UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

Sistema resumidor-traductor automático

T E S I S

QUE PARA OBTENER EL TÍTULO DE

INGENIERO EN COMPUTACIÓN

P R E S E N T A

JOSUE ANTONIO CAREAGA MOYA

No. De Cuenta: 30300974 -0

CON LA DIRECCIÓN DEL

DR. ALFONSO MEDINA URREA

MÉXICO D.F. 2012

Esta tesis se enmarca en el proyecto **CONACyT**
Extracción de conocimiento lexicográfico
a partir de textos en internet
(Clave de registro: 105711).

Responsable: Dr. Alfonso Medina Urrea

Investigador del Instituto de Ingeniería de la UNAM.

Contenido

Sistema resumidor-traductor automático

I. INTRODUCCIÓN

1.1 Antecedentes

Este trabajo se llevó a caboen el marco del proyecto "Extracción de conocimiento lexicográfico a partir de textos de Internet" realizado en el Grupo de Ingeniería Lingüística (GIL) que forma parte del Instituto de Ingeniería (II) de la Universidad Nacional Autónoma de México (UNAM) con el patrocinio de una beca del Consejo Nacional de Ciencia y Tecnología (CONACyT, proyecto 105711, Convocatoria Ciencia Básica 2008, bajo la responsabilidad del Dr. Alfonso Medina Urrea).

1.2 Planteamiento del problema

Hoy en día, la búsqueda de información disponible en la web, requiere que las consultas sean específicas para obtener resultados que concuerden con los intereses del usuario. Algunos de los factores que influyen en estas búsquedas son: la gran diversidad de fuentes de información, las diversas lenguas en las que se encuentra disponible dicha información y la extensión de los documentos que la contienen. Por tales motivos, la influencia de los motores de búsqueda para predecir y hacer más sencilla la búsqueda al usuario, los diccionarios, los traductores automáticos, los resumidores, entre otras herramientas; se han convertido en aplicaciones altamente demandadas.

Dentro de este contexto, el objetivo de este trabajo es exponer la propuesta de un sistema que combine la función de dos de las aplicaciones previamente mencionadas: la traducción y el resumen automáticos. Parasimplificar al usuario la tarea de leer todo un documento que se desee resumir y posteriormente traducir, o viceversa, y así conjuntar dicha tarea en un solo proceso ejecutado por un sistema automático. Para ello, es necesario

determinar la manera en la cual ambas tareas se complementarán, es decir, de qué forma las dos serán vinculadas para que el resultado sea el mejor.

1.3 Justificación

Las necesidades de todo usuario respecto a las tecnologías de la información evolucionan rápidamente. A menudo combinar procesos resulta en su simplificación. Vale la pena investigar cómo desarrollar un sistema que combine estos dos procesos y que ataque el problema al que se enfrentan las personas que no dominan el idioma Inglés y que se encuentran con una vasta cantidad de información disponible en esa lengua, la cual, no es su lengua materna.

1.4 Objetivos

El objetivo general de este trabajo es generar una herramienta que sea considerada como una propuesta útil para la revisión de documentos en idioma inglés y que conjunte ambas tareas: no sólo resumir sino también traducir textos de manera fiable, es decir, poder predecir que la parte de la traducción será la mejor de acuerdo con el contexto del documento dados ciertos marcadores de traducción y por supuesto la calidad del resumen.

Los objetivos específicos son los siguientes:

1. Explicar la manera en la que un sistema de resumen automático funciona, para qué es útil y qué trabajos se han hecho al respecto. Mostrar algunas técnicas existentes para resumir un texto de manera automática.
2. Mostrar un panorama acerca de la traducción automática.
3. Desarrollar un resumidor para el español.
4. Implementar un traductor inglés-español.
5. Integrar el resumidor y el traductor.

Se espera obtener un resumen en español de calidad, proveniente de un documento en inglés. Que al momento de evaluarlo, arroje valores en las medidas (divergencias de Jensen-Shannon y Kullback-Liebler) que son consideradas por las herramientas de evaluación como buenas y que el contenido del resumen sea coherente y conciso con respecto a la información contenida en el documento original.

1.5 Estructura de la tesis

La presente tesis abarca diversos temas que se tratan tanto en el marco teórico como en el desarrollo de la investigación. En este apartado, se muestra brevemente la estructura de esta tesis con una breve descripción de cada sección.

En la *Introducción* se muestra una breve descripción del trabajo: el marco dentro del cual se llevó a cabo, qué antecedentes son importantes, el planteamiento del problema, la justificación del trabajo y los objetivos a alcanzar.

En el capítulo siguiente se presenta el *Marco teórico;*que proporciona un panorama de los conceptos y áreas bajo las cuales fue desarrollado el trabajo; como el procesamiento del lenguaje natural (PLN), la recuperación de información (RI), etc. Aborda el principio de funcionamiento de los sistemas de resumen y traducción automáticos, a través de diagramas estructurales y en la descripción de algunas de las técnicas existentes para realizar ambas tareas, así mismo el estado del arte de los temas relacionados con la investigación.

En el capítulo *Desarrollo del sistema resumidor-traductor* se describe el desarrollo de la tesis. Se enfoca en el sistema propuesto y su desarrollo. Se explicacómo fue desarrollada la herramienta, qué tecnologías se utilizaron (lenguajes de programación, frameworks, herramientas extras, librerías, etc.);cuál es la arquitectura del sistema, la descripción detallada de la misma; y cuál es el tipo de entrada al sistema. Se describe el proceso de recolección de datos de entrada al sistema.

El capítulo *Metodología*describe de manera detallada el método que siguió la investigación propuesta. Se muestra un diagrama estructural de la metodología seguida; así

como la descripción de los conceptos que son empleados a través del contenido del presente trabajo.

En el capítulo *Resultados* se aborda la problemática de cómo evaluar un sistema de resumen automático, como están formados los resultados del sistema, cuál es el análisis y la interpretación de los datos obtenidos y la descripción de los datos estadísticos, comparaciones o marcadores empleados.También se incluye una descripción de los resultados esperados.

Finalmente en el capítulo *Conclusiones*se hace una reflexión e interpretación de los resultados obtenidos, concluyendo cual fue el desempeño del sistema y de la investigación en general, y cuáles fueron los logros alcanzados. También se plantean las posibilidades de mejora del sistema y trabajo futuro.

II. MARCO TEÓRICO

2.1 Procesamiento del lenguaje natural (PLN)

El procesamiento del lenguaje natural, o PLN, es una disciplina que mezcla dos perfiles profesionales para llevar a cabo diversas tareas con datos de texto o voz: por un lado, la inteligencia artificial como rama de la ingeniería en computación y,por el otro, la lingüística como rama del estudio de la lengua.

Como su mismo nombre lo indica, combina técnicas y tareas que son realizadas por ambos perfiles para procesar de manera natural el lenguaje como es; y de este modo, hacer posible que el lenguaje hablado y escrito por los seres humanos sea manejable por las máquinas.

La tarea de crear un puente entre la manera en la que se comunica el ser humano mediante un lenguaje natural (español, tarahumara, inglés, griego, maya, rumano, francés, holandés,sueco, mandarín, etc.) y la forma en la que la computadora hace una representación de dicho lenguaje no es una tarea trivial. De hecho,esta tarea requiere que las personas que se vean involucradas estén familiarizadas con conceptos de ambos perfiles

(lingüístico y computacional) para llevar a cabo con una computadora actividades que comúnmente sería indispensable que realizara un humano.

Ahora bien, el objetivo de realizar dicha comunicación es para poder solicitar a una computadora que lleve a cabo ciertas tareas que los seres humanos realizamos con frecuencia y facilidad; por ejemplo,obtener respuestas desde el contenido de un archivo, generalizar la idea principal de un archivo, resumirlo, obtener palabras clave, encontrar parecido entre dos o más textos, entre otras numerosas tareas que se pueden realizar con la información.

En sí, se desea procesar el texto contenido en un archivo por su sentido y no sólo por su forma, que no necesariamente refleja el significado de las palabras que contiene.Típicamente, a los textos se les puede agregar información lingüística que permite que las computadoras reconozcan información lingüística; por ejemplo, reconocer qué tipo de palabra es (verbo, sujeto, artículo, etc.) o cuándo una oración ha sido formada, en lugar de simplemente reconocer cadenas de caracteres o *tokens*.

Jurafsky y Manning (2009) dicenacerca del PLN lo siguiente:

> Natural language processing is the technology for dealing with our most ubiquitous product: human language, as it appears in emails, web pages, tweets, product descriptions, newspaper stories, social media, and scientific articles, in thousands of languages and varieties. In the past decade, successful natural language processing applications have become part of our everyday experience, from spelling and grammar correction in word processors to machine translation on the web, from email spam detection to automatic question answering, from detecting people's opinions about products or services to extracting appointments from your email.

Como es mencionado por estos autores, el procesamiento del lenguaje natural tiene grandes alcancesdentro de todo contexto en el que el lenguaje se encuentre relacionado con la computadora. Se implica de este modo que donde se encuentre el lenguaje humano, se encontrará un reto para lograr procesar dicha información de acuerdo a las necesidades o curiosidades de aquellos que procesamos el lenguaje natural.

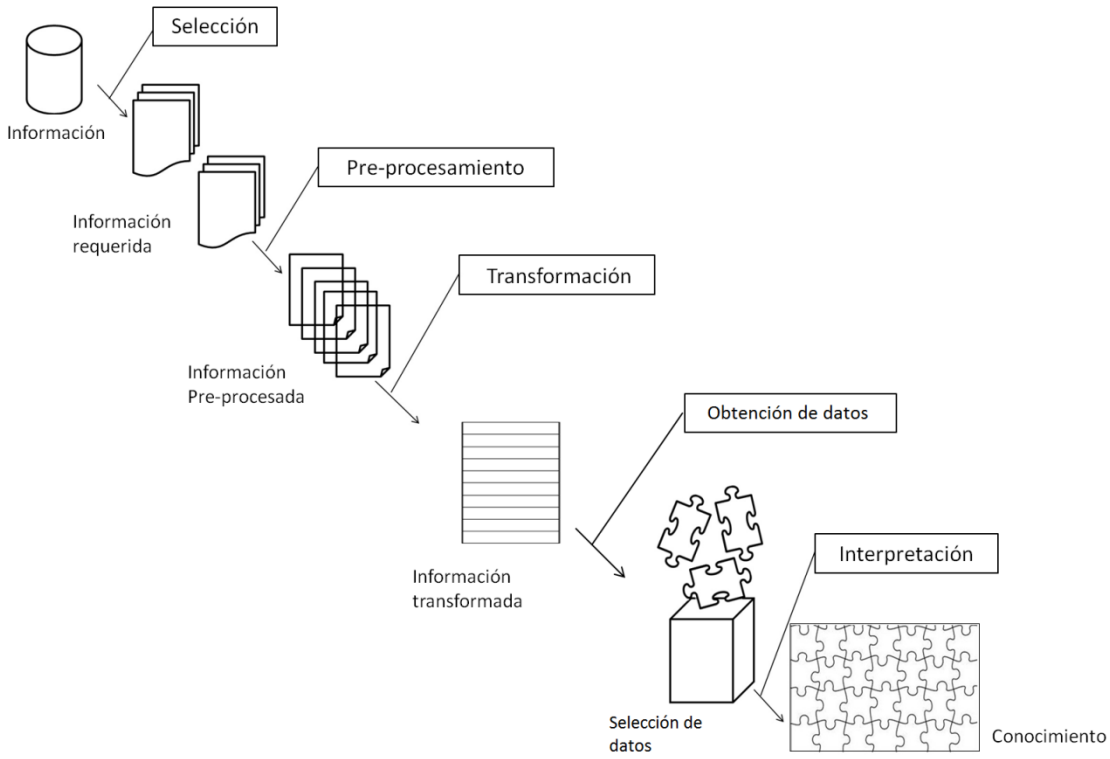
Dentro de la disciplina del PLN, se encuentran distintas líneas de investigación, las cuales se pueden clasificar de la siguiente manera: minería de textos, extracción o recuperación de información, sistemas pregunta-respuesta (*Question-Answering*), generación de taxonomías, entre muchas otras. Este trabajose enfoca en la extracción o recuperación de información; ya que cubre tanto la traducción, como el resumen automáticos.

2.2 Recuperación de información (RI)

Este subtema o línea de investigación del PLN guarda una relación con lo que se expresó anteriormente en el planteamiento del problema de esta tesis; donde se habla del crecimiento de la información y de las consultas que los usuarios realizan a través de motores de búsqueda disponibles en internet, así como de la misma información disponible en la red.

La recuperación de información implica tareas más complejas que el utilizar un motor de búsqueda. Estas tareas de análisis, filtrado, normalizado y extracción de datos se realizan de principio a fin.Méndez y Medina (2005) enuncian que son muchas las metodologías desarrolladas para obtener información a través de los distintos enfoques que existen y que a su vez, esto refleja la complejidad del realizar una extracción de información de manera exitosa.

Figura (1). Procesos implícitos en la recuperación de información (RI).



Actualmente los usuarios prefieren obtener información a través de un sistema de RI que hacerlo mediante otra personafamiliarizada con el tema de interés (como anteriormente se hacía); debido a que los resultados entregados por un sistema de RI han sido llevados a un nuevo nivel de calidad,a través dela optimización de la efectividad con la que realizan sus búsquedas.

Manning(2009) dice que la RI no comenzó a partir de internet, sino con la necesidad de resolver los distintos retos de brindar al usuario acceso a la información. De este modo, dio principio a enfoques de búsqueda endiversas formas de contenido.El campo comenzó con publicaciones científicas y registros de bibliotecas. Pronto se esparció a otras formas de contenido, particularmente aquellas de profesionales de la información; como periodistas, abogados y médicos. Mucha de la investigación científica en el campo de la RI se ha realizado en estas disciplinas y mucha de la continua práctica de la RI se enfrenta con el

reto de brindar acceso a la información no estructurada en varias empresas y dominios gubernamentales.

A pesar de ello, no se puede dejar de considerar que en años recientes, internet ha sido el manejador principal de información, desencadenando publicaciones de la escala de los diez-millones de creadores de contenido. Cerca de los años 90, mucha gente se dio cuenta que el seguir tratando de indexar toda la internet rápidamente se volvería imposible debido al crecimiento exponencial de la misma.

De esta manera, las instituciones que desean explotar la riqueza de lainformación que se encuentra en internet han apostado en mejorar los motores de búsqueda; que ya son capaces de proporcionar resultados de gran calidad en cuestión de milisegundos de respuesta a cientos de millones de búsquedas al día sobre miles de millones de sitios de internet.

La RI contempla diversas técnicas para llevar a cabo el rastreo, recuperación y en cierto modo estructuración de la información que se desea conseguir. Dichas técnicas, se pueden clasificar en:

El modelo booleano. En este modelo, las peticiones o búsquedas son términos indexados unidos por operadores lógicos (AND, OR y NOT) que posteriormente son llevados a su forma normal disyuntiva donde cada parte en la que es descompuesta se refiere a un vector binario con cierto peso. Comoel peso otorgado a cada palabra de los vectores es binario, se obtiene mucha información no relevante debido a que las palabras clave "están o no están", lo cual implica que los documentos son relevantes o no.De esta manera, en este modelo es difícil medir el grado de relevancia de las palabras clave.

El modelo vectorial. A diferencia del modelo anterior, en una búsqueda, la relevancia de los términos indexados se cuantifica con una escala ponderada de pesos. Estos pesos son de dos tipos: asociados a la búsqueda y a los documentos que contienen términos indexados en el sistema, o asociados a varios documentos que contienen términos

indexados en el sistema. En este segundo tipo de peso,la asignación se realiza sin considerar la búsqueda.El concepto principal de este modelo es la medida tf-idf (*total frequency-inversedocumentfrequency*) definido bajo el principio que el peso de un término indexado es proporcional a su frecuencia en el documento (tf) e inversamente proporcional a su frecuencia entre todos los documentos del sistema (idf).

El modelo probabilístico. Este modeloes también conocido como de independencia binaria, debido a que los pesos de los términos indexados para los documentos y las búsquedas son 1 ó 0. Calcula la probabilidad de que un término indexado en una búsqueda se encuentre en un conjunto de documentos recuperados. Entonces utiliza procesos recursivos en los documentos recuperados para mejorar ese cálculo de probabilidad. De este modo, los documentos son ordenados o ponderados en orden decreciente de su probabilidad de ser relevantes para la búsqueda. Este modelo no emplea la frecuencia de aparición de los términos indexados en los documentos por la asignación binaria de los pesos a las mismas.

La aplicación de estas técnicas de RI muchas veces depende del objetivo de la RI, así como de la conformación del conjunto de documentos para recuperar información (o en su caso de su disponibilidad en internet). De acuerdo con el objetivo de la RI, se puede elegir qué método o técnica es la que conviene utilizar y si no es claro, se pueden realizar pruebas con más de una técnica.De acuerdo con Medina (2011), se puede ver una tendencia global dentro de las aplicaciones de la RI hacia resumen automático y al reconocimiento de entidades (name entity recognition).

