Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 303-311 recibido 1-05-2009; aceptado 5-06-2009

Comparación de Tres Modelos de Texto para la Generación Automática de Resúmenes

Comparison of Three Text Models for Automatic Generation of Summaries

Romyna Montiel Soto
Laboratorio de Reconocimiento de Patrones
Instituto Tecnológico de Toluca
Av. Tecnológico s/n, Ex Rancho La Virgen,
Metepec, México C.P. 52140
romyna.montiel@yahoo.com.mx

René Arnulfo García-Hernández
Departamento de Ingeniería de Software
Unidad Académica Profesional Tianguistenco
Universidad Autónoma del Estado de México,
http://scfi.uaemex.mx/~ragarcia/
renearnulfo@hotmail.com

Yulia Ledeneva
Departamento de Ingeniería de Software
Unidad Académica Profesional Tianguistenco
Universidad Autónoma del Estado de México,
yledeneva@yahoo.com
http://ledeneva.com

Rafael Cruz Reyes
Laboratorio de Reconocimiento de Patrones
Instituto Tecnológico de Toluca
Av. Tecnológico s/n, Ex Rancho La Virgen,
Metepec, México C.P. 52140
tectoluca@hotmail.com

Resumen: Uno de los principales problemas en la generación automática de resúmenes de texto consiste en identificar, independientemente del idioma y dominio, la información más importante en el documento origen. Para este problema, una gran cantidad de trabajos han aplicado el modelo espacio-vectorial basado en *n*-gramas (secuencias de palabras de tamaño fijo). Una alternativa al modelo de n-gramas es emplear solo las Secuencias de palabras que son Frecuentes y además Maximales (SFM's), las cuales permiten disminuir la dimensionalidad del modelo, a la vez que brindan información más relevante, puesto que el tamaño de cada SFM no es determinado previamente como sucede con *n*-gramas. Este artículo presenta un estudio comparativo de estos modelos de textos para la generación automática de resúmenes extractivos con 567 documentos, empleando un algoritmo de aprendizaje no supervisado.
Palabras clave: Resúmenes de texto, secuencias frecuentes maximales, aprendizaje no supervisado, modelo espacio vectorial.

Abstract: One of the main issues on automatic text summarization consists in identify, independent from the language and domain, the most important information from the source document. For this problem many works have been applied the vector-space model, which is based on *n*-grams (word sequences of predetermined size). Other alternative to *n*-grams model is to use only Maximal Frequent Sequences (MFS's), which let decrease the dimensionality of the model at the time, while give more relevant information, because the size of each MFS is not previously determined as it does with *n*-grams. This work presents a comparative study of these text models for generating automatic extractive summaries with 567 documents, using an unsupervised learning algorithm.
Keywords: Text Summarization, Maximal Frequent Sequences, Unsupervised Learning, Vector-Space Model.

1 Introducción

En los últimos años se ha visto un crecimiento acelerado en la cantidad de información electrónica que existe a nivel mundial. También este incremento se observa en el número de sitios web. De acuerdo con la revista Netcraft (2009), en marzo del 2009 se obtuvo respuesta de 224,749,695 sitios web. En la gráfica de la

ISSN: 1135-5948

figura 1 se muestra el incremento de sitios web entre junio de 2000 y marzo de 2009.

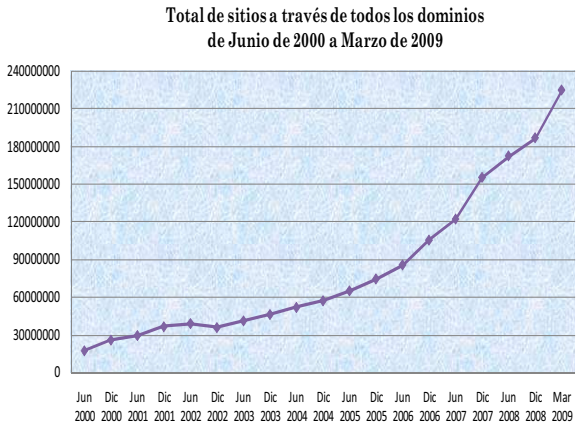

Figura 1: Incremento de sitios web entre junio de 2000 y marzo de 2009

En la actualidad, para realizar una consulta de páginas web, es común utilizar buscadores como *Google*, el cual encuentra una serie de enlaces, con dos o tres líneas del contenido original del documento o página web asociada. Por tanto, el usuario debe escoger, con base en esa poca información, si el documento es de utilidad. Para ello, es necesario descargar y examinar cada documento, hasta encontrar aquél que satisfaga su necesidad de información. Este proceso podría mejorarse si junto al enlace se muestra un resumen que describa de mejor manera el contenido del documento, sin necesidad de descargarlo completamente. Esto nos impulsa a desarrollar técnicas que permitan generar en forma automática resúmenes de documentos.

