INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN
LABORATORIO DE LENGUAJE NATURAL Y PROCESAMIENTO DE TEXTO

MODELO PARA LA GENERACIÓN AUTOMÁTICA DE RESÚMENES ABSTRACTIVOS BASADO EN GRAFOS CONCEPTUALES

TESIS

QUE PARA OBTENER EL GRADO DE

DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

P R E S E N T A

SABINO MIRANDA JIMÉNEZ

DIRECTORES DE TESIS:

DR. ALEXANDER GELBUKH

DR. GRIGORI SIDOROV

México, D. F., julio 2013

SIP-14 BIS

INSTITUTO POLITÉCNICO NACIONAL
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D.F. siendo las 13:00 horas del día 19 del mes de junio de 2013 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis titulada:

"Modelo para la generación automática de resúmenes abstractivos basado en grafos conceptuales"

Presentada por el alumno:

MIRANDA	**JIMÉNEZ**	**SABINO**
Apellido paterno	Apellido materno	Nombre(s)

Con registro: B 0 9 1 6 9 4

aspirante de: **DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN**

Después de intercambiar opiniones los miembros de la Comisión manifestaron ***APROBAR LA TESIS***, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Directores de tesis

Dr. Alexander Gelbukh

Dr. Grigori Sidorov

Dr. Sergio Suárez Guerra

Dr. Marco Antonio Moreno Ibarra

Dr. Gerardo Eugenio Sierra Martínez

Dr. Miguel Jesús Torres Ruiz

PRESIDENTE DEL COLEGIO DE PROFESORES

Dr. Luis Alfonso Villa Vargas

INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION EN COMPUTACION
DIRECCION

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESIÓN DE DERECHOS

En la Ciudad de México, D.F. el día 24 del mes de junio del año 2013, el (la) que suscribe Sabino Miranda Jiménez alumno(a) del Programa de Doctorado en Ciencias de la Computación, con número de registro B091694, adscrito(a) al Centro de Investigación en Computación, manifiesta que es autor(a) intelectual del presente trabajo de Tesis bajo la dirección de Dr. Alexander Gelbukh y Dr. Grigori Sidorov y cede los derechos del trabajo titulado **"Modelo para la generación automática de resúmenes abstractivos basado en grafos conceptuales"**, al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor(a) y/o director(es) del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección **sabino_m@hotmail.com**. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Sabino Miranda Jiménez

Nombre y firma del alumno(a)

Resumen

Actualmente, se cuenta con grandes cantidades de información textual, pero las personas no tienen suficiente tiempo para leer toda la información disponible y tomar decisiones importantes basadas en ella. De ahí, que las tecnologías para la generación de resúmenes, capaces de presentar la información de manera concisa, están adquiriendo suma importancia.

El resumen puede crearse a partir de uno o más documentos, el cual integra el contenido importante de las fuentes originales por medio de procesos de selección, combinación o generalización de las oraciones.

Hay dos enfoques principales para la generación automática de resúmenes: extractivo y abstractivo. En el enfoque extractivo, las oraciones que forman al resumen se seleccionan y se extraen literalmente de los documentos fuente; a saber, se hace un análisis superficial de los textos. Este enfoque se ha estudiado extensamente y ahora se conocen sus limitaciones. Por otro lado, en el enfoque abstractivo, se requiere de un análisis profundo del texto, por lo que, se usan representaciones intermedias para estructurarlo. Las representaciones usadas han sido, por lo general, sintácticas; y las representaciones semánticas para este propósito no han sido suficientemente investigadas.

En esta tesis se explora la generación de resúmenes abstractivos a partir de los textos representados como grafos conceptuales, ya que hemos encontrado que proporcionan una forma simple, natural y un nivel detallado para representar los textos. En este contexto, el resumen se obtiene por medio de operaciones (selección, poda, unión y generalización) sobre los grafos conceptuales. La selección de los nodos relevantes es de suma importancia, ya que el resumen se obtiene a partir de esos resultados. Se usa el algoritmo HITS para la ponderación de nodos, combinado con las restricciones sintácticas asociadas a los conceptos verbales.

Las operaciones de poda, unión y generalización reducen los grafos y se infiere nuevos conceptos que, en principio, no existen en el texto original. Los grafos reducidos representan el resumen del texto a nivel conceptual.

Abstract

Nowadays, we have a huge amount of textual information, but we do not have enough time to read all available information in order to make important decisions based on it. Hence, technologies for text summarization are gaining extreme importance.

A summary can be created from one or more documents, which consists of the important content from the original sources by means of sentence selection, sentence fusion, or sentence generalization.

In the field of automatic text summarization, there are two main approaches: extractive and abstractive. In the extractive approach, sentences are selected and extracted verbatim from source documents; namely, a shallow text analysis is conducted. After many years of study, it is well known that extractive summaries have important limitations. On the other hand, the abstractive approach needs a deep analysis of text; thus, intermediate representations are used to structure it. Generally, syntactic representations of text have been used, and semantic representations for this purpose have not been sufficiently explored.

In this dissertation, we explore the abstractive text summarization. Texts are represented as conceptual graphs. We have found that conceptual graphs provide us a simple, natural and a fine-grained level to represent texts. In this context, the summary is generated by means of several operations (selection, pruning, joining, and generalization) on conceptual graphs. The selection of salient nodes is great important because the summary is created based on those results. HITS algorithm is used for ranking nodes, combined with syntactic constraints related to verbal concepts.

Pruning, joining and generalization operations are used to reduce graphs, and new concepts can be inferred that, in principle, they do not exist in the original text. The reduced graphs represent the summary of the text at conceptual level.

Agradecimientos

Agradezco a mi madre, hermanos y familiares por su comprensión y apoyo durante estos años de trabajo.

A mis asesores el Dr. Alexander Gelbukh y el Dr. Grigori Sidorov por compartir conmigo sus conocimientos, sus experiencias y apoyarme durante mi estancia en el centro.

A los Dres. Gerardo Sierra Martínez, Sergio Suarez Guerra, Marco Antonio Moreno Ibarra y Miguel Jesús Torres Ruíz, miembros del comité tutorial, por sus valiosas aportaciones para el mejoramiento de esta tesis; y al Dr. Manuel Montes y Gómez por proporcionarme los resultados de su investigación.

No podía dejar de agradecer a todos mis amigos y amigas por la amistad que me brindan, los consejos y las discusiones sobre mi trabajo y otros temas.

Finalmente, mi reconocimiento al Consejo Nacional de Ciencia y Tecnología (CONACYT) y al Programa Institucional de Formación de Investigadores (PIFI-IPN) por el apoyo económico que me proporcionaron para el desarrollo de este proyecto.

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Capítulo 1

Introducción

En este capítulo, se discute la necesidad de explorar nuevas estrategias que permitan moverse hacia la generación automática de resúmenes abstractivos. Basándonos en esta necesidad, proponemos los objetivos de esta investigación, los cuales se orientan al desarrollo de un modelo para la generación de resúmenes considerando como representación intermedia de los textos a los grafos conceptuales, de esta manera los resúmenes obtenidos dentro de nuestro marco de trabajo serán a nivel conceptual.

Al final del capítulo se presenta la organización de esta tesis.

1.1 Motivación

Con el incremento exponencial de la información que ahora está disponible, surge un nuevo problema: el exceso de ésta. Actualmente, se cuenta con grandes cantidades de información textual, pero la gente no tiene suficiente tiempo para leer toda la información disponible, analizarla y tomar decisiones importantes con base en ella. De ahí que las tecnologías para la generación automática de resúmenes, capaces de presentar la información de manera concisa, están adquiriendo gran importancia.

El resumen se puede crear a partir de uno o más documentos, el cual integra el contenido más importante de las fuentes originales por medio de procesos de selección, combinación y generalización (Spärck Jones, 1999; Hovy & Chin-Yew, 1999; Spärck Jones, 2007). Por un lado, el contenido involucra a la información misma, así como a la forma en que se expresa dicha información; por otro lado, la importancia concierne a lo que es esencial y sobresaliente. Tales consideraciones se deben tomar en cuenta para que los resúmenes generados automáticamente sean útiles para las personas.

Los resúmenes se pueden clasificar de diversas formas (Spärck Jones, 1999; Hovy *et al*., 1999; Hahn & Mani, 2000; Spärck Jones, 2007; Torres-Moreno, 2011), dos de las clasificaciones más comunes que guían las investigaciones actuales son las siguientes: 1) por la cantidad de documentos que se usan para generar un resumen: monodocumento o multidocumento, y 2) por el tipo de resumen que se genera: extractivo o abstractivo.

Los resúmenes extractivos se crean a partir de la selección de varias oraciones consideradas sobresalientes en el texto original. Las oraciones se obtienen literalmente, se unen libremente, y se presentan como el resumen del texto. En este enfoque se hace un análisis superficial de los textos, a nivel de palabras u oraciones; por lo que, en general, los resúmenes no tienen coherencia y sólo se da una idea de lo que es sobresaliente en el texto.

Por el contrario, los resúmenes abstractivos se crean regenerando el contenido extraído del texto original; esto es, se reformulan las frases por medio de procesos de fusión, combinación o supresión de términos. De esta manera, se obtienen frases que en principio

no estaban en el texto de origen, similar al modo en que lo harían las personas; por lo que se requiere de una "comprensión" del texto para generar este tipo de resúmenes.

En el campo de la generación automática de resúmenes, se han desarrollado muchos trabajos y diversas técnicas desde finales de 1960 hasta ahora. Se han usado tradicionalmente la frecuencia de palabras o frases introductorias para identificar las oraciones más sobresalientes del texto ('*en resumen'*, '*lo más importante es'*) (Luhn, 1958; Baxendale, 1958; Edmundson, 1969; Chin-Yew & Hovy, 1997). También, se han desarrollado modelos estadísticos basados en corpus de entrenamiento para combinar varias heurísticas: palabras clave, posición de las oraciones, longitud de las oraciones, frecuencia de palabras y palabras contenidas en los títulos (Hovy *et al*., 1999).

Otros enfoques se basan en la representación del texto en forma de grafo, por ejemplo, las cadenas léxicas. En las cadenas léxicas, las oraciones importantes y los conceptos son las entidades altamente conectadas. Las oraciones que contienen a estas entidades altamente conectadas forman parte del resumen (Barzilay & Elhadad, 1997; Barzilay & McKeown, 2005).

De igual modo, se ha propuesto analizar la estructura discursiva y extraer las relaciones retóricas entre las diferentes unidades textuales, y así separar las unidades núcleo de las satélites para descubrir las unidades que juegan un papel principal dentro de la estructura discursiva (Marcu, 1997; Marcu, 1999; Marcu, 2000; Soricut & Marcu, 2003; da Cunha I. , 2008).

