Universidad Politécnica Metropolitana de Puebla

Informe final del proyecto

Desambiguación del Sentido de la Palabra Aplicada en Recuperación de Información

Financiado por:

PRODEP

Informe Presentado por:

Dr. Franco Rojas López

Puebla, Puebla, México. 21 de febrero de 2018

Agradecimientos

Estoy especialmente agradecido con el Programa de Mejoramiento del Profesorado por el apoyo ecónomico con número 170857, sin el cual no hubiera sido posible la finalización del proyecto de investigación "Desambiguación del Sentido de la Palabra Aplicada en Recuperación de Información". Gracias por contribuir en el desarrollo científico y tecnológico en favor de la investigación y la docencia.

Índice

Índice de figuras

Índice de tablas

Resumen

Abordar el problema de la ambigüedad de consultas web enviadas a un motor de búsqueda, es un problema desafiante en Recuperación de Información, debido a la ambigüedad intrínsica del lenguaje natural y a la longitud de las consultas enviadas por los usuarios. En el presente proyecto de investigación se abordó el problema antes mencionado, el cual esta dividido en 3 fases. La primera fase consiste en extraer términos y frases clave a partir de un conjunto de entrenamiento. La segunda fase consiste en recuperar documentos web relacionados con el conjunto de entrenamiento y la tercera fase en la implementación del Sistema de Recuperación de Información y su validación. El informe presenta el trabajo realizado en la primera y segunda fase, cabe mencionar que los experimentos fueron realizados en conjuntos de datos estándar usados por otros investigadores. Los resultados obtenidos son competentes con otros enfoques reportados en la literatura.

1. Introducción

Existen billones de páginas de información en la Web[1], y encontrar información confiable y relevante a nuestro interés puede ser un desafío. Para apoyarnos en esta tarea surge el concepto de motor de búsqueda, el cual es una herramienta poderosa que colecciona y organiza contenido desde todo el Internet indexando millones de sitios web.
Buscar información en la Web, es una tarea diaria e indispensable para que la mayoría de las personas puedan realizar sus actividades cotidianas tal como la búsqueda de información, de un producto o un servicio. En un motor de búsqueda el usuario ingresa una consulta compuesta por una o más palabras, llamadas palabras clave a continuación el motor de búsqueda devuelve una lista de enlaces a páginas web que contienen las palabras ingresadas en la consulta.

De acuerdo con Salton [3], recuperación de información es el campo relacionado con la estructura, análisis, organización, almacenamiento, búsqueda y recuperación de información.

En el contexto de Recuperación de Información (RI) es claro que una cantidad inmensa de consultas enviadas a un motor de búsqueda son inherentemente ambiguas. Entender el significado de las consultas enviadas a un motor de búsqueda z n así como la relevancia de páginas web, es un problema desafiante y por lo tanto

[1] `http://www.worldwidewebsize.com/`

difícil de resolver. Un gran problema que presentan los motores de búsqueda actuales, es que las consultas de búsquedas son usualmente cortas, ambiguas y por lo tanto insuficientes para especificar las necesidades precisas de los usuarios. De acuerdo con Enge *et al.* [2] las consultas de longitud 1, es decir formadas por una palabra superan en uso a las consultas de mayor longitud. En la Tabla 1 se muestra la longitud de las consultas (columna 1) y el porcentaje de su uso en la Web (Columna 2). Como se puede observar en la Tabla 1 las consultas formadas por una palabra representan el 25.8 % del total de consultas en la web. Por ejemplo, si el usuario

Tabla 1: *Búsquedas por longitud de consultas, datos extraídos desde [2]*

Palabras	Porcentajes de búsquedas
1	25.8
2	22.8
3	18.7
4	13.2
5+	19.5

ingresa la consulta "panda", el motor de búsqueda Google muestra en primer lugar resultados relacionados con *panda antivirus 2017* y *panda como mamífero*. Cuando es muy probable que el usuario este interesado en panda como grupo musical. Aunado a lo anterior posiblemente el usuario desea que la información consultada le sea mostrada en inglés a pesar de realizar la consulta en un país de habla hispana. Para solventar el problema antes mencionado los motores de búsqueda han implementado nuevas técnicas basadas en búsquedas semánticas. La semántica es el estudio del significado que entiende la relación entre frases, símbolos, signos, antónimos, sinónimos, etc. Los motores de búsqueda modernos tal como: Google, Yahoo y Bing hacen uso de la semántica especialmente para consultas ambiguas.

De acuerdo con lo anteriormente expuesto en el presente proyecto abordamos el problema de ambigüedad de las palabras y su aplicación en RI. En el proyecto se usaron datasets estándar, los cuales han sido empleados por otros investigadores en el área de RI. Los experimentos llevados a cabo demuestran que el enfoque propuesto implementada es competitiva con lo reportado en la literatura. En este contexto el proyecto contribuye con una técnica para la extracción de términos clave desde textos cortos, los términos clave pueden fungir como un conjunto de consultas enviadas a un motor de búsqueda. Aunado a lo anterior también se contribuye con una técnica para obtener el grado de similitud semántica entre un par de textos cortos, la técnica esta basada en Hipótesis Distribucional (HD). HD asume que palabras similares ocurren en contextos similares, esto permite recuperar contextos en los que ocurre una palabra para solventar el problema de la ambigüedad. Finalmente se implementó

un prototipo del Sistema de Recuperación de Información (SRI), en el lenguaje de programación Java[2].

El presente documento está organizado como sigue: en la sección 2 se presenta el estado del arte relacionado con el proyecto de investigación. En la sección **??** se describe la metodología propuesta para abordar el problema de la ambigüedad de las consultas web. La sección 4 muestra la posible publicación en revista del trabajo realizado. Finalmente las conclusiones y trabajo futuro se presentan en la sección 5.

2. Estado del arte

En esta sección se describe la arquitectura general de los motores de búsqueda, así como trabajos relacionados que presentan avances en el tratamiento de consultas web.

2.1. Arquitectura de un motor de búsqueda

La búsqueda de información en la Web es una actividad diaria para la mayoría de las personas, lo cual convierte a una computadora en uno de los usos más populares en nuestros días. Dada la importancia de búsqueda de información los motores de búsqueda intentan mejorar este proceso haciendolo más fácil y rápido para encontrar la información correcta, este proceso es conocido como *Recuperación de Información*. Desde sus inicios en 1950 el objetivo de RI ha sido el texto y documentos de texto, por ejemplo páginas web, correo electrónico, libros digitales, etc., esto se conoce como información no estructurada. En el proceso de RI debemos diseñar algoritmos que nos permitan comparar el texto en las consultas con el texto de cada uno de los documentos y decidir que documento contiene la información buscada, esta no es una tarea fácil en si misma, es necesario entender y modelar como las personas comparan textos, y diseñar algoritmos computacionales para llevar a cabo esta comparación, este es el núcleo de RI. Actualmente el objetivo no es solo recuperar texto sino también imagenes, videos, audio incluyendo musica, este tipo de información hace más complicado el proceso de RI.
En la Figura 1 Arasu *et al.* [1] muestra la arquitectura general de un motor de búsqueda. Motores de búsqueda como Google, Bing, Yahoo deben ser capaces de almacenar muchos terabytes de datos y mostrar una respuesta a millones de consultas

[2] `http://www.oracle.com/technetwork/es/java/javase/downloads/index.html`

enviadas cada día alrededor del mundo. De acuerdo con Arasu los dos principales objetivos de los motores de búsqueda son:

- Efectividad (calidad): se desea recuperar el conjunto de documentos más relevantes para nuestra consulta.
- Eficiencia (velocidad): se desea procesar consultas de los usuarios tan rápido como sea posible.

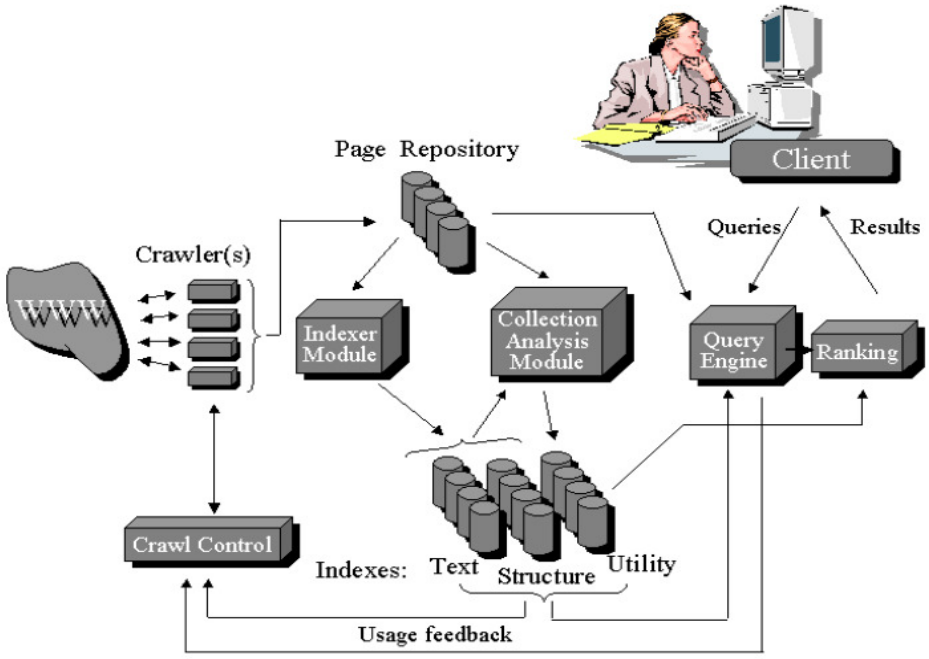

Figura 1: *Arquitectura general de un motor de búsqueda. Imagen extraída desde [1]*

De acuerdo con Arusu [1] se dará una breve explicación de los módulos que forman parte de la arquitectura general de un motor de búsqueda (ver Figura 1).