De acuerdo con Lloret *et. al.* (2008), un resumen se define como "un texto que se genera a partir de uno o más textos, que contiene la información más significativa y que no es más extenso que la mitad de los textos originales".

Las tareas de generación automática de resúmenes se han aplicado a un solo documento y a colecciones completas de documentos.

Existen diversos métodos para generar resúmenes automáticamente. En forma general, los métodos se pueden clasificar en *abstractivos* y *extractivos*. Esta clasificación depende del nivel de análisis lingüístico realizado sobre el documento fuente (Maña, 2003).

De acuerdo con Ledeneva (2008), *los métodos abstractivos* consisten en "comprender" el texto original, y reescribir su contenido en menos palabras. En la generación automática de resúmenes abstractivos, se utilizan métodos lingüísticos, de manera que permitan describir mejor al documento (Lin, 2002). La desventaja es que se debe llevar a cabo un análisis de mayor profundidad en comparación con los métodos extractivos. Además, se requiere un amplio conocimiento del dominio del texto, y es aplicable solo en ámbitos muy concretos y enteramente conocidos (Maña, 2003).

Por su parte, *los métodos extractivos* consisten en una selección de oraciones (frases, párrafos, etc.) del texto original. Por lo general, en ellos se decide, para cada oración, si ésta será incluida o no en el resumen (Ledeneva, 2008). La mayor ventaja de este enfoque es que resulta muy robusto y fácilmente aplicable a contextos de propósito general, por su independencia del dominio y género (Maña, 2003). Lo anterior hace muy interesante la idea de utilizar métodos extractivos en la generación automática de resúmenes.

Uno de los principales problemas en la generación automática de resúmenes consiste en identificar, bajo la independencia mencionada, la información más relevante del documento original. Para esto, existen diversos enfoques utilizados en la generación automática de resúmenes. Algunos de ellos utilizan técnicas tales como la selección de frases, basadas en palabras clave y ubicación del texto (Acero, 2001), frecuencia de las palabras e índices estadísticos (Cunha, 2007), grado de importancia de las oraciones (Lee, 2006), similitud de oraciones (Márquez, 2007) y análisis lingüísticos de las frases (Mateo, 2003). Sin embargo, en algunos de estos trabajos, el resumen está sujeto a una adaptación al perfil del usuario.

Otros enfoques utilizan métodos que son dependientes, ya sea del dominio de los documentos (Lee, 2006), o del lenguaje (Mateo, 2003), (Liu, 2006) y (Cunha, 2007).

En nuestro trabajo (García *et al.*, 2008), hemos propuesto un método para la generación automática de resúmenes, mediante la extracción de frases clave. A partir de ese trabajo, ahora se presenta la comparación de tres modelos de representación de texto, que también emplean un algoritmo de aprendizaje no supervisado para la generación del resumen. Nuestra hipótesis fue que un algoritmo de agrupamiento puede formar grupos de oraciones similares. Por lo cual, el resumen es conformado tomando de cada grupo la oración

más representativa. Una ventaja, es que el enfoque resultante es independiente del dominio y del lenguaje de los documentos.

El resto del documento está organizado de la siguiente forma. En la sección 2 se presenta un resumen del estado del arte sobre los métodos de generación automática de resúmenes. En la sección 3 se describe el esquema general, del enfoque propuesto. En la sección 4 se presenta la configuración aplicada en la etapa de experimentación. En la sección 5 se plantea la comparación de los tres modelos de representación de textos, utilizados para generar los resúmenes, y la conclusión que es alcanzada en este trabajo.

2 Trabajo relacionado

Eduard Hovy *et al.* (1999), presentó un sistema automatizado para la generación de resúmenes, llamado SUMMARIST. Este sistema está basado en la ecuación: *resumen = identificación del tema + interpretación + generación*. Cada elemento de la ecuación representa una etapa del sistema. En la primera etapa, los módulos de posición, tipos de palabras y frases significativas evalúan a cada oración. Después, un módulo de integración combina las puntuaciones para producir una clasificación global. En la segunda etapa se realiza la interpretación del tema. Lo anterior consiste en 'fusionar' dos o más tópicos extraídos dentro de uno o más conceptos unificadores. La etapa final del proceso es la generación del resumen. En esta fase, se consideran tres módulos para la generación de resúmenes: a) *Extracción*, en el cuál solo se reproducen las oraciones seleccionadas en la primer etapa; *b) Listas de tópicos,* donde una simple lista de los temas del resumen es suficiente; y *c) Concatenación de frases*, que forma unidades de frases y cláusulas dentro de simples oraciones.