En la misma línea de representación del texto como un grafo conectado, se usan técnicas de Recuperación de Información para identificar oraciones similares y determinar las más importantes, la cuales formarán al resumen final (Salton, Singhal, Mitra, & Buckley, 1997; Carbonell & Goldstein, 1998; Radev D. , 1999; Torres-Moreno, Velázquez-Morales, & Meunier, 2001; Torres-Moreno, 2012).

También, se ha usado el "prestigio" de las unidades léxicas (oraciones o palabras) dentro del grafo a partir de las conexiones de similitud que hay entre el vocabulario traslapado. En estos casos, las oraciones se recomiendan entre sí a través de una función de similitud; de este modo, los nodos ganan puntos cada vez que son recomendados. Los

nodos que adquieren mayor puntaje al final de realizar la ponderación de los nodos se consideran candidatos para formar parte del resumen (Mihalcea & Tarau, 2004; Erkan & Radev, 2004b).

La mayoría de las investigaciones se han centrado en la generación de resúmenes extractivos; pero, actualmente, se ha puesto énfasis en la construcción de resúmenes abstractivos, los principales trabajos se han enfocado a la fusión y compresión de sentencias (Knight & Marcu, 2000; Knight & Marcu, 2002; Barzilay, 2003; Barzilay *et al.*, 2005; Ganesan, Zhai, & Han, 2010; Molina A. , Torres-Moreno, SanJuan, da Cunha, & Sierra, 2013). Se han propuesto modelos estadísticos y lingüísticos para calcular la compresión óptima de un texto, o para eliminar los constituyentes sintácticos hasta alcanzar un umbral predefinido por el usuario. Sin embargo, no se ha explorado suficientemente el uso de representaciones semánticas para la creación de resúmenes.

Se ha demostrado que para propósitos indicativos (idea de lo que contiene el texto) los resúmenes extractivos han sido adecuados, pero para otros propósitos como informativos es necesario generar resúmenes abstractivos (Hovy, 2005; Spärck Jones, 2007; Nenkova & McKeown, 2011).

En este trabajo de investigación, se explora la generación de resúmenes de un solo documento y de tipo abstractivo. Se propone un modelo para la generación de resúmenes, así como un conjunto de operaciones adaptadas para la identificación y reducción de nodos de los grafos conceptuales, los cuales formarán el resumen. Usamos grafos conceptuales como representación intermedia de los textos, ya que se aprovechan la simplicidad, naturalidad y granularidad que poseen estas estructuras para representar la información textual (Sowa J. F., 1984; Sowa J. F., 1999).

1.2 Descripción del problema

Actualmente, se ha puesto nuevamente atención a las tecnologías para la generación automática de resúmenes, debido a las grandes cantidades de información que ahora se

poseen. Tanta información es inútil si las personas no pueden tomar decisiones importantes con base en ésta.

El principal enfoque que se ha venido investigando para la generación de resúmenes ha sido el extractivo y poca atención se le ha dado al abstractivo, debido a que se requiere de un análisis profundo del lenguaje y diversos recursos lingüísticos auxiliares. Sin embargo, para que los resúmenes generados sean de utilidad para las personas, se requiere que dichos resúmenes sean coherentes y estructurados. En el enfoque extractivo, la coherencia se espera obtener de la concatenación de las oraciones seleccionadas, generalmente no sucede así y sólo se tienen las oraciones unidades libremente, lo cual da una idea del contenido del texto.

Se ha planteado que donde los resúmenes extractivos parecen ser inadecuados para ciertos propósitos, es necesario moverse a los resúmenes abstractivos, esto es, para identificar y representar adecuadamente el contenido de la fuente y crear un documento informativo en lugar de indicativo (Spärck Jones, 2007; Nenkova *et al*., 2011).

Los sistemas para generar resúmenes tienen un esquema general similar al de la figura 1.1 (Spärck Jones, 1999; Spärck Jones, 2007). En este esquema, se resaltan las etapas donde los sistemas difieren. En la manera en que se representan los documentos u oraciones (Interpretación-Representación), por ejemplo, *bolsa de palabras* usadas en el modelo de espacio vectorial; o nodos y aristas (representación basada en grafos). Por lo que es esencial el tipo de representación que se utiliza al interpretar los textos ya que redundará en los mecanismos que implementan la generación del resumen.

De acuerdo a los métodos revisados para la generación de resúmenes (véase el capítulo 2), las representaciones usadas no son suficientemente expresivas y flexibles para construir un resumen abstractivo.

Nosotros proponemos usar una representación simple y flexible, tanto estructural como semántica, de las unidades textuales consideradas más importantes en un documento y con ellas formular el resumen a través de varias operaciones sobre dichas estructuras.

Para lograr esta expresividad, nos basamos en el formalismo de grafos conceptuales introducido por Sowa (Sowa J. F., 1984). Dicho formalismo nos proporciona una muy

buena expresividad para representar las relaciones conceptuales y contextuales entre los conceptos representados (véase el capítulo 3).

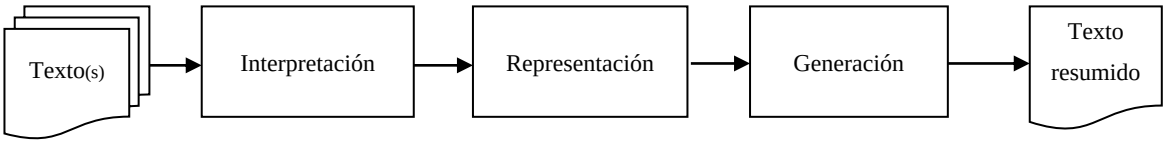


Figura 1.1 Esquema general para la generación de resúmenes automáticos

Entre las principales razones para seleccionar los gafos conceptuales de entre otras representaciones como lógica de predicados, árboles sintácticos o KL-ONE son las siguientes:

1. Los grafos conceptuales permiten representar adecuadamente —en términos de expresividad y eficiencia notacional— la información en lenguaje natural.
2. Los grafos conceptuales disponen de mecanismos formales que facilitan su manipulación, transformación y análisis.
3. Los grafos conceptuales se han usado en áreas como minerías de textos; por ejemplo, en la Recuperación de Información y en el agrupamiento de textos donde se han obtenido buenos resultados (Montes-y-Gómez, López-López, & Gelbukh, 2000a; Montes-y-Gómez, Gelbukh, & López-López, 2000b; Montes-y-Gómez M. , Gelbukh, López-López, & Baeza-Yates, 2001).

Como se ve en la figura 1.1, para lograr la representación es necesario interpretar el texto adecuadamente y transformarlo en un grafo conceptual. Esta transformación de los textos en grafos conceptuales es un problema complejo vinculado con su análisis sintáctico y semántico (Hensman & Dunnion, 2004; Hensman, 2005; Ordoñez-Salinas & Gelbukh , 2010).

Actualmente, esta tarea tiene sus propios retos y es un campo abierto para la investigación. De ahí que en este trabajo se asume que los textos están adecuadamente representados como grafos conceptuales, y los métodos propuestos se enfocan al tratamiento de tales grafos.

La creación de los grafos a partir del texto así como la generación del texto está fuera del alcance de este trabajo.

1.3 Hipótesis

El formalismo de grafos conceptuales tiene la suficiente expresividad para representar la semántica y estructura de los textos. De esta manera, el desarrollo de operaciones adaptadas para operar sobre los componentes de los grafos permitirá reducir las estructuras y mantener la coherencia entre sus componentes, los cuales representan al resumen.

1.4 Objetivos

Asumimos que tener mejores representaciones del contenido del texto permite "entender" los componentes y las relaciones que los vinculan y descubrir nueva información que aparentemente no existe. En este sentido los objetivos son los siguientes:

Objetivo general

- Proporcionar un modelo para la generación de resúmenes abstractivos basado en la representación del texto por medio de grafos conceptuales.

Objetivos específicos

- Proporcionar un método de ponderación de los elementos de los grafos conceptuales.
- Proporcionar un conjunto de operaciones para la síntesis de grafos conceptuales (generalización, unión, poda).

1.5 Organización del documento

El resto de este trabajo está organizado como sigue.

En el capítulo 2 se presenta una revisión del estado del arte sobre la generación automática de resúmenes. Aquí se discuten los principales enfoques y métodos representativos para la generación de resúmenes.

En el capítulo 3 se introducen los elementos básicos de la teoría de grafos conceptuales. Principalmente, se presenta la terminología elemental y se describen las operaciones canónicas para los grafos conceptuales. Así como las medidas de semejanza que se utilizarán.

En el capítulo 4 se presenta el método propuesto para la generación de resúmenes basado en la teoría de grafos conceptuales.

En el capítulo 5 se presentan los resultados. Por último, en el capítulo 6, se presentan las conclusiones, las limitaciones del modelo y las líneas de trabajo futuras.

Capítulo 2

Antecedentes

En este capítulo se detalla el estado del arte con relación a la generación automática de resúmenes. Se da una clasificación de los tipos de resúmenes de acuerdo a ciertas características del texto de origen, las cuales influyen al diseñar los métodos que pretenden resolver la tarea de generación de resumen.

También, se describen las dos líneas principales de investigación sobre el área en cuestión. Primeramente, se presenta el enfoque extractivo así como los métodos representativos que se han desarrollado en esta área. Posteriormente, se presenta el enfoque abstractivo y los métodos representativos de este enfoque.

También se describen las competencias actuales donde se evalúan los sistemas de generación automática de resúmenes, así como los métodos que se usan para evaluar los resúmenes generados.

Por último, se describen los métodos populares de ponderación en un contexto de estructuras en forma de grafos.

2.1 Generación automática resúmenes

Actualmente un problema que prevalece en la *Web* es la gran cantidad de información con que se cuenta. Esta información no es útil si no existen filtros que la depuren y ayuden a las personas a tomar decisiones basándose en ella.

Las tecnologías para la generación automáticamente resúmenes son herramientas de filtrado que pueden ayudar a las personas a lidiar con el problema de la sobrecarga de información.

Spärck Jones caracteriza el proceso de generación de resúmenes como la transformación de un texto fuente a un texto reducido por medio de la condensación del contenido importante por selección y generalización (Spärck Jones, 1999; Spärck Jones, 2007). El modelo general que propone para la mayoría de los sistemas de generación de resúmenes se mostró en la figura 1.1. Estas tres etapas generales (interpretación, representación y generación) se implementan a diferentes niveles y marcan las diferencias entre los sistemas que intentan resolver la generación automática de resúmenes.

En el esquema general, en la primera etapa, el texto se interpreta para tener una representación adecuada y enriquecida de su contenido para su manipulación; a partir de la representación del contenido del texto, las oraciones se regeneran de acuerdo a los procesos y métodos particulares implementados, para finalmente obtener un resumen.