- **Crawlers**. Los crawlers tienen la responsabilidad de identificar y adquirir documentos para el motor de búsqueda, es decir un crawler es diseñado para seguir los enlaces de las páginas web para descubrir y descargar nuevas páginas. Esta información es transferida al módulo control del crawler (*Crawl Control*).
- **Módulo de indización** (*Indexer module*). Este módulo extrae todas las palabras de cada una de las páginas y registra la URL donde cada palabra ocurrió.

- **El índice de utilidad** (*utility index*) es creado por el módulo análisis de colección (*Collection analysis module*), el cual es responsable de crear una colección de otros índices. El índice de utilidad puede proporcionar acceso a páginas de una longitud dada, páginas de cierta importancia, o páginas con un cierto número de imágenes en ellas.
- **Repositorio de páginas** (*Page Repository*). Durante la ejecución del crawler e indización los motores de búsqueda deben almacenar las páginas que ellos recuperan desde la Web. El repositorio de páginas representa esta colección posiblemente temporal. Algunas veces los motores de búsqueda almacenan las páginas web que han sido visitadas en la cache por cierto tiempo para construir el índice.
- **Motor de búsqueda** (*Query Engine*). El módulo motor de búsqueda es responsable de recibir y enviar las solicitudes de búsqueda de los usuarios. El motor de búsqueda depende del índice creado y algunas veces del repositorio de páginas web.
- **Módulo de ranking** (*Ranking Module*). Este módulo tiene la tarea de ordenar los resultados de tal manera que los más revelantes para la consulta se encuentren en la parte superior de la lista retornada.

En esta sección se ha presentado una descripción de los módulos que forman parte de la arquitectura general de un motor de búsqueda web. En el presente proyecto se abordarán los problemas que se presentan en los módulos: motor de búsqueda y módulo de ranking.

2.2. Trabajo relacionado

En esta sección se presentan trabajos relevantes, relacionados con el proyecto de investigación propuesto.

2.2.1. Extracción de términos clave

La extracción de palabras o frases clave sigue siendo una tarea intermedia imprescindible para el éxito de otras aplicaciones en el Procesamiento del Lenguaje Natural (PLN). Por ejemplo, las palabras clave han sido empleadas para mostrar anuncios en una página web de acuerdo con su contenido [4, 5, 6]. En este contexto

Yih *et al.* [7] propusieron un método para extraer palabras clave desde una página web y a continuación mostrar anuncios en esa página de acuerdo con las palabras clave recuperadas. El método propuesto consiste en extraer características relevantes usando TF-IDF, metadatos desde la página web e información recuperada desde archivos log de consultas web. Después de llevar a cabo una fase de pre-proceso de la información recuperada, ellos consideran una frase de longitud máxima 5 como término clave. A continuación, implementan un clasificador entrenado usando aprendizaje automático para predecir si la frase es un término clave o no. Chen *et al.* [6] intentaron refinar la extracción de palabras clave desde páginas web. El método propuesto recomienda palabras clave con la ayuda directa de una jerarquía de conceptos la cual usa información semántica para mejorar la precisión y cobertura de las palabras clave sugeridas. El método recomienda nuevas palabras clave de acuerdo con los objetivos de los anunciantes y no con las palabras clave de las consultas.

2.2.2. Similitud semántica entre oraciones de texto

Medir la similitud semántica entre oraciones o conceptos es una tarea fundamental en varias aplicaciones del PLN, por ejemplo, en sistemas de recomendación para filtrar información y guiar a los usuarios para descubrir productos o servicios en una forma personalizada [8, 9]. Para ofrecer explicaciones verbales a partir de un par de oraciones [10] que puede ser aplicado en sistemas de tutorado inteligente [11], en recuperación de información para medir la similitud entre la consulta y textos almacenados en una colección de documentos, entre otras aplicaciones.
La similitud semántica entre textos, oraciones o palabras tiene como objetivo capturar cuando el sentido de los textos, oraciones o palabras es similar. Esta es una tarea que ha llamado la atención en los últimos años dada su importancia en varias tareas del PLN, por ejemplo en traducción autómatica, extracción de resúmenes, recuperación de información, sistemas de recomendación, entre otras.
Otra aplicación de la similitud semántica es determinar si una página web es plagio de otra, es decir; una página web puede ser espejo de otra que tiene casi el mismo contenido pero diferente información.
Dada la importancia de la tarea en la literatura se ha propuesto un gran número de técnicas para encontrar la similitud semántica entre fragmentos de textos y conceptos, las cuales se dividen básicamente en similitud semántica basada en corpus y similitud semántica basada en conocimiento.

- **Similitud semántica basada en corpus**. Los enfoques basados en corpus implementan métodos distribucionales para representar los significados conceptuales en vectores como, por ejemplo, Análisis Semántico Latente [12]. En

este contexto, Mikolov *et al.* [13] desarrollaron un modelo para la representación distribuida de palabras y frases y su composición. El objetivo del modelo es encontrar representaciones de palabras que son útiles para predecir palabras en el contexto de una sentencia o documento. Por otro lado, en el trabajo presentado por Ishizuka *et al.* [14] estimaron la similitud semántica entre pares de palabras usando motores de búsqueda. El método considera los hits retornados y patrones léxicos sintácticos extraídos desde los fragmentos de texto recuperados por el motor de búsqueda dada una consulta web. Usando una representación de las fuentes de información antes mencionadas implementaron el algoritmo Máquinas de Vectores Soporte (MVS) para encontrar pares de palabras sinónimos y antónimos. La función de similitud entre dos palabras P y Q es aproximada por un score de confianza del algoritmo MVS entrenado.

- **Similitud semántica basada en grafos de conocimiento**. Zhu e iglesias [15], propusieron un método para calcular la similitud semántica entre conceptos basado en grafos de conocimiento tal como WordNet y BDpedia. El método nombrado wpath, combina el Contenido de Información (CI) de conceptos para asignar un peso a la longitud de la ruta más corta entre ambos conceptos. La idea de usar CI para calcular la similitud semántica es que mientras más información comparte dos conceptos, más similares son. Majid y Alireza [16] implementaron un método no supervisado basado en el conocimiento para medir la similitud semántica de los textos en los que se tienen en cuenta las similitudes específicas de palabra a palabra. El método esta implementado en un grafo bipartito sin pesos y no dirigido. Para un par de segmentos de texto, se producen conjuntos de palabras de clase abierta, con un conjunto distinto creado para sustantivos, verbos, adjetivos y adverbios. A continuación, determinan la similitud de los pares de palabras en los conjuntos correspondientes en la misma clase abierta en los segmentos de texto. Para sustantivos y verbos usan una medida de similitud semántica basada en WordNet. Mientras que para las otras clases de palabras usan un emparejamiento léxico.

En el proyecto se presenta un enfoque basado en corpus el cual tiene la ventaja de tener una mayor cobertura de vocabulario porque, el modelo computacional puede ser efectivamente aplicado en un corpus actualizado o enriquecido.

3. Enfoque propuesto

En esta sección se describe la metodología propuesta para abordar el problema de desambiguación del sentido de la palabra aplicada en Recuperación de Información.

La Figura 2 muestra la metodología propuesta dividida en tres fases. La primera fase consiste en la extracción de términos y frases clave, así como la recuperación de documentos web relacionados con el conjunto de entrenamiento. La segunda fase consiste en el procesamiento de la información, extracción de contextos de acuerdo al conjunto de entrenamiento así como la representación de la información. Finalmente la tercera fase consiste en implementar el Sistema de Recuperación de Información para su posterior validación.

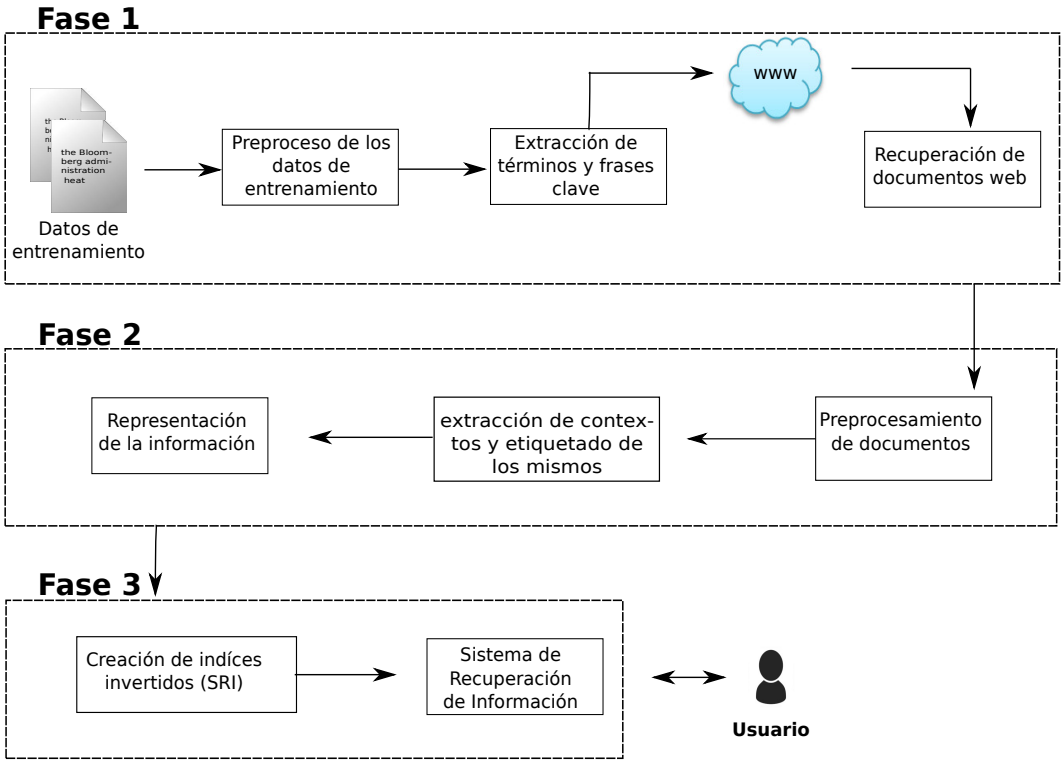

Figura 2: *Metodología propuesta*

A continuación se describen las actividades desarrolladas por el alumno de acuerdo con la metodología propuesta.