Radev *et al.* (1998), desarrolló un sistema llamado SIMMONS, el cual utiliza la salida de sistemas desarrollados para la DARPA Message Understanding Conferences, para generar resúmenes de múltiples documentos sobre el mismo evento. La arquitectura global de SIMMONS apunta a la investigación sobre agentes de software, a fin de permitir las conexiones con diferentes tipos de fuentes de información. El sistema extrae información de las diferentes fuentes y luego la combina dentro de una representación conceptual del resumen. El generador del resumen combina información de múltiples artículos de entrada y organiza esa información usando un planificador de párrafo. La representación conceptual estructurada, es llevada al selector léxico, que también recibe información de otras fuentes, como posibles descripciones de personas u organizaciones para aumentar el resumen. El contenido completo se pasa entonces a través de un generador de oraciones. SIMMONS produce un resumen desde plantillas que contienen datos sobresalientes reportados en los artículos de entrada.

Mihalcea (2004), propone un método en el cual se construye un grafo para representar el texto. Los nodos del grafo son palabras interconectadas mediante vértices con relaciones significativas. En la extracción de oraciones, se califican oraciones enteras y se ordenan de mayor a menor calificación. Por cada oración en el texto, se agrega un vértice al grafo. Para establecer las conexiones entre oraciones, se define una relación de similitud. En dicha relación, una oración que señala a cierto concepto en el texto, da al lector una "recomendación" para referirse a otras oraciones que señalan a los mismos conceptos. Por tanto, se establece un vínculo entre oraciones que compartan un contenido común.

Lee *et al.* (2006), propone un método llamado "modelo de sumarización fractal", para la obtención de resúmenes en forma automática. El método consiste en dividir el documento y transformarlo en una estructura de árbol. Para cada nodo, se calcula el grado de importancia del bloque, sumando los grados de importancia de las oraciones que se encuentran bajo tal bloque. Posteriormente, se calcula el número de oraciones que son extraídas para el resumen. Este enfoque obtuvo resultados favorables, aunque solo se analizó para dos documentos sobre ataques terroristas.

En el trabajo de Cunha *et al.* (2007), se examinaron tres métodos para la generación automática de resúmenes. El primer método, llamado *CORTEX*, está basado en el modelo de espacio vectorial y consiste en representar el texto mediante el modelo de bolsa de palabras. El segundo método, denominado *ENERTEX*, es un enfoque de redes neuronales, cuya función es traducir el documento en un sistema de unidades que permitan calificar las oraciones del texto de acuerdo a su relevancia. El tercer método, llamado *DISICOSUM*, combina diversos aspectos lingüísticos. Las pruebas se realizaron sobre 10 artículos médicos en español, alcanzando los mejores resultados con *DISICOSUM*.

Mateo *et al.* (2003) presenta un generador de resúmenes de textos que consta de cinco módulos.

El primero de ellos analiza cada palabra, determinando si es sustantivo, verbo, artículo, etcétera. El segundo método asigna puntuaciones a las frases del texto según su importancia. El tercer módulo detecta si en las oraciones con altas puntuaciones existen frases que hagan referencia a otras oraciones que no estén consideradas para el resumen (anáforas). El cuarto módulo selecciona las oraciones candidatas a formar parte del resumen, con base en sus puntuaciones. El último módulo elimina las oraciones conectadas a otras no seleccionadas, y entrega el resumen. Con este enfoque se obtuvieron buenos resultados, sin embargo, sólo fue aplicado a textos en castellano. Además, depende del lenguaje, pues requiere bases de conocimiento léxico para la detección de anáforas.

Manuel Maña (2003) aplica un método de selección y extracción de frases para generar resúmenes en forma automática. La selección de frases se realiza bajo tres criterios: palabras clave, título y localización. Las palabras clave son aquéllas que aparecen con mayor frecuencia en el texto, por tanto son seleccionadas, al igual que las frases contenidas en el título, para formar parte del resumen. El tercer criterio consiste en seleccionar las primeras frases del texto, dado que la colección de documentos utilizada consiste en artículos periodísticos, que usualmente contienen la información más importante al inicio. Posteriormente, el resumen se adapta a los requerimientos del usuario. Por tanto, este enfoque depende de un modelo de usuario, construido previamente con base en las consultas realizadas por él mismo.