El modelo que propone Spärck Jones describe, en general, a la mayoría de sistemas desarrollados, y al mismo tiempo remarca las diferencias en la forma en que se implementan las etapas en tales sistemas. Por ejemplo, algunos modelos propuestos representan las frases u oraciones como nodos en una representación basada en grafos (Erkan *et al*., 2004; Mihalcea & Tarau, 2004); otros, usan las palabras de la oración como características que la identifican y aprovechan el modelo de espacio vectorial para determinar su importancia dentro del documento (Salton *et al*., 1997; Carbonell *et al*., 1998; Torres-Moreno, 2012). Los métodos más representativos se presentan en la sección 2.3 y 2.4.

2.2 Clasificación de los resúmenes

Es difícil diseñar un sistema de generación de resúmenes si se desconocen las características de los textos, el uso que se va a dar a los resúmenes y la forma que se espera que tenga el texto final. Spärck Jones denominó a estas características factores de entrada, propósito y salida (Spärck Jones, 1999). Hovy (Hovy *et al*., 1999) extienden la clasificación de Spärck Jones e indican que cualquier resumen puede determinarse por al menos tres de las características de los siguientes grupos.

I. De acuerdo a la entrada del texto fuente.

1. Tamaño.
   - ***Monodocumento***. Se procesa un solo documento.
   - ***Multidocumento***. Se procesa más de un documento, generalmente textos que están temáticamente relacionados.
2. Especificidad.
   - ***Dominio específico***. Los temas pertenecen a un dominio restringido. Se considera menos ambigüedad de términos, el formato es especializado, entre otros.
   - ***Dominio general.*** No se puede considerar aseveraciones como en el específico.
3. Género y escala.
   - ***Género***. Los géneros pueden ser artículos periodísticos, editoriales, piezas de opinión, novelas, etc.
   - ***Escala***. Representa la longitud del texto. Los textos pueden ser libros, capítulos, párrafos, oraciones.

II. De acuerdo a la salida del texto fuente

1. Derivación.
   - ***Extractivo***. Se extraen fragmentos del texto de entrada (desde palabras hasta párrafos completos) y se presentan literalmente como el resumen generado.

- ***Abstractivo***. Es un nuevo texto generado, es decir, se toman como base los textos de origen y se reducen por medio de procesos de fusión generalización y compresión de las oraciones. Esto es el resultado de una representación y análisis interno del texto original.

2. Coherencia.
   - ***Fluidez***. Las oraciones que conforman el resumen son gramáticamente escritas.
   - ***Ininteligible***. El resumen está fragmentado, no hay una estructura coherente, las oraciones no son gramaticales.
3. Parcialidad.
   - ***Neutral***. Se refleja el contenido del texto de entrada, es imparcial.
   - ***Evaluativo***. Se incluyen frases de opinión, o se omite o agrega material.
4. Convencionalidad.
   - ***Fijo***. Se crean para uso específico: para los lectores o situaciones tales como convenciones internas de formato, etc.
   - ***Variable***. Se usan configuraciones variables del lector de acuerdo a diversos propósitos.

III. De acuerdo al <u>propósito</u> o uso del resumen

1. Audiencia.
   - ***Genérico***. Se proporciona el punto de vista del autor del texto, todos los temas son igualmente importantes.
   - ***Orientado al usuario***. Se favorecen los temas específicos o aspectos del texto de acuerdo al deseo del usuario.
2. Uso.
   - ***Indicativo***. Se proporciona simplemente el tema principal o el dominio del texto, sin incluir su contenido. Se puede entender de qué tema trata pero no necesariamente que contiene.
   - ***Informativo***. Se refleja el contenido (o parte) y se describen partes de lo que contiene el texto.
3. Expansión

- ***El fondo***. Se incluye material explicativo como circunstancias de tiempo, lugar y actores principales, ya que se asume que el conocimiento previo del lector es pobre en contenido.
- ***Lo novedoso***. Se presenta sólo el contenido nuevo, ya que se asume que el lector tiene el conocimiento suficiente para interpretarlo y contextualizarlo.

De acuerdo a las características anteriores, nuestro trabajo se centra en la generación de resúmenes abstractivos, informativos y de un solo documento. Estas características son las que guían el diseño de nuestro modelo.

Como hemos mencionado, los dos enfoques que han guiado las investigaciones actuales son el enfoque extractivo y el abstractivo. De ahí nuestro interés en analizar estos dos enfoques. En el siguiente aparatado se detallan estos enfoques.

2.3 Enfoque extractivo

Un resumen extractivo se crea a partir de la selección de las unidades textuales más importantes del texto original y se copian al texto destino (el resumen). Las unidades textuales que se extraen pueden variar de longitud, éstas pueden ser palabras, frases, oraciones o párrafos completos. Generalmente, se utilizan oraciones como unidades textuales, ya que tienen significado autocontenido.

El objetivo primordial de los sistemas para generar resúmenes extractivos es identificar las oraciones más importantes en el texto o textos. Por lo que se requiere de mecanismos que evalúen las oraciones individuales, les asignen un peso, y determinen el conjunto de oraciones más relevantes del texto. Esta etapa es clave debido a que es en la que se diferencian la mayoría de los sistemas extractivos.

En la siguiente sección presentamos los métodos más representativos para el enfoque extractivo. Aunque es difícil dar una clasificación de los métodos que se han desarrollado para la generación de resúmenes ya que a veces usan técnicas combinadas, trataremos de

ubicar a los métodos de acuerdo a la representación subyacente que predomina al modelar el contenido de los documentos.

2.3.1 Métodos

A finales de los años 1950, se desarrollaron los primeros métodos para la generación de resúmenes, se aprovechaba la distribución de las palabras en el texto y por medio de la frecuencia de las palabras o posición de las oraciones se determinaban las oraciones relevantes. En estos métodos se hacía un análisis superficial del texto. Estos primeros intentos demostraron la complejidad que tiene la tarea de la generación automática de resúmenes.

A partir de entonces, se han propuesto diferentes métodos a lo largo de los años. Algunos métodos retoman ideas de métodos anteriores con algunas variantes o combinan varios métodos en una nueva propuesta.

En los siguientes párrafos mencionamos los métodos que consideramos representativos; así como a otros que de alguna manera están relacionados con el enfoque que proponemos.

Métodos basados en frecuencia de palabras, palabras clave, posición

Luhn (Luhn, 1958) extraía las oraciones que contenían las palabras más frecuentes en el texto. Usó límites superiores e inferiores de las frecuencias de las palabras para determinar las palabras informativas frecuentes. Las oraciones se ponderaban de acuerdo al número de palabras frecuentes que éstas contenían. Las oraciones con mayor puntaje era las que formaban el resumen.

Baxendale (Baxendale, 1958), en sus investigaciones, identificó que la primera oración del párrafo contiene información más relacionada con el tema del párrafo. Sin embargo, las posiciones relevantes de las oraciones dependen del género del documento que se está analizando.

Edmudson (Edmundson, 1969) también utilizaba un método similar con algunas variantes, les daba mayor importancia a las palabras que eran usadas en los encabezados y

en los títulos de sección. También se le daba importancia a la posición de las oraciones si éstas aparecían al inicio o final de un párrafo. También incluyó el uso las palabras clave en dos conjuntos de frases, uno de *frases de bonificación* y otro con *frases estigma.* Estos conjuntos auxiliaban al momento de decidir si una oración era candidata para ser incluida en el resumen.

Las *frases de bonificación* se usaban como indicadores del contenido importante como '*en resumen'*, '*lo más importante es*', '*en conclusión*', etc. Sin embargo, estos indicadores son dependientes del dominio, por ejemplo '*en conclusión',* '*en resumen'* son más comunes en artículos científicos que en otros dominios como noticias. Una de las desventajas de usar este tipo de indicadores de entrada es que dependen del género del documento y es necesario identificar estos indicadores para el tipo de documento que se desea resumir.

Kupiec *et al.* (Kupiec, Pedersen, & Chen, 1995) usaban un clasificador Bayesiano para identificar las oraciones importantes que conformarían el resumen final. Para el entrenamiento del clasificador, se usaban como atributos la posición de las oraciones en el texto, las frecuencias de las palabras, el uso de palabras en mayúsculas y palabras temáticas. También se usaban palabras clave como '*resumen',* '*resultados'* o frases indicadoras como '*El objetivo principal de este escrito*', '*El propósito de este artículo*'.

En sus resultados mostraron que con el atributo posición se obtiene un 33% de precisión, con las palabras o frases fijas se obtiene 29%, y con la longitud de oraciones se obtiene 24%. Con la combinación de estos tres atributos se obtiene un 44%.

Uno de los problemas con este enfoque es que se usan atributos que dependen del dominio como son las palabras o frases fija y las palabras temáticas. Los experimentos realizados fueron con documentos científicos, en los cuales las frases fijas y las posiciones de las oraciones se usan, en cierta forma, sistemáticamente debido a la naturaleza del documento. No obstante, con los atributos de posición y longitud obtuvieron buenos resultados los cuales son independientes del dominio.

Estas características básicas de frecuencia, posición y palabras claves se han usado en distintos enfoques, ya que son buenas características que pueden identificarse por medio de un análisis superficial y sencillo en los textos.

Métodos basados en el modelo de espacio vectorial

Una de las estrategias más usadas se basada en el modelo de espacio vectorial (*Vector Space Model*, VSM), el cual se usa en Recuperación de Información (*Information Retrieval*, IR) (Salton & McGill, 1983). En este enfoque el texto se representa en forma de vectores, ya sea a nivel de documento, de párrafo o de oraciones. Es decir, los componentes de los vectores son los términos (generalmente palabras), y cada vector define a la unidad de trabajo elegida (documento, párrafo u oración). Se usan las frecuencias de aparición de los términos en los textos como las componentes del vector en un espacio multidimensional. En la figura 2.1, se muestra la representación de los textos a nivel de documentos en un modelo de espacio vectorial. Cada vector representa a un documento.

Generalmente, no sólo se usan las frecuencias de aparición de los términos en el texto, sino que se usa una estrategia de ponderación de los términos del vector basado en su importancia relativa (referente al texto) y su importancia global (referente a una colección de textos), esto es, la combinación conocida de frecuencia del término (*Term Frequency*, TF) y la frecuencia inversa del documento (*Inverse Document Frequency*, IDF). Esta combinación definida como TF*IDF refleja qué tan importante es un término con relación al texto en cuestión dentro de un conjunto de textos. De esta manera, las componentes del vector (las frecuencias de los términos) se recalculan con la fórmula TF*IDF (Manning, Raghavan, & Schütze, 2008), y se obtiene la "importancia real" de cada componente del vector.