3.1. Datos de entrenamiento

Antes de investigar los conjuntos de entrenamiento para la evaluación de los algoritmos propuestos en cada fase de la metodología, el alumno estudió la arquitectura general de los motores de búsqueda para comprender en que módulo se trabajaría con el proyecto, extracción de términos y frases clave, similitud semántica entre pa-

res de oraciones así como algoritmos de clustering. Posteriormente se investigó y se comprendió la estructura de los conjuntos de entrenamiento disponibles en la web para la extracción de términos y frases clave. Los conjuntos de entrenamiento se describen a continuación.

- Datasets usados en la extracción de términos clave
  - **Hulth 2003**. El dataset (conjunto de textos cortos) empleados en los experimentos fue usado por Hulth[3] [17]. El dataset contiene 2000 resúmenes de artículos científicos en el idioma inglés. Los resúmenes están divididos en tres conjuntos: un conjunto de entrenamiento que contiene 1000 resúmenes, un conjunto de validación el cual contiene 500 resúmenes y un conjunto de prueba que contiene los restantes 500 resúmenes.
- Dataset usado en la similitud semántica de sentencias
  - **Microsoft Research Paraphrase**. El corpus usado en los experimentos fue liberado por Microsoft Research Paraphrase, el archivo de texto contiene 4,077 pares de sentencias extraídas de fuentes de noticias en la Web con anotaciones humanas que indican si cada par de oraciones posee una relación de equivalencia semántica [18].
- Dataset usado en el SRI
  - **LISA**[4]. Es un dataset proporcionado por Peter Willett de la universidad de Sheffield, Inglaterra. Contiene 6,004 documentos y 35 consultas con un juicio de relevancia.

3.2. Preproceso de los datos de entrenamiento

Para implementar este módulo se desarrolló un programa en el lenguaje de programación Java[5] para extraer las oraciones desde el conjunto de entrenamiento. El programa implementado hace uso de expresiones regulares para eliminar símbolos desde el conjunto de datos. Cabe mencionar que también fueron removidos los artículos, preprosiciones, conjunciones, etc.

[3] `https://github.com/snkim/AutomaticKeyphraseExtraction`

[4] `http://ir.dcs.gla.ac.uk/resources/test_collections/`

[5] `http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html`

3.3. Identificando Palabras Clave en Resúmenes de Texto Aplicando Chi-cuadrado

Una vez realizado el preproceso de la fase anterior se propuso implementar una variante del estadístico chi-cuadrado (Ecuación 1) para la extracción de términos clave.

$$x^2 = \frac{(O - E)^2}{E} \tag{1}$$

Donde O denota la frecuencia observada y E denota la frecuencia esperada entre los términos w_i y $t_k = w_i + 1$. Sea $S = \{w_1, w_2, w_3, \cdots, w_n\}$ una sentencia en el documento de texto, la extracción de n-gramas clave (bigramas y trigramas) es como sigue.

Para obtener la frecuencia esperada de bigramas usamos la Ecuación 2 donde $f(w_i, t_k)$ es la frecuencia de aparición del bigrama w_i y t_k en la sentencia S. La probabilidad de ocurrencia de los términos w_i y t_k se obtiene dividiendo su valor de frecuencia entre el número de ocurrencia ($\sum$ t) de todos los pares de bigramas encontrados en el documento.

$$E = \frac{\sum f(w_i, t_k)}{\sum t} \tag{2}$$

La frecuencia observada de los términos w_i y t_k se obtiene multiplicando los factores a y b de la Ecuación 3 $f(w_i, t_k)$ es la frecuencia del bigrama w_i y t_k en la sentencia S y $f(w_i)$, $f(t_k)$ es la frecuencia del término en el documento de texto.

$$a = \frac{\sum f(w_i, t_k)}{\sum f(w_i)}, b = \frac{\sum f(w_i, t_k)}{\sum f(t_k)} \tag{3}$$

Para obtener trigramas clave se sigue el procedimiento descrito anteriormente con la diferencia que $w_i = (w_i, w_{i+1})$ y $t_k = wi + 2$.

3.3.1. Experimentos y resultados

El dataset (conjunto de textos cortos) empleados en los experimentos fue usado por [17] (ver Sección 3.1). El dataset incluye dos conjuntos de sintagmas nominales clave, uno controlado y no controlado. En los experimentos se usó el conjunto de prueba con sintagmas no controlados indexado por profesionales.

Antes de realizar el experimento se aplicó el siguiente pre-proceso. Se usó la biblioteca Apache OpenNLP para segmentar el texto en sentencias. Después eliminamos el conjunto de palabras cerradas tal como: artículos, pre-posiciones, conjunciones, etc. Es importante mencionar que no usamos un algoritmo de stemm para reducir las palabras a su raíz esto es así porque un stemming agresivo puede causar problemas

de búsqueda [19].
Después de la fase de pre-proceso aplicamos el estadístico chi-cuadrado descrito anteriormente. La columna 1 de la Tabla 2 muestra algunos trigramas recuperados por el enfoque propuesto. La columna 2 muestra las palabras únicas recuperadas a partir de los trigramas de la columna 1. Cabe mencionar que recuperamos los primeros 10 trigramas con mayor valor chi-cuadrado.

Tabla 2: *Palabras clave recuperadas por el sistema*

Trigramas	**Palabras clave**
Materials materials recurring	Materials
Materials materials practices	Recurring
Print materials print	Practices

La Tabla 3 muestra los resultados extraídos desde el conjunto de prueba.

Tabla 3: *Resultados extraídos desde el conjunto de prueba*

Sintagmas nominales
recurring issues
books
changing practices
out-of-print
library materials
acquisition

En la primera fase del proyecto se ha contribuído con un método para extraer términos clave a partir de textos cortos, los resultados de la investigación serán publicados en el Congreso Internacional de Investigación e Innovación 2018[6].

[6] `http://www.congresoucec.com.mx/`

3.4. Midiendo la Similitud Semántica en Textos Cortos usando el Contexto Relacionado y DISCO

El algoritmo de extracción de términos clave implementado en la fase 1 es usado en la fase 2 (ver Figura 2), donde el objetivo es medir la simitud semántica entre pares de oraciones. El conjunto de entrenamiento usado para medir la similitud semántica en textos cortos fue liberado por Microsoft Research ver la Sección 3.1

3.4.1. Recuperación de documentos web

El uso de la Web ha mostrado su utilidad en varias tareas del PLN, por ejemplo, en Desambiguación del Sentido de la Palabra. Motivados por la utilidad de la web en el PLN en el presente trabajo se sigue un enfoque similar al presentado por Lopez-Arevalo *et al.* [20].
En la propuesta se recuperaron las primeras 20 respuestas retornadas por el motor de búsqueda Google en formato HTML, docx, y pdf. Se usó la Api boilerpipe de Google para extraer el contenido principal de texto de una página web; para convertir a texto plano los archivos en formato pdf, docx, se usó la herramienta Tika de apache. Una vez extraído el texto plano se empleó la herramienta OpenNLP para dividir el texto en sentencias. A este conjunto de sentencias le denominamos conjunto de texto relacionado (CTR). Cabe mencionar que una sentencia forma parte del Conjunto de Texto Relacionado si contiene un verbo, sustantivo, adjetivo o adverbio del Conjunto de Entrenamiento (CE).

3.4.2. Preproceso de documentos web

Una vez obtenido el CTR el texto fue dividido en sentencias, se eliminaron las palabras cerradas tal como: artículos, preposiciones, conjunciones, símbolos de puntuación, etc., es decir; el conjunto de palabras que carece de significado. En el presente trabajo se usó toda la oración para recuperar el contexto en el cual aparece cada una de las palabras (verbos, sustantivos, adjetivos y adverbios) contenidas en el conjunto de entrenamiento.

3.4.3. Representación de la información

Es importante mencionar que para cada una de las palabras contenidas en las oraciones del CE ($wt_i \epsilon\ CE$) se recuperaron sus contextos relacionados desde el CTR obtenido desde la Web. Cada una de las palabras wt_i y su contexto relacionado es representada por una matriz (M) como se muestra en la Tabla 1. Donde Wt_i puede ser un verbo, sustantivo, adjetivo o adverbio del CE. Wj son las palabras extraídas del CTR.
Así cada palabra wt_i tiene asociado un vector de características $wt_i = w_0, w_1, \cdots, w_n$

Tabla 4: *Representación del contexto relacionado*

M	w_0	w_1	$\cdots$	w_n
wt_0	v_{im00}	v_{im01}	$\cdots$	v_{im0n}
wt_1	v_{im10}	v_{im11}	$\cdots$	v_{im1n}
wt_2	v_{im20}	v_{im21}	$\cdots$	v_{im2n}
.	$\cdots$	$\cdots$	$\cdots$	$\vdots$
wt_m	v_{imm0}	v_{imm1}	$\cdots$	v_{immn}

Cada elemento m_{ij} en la matriz M tiene una frecuencia asociada. El valor de frecuencia entre una palabra wt_i y w_j es transformado en un valor de correlación usando la función Información Mutua [21], implementando la Ecuación 4. La frecuencia es usada por la función para asignar valores más altos a contextos que son más indicativos del significado de una palabra.

$$PMI(wt_i, wt_j) = log_2 \frac{P(wt_i, w_j)}{P(wt_i)P(W_j)} \tag{4}$$

Donde $P(wt_i, w_j)$ es la probabilidad de que la palabra wt_i co-ocurra en el contexto w_j. Entre la probabilidad de observar wt_i y w_j independientemente. El peso asignado por la función Información Mutua es denotado por v_{imij} como se muestra en la matriz M de la Tabla 4.

3.5. Similitud semántica entre pares de oraciones

La similitud semántica, es un valor que refleja el grado de correlación entre dos conceptos u oraciones. El enfoque propuesto usa una técnica basada en corpus o asociaciones estadísticas para calcular la similitud semántica entre palabras. Tal

información es extraída desde la matriz de pesos M del conjunto de textos relacionados. Para calcular la similitud semántica entre cada par de palabras se usó la fórmula del Coseno de Salton Ecuación 5.

$$Sim(a_1, b_2) = \frac{\sum_{j=1}^{n} PMI(a_1, w_j)PMI(a_2, w_j)}{\sqrt{\sum_{j=1}^{n} PMI(a_1, w_j)^2}\sqrt{\sum_{j=1}^{n} PMI(a_2, w_j)^2}} \quad (5)$$

donde a_1, a_2 son cualquier par de palabras que pertenecen al CE.