2.3 Resumen automático

Para comprender el concepto de resumen automático, es vital proporcionar una definición de resumen en sí. Un resumen es una versión condensada de un documento fuente que tiene un propósito específico: dar al lector una idea concisa del contenido de la fuente (Saggion y Lapalme 2002). Ahora bien, la tarea del resumen automático es una disciplina del procesamiento del lenguaje natural que implica generar, mediante un programa, un

documento que contenga la información más relevante en el contexto que se encuentra en el documento fuente, sin intervención humana (salvo, por supuesto, la ejecución del programa). Dentro de este escenario no se puede prescindir de la participación del ser humano a la hora de generar las listas que servirán para filtrar las palabras contenidas en los documentos para poder diferenciar las "palabras de contenido" de los artículos, conjunciones, etc.

Los resumidores automáticos se pueden clasificar de acuerdo con su tipo (sección 2.5), la técnica implementada (sección 2.4), y el tipo de documento de entrada, entre otros. Enseguida se presenta una breve reseña de las técnicas implementadas para realizar resumen automático, así como los tipos, estructura y algunos desarrollos.

2.4 Técnicas de resumen automático

Para hablar sobre las técnicas existentes para realizar resúmenes automáticos, considerouna clasificación hecha por Mani y Maybury (1999) que agrupa las técnicas en: enfoques clásicos, enfoques basados en corpus, enfoques ricos en conocimiento y el análisis de estructuras discursivas. Cada una de ellas se describirá brevemente por separado.

Losenfoques clásicos comprenden los primeros acercamientos con el resumen automático, como los que generan resúmenes automáticos de ámbitos específicos, los que proponen metodologías de extracción de enunciados, así como los que tomaron la tarea de resumir textos como un área de investigación en centros especializados. Trabajos que se pueden incluir en este enfoque son los de Luhn (1959), Edmundson(1969) y Pollock (1975).

Los enfoques basados en corpus utilizan, como su nombre lo indica, un corpus de entrada para probar, entrenar y, algunos de ellos, evaluar sus propios sistemas. Esto permite que adquieran conocimiento y lleven a cabo metodologías estadísticas. Finalmente, les permite basar sus sistemas en comparaciones con resultados obtenidos del procesamiento de distintos documentos, así como evaluarlos. Se lleva a cabo un aprendizaje no

supervisado.Autores que han trabajado bajo ese enfoque son: Xie (2010) y Suneetha (2011).

Los enfoques basados en el análisis de estructuras discursivas son aquellos que analizan la estructura discursiva de los textos que se desean resumir empleando cadenas léxicas, árboles discursivos y clasificación argumentativa de las oraciones extraídas para obtener indicadores de importancia en el texto, y así asegurar que la información relevante sea contemplada para formar el resumen. La codificación de los textos es a priori.Se tienen autores como Marcu(2000) y daCunha (2006) que han trabajado bajo este enfoque.

En general, el agrupamiento de las técnicas existentes para realizar resumen automático, considera las áreas de la computación y la lingüística que se han visto involucrados en la tarea de resumen automático. A pesar de que no han sido descritas a detalle, se tiene un panorama general de cómo se asocian ambas disciplinas para que, trabajando en conjunto el resultado del análisis textual tenga un sentido para el usuario final: un ser humano. Y de este modo, cumpla su propósito de resumir textos.

2.5 Tipos de resumidores automáticos

Los tipos de resumidores automáticos se pueden describir de manera general como sigue:

Resumidores por extracción. El sistema realiza el resumen utilizando fragmentos previamente extraídos del documento fuente mediante diversas metodologías.Emplea técnicas no supervisadas ya que no requiere que el sistema tenga conocimiento previo alguno de los temas tratados en los documentos a resumir.

Resumidores por compresión. El resumen contiene el mismo número de enunciados que el texto original, pero esas oraciones han sido reducidas comprimiéndolas.

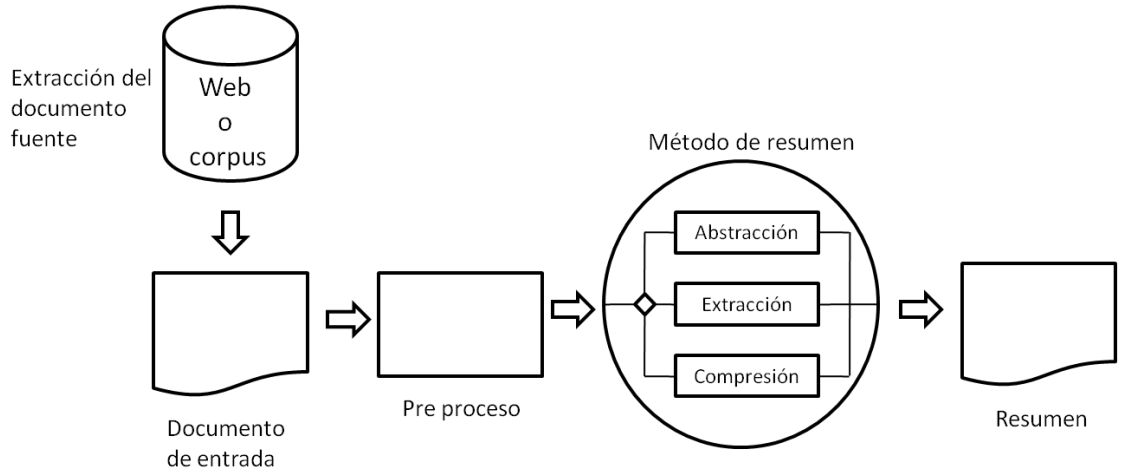
Resumidores por abstracción. En este tipo de sistemas, el resumen está conformado por un texto reducido y corregido y/o parafraseado del texto original.

Del mismo modo, también existe la clasificación de resumidores automáticos de acuerdo con el tipo de documentos que reciben como entrada a procesar: mono-documento, multi-documento, mono-lengua, bilingüe, multilingüe.

2.6 Estructura de un resumidor automático

Para definir la estructura de un resumidor automático se presenta el diagrama siguiente:

Figura(2). Estructura de un sistema resumidor automático.



En el diagrama anterior se representa de manera modular el funcionamiento de un sistema resumidor automático. Para comenzar, se obtiene el documento que se desea resumir de la fuente, ya sea la web o un corpus de alguna temática en especial. Ese texto se pre-procesa para ser filtrado, segmentado y normalizado. El siguiente paso es resumir dicho texto empleando cualquiera de los tres métodos (abstracción, extracción o compresión). El resultado es el resumen del documento de entrada.

Luhn (1959) fue el primero en tocar el tema de resumir textos de manera automática, para ser más precisos, textos de carácter literario. Su técnica de resumen automático fue uno de sus grandes aportes al campo de la computación, la cual inspiró y fue un parteaguas para que distintos investigadores desarrollaran más trabajos al respecto, basados en diferentes metodologías e incursionando en un área prácticamente desconocida. Posteriormente,Edmundson (1969) propuso una metodología para seleccionar las oraciones

más relevantes del texto y así entregárselas al lector. Más tarde, Kupiec*et al.* (1995) generaron un sistema resumidor automático multi-documento capaz de mejorar sus propios resultados basándose en el entrenamiento del mismo sistema.

Dos años después, Marcu realizó el análisis del discurso para obtener resúmenes de documentos. Torres-Moreno señala en *Résuméautomatique de documents* (2011) que para el año 2000, se propuso el desarrollo de resúmenes mediante la compresión de oraciones; que, en el año 2001, varias instituciones de investigación en el área del procesamiento del lenguaje natural crearon herramientas de evaluación de sistemas de resumen automático (entre las que se encuentran *ROUGE, TextRank* y *LexRank* que son herramientas ampliamente reconocidas en el ámbito de PLN); y que en ese mismo año se generaron sistemas basados en la representación del texto mediante grafos.

Asimismo, en el 2002, se inicia la investigación en sistemas resumidores-traductores multilingües. Para el año 2003 se realizan resúmenes a gran escala. Cinco años más tarde, se comenzó a trabajar con las medidas de divergencia de Jensen-Shannon para comparar qué tan bien se había realizado el resumen. En el año 2010 se genera la herramienta de evaluación llamada FRESA[1](la cual mencionaré en la sección 8.2 del capítulo VIII Resultados).

Hoy en día existen diversas propuestas para llevar a cabo resúmenes de manera automática, pero lo que coincide en cada una de ellas es realizar la tarea de extraer la información más relevante de los documentos de entrada sin tener pérdida de información. Algunas de esas propuestas están basadas en el análisis del discurso, en la segmentación en cadenas léxicas del texto, en la representación de las palabras del documento mediante grafos, en la indexación aleatoria, así como en la posición de las palabras en el documento.

Una de las propuestas es la que hace (Smith 2011) realizando una selección de enunciados importantes del texto mediante el uso de indexación aleatoria y posición en la página. De este modo, una vez que el texto se procesó, se extraen las oraciones más

[1] FRESA-Framework for Evaluating Summaries Automatically
http://daniel.iut.univ--metz.fr/~LIA_TALNE/FRESA/

importantes. Es importante destacar que el algoritmo utilizado sólo considera el documento en cuestión como contexto; es decir, sólo la información contenida en el documento, sin ningún conocimiento de un corpus externo, lo cual lo hace funcionar independientemente del área, dominio e, inclusive, idioma.

La metodología que se siguió para realizar este trabajo es una adaptación de la propuesta de calcular la energía textual descrita por Fernández *et al*., (2007) donde se representa el texto, desde un enfoque de redes neuronales inspirado en la física estadística, como un modelo de espacio vectorial. En este enfoque, las palabras del documento a procesar son vistas como un conjunto de unidades interactivas, donde cada una de ellas se ve afectada por el campo creado por las demás palabras que conforman el texto.

ENERTEX calcula la interacción entre los términos y la energía textual entre las frases mediante la proyección de una expresión física en una expresión matricial que permite operar la representación vectorial del texto. Esa expresión física, implementada por Hopfield (1982) en memorias asociativas, permite mediante el modelo magnético de Ising construir una red neuronal capaz de almacenar patrones. Dichos patrones permiten realizar el aprendizaje al sistema mediante la regla de Hebb (Hertz, *et al*. 1991). Posteriormente, la recuperación de información se realiza por minimización de la energía del modelo de Ising.

2.7Traducción automática

La traducción es la tarea que realiza la transferencia de pensamientos e ideas de un idioma a otro. Valdés (1989) señala que la traducción no se refiere solamente a la búsqueda de palabras en un diccionario, sino a la difícil y hasta el momento no perfeccionada tarea de establecer una comunicación entre el pensamiento de los habitantes y los distintos lugares del mundo. De este modo, la traducción obedece a la transferencia de pensamientos, acepciones e ideas de un idioma (origen) a otro (destino). Así como al establecimiento del puente que conecta la expresión de una idea en un idioma con su equivalente, en otro.

Ahora bien, la traducción automática (TA) es una disciplina que forma parte de la lingüística computacional y del procesamiento del lenguaje natural. A través de la

implementación de un sistema de software, lleva a cabo traducciones de un texto o habla, de un idioma a otro sin intervención humana. Esta traducción conlleva la tarea de interpretar y analizar los elementos del texto, así como detectar la influencia de unas palabras en otras, es decir, realizar un análisis sintáctico y semántico, así como poseer conocimientos de gramática respecto a ambos idiomas (origen y destino).

2.8Tipos de traducción automática

Existen dos tipos de TA, la que se realiza a través del uso de reglas y la que se basa en análisis estadístico. En la TA basada en reglas, se emplean numerosas reglas lingüísticas además de diversos diccionarios bilingües para cada par de idiomas en los que se desee trabajar la traducción. Este tipo de sistema lleva a cabo un análisis sintáctico del texto, así como una representación temporal en el idioma en el que éste se traducirá.

Por otro lado, la TA estadística emplea modelos estadísticos. Los parámetros para realizar la traducción se obtienen del análisis de corpus monolingües y bilingües. Básicamente, para poder llevar a cabo la TA, es necesario crear modelos de traducción estadísticos con base en los parámetros previamente mencionados; lo cual no representa un gran problema como lo es la cantidad de palabras o corpus paralelos[2] existentes para así realizar el modelo estadístico pertinente. Esos corpus paralelos demandan como mínimo una extensión de dos millones de palabras para cubrir un dominio específico. Por ende, son pocos los corpus paralelos existentes que permiten la creación de modelos de traducción. Google posee una gran cantidad de información en diversos idiomas (incluyendo corpus paralelos). De ahí que su herramienta de TA se ubique entre los mejores traductores automáticos basados en estadística.

Partiendo de la definición previa de ambos tipos de TA, del desempeño de las herramientas existentes y de la evaluación de los mismos hecha por algunos autores, se puede hablar a grandes rasgos de sus ventajas y desventajas. La calidad de la TA basada en reglas es buena cuando se trabaja independientemente del dominio; de hecho, si se emplean

[2] Donde un corpus paralelo se refiere a la existencia de los mismostextos en dos idiomas.

diccionarios personalizados, la calidad mejora (lo cual no implica que sea lo más conveniente respecto al tiempo de adaptación al diccionario personalizado). En cuanto a la fluidez o cohesión de las palabras en las oraciones o párrafos que constituyen al texto, la TA basada en reglas muestra debilidades, ya que traduce los textos no por oraciones, sino por palabra. Hablando a nivel de hardware, comúnmente este tipo no demanda un hardware específico, es decir, se puede llevar a cabo con un equipo de hardware estándar.

En contraste a la TA basada en reglas, la TA estadística tiene mejores resultados en la fluidez, debido a que realiza la comparación entre dos corpus de igual contenido pero en idiomas distintos. Por ello, la probabilidad de que las palabras estén juntas y se encuentre una transliteración de las ideas completas en lugar de palabra por palabra, es mayor. En general, la calidad de la traducción es buena cuando se cuenta con corpus grandes o en caso de no ser un corpus grande, cuando se generan modelos del lenguaje que abarquen mayores rasgos de los idiomas a trabajar.

Un punto en contra o una debilidad que presenta este tipo de traducción es que se requiere un entrenamiento con los corpus, lo cual es relativamente sencillo, pero a menudo implica el utilizar hardware específico para poder crear y modificar los modelos de traducción. Además de que los entrenamientos con los corpus generales o con los que no pertenecen a un dominio o área en específico, estos sistemas no arrojan los mejores resultados para considerar una buena calidad de traducción.

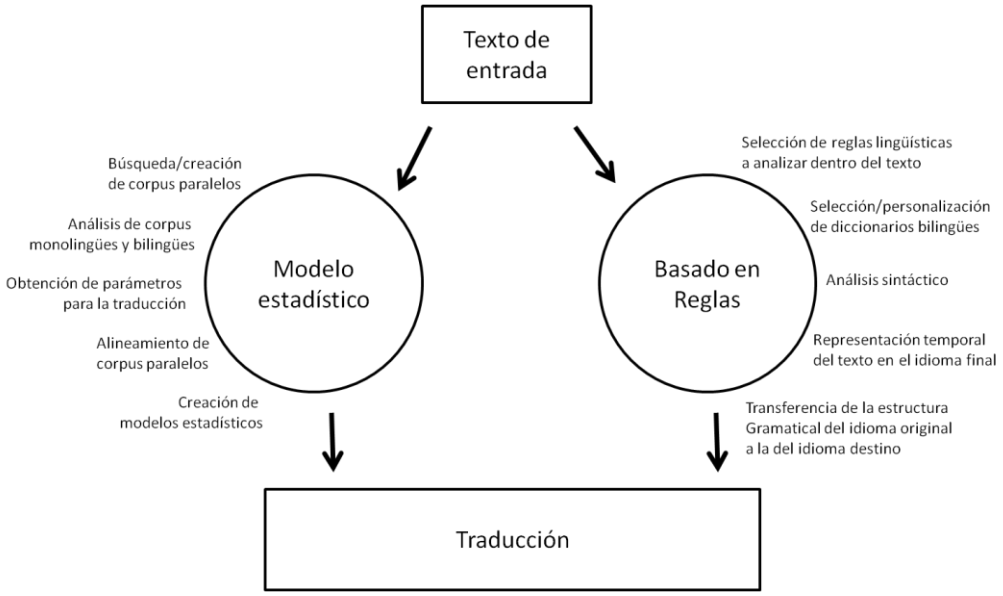
Se puede decir que la existencia de ambos tipos se debe a que las debilidades de un tipo son las fortalezas del otro y que simplemente las técnicas para realizar la TA empleadas en cada tipo, a pesar de ser distintas, pueden complementarse.

2.9 Estructura de la traducción automática

Existen diversos diagramas estructurales de los sistemas de TA existentes (tanto de los basados en reglas como de los estadísticos) que describen de manera modular el proceso que se le aplica al texto de acuerdo con las técnicas implementadas para lograr la traducción.

Enseguida se muestra un diagrama general que contempla ambos tipos de TA, enlistando los procesos que se deben realizar en cada tipo de traducción, ya sea por reglas o estadístico.

Figura (3). Tipos de traductores automáticos



Los sistemas estadísticos (basados en corpus de textos bilingües), son muy rápidos y otorgan resultados considerablemente adecuados. Un ejemplo de este tipo de traductor es MOSES (http://www.statmt.org/moses/), que permite entrenar modelos de traducción para cualquier par de idiomas mediante un algoritmo eficiente de búsqueda que encuentra

rápidamente la probabilidad de traducción más alta entre el número exponencial de opciones.

Los sistemas simbólicos (basados en reglas y en el contexto), requieren más tiempo por el análisis lingüístico (sintáctico-semántico) que se hace al texto previo a la comparación con las posibles traducciones, pero sus resultados son buenos. No existe un sistema de TA que sea puramente simbólico dado que se necesita emplear estadística para comparar los resultados del análisis semántico y sintáctico del texto.

