar, Ciudad, Cultura, Deportes, Espectáculos, Estados, Mundo, Nacional, negocios, Opinión y Sociedad. Dando un total de 240 noticias. Las noticias seleccionadas fueron noticias del mes de abril de 2014. Una de las consideraciones más importantes para la selección de las noticias fue que tuvieran diferentes longitudes, pero siempre más de 100 palabras.

1.1.2. Pre-procesamiento

Las noticias se descargaron de la web en un formato .html, por lo que se realizó el siguiente proceso de limpieza y normalización.

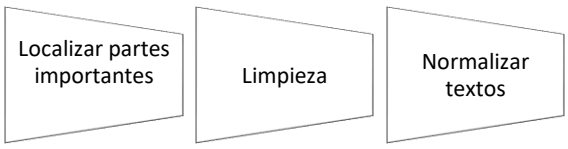

Figura 1. Fases del pre-procesamiento

- Localizar partes importantes. Las noticias que están disponibles en la red, además del texto de la noticia, pueden contener más información como, anuncios, fotografías, ligas a otras páginas, etc. Por esto fue necesario detectar las partes que nos brindaran información relevante y necesaria para la construcción del corpus. Los segmentos que se eligieron son, la clave de la noticia, la cual consiste en un número único que la identifica y que también forma

parte del nombre del archivo, el título de la noticia, la categoría a la que pertenece, la fecha de publicación y el texto de la noticia.

- Limpieza. El proceso de limpieza consiste en eliminar todas las etiquetas html, texto, ligas, imágenes, etc. dejando solamente el título de la noticia, la categoría a la que pertenece, la fecha de publicación y el texto de la noticia. La limpieza de los textos se realizó mediante un programa en java utilizando expresiones regulares para que se realizará de manera automática para todos los textos.
- Normalizar textos. Una vez limpios los textos, se realizó el etiquetado para poder identificar de manera más sencilla las partes de la noticia. A continuación en la tabla 1 se describen las etiquetas utilizadas y posterior de se da un ejemplo de etiquetado.

Tabla 1. Descripción de etiquetas utilizadas para etiquetar los textos completos.

Etiquetas	Descripción
<DOC></DOC>	Etiqueta que indica el inicio y final de documento
<DOCNO> </DOCNO>	Etiqueta que indica el nombre del documento
<FILEID></FILEID>	Etiqueta que indica un número único del documento
<TITLE></TITLE>	Etiqueta que indica el título del documento
<CATEGORY> </ CATEGORY >	Etiqueta que indica la categoría a la que pertenece el documento
<DATE></DATE>	Etiqueta que indica la fecha de expedición del documento
<TEXT></TEXT>	Etiqueta que indica cual es el texto a resumir
<s><\s>	Etiquetas que indican el inicio y fin de una oración

Tabla 2. Ejemplo de texto completo etiquetado

```
<DOC>
<DOCNO>
<s docid="09ED020414_825542" num="1" wdcount="1"> 825542 </s>
</DOCNO>
<FILEID>
<s docid="09ED020414_825542" num="2" wdcount="1">09ED020414_825542</s>
</FILEID>
<HEAD>
<s docid="09ED020414_825542" num="3" wdcount="10"> Atención digna a grupos vulnerables distingue a Toluca, afirma alcaldesa </s>
</HEAD>
<CATEGORY>
```

<s docid="09ED020414_825542" num="4" wdcount="1"> Estados </s>
</CATEGORY>
<DATE>
<s docid="09ED020414_825542" num="5" wdcount="1"> 2014-04-02 </s>
</DATE>
<TEXT>
<s docid="09ED020414_825542" num="6" wdcount="61"> La alcaldesa de Toluca, Martha Hilda González Calderón, encabezó la entrega de auxiliares auditivos, sillas de ruedas y lentes, en beneficio de 240 personas del municipio, acompañada por la presidenta del sistema DIF Toluca, Diana Elisa González Calderón, y el representante del secretario de Salud de la entidad, César Nomar Gómez Monge, Pedro Montoya Moreno, coordinador de Salud de dicha secretaría</s>
<s docid="09ED020414_825542" num="7" wdcount="82"> En su mensaje, González Calderón refrendó su compromiso con las personas con discapacidad y dijo que la atención a grupos vulnerables distingue a Toluca como Municipio Educador, además de que hay la responsabilidad de brindar las herramientas necesarias a aquellos grupos con algún grado de vulnerabilidad para contribuir a mejorar su calidad de vida Reconoció el apoyo y coordinación de la Secretaría de Salud estatal que, dijo, se traduce en mejores condiciones para los toluqueños y las toluqueñas que más lo necesitan</s>
<s docid="09ED020414_825542" num="8" wdcount="40"> En presencia de miembros del Cabildo y de autoridades municipales, la presidenta del sistema DIF de Toluca, Diana Elisa González Calderón, indicó que en esta ocasión se entregaron 110 auxiliares auditivos, 100 juegos de lentes y 30 sillas de ruedas.</s>
</TEXT>
</DOC>

1.1.3. Almacenamiento

Para nombrar a cada uno de los archivos se siguieron las siguientes consideraciones.