En esta representación para determinar la similitud entre los documentos (vectores) se calcula el ángulo θ entre los vectores. Generalmente se usa una función de comparación de *similitud coseno* entre los vectores: $Cos\theta$. Los valores de la función están entre 0 y 1, siendo 1 el valor máximo el cual representa la máxima semejanza.

$$Cos\theta = \frac{V_1 \cdot V_2}{\|V_1\| \, \|V_2\|}$$

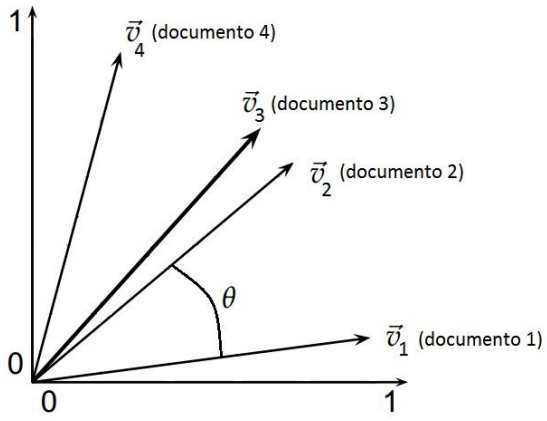


Figura 2.1 Representación de documentos en un modelo vectorial

El VSM se usa ampliamente en IR, y dado sus características para representar los documentos en forma de vectores se aprovecha ésta para la generación de resúmenes. En este contexto, la generación del resumen se puede ver como la extracción de segmentos de texto o pasajes relevantes.

También se puede identificar en esta representación que los pasajes semejantes forman subyacentemente un grafo conectado por las similitudes de tales pasajes, lo cual es otra característica que se ha usado para identificar la relevancia entre los segmentos de textos. Esta representación, el VSM, se ha usado en diferentes estrategias para la generación de resúmenes, algunas de ellas son las siguientes.

Salton *et al*. (997) usaban el VSM, para identificar los textos o fragmentos de textos que tienen características similares, los textos o fragmentos se representan en vectores y se calcula la similitud por medio de la función *similitud coseno.* Aquí la colección de palabras o textos no tiene orden, ya que se tiene una representación de *bolsa de palabras*. En este esquema se definen unidades léxicas de trabajo ya sean oraciones o párrafos, las cuales se consideran nodos del grafo que se forma. Las relaciones que se establecen entre las unidades léxicas (nodos) son de acuerdo al número de palabras que comparten. De esta manera, los nodos que están altamente conectados son buenos candidatos para formar parte del resumen.

La Relevancia Marginal Máxima (*Maximal Marginal Relevance,* MMR) (Carbonell *et al*., 1998) es un método enfocado a la generación de resúmenes monodocumento y

multidocumento, especialmente para generar resúmenes orientados a consultas, y usa el VSM como representación de los textos.

En la MMR se seleccionan segmentos de texto candidatos que se incluirán en el resumen final, balanceando la relevancia y la redundancia en cada iteración. La redundancia se calcula por medio de la similitud del contenido entre cada segmento candidato y el estado del resumen actual.

Para obtener la "novedad relevante", se miden independientemente la relevancia y la novedad, y se proporciona como métrica una combinación lineal de ambas, a dicha combinación lineal le llaman *relevancia marginal*.

Para la generación del resumen se considera que la información redundante no debe formar parte del resumen, por lo que los candidatos que contienen palabras que ya están en el resumen son penalizados. El texto se segmenta en oraciones, y se usa la MMR con una métrica de *similitud coseno* para reclasificar las oraciones. Se obtienen las oraciones con mayor puntaje de acuerdo con la clasificación y se presentan en el orden en el cual aparecían en el texto original.

Otros sistemas básicamente estadísticos basados en el VSM son el sistema CORTEX (Torres-Moreno, Velázquez-Morales, & Meunier, 2001) y su versión adaptada orientada al usuario NEO-CORTEX (Boudin & Torres-Moreno, 2007), el sistema ENERTEX (Fernández, SanJuan, & Torres-Moreno, 2007), y el sistema ARTEX (Torres-Moreno, 2012)

CORTEX es un sistema que genera resúmenes extractivos utilizando un algoritmo de decisión que combina varias métricas: frecuencias, TF*IDF, entropía, varias distancias de de Hamming para los párrafos y oraciones, títulos y subtítulos, además de la posición. Cada métrica tiene un grado diferente de importancia informativa. El sistema decide si un segmento de texto es importante de acuerdo a la información que proporciona cada una de las métricas.

ENERTEX es un sistema de generación de resúmenes extractivos inspirado en la física estadística que codifica un documento como un sistema de espines; posteriormente, se computa la energía textual entre oraciones para asignarles una puntuación. El método se

basa en la representación de las oraciones del documento como vectores, cada uno de estos vectores se estudia como una red neuronal de Hopfield, y se determina la energía textual entre los vectores, es decir, el grado de similitud a través del nivel de interacción entre los términos y las oraciones, ya que los vectores están correlacionados de acuerdo con las palabras que comparten. En este contexto, la energía textual se usa para ponderar las oraciones de un documento, y seleccionar las oraciones más relevantes de las que no lo son.

ARTEX este sistema, al igual que los anteriores, se basa en la representación del documento en vectores. La diferencia es que se usan vectores de pseudo-oración promedio (el tópico global) y pseudo-palabras promedio (peso léxico). Esto es, se construye un vector de documento promedio que representa el promedio de todos los vectores a nivel de oración (el tópico global). También se obtiene el peso léxico de cada oración, esto es, el número de palabras en la oración. De esta manera, se calculan las similitudes entre los vectores oración y el vector documento promedio. Los vectores más similares al vector documento promedio se extraen y son los que representan como el resumen.

Métodos basados en grafos

Los métodos que se basan en grafos se han popularizado y son buenos para localizar las unidades textuales sobresalientes en el documento, y se cree que al usar estructuras como grafos se puede tener mejor comprensión de las relaciones y asociaciones entre las unidades textuales. Enfoques similares al LexRank (Erkan *et al*., 2004), LexPageRank (Erkan & Radev, 2004a) y TextRank (Mihalcea *et al*., 2004) se han aplicado para la generación de resúmenes (Litvak & Last, 2008; Wang, Wei, Li, & Li, 2009; Thakkar, Dharaskar, & Chandak, 2010). LexRank y LexPageRank se han aplicado a la generación de resúmenes multidocumento, mientras que el TextRank se ha aplicado a la extracción de términos y la generación de resúmenes de un documento.

Mihalcea *et al*. (2004) exploraron el uso de grafos para la identificación de las oraciones más importantes (TextRank), la idea es similar a la que propuso Salton (Salton *et al*., 1997) en representar el texto por medio de la conectividad de sus oraciones, la diferencia es que se usa el modelo PageRank (modelo para para identificar páginas web populares) (Page & Brin, 1998; Page, Brin, Motwani, & Winograd, 1999) y el modelo

HITS (para identificación de temas en la web) (Kleinberg, 1999) para ponderar las oraciones.

En este enfoque, los nodos del grafo representan las oraciones; y las aristas, a las similitudes entre las oraciones. Dos nodos están relacionados si existe una relación de similitud (o recomendación). La función de similitud mide el traslape del contenido entre las oraciones. Este proceso de recomendación se aplica a todas las oraciones por lo cual se obtiene, generalmente, un grafo conectado. Durante el proceso de ponderación cada nodo incrementa su puntaje, si es recomendado, y al final los nodos con mayor puntaje formarán el resumen.

LexRank (Erkan *et al*., 2004b) y LexPageRank (Erkan *et al*., 2004a) son métodos para identificar las frases principales en un conjunto de documentos, se basan en la centralidad de vectores propios con la representación de las oraciones como grafo y en el prestigio de éstas. Las oraciones se consideran nodos, y los enlaces entre los nodos se representan por un peso; el peso es el valor de la función de *similitud coseno* entre el par de oraciones relacionadas. Las oraciones se representan en un modelo de espacio vectorial. Este enfoque busca identificar la centralidad de cada oración dentro de un *cluster* de oraciones (o documentos) con relación a un mismo tema y extraer las oraciones más importantes basadas su prestigio (recomendación entre las mismas oraciones) para que conformen el resumen. Para determinar el prestigio se usa el método PageRank.

La centralidad de la oración se calcula en términos de las propiedades léxicas de la misma oración. Al igual que TextRank usa el concepto de recomendación entre las oraciones; se considera una razón de compresión (umbral) definido por el usuario para reducir el grafo y quedarse sólo con los nodos más sobresalientes.

Las cadenas léxicas (L*exical Chains*) se han usado también (Barzilay & Elhadad, 1997; Barzilay *et al*., 2005) como método para identificar las oraciones significativas en el texto. Las cadenas léxicas se forman en los textos coherentes y cohesivos a través de sus oraciones, ya que las oraciones refieren a conceptos ya mencionados, en oraciones previas, o a conceptos relacionados entre ellos (Halliday & Hasan, 1976; Hirst & St-Onge, 1998). Cada palabra en la cadena léxica está relacionada con las demás ya sea porque es el mismo

concepto referido o está semánticamente cercano o asociado por medio de algún criterio de asociación (sinonimia, antonimia, hiperonimia u holonimia). El cálculo del puntaje de la cadena léxica se determina por el número y tipo de relaciones en la cadena. Las oraciones que contienen a cadenas léxicas representativas se extraen para formar parte del resumen. Este enfoque se hace uso de las relaciones y jerarquía de bases de datos léxicas, por ejemplo, WordNet (Fellbaum, 1998) para determinar las correspondencias entre las oraciones utilizando las palabras que contiene.

MEAD[1] (Radev *et al*., 2004) es un sistema muy popular orientado a la generación de resúmenes multidocumento que implementa varios métodos como los que se basan en la posición de las oraciones, palabras clave, en el *centroide* de los textos y el prestigio de las oraciones (LexPageRank), en las similitudes de *función coseno* entre oraciones, ponderación por medio de MMR; además de poseer otras características.

Básicamente, el sistema realiza cuatro etapas. Primero, los documentos de un *cluster* (colección) se convierten en un formato específico; segundo, se extraen las características de cada oración del *cluster*; tecero, para cada oración las características se combinan para obtener un puntaje compuesto; cuarto, los puntajes puede refinarse después de considerar posibles dependencias entre las oraciones como oraciones repetidas, orden cronológico, preferencias, etc. Este sistema está disponible para el público.

Teoría de la Estructura Retórica

Otros métodos que usan técnicas lingüísticas y representaciones en forma de grafos son basados en la Teoría de la Estructura Retórica (*Rhetorical Structure Theory,* RST) (Mann & Thompson, 1988).