A pesar de que existen diversas combinaciones de ambas metodologías para la TA, en muy pocas se considera el medir qué tan bien fue realizada la traducción. Blatz*et al*. (2003) establecen que la calidad de la traducción automática es un problema de clasificación binaria para distinguir las buenas traducciones de las malas. Por otro lado, Raybaud*et al*. (2009) han hecho estudios recientemente para estimar una puntuación continua (*scoring*) de la calidad de la traducción automática a nivel palabra y a nivel oración.

El sistema de TA utilizado en este proyecto, REVERSO[3], fue desarrollado por *Softissimo*, una compañía de desarrollo de software de origen francés que desarrolla numerosas aplicaciones de TA de alto rendimiento, entre las que destacan los traductores automáticos y los diccionarios electrónicos. Estas aplicaciones permiten realizar las traducciones con gran rapidez y precisión en diversos idiomas. La firma *Softissimo* cuenta con servicios lingüísticos profesionales que garantizan una calidad de resultados reconocida por publicaciones como: *PC expert*, *Windows News* y además ganó el *IST prize*.

El enfoque del traductor automático en dirección inglés al español tiene resultados muy buenos aun considerando que en los diccionarios que emplea, se trabaja el español de España. Dicha traducción no está limitada a trabajar con algún tipo específico de textos. De hecho, son admisibles: e-mails, cartas, informes, e incluso *websites* completos.

REVERSO es un sistema de TA que funciona empleando el enfoque de TA estadístico para generar los resultados traducidos del texto fuente al idioma destino. Existe una versión gratuita en línea del sistema, la cual entrega la misma calidad que el sistema completo (cuando se compra); sólo con la limitante del tamaño del texto a traducir.

[3]El sistema de traducción automática (TA) REVERSO está disponible en: http://www.reverso.net

REVERSO considera sus propios modelos basados en los parámetros generados al analizar corpus paralelos. Además de considerar una funcionalidad de retroalimentación para que los usuarios sugieran la mejor traducción de acuerdo a la cohesión que por su naturaleza el enfoque estadístico no contempla.

En cuanto al trabajo de ambas tareas (resumen y traducción automáticos) en una misma herramienta, éste puede ser abordado mediante diferentes enfoques, tales como mono-documento, multi-documento, documentos bilingües, multi-idioma, entre otros. Hay variaciones que consideran diferentes documentos escritos en diferentes idiomas o la posibilidad de generar el resumen resultante en más de un idioma.

Como fue mencionado, dependiendo de cómo esté escrito el documento de entrada, hay diferentes maneras de llevar a cabo el resumen y la traducción. Para la tarea de resumir una entrada que sea multi-documento y que esos documentos estén escritos en varios idiomas (multi-idiomas), el sistema Columbia Newsblaster (Evans 2004) extrae, traduce al inglés, clasifica y resume documentos de internet (específicamente noticias) escritos en diferentes idiomas.

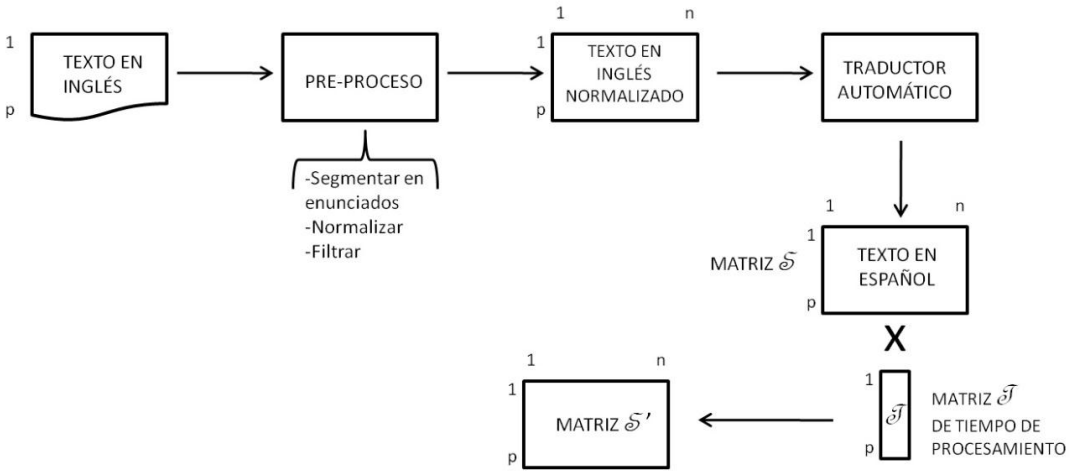
Otra propuesta de sistema de resumen y traducción automáticos es la basada en grafos (Boudin 2011) que considera la calidad de la traducción de las oraciones durante el proceso de selección de las mismas, aplicando una técnica de aprendizaje determinado. El proceso de resumen es elaborado en dos pasos, otorgando una calificación a cada oración del documento y posteriormente, seleccionando las que tengan calificación más alta para incluirlas en el resumen. El proceso de traducción se añade como un paso previo y es hecho a través del sistema traductor de Google.

III. DESARROLLO DEL RESUMIDOR-TRADUCTOR

3.1 Metodología

La metodología a seguir está basada fielmente en el diseño del sistema. La figura 4 muestra mediante módulos la composición de tal sistema para que sea posible describir los procedimientos realizados en cada fase.

Figura (4). Arquitectura general del sistema resumidor-traductor automático



La entrada al sistema es un documento informativo (noticia) en inglés, el cual es segmentado en enunciados, normalizados y filtrados para que sean traducidos por el traductor automático (TA). La salida resultante es una matriz S, con enunciados en los renglones y palabras en las columnas, que contiene el texto traducido al español. Así como una matriz columna T de tiempos de procesamiento de la traducción de cada enunciado del documento de entrada. La matriz S es multiplicada por la matriz T, la cual está formada por los valores inversos de los tiempos obtenidos en la matriz T. Esto para considerar el tiempo de procesamiento de cada enunciado, formando así, la matriz S'. La operación se expresa como sigue.

$$S' = S\ x\ T^{-1}$$

Una vez que se obtuvo la matriz S', se opera con ella para obtener la energía textual referente al texto previamente procesado y así generar una lista de enunciados relevantes del texto que conformarán el resumen. Para obtener la energía textual (matriz E), se multiplica la matriz S' por su matriz transpuesta S'^{T} y el producto se eleva al cuadrado.

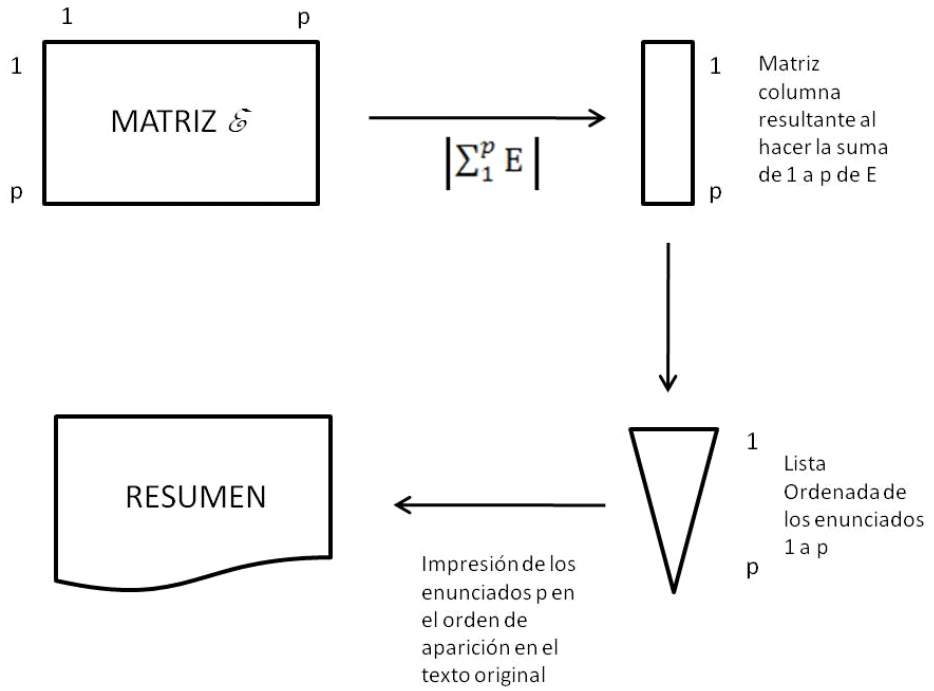
$$E = (S'x\ S'^{T})^2$$

Ya obtenida la matriz E, la cual sigue estando formada por los enunciados en los renglones y por las palabras en las columnas, se hace una suma de los elementos de dicha matriz para obtener una matriz columna que contenga los enunciados ya ordenados por importancia.

$$|\sum E\ |$$

La lista ordenada que se obtuvo es la que se emplea para tomar el número de enunciados que el usuario desee conformen el resumen respecto al texto original. Es decir, si el usuario requiere un extracto del 10% del texto de entrada, se tomarán de la lista aquellos enunciados que conformen el 10% del texto original, y se entregarán en el archivo de salida ordenados respecto a la aparición original en el texto.

Figura (5). Procesos en el algoritmo basado en ENERTEX



3.2 Arquitectura del sistema

El sistema resumidor-traductor automático está estructurado como se muestra en la figura 4, donde se puede resaltar que son dos módulos esenciales (pre-procesamiento [PREP] y generación del resumen mediante ENERTEX [MAT]), los que llevan a cabo el procesamiento del texto y otorgan resultados que son utilizados como entradas para el módulo siguiente, respectivamente. Cabe mencionar que hay tareas como la medición del tiempo de procesamiento de la traducción de cada oración que conforman los textos (tiempo que aquí llamaré *benchmark*), las cuales se llevan a cabo de manera independiente a los módulos debido a que para esta versión del sistema se realizan de manera manual. La arquitectura del sistema obedece a la conformación de cada módulo, ya que a su vez, cada uno de ellos está formado por un conjunto de programas que realizan una tarea específica con el texto de entrada.

El primero de esos módulos (PREP) es el que realiza el pre-procesamiento del texto. Se encuentra situado antes del proceso de traducción de los textos. Este módulo está constituido por un programa escrito en Python. Este programa realiza el normalizado del texto de entrada para que pueda ser manipulado por el traductor automático (y sea posible medir el tiempo de traducción) así como por los programas del módulo siguiente.

El segundo módulo (MAT) está situado después de la medición del tiempo de traducción de los textos a resumir; es decir, cuando ya se tienen los textos en español. Este módulo está formado por tres programas, escritos en PERL, que llevan a cabo el filtrado y la representación vectorial del texto así como las operaciones matriciales pertinentes para la obtención de la matriz E de energía textual. A su vez, este módulo realiza el ordenamiento final de las oraciones que conformarán el resumen objetivo del sistema.

3.3 Desarrollo de las partes de la arquitectura

En esta sección se explica a manera detallada la composición de los módulos del sistema; es decir, se describe el funcionamiento así como las herramientas extras empleadas para que los programas que forman los módulos y los algoritmos implementados desempeñen su función de la manera deseada. Además, de incluir una explicación a detalle de cómo se estructuran los módulos extras empleados.

Los algoritmos empleados en el diseño y estructuración de este sistema están basados en la metodología del VSM (Modelo de Espacio Vectorial por sus siglas en inglés) así como la de la energía textual. Dichos algoritmos logran poner en conjunto la unión y automatización de todos los procesos que culminan con la salida deseada.

Para el módulo PREP se tiene el programa: tokenizer.py. Este programa realiza, como se mencionó anteriormente, la tarea de segmentar y normalizar el texto de los archivos de entrada al sistema. Donde normalizar significa que: dentro del contenido del

texto, aquel que esté en mayúsculas y minúsculas se normaliza a fin de que todo el texto esté escrito en minúsculas y segmentar: dividir el contenido del texto en enunciados.

Por su parte, tokenizer.py utiliza la librería NLTK (*Natural LanguageTool Kit*) de Python para emplear el poder y entrenamiento de su segmentador para dividir el texto en enunciados. Otra razón de usar el NLTK es que los textos se encuentran originalmente en inglés y los recursos que ofrece han sido sobradamente evaluados para dicho idioma. Ese segmentador no es sino la función llamada Tokenize, la cual se puede invocar de la siguiente manera:

```
text = open('/Users/Josh/Desktop/001S.txt', "r").read()
sentences = nltk.sent_tokenize(text)
```

En la primera línea, se indica la ruta donde se encuentra ubicado el archivo que contiene el texto a segmentar. Por su parte, la segunda línea, llama a la función Tokenize del NLTK.

La función tokenize realiza una segmentación del texto en enunciados (aunque en la documentación de NLTK se indica que la segmentación es en oraciones; esto es, *sentences*). El segmentador se entrenó con una muestra de textos en inglés compilada por los desarrolladores de la función. La segmentación que realiza está basada en las probabilidades calculadas de la longitud de oraciones en las que las cadenas del texto son divididas en sub-cadenas. Este principio de funcionamiento permite que el segmentador genere una lista de enunciados contenidos en un archivo de texto.

Dependiendo de los atributos elegidos al momento de llamar la función, y a los fines que tenga la segmentación del texto, ésta puede realizarse a nivel palabra, a nivel punto, o salto de línea. Para fines de este trabajo, lo que se requiere es obtener el texto segmentado por enunciados (cadenas de caracteres entre puntos). La entrada a este programa es el texto original (como fue recopilado de la fuente en inglés) y la salida generada es un archivo que

contiene los enunciados del texto enlistados para su fácil manipulación en los procesos siguientes.

El primer programa del módulo MAT es *filter.pl*; éste realiza el filtrado del texto previamente segmentado (como resulta del traductor automático, ya que éste se alimentó de enunciados). El programa inicia con la normalización del texto para seguir con el etiquetado POS que identifica las partes de la oración (adjetivos, sustantivos, etc.) y da paso al proceso de filtrado final que se realiza con una *stoplist*.

Ese archivo de *stoplist* es un archivo que contiene palabras funcionales como artículos, conjunciones, preposiciones y demás palabras que carecen de significado por sí mismas y que son utilizadas para dar sentido a las palabras del contenido del discurso presente en los documentos.

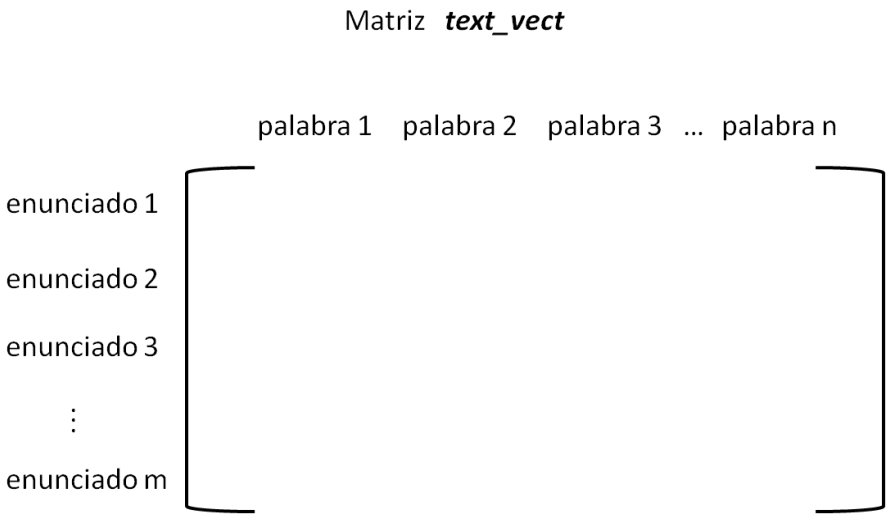
El programa *filter.pl*, que emplea la *stoplist*, utiliza funciones sencillas de PERL, así como expresiones regulares para poder realizar el análisis del texto y lograr obtener las palabras que serán las que alimenten la matriz de frecuencias binarias. Este filtrado se realiza en la mayoría de los pre-procesamientos de lenguaje natural debido a que las palabras que conforman un texto, dependiendo del fin de la tarea, no son todas útiles para realizar una tarea de extracción de información factual.

De manera general, las tareas que impliquen un análisis de lenguaje natural con un fin en específico como extraer, resumir o traducir (o hasta el simple hecho de comparar) requieren contemplar sólo aquellas palabras que tienen "relevancia" y que pueden reflejar o poseer un sentido o significado que tenga relación con las otras palabras o que explique el tema del cual trata el texto.