3.5.1. Experimentos y resultados

Para llevar a cabo la fase de experimentación el sistema recibe como entrada un par de oraciones. Por ejemplo, sea $s_1 = \{p_1, p_2, \cdots, p_n\}$ y $s_2 = \{q_1, q_2, q_3, \cdots, q_n\}$ un par de oraciones extraídas desde el CE, el cálculo de similitud semántica entre un par de oraciones se lleva acabo implementando la Ecuación .

$$sim(s_1, s_2) = \frac{\sum_{i=1}^{l_{s1}} \sum_{j=1}^{l_{s2}} sim_{coseno}(p_i, q_j) + DISCO(p_i, q_j)}{\sum_{k=1}^{2} l_{sk}} \quad (6)$$

donde l_{s1} y l_{s2}, es la longitud de la sentencia 1 y la sentencia 2. l_{sk} es la suma de la longitud de la sentencia 1 y 2. En el enfoque propuesto dividimos el valor de similitud entre la suma de la longitud de las sentencias ($\sum_{k=1}^{2} l_{sk}$) para no favorecer la sentencia que contiene más palabras.
DISCO, es una aplicación Java que permite recuperar la similitud semántica entre palabras y frases. La similitud está basada en análisis estadístico de grandes colecciones de texto. En la Tabla se muestran los experimentos llevados a cabo. La primera fila muestra el resultado del algoritmo de Lesk (baseline) el cual mide el grado de traslape que existe entre un par de oraciones. Cabe mencionar que en las oraciones se removieron los stopwords y se obtuvo la raíz de cada palabra usando WordNet y JWNL. La segunda fila muestra el resultado del método propuesto por

Tabla 5: *Resultados*

No. Técnica	Técnica	Precisión
1	Lesk	56.6
2	Kenter [22]	78.1
3	Shrestha [23]	71.0
4	Enfoque propuesto	78.4

Kenter, basado en representaciones vectoriales de términos computados desde datos no etiquetados. En la tercera fila se muestra la precisión obtenida por el enfoque

propuesto por Shrestha, que consiste en un método basado en corpus no supervisado basado en el modelo de espacio vectorial.

En esta fase se ha contribuido con un método para medir el grado de correlación entre un par de oraciones, esto permitirá medir la similitud entre una consulta y los documentos almacenados en el índice invertido. Los resultados serán publicados en la revista ingeniantes "*indexada ante el Sistema Regional de Información en Línea para Revistas Científicas de América Latina, el Caribe, España y Portugal LATINDEX, con número de registro 22843, ISSN 2395-9452 , siendo reconocida como una revista arbitrada de calidad*".

3.6. Prototipo

En aplicaciones del mundo real, los usuarios demandan resultados rápidamente desde un motor de búsqueda para satisfacer sus necesidades de información. Para cubrir esa necesidad los motores de búsqueda de recuperación de información hoy en día se basan en una estructura de datos llamada índice invertido que dado un término proporciona acceso a una lista de documentos que contienen el término. Dada una consulta enviada por el usuario el SRI utiliza el índice invertido para asignar un score a los documentos que contienen los términos de la consulta con respecto a algún modelo de ranking.
La Figura 3 muestra las fases del índice invertido.

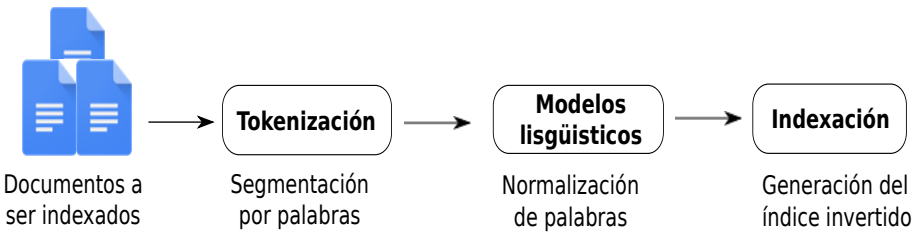

Figura 3: *Construcción del índice invertido*

Antes de construir el índice invertido los documentos a ser indexados (dataset LISA ver subsección 3.1) fueron pre-procesados, es decir; se eliminaron los stop-words o palabras cerradas, las cuales carecen de significado para la tarea de RI. A continuación se muestran dos documentos pre-procesados del dataset LISA.

1. **Documento 1.** xxv india library conference trivandrum 14 18 may 1979. papers proceedings conference papers relate indian library situation general libraries states libraries movement kerala.

2. **Documento 2.** lingering fragrance proceedings xxiv india library conference bangalore papers proceedings conference summary bagari transcripts inaugural speech introductory papers main presentations conference held 29 jan 1 feb 78

La Figura 4 muestra la construcción del índice invertido dado como entrada dos fragmentos de documentos extraídos desde el dataset LISA.

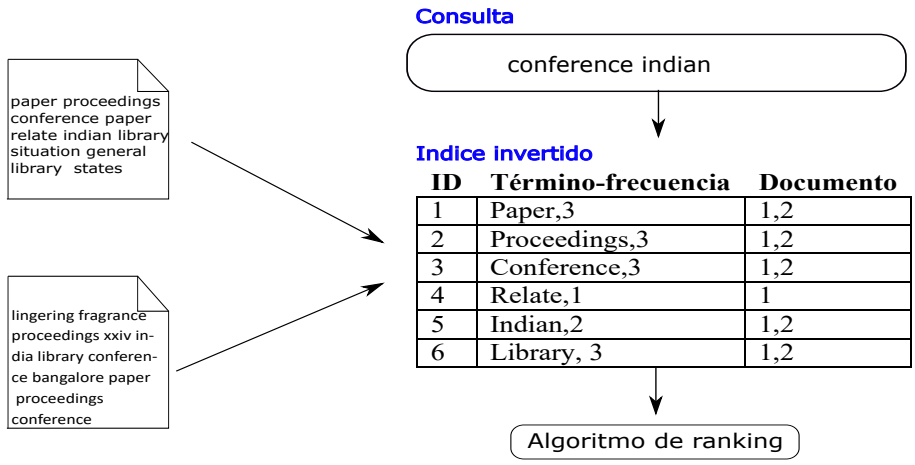

Figura 4: *Implementación del índice invertido*

4. Contribuciones

- **Extracción de términos clave**. Se ha contribuído en el estado del arte con una técnica para extraer términos clave desde textos cortos. La extracción de términos clave en textos representa un desafío dada la baja frecuencia de las palabras. La Figura 5 muestra la portada del artículo que se publicará en el "congreso internacional de investigación e innovación 2018".

  La Figura 6 muestra el dictamen que avala la aceptación del trabajo de investigación para presentarse y publicarse en el congreso internacional de investigación e innovación 2018, **ISSN 2448-6035**.

  La Figura 7 muestra la portada de la memoria en la que se publicará el artículo.

- **Similitus semántica entre un par de oraciones**. Aunado a lo anterior también se ha contribuído con una técnica para medir la similitud semántica entre un par de oraciones. La Figura

5. Conclusiones y trabajo futuro

En el presente informe se reportaron las actividades desarrolladas a partir de diciembre de 2016 al 24 de abril de 2017. En este periodo de tiempo se obtuvieron avances significativos del proyecto de investigación propuesto, el cual contribuye en el estado del arte con un algoritmo para la extracción de frases clave a partir de resúmenes cientifícos. Como trabajo futuro se reportará en el estado del arte un algoritmo para medir la similitud semántica entre pares de oraciones, este trabajo ha dado origen a la implementación de un sistema de recomendación de información en diferentes dominios basado en técnicas de recuperación de información.

Referencias

[1] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan. Searching the web. *ACM Trans. Internet Technol.*, 1(1):2–43, August 2001. ISSN 1533-5399. doi: 10.1145/383034.383035.

[2] Eric Enge, Stephan Spencer, Jessie Stricchiola, and Rand Fishkin. *The Art of SEO*. O'Reilly Media, Inc., 2nd edition, 2012.

[3] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November 1975. ISSN 0001-0782. doi: 10.1145/361219.361220. URL `http://doi.acm.org/10.1145/361219.361220`.

[4] Weinan Zhang, Dingquan Wang, Gui-Rong Xue, and Hongyuan Zha. Advertising keywords recommendation for short-text web pages using wikipedia. *ACM Trans. Intell. Syst. Technol.*, 3(2):36:1–36:25, February 2012.

[5] Stamatina Thomaidou and Michalis Vazirgiannis. Multiword keyword recommendation system for online advertising. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 423–427. IEEE, 2011.

[6] Yifan Chen, Gui-Rong Xue, and Yong Yu. Advertising keyword suggestion based on concept hierarchy. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 251–260, New York, NY, USA, 2008. ACM.

[7] Wen-tau Yih, Joshua Goodman, and Vitor R. Carvalho. Finding advertising keywords on web pages. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 213–222, New York, NY, USA, 2006. ACM.

[8] Kazem Qazanfari, Abdou Youssef, Kai Keane, and Joseph Nelson. A novel recommendation system to match college events and groups to students. *CoRR*, abs/1709.08226, 2017. URL `http://arxiv.org/abs/1709.08226`.

[9] Shuai Zhang, Lina Yao, and Aixin Sun. Deep learning based recommender system: A survey and new perspectives. *CoRR*, abs/1707.07435, 2017. URL `http://arxiv.org/abs/1707.07435`.

[10] Iñigo Lopez-Gazpio, Montse Maritxalar, Aitor Gonzalez-Agirre, German Rigau, Larraitz Uria, and Eneko Agirre. Interpretable semantic textual similarity: Finding and explaining differences between sentences. *CoRR*, abs/1612.04868, 2016. URL `http://arxiv.org/abs/1612.04868`.