Liu *et al.* (2006) propone la generación automática de resúmenes mediante una estrategia que combina agrupamiento y extracción de oraciones. En el agrupamiento de oraciones, se plantean dos heurísticas para determinar el número de grupos automáticamente. La primera utiliza la longitud del resumen, definida por el usuario. La segunda heurística se basa en un método de *estabilidad*, para deducir el número óptimo de grupos. La selección de las oraciones que formarán el resumen se lleva a cabo mediante una búsqueda, cuyo objetivo es encontrar la oración que mejor contribuya a la interpretación del resumen. Este enfoque brindó buenos resultados, sin embargo, está diseñado solo para documentos en el idioma chino.

Márquez *et al.* (2007) compara dos métodos para la obtención de resúmenes automáticos. El primer método permite seleccionar las oraciones relevantes, mediante las técnicas de *Párrafo Virtual (PV)* y *Punto de Transición (PT)*. El Punto de Transición divide al documento en términos de alta y baja frecuencia. El Párrafo Virtual se construye seleccionando el 25% de los términos alrededor del PT. Posteriormente, se seleccionan del texto las cinco oraciones que presenten mayor similitud con el PV. En el segundo método, se aplica el algoritmo de agrupamiento ROCK, con el objetivo de construir grupos de términos con cierto grado de similitud. Los resultados mostraron que la obtención de resúmenes automáticos mediante el PV superó al método ROCK.

Ledeneva *et al.* (2008) presenta un enfoque independiente del dominio y del idioma de los documentos, para la generación automática de resúmenes. En él, se aplican cuatro pasos: selección de términos, pesado de términos, pesado de oraciones y selección de oraciones. En la selección de términos, se aplica el enfoque de *secuencias frecuentes maximales (SFM´s)*, que consiste en seleccionar secuencias frecuentes de palabras, que no estén contenidas en otra secuencia frecuente. También son seleccionados los bigramas repetitivos y las palabras. Para el pesado de términos, se utiliza la frecuencia del término dentro de una SFM. En el pesado de oraciones, únicamente se suma el peso de todos los términos contenidos en la oración. Finalmente, la selección de las oraciones que conformarán el resumen se realiza seleccionando las oraciones con mayor peso, así como las primeras oraciones que aparecen en el documento, hasta alcanzar la longitud del resumen deseada. Los resúmenes se realizaron sobre la colección DUC 2002 (DUC, 2002). Los mejores resúmenes alcanzaron un 47% de similitud con los realizados por un humano.

3 Resumen mediante agrupamiento

Normalmente, en la generación automática de resúmenes extractivos se llevan a cabo cuatro pasos: selección de términos, pesado de términos, pesado de oraciones y selección de oraciones (Kupiec, 1995), (Ledeneva, 2008). Sin embargo, la selección de oraciones se reduce simplemente a tomar las oraciones con mayor peso. Aún cuando esta técnica funciona bien para las primeras oraciones seleccionadas, es probable que otras oraciones similares a las ya elegidas, sean seleccionadas, lo cual produciría redundancia en el resumen. Para contrarrestar este problema, en nuestro trabajo (García *et al.*, 2008), se propuso sustituir las etapas de pesado y selección de oraciones por un algoritmo de aprendizaje no

supervisado. La idea es que un algoritmo de agrupamiento puede ayudar a detectar automáticamente los grupos de oraciones similares a partir de los cuales se puede seleccionar la oración más representativa como resumen del documento; reduciendo de esta forma la redundancia en el resumen. Además el método propuesto por García *et al.* (2008) es independiente del lenguaje y del dominio del texto. Por sus características, este último método es el que se va a utilizar al realizar la comparación entre los diferentes modelos de texto. En este apartado se describen las etapas generales que se llevan a cabo en este método.

3.1 Etapa de selección de términos

Un *n*-grama es una secuencia de *n* palabras consecutivas. Decimos que un *n*-grama ocurre en un texto si esas palabras aparecen en el mismo orden, inmediatamente una después de la otra (Schneider, 2002).

Llamamos *n*-grama frecuente a una secuencia de palabras que aparece al menos β veces en el documento. El índice β es el umbral de frecuencia, previamente definido. Lo anterior significa que, aquéllos términos o secuencias de palabras que se repitan al menos β veces en el texto, serán consideradas como secuencias frecuentes (SF's). Asimismo, las SF's que no son subsecuencia de otra SF son por lo tanto *Secuencias Frecuentes Maximales* (SFM's) (García *et al.*, 2004), (Ledeneva, 2008), (García *et al.*, 2008a).

En este trabajo se compara el modelo de texto basado en SFM's con los dos modelos comúnmente utilizados, 1-gramas también conocido como Bolsa De Palabras (BDP) y 2-gramas conocido como bigramas.