Siguiendo la descripción modular del sistema; el segundo programa del módulo MAT (*fillmat.pl*) realiza la representación vectorial del texto mediante una matriz de enunciados contra palabras (llamada @text_vect) a través de la definición de arreglos en PERL. Al tiempo que son pre-procesados los textos, los resultados de esos procesamientos

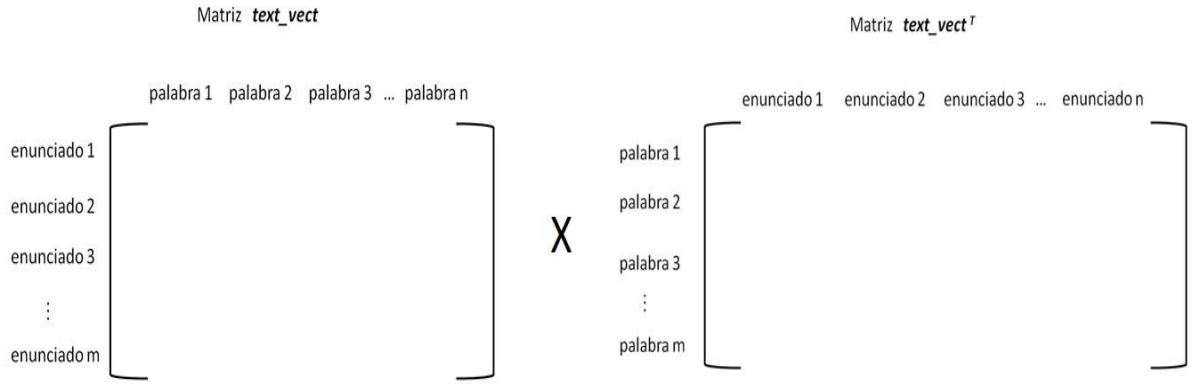
son almacenados en arreglos. En este caso, son esos mismos arreglos los renglones de la matriz.

Figura(6). Estructura de la matriz *text_vect*que almacena la representación vectorial del texto.



El tercer programa del módulo MAT: *ener.pl* toma la matriz generada anteriormente (@text_vect) y realiza las operaciones matriciales correspondientes a la ecuación de la energía textual (ENERTEX). El siguiente paso es ordenar los enunciados respecto a la ponderación obtenida por la energía textual.

Figura (7). Representación de la primera parte de la operación matricial con la matriz text_vect que se representa como la matriz S en la metodología



Las operaciones que implican el cálculo de la energía textual se llevan a cabo con apoyo de las funciones predefinidas en PERL del módulo *Matrix* de la librería *Math*. Estas operaciones son: la transposición y la multiplicación de matrices. Dichas operaciones matriciales reflejan la proyección del modelo físico de Hopfield que mide la manera en cómo interactúan las unidades de los textos (palabras) entre sí y cómo se afectan unas a otras respecto a esa interacción.

Posteriormente se genera una matriz (llamada @ener_index) que sólo contiene la numeración de los enunciados y su frecuencia binaria. Una vez que se tienen los resultados de la energía textual en la matriz recientemente generada, se establece una relación entre la matriz original y la matriz de frecuencias binarias (que es la matriz resultante de hacer los cálculos de la energía textual) mediante la adición de una columna que indica el valor de la frecuencia binaria de cada enunciado.

El último paso de este programa es ordenar esos enunciados o renglones de la matriz original por orden de importancia; es decir, mientras mayor sea el valor del índice de la energía textual, más importante o relevante es el enunciado (respecto a los enunciados restantes).

Este ordenamiento es generado al momento en que el usuario hace la petición del resumen. El usuario especifica a qué porcentaje del texto original desea que el resumen se genere, ya que el ordenamiento es llevado a cabo comparando los índices de energía textual que son asociados a cada enunciado en la matriz (@ener_index), para poder diferenciar los que tengan un índice alto. La última fase de este ordenamiento se realiza cuando se compara la matriz (@ener_index) con el arreglo que almacena los enunciados por primera vez (al salir del proceso de traducción).

Por ejemplo, cuando el usuario desee obtener un resumen al 20% del texto original, el sistema generará el archivo de resumen conteniendo el 20% del texto ordenado como

originalmente apareció en el archivo de entrada, pero sólo conformará ese 20% con los enunciados más importantes.

Una vez realizado este ordenamiento, el archivo de salida (el resumen) está listo para ser creado. Este archivo tendrá contenida la cantidad de enunciados que correspondan al porcentaje que fue solicitado. Cabe recalcar que dichos enunciados estarán ordenados de acuerdo con el orden de aparición en el texto de entrada al sistema. Otra acotación que es importante, es que la generación de este archivo se da hasta que el sistema ha culminado el procesamiento del texto.

3.4 Entrada al sistema

Para que el sistema lleve a cabo con éxito cada uno de los procedimientos que se tienen planteados para la traducción y la generación de un resumen, es necesario que los documentos de entrada tengan ciertas características como son:

- Tipo

El tipo de los documentos debe ser informativo (noticias), debido a que éste fue el enfoque que se planteó para el sistema. Y también a que se conoce de antemano que los sistemas de resumen automático generan resúmenes con mejores resultados cuando trabajan con noticias o textos informativos.

- Longitud

La longitud óptima de los documentos es de una página, la cual se determinó gracias a las pruebas realizadas con documentos de distinta longitud; variando ésta entre corta, larga y mediana. Atribuyendo esos términos a los tamaños de manera particular para este trabajo.

- Formato

El formato de los documentos debe de ser *.txt*, es decir, texto plano. La codificación, de manera idónea para evitar problemas con acentos y caracteres especiales, es ANSI, aunque también permite manipular documentos en codificación UTF-8.

- Idioma

El idioma de los documentos, dado el sentido de la traducción, es inglés. Inglés de Estados Unidos ya que, por un lado, los entrenamientos realizados con el segmentador se realizaron con ese idioma y, por el otro, la traducción se obtiene con mejores resultados contemplando ese inglés (hablando con respecto al traductor utilizado).

3.5 Proceso de recolección de información (datos de entrada al sistema)

La recolección de los documentos de entrada se realizó a través del sitio de noticias en internet: *http://www.cnn.com* del cual se extrajeron 25 noticias de temas aleatorios para comprobar la eficiencia del sistema independientemente de la temática. La recolección de documentos se hizo de manera gradual durante un periodo de tres semanas para que las noticias no estuvieran relacionadas entre sí (aunque esto no garantizara en un 100% que no existiera dicha relación).

El idioma original de las noticias es el inglés de Estados Unidos ya que la página visitada fue la de ese país. Se llevó a cabo una revisión de las publicaciones de noticias en el sitio para saber en qué áreas podrían clasificarse.

Las áreas de las noticias son: política, religión, economía y finanzas, espectáculos, deportes, calentamiento global, entre otras. La distribución de esas áreas en la muestra fue aleatoria. Las áreas definidas en estas líneas fueron determinadas a través del contenido de los textos y basadas en la guía de contenidos de la misma página.

La extensión de las noticias fue de dos páginas y media (a renglón seguido) como máximo y tres cuartos de página como mínimo. La recolección de documentos (noticias) se hizo de manera aleatoria. Sin embargo, se hicieron pruebas previas con noticias muy cortas y con muy largas. Es decir, se obtenían buenos valores aún con noticias cortas y largas, pero se logró identificar que las noticias cortas o aquellas que fueran muy largas producían resúmenes pobres de contenido.

3.6 Características físicas del equipo donde se ejecutó el sistema

El equipo que se empleó para programar y generar las pruebas del sistema se describe con las siguientes especificaciones técnicas:

Marca: Apple Macbook

Procesador: Intel Core 2 Duo a 2.4 Ghz

Memoria RAM: 2 GB de 667 Mhz DDR2 SDRAM

Disco duro: 250 GB Serial ATA

Unidad óptica: SuperDrive (DVD ±R / CD-RW)

Tarjeta de video: Intel GMA X3100 ntegrada de 144 MB de DDR2 SDRAM

Cabe mencionar que las ejecuciones de prueba y evaluación se realizaron solamente en este equipo. De cualquier modo, el sistema se puede ejecutar en cualquier equipo con PERL, Python y las librerías mencionadas instalados. Obedeciendo a que si las características físicas del equipo donde se pruebe, sean menores a las del equipo mencionado arriba, el tiempo de procesamiento puede elevarse.

Los tiempos de ejecución oscilaron desde los 3 hasta los 9.5 minutos. Siendo este último valor, el que se obtuvo con el documento más extenso (269 palabras).

IV. RESULTADOS

4.1 ¿Cómo evaluar un sistema de resumen automático?

Existen artículos que tratan en su totalidad del cómo evaluar un sistema de resumen automático y en general sistemas de procesamiento del lenguaje natural. Dichos artículos como el de Inderjeet (SummarizationEvaluation: AnOverview, 2001) hablan de las diversas técnicas existentes para evaluar un sistema de resumen automático. Mencionan las diferentes métricas que han sido consideradas a través del tiempo en el que se han desarrollado dichos sistemas; algunas de las distintas herramientas de evaluación que existen y qué consideran para evaluarlos.

Inderjeet considera el trabajo de Sparck-Jones y Galliers (1996), donde se expone una clasificación para los métodos de evaluación:

> Methods for evaluating text summarization can be broadly classified into two categories. The first, an intrinsic evaluation, tests the summarization system in of itself. The second, an extrinsic evaluation, tests the summarization based on how it affects the completion of some other task. Intrinsic evaluations have assessed mainly the coherence and informativeness of summaries. Extrinsic evaluations, on the other hand, have tested the impact of summarization on tasks like relevance assessment, reading comprehension, etc.

Dentro de esa clasificación, los puntos a contemplar para la evaluación de los sistemas son, entre otros:

Para los métodos intrínsecos: la coherencia del resumen, la informatividad del mismo, informatividad VS coherencia, la comparación con resúmenes de referencia, evaluación automática, métodos semánticos, así como métodos superficiales.Para los métodos extrínsecos: la evaluación de la relevancia del contenido del resumen respecto a cierto tema, así como tareas de comprensión de lectura.

Sin embargo hoy, después de varias décadas de trabajo en el área de resumen automático, es difícil evaluar la eficiencia de un sistema generador de resúmenesautomático a pesar de la existencia de las diversas técnicas y herramientas de evaluación de estos sistemas. Se dice que esta tarea es complicada, ya que el simple hecho de determinar cuál es un buen resumen, y porqué lo es, es ya un tanto difícil.

De las diferentes técnicas existentes para evaluar un resumen, la más rudimentaria (y por algunos autores la más fidedigna) es la revisión manual. Cuando se pide a cierto número de personas que realicen la evaluación del sistema mediante la lectura del texto y la generación de un resumen por sí mismos para compararlo con el resumen generado por el sistema.O simplemente leer el texto y luego el resumen del sistema para evaluar éste. Dicho procedimiento es por un lado costoso,dado el tiempo que invierten los sujetos en la evaluación(que se refleja en el tiempo en que el desarrollador del sistema evalúa cómo fue el desempeño del mismo) y por otro lado, subjetivo dado que las personas que son elegidas

para evaluar los resúmenes son consideradas pertinentes para hacer dicha tarea según el criterio de quien las elije.

Otra de las técnicas queforma parte de los métodos intrínsecos, de acuerdo con la clasificación propuesta por Sparck-Jones y Galliers (1996) es la evaluación automática. Dicha evaluación contempla numerosos rasgos del resumen respecto al texto resumido, como lo son: *recall* y *rank* de oraciones, medidas de utilidad y de contenido, entre otras.

4.2 Procedimiento para evaluar el sistema

El procedimiento que se llevó a cabo en este trabajo para evaluar el sistema de resumen automático propuesto, basado en ENERTEX, se realiza mediante la implementación de la herramienta FRESA desarrollada en el Laboratorio de Informática de Aviñón (LIA) de la Universidad de Aviñón[4].

FRESA es un *framework,* como su nombre lo dice (*FRameworkforEvaluatingSummariesAutomatically*), que se emplea para evaluar resúmenes de manera automática, es decir, al evaluar el resumen resultante de un sistema generador de resúmenes automático se está evaluando implícitamente el desempeño del sistema.

FRESA toma como entrada el mismo texto del cual el sistema a evaluar generó un resumen y genera un resumen por sí mismo.Posteriormente, le pide al usuario ingresar el resumen generado por el sistema a evaluar. De este modo, se calculan divergencias estadísticas entre los dos resúmenes; el que fue generado por FRESA y el que generó el sistema que se desea evaluar.Estas medidas de divergencia son: Kullback-Liebler (KL) y Jensen-Shannon (JS).

[4] Laboratoire Informatique d'Avignon http://lia.univ-avignon.fr/

La razón por la cual FRESA emplea estas divergencias es queintegra un conjunto de estadísticas que permiten realizar una comparación de distribuciones de probabilidad (a diferencia de losdistintos sistemas de evaluación existentes que utilizan medidas que dependen del uso de n-gramas y del procesamiento aplicado al texto de entrada como coocurrencias de n-gramas, lematización o quitar palabras de una *stop-list*). El resultado de esta comparación es un valor que puede ser usado para calificar el resumen del sistema.

Dado que el desempeño de FRESA ha sido comparado con otras herramientas como Rouge, es garantía el considerar que si el resultado de la divergencia de los resúmenes sometidos a evaluación es más bajo que los resultados otorgados para los resúmenes que el mismo FRESA genera, el resumen evaluado es bueno y de hecho se considera mejor que el que genera FRESA.

Las dos medidas de divergencia de la teoría de la información empleadas por FRESA son descritas a continuación:

$$D_{KL}(P||Q) = \frac{1}{2}\sum_{\omega} P_\omega \log_2 \frac{P_\omega}{Q_\omega}$$

Como se aprecia en esta ecuación, la divergencia de Kullback-Liebler calcula la distancia entre dos distribuciones de probabilidad P y Q, donde P es la distribución de probabilidad de los resúmenes del sistema y Q es la distribución de probabilidad de los resúmenes de referencia.

Por otro lado, la divergencia de Jensen-Shannon realiza la comparación entre las mismas distribuciones de probabilidad pero mediante la siguiente expresión:

$$D_{JS}(P||Q) = \frac{1}{2}\sum_{\omega} P_\omega \log_2 \frac{2P_\omega}{P_\omega + Q_\omega} + Q_\omega \log_2 \frac{2Q_\omega}{P_\omega + Q_\omega}$$

Las medidas de Kullback-Liebler (KL) y Jensen-Shannon (JS) son medidas estadísitcas que permiten calcular divergencias entre distribuciones de probabilidad. Para este caso, estas distancias entre las distribuciones hacen referencia a la distribución de unidades en el resumen generado automáticamente y la distribución de unidades en el resumen modelo generado por FRESA.

Una divergencia mide de manera estadística cuán diferente es un objeto de otro, lo cual, para el caso de FRESA, es útil ya que emplea las medidas de estas divergencias para diferenciar qué tan parecido es un resumen de otro.Bajo numerosas pruebas usando otros sistemas de evaluación (ROUGE, PYRAMIDS, entre otros),el sistema resumidor automático que posee FRESA es un buen parámetro para comparar un resumen.

El conocer la divergencia entre los resúmenes generados por FRESA y los resúmenes generados por el sistema basado en ENERTEX permite evaluar la calidad con la que el sistema genera un resumen de manera automática, partiendo del principio: mientras menor sea la divergencia, mejor calificación obtendrá el resumen generado por el sistema.

Es importante mencionar que el conjunto de noticias que fue utilizado para la evaluación del sistema fue el mismo que el que conformó el corpus de entrada. Este conjunto está constituido por noticias que no tienen un resumen proporcionado por los autores de los textos.

Para tener un marco de comparación mayor respecto a la calidad de los resúmenes; adicionalmente a la implementación de FRESA,se generaron resúmenes de referencia de los documentos de entrada de manera automáticaconsiderando las primeras expresiones de cada uno de los documentos a tres porcentajes: 10%, 15% y 20% (los cuales en adelante serán llamados RA). Cada uno de esos resúmenes de referencia fue evaluado del mismo modo con FRESA.

De esta manera, se obtuvieron seis tipos de resúmenes por documento; tres de ellos eran RA(a 10%, 15% y 20%) y los otros tres habían sido obtenidos del sistema propuesto en este trabajo (en los mismos porcentajes).

Respecto a la evaluación de la traducción automática,se midió manualmente el tiempo de procesamiento al momento de traducircada enunciado y se analizó el contenido de la traducción resultante.

4.3 Datos estadísticos y comparaciones

- Distribución de probabilidad

Se define como una función de una variable aleatoria que asigna a cada evento definido sobre la misma variable la probabilidad de que dicho evento ocurra. De esta manera, la distribución de probabilidad está definida plenamente por la función de distribución cuyo valor en cada x_i es la probabilidad de que la variable aleatoria sea menor o igual que *x*.

$$F_X(x) = P(X \leq x)$$

- Frecuencia de aparición

Representa el número de veces que aparece una palabra en un documento o en un conjunto de documentos. De este modo, se evalúan las frecuencias de aparición de algunas palabras en los documentos con la medida TF-IDF (explicada en el capítulo Marco teórico).

- Media

Se define como un promedio sobre un conjunto de valores que representan algún evento. Para este trabajo se utiliza la media muestral que es la media aritmética de los valores de una muestra de una variable aleatoria.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} a_i \frac{a_1 + a_2 + \ldots + a_n}{n}$$

- Divergencia

Una divergencia, matemáticamente hablando, mide la diferencia entre flujo entrante y saliente de un campo vectorial sobre una superficie. Dentro del contexto de la probabilidad, el concepto se define como un indicador de la similitud entre dos funciones de distribución de probabilidad. En este trabajo se expresa como la similitud o diferencia entre las distribuciones de probabilidad generadas por los resúmenes del sistema ENERTEX y por los resúmenes generados automáticamente.Los valores de esta divergencia son calculados por la herramienta de evaluación FRESA.