El enfoque propuesto por Marcu para la generación de resúmenes (Marcu, 1997; Marcu, 1999; Marcu, 2000) consiste en obtener el resumen a partir del árbol generado del análisis de un texto basado en su estructura retórica. En la RST, se propone que existen

[1] http://www.summarization.com/mead/

unidades textuales (frases o párrafos) que juegan un papel más importante que otras, a éstas se les llama *núcleos* y las que no son esenciales se les nombra *satélites.*

En esta teoría, se cuenta con varias relaciones que unen a los núcleos con sus satélites y con otros núcleos formando una estructura jerárquica. Con esto se llega a crear un árbol conectado por las relaciones retóricas como *justificación*, *elaboración*, *preparación*, *fondo*, entre otras. A partir del árbol formado se extraen los núcleos importantes excluyendo a los satélites; además se considera que los elementos superiores de la estructura jerárquica son más importantes que los inferiores.

Para determinar la importancia de las unidades discursivas se ponderan recursivamente los nodos tomando en cuenta la importancia de sus nodos dependientes y el nivel de profundidad donde se identifica la unidad que se detectó importante. Las unidades discursivas que obtienen mayor puntaje se consideran como parte del resumen.

En la figura 2.2, se muestra un ejemplo de un documento representado por su estructura retórica. Ejemplo extraído del corpus en español anotado con relaciones de la RST (da Cunha I. , Torres-Moreno, Sierra, Cabrera-Diego, Castro Rolón, & Rolland Bartilotti, 2011).

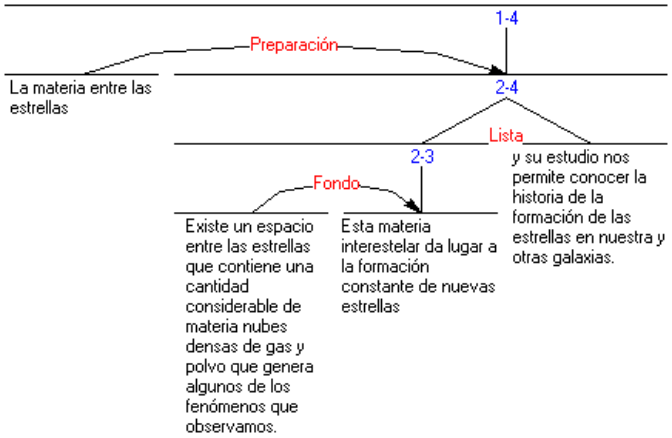


Figura 2.2. Árbol formado a partir de la estructura discursiva un texo

A partir del trabajo influyente de Marcu (2000), otros trabajos se han desarrollado en la misma dirección con algunas variantes para otros idiomas como el español (da Cunha, Wanner, & Cabré, 2007) y el portugués (Pardo, Rino, & Nunes, 2003).

En el enfoque que propone da Cunha *et al*. (2007; 2008) para resumir documentos en español en el campo médico, se hace uso de la estructura discursiva de los textos basada en la formalización mediante relaciones de la RST. Se usa un conjunto de relaciones discursivas seleccionadas para identificar el contenido relevante del texto en un dominio especializado; adicionalmente, el método se basa en reglas que consideran los rasgos superficiales de la estructura discursiva y algunas estructuras sintácticas que ayudan a aumentar la importancia de los fragmentos del texto donde aparecen éstas; por ejemplo, si una oración está dentro de las dos primeras oraciones de la sección de *resultados*, entonces la oración aumenta su valor de importancia. Esta combinación de información permite obtener una representación parcial del discurso y determinar los fragmentos de texto relevantes.

También se han realizado estudios del beneficio de usar la compresión de frases basada en la RST para optimizar un resumen (Molina, da Cunha, Torres-Moreno, & Velázquez-Morales, 2011). Lo que indican estos estudios es que en ciertas circunstancias ayuda a algunos sistemas de generación de resúmenes y es poca la mejora basado en las métricas ROUGE.

Métodos híbridos

También se han propuesto métodos híbridos, es decir, que implementan métodos estadísticos combinados con métodos lingüísticos. A continuación mencionamos algunos de ellos.

El enfoque propuesto en (da Cunha I. , Torres-Moreno, Velázquez-Morales, & Vivaldi, 2009) implementa un modelo de selección para determinar qué oraciones se incluirá en el resumen final. El modelo de selección se basa en los resúmenes obtenidos por otros sistemas de generación de resúmenes básicamente estadísticos (CORTEX, ENERTEX), y lingüísticos YATE (Vivaldi & Rodríguez, 2001) y DICOSUM (da Cunha, 2008).

La selección de oraciones se realiza en diferentes fases tomando en cuenta si una oración fue seleccionada por todos los sistemas o sólo por algunos, considerando el puntaje que se obtiene por cada sistema, además de la posición en el texto original. La combinación

de varios enfoques dio mejores resultados que los resultados independientes de cada sistema.

Otra enfoque es SummTerm (Vivaldi, da Cunha, Torres-Moreno, & Velázquez-Morales, 2010), el cual es un sistema para la generación de resúmenes especializados con características semánticas. Este sistema se basa en la obtención de los puntajes relevantes de cada oración tomando en cuenta el grado de importancia de los términos que contiene la oración en relación con un grupo de términos definidos para un dominio específico, y la similitud entre tales términos y los términos que aparecen en el documento.

El conjunto de términos se obtienen de una base de datos léxica, EuroWordNet[2], como fuente de información del dominio. La información temática se combina con el sistema YATE y CORTEX para la generación del resumen.

2.4 Enfoque abstractivo

Varios de los métodos implementados con enfoque abstractivo han sido aplicados a múltiples documentos. Se sabe que las estrategias para reducir un solo documento son diferentes a las que intentan combinar más de un documento, esto es, redundancia de información, algún orden en presentar los eventos, por ejemplo, cronológico, etc. Sin embargo existen retos comunes como la combinación de frases que mantenga la relevancia y la redundancia mínima así como la coherencia (Carbonell *et al*., 1998; Barzilay *et al*., 2005). En los siguientes párrafos, se presentan algunos métodos representativos y estrategias que se han implementado para crear el resumen abstractivo, ya sea por generalización, unión, eliminación o compresión de oraciones.

[2] http://www.illc.uva.nl/EuroWordNet

2.4.1 Métodos

Sistema SUMMONS

SUMMONS (McKeown & Radev, 1995; Radev D. , 1999) es un generador de resúmenes multidocumento sobre noticias de eventos terroristas. Su método se basa en preprocesar los documentos de noticias y se crean plantillas para todos los artículos relacionados, éstas contienen información del evento sobre el terrorista, las víctimas, el lugar, etc. Se realiza *clustering* con tales plantillas para identificar los temas principales, cada uno de los *clusters* pasan a la etapa de generación para combinarse.

En la etapa de generación del resumen se utilizan dos componentes: un planeador de contenido y un componente lingüístico. El planeador de contenido genera una representación conceptual del significado del texto, se usa un formalismo de unificación funcional y una gramática funcional (Halliday M. , 1985). El componente lingüístico selecciona las palabras adecuadas para referirse a los conceptos contenidos en la información seleccionada.

También, se utilizan varios recursos lingüísticos como diccionarios léxicos, gramáticas del inglés, ontología de dominio, bases de conocimiento de eventos previos y bases de datos con frases hechas para el "reuso" del lenguaje. Los diccionarios léxicos contienen el vocabulario del sistema, restricciones codificadas, de cómo cada palabra puede usarse, etc.

La ontología se usa para describir las relaciones entre las entidades y los eventos en el dominio del terrorismo, por ejemplo, un atentado en *Tel Aviv* con otro en *Jerusalén* para indicar que ambos se realizaron en el mismo país. La base de conocimiento de eventos previos se usa para agregar información conocida anteriormente del evento y que no se encuentra presente en el resumen actual. La base de datos para el reuso del lenguaje consiste en un diccionario de frases que mapea una entidad nombrada (persona, lugar, organización) con todas las posibles frases sustantivas usadas para describirla en noticias anteriores. Con todos los recursos anteriores, el sistema integra información ontológica, histórica y la aplica sobre el resumen actual para mejorarlo.

Cut and Paste

Otro método orientado a la abstracción de resúmenes es el denominado *Cut and Paste* (Jing, 2001). Es independiente del dominio y orientado a un solo documento. En la primera parte, se extraen las oraciones más importantes del documento: se utilizan las relaciones léxicas entre las palabras para identificarlas, también se incorpora otro tipo de información tal como medidas estadísticas usadas en Recuperación de Información, la posición de las oraciones, palabras clave, etc. En la segunda parte, se basa en dos módulos para reducir las frases extraídas: reducción de oraciones y combinación de oraciones.

El primer módulo se encarga de remover frases extrañas de las oraciones extraídas; el segundo, se encarga de combinar las oraciones extraídas o las oraciones reducidas previamente. El módulo de reducción utiliza varios recursos de información como información léxica (WordNet), conocimiento sintáctico, datos estadísticos obtenidos de corpus entrenados que contienen frases de ejemplo que fueron reducidas por humanos. El módulo de combinación usa reglas que se identificaron en resúmenes de ejemplo escritos por profesionales.

La hipótesis que se asume en este enfoque es que los profesionales del resumen siguen al autor muy de cerca para reintegrar los puntos más importantes en un texto más corto. Este enfoque define seis operaciones para reducir un texto:

1) R*educción de sentencias*. Remueve frases extrañas de la oración.

2) *Combinación de sentencias*. Une información de varias frases.

3) *Transformación sintáctica*. Cambia la estructura sintáctica de la oración.

4) *Parafraseo léxico*. Transforma una frase por su paráfrasis, '*point out*' por '*note*'

5) *Generalización/especificación.* Reemplaza frases u oraciones con descripciones más generales o específicas

6) *Reordenamiento*. Cambia el orden de las frases extraídas.

Con esta combinación de recursos, y las operaciones indicadas se genera el texto resumido.

Fusión de Información

En Barzilay *et al.* (2003; 2005) proponen la fusión de información para la generación de resúmenes multidocumento. La estrategia consiste en analizar las frases de un mismo grupo de documentos temáticos (*cluster*) y regenerar una nueva frase que contiene sólo la información común de la mayoría de las frases del grupo. Este método opera en tres etapas: se analizan sintácticamente las frases de cada grupo, se alinean los árboles de dependencias resultantes, y se genera la nueva frase a partir de los elementos que coincidieron en los árboles sintácticos.

La generación de texto se obtiene mediante la selección de una oración del grupo como oración clave, a ésta se le agrega la información de las frases que coinciden a lo largo del procesado del tema. La parte central para la reducción del texto es el alineamiento de los árboles de dependencias (Mel'čuk, 1988) que representan a las oraciones. En estos árboles no se consideran la representación de palabras auxiliares. El alineamiento se basa en dos fuentes de información: medida de similitud entre dos palabras, y similitud entre los árboles de dependencias. Adicionalmente, se usan bases de datos léxicas como WordNet y diccionarios que contienen frases con sus parafraseos asociados. También se determina la similitud estructural entre los árboles, se toma en cuenta el tipo de relación entre nodos tal como sujeto-verbo, adjetivo-sustantivo, etc.