- Considerando que son 20 archivos por categoría se asignó un número consecutivo (1-20), posterior a esto se tomaron dos letras para cada categoría, a continuación se describen, Academia (AC), Bienestar (BI), Ciudad (CI), Cultura (CU), Deportes (DE), Espectáculos (ES), Estados (ED), Mundo (MU), Nacional (NA), Negocios (NE), Opinión (OP) y Sociedad (SO). Seguido de la abreviación de la categoría se colocó la fecha de la noticia y finalmente separado por un guion bajo la clave de la noticia.

Ejemplo de nombre de archivo. 01AC010414_825278.txt.

Finalmente, lo que se tiene son 12 carpetas con 20 archivos cada una, dando un total de 240 archivos.

2 Construcción de los resúmenes

Una vez construido el corpus de noticias en español, se crearon para cada archivo dos resúmenes hechos por dos humanos.

2. Selección de los humanos

Las consideraciones tomadas para seleccionar a un humano fueron, que tuviera nacionalidad mexicana, educación mínima de licenciatura y se le dieron las siguientes indicaciones.

2.1. Construcción de los resúmenes

A los humanos se les proporciono la noticia dividida en oraciones con el número de palabras de cada una de ellas. Se le pidió leyera completamente la noticia y seleccionara las oraciones que consideraba importantes. De las oraciones seleccionadas, se le pidió que creara un resumen mayor a 100 palabras. En el apartado de Anexos se muestra un ejemplo de las instrucciones dadas a los humanos, además de la lista con los nombres de cada uno de ellos.

A continuación en la tabla 2 se describen las etiquetas utilizadas para etiquetar los resúmenes generados por los humanos y posterior de se da un ejemplo de etiquetado.

Tabla 3. Descripción de etiquetas utilizadas para etiquetar los resúmenes.

Etiquetas	Descripción
<SUM></SUM>	Etiqueta que indica el inicio y final Del resumen hecho por el humano
CATEGORY	Indica la categoría a la que pertenece la noticia
TYPE	Indica el tipo de resumen, en este caso es por documento
SIZE	Indica el tamaño mínimo de palabras que debe tener el resumen
DOCREF	Muestra el nombre del documento base para la generación del resumen extractivo
SELECTOR	Indica las iniciales del humano que realizó el resumen
SUMMARIZER	Indica cuál de los dos resúmenes generados es. A- el primero, B- el segundo.

Tabla 4. Ejemplo de resumen etiquetado

<SUM
CATEGORY="ESTADOS"
TYPE="PERDOC"
SIZE="100"
DOCREF="09ED020414_825542"
SELECTOR="EX"
SUMMARIZER="B">
La alcaldesa de Toluca, Martha Hilda González Calderón, encabezó la entrega de auxiliares auditivos, sillas de ruedas y lentes, en beneficio de 240 personas del municipio, acompañada por la presidenta del sistema DIF Toluca, Diana Elisa González Calderón, y el representante del secretario de Salud de la entidad, César Nomar Gómez

Monge, Pedro Montoya Moreno, coordinador de Salud de dicha secretaría. En su mensaje, González Calderón refrendó su compromiso con las personas con discapacidad y dijo que la atención a grupos vulnerables distingue a Toluca como Municipio Educador, además de que hay la responsabilidad de brindar las herramientas necesarias a aquellos grupos con algún grado de vulnerabilidad para contribuir a mejorar su calidad de vida Reconoció el apoyo y coordinación de la Secretaría de Salud estatal que, dijo, se traduce en mejores condiciones para los toluqueños y las toluqueñas que más lo necesitan.
</SUM>

2.2. Recopilación de los resúmenes generados por los humanos

Una vez creado el resumen extractivo por el humano, se asignó una clave a cada uno de ellos, para nombrar los archivos de resumen de la siguiente manera.