- Comparación entre dos medidas de divergencia

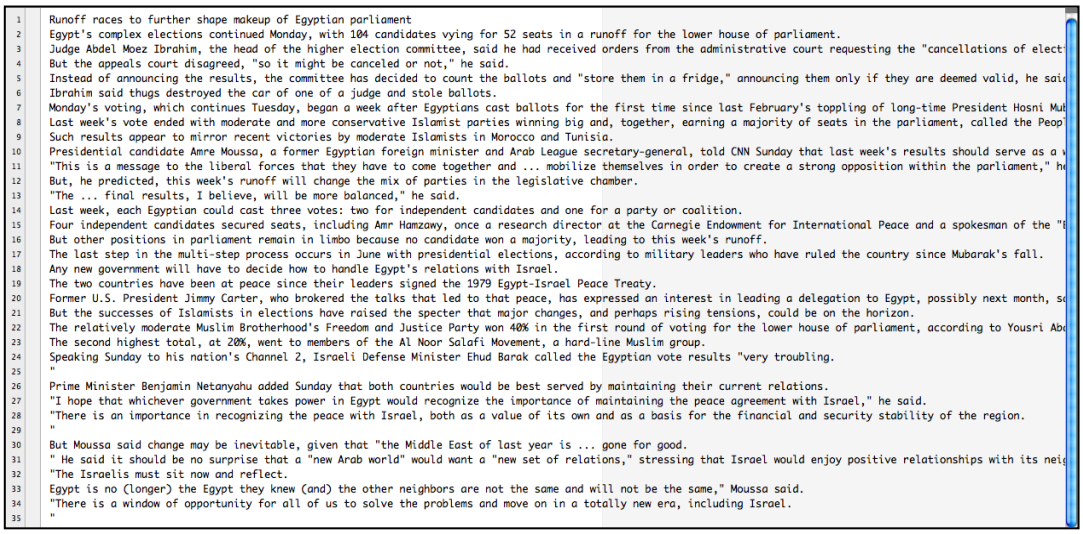
La comparación de dos o más valores de divergencia permite saber qué tan similar o distinta es una distribución de probabilidad de otra; usualmente estas comparaciones son usadas cuando se quiere comparar el comportamiento del resultado de un sistema vectorial.

4.4 Análisis e interpretación de los datos obtenidos

Para comenzar a analizar los resultados se presentará con su respectiva descripción una muestra de los resultados del sistema cada vez que el mismo genera alguna salida durante su proceso de ejecución. Cabe mencionar que algunos resultados harán referencia a un análisis individual de documentos y otros a un análisis general (todos los documentos del corpus). De cualquier manera, se señalará cuando sea el caso respectivamente.

En el primer módulo del sistema, se inserta el texto en inglés, el cual es filtrado y normalizado para ser procesado por el traductor automático como expresiones, las cuales deseablemente son oraciones. El contenido del archivo normalizado se muestra en la siguiente figura:

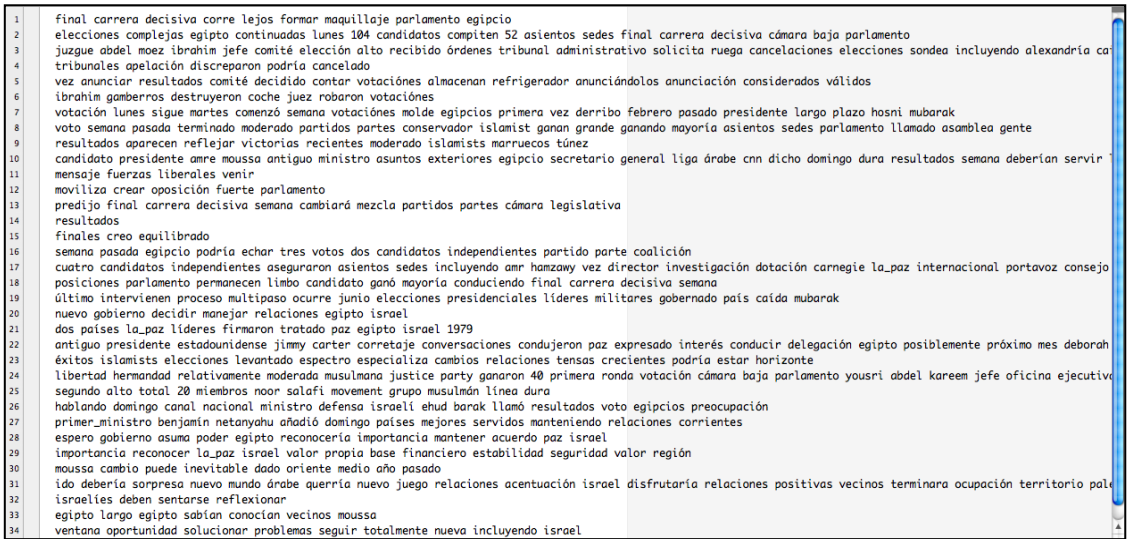
Figura (8). Contenido del documento segmentado



```
Runoff races to further shape makeup of Egyptian parliament
Egypt's complex elections continued Monday, with 104 candidates vying for 52 seats in a runoff for the lower house of parliament.
Judge Abdel Moez Ibrahim, the head of the higher election committee, said he had received orders from the administrative court requesting the "cancellations of elect
But the appeals court disagreed, "so it might be canceled or not," he said.
Instead of announcing the results, the committee has decided to count the ballots and "store them in a fridge," announcing them only if they are deemed valid, he sai
Ibrahim said thugs destroyed the car of one of a judge and stole ballots.
Monday's voting, which continues Tuesday, began a week after Egyptians cast ballots for the first time since last February's toppling of long-time President Hosni Mu
Last week's vote ended with moderate and more conservative Islamist parties winning big and, together, earning a majority of seats in the parliament, called the Peop
Such results appear to mirror recent victories by moderate Islamists in Morocco and Tunisia.
Presidential candidate Amre Moussa, a former Egyptian foreign minister and Arab League secretary-general, told CNN Sunday that last week's results should serve as a w
"This is a message to the liberal forces that they have to come together and ... mobilize themselves in order to create a strong opposition within the parliament," he
But, he predicted, this week's runoff will change the mix of parties in the legislative chamber.
"The ... final results, I believe, will be more balanced," he said.
Last week, each Egyptian could cast three votes: two for independent candidates and one for a party or coalition.
Four independent candidates secured seats, including Amr Hamzawy, once a research director at the Carnegie Endowment for International Peace and a spokesman of the "E
But other positions in parliament remain in limbo because no candidate won a majority, leading to this week's runoff.
The last step in the multi-step process occurs in June with presidential elections, according to military leaders who have ruled the country since Mubarak's fall.
Any new government will have to decide how to handle Egypt's relations with Israel.
The two countries have been at peace since their leaders signed the 1979 Egypt-Israel Peace Treaty.
Former U.S. President Jimmy Carter, who brokered the talks that led to that peace, has expressed an interest in leading a delegation to Egypt, possibly next month, s
But the successes of Islamists in elections have raised the specter that major changes, and perhaps rising tensions, could be on the horizon.
The relatively moderate Muslim Brotherhood's Freedom and Justice Party won 40% in the first round of voting for the lower house of parliament, according to Yousri Abo
The second highest total, at 20%, went to members of the Al Noor Salafi Movement, a hard-line Muslim group.
Speaking Sunday to his nation's Channel 2, Israeli Defense Minister Ehud Barak called the Egyptian vote results "very troubling.
"
Prime Minister Benjamin Netanyahu added Sunday that both countries would be best served by maintaining their current relations.
"I hope that whichever government takes power in Egypt would recognize the importance of maintaining the peace agreement with Israel," he said.
"There is an importance in recognizing the peace with Israel, both as a value of its own and as a basis for the financial and security stability of the region.
"
But Moussa said change may be inevitable, given that "the Middle East of last year is ... gone for good.
" He said it should be no surprise that a "new Arab world" would want a "new set of relations," stressing that Israel would enjoy positive relationships with its neig
"The Israelis must sit now and reflect.
Egypt is no (longer) the Egypt they knew (and) the other neighbors are not the same and will not be the same," Moussa said.
"There is a window of opportunity for all of us to solve the problems and move on in a totally new era, including Israel.
"
```

El siguiente paso enel sistema consiste enfiltrar y normalizar los documentos en español quitando palabras como artículos, conjunciones, etc. de acuerdo con una *stop-list*para que no se les otorgue más peso del que deberían tener por su repetida aparición en cada documento y se otorgue, de este modo, peso a las palabras que tienen mayor impacto en el contexto del documento.Además de remover esas palabras, se hace una segmentación en expresiones separadas por puntos, quedando el texto de la siguiente forma:

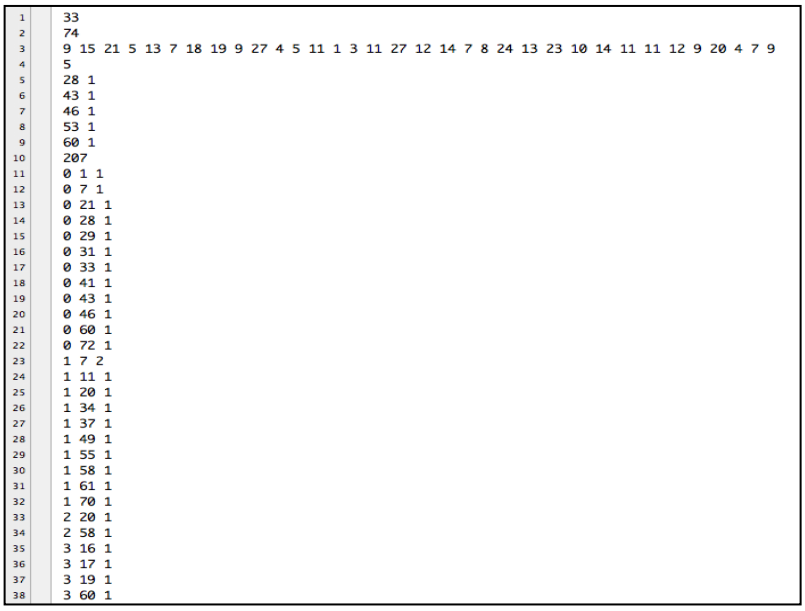
Figura(9).Contenido del documento traducido, segmentado y normalizado



```
final carrera decisiva corre lejos formar maquillaje parlamento egipcio
elecciones complejas egipto continuadas lunes 104 candidatos compiten 52 asientos sedes final carrera decisiva cámara baja parlamento
juzgue abdel moez ibrahim jefe comité elección alto recibido órdenes tribunal administrativo solicita ruega cancelaciones elecciones sondea incluyendo alexandría ca
tribunales apelación discreparon podría cancelado
vez anunciar resultados comité decidido contar votaciónes almacenan refrigerador anunciándolos anunciación considerados válidos
ibrahim gamberros destruyeron coche juez robaron votaciónes
votación lunes sigue martes comenzó semana votaciónes molde egipcios primera vez derribo febrero pasado presidente largo plazo hosni mubarak
voto semana pasada terminado moderado partidos partes conservador islamist ganan grande ganando mayoría asientos sedes parlamento llamado asamblea gente
resultados aparecen reflejar victorias recientes moderado islamists marruecos túnez
candidato presidente amre moussa antiguo ministro asuntos exteriores egipcio secretario general liga árabe cnn dicho domingo dura resultados semana deberían servir
mensaje fuerzas liberales venir
moviliza crear oposición fuerte parlamento
predijo final carrera decisiva semana cambiará mezcla partidos partes cámara legislativa
resultados
finales creo equilibrado
semana pasada egipcio podría echar tres votos dos candidatos independientes partido parte coalición
cuatro candidatos independientes aseguraron asientos sedes incluyendo amr hamzawy vez director investigación dotación carnegie la_paz internacional portavoz consejo
posiciones parlamento permanecen limbo candidato ganó mayoría conduciendo final carrera decisiva semana
último intervienen proceso multipaso ocurre junio elecciones presidenciales líderes militares gobernado país caída mubarak
nuevo gobierno decidir manejar relaciones egipto israel
dos países la_paz líderes firmaron tratado paz egipto israel 1979
antiguo presidente estadounidense jimmy carter corretaje conversaciones condujeron paz expresado interés conducir delegación egipto posiblemente próximo mes deborah
éxitos islamists elecciones levantado espectro especializa cambios relaciones tensas crecientes podría estar horizonte
libertad hermandad relativamente moderada musulmana justice party ganaron 40 primera ronda votación cámara baja parlamento yousri abdel kareem jefe oficina ejecutiv
segundo alto total 20 miembros noor salafi movement grupo musulmán línea dura
hablando domingo canal nacional ministro defensa israelí ehud barak llamó resultados voto egipcios preocupación
primer_ministro benjamín netanyahu añadió domingo países mejores servidos manteniendo relaciones corrientes
espero gobierno asuma poder egipto reconocería importancia mantener acuerdo paz israel
importancia reconocer la_paz israel valor propia base financiero estabilidad seguridad valor región
moussa cambio puede inevitable dado oriente medio año pasado
ido debería sorpresa nuevo mundo árabe querría nuevo juego relaciones acentuación israel disfrutaría relaciones positivas vecinos terminara ocupación territorio pal
israelies deben sentarse reflexionar
egipto largo egipto sabían conocian vecinos moussa
ventana oportunidad solucionar problemas seguir totalmente nueva incluyendo israel
```

El siguiente paso en el sistema es realizar la representación vectorial del texto. Formar la matriz que represente las palabras y oraciones de un documento.

Figura (10).Representación vectorial del documento traducido, segmentado y normalizado



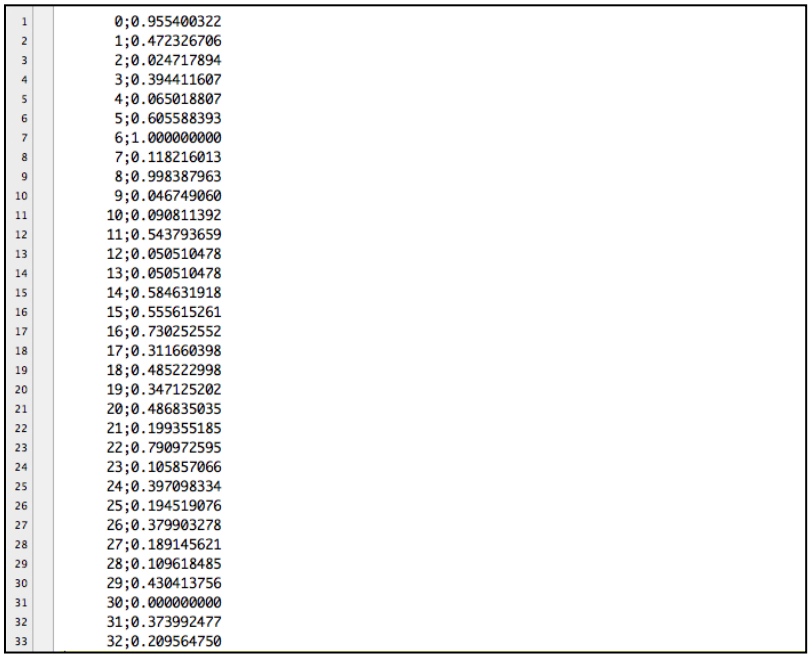
```
33
74
9 15 21 5 13 7 18 19 9 27 4 5 11 1 3 11 27 12 14 7 8 24 13 23 10 14 11 11 12 9 20 4 7 9
5
28 1
43 1
46 1
53 1
60 1
207
0 1 1
0 7 1
0 21 1
0 28 1
0 29 1
0 31 1
0 33 1
0 41 1
0 43 1
0 46 1
0 60 1
0 72 1
1 7 2
1 11 1
1 20 1
1 34 1
1 37 1
1 49 1
1 55 1
1 58 1
1 61 1
1 70 1
2 20 1
2 58 1
3 16 1
3 17 1
3 19 1
3 60 1
```

Una vez que el texto ha sido filtrado y posteriormente representado como vectores en la matriz de documentos, se asignan pesos a las palabras con respecto a su aparición en el listado de palabras esenciales[4] del documento, así como a su aparición en el mismo. Posteriormente se realizan las operaciones matriciales correspondientes al cálculo de la energía textual (expuestas en el capítulo: Metodología).

Al finalizar esta operación con los pesos de las palabras de cada expresión, se procede a sumar los renglones de la matrizde energía textual para obtener una cantidad por renglón (por expresión) del documento.Dichos enunciados son las que conforman el resumen del documento original.

[4] Una palabra esencial es una palabra que se encuentra en la lista del vocabulario esencial que se obtuvo de cada documento en el filtrado inicial del sistema.

Figura (11).Valores de energía textual asociado a los enunciados del documento



```
0;0.955400322
1;0.472326706
2;0.024717894
3;0.394411607
4;0.065018807
5;0.605588393
6;1.000000000
7;0.118216013
8;0.998387963
9;0.046749060
10;0.090811392
11;0.543793659
12;0.050510478
13;0.050510478
14;0.584631918
15;0.555615261
16;0.730252552
17;0.311660398
18;0.485222998
19;0.347125202
20;0.486835035
21;0.199355185
22;0.790972595
23;0.105857066
24;0.397098334
25;0.194519076
26;0.379903278
27;0.189145621
28;0.109618485
29;0.430413756
30;0.000000000
31;0.373992477
32;0.209564750
```

Después de haber obtenido los resúmenes de los documentos, se insertan los textos originales y los resúmenes generados por ENERTEXen FRESA para obtener resultados de las divergencias con los resúmenes que genera FRESA.