En el sistema *NewsBlaster* (McKeown *et al*., 2002) se implementan métodos como *Cut and Paste* y algunos procesos de fusión de información, para múltiples documentos con base en estas ideas. Este sistema es básicamente estadístico que usa también elementos simbólicos para resumir noticias. Es su estructura general, primero se identifican las artículos de noticia y se agrupan; el agrupamiento de eventos se hace multinivel usando TF*IDF y características sintácticas como términos, frases sustantivas que encabezan a sustantivos propios. Los eventos se clasifican en una de seis categorías predeterminadas, para posteriormente generar un resumen para cada *cluster*.

Compresión de frases

Otro enfoque que ha tomado impulso es la compresión de frases, ya que es parte fundamental para la generación de un resumen (Knight *et al*., 2000; Knight *et al*., 2002; Vandeghinste & Pan, 2004; Steinberger & Jezek, 2006; Madnani, Zajic, & Dorr, 2007; Cohn & Lapata, 2009; Molina *et al*., 2011; Molina A. , Torres-Moreno, da Cunha, SanJuan, & Sierra, 2012; Molina A. *et al*., 2013). En Knight *et al.* (2000) proponen el modelo *noisy-channel* para la compresión de frases. En el modelo *noisy-channel* se considera que se tiene una cadena de texto larga y que era corta inicialmente, a través de algún proceso a la cadena corta se fue agregando texto adicional y opcional hasta que se obtuvo dicha cadena de texto larga. Con esta perspectiva la compresión de texto consiste en determinar el texto esencial y eliminar el material textual opcional considerado como ruido.

En el modelo *noisy-chanel* se asignan probabilidades a cada cadena candidata ***s***: P(s) con valor bajo si ***s*** no es gramatical; y probabilidades a cada par de cadenas <s, t> P(t | s), la cual proporciona la oportunidad que cuando la cadena corta ***s*** se expande, el resultado es la cadena larga ***t***; lo que se busca es maximizar P(s | t) que equivale a maximizar P(S) P(t| s). Se usan gramáticas libres de contexto probabilísticas para crear los árboles ***t*** y ***s***. Otros enfoques con relación a la compresión de frases usan reglas de rescritura para indicar qué palabras se deben borrar en un contexto determinado como el propuesto en (Knight *et al*., 2002; Cohn *et al*., 2009).

Steinberger *et al*. (2006) presentan un enfoque de compresión basado en el Análisis de Semántica Latente (*Latent Semantic Analysis*, LSA) (Landauer & Dumais, 1997). El método se basa en extraer los posibles candidatos de compresión separando las oraciones completas en cláusulas por medio de un analizador sintáctico. Posteriormente, se transforma cada oración a una representación en espacio semántico donde se puede calcular el puntaje LSA de cada oración de acuerdo al peso combinado a través de todos los tópicos. Basándose en este puntaje, el algoritmo de compresión remueve las clausulas menos importantes de una oración completa. Sus resultados superan a su *línea base*, pero no alcanzan a los valores de compresiones hechas por humanos.

Molina *et al*. (2013) usan una estrategia lingüística basándose en la RST para eliminar segmentos discursivos al interior de las oraciones. Su enfoque se basa en separar las oraciones completas en unidades discursivas elementales (*Elementary Discourse Units*, EDUs), y por medio de la energía textual (Fernández *et al*., 2007) de cada EDU, se determinan las EDUs menos informativas respecto a todo el documento como contexto. Los resultados obtenidos por su método, en comparación con los resultados obtenidos por los humanos para la misma colección de datos, indican que hay una buena correlación entre los segmentos que eliminan los humanos y los que elige el método.

Elementos de información

Genest y Lapalme (Genest & Lapalme, 2011; Genest & Lapalme, 2012) proponen el concepto de elementos de información (I*nformation Items,* InIts) como forma de representación abstracta, la cual la definen como la unidad mínima de información coherente en un texto o una oración. Su representación se orienta a responder a consultas o aspectos de tópicos inducidos generando oraciones que cubran una necesidad de información específica. Un elemento de información se define como una tripleta sujeto–verbo–objeto, por ejemplo, *persona–mata–mujer.*

Para obtener los InIts en un documento, se identifica el sujeto del verbo y su objeto, por medio de los árboles generados por un analizador sintáctico de dependencias. Se filtran los InIts por medio de reglas para rechazar a los que no cumplen con la estructura. Por ejemplo, se rechaza el InIt si es el verbo está en participio o en infinitivo, si el sujeto o el objeto es un pronombre relativo, etc.

A partir de los InIts obtenidos, se generan las nuevas oraciones tomando como base el árbol de dependencias original de las oraciones de donde se extrajeron los InIts. Con este proceso se seleccionan las partes que se desean del árbol de dependencias. Se seleccionan, por ejemplo, los InIts más frecuentes, o los InIts que contienen los más frecuentes pares sujeto–verbo. En este enfoque, la calidad lingüística de los resúmenes generados fue pobre, pero se obtuvieron resultados satisfactorios respecto a la selección del contenido y a la sensibilidad global.

Todos los enfoques anteriores no consideran una representación semántica de los textos de gránulo fino como la que se puede obtener con una representación más expresivas como grafos conceptuales (Sowa J. F., 1984) que aprovechan los roles temáticos como agente, paciente, tema, etc. (Jackendoff, 1972; Fillmore & Atkins, 1992) y forman parte de la estructura semántica del texto representado.

Nuestro interés en esta tesis es explorar el uso de representaciones detalladas que representen mejor a los textos para la generación de resúmenes, que por lo revisado y publicado (véase el punto 6.4) no se han usado para esta tarea.

2.5 Evaluación del resumen

Evaluar la calidad de un resumen resulta una tarea desafiante, y surgen diversas preguntas sobre los métodos y los tipos apropiados para la evaluación.

Existen importantes retos al evaluar los resúmenes (Mani, 2001):

1) La generación del resumen involucra una salida automática de datos que debe estar codifica en una comunicación en lenguaje natural. En los casos donde la salida es una respuesta a una pregunta realizada (área de estudio *Question-Answering*), puede haber una respuesta correcta; pero en otros casos es difícil evaluar si la salida es correcta. Esto es que un sistema generador de resúmenes puede obtener un buen resumen que difiera totalmente de cualquier otro resumen (sobre el mismo grupo de documentos) realizado por un humano, el cual es usado como modelo, considerado como una salida correcta.

2) La evaluación es costosa debido a que se requieren los juicios humanos para evaluar los resúmenes generados por los sistemas, además de ser subjetiva. Por lo que es preferible usar programas que valoren los resúmenes en lugar de usar juicios humanos; además, de esta manera, el procedimiento se puede repetir.

3) La evaluación a diferentes niveles es necesaria, ya que los resúmenes pueden generarse a diferentes razones de compresión. Esto incrementa la complejidad de la evaluación.

4) La presentación del resumen de acuerdo a las necesidades del usuario o aplicación es un factor más que aumenta la complejidad de la evaluación, ya que el resumen es personalizado y el usuario mismo o la aplicación determinará si es útil o no de acuerdo a sus necesidades de información.

Los métodos de evaluación se pueden clasificar en dos tipos: intrínseca y extrínseca (Sparck-Jones & Galliers, 1996). La evaluación intrínseca se ha usado principalmente en las competencias actuales.

La evaluación intrínseca valora las características del resumen mismo; y principalmente se evalúa la coherencia y la informatividad del resumen. Por otro lado, la evaluación extrínseca decide sobre la calidad de un resumen dependiendo de cómo afecta éste a la finalización de otra tarea específica, por ejemplo, cálculo de relevancia en tareas de Recuperación de Información, o compresión de lectura en tareas de preguntas-respuestas, etc.

Debido a la importancia que se aprecia para la tarea de generación de resúmenes, surgió la necesidad de elaborar estrategias para evaluar los trabajos sobre la generación del resumen.

En el año 2001 el NIST (*National Institute of Standards and Technology*) inició la organización del foro DUC[3] (*Document Understanding Conference*) donde se llevaba a cabo una competencia para evaluar los resúmenes generados automáticamente por los sistemas, este foro se realizó hasta el 2007. A partir del 2008 y a la fecha, DUC forma parte del TAC[4] (*Text Analysis Conference*) donde se continúa la evaluación de los sistemas de

[3] http://duc.nist.gov/

[4] http://www.nist.gov/tac/

generación de resúmenes. La competencia es abierta para cualquier equipo que quiera participar.

En general el proceso para evaluar los resúmenes en DUC se puede dividir en tres etapas: preparación, generación y evaluación. En la primera etapa, los organizadores deciden las tareas que se llevarán a cabo y preparan la información con la que se evaluarán los sistemas de generación de resúmenes. En la segunda etapa, los participantes envían los resúmenes generados automáticamente por sus sistemas. En la última etapa, se evalúan los resúmenes enviados y se publican los resultados.

En la competencia DUC, se trabaja con resúmenes multidocumento y monodocumento, aunque no son las mismas tareas para todos los años, cada año se publican las tareas y objetivos del resumen (véase el sitio web para mayores detalles). La generación de resúmenes genéricos fue la tarea principal, pero al paso del tiempo se agregaron otras tareas, incluyendo variedades de resúmenes, por ejemplo, basados en consultas: dado un conjunto de documentos referentes a una persona, la tarea es construir un resumen (tipo biografía) con la información de la persona.

La generación de resúmenes de noticias monodocumento (un solo documento) ya no se continuó realizando en la competencia DUC, debido a que la mayoría de los sistemas de generación de resúmenes no superaban a la "*línea base*", la cual consistía del inicio de los párrafos de los artículos de noticia. De forma similar, para la generación de encabezados (resumen de hasta 10 palabras), ninguno de los sistemas superaba a la línea base la cual utilizaba el encabezado del artículo original.

Para ambas tareas, el desempeño de los humanos, en general, fue significativamente más alto que las *líneas base*. En los experimentos realizados se mostró que el grupo de humanos que generaron los resúmenes son mucho mejor que los sistemas de generación de resúmenes. Tales hechos indican que aunque la generación de resúmenes monodocumento ya no se continuó realizando en los años siguientes de la competencia DUC, sigue siendo un problema abierto (Nenkova A. , 2005; Nenkova *et al*., 2011).

Dada la problemática para evaluar los resúmenes, se ha avanzado poco en el desarrollo de métodos para la evaluación automática de resúmenes. Por otro lado, las métricas

utilizadas en la competencia DUC se centran en la selección del contenido, y se le ha dado poco interés a las investigaciones sobre evaluación automática de la calidad lingüística del resumen. Aunado a esto, los participantes en la competencia DUC se vieron obligados a desarrollar sistemas rápidos de implementar debido a los periodos cortos de tiempo que se tienen para registrarse y participar en la competencia; además, con el fin de ganar la competencia, los participantes implementaban enfoques incrementales y "seguros". Actualmente, la competencia DUC/TAC sigue promoviendo la creación de enfoques seguros, así como el desarrollo de nuevos métodos de evaluación y generación de resúmenes.