[11] Rodney d. Nielsen, Wayne Ward, and James h. Martin. Recognizing entailment in intelligent tutoring systems*. *Nat. Lang. Eng.*, 15(4):479–501, October 2009. ISSN 1351-3249. doi: 10.1017/S135132490999012X. URL http://dx.doi.org/10.1017/S135132490999012X.

[12] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[14] D. Bollegala, Y. Matsuo, and M. Ishizuka. A web search engine-based approach to measure semantic similarity between words. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):977–990, July 2011. ISSN 1041-4347. doi: 10.1109/TKDE.2010.172.

[15] Ganggao Zhu and Carlos A Iglesias. Computing semantic similarity of concepts in knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):72–85, 2017.

[16] Majid Mohebbi and Alireza Talebpour. Texts semantic similarity detection based graph approach. *International Arab Journal of Information Technology (IAJIT)*, 13(2), 2016.

[17] Anette Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 216–223, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1119355.1119383.

[18] Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. doi: 10.3115/1220355.1220406. URL https://doi.org/10.3115/1220355.1220406.

[19] Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 1st edition, 2009. ISBN 0136072240, 9780136072249.

[20] Ivan Lopez-Arevalo, Victor J Sosa-Sosa, Franco Rojas-Lopez, and Edgar Tello-Leal. Improving selection of synsets from wordnet for domain-specific word sense disambiguation. *Computer Speech & Language*, 41:128–145, 2017.

[21] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[22] Tom Kenter and Maarten De Rijke. Short text similarity with word embeddings. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1411–1420. ACM, 2015.

[23] Prajol Shrestha. Corpus-based methods for short text similarity. In *Rencontre des Étudiants Chercheurs en Informatique pour le Traitement automatique des Langues*, volume 2, page 297, 2011.

"CONGRESO INTERNACIONAL DE INVESTIGACIÓN E INNOVACIÓN 2018"
Multidisciplinario
19 y 20 de abril de 2018, Cortazar, Guanajuato, México
ISSN 2448-6035

Identificando Palabras Clave en Resúmenes de Texto Aplicando Chi-cuadrado

Rosas Colula Luís Ángel, estudiante, angelrosassf@gmail.com, Universidad Politécnica Metropolitana de Puebla.

Franco Rojas López, Dr. En ciencias de la computación, frlb99@gmail.com, Universidad Politécnica Metropolitana de Puebla.

José Luis Hugo Díaz Biffano, Mtro. En administración, hugo.diaz@metropoli.edu.mx, Universidad Politécnica Metropolitana de Puebla.

Adriana Hernández Beristain, M.C. Ciencias de la computación, adrianah.beristain@correo.buap.mx, Benemérita Universidad Autónoma De Puebla.

Resumen: *La extracción de términos o frases clave sigue siendo una tarea desafiante e importante para el éxito en varias tareas del Procesamiento del Lenguaje Natural tal como en: búsquedas web, anuncios en páginas web o en dispositivos móviles, filtrado de contenido, entre otras.* Extraer términos o frases clave desde resúmenes representa un reto por la baja frecuencia de las palabras. Por esta razón en este artículo se propone una técnica para extraer n-gramas clave desde resúmenes usando el estadístico Chi-cuadrado. La técnica fue evaluada sobre un conjunto de textos estándar usado por otros investigadores. Los resultados obtenidos por el sistema son alentadores dada la complejidad de la tarea.

PALABRAS CLAVE: extracción de frases clave, chi-cuadrada, n-gramas, procesamiento del lenguaje natural, sintagma nominal.

Figura 5: *Portada del artículo a publicar*

"CONGRESO INTERNACIONAL DE INVESTIGACIÓN E INNOVACIÓN 2018"
Multidisciplinario
19 y 20 de abril de 2018, Cortazar, Guanajuato, México
ISSN 2448-6035

Cortazar, Guanajuato México a 20 de enero de 2018

ASUNTO: Carta de Aprobación
CLAVE: P-UCEC366

Rosas Colula Luís Ángel, Dr. Franco Rojas López, Mtro. José Luis Hugo Díaz Biffano, M.C. Adriana Hernández Beristain.

PRESENTE

El que suscribe Director de Investigación de la Universidad del Centro de Estudios Cortazar, incorporada a la Universidad de Guanajuato (Oficio No. 253-04 DEL 25-XI-2004), hacemos constar que su proyecto de investigación:

"Identificando Palabras Clave en Resúmenes de Texto Aplicando Chi-cuadrado"

en **Modalidad de Ponencia** fue **Aprobado** según el Criterio Arbitral por nuestro Comité de Investigadores para presentarse en el

"CONGRESO INTERNACIONAL DE INVESTIGACIÓN E INNOVACIÓN 2018"
Multidisciplinario
19 y 20 de abril de 2018, Cortazar, Guanajuato, México

Se extiende la presente para los fines que al interesado convenga en la ciudad de Cortazar, Guanajuato, México.

"La calidad no está en las cosas que hace el hombre, sino en el hombre que hace las cosas"

Atentamente

Dr. Florentino Vázquez Puente
Director de Investigación (investigacion@ucec.edu.mx)
Universidad Centro de Estudios Cortazar
Inc. Universidad de Guanajuato
Tel (411) 155 46 47/48

c.c.p. archivo

Figura 6: *Carta de aceptación*

Figura 7: *Portada del congreso donde se publicará el artículo*

"CONGRESO INTERNACIONAL DE INVESTIGACIÓN E INNOVACIÓN 2018"
Multidisciplinario
19 y 20 de abril de 2018, Cortazar, Guanajuato, México
ISSN 2448-6035

Identificando Palabras Clave en Resúmenes de Texto Aplicando Chi-cuadrado

Rosas Colula Luís Ángel, estudiante, angelrosassf@gmail.com, Universidad Politécnica Metropolitana de Puebla.

Franco Rojas López, Dr. En ciencias de la computación, frlb99@gmail.com, Universidad Politécnica Metropolitana de Puebla.

José Luis Hugo Díaz Biffano, Mtro. En administración, hugo.diaz@metropoli.edu.mx, Universidad Politécnica Metropolitana de Puebla.

Adriana Hernández Beristain, M.C. Ciencias de la computación, adrianah.beristain@correo.buap.mx, Benemérita Universidad Autónoma De Puebla.

Resumen: *La extracción de términos o frases clave sigue siendo una tarea desafiante e importante para el éxito en varias tareas del Procesamiento del Lenguaje Natural tal como en: búsquedas web, anuncios en páginas web o en dispositivos móviles, filtrado de contenido, entre otras.* Extraer términos o frases clave desde resúmenes representa un reto por la baja frecuencia de las palabras. Por esta razón en este artículo se propone una técnica para extraer n-gramas clave desde resúmenes usando el estadístico Chi-cuadrado. La técnica fue evaluada sobre un conjunto de textos estándar usado por otros investigadores. Los resultados obtenidos por el sistema son alentadores dada la complejidad de la tarea.

PALABRAS CLAVE: extracción de frases clave, chi-cuadrada, n-gramas, procesamiento del lenguaje natural, sintagma nominal.

ISSN 2448-6035

ABSTRACT: Keyword or keyphrase extraction is an important and challenging useful task in several natural language processing tasks such as: web search, content filtering, and advertising on web pages or on mobile devices. Finding key terms/phrases from abstracts is a challenging task due to the low frequency of the words. In this paper, we propose a technique to extract n-grams from short texts (abstracts) applying the chi-square statistical. The technique was evaluated on a standard dataset used by other researchers. The Results are encouraging because of the complex task.

Keywords: keyphrase extraction, chi-square, n-grams. Natural processing language, noun phrase.

I. INTRODUCCIÓN

En recientes años la publicación de información ha crecido exponencialmente en todos los dominios, como en redes sociales, noticias, educación etc. Debido a este crecimiento, es imprescindible la extracción de palabras o frases clave para extraer información relevante desde diferentes medios de publicación de información. La extracción de palabras clave consiste en extraer las frases que mejor describen el tema de un documento. Diferentes tareas del Procesamiento del Lenguaje Natural (PLN) han sido beneficiadas de la extracción de términos clave por ejemplo, las frases clave de un documento han permitido la rápida y precisa búsqueda de un documento en una gran colección de textos. Aunado a lo anterior otras tareas del PLN y recuperación de información han mostrado una gran mejora tal como desambiguación del sentido de la palabra (Rojas-López, López-Arévalo, & Sosa-Sosa, 2012), identificar palabras clave relacionadas a un tema de interés para encontrar post relevantes en redes sociales (Shuai, Zhiyuan, Bing, & Sherry, 2016), así como para promover productos o servicios en sitios web

ISSN 2448-6035

a través de anuncios textuales (Zhang, Wang, Xue, & Zha, 2012), (Stamatina & Michalis, 2011).

El uso de redes sociales como Twitter y Facebook ha propiciado la publicación en línea de textos cortos y además informales, especialmente en Twitter cuyo contenido está limitado a 140 caracteres. Varias de las técnicas que se ha propuesto tienden a desempeñarse deficientemente en este dominio (Marujo, y otros, 2015). Esto nos obliga a proponer nuevos algoritmos para minar información útil usada por negocios, consumidores y para investigaciones relacionadas con la economía, salud, política, entre otras.

En este artículo se propone una técnica para extraer n-gramas clave a partir de textos cortos (resúmenes) usando el estadístico chi-cuadrado. La técnica propuesta tiene su origen en el contexto de desambiguar consultas web para obtener respuestas más precisas de acuerdo al interés de los usuarios.

El presente artículo está organizado como sigue. En la sección trabajo relacionado se presentan algunos trabajos afines con las aplicaciones de la tarea extracción de términos clave. En la sección metodología se describe la técnica propuesta. En la sección resultados se describe el conjunto de textos (dataset) usado en los experimentos llevados a cabo y los resultados obtenidos. El artículo concluye con la sección conclusiones, donde se presenta una discusión general acerca del trabajo, así como un panorama general para investigaciones futuras.