3.2 Etapa de pesado de términos

Los tres modelos mencionados anteriormente pueden ser ponderados de diferentes maneras:

3.2.1 Pesado Booleano (BOOL)

Es la forma más sencilla de pesar un término. Evalúa la presencia o ausencia de un término en el documento. Está definido por la fórmula (1):

$$p_i(t_j) = \begin{cases} 1, & \text{si el término } t_j \text{ aparece en el documento } i \\ 0, & \text{en otro caso} \end{cases} \quad (1)$$

3.2.2 Frecuencia del término (TF)

Este pesado fue propuesto por Lunh (1957). TF otorga mayor relevancia a los términos con mayor frecuencia, evaluando el número de veces que el término aparece en el documento. Lo anterior se expresa en la fórmula (2):

$$p_i(t_j) = f_{ij} \quad (2)$$

Donde f_{ij} es la frecuencia del término *j* en el documento *i*.

3.2.3 Frecuencia inversa del documento (IDF)

El pesado IDF fue propuesto por Salton (1988) con el objetivo de mejorar los sistemas de Recuperación de Información (RI). El problema del pesado TF es que, cuando un término aparece en casi todos los documentos de la colección, dicho término no es útil para discriminar los documentos relevantes. Como una alternativa, surge el pesado IDF, que está definido por la fórmula (3):

$$p_i(t_j) = \log\left(\frac{N}{n_j}\right) \quad (3)$$

Donde N es el número de documentos en la colección y n_j es el número de documentos donde el término *j* aparece.

3.2.4 TF-IDF

El problema del pesado IDF en RI es que no es posible distinguir entre dos documentos con el mismo vocabulario, aún cuando el término es muy frecuente en un documento. El pesado TF-IDF (Brunzel, 2007), concede mayor relevancia a los términos que son menos frecuentes en la colección, pero más frecuentes en el documento. El pesado TF-IDF está dado por la fórmula (4):

$$p_i(t_j) = f_{ij} \times \log\left(\frac{N}{n_j}\right) \quad (4)$$

Cabe señalar, que en este trabajo se utilizan los pesados de términos descritos anteriormente para generar el resumen de un solo documento. Por lo tanto, para la aplicación de estos pesados de términos se considera al documento como una colección de oraciones en lugar de una colección de documentos.

3.3 Etapa de selección de oraciones utilizando una técnica de aprendizaje no supervisado

La principal característica de las técnicas de aprendizaje no supervisado es que no se necesita un conocimiento previo de los datos para analizarlos y procesarlos. En nuestro caso, se aplica el algoritmo K-medias con el objetivo de descubrir grupos de oraciones con significado semejante, para después conformar el resumen seleccionando de cada grupo la oración más representativa. El algoritmo K-medias asume que el número de grupos se define previamente. Aunque esta característica suele verse como un inconveniente, en este caso es una ventaja, ya que permite especificar el número de grupos que serán creados, lo que a su vez, permite estimar el número de palabras que contendrá el resumen final. Por ejemplo, si el promedio de palabras por oración es de 20, y el usuario requiere un resumen de 100 palabras, K-medias deberá crear 5 grupos. Por supuesto, esto es solo una estimación del número de palabras en el resumen final. En K-medias, cada oración está representada en un modelo de espacio vectorial. Por lo tanto, cada documento es representado como un vector de características, donde las características corresponden a los diferentes términos en el documento. En este caso, los términos son *n*-gramas y SFM's.

En principio, el algoritmo K-medias necesita semillas como centroides iniciales para cada grupo. Por lo tanto, el éxito de K-medias depende de elegir buenas semillas iniciales. Normalmente, las semillas iniciales son seleccionadas de manera aleatoria. En nuestra comparación, las primeras oraciones son consideradas como semillas iniciales, dado que las oraciones *Baseline* (heurística que ha demostrado que las primeras oraciones de una noticia pueden considerarse como un buen resumen) son buenas candidatas para conformar el resumen. Para medir la similitud entre dos oraciones, se utiliza la distancia Euclidiana.

4 Procedimiento y Resultados experimentales

En esta sección se presentan los resultados obtenidos en la etapa de experimentación.

Algoritmo. Cada experimento se realizó mediante la misma secuencia de pasos:

- *Preprocesamiento*. Se eliminaron las palabras vacías y se aplicó el algoritmo de *Stemming* de Porter (Sparck, 1997).
- *Selección de términos.* Se eligió uno de los tres modelos para representar el texto: bolsa de palabras, *n*-gramas y SFM's.
- *Pesado de términos.* Se eligió uno de los cuatro métodos para calcular la importancia de cada término: BOOL, TF, IDF y TF-IDF.
- *Agrupamiento de oraciones.* Se eligió entre dos opciones la forma en que se calcularían los centroides iniciales para el algoritmo K-medias: aleatoriamente o con las oraciones Baseline.
- *Selección de oraciones.* Al terminar el agrupamiento, se seleccionó la oración más cercana al centroide de cada grupo (oración más representativa del grupo), para conformar el resumen.
- Los parámetros específicos para cada paso varían entre los experimentos.