4.5 Análisis individual de los resultados

Como un análisis particular de los resultados para evaluar el desempeño del sistema propuesto,a continuación se mostraránlos valores de las divergencias respecto a algunos documentos del corpus, los cuales arrojaron los resultados más significativos.Se unen en una tabla, para comparar dicha divergencia, con las de los RAen distintos porcentajes.Más adelante, se presentará un análisis general del desempeño del sistema de acuerdo con los resultados obtenidos.

En las siguientes tablas se muestran los resultados de las medidas de divergencia Jensen-Shannon (JS) y Kullback-Leibler (KL) obtenidos por FRESA, de los resúmenes

generados al 10%, al 15% y al 20% de tres documentos distintos. De igual modo y como se describió en la sección de evaluación, se comparan con los valores de las medidas de divergencia obtenidos para los resúmenes RA. En esencia, cada tabla muestra en el primer renglón los resultados de los tres resúmenes RA; en el segundo renglón los resultados de los resúmenes generados por ENERTEX y en el tercer renglón aquellos calculados por FRESA.

En la primera tabla se muestran los resultados de la medida de divergencia Jensen-Shannon (JS) para un documento que reflejó que el desempeño del sistema basado en ENERTEX no fue el esperado para dos de los porcentajesen los que se generaron los resúmenes (los valores de la divergencia fueron mayores de lo esperado).

Tabla (1). Resultados de la divergencia Jensen-Shannon (JS) evaluados por FRESA.

Tipo de resumen	10%	15%	20%
Referencia FRESA	3.12057	3.12057	3.12057
Referencia RA	2.46236	2.46236	2.46236
Sistema ENERTEX	2.38968	2.57715	2.57715

Para el resumen generado por el sistema propuesto (ENERTEX) al 10% del texto, se observa que el valor de la divergencia es menor, es decir, tiene una diferencia positiva de 0.08 (el valor de ENERTEX es menor).En cuanto a los otros porcentajes de resumen, podemos ver que ENERTEX presenta una diferencia negativaen los resultados de la divergencia de 0.11 (el valor de ENERTEX es mayor). De modo que ENERTEX no fue lo suficientemente bueno cuando se le solicitóque generara resúmenes en esos porcentajes.

Cabe mencionar que los resultados de la Tabla (1)corresponden a uno de los documentos más cortos(105 palabras).

Por otro lado, se tienen los resultados de la divergencia Kullback-Leibler comparando los resúmenes del mismo documento generados por ENERTEX con aquellos de referencia.

Tabla (2). Resultados de la divergencia Kullback-Leibler (KL) evaluados por FRESA.

Tipo de resumen	10%	15%	20%
Referencia FRESA	14.22204	14.22204	14.22204
Referencia RA	11.59062	11.59062	11.59062
Sistema ENERTEX	11.48663	12.03997	12.03997

Al revisar las divergencias KL se puede ver de nuevo que en los resúmenes generados al 15% y al 20% del texto original, ENERTEX no se desempeñó bien, mostrando una diferencia negativa de 0.44 con respecto a los resúmenes RA(la divergencia de ENERTEX es mayor).De nueva cuenta, estos resultados hacen referencia al mismo documento de la tabla anterior.

En la siguiente tabla se puedenver resultados de ENERTEX un poco mejores; esto es, las diferencias positivas son mayores (la divergencia de ENERTEX es considerablemente menor que las de RA y FRESA) y las diferencias negativas son menores (la divergencia de ENERTEX es un poco mayor que la de los valores de RA y de FRESA).

Tabla (3). Resultados de la divergencia Jensen-Shannon (JS) evaluados por FRESA.

Tipo de resumen	10%	15%	20%

Referencia FRESA	2.54579	2.54579	2.54579
Referencia RA	2.4129	2.1071	2.4129
Sistema ENERTEX	2.33517	2.37228	2.37228

Es notable la persistencia del bajo desempeño deENERTEX cuando se genera un resumen al 15%. Sin embargo, es destacable que bajo dos porcentajes de tres, el resultado es mejor. Por ello, se consideró que el desempeño del sistema para este documento mostró una mejoría respecto al mostrado en la Tabla (2).

Similarmente, los resultados de la medida de divergencia Kullback-Leibler (KL), para este mismo documento, muestran lo siguiente:

Tabla (4). Resultados de la divergencia Kullback-Leibler (KL) evaluados por FRESA.

Tipo de resumen	10%	15%	20%
Referencia FRESA	12.99419	12.99419	12.99419
Referencia RA	12.41901	11.0308	12.41901
Sistema ENERTEX	11.64875	12.31363	12.31363

Se observa una diferencia entre los dos tipos de resumen positiva de 0.77entre los resúmenes generados al 10%, negativa de 1.28entre aquellos generados al 15% y positiva de 0.09entre los que fueron generados al 20%. En comparación con la divergencia JS, sí hubo una diferencia considerable entre el valor obtenido por ENERTEX y el valor asociado al resumen RA al evaluar el resumen generado al 15% (1.28).No obstante, los resultados de

ENERTEX fueron mejores que los resultados del resumen RAen la evaluación al 10% y al 20% (tuvieron menores divergencias).

A continuación se muestra otro análisis como parte de la evaluación individual para el documentocon el cual el sistema mostró el mejor desempeño.La siguiente tabla muestra que los resultados de la medida de divergencia favorecieron a ENERTEX (fueron aún menores) en los tres porcentajes.

Tabla (5). Resultados de la divergencia Jensen-Shannon (JS) evaluados por FRESA.

Tipo de resumen	10%	15%	20%
Referencia FRESA	2.26734	2.26734	2.26734
Referencia RA	2.27055	2.27055	2.27055
Sistema ENERTEX	1.72643	2.20994	2.08431

De la tabla anterior se observa que la medida de divergencia correspondiente a ENERTEX permaneció por debajo de los valores obtenidos para los resúmenes RA. La mayor diferencia se dio en el caso del resumen generado al 10% (0.55) mientras que la menor ocurrió al 15% (0.06). De cualquier manera, los tres resúmenes generados en los distintos porcentajes obtuvieron mejores valores de divergencia que los obtenidos por los resúmenes RA; calificando al resumen de este documento como el mejor obtenido.

En cuanto a la medida de divergencia KL se obtuvieron resultados similares a los de la divergencia JS.

Tabla(6). Resultados de la divergencia Kullback-Leibler (KL) evaluados por FRESA.

Tipo de resumen	10%	15%	20%
Referencia FRESA	10.49238	10.4938	10.49238

Referencia RA	10.51794	10.51794	10.51794
Sistema ENERTEX	8.73392	10.37572	9.72557

La tabla permite observar que la mayor diferencia corresponde al resumen generado al 10% (1.78). Las diferencias del 15% y 20% son 0.14 y 0.78 respectivamente. Para este documento, todas las diferencias son positivas, lo cual reitera que fue aquél con el cual el sistema logró su mejor desempeño.

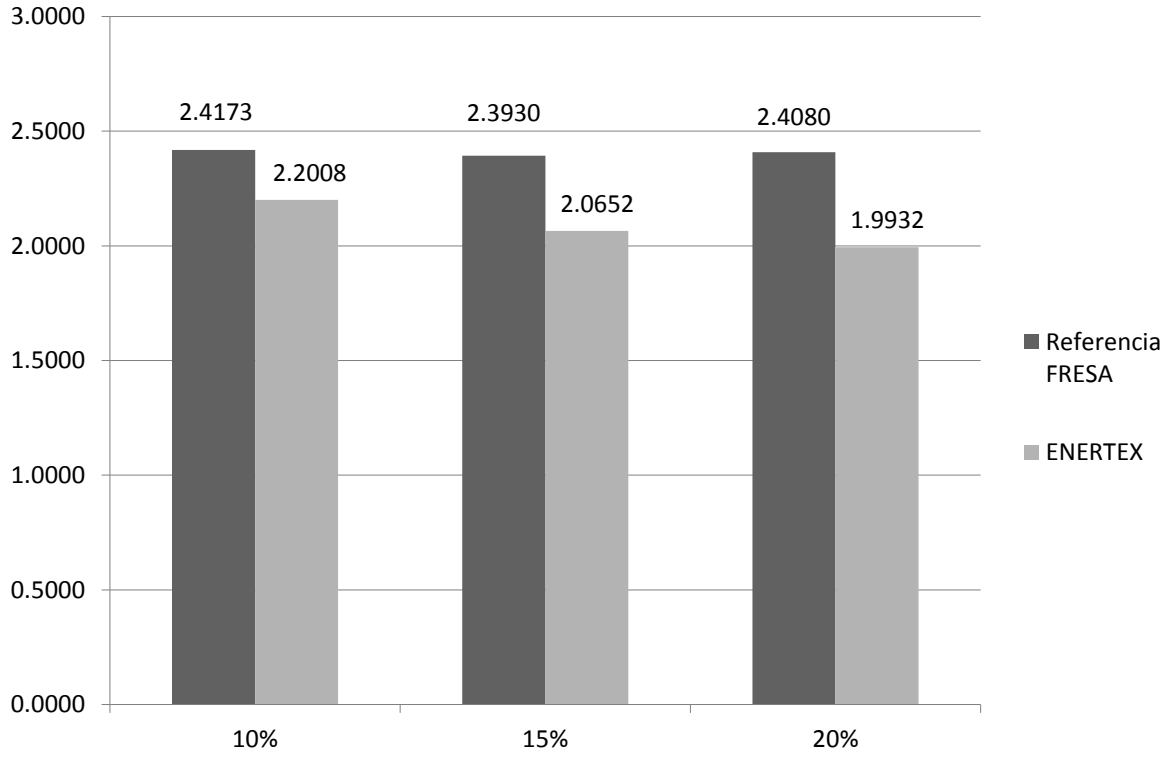
4.6 Análisis general de los resultados

A manera de análisis general del desempeñodel sistema de resumen automático propuesto (ENERTEX) se muestran las siguientes gráficas para analizar el comportamiento general del sistema (sobre todos los documentos) al calcular la media de las medidas de divergencia obtenidas en la evaluación realizada por FRESA.

Cada gráfica, por separado, muestralas medias de las medidas de divergencia calculadas para el sistema propuesto y para los resúmenes RA(generados automáticamente con las primeras enunciados de los documentos)junto con las medias para los resúmenes generados por FRESA.

Primero se realiza una comparación entre los resultados de la evaluación de la divergencia Jensen-Shannon (JS) entre ENERTEX y los resúmenes RA(ambos son comparados por separado con los resúmenes de referencia generados por FRESA). Posteriormente se mostrarán las gráficas que complementan el análisis, mediante la evaluación de la divergencia Kullback-Leibler (KL).

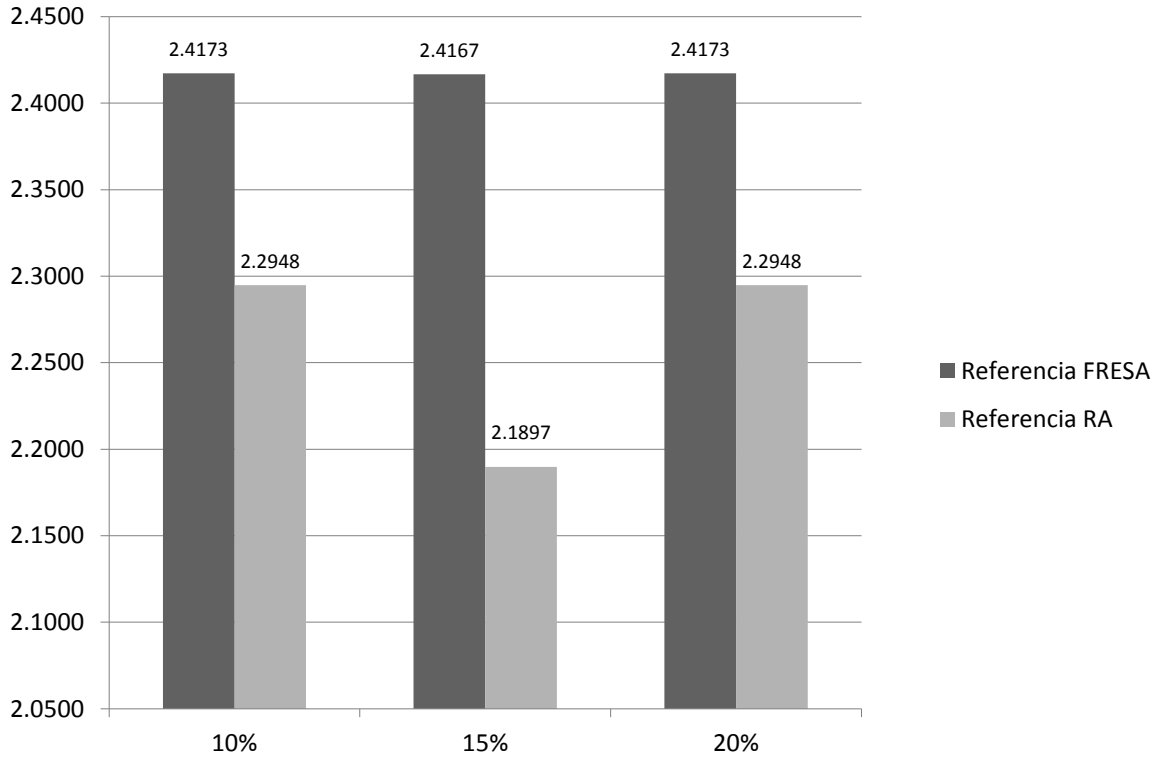
Gráfica (1). Media de la medida de divergencia Jensen-Shannon (JS) con el sistema ENERTEX y los resúmenes de referencia de FRESA



De la gráfica anterior se puede observar que las medidas de divergencia JS fueron mejores para el sistema ENERTEX, comparado con las medidas de divergencia obtenidas para el resumen que genera FRESA. En los tres porcentajes, la media (correspondiente al valor de la medida de divergencia en todos los resúmenes) fue más baja para ENERTEX con una diferencia positiva máxima de 0.41 y la mínima de 0.32.

Enseguida se muestra una gráfica que contiene las medias de las medidas de divergencia obtenidas por FRESA de los resúmenesRA y de los resúmenes que genera FRESA.Esta gráfica se compara con la anterior ya que ambas contienen las medias de las medidas de divergencia JS.

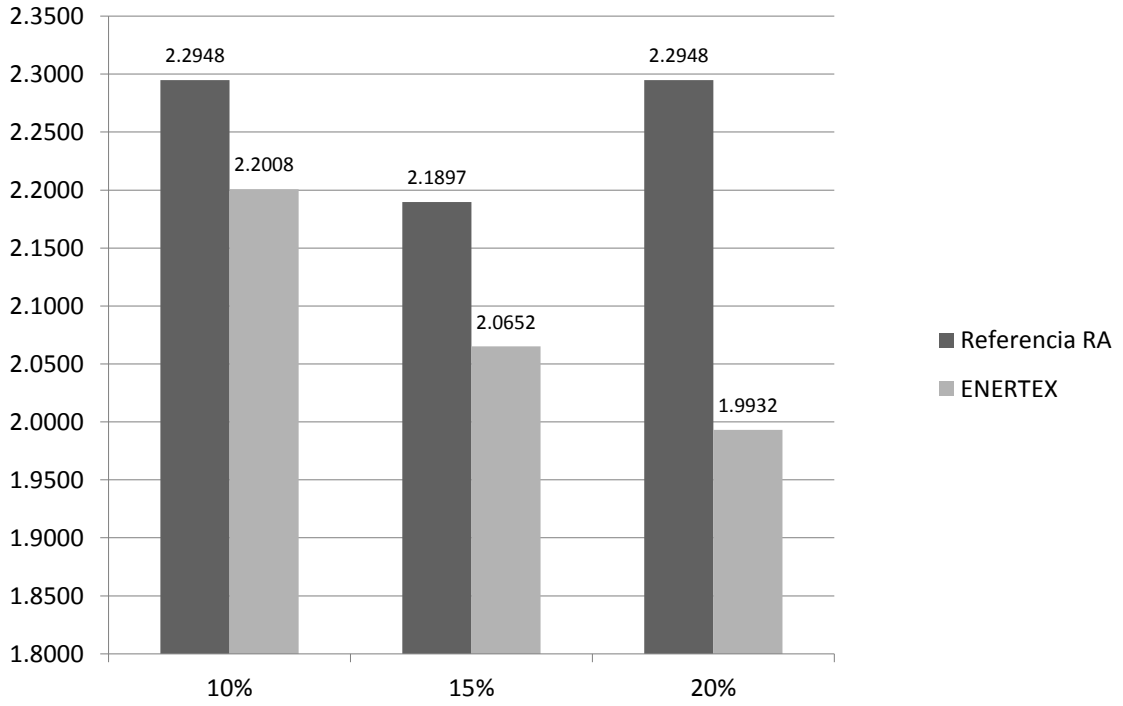
Gráfica (2). Media de la medida de divergencia Jensen-Shannon (JS) con los resúmenes RA y los resúmenes de referencia de FRESA.



La gráfica muestra que las medias de la medida de divergencia JS entre los resúmenes RAtambién se encuentran por debajo de las medias obtenidas para los resúmenes que genera FRESA, obteniendo una diferencia máxima de 0.22 y una mínima de 0.12 repetida para dos de los tres porcentajes (10% y 20%).

A continuación se presenta una gráfica que compara directamente las medias de la medida de divergencia JS para el sistema ENERTEX y los resúmenes generados de manera automática.