En los foros tal como DUC y, ahora, TAC las evaluaciones de los resúmenes generados automáticamente se realizan principalmente por medio de la comparación contra un "estándar de oro" (*gold standard*); esto es, el resumen generado automáticamente se compara contra un modelo del resumen que se considera correcto. El resumen o resúmenes modelo son hechos por humanos. Sin embargo, el objetivo final del desarrollo de sistemas de generación de rúmenes es ayudar al usuario a desempeñar mejor sus tareas basándose en la información que proporcionan los resúmenes generados automáticamente.

Los métodos que se han desarrollado para la evaluación intrínseca y extrínseca se presentan en las siguientes secciones.

2.5.1 Precisión y *Recall*

La mayoría de los sistemas de generación de resúmenes seleccionan oraciones representativas del documento o documentos de entrada para formar un resumen.

Por ejemplo, en el resumen extractivo, las oraciones extraídas se disponen juntas para formar el resumen sin realizar ninguna modificación en la redacción original. Por otro lado, un resumen modelo (creado por un humano) se puede basar en estrategias similares o modificar las oraciones.

Para el escenario anterior, teniendo estos dos resúmenes se pueden aplicar medidas de similitud entre ambos documentos para determinar su semejanza. Si el resumen obtenido

automáticamente es muy similar al resumen modelo, entonces el sistema están bueno al resumir como un humano.

Las principales métricas de evaluación para la selección del contenido son precisión, *recall*[5] y la medida-F.

La evaluación se basa en la comparación de los resúmenes modelos, los cuales los realizan asesores humanos y se consideran correctos, contra los resúmenes que generan los sistemas. Por lo que las métricas se basan en las oraciones seleccionadas por los humanos y las oraciones seleccionadas por los sistemas. De esta manera, las medidas mencionadas quedan definidas de la siguiente forma.

Recall es la fracción de las oraciones que seleccionó el humano y eligió correctamente el método, ecuación (2.1). La **Precisión** es la fracción de las oraciones que el sistema eligió correctas, ecuación (2.2). La **Medida-F** es la media harmónica de la Precisión y *Recall*, ecuación (2.3).

$$Recall = \frac{\text{Traslape de num. oraciones elegidas por el método y el humano}}{\text{Num. oraciones elegidas por el humano}} \quad (2.1)$$

$$\text{Precisión} = \frac{\text{Traslape de num. oraciones elegidas por el método y el humano}}{\text{Num. oraciones elegidas por el método}} \quad (2.2)$$

$$\text{Medida} - \text{F} = 2 \text{ x} \frac{\text{Precisión x } Recall}{\text{Precisión} + Recall} \quad (2.3)$$

El principal problema con las métricas es que personas diferentes tienden a seleccionar diferentes oraciones cuando se les solicita construir un resumen referente a un texto. Por lo que se recomienda usar un conjunto de resúmenes modelo del mismo documento para reducir el error.

2.5.2 Evaluación intrínseca

La evaluación intrínseca, como mencionamos, valora las características del resumen mismo. Esta clase de evaluación es la que, generalmente, se usa ya sea para comparar a los

[5] Se usa el término *recall* (del inglés) debido a que no hay un término comúnmente aceptado en español.

sistemas de generación de resúmenes contra los resúmenes modelo considerados correctos, o para determinar si un resumen cumple con los juicios humanos de calidad y utilidad.

En el caso del uso de resúmenes modelo, se espera que las evaluaciones sean realizadas automáticamente y se reduzca la participación humana.

Evaluación manual

La evaluación manual realizada por los jueces humanos usualmente se centra en dos aspectos de la calidad del resumen: el contenido y la forma.

SEE (Summary Evaluation Environment)

En la competencia DUC del 2001 al 2004, los resúmenes se evaluaron manualmente tanto para el contenido como para la legibilidad. Para evaluar el contenido, cada resumen (llamado *peer*) se comparó contra un resumen modelo (llamado *model*) hecho por un humano. Se usó el software SEE, el cual era una interfaz para el asesor que evaluaba el resumen *peer* contra el resumen modelo para determinar la cantidad de información que el resumen *peer* cubría del resumen modelo. La calidad se evaluaba en términos de gramaticalidad, cohesión, coherencia para cada unidad segmentada (oraciones, clausulas, frases, etc.). Se tenían cinco niveles de calidad (*all*, *most*, *some*, *hardly any*, o *none*), con los cuales el asesor podía elegir.

Método de la pirámide

La suposición fundamental del método de la pirámide (Nenkova & Passonneau, 2004) es que se requieren múltiples modelos humanos, los cuales todos juntos conforman un *gold standard* para la salida del sistema. En este método, se realiza un procedimiento manual para la identificación de las equivalencias semánticas de los resúmenes, lo que permite una variabilidad en la granularidad del análisis. Tales equivalencias son unidades de significado las cuales las llaman SCUs (*Summary Content Units*).

El análisis en el método es semánticamente dirigido, esto es, información con el mismo significado, incluso, cuando se expresa con diferentes palabras en los diferentes

resúmenes, se marcan como unidades con el mismo contenido de resumen (SCUs). La pirámide representa las opiniones comunes de múltiples escritores de resúmenes, es decir, las SCUs representan el acuerdo entre los resúmenes humanos. Las SCUs que aparecen más veces en los resúmenes humanos, se les asigna un peso mayor lo que permite identificar el contenido importante.

La desventaja de este método es que requiere de un trabajo humano para identificar las SCUs en los documentos.

Evaluación automática

La etapa de evaluación del resumen en los sistemas de generación de resúmenes es muy importante, ya que permite identificar errores y reformular algunos aspectos del proceso para optimizarlo.

Actualmente, la evaluación automática de resúmenes ha sido dirigida sólo superficialmente, debido a que muchas de las cualidades requeridas por el resumen no pueden medirse automáticamente. Por lo que jueces humanos se han utilizado ampliamente para evaluar o revisar el proceso de generación del resumen (Dang, 2005; Elhadad, Miranda-Jiménez, Steinberger, & Giannakopoulos, 2013).

En los siguientes apartados, presentamos los métodos comúnmente usados en las competencias de sistemas para la generación de resúmenes.

ROUGE

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Chin-Yew, 2004) es el método más popular usado para la evaluación de resúmenes debido a su simplicidad y alta correlación con los juicios humanos, se usa en la competencia TAC. Sus métricas comparan el resumen candidato contra un resumen o conjunto de resúmenes modelo creados manualmente por personas.

El método se basa en la coocurrencia de n-gramas entre los resúmenes (candidato y modelo), secuencias de palabras y pares de palabras. ROUGE-N es una métrica orientada a *recall* debido a que se basa en la suma total de las ocurrencias de los n-gramas (secuencias

de palabras, n representa la cantidad de secuencias; por ejemplo, 2-grama, secuencias de dos palabras) o *skip* n-gramas (n-gramas con saltos, es decir, no secuenciales) entre los resúmenes candidato y modelo. El método hace un análisis superficial de las palabras aunque tiene una variante BE (*Basic Elements*) la cual hace uso de la información sintáctica de las relaciones de dependencia entre los elementos del texto.

AutoSummENG

AutoSummENG (*AUTOmatic SUMMary Evaluation based on N-gram Graphs*) (Giannakopoulos & Karkaletsis, 2010; Giannakopoulos & Karkaletsis, 2011) es un método que intenta implementar las siguientes características deseables para los métodos de evaluación automática:

- **Neutralidad del lenguaje**. Un método que no requieran de recursos dependientes del lenguaje (tesauros, diccionarios léxicos, etc.) y se pueda aplicar directamente a diferentes lenguajes.
- **Automatización completa**. Un método que no requiera de intervención humana, fuera de los resúmenes modelo que generan los humanos.
- **Sensibilidad al contexto**. Un método que considere la información contextual, con el fin de que los textos "bien formados" se consideren. Se puede decir que un texto bien formado describe la calidad del texto, lo que permite una fácil lectura. Un texto con una secuencia aleatoria de palabras carecen de esta calidad, incluso si las palabras se refieren al tópico en cuestión.

El método se basa en extraer estadísticamente información textual de los resúmenes, la información se integra en una representación enriquecida para evaluar la similitud entre los resúmenes generados y el conjunto de los resúmenes modelo.

La información extraída de los textos fuente es un conjunto de indicadores de la vecindad entre los n-gramas contenidos dentro de éstos. Esto es, se extraen las relaciones entre los n-gramas. El método considera dos tipos de representaciones de n-gramas: palabras y caracteres. Además, puede usar o no usar la información de conectividad entre

los datos, esto es, representación como grafo o como histograma. El grafo que se construye usa la proximidad de los n-gramas.

Para llevar a cabo la evaluación, se realiza una comparación entre la representación del grafo del resumen generado automáticamente y el del modelo. Lo que se obtiene es una medida de similitud entre los grafos. En este contexto, los resúmenes generados automáticamente que se parecen más a los resúmenes modelo se consideran mejores.

El método se aplicó a la competencia TAC 2010 y 2011 para la tarea de generación de resúmenes multilingüe.

FRESA

FRESA (FRamework for Evaluating Summaries Automatically) (Torres-Moreno, Saggion, da Cunha, SanJuan, & Velázquez-Morales, 2010; Saggion, Torres-Moreno, da Cunha, & SanJuan, 2010) es un marco de trabajo para la evaluación de resúmenes, el cual incluye medidas de evaluación de resúmenes de documentos basadas en distribución de probabilidades, específicamente la divergencia de Jensen-Shanon. Similar a ROUGE, FRESA soporta diferentes probabilidades de n-gramas y *skip* n-gramas para el cálculo de las divergencias.

ROUGE usa modelos de resúmenes (creados por humanos) para poder evaluar la calidad del resumen generado por un sistema. En cambio, FRESA toma como modelo al texto original (puede usarse también un resumen de referencia), lo que no requiere de la intervención del humano, y lo compara contra el resumen obtenido automáticamente. Los autores puntualizan que no siempre es recomendable usar los documentos originales como modelos, ya que se obtienen correlaciones débiles en tareas complejas de resúmenes como la generación de resúmenes de opinión sobre una entidad y de información biográfica.

Esta herramienta que todavía no es tan popular como ROUGE, pero puede orientar y dar guía a las investigaciones sobre la evaluación de resúmenes donde no se cuentan con los modelos de resúmenes para evaluar los sistemas.