Colección de documentos. Se utilizó la colección de documentos estándar DUC 2002 (DUC, 2002). Específicamente, se utilizó un conjunto de 567 noticias sobre diversos temas y de diferentes longitudes. Para cada documento en la colección DUC, se dispuso de un conjunto de resúmenes generados por humanos, proporcionados por dos expertos diferentes. Mientras que a cada experto se le solicitaron resúmenes con diferentes longitudes, en nuestro caso sólo se utilizaron variantes de 100 palabras.

Método de evaluación. La evaluación de los experimentos se llevó a cabo mediante la herramienta ROUGE, un sistema automático para la evaluación de resúmenes, propuesto por Lin (2004). ROUGE tiene la capacidad de medir la correlación entre los resúmenes creados por humanos y los resúmenes generados automáticamente (Lin, 2003). Los valores obtenidos mediante la evaluación con ROUGE-n, con n=1, corresponden a las métricas de Recuerdo, Precisión y F-measure. Consideramos principalmente el valor de F-measure, por representar un balance (no un promedio) entre los resultados de Recuerdo y Precisión.

A continuación se muestran los resultados obtenidos con ROUGE para los tres modelos de representación de texto, cuya comparación se ha propuesto en este trabajo. Las tablas 1, 2 y 3 muestran los valores de Recuerdo, Precisión y F-measure, respectivamente. En cada tabla se

presentan los valores obtenidos por cada modelo, aplicando los cuatro diferentes pesados de términos, mencionados en la sección 3, con preprocesamiento (CP) y sin preprocesamiento (SP).

Como se puede observar en las tablas 1, 2 y 3, los valores más altos de Recuerdo, Precisión y F-measure fueron obtenidos con el modelo de SFM's y pesado booleano. Sin embargo, al omitir la etapa de preprocesamiento, los resultados mejoraron ligeramente, obteniéndose los índices más altos con el modelo de bolsa de palabras y frecuencia del término (*tf*).

Modelo de texto		Pesado de términos			
		BOOL	TF	IDF	TF-IDF
CP	BDP	0.4456	0.4420	0.4413	0.4419
	2-gramas	0.4400	0.4399	0.4409	0.4412
	SFM's	**0.4469**	0.4412	0.4412	0.4422
SP	BDP	0.4434	**0.4488**	0.4443	0.4447
	2-gramas	0.4420	0.4386	0.4440	0.4408
	SFM's	0.4429	0.4424	0.4387	0.4406

Tabla 1: Valores de recuerdo de los resúmenes obtenidos con los tres modelos y los cuatro pesados, con y sin preprocesamiento

Modelo de texto		Pesado de términos			
		BOOL	TF	IDF	TF-IDF
CP	BDP	0.4420	0.4389	0.4379	0.4382
	2-gramas	0.4364	0.4368	0.4378	0.4380
	SFM's	**0.4437**	0.4385	0.4380	0.4391
SP	BDP	0.4402	**0.4459**	0.4406	0.4418
	2-gramas	0.4394	0.4359	0.4412	0.4374
	SFM's	0.4398	0.4396	0.4354	0.4376

Tabla 2: Valores de precisión de los resúmenes obtenidos con los tres modelos y los cuatro pesados, con y sin preprocesamiento.

Modelo de texto		Pesado de términos			
		BOOL	TF	IDF	TF-IDF
CP	BDP	0.4437	0.4403	0.4394	0.4399
	2-gramas	0.4381	0.4382	0.4392	0.4395
	SFM's	**0.4452**	0.4397	0.4395	0.4405
SP	BDP	0.4417	**0.4472**	0.4423	0.4432
	2-gramas	0.4406	0.4372	0.4425	0.4390
	SFM's	0.4412	0.4409	0.4369	0.4390

Tabla 3: Valores de F-measure de los resúmenes obtenidos con los tres modelos y los cuatro pesados, con y sin preprocesamiento

Con el objetivo de incrementar la calidad de los resúmenes, se realizaron más experimentos, con una variante en la etapa de agrupamiento de oraciones: se tomaron las primeras oraciones de los documentos (Baseline) como semillas iniciales para el algoritmo K-medias.