Ejemplo de nombre de archivo: SUM_01AC010414_825278_LX.sum

Como se pude observar, para identificar que el archivo pertenece a los resúmenes modelos se agregó la etiqueta SUM a cada uno de ellos, seguido de esto se colocó el nombre de la noticia original y finalmente se agregó la clave asignada al humano. La extensión de estos archivos es .sum.

Finalmente lo que se tiene son 12 carpetas con 40 archivos cada una, dando un total de 480 archivos de resúmenes modelo.

3 Descripción del corpus

Como se había mencionado el corpus está compuesto por 240 noticias de diferentes categorías. La colección se presenta de forma etiquetada, las cuales describen en que consiste cada parte que compone el texto. Es importante mencionar que el corpus fue dividido en oraciones, las cuales también son etiquetadas, para facilitar el análisis del texto.

A continuación en la tabla 3 se muestra, las categorías en las que está dividido el corpus, el número de documentos que lo componen y el número de oraciones.

Tabla 5. Características de los textos completos del corpus.

Periódico	Categoría	Número de textos	Número de palabras	Promedio de palabras	Número de Oraciones	Promedio de oraciones
Crónica	Academia	20	10966	548,3	382	19,1
	Bienestar	20	11801	590,05	405	20,25
	Ciudad	20	7568	378,4	219	10,95
	Cultura	20	8631	431,55	297	14,85
	Deportes	20	9519	475,95	363	18,15
	Espectáculos	20	8869	443,45	311	15,55
	Estados	20	7471	373,55	185	9,25
	Mundo	20	7108	355,4	247	12,35
	Nacional	20	7533	376,65	186	9,3

	Negocios	20	7523	376,15	229	11,45
	Opinión	20	12716	635,8	443	22,15
	Sociedad	20	6507	325,35	228	11,4
	Total	240	106212		3495	
	Media			442,55		14,5625

Los resúmenes generados por los humanos, son de más de 100 palabras. Sin embargo, para la evaluación de los métodos y las herramientas para la generación de resúmenes se recomienda que la evaluación se realice a 100 palabras.

4 Organización del corpus

El corpus está organizado de la siguiente manera.

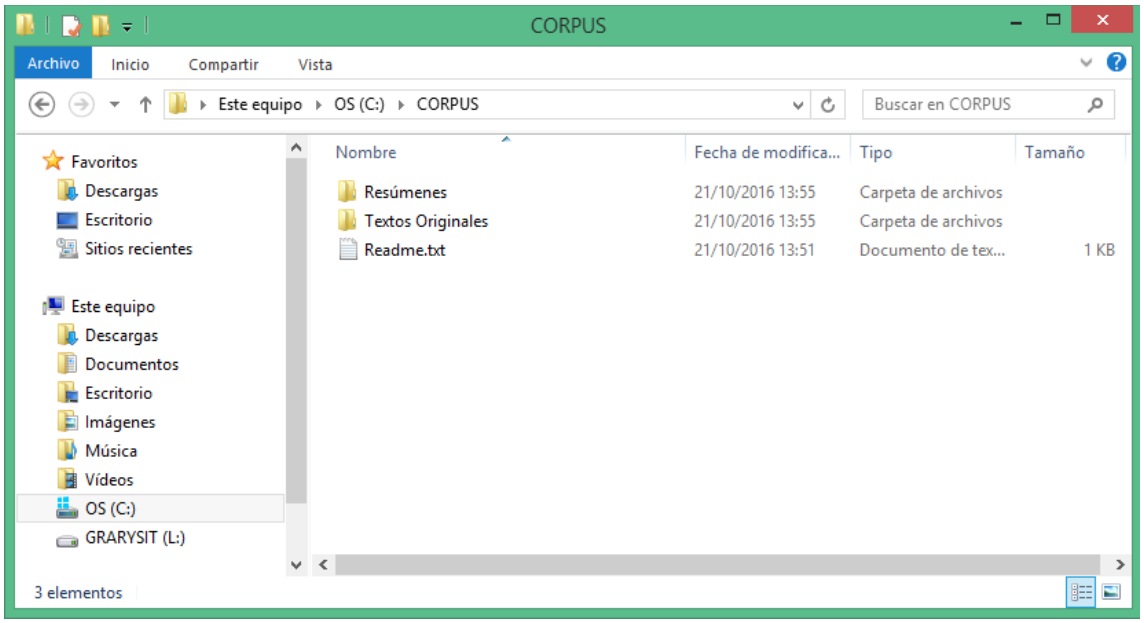

Figura 2. Directorio de corpus.