II. Trabajo relacionado

En nuestros días la extracción de palabras o frases clave sigue siendo una tarea intermedia imprescindible para el éxito de otras aplicaciones en el PLN. Por ejemplo, las palabras clave han sido empleadas para mostrar anuncios en una página web de acuerdo con su contenido (Zhang, Wang, Xue, & Zha, 2012),

ISSN 2448-6035

(Stamatina & Michalis, 2011). En este contexto (Yih, Goodman, & Carvalho, 2006) propusieron un método para extraer palabras clave desde una página web y a continuación mostrar anuncios en esa página de acuerdo con las palabras clave recuperadas. El método propuesto consiste en extraer características relevantes usando TF-IDF, metadatos desde la página web e información recuperada desde archivos log de consultas web. Después de llevar a cabo una fase de pre-proceso de la información recuperada, ellos consideran una frase de longitud máxima 5 como término clave. A continuación, implementan un clasificador entrenado usando aprendizaje automático para predecir si la frase es un término clave o no. Otra aplicación de extracción de términos clave es la clasificación o clustering de textos relacionados semánticamente. (Zhan & Dahal, 2017), propusieron un algoritmo basado en una red neuronal profunda para clasificar y agrupar textos relacionados semánticamente. La red neuronal produce representaciones binarias compactas de textos cortos y puede asignar la misma categoría a textos que tiene representaciones binarias similares. Para solventar el problema de falta de información contextual y ambigüedad en textos cortos usaron sustantivos y verbos como enriquecimiento semántico. Por otra parte (Habibi & Popescu-Belis, 2015) abordaron el problema de extraer palabras clave desde conversaciones con el propósito de recuperar para cada conversación corta, un número pequeño de documentos relevantes que pueden ser recomendados a los participantes. El enfoque propuesto usa *Latent Dirichlet Allocation* como técnica para modelar temas, por cada tema un conjunto de palabras clave es recuperado, a continuación, aplican una técnica de *clustering* para dividir las palabras clave en conjuntos más pequeños constituyendo consultas implícitas. Estas consultas son enviadas a un sistema de recuperación de documentos los cuales son recomendados al usuario.

ISSN 2448-6035

Dada la importancia de las aplicaciones de la extracción de términos clave, en el presente trabajo se propone una técnica para extraer n-gramas clave desde textos cortos (resúmenes) usando el estadístico chi-cuadrado. La metodología propuesta se explica en la siguiente sección.

III. Metodología

En este artículo se presenta una técnica para la extracción de n-gramas clave desde textos cortos. Tal técnica consiste en medir el grado de correlación entre un par de palabras para encontrar n-gramas clave a partir de textos. Para lograr este objetivo se implementó la técnica x^2 la cual ha sido usada por su efectividad en tareas del PLN (Rojas-López, López-Arévalo, & Sosa-Sosa, 2012).

Chi-cuadrado

Una alternativa para medir la independencia entre dos variables es el estadístico Chi-cuadrado el cual no asume probabilidades distribuidas normalmente. Cabe mencionar que no es de nuestro interés discutir los aspectos estadísticos de la x^2 una explicación más extensa puede verse en el libro de Christopher y Hinrich (Manning & Schütze, 1999). El valor x^2 entre dos términos se obtiene empleando la Ec. 1.

$$x^2 = \frac{(O - E)^2}{E} \qquad \text{Ec 1.}$$

Donde *O* denota la frecuencia observada y *E* denota la frecuencia esperada entre los términos w_i y $t_k = w_{i+1}$. Sea S = {w_1, w_2, w_3,..., w_n} una sentencia en el documento de texto, la extracción de n-gramas clave (bigramas y trigramas) es como sigue.

Para obtener la frecuencia esperada de bigramas usamos la Ec. 2. Donde $f(w_i, t_k)$ es la frecuencia de aparición del bigrama w_i y t_k en la sentencia S. La probabilidad

ISSN 2448-6035

de ocurrencia de los términos w_i y t_k se obtiene dividiendo su valor de frecuencia entre el número de ocurrencia (Σt) de todos los pares de bigramas encontrados en el documento.

$$E = \frac{\sum f(w_i, t_k)}{\sum t} \qquad \text{Ec 2.}$$

La frecuencia observada de los términos w_i y t_k se obtiene multiplicando los factores a y b de la Ec. 3. $f(w_i, t_k)$ es la frecuencia del bigrama w_i y t_k en la sentencia S y $f(w_i)$, $f(t_k)$ es la frecuencia del término en el documento de texto.

$$a = \frac{\sum f(w_i, t_k)}{\sum f(w_i)} \quad b = \frac{\sum f(w_i, t_k)}{\sum f(t_k)} \qquad \text{Ec 3.}$$

para obtener trigramas clave se sigue el procedimiento descrito anteriormente con la diferencia que $w_i = (w_i, w_{i+1})$ y $t_k = w_{i+2}$.

Experimentos y resultados

En esta subsección se presentan los resultados de evaluación de la Metodología propuesta.

DATASET

El dataset (conjunto de textos cortos) empleados en los experimentos fue usado por (Hulth, 2003). El dataset contiene 2000 resúmenes de artículos científicos en el idioma inglés. Los resúmenes están divididos en tres conjuntos: un conjunto de entrenamiento que contiene 1000 resúmenes, un conjunto de validación el cual contiene 500 resúmenes y un conjunto de prueba que contiene los restantes 500 resúmenes. El dataset incluye dos conjuntos de sintagmas nominales clave, uno controlado y no controlado. Nosotros usamos el conjunto de prueba con sintagmas no controlados indexado por profesionales.

ISSN 2448-6035

IV. Resultados

Antes de realizar el experimento aplicamos el siguiente pre-proceso. Primero usamos la biblioteca Apache OpenNLP para segmentar el texto en sentencias. Después eliminamos el conjunto de palabras cerradas tal como: artículos, pre-posiciones, conjunciones, etc. Es importante mencionar que no usamos un algoritmo de *stemm* para reducir las palabras a su raíz esto es así porque un *stemming* agresivo puede causar problemas de búsqueda (Croft, Metzler, & Strohman, 2009).

Después de la fase de pre-proceso aplicamos el estadístico chi-cuadrado descrito anteriormente. La columna 1 de la Tabla 1 muestra algunos trigramas recuperados por el enfoque propuesto. La columna 2 muestra las palabras únicas recuperadas a partir de los trigramas de la columna 1. Cabe mencionar que recuperamos los primeros 10 trigramas con mayor valor chi-chicuadrado.

Tabla 1. Respuesta del enfoque propuesto.

Trigramas	**Palabras clave**
Materials materials recurring Materials materials practices Print materials print …	Materials Recurring Practices Print …

La Tabla 2 muestra los resultados extraídos desde el conjunto de prueba.

ISSN 2448-6035

Tabla 2. Resultados extraídos desde el conjunto de prueba.

Sintagmas nominales
recurring issues
books
changing practices
out-of-print
library materials
acquisition
out-of-print materials

De acuerdo con los resultados mostrados en la Tabla 1 y Tabla 2, las palabras clave mostradas en la Tabla 1 se encuentran en los sintagmas nominales mostrados en la Tabla 2.

Los resultados mostrados son alentadores y creemos que si recuperamos sintagmas nominales con una longitud máxima de 5 (Yih, Goodman, & Carvalho, 2006) el enfoque propuesto puede obtener una buena precisión y cobertura. Es importante mencionar que los mejores resultados fueron obtenidos usando trigramas, los cuales son reportados en la Tabla 1 y 2. Aunado a lo anterior el conjunto de entrenamiento posee un vocabulario de tamaño 13,835 palabras el enfoque propuesto recupera un total de 2,456 palabras, es decir; reduce el vocabulario en un 82%.

V. Conclusiones

En el presente trabajo de investigación se presentaron resultados preliminares de la metodología, extracción de n-gramas clave a partir de textos cortos usando el estadístico chi-cuadrado. Los resultados obtenidos son alentadores dada la

ISSN 2448-6035

complejidad de la tarea para extraer términos clave a partir de resúmenes. Como trabajo futuro es nuestro objetivo extraer sintagmas nominales para comparar nuestro sistema con los trabajos reportados en la literatura. Aunado a lo anterior el trabajo presentado es la primera etapa de nuestro proyecto ya que nuestro objetivo es identificar las palabras clave que usa un usuario para acceder a una página web, así como usar los archivos log de consultas para desarrollar una aplicación móvil o web que permita dar a conocer promociones de tiendas departamentales en centros comerciales de acuerdo con la consulta ingresada. De acuerdo con lo anterior el resultado esperado es un sistema de recomendación de información de acuerdo con los intereses del usuario.

Agradecimientos

Los autores agradecen al Programa de Mejoramiento del Profesorado (PROMEP) por su apoyo financiero a través del proyecto F-PROMEP-39/Rev-04.

VI. Bibliografía

Croft, B., Metzler, D., & Strohman, T. (2009). *Search Engines: Information Retrieval in Practice.* Addison-Wesley Publishing Company.

Habibi, M., & Popescu-Belis, A. (2015). Keyword Extraction and Clustering for Document Recommendation in Conversations. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 746 - 759.

Hulth, A. (2003). Improved Automatic Keyword Extraction Given More Linguistic Knowledge. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 216-223.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing.* MIT Press.

Marujo, L., Ling, W., Trancoso , I., Dyer , C., Black , A., Gershman , A., . . . Carbonell, J. (2015). Automatic Keyword Extraction on Twitter. *Association for Computational Linguistics*, 637-643.

ISSN 2448-6035

Rojas-López, F., López-Arévalo, I., & Sosa-Sosa, V. (2012). Combining Local and Related Context for Word Sense Disambiguation on Specific Domains. *Data Science, Technology and Applications* , 135-140.

Shuai, W., Zhiyuan, C., Bing, L., & Sherry, E. (12-17 de February de 2016). Identifying Search Keywords for Finding Relevant Social Media Posts. *Proceedings of the Thirtieth Conference on Artificial Intelligence*, 3052-3058.

Stamatina, T., & Michalis, V. (2011). Multiword Keyword Recommendation System for Online Advertising. *International Conference on Advances in Social Networks Analysis and Mining*, 423-427.

Yih, W.-t., Goodman, J., & Carvalho, V. (2006). Finding Advertising Keywords on Web Pages. *Proceedings of the 15th International Conference on World Wide Web*, 213-222.