Las tablas 4, 5 y 6 muestran los valores de Recuerdo, Precisión y F-measure respectivamente, obtenidos en los experimentos con Baseline.

Modelo de texto		Pesado de términos			
		BOOL	TF	IDF	TF-IDF
CP	BDP	0.4751	0.4768	0.4763	0.4754
	2-gramas	0.4770	0.4769	**0.4777**	**0.4777**
	SFM's	0.4725	0.4709	0.4728	0.4722
SP	BDP	0.4754	0.4745	0.4745	0.4762
	2-gramas	**0.4782**	0.4778	0.4779	0.4778
	SFM's	0.4711	0.4705	0.4730	0.4759

Tabla 4: Valores de recuerdo de los resúmenes obtenidos con los tres modelos, los cuatro pesados, con Baseline, con y sin preprocesamiento

Modelo de texto		Pesado de términos			
		BOOL	TF	IDF	TF-IDF
CP	BDP	0.4703	0.4721	0.4716	0.4707
	2-gramas	0.4721	0.4720	**0.4728**	**0.4728**
	SFM's	0.4682	0.4665	0.4684	0.4677
SP	BDP	0.4706	0.4700	0.4696	0.4714
	2-gramas	**0.4734**	0.4729	0.4730	0.4729
	SFM's	0.4666	0.4664	0.4687	0.4714

Tabla 5: Valores de precisión de los resúmenes obtenidos con los tres modelos, los cuatro pesados, con Baseline, con y sin preprocesamiento

Modelo de texto		Pesado de términos			
		BOOL	TF	IDF	TF-IDF
CP	BDP	0.4726	0.4743	0.4738	0.4729
	2-gramas	0.4744	0.4743	**0.4751**	**0.4751**
	SFM's	0.4702	0.4686	0.4705	0.4698
SP	BDP	0.4729	0.4721	0.4719	0.4737
	2-gramas	**0.4757**	0.4752	0.4753	0.4752
	SFM's	0.4687	0.4683	0.4707	0.4736

Tabla 6: Valores de F-measure de los resúmenes obtenidos con los tres modelos, los cuatro pesados, con Baseline, con y sin preprocesamiento

De acuerdo con los datos mostrados en las tablas 4, 5 y 6, en todos los experimentos donde se aplicó Baseline, los valores más altos de Precisión, Recuerdo y F-measure se obtuvieron con el modelo de 2-gramas y los pesados *idf* y *tf-idf*. Omitiendo la etapa de pre-procesamiento, el pesado de términos booleano alcanzó los índices más altos de dichas métricas, superando en calidad a los resúmenes obtenidos con pre-procesamiento.

5 Conclusiones

En este trabajo se realizó la comparación de tres modelos de representación de textos para la generación automática de resúmenes, aplicando un algoritmo de aprendizaje no supervisado. En particular, se compararon los modelos de bolsa de palabras, bigramas y SFM's; para lo cual se utilizó el algoritmo K-medias.

Los resultados de los experimentos realizados con el enfoque propuesto fueron interesantes:

- Originalmente, los mejores resultados de Recuerdo, Precisión y F-measure se lograron con el modelo de SFM's y pesado booleano.
- Al omitir la etapa de preprocesamiento, el modelo que aportó mejores resultados fue bolsa de palabras con frecuencia del término (pesado *tf*). Los resúmenes generados con este modelo superaron a los realizados con SFM's.
- Aplicando Baseline al algoritmo K-medias, los resultados de Recuerdo, Precisión y F-measure se elevaron notablemente. En este enfoque, los valores más altos fueron alcanzados con el modelo de 2-gramas y los pesados *idf* y *tf-idf*, superando también los valores obtenidos con los experimentos donde las semillas iniciales de K-medias se calcularon en forma aleatoria.
- Finalmente, se eliminó la etapa de preprocesamiento y se aplicó Baseline con K-medias. La calidad de los resúmenes obtenidos con este esquema fue mayor que la de todos los resúmenes generados en los experimentos anteriores. El modelo que aportó mayores índices de Recuerdo, Precisión y F-Measure fue 2-gramas con pesado booleano.

De acuerdo con los resultados experimentales, se ha observado que el modelo de bigramas con pesado booleano obtuvo mejores resultados que los obtenidos con los modelos de bolsa de palabras y SFM's. Además, los experimentos mostraron que el pre-procesamiento de los documentos no es tan relevante en la generación automática de resúmenes extractivos.

Bibliografía

Acero, I., M. Alcojor, A. Díaz, J. M. Gómez y M. Maña. 2001. Generación automática de resúmenes personalizados. Procesamiento del Lenguaje Natural, no. 27, pp. 281-290.