Como se puede observar en la figura 2, el corpus está compuesto por dos carpetas. Resúmenes, donde se localizan los dos resúmenes hechos por los dos humanos para cada uno de los documentos originales. Finalmente como se muestran en la figura 3, la carpeta Textos Originales, dividida en dos carpetas, Textos por archivos y Textos por categoría, donde se encuentras los textos completos originales etiquetados.

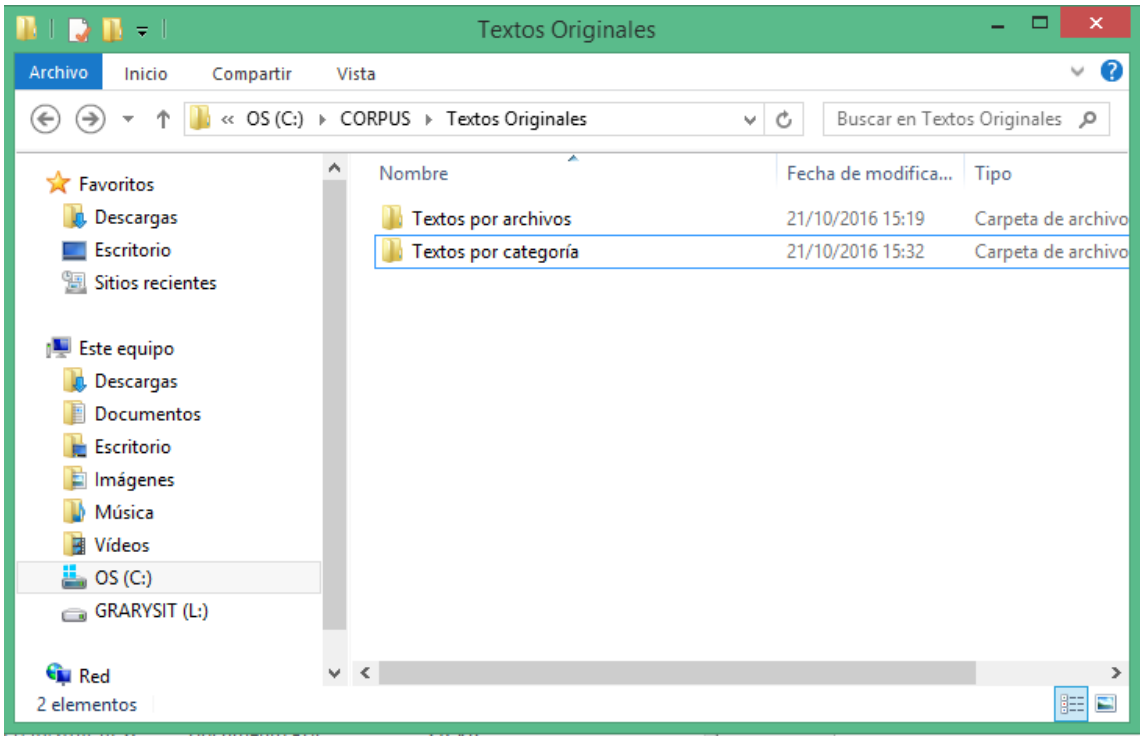

Figura 3. Directorio de Textos Originales.

La única diferencia entre estas dos últimas carpetas es que en la carpeta Textos por categoría, se encuentran 12 carpetas, una para cada categoría del corpus, en la que se encuentran 20 archivos pertenecientes a la categoría. Mientras que en la carpeta Texto por archivos, se encuentra los 240 archivos. En la figura 4. se muestra el contenido de estas carpetas.

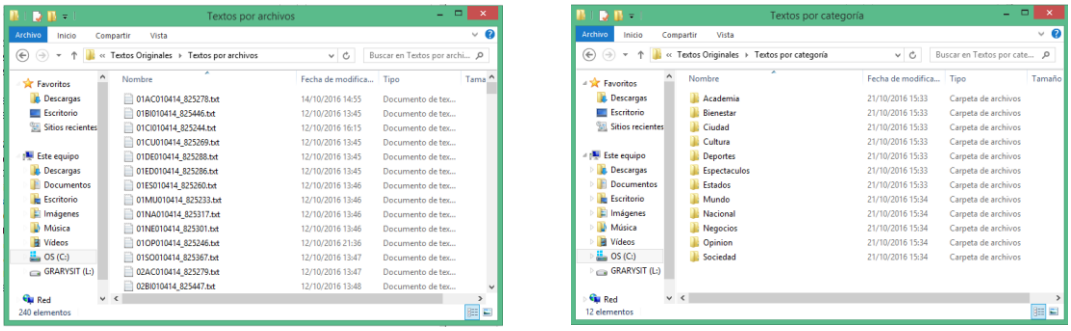

Figura 4. Directorio de textos por archivos y textos por categoría

Consideraciones finales

Como se mencionó anteriormente el corpus propuesto en este trabajo esta contribuido para ser utilizado principalmente en el estudio de la generación de resúmenes extractivos para el lenguaje español. El corpus presentado esta etiquetado de tan manera que sólo se presentan los textos completos y los resúmenes en diferentes