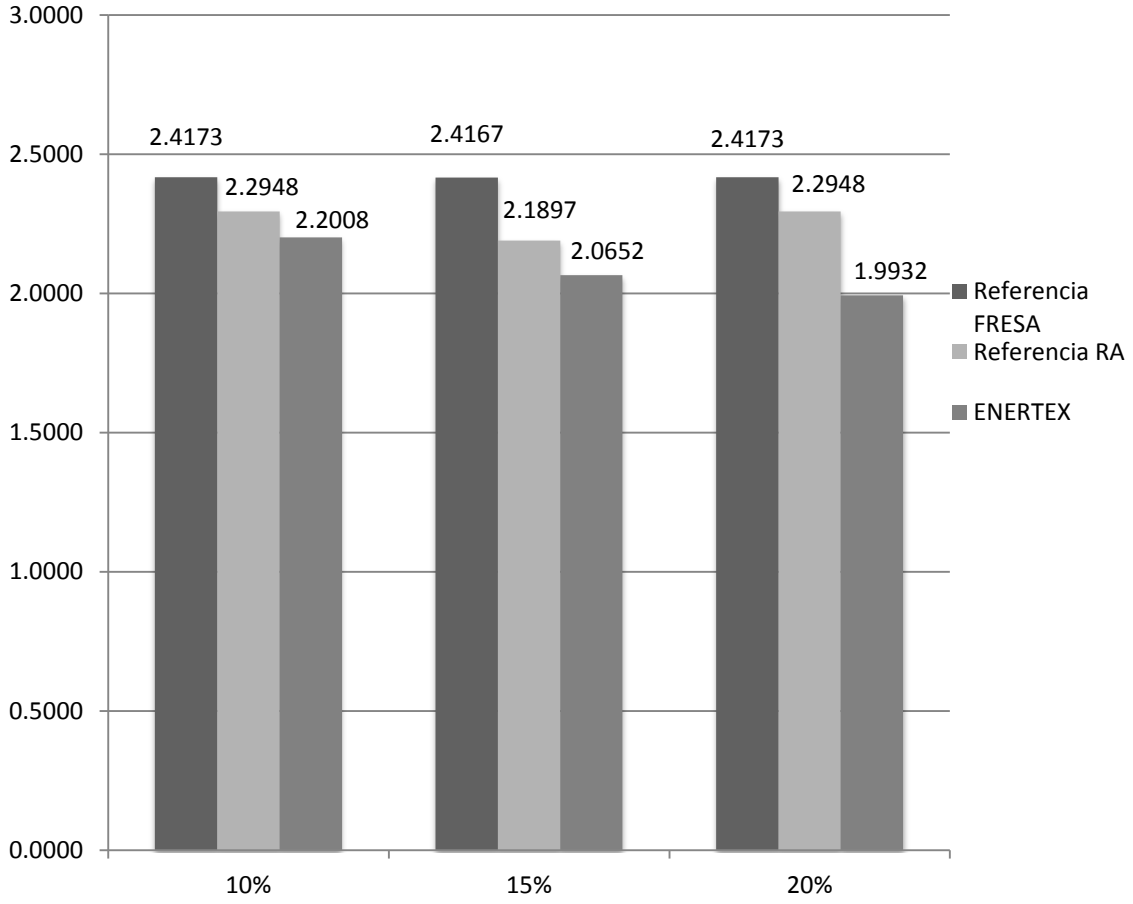
Gráfica (3). Media de la medida de divergencia Jensen-Shannon (JS) con ENERTEX y los resúmenes RA.



La gráfica anterior permite ver que la diferencia entre las medias de las medidas de divergencia calculadas para ENERTEX son más bajas que las que fueron calculadas para los resúmenes RA. Demostrando que el desempeño del sistema fue bueno dadas las diferencias positivas de 0.09, 0.12 y 0.30 respectivamente para los tres porcentajes (10%, 15% y 20%).Siendo la última la máxima diferencia asociada a los resúmenes generados al 20% del texto.

Dados estos resultados, hasta este punto el desempeño del sistema fue mejor que los resúmenes RA. A continuación se muestra una gráfica que permite comparar el desempeño del sistema propuesto (ENERTEX) con ambos tipos de resúmenes de referencia (RA y FRESA).

Gráfica (4). Media de la medida de divergencia Jensen-Shannon (JS) para ENERTEX, los resúmenes de referencia RA y los resúmenes de referencia de FRESA.

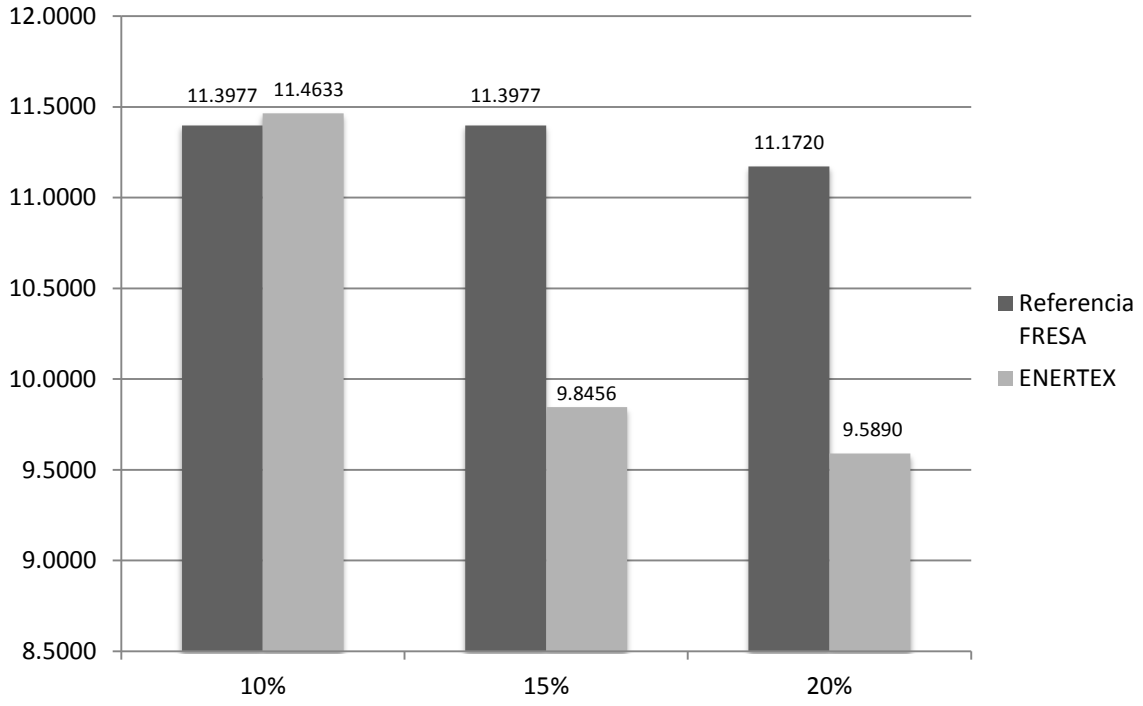


En la gráfica se observa que para los tres porcentajes las medidas de divergencia obtenidas para ENERTEX son las más bajas. Demostrando así que su desempeño, evaluado con la divergencia Jensen-Shannon fue muy bueno y de hecho, el mejor de los tres tipos de resúmenes evaluados. Para el 10% se tiene una diferencia positiva de 0.12, para el 15% una de 0.23 y para el 20% una de 0.30. Se puede concluir que el mejor resultado fue obtenido al generar resúmenes al 20% del texto original.

A continuación se muestran las evaluaciones generales de la medida de divergencia Kullback-Leibler siguiendo la misma estructura, primero comparando los valores de ENERTEX con los valores de los resúmenes de referencia generados por FRESA, después comparando losvalores de los resúmenes de referencia generados automáticamente por

FRESA con los resúmenes RA y finalmente, se comparan los valores de los resúmenes asociados a ENERTEX con los valores asociados con los resúmenes RA.

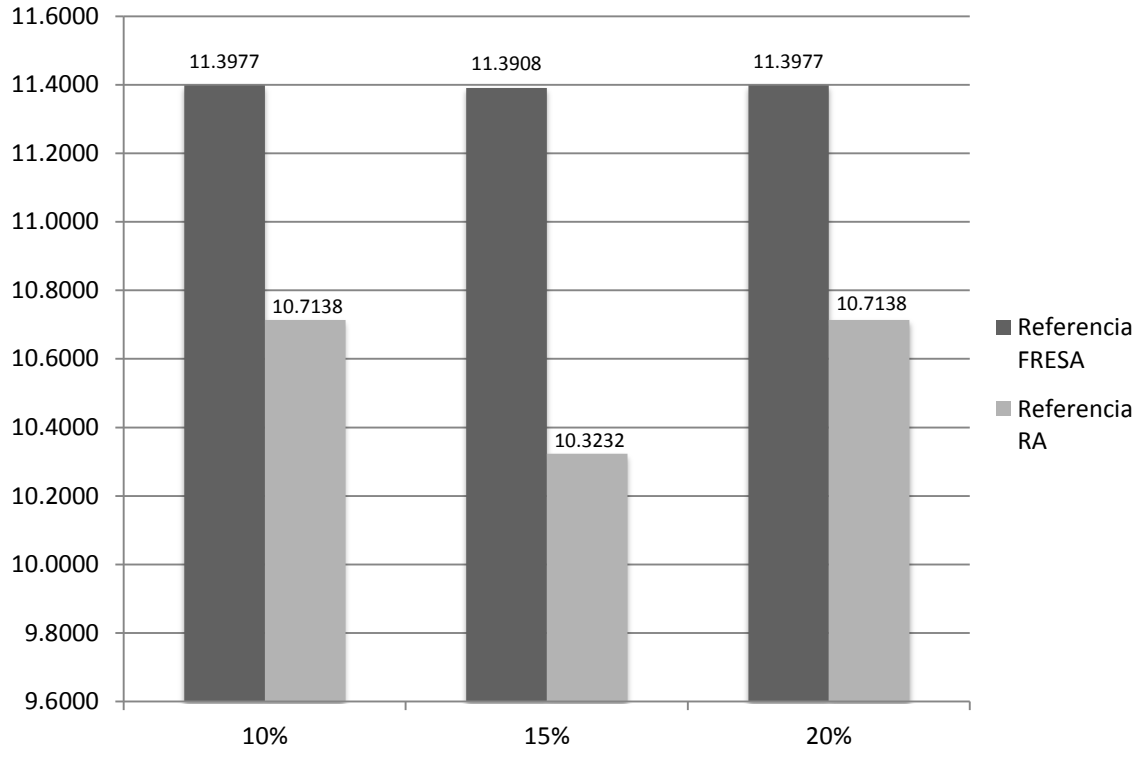
Gráfica (5). Media de la medida de divergencia Kullback-Leibler (KL) con el sistema ENERTEX y los resúmenes de referencia de FRESA.



En la gráfica de arriba se observa que la media de la medida de divergencia correspondiente a los resúmenes generados al 10% por ENERTEX es más alta (peor) que la de los resúmenes de referencia de FRESA con una diferencia negativa de 0.07; lo cual no es una gran diferencia, pero existe. Y al estar considerando medias de las medidas de divergencia, esto permite observar que el desempeño del sistema siempre tuvo problemas al generar resúmenes en este porcentaje.

Enseguida se muestra la gráfica de las medias de las medidas de divergencia KL correspondientes a los resúmenes RA y a los resúmenes de referencia generados por FRESA.

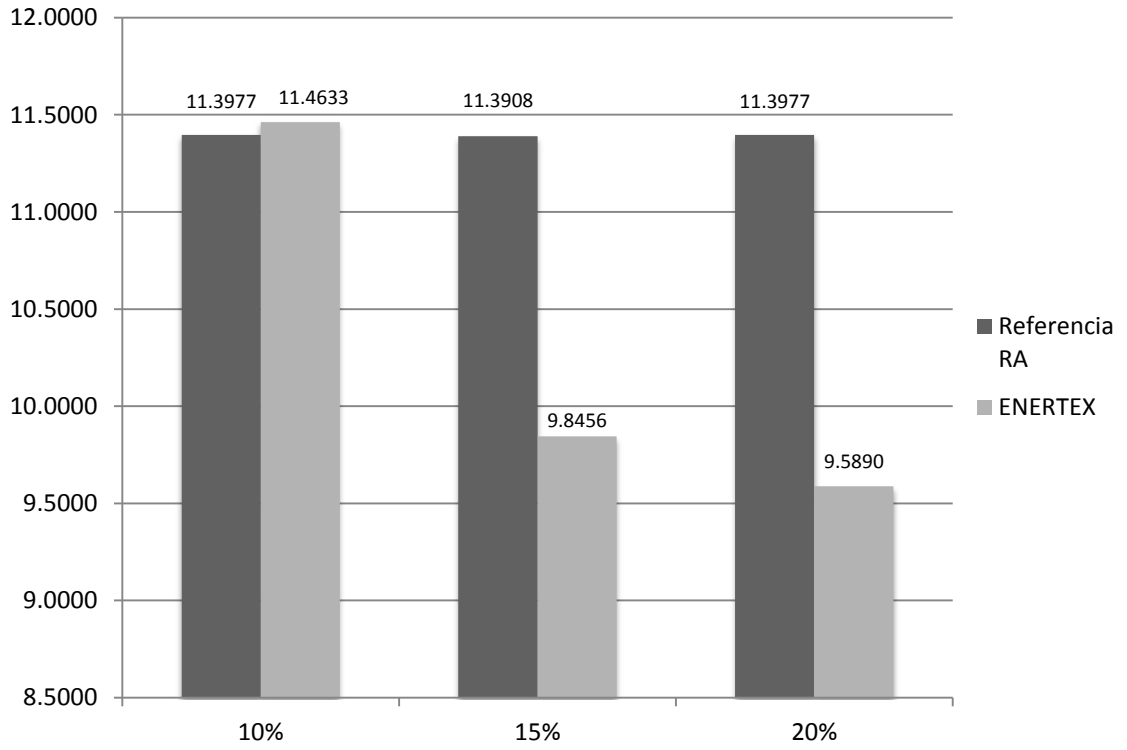
Gráfica (6). Media de la medida de divergencia Kullback-Leibler (KL) con los resúmenes RA y los resúmenes de referencia de FRESA.



Por un lado, se observa que las medias de las medidas de divergencia correspondientes a los resúmenes generados automáticamente se encuentran en su totalidad por debajo de las medias que corresponden a los resúmenes de referencia generados por FRESA, lo que indica que los resúmenes generados automáticamente con los primeros enunciados de los textos son buenos.

Por otro lado, y para concluir el análisis general del desempeño del sistema con todos los documentos que formaron el corpus de entrada, se presenta la gráfica que compara las medias de las medidas de divergencia entre los resúmenes generados con ENERTEX con los resúmenes de RA.

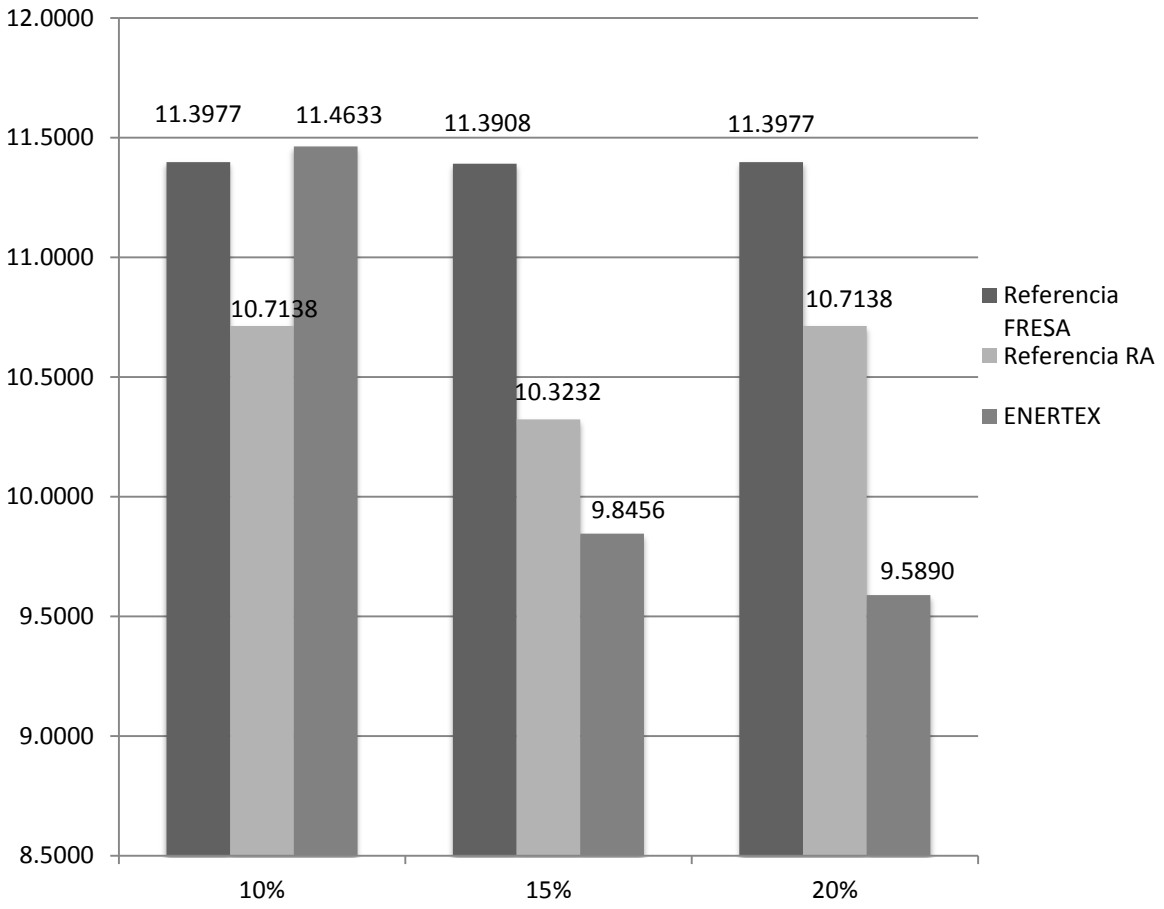
Gráfica (7). Media de la medida de divergencia Kullback-Leibler (KL) con ENERTEX y los resúmenes RA.