Brunzel, M. y M. Spiliopoulou. 2007. Domain Relevance on Term Weighting. Lecture Notes in Computer Science 4592, Springer, 2007, pp. 427-432.

Cunha, I., S. Fernández, P. Velázquez, J. Vivaldi, E. SanJuan y J. Torres. 2007. A new hybrid summarizer based on vector space model, statistical physics and linguistics. LNCS 4827, pp. 872-882.

DUC, 2002. Document Understanding Conference 2002.

García, R. A., J. Martínez, y J. Carrasco. 2004. A Fast Algorithm to Find All the Maximal Frequent Sequences in a Text. 9th Iberoamerican Congress on Pattern Recognition (CIARP), LNCS Vol. 3287, pp. 478-486, Springer-Verlag, 2004.

García, R. A., R. Montiel, Y. Ledeneva, E. Rendón, A. Gelbukh y R. Cruz. 2008. Text Summarization by Sentence Extraction using Unsupervised Learning. Lecture Notes in Computer Science 5317, Springer, pp. 133-143.

García, R. A., Y. Ledeneva, A. Gelbukh, y C. Gutiérrez. 2008. An Assessment of Word Sequence Models for Extractive Text Summarization. Research in Computing Science (a), N 38, pp. 253-262.

Hovy, E., y C. Lin. 1999. Automated Text Summarization in SUMMARIST. In 'Advances in Automatic Text Summarization', I. Mani and M. Maybury (editors).

Kupiec, J., J. Pedersen, y F. Chen. 1995. A Trainable Document Summarizer. In Proceedings of the 18th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, pp. 68-73.

Ledeneva, Y., A. Gelbukh, y R. García. 2008. Terms derived from frequent sequences for extractive text summarization. CICLing, LNCS 4919, pp. 593-604.

Lee, F., C. Yang, y X. Shi. 2006. Multi-document summarization for terrorism information extraction. LNCS 3975, pp. 602-608.

Lin, C., E. Hovy. 2002. Manual and Automatic evaluation of summaries. Proceedings of the Workshop on Automatic Summarization (including DUC 2002), Philadelphia, July 2, Asociation for Computational Linguistics, pp. 45-51.

Lin, C., y E. Hovy. 2003. Automatic evaluation of summaries using N-gram co-occurrence statistics. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Vol. 1, pp. 71-78.

Lin, C. 2004. ROUGE: A package for automatic evaluation of summaries. Proceedings of the Association for Computational Linguistics Workshop, pp. 74-81. España.

Liu, D., Y. He, D. Ji, H. Yang y Z. Wu. 2006. Chinese multi-document summarization using adaptive clustering and global search strategy. LNCS 4099, pp. 1135-1139.

Lloret, E., O. Ferrández, R. Muñoz y M. Palomar. 2008. Integración del Reconocimiento de la implicación textual en tareas automáticas de resúmenes de textos. Procesamiento del Lenguaje Natural, no. 41, pp. 183-190.

Lunh, H. P. 1975. A statistical approach to mechanical encoding and searching of literary information. IBM Journal of Research and Development, pp. 309-317.

Maña, M. 2003. Generación automática de resúmenes de texto para el acceso a la información. Universidad de Vigo, Departamento de Informática, España, Septiembre.

Márquez, J., P. Rendón, R. Rodríguez, D. Vilariño y B. Beltrán. 2007. Comparación de dos métodos para la obtención de resúmenes automáticos. IEEE Congreso Internacional en Innovación y Desarrollo Tecnológico, México.

Mateo, P., J. González, J. Villena y J. Martínez. 2003. Un sistema para resumen automático de textos en castellano. Procesamiento del Lenguaje Natural, no. 31, pp. 29-36.

Mihalcea, R. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. Proceedings of the ACL on Interactive poster and demonstration sessions, artículo no. 20.

Netcraft. 2009. March 2009 Web Server Survey. Inglaterra.

Radev, D. y K. McKeown. 1998. Generating Natural Language Summaries from Multiple on-line sources. Computational Linguistics, Special issue on natural language generation, pp. 470-500.

Salton, G., C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. Information Processing and Management: an International Journal. Vol. 24, pp. 513-523.

Schneider, R. 2002. n-grams of sedes: A hybrid System for Corpus-based Text Summarization. In proceedings of LREC, Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, España, pp. 29-31.

Sparck, K., P. Willett. 1997. Readings in information retrieval. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

Los autores agradecemos al gobierno mexicano (PROMEP, CONACyT, SNI, SIP) por su apoyo y financiamiento. También agradecemos a la Universidad Autónoma del Estado de México y al Instituto Tecnológico de Toluca.