carpetas. Sin embargo, el etiquetado permite eliminar las partes que no se consideren importantes para el análisis de esta colección, por ejemplo si se desea sólo trabajar con el texto se considera sólo lo contenido en la etiqueta <TEXT></TEXT>. Una de las aportaciones importantes de este trabajo es que el texto está separado por oraciones, lo que muestra una estandarización para futuros usos.

Agradecimientos

El presente trabajo fue realizado bajo la supervisión de la Dra. Yulia Ledeneva, bajo el proyecto de la Red Temática en Tecnologías del Lenguaje (Red TTL). Gracias al apoyo de Consejo de Ciencia y Tecnología (CONACYT).

Referencias Bibliográficas

Bhargava, 2016	Bhargava, R., Sharma, Y., & Sharma, G. (2016). ATSSI: Abstractive Text Summarization Using Sentiment Infusion. Procedia Computer Science, 89, 404-411.
Crónica. (s.f.)	© La Crónica Diaria S.A. de C.V. Obtenido de © La Crónica Diaria S.A. de C.V: http://www.cronica.com.mx/noticias.php
García, 2008	arcía R., Montiel, R., Ledeneva, Y., Rendón, e., Gelbukh, A. & Cruz, R. (2008). Text Summarization by Senten ce Extraction Using Unsupervised Learning. 7° Conferencia Internacional Mexicana de Inteligencia Artificial (MICAI08); Notas de la conferencia de Inteligencia Artificial, Springer-Verlag, Vol 5317, pp133-143.
García, 2013	García-Hernández, R. A., & Ledeneva, Y. (2013, June). Single extractive text summarization based on a genetic algorithm. In Mexican Conference on Pattern Recognition (pp. 374-383). Springer Berlin Heidelberg.
Last, 2010	Last, M. & Litvak, M. (2010). Language-independent Techniques for Automated Text Summarization. NATO Science for Peace and Security Series - D: Information and Communication Security. Vol. 27: Web Intelligence and Security, pp. 207-237.

Ledeneva, 2008	Ledevena, Y. N. (2008). Automatic Language-Independent Detection of Multiword Descriptions for Text Summarization. México, D.F.: Presentada en el Instituto Politécnico Nacional, para obtención del grado de Doctor.
Ledeneva, 2008a	Ledeneva, Y., Gelbukh, A. & García, R. (2008). Keeping Maximal Frequent Sequences Facilitates Extractive Summarization. Research in Computing Science, Vol. 34, pp.163-174.
Meena, 2015	Meena, Y. K., & Gopalani, D. (2015). Feature Priority Based Sentence Filtering Method for Extractive Automatic Text Summarization. Procedia Computer Science, 48, 728-734.
Mendoza, 2014	Mendoza, M., Bonilla, S., Noguera, C., Cobos, C., & León, E. (2014). Extractive single-document summarization based on genetic operators and guided local search. Expert Systems with Applications, 41(9), 4158-4169.
Mihalcea, 2005	Mihalcea, R. & Taran, P.. (2005). A Language Independent Algorithm for Single and Multiple Document Summarization. Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Vol. 1, pp. 602-607.
Montiel, 2009	Montiel, R. (2009). Generación automática de resúmenes mediante aprendizaje no supervisado. Edo. de México: Presentada en el Instituto Tecnológico de Toluca, para obtención del Título de Ingeniero en Sistemas Computacionales.
Patel, 2007	Patel, A., Siddiqui, T & Tiwary, U. (2007). A language independent approach to multilingual text summarization. Conference RIA2007, Pittsburgh PA, U.S.A., 123-132.
Saggion, 2011	Saggion, H., Szasz, S., & Grupo, T. A. L. N. (2011). A Bilingual Summary Corpus for Information Extraction and other Natural

Language Processing Applications. on Iberian Cross-Language Natural Language Processings Tasks (ICL 2011), 28.