En la gráfica anterior se observa que la media de la medida de divergencia se encuentra elevada para ENERTEX en los resúmenes generados al 10%. Demostrando que el desempeño de ENERTEX no fue el mejor al momento de generar resúmenes al 10% del texto original. Para los otros dos porcentajes, se obtuvo una media más baja. Con una diferencia positiva de 1.55 al 15% y de 1.59 al 20%, siendo ambas diferencias considerables ya que son mayores a la unidad.

Enseguida se mostrará una gráfica que conjunta los valores de las divergencias obtenidos para los tres tipos de resúmenes (referencia FRESA, referencia RA y ENERTEX).

Gráfica (8). Media de la medida de divergencia Kullback-Leibler (KL) con ENERTEX, los resúmenes RA y los resúmenes de referencia de FRESA.



En la tabla anterior se puede observar mediante un análisis global, comparando los valores de divergencia para los tres tipos de resumen (referencia FRESA, referencia RA y ENERTEX) que ENERTEX obtiene el mejor valor (el menor) para dos de los porcentajes y mayor incluso que el obtenido por los resúmenes de referencia de FRESA.

V. CONCLUSIONES

A lo largo del desarrollo de este trabajo se observó el seguimiento de la arquitectura del sistema automático de resumen y traducción planeada inicialmente. Dicha arquitectura está basada en el flujo que debe tomar el procesamiento de documentos para poder realizar un

análisis lingüístico y prepararlo para obtener la información que se desea conocer a través de procesos automatizados.

El módulo de pre-procesamiento empleadopara formar la parte inicial del flujo que siguió al análisis de los documentos fue desarrollado bajo técnicas y algoritmos previamente probados y entrenados, como lo es *NLTK*, del lenguaje de programación*Python*. Dada la base de la cual se partió al desarrollar este módulo, los enunciados fueron segmentados de manera eficiente. La segmentación de los documentos se realizó por completo y sin ninguna complicación.

Los filtros de normalización fueron desarrolladosen su totalidad sin ningún *framework* o *tool kit* empleado como marco de referencia; considerando el análisis del contenido de los documentos de entrada.La normalización fue exitosa para todos los documentos.

La implementación del sistema traductor automático se llevó a cabo empleando el traductor en línea REVERSO. Este traductor generó buenos resultados de traducción de las palabras de acuerdo con su contexto en los enunciados; es decir, la traducción generada de las palabras era mejor con respecto a otros traductores, gracias a la posición donde éstas se encontraban en los enunciados.

El módulo que realiza el procesamiento correspondiente a la energía textual fue implementado con éxito en el lenguaje de programación *PERL*. La obtención de la energía textual asociada a cada enunciado fue la medida que ponderó el ordenamiento de los enunciados como los más importantes y por tanto, aquellos que conformarían el resumen resultante.

En general el desempeño del sistema fue mejor que el esperado al momento de realizar la evaluación y compararla con los resultados de los resúmenes RA,los cuales fueron generados tomando los primeros enunciados de los documentos de entrada y con los resúmenes de referencia generados por FRESA.

Comparando las medidas de divergencia, que fue el método de evaluación empleado para medir el desempeño del sistema, se observó que el sistema basado en ENERTEX generó una evaluación aún mejor que los resúmenes generados por la herramienta de evaluación FRESA.

Del total de documentos, se observó un bajo desempeño del sistema para aquellos que eran de pequeñas dimensiones(menores a 150 palabras de longitud), además de los resúmenes generados al 15% del contenido del documento original.Los resultados del sistema para los resúmenes generados al 10% y al 20% fueron, en notable diferencia, mejores que los resúmenes generados por FRESA y los resúmenes RA. Cabe aclarar que la longitud de los documentos era mayor o igual a 200 palabras.

A manera de hallazgo en esta investigación, se tiene que el enfoque de resumen automático basado en energía textual resultó ser eficiente debido a que respecto a las evaluaciones realizadas con la herramienta de evaluación automática FRESA, el sistema propuesto fue el que obtuvo la mejor evaluación. Se observó que el sistema es capaz de generar mejores resúmenes cuando: a) los documentos fuente no son tan cortos (menores a 150 palabras) y b) cuando el resumen deseado se genera al 15% o al 20% del contenido original de los documentos.

Lo anterior permite concluir que el enfoque de energía textual no presenta un desempeño eficiente para textos pequeños. Al calcular la energía textual en los enunciados que conforman un documento corto, la ponderación realizada de los mismos no es suficiente para que se entreguen los enunciados más relevantes del documento fuente.Por otra parte, al generar resúmenes del 15% o el 20% y con documentos no tan pequeños, el sistema propuesto genera resúmenes de calidad.

De manera general, el resumir documentos en un idioma distinto al del documento fuente enfrentará dificultades con la traducción automática. A pesar de los obstáculos considerables, la investigación y el trabajo realizado sobre el tema es bastante extenso y complejo.

La implementación del sistema traductor automático presentó varias desventajas al haber contemplado en primera instancia al sistema propuesto por Google. Cuando su API fue cerrada debido al abuso ensu implementación.Google detectó que tanto robots, como equipos particulares estaban destinados a ofrecer el servicio de traducción implementando su traductor evadiendo cualquier pago por uso de la licencia del mismo.

Para obtener mejores resultados en la tarea de resumen automático multi-idioma se pueden combinar técnicas para mejorar la traducción automática y así formar mejores resúmenes en diferentes idiomas.

Como trabajo futuro se propone realizar un análisis más detallado del desempeño del sistema con documentos de longitud corta, así como realizar pruebas con los mismos implementando distintos enfoques de resumen automático, como lo son: el algoritmo basado en grafos o la compresión de frases.

Además de evaluar los resultados del sistema con ROUGE empleando documentos que tengan resúmenes de referencia, también utilizar un traductor automático abierto como MOSES, APERTIUM, TAUS, entre otros, para desarrollar un sistema completo de traducción y resumen automáticos.

REFERENCIAS

Blatz, J. Fitzgerald, E. Foster, G. Gandrabur, S. Goutte, C. Kulesza, A. Sanchis, A. y Ueffing, N.2009"Confidence Estimation for Machine Translation". *Reporte Técnico de la Universidad Johns Hopkins*. Baltimore, Estados Unidos.

Boudin, F., Huet, S. y Torres-Moreno, J. 2011 "A Graph-based Approach to Cross-language Multi-document Summarization".*Polibitis* 43:113-118.

daCunha, I., Torres-Moreno, J. M., Velázquez-Morales, P. y Vivaldi, J. 2009 "Un algoritmo lingüístico-estadístico para resumen automático de textos especializados". *LinguaMÁTICA*,2:67-80.

daCunha, I. y Wanner, L. 2006 "Resumen automático de artículos médicos en castellano: integración de técnicas de análisis textual, léxico, discursivo y sintáctico.comunicativo" Congreso de Lingüística General. Barcelona

Edmundson, H. P. 1969 "New Methods in Automatic Extracting"*Journal of the ACM*,Volumen 16

Evans, D.,Klavans, J. y Mckeown, K. 2004 "Columbia Newsblaster: Multilingual News Summarization on the Web"*Association for Computational Linguistics*, 1-4.

Fernández, S., SanJuan, E., y Torres-Moreno, J. M. 2007 "Energie textuelle de m´emoiresassociatives"*Traitement Automatique des Langues Naturelles*, 25–34. Toulouse, Francia.

Gutiérrez, M. X. 2010*Sistema de resumen extractivo automático*, Facultad de Ingeniería, UNAM, ciudad de México, tesis de licenciatura.

Hertz, J., Krogh, A., and Palmer G. 1991 "Introductiontothetheory of Neural Computation" *apud* da Cunha, I., Torres-Moreno, J., Velázquez-Morales, P. y Vivaldi, J. 2009.

Hopfield, J. 1982 "Neural networks and physical systems with emergent collective computational abilities" *National Academy of Sciences*, 9:2554–2558 *apud* da Cunha, I., Torres-Moreno, J., Velázquez-Morales, P. y Vivaldi, J. 2009.

Kupiec, J., Pedersen, J., CHENF. 1995 "A trainable document summarizer" In *Proceedings of the 18th ACM Special Interest Group on Information Retrieval* (SIGIR), ACM Press, 68–73. Nueva York.

Luhn, H. P. 1959 "The automatic creation of Literature abstracts". *IBM Journal of research and development*, 2(2).

Mani, I. y M.T. Maybury.1999*Advances in automatic text summarization*. MA: MIT Press. Cambridge.

Mani, I. 2001*Automatic summarization*.John Benjamins Publishing. Amsterdam.

Mani, I. 2001 "Summarization evaluation: An overview". North American Chapter of the Association for Computational Linguistics – NAACL 2001.

Manning, C.D., Raghavan, P. y Schütze, H., 2008 *Introduction to Information Retrieval*. Cambridge University Press.

Manning, C.D. y Schütze, H. *Foundations of Statistical Natural Language Processing* (Stanford University and Xerox Palo Alto Research Center) Cambridge, MA: The MIT Press

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

Medina, A.2011 "De la palabra gráfica al texto: sobre la extracción de enunciados para el resumen automático". En: VÁZQUEZ LASLOP, María Eugenia, Klaus ZIMMERMANN y Francisco SEGOVIA, eds., *De la lengua por sólo la extrañeza. Volumen 2. Estudios de lexicología, norma lingüística, historia y literatura en homenaje a Luis Fernando Lara* (ISBN 978-607-462-319-2), El Colegio de México, México.

Méndez, C. F. y Medina, A. 2005"Extractive Summarization Based on Word Information and Sentence Position", *Lecture Notes in Computer Science*, 3406, pp. 653-656. En: GELBUKH, Alexander, ed., *Computational Linguistics and Intelligent Text Processing*, Springer, Berlín.

Ogden, W., Cowie, J., Davis, M., Ludovik, E., Molina-Salgado, H. y Shin, H. "Getting Information from Documents You Cannot Read: An Interactive Cross-Language Text Retrieval and Summarization System".

Pollock, J. y Zamora, A. 1975 "Automatic abstracting research at the chemical abstracts service". Journal of Chemical Information and Computer Sciences, 226–232.

Raybaud, S. Langlois, D. y Smaïli, K. 2009 "Efficient Combination of Confidence Measures for Machine Translation". *Interspeech 2009*. Brighton, ReinoUnido.

Saggion, H. 2008 "Automatic Summarization: An Overview". *Revue française de linguistique appliquée,*Volumen 13. Francia

Smith, Christian 2011 *Automatic Summarization and Readability*. Tesis de Maestría. Universidad Linköpings

Sparck-Jones y Galliers 1996 *Evaluating Natural Language Processing Systems: An analysis and Review*, en, http://acl.ldc.upenn.edu/J/J98/J98-2013.pdf (fecha de consultajulio 2012).

Sunnetha, M. y Sameen, S. 2011 "Corpus based Automatic Text Summarization System with HMM Tagger" International Journal of Soft Computing and Engineering 2011, 118–123.

Torres-Moreno, J. M., Velázquez-Morales, P. y Meunier. J.G. 2001 "Cortex : un algorithme pour la condensation automatique des textes". *Proceedings of ARCo 2001*, 65–75.

Torres-Moreno, Juan-Manuel Boudin, Florian y Huet, Stéphane. 2011 "A Graph-based Approach to Cross-language Multi-document Summarization".*CICLing 2011*, Tokio.

Xie, S. 2010 *Automatic extractive summarization on meeting corpus.*Tesis doctoral Universidad de Texas en Dallas Richardson, en, http://www.hlt.utdallas.edu/~shasha/dissertation/dissertation.pdf (fecha de consulta agosto 2012)

Yang, J., Cohen, A. and Hersh, W. 2007 "Automatic Summarization of Mouse Gene Information by Clustering and Sentence Extraction".*MEDLINE Abstracts*.AMIA AnnuSympProc. 2007; 831–835.

Páginas consultadas

http://www.systransoft.com/systran/corporate-profile/translation-technology/what-is-machine-translation

http://acl.ldc.upenn.edu/J/J98/J98-2013.pdf

http://www.seas.upenn.edu/~zives/03s/cis650/ir.pdf

http://nlp.stanford.edu/IR-book/html/htmledition/irbook.html

ANEXOS Y APÉNDICES

Enseguida se muestran las tablas de resultados de divergencia para cada documento. La primera tabla muestra los resultados de la medida de divergencia Jensen-Shannon.

Doc	Resumen 10% para ENERTEX	Resumen 10% para RRA	Resumen 15% para ENERTEX	Resumen 15% para RRA	Resumen 20% para ENERTEX	Resumen 20% para RRA
1	2.59422	2.44	2.59422	2.44	2.29967	2.44
2	2.30064	2.28458	2.09405	2.78776	2.09405	2.28458
4	2.33517	2.4129	2.37228	2.1071	2.37228	2.4129
5	2.38968	2.46236	2.57715	2.46236	2.57715	2.46236
6	2.26529	2.4466	2.02019	2.33118	2.02019	2.4466
7	1.72643	2.27055	2.20994	2.27055	2.08431	2.27055
8	2.17791	2.17718	1.95649	2.17718	2.05092	2.17718
9	2.4295	2.24691	1.99622	2.11628	1.99622	2.24691
11	2.37768	2.65651	2.4453	2.65651	2.63376	2.65651
12	2.35942	2.39446	1.85249	2.39446	1.85249	2.39446
13	1.92668	2.3575	1.92668	1.65296	1.60662	2.3575
14	3.80309	2.90013	3.03163	2.90013	2.77563	2.90013
15	2.26605	2.44034	1.96412	2.44034	1.92381	2.44034
16	2.2411	2.13527	1.79008	2.13527	1.79008	2.13527
17	2.55827	2.4398	2.27525	2.1321	1.92749	2.4398
18	2.5416	2.89603	2.5416	2.88405	2.5416	2.89603
19	1.97768	2.34245	1.97768	1.91627	1.97768	2.34245
20	2.40739	2.48236	1.88684	2.48236	1.88684	2.48236
21	2.2169	2.98983	2.2169	2.39503	1.99704	2.98983

22	2.89369	3.35738	2.95256	3.35738	2.88875	3.35738
23	2.72067	2.22821	2.58413	2.22821	2.16925	2.22821
24	2.17204	2.16651	1.89591	2.16651	1.89591	2.16651
25	2.3389	2.84209	2.46704	2.30975	2.46704	2.84209

La siguiente tabla muestra los resultados de la medida de divergencia Kullback-Leibler.

Doc	Resumen 10% para ENERTEX	Resumen 10% para RRA	Resumen 15% para ENERTEX	Resumen 15% para RRA	Resumen 20% para ENERTEX	Resumen 20% para RRA
1	12.11041	11.72607	12.11041	11.72607	11.05547	11.72607
2	9.18188	10.204	9.18188	11.90438	9.18188	10.204
4	11.64875	12.41901	12.31363	11.0308	12.31363	12.41901
5	11.48663	11.59062	12.03997	11.59062	12.03997	11.59062
6	10.9074	11.73288	9.98791	11.38087	9.98791	11.73288
7	8.73392	10.51794	10.37572	10.51794	9.72557	10.51794
8	10.08263	9.87626	9.08467	9.87626	9.51457	9.87626
9	10.7402	10.69418	9.59936	10.23938	9.59936	10.69418
11	10.45341	11.58492	10.57119	11.58492	11.57922	11.58492
12	11.05863	10.56276	8.79145	10.56276	8.79145	10.56276
13	9.44343	11.25872	9.44343	8.29014	7.95325	11.25872
14	19.04764	15.60014	14.29618	15.60014	13.75566	15.60014
15	11.25238	11.9332	9.59471	11.9332	9.38531	11.9332
16	10.83928	10.03695	9.00712	10.03695	9.00712	10.03695
17	10.5132	9.87382	9.43814	9.127	8.40444	9.87382
18	12.60104	13.77835	12.60104	13.66321	12.60104	13.77835
19	9.32492	10.50799	9.32492	8.92352	9.32492	10.50799
20	10.6347	10.29195	8.7782	10.29195	8.7782	10.29195
21	10.59469	12.25609	10.59469	10.57556	9.58702	12.25609
22	17.16058	16.34028	14.99348	16.34028	14.72897	16.34028

23	14.45945	12.12957	13.47028	12.12957	11.86829	12.12957
24	10.00446	9.95681	8.80023	9.95681	8.80023	9.95681
25	11.37704	12.9733	11.74211	10.79881	11.74211	12.9733