Zhan, J., & Dahal, B. (2017). Using deep learning for short text. *Journal of Big Data*, 34.

Zhang, W., Wang, D., Xue, G.-R., & Zha, H. (2012). Advertising keywords recommendation for short-text web pages using Wikipedia. *ACM Transactions on Intelligent Systems and Technology*, 36.

Ingeniantes

Revista de Investigación · Instituto Tecnológico Superior de Misantla

Midiendo la Similitud Semántica en Textos Cortos usando el Contexto Relacionado y DISCO

Franco Rojas López [1]
franco.rojas@metropoli.edu.mx

Jorge Jaime Juárez Lucero [2]
jorge.jaime@metropoli.edu.mx

Adriana Hernandez Beristain [3]
adrianah.beristain@correo.buap.mx

Vidal Armas [4]
vidal.armas@metropoli.edu.mx

Contacto:
franco.rojas@metropoli.edu.mx
2226854829

Inteligencia artificial

Franco Rojas-López, Dr. en ciencias de la computación, Ingeniería en Sistemas Computacionales, Universidad Politécnica Metropolitana de Puebla.
Jorge Jaime Juárez Lucero, M.C. en Optoelectrónica, Ingeniería en Sistemas Computacionales, Universidad Politécnica Metropolitana de Puebla.
Adriana Hernández Beristain, M.C. de la computación, Ingeniería en Sistemas Computacionales, Facultad de ciencias de la computación.
Vidal Armas Torres, Dr. En dirección y mercadotecnia, Administración y gestión de PYMES, Universidad Politécnica Metropolitana de Puebla.

***RESUMEN**: Medir el grado de similitud semántica entre texto o conceptos es una tarea desafiante e importante en varias tareas en Recuperación de Información y Procesamiento del Lenguaje Natural. Dada la importancia de la tarea, en este artículo se propone un método para medir la similitud semántica entre un par de oraciones usando la técnica hipótesis distribucional, para recuperar desde la Web, contextos relacionados con el conjunto de entrenamiento. Los contextos relacionados son un componente importante para calcular la similitud semántica entre pares de oraciones. En el artículo se presentan los resultados obtenidos desde un conjunto de entrenamiento estándar. La evaluación empírica muestra que el enfoque propuesto supera el baseline, así como algunos métodos propuestos previamente en el conjunto de entrenamiento estándar.*

***PALABRAS CLAVE:** contexto relacionado, información mutua, hipótesis distribucional, procesamiento*

del lenguaje natural, Similitud semántica.

ABSTRACT: ***Measuring the degree of semantic similarity between texts or concepts is a challenge task and important in several tasks in Information Retrieval and Natural Language Processing. Given the importance of the task in this paper proposes a method to measure the semantic similarity between a pair of sentences using the technique distributional hypothesis to extract from the web related contexts to the training set.***
The related contexts are an important component to calculate the semantic similarity between pairs of sentences. The article presents the results obtained from a standard training set. The empirical evaluation shows that the proposed approach exceeds the baseline, as well as some methods previously proposed in the standard training set.

KEYWORDS: distributional hypothesis, mutual information, natural language processing, related context, semantic similarity.

INTRODUCCIÓN

Medir la similitud semántica entre oraciones o conceptos es una tarea fundamental en varias aplicaciones del Procesamiento del Lenguaje Natural (PLN), por ejemplo, en sistemas de recomendación para filtrar información y guiar a los usuarios para descubrir productos o servicios en una forma personalizada (Kazem, Youssef, Keane, & Nelson, 2017), (Zhang, Yao, & Sun, 2017). Para ofrecer explicaciones verbales a partir de un par de oraciones (Lopez-Gazpio, y otros, 2016), lo cual puede ser aplicado en sistemas de tutorado inteligente (Rodney , Wayne, & James , 2009), en recuperación de información para medir la similitud entre la consulta y textos almacenados en una colección de documentos (Pilsen & Ptacek, 2012), entre otras aplicaciones. La tarea de medir la similitud semántica entre un par de oraciones se define como la determinación de cuán similares son los significados de dos oraciones. Medir la similitud no es una tarea trivial debido a la variabilidad del lenguaje y la ambigüedad del mismo, la cual es una característica intrínseca del lenguaje natural.
Por ejemplo, dado las siguientes dos sentencias extraídas de un corpus de encabezados de noticias los autores consideraron su similitud como "aproximadamente equivalente, pero difiere en alguna información menor" (Lopez-Gazpio, y otros, 2016).

killed in bus accident in Pakistan
killed in road accident in NW Pakistan

Para medir el grado de similitud entre ambas sentencias básicamente se han propuesto dos enfoques, el primero se basa en corpus y el segundo se basa en grafos de conocimiento. El primero mide la similitud semántica usando modelos de similitud distribucional aprendidos desde grandes colecciones de texto plano confiando en la distribución de las palabras. Dos palabras son similares si aparecen en contextos similares (Curran, 2004). El segundo enfoque mide la similitud semántica de conceptos usando grafos de conocimiento, las cuales capturan la similitud semántica entre dos oraciones usando un diccionario semántico tal como WordNet (Lingling, Runqing, & Junzhong, 2013).
En el presente artículo se propone un método para medir la similitud semántica entre un par de oraciones. El enfoque está basado en la extracción, representación e integración de información proporcionada por los verbos, sustantivos, adjetivos, adverbios recuperados desde el corpus de entrenamiento, los cuales

son usados para recuperar contextos relacionados desde la Web.

El resto del artículo está organizado como sigue, en la sección 2 se presentan trabajos relevantes relacionados con la similitud semántica entre conceptos y oraciones. En la sección 3 se describe la metodología propuesta para determinar la similitud entre un par de oraciones de texto. En la sección 4 se describen los experimentos y resultados, así como el conjunto de textos usados en la experimentación. Finalmente, en la sección 5, se discuten las conclusiones y temas prominentes de investigación futura.

TRABAJO RELACIONADO

Encontrar el grado de correlación entre pares de conceptos u oraciones es una tarea de suma importancia en tareas de recuperación de información y procesamiento del lenguaje natural. Por ejemplo, para determinar si una página web es plagio de otra, es decir; una página web puede ser espejo de otra que tiene casi el mismo contenido, pero diferente información. Dada la importancia de la tarea en la literatura se han propuesto un gran número de técnicas para encontrar la similitud semántica entre fragmentos de textos y conceptos, las cuales se dividen básicamente en similitud semántica basada en corpus y similitud semántica basada en conocimiento.

Similitud semántica basada en corpus

Los enfoques basados en corpus implementan métodos distribucionales para representar los significados conceptuales en vectores como, por ejemplo, Análisis Semántico Latente (Landauer & Dumais, 1997). En este contexto, Mikolov et al. (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) Desarrollaron un modelo para la representación distribuida de palabras y frases y su composición. El objetivo del modelo es encontrar representaciones de palabras que son útiles para predecir palabras en el contexto de una sentencia o documento.

Por otro lado, en el trabajo presentado por Ishizuka et al. (Ishizuka, Matsuo, & Bollegala, 2010) estimaron la similitud semántica entre pares de palabras usando motores de búsqueda. El método considera los hits retornados y patrones léxicos sintácticos extraídos desde los fragmentos de texto recuperados por el motor de búsqueda dada una consulta web. Usando una representación de las fuentes de información antes mencionadas implementaron el algoritmo Máquinas de Vectores Soporte (MVS) para encontrar pares de palabras sinónimos y antónimos. La función de similitud entre dos palabras P y Q es aproximada por un score de confianza del algoritmo MVS entrenado.

Similitud semántica basada en grafos de conocimiento

Zhu e iglesias (Zhu & Iglesias, 2017), propusieron un método para calcular la similitud semántica entre conceptos basado en grafos de conocimiento tal como WordNet y BDpedia. El método nombrado wpath, combina el Contenido de Información (CI) de conceptos para asignar un peso a la longitud de la ruta más corta entre ambos conceptos. La idea de usar CI para calcular la similitud semántica es que mientras más información comparte dos conceptos, más similares son. Majid y Alireza (Majid & Alireza, 2016) implementaron un método no supervisado basado en el conocimiento para medir la similitud semántica de los textos en los que se tienen en cuenta las similitudes específicas de palabra a palabra. El método esta implementado en un grafo bipartito sin pesos y no dirigido. Para un par de segmentos de texto, se producen conjuntos de palabras de clase abierta, con un conjunto distinto creado para sustantivos, verbos, adjetivos y adverbios. A continuación, determinan la similitud de los pares de palabras en los conjuntos correspondientes en la misma clase abierta en los segmentos de texto. Para sustantivos y verbos usan una medida de similitud semántica basada en WordNet.

Mientras que para las otras clases de palabras usan un emparejamiento léxico.
En el presente artículo se presenta un enfoque basado en corpus el cual tiene la ventaja de tener una mayor cobertura de vocabulario porque el modelo computacional puede ser efectivamente aplicado en un corpus actualizado o enriquecido. El enfoque propuesto de describe a continuación.

ENFOQUE PROPUESTO

Para determinar la similitud semántica entre un par de oraciones usamos el enfoque hipótesis distribucional, el cual está basado en la intuición de que las palabras que aparecen en contextos similares tienden a tener significados similares. En la Figura 1 se muestra la metodología propuesta la cual se explica a continuación.

Figura 1. Enfoque propuesto

Datos de entrenamiento

El corpus usado en los experimentos fue liberado por Microsoft Research Paraphrase, el archivo de texto contiene 4,077 pares de sentencias extraídas de fuentes de noticias en la Web con anotaciones humanas que indican si cada par de oraciones posee una relación de equivalencia semántica (Bill , Chris , & Chris, 2004).

Extracción de información relacionada

El sistema propuesto recibe como entrada un conjunto de entrenamiento, desde el cual fueron extraídas 8,154 sentencias. A continuación, las sentencias extraídas son procesadas por la biblioteca OpenNLP de apache. OpenNLP, es una herramienta basada en aprendizaje automático para el procesamiento de texto en lenguaje natural, la cual es usada para identificar frases desde las sentencias. Sea kp una frase identificada por OpenNLP el filtrado de las mismas es como sigue: si la longitud de kp es mayor o igual a 2 y menor igual que 5 las frases son almacenadas en una lista, esto así para tratar de evitar formular consultas ambiguas. Las frases almacenadas en la lista son enviadas como consultas al motor de búsqueda Google para recuperar documentos relacionados con el conjunto de entrenamiento.

Recuperación de documentos web

El uso de la Web ha mostrado su utilidad en varias tareas del PLN, por ejemplo, en Desambiguación del Sentido de la Palabra. Motivados por la utilidad de la web en el PLN en el presente trabajo se sigue un enfoque similar al presentado por Lopez-Arevalo *et al*. (Lopez-Arevalo, Sosa-Sosa, Rojas-Lopez, & Tello-Leal, 2017).
En nuestra propuesta nosotros recuperamos las primeras 20 respuestas retornadas por el motor de búsqueda Google en formato HTML, docx, y pdf. Se usó la Api boilerpipe de Google para extraer el contenido principal de texto de una página web; para convertir a texto plano los archivos en formato pdf, docx, se usó la herramienta Tika de apache. Una vez extraído el texto plano se empleó la herramienta OpenNLP para dividir el texto en sentencias. A este conjunto de sentencias le denominamos conjunto de texto relacionado (CTR). Cabe mencionar que una sentencia forma parte del Conjunto de Texto Relacionado si contiene un verbo, sustantivo, adjetivo o adverbio del Conjunto de Entrenamiento (CE).

Procesamiento de documentos

Una vez obtenido el CTR el texto fue dividido en sentencias, se eliminaron las palabras

cerradas tal como: artículos, preposiciones, conjunciones, símbolos de puntuación, etc., es decir; el conjunto de palabras que carece de significado. En el presente trabajo se usó toda la oración para recuperar el contexto en el cual aparece cada una de las palabras (verbos, sustantivos, adjetivos y adverbios) contenidas en el conjunto de entrenamiento.

Representación de la información

Es importante mencionar que para cada una de las palabras contenidas en las oraciones del CE ($wt_i \in$ CE) se recuperaron sus contextos relacionados desde el CTR obtenido desde la Web. Cada una de las palabras wt_i y su contexto relacionado es representada por una matriz (M) como se muestra en la Tabla 1. Donde Wt_i puede ser un verbo, sustantivo, adjetivo o adverbio del CE. W_j son las palabras extraídas del CTR.
Así cada palabra wt_i tiene asociado un vector de características.
$wt_i = \{w_0, w_1, \ldots, w_n\}$

Tabla 1. Representación del contexto relacionado

Palabra	**Contexto relacionado**				
	$\mathbf{w_0}$	$\mathbf{w_1}$	$\mathbf{w_2}$	**…**	$\mathbf{w_n}$
$\mathbf{Wt_0}$	v_{im00}	v_{im01}	v_{im02}	…	v_{im0n}
$\mathbf{Wt_1}$	v_{im10}	v_{im11}	v_{im12}	…	v_{im1n}
…	…	…	…	…	…
$\mathbf{wt_m}$	v_{imm0}	v_{imm1}	v_{imm2}	…	v_{immn}

Cada elemento m_{ij} en la matriz M tiene una frecuencia asociada. El valor de frecuencia entre una palabra wt_i y w_j es transformado en un valor de correlación usando la función Información Mutua (Church & Hanks, 1990), implementando la Ecuación 1. La frecuencia es usada por la función para asignar valores más altos a contextos que son más indicativos del significado de una palabra.

$$\mathrm{PMI}(\mathrm{wt_i}, \mathrm{wt_j}) = \log_2 \frac{\mathrm{P(wt_i, w_j)}}{\mathrm{P(wt_i)P(w_j)}} \qquad \text{Ec. (1)}$$

Donde $P(wt_i, w_j)$ es la probabilidad de que la palabra wt_i co-ocurra en el contexto w_j. Entre la probabilidad de observar wt_i y w_j independientemente. El peso asignado por la función Información Mutua es denotado por v_{imij} como se muestra en la matriz M de la Tabla 1.

Similitud semántica

La similitud semántica, es un valor que refleja el grado de correlación entre dos conceptos u oraciones. El enfoque propuesto usa una técnica basada en corpus o asociaciones estadísticas para calcular la similitud semántica entre palabras. Tal información es extraída desde la matriz de pesos M (ver la Tabla 1) del conjunto de textos relacionados. Para calcular la similitud semántica entre cada par de palabras se usó la fórmula del Coseno de Salton Ecuación 2.

$$sim_{coseno}(a_1, a_2) = \frac{\sum_{j=1}^{n} PMI(a_1, w_j) PMI(a_2, w_j)}{\sqrt{\sum_{j=1}^{n} PMI(a_1, w_j)^2 \sum_{j=1}^{n} PMI(a_2, w_j)^2}} \qquad \text{Ec. (2)}$$

donde a_1, a_2 son cualquier par de palabras que pertenecen al CE.

EXPERIMENTOS Y RESULTADOS

Para llevar a cabo la fase de experimentación el sistema recibe como entrada un par de oraciones. Por ejemplo, sea $s_1 = \{p_1, p_2, \ldots, p_n\}$ y $s_2 = \{q_1, q_2, q_3, \ldots, q_n\}$ un par de oraciones extraídas desde el CE, el cálculo de similitud semántica entre un par de oraciones se lleva acabo implementando la Ecuación 3.

$$sim(s_1, s_2) = \frac{\sum_{i=1}^{l_{s1}} \sum_{j=1}^{l_{s2}} sim(p_i, q_j) + DISCO(p_i, q_j)}{\sum_{k=1}^{2} l_{sk}} \qquad \text{Ec. (3)}$$

donde l_{s1}y l_{s2}, es la longitud de la sentencia 1 y la sentencia 2. l_{sk}es la suma de la longitud de la sentencia 1 y 2. En el enfoque propuesto dividimos el valor de similitud entre la suma de la longitud de las sentencias ($\sum_{k=1}^{2} l_{sk}$) para no favorecer aquella que contiene más palabras.

DISCO, es una aplicación Java que permite recuperar la similitud semántica entre palabras y frases. La similitud está basada en análisis estadístico de grandes colecciones de texto.
En la Tabla 2 se muestran los experimentos llevados a cabo. La primera fila muestra el resultado del algoritmo de Lesk (*baseline*) el cual mide el grado de traslape que existe entre un par de oraciones. Cabe mencionar que en las oraciones se removieron los stopwords y se obtuvo la raíz (stem) de cada palabra usando WordNet y JWNL.

Tabla 2. Resultados

No. Técnica	Técnica	Precisión %
1	Lesk	56.6
2	Kenter (Kenter & de Rijke, 2015) OoB + both aux unwghtd + swsn	78.1
3	Prajol (Shrestha, 2011)	71.0
4	Enfoque propuesto	78.4

La segunda fila muestra el resultado del método propuesto por Kenter, basado en representaciones vectoriales de términos computados desde datos no etiquetados.
En la tercera fila se muestra la precisión obtenida por el enfoque propuesto por Prajol (Shrestha, 2011), el cual consiste en un método basado en corpus no supervisado basado en el modelo de espacio vectorial.

CONCLUSIONES Y TRABAJO FUTURO

En el presente artículo se presentó un método para obtener el grado de similitud semántica entre un par de sentencias. El método propuesto obtuvo un buen desempeño comparado con otros enfoques reportados en la literatura. Nosotros creemos que sí, enriquecemos más el conjunto de contextos relacionados podemos incrementar el valor de precisión en el enfoque propuesto. Como trabajo futuro es de nuestro interés integrar el enfoque propuesto en un sistema de recomendación basado en técnicas de recuperación de información

BIBLIOGRAFÍA

Bill , D., Chris , Q., & Chris, B. (2004). Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. Proceedings of the 20th International Conference on Computational Linguistics.

Chiranjibi, S., & Raj Ojha, Y. (2013). Semantic Sentence Similarity Using Finite State Machine. Intelligent Information Management, 171-174.

Church, W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational linguistics, 22-29.

Curran, J. (2004). From distributional to semantic similarity. University of Edinburgh. College of Science and Engineering. School of Informatics.

Ishizuka, M., Matsuo, Y., & Bollegala, D. (2010). A Web Search Engine-Based Approach to Measure Semantic Similarity between Words. IEEE Transactions on Knowledge & Data Engineering, 1041-4347.

Kazem, Q., Youssef, A., Keane, K., & Nelson, J. (2017). A novel recommendation system to match college events and groups to. Computing Research Repository.

Kenter, T., & de Rijke, M. (2015). Short Text Similarity with Word Embeddings. Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 1411-1420.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 211.

Lingling, M., Runqing, H., & Junzhong, G. (2013). A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology.

Lopez-Arevalo, I., Sosa-Sosa, V., Rojas-Lopez, F., & Tello-Leal, E. (2017). Improving selection of synsets from WordNet for domain-specific word sense disambiguation. Computer Speech & Language, 128-145.

Lopez-Gazpio, I., Maritxalar, M., Gonzalez-Agirre, A., Rigau, G., Uria, L., & Agirre, E. (2016). Interpretable Semantic Textual Similarity: Finding and explaining. Computing Research Repository.

Majid, M., & Alireza, T. (2016). Texts semantic similarity detection based graph approach. The International Arab Journal of Information Technology.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 3111-3119.

Pilsen, 1., & Ptacek, T. (2012). Advanced Methods for Sentence Semantic Similarity. Tesis.

Rodney , N., Wayne, W., & James , M. (2009). Recognizing entailment in intelligent tutoring systems. Natural Language Engineering, 479-501.

Shrestha, Prajol. (2011). Corpus-Based methods for Short Text Similarity. Rencontre des Etudiants Chercheurs en Informatique pour le Traitement automatique des Langues

Zhang, S., Yao, L., & Sun, A. (2017). Deep Learning based Recommender System: A Survey and New Perspectives. Computing Research Repository.

Zhu, G., & Iglesias, C. (2017). Computing Semantic Similarity. IEEE Transactions on Knowledge & Data Engineering, 72-85.