2.5.3 Evaluación extrínseca

La evaluación extrínseca decide sobre la calidad de un resumen dependiendo de la efectividad de usarlo en una tarea específica. Las principales tareas donde se ha usado el resumen son en la Recuperación de Información, clasificación de documentos y preguntas y respuestas.

Un ejemplo de este tipo de evaluación es el realizado en la evaluación de SUMMAC (Mani, Klein, House, & Hirschmans, 2002). La tarea se enfocó en los resúmenes indicativos, donde se realizaban búsquedas con el texto completo usando un sistema de recuperación de información para determinar la relevancia de los documentos recuperados. Esto es, dado un documento y la descripción de un tópico, se le preguntaba a una persona para determinar si el documento era relevante para el tópico.

Para el caso de la tarea de clasificación, la evaluación buscaba encontrar si un resumen indicativo genérico podría presentar la suficiente información que permita un análisis para la clasificación correcta de un documento. En este caso, el tópico no era conocido y el sujeto humano podía escoger una categoría de cinco para la cual el documento era relevante, cada una de ellas tenía una descripción del tópico asociada.

En el caso de la tarea de preguntas y respuestas, Se midió cuántas de las preguntas un individuo responde correctamente bajo condiciones diferentes. Primeramente, se les muestran los pasajes originales, después un resumen generado automáticamente, a continuación, un resumen generado por una persona profesional en generación de resúmenes; por último, los individuos deben seleccionar las respuestas correctas de acuerdo a la información presentada. Si eran capaces de responder las repuestas correctamente, entonces el material presentado era útil para esta tarea.

2.6 Algoritmos de ponderación basados en grafos

Las estructuras de grafos se han usado para determinar las partes sobresalientes del texto. También se ha usado los grafos para representar las relaciones léxicas o estructuras retóricas.

Una ventaja de representar los textos en estructuras en forma de grafos es que se puede aplicar algoritmos y métodos conocidos en el campo estudio de la teoría de grafos y las redes.

En el contexto de representación del texto como grafo, se han propuesto varios enfoques para ponderar las oraciones. Entre los algoritmos que adaptan algunos métodos de ponderación para páginas web como el PageRank (Page *et al*., 1998) y HITS (Kleinberg, 1999), se encuentra el TextRank (Mihalcea *et al*., 2004), LexRank (Erkan *et al.*, 2004b) o SemanticRank (Tsatsaronis, Varlamis, & Nørvåg, 2010) entre otros; donde el grafo se construye agregando un vértice por cada oración o término del texto, y las aristas entre los vértices se establecen usando las interconexiones entre las oraciones. Las conexiones se definen usando una relación de similitud, usualmente se usa una función del traslape del contenido; dicho traslape se realiza por medio de los unidades léxicas comunes entre dos oraciones, o se usa una jerarquía de conceptos para identificar la semántica entre los términos.

La mayoría de los algoritmos de ponderación basados en grafos usan algoritmos conocidos para determinar la importancia de los nodos. Entre los algoritmos más populares, y que se emplean para el área de Recuperación de Información son los métodos PageRank y HITS.

2.6.1 PageRank

PageRank (Page *et al*., 1998; Page, Brin, Motwani, & Winograd, 1999) es un método para determinar el prestigio de las páginas web, el cual se basa en el voto que emite una página hacia otra por medio de los enlaces que posee. Por ejemplo, una página A apunta a una página B, entonces A vota por la página B, pero no solamente se considera esto, sino que además transmite la importancia de la página que emite el voto, esto es, si las páginas que emiten un voto son importantes, entonces las páginas por las que se vota también debe ser importantes. En la figura 2.3, se muestra el esquema general en el que se basa este enfoque.

Esta medida se calcula recursivamente para todas las páginas y depende del número y la métrica misma del PageRank de las páginas a las cuales se enlaza la página analizada. La ecuación 2.4 describe el modelo PageRank.

$$\mathrm{PR}(V_i) = (1-d) + d \sum_{k \in B(V_i)} \frac{\mathrm{PR}(V_k)}{\mathrm{L}(V_k)} \qquad (2.4)$$

Donde B(V_i) es el conjunto de páginas que apuntan hacia la página analizada V_i. L(V_k) representa al conjunto de páginas a las que apunta V_k. '*d*' es un factor de amortiguamiento que puede estar entre 0 y 1, generalmente, se establece en 0.85 (Page & Brin, 1998)

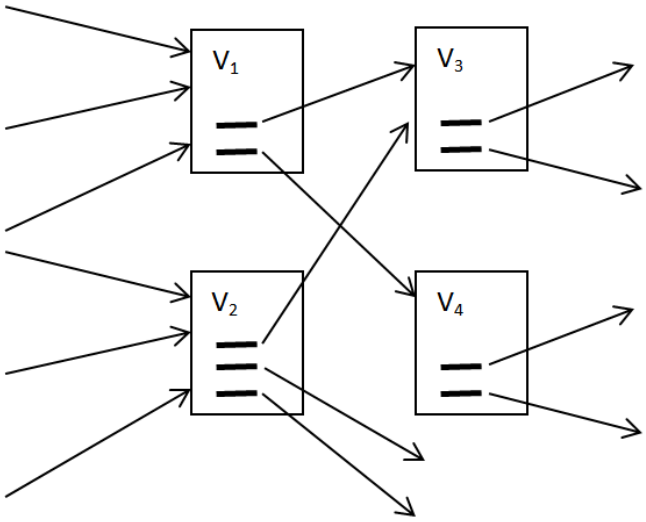


Figura 2.3 Esquema básico de relaciones entre páginas para el cálculo de PageRank

Para este tipo de algoritmos se establecen los valores iniciales de PageRank arbitrariamente y se recalculan recursivamente hasta que el algoritmo converja o hasta un determinado número de iteraciones. Después de terminar de ejecutar el procedimiento, cada nodo tiene asociada una puntuación, la cual representa la importancia del nodo dentro del grafo. Los nodos con mayor puntuación son los nodos considerados más importantes.

2.6.2 HITS

Otro algoritmo popular para determinar la importancia, en el contexto de las páginas web, es el método HITS (Kleinberg, 1999). Se basa en el cálculo de dos métricas para cada página web: autoridad y concentración. En la figura 2.4, podemos ver el esquema de relación entre los concentradores y autoridades.

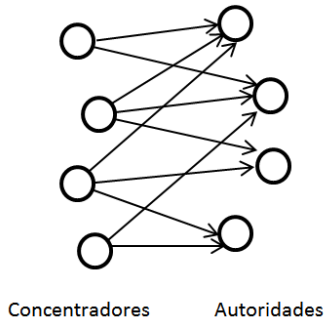


Figura 2.4. Esquema del modelo autoridad-concentrador

El modelo de autoridad-concentrador es una relación de reforzamiento mutuo. Un buen concentrador es una página que apunta a muchas páginas con buena autoridad; y una buena autoridad es una página a la cual apuntan muchas páginas que son buenas concentradoras.

Para determinar la importancia se usan las ecuaciones (2.5) y (2.6). La figura 2.5 ilustra el cálculo de cada una de las métricas.

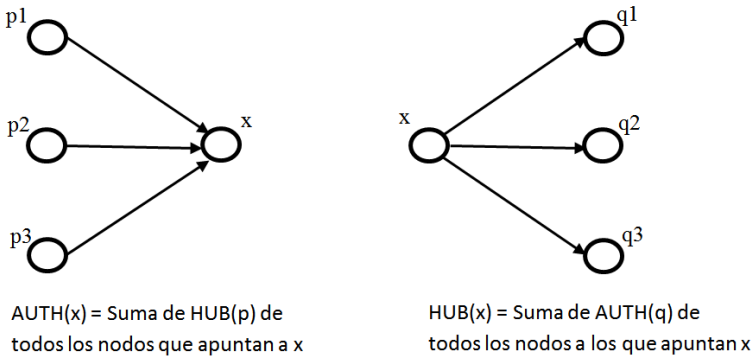


Figura 2.5. Operaciones para el cálculo de métricas HITS

$$AUTH(P) = \sum_{q \in In(P)} HUB(q) \qquad (2.5)$$

$$HUB(P) = \sum_{q \in Out(P)} AUTH(q) \qquad (2.6)$$

Donde *In*(*P*) representa el conjunto de páginas que apuntan a la página *P*; *Out*(*P*) representa el conjunto de las páginas a las que apunta *P*.

AUTH(*P*) y *HUB*(*P*) representan la puntuación de autoridad y concentración para la página P.

La ventaja de este método es que calcula dos tipos de importancia: 1) que tan bueno es el nodo como fuente de información, métrica AUTH; y 2) que tan bueno es el nodo para obtener información por medio de él, métrica HUB.

2.6.3 TextRank

La idea básica en los algoritmos de ponderación es recomendación o voto entres los nodos. El modelo TextRank usa esta idea aplicada a los textos. Este modelo se aplicó a dos tareas: la extracción de oraciones y la extracción de palabras clave.

En el contexto de la navegación entre las páginas web, es inusual usar enlaces múltiples o parciales hacia otra página, por lo que el modelo original del PageRank para la ponderación basada en grafos se considera sin pesos.

Una aportación importante en el TextRank es la asignación de pesos al modelo original de PageRank y HITS, ya que en el contexto de texto representado como grafo es factible considerar la "fuerza" de conexión entre los nodos como un peso que se anexa a la arista que los conecta.

Para poder usar los algoritmos de ponderación basados en grafos, se debe transformar el texto a una estructura en forma de grafo. Los nodos del grafo se puede construir a diferentes niveles, dependiendo de la aplicación: palabras, colocaciones, oraciones completas o cualquier otra unidad. De igual manera, dependiendo de la aplicación, son las relaciones que se establecen entre cualquier par de vértices, por ejemplo, relaciones léxicas, semánticas, traslapes contextuales, etc.

En los experimentos para la extracción de oraciones se utilizó una medida de similitud entre las oraciones de traslape del contenido. Los resultados obtenidos demostraron que es factible el uso de algoritmos de ponderación para el análisis de textos (Mihalcea & Tarau, 2004; Mihalcea & Tarau, 2005).

2.6.4 LexRank

LexRank (Erkan *et al*., 2004b), similar al método TextRank, se basa en la idea de identificar el prestigio o importancia de un nodo, concepto usado en las redes sociales.

Una red social es un mapeo de relaciones entre entidades relacionadas (personas, organizaciones, etc.). En la red social los nodos representan las entidades y los enlaces representan las relaciones.

De manera análoga a la red social, un conjunto de documentos se pueden ver como una red de oraciones que están relacionadas entre ellas. Algunas oraciones están más relacionadas que otras.

Para identificar la importancia de los oraciones (nodos), se calcula la similitud entre las oraciones, para ello, se usa un modelo de espacio vectorial y la medida coseno similar a al enfoque presentado por Salton (Salton *et al*., 1997), y para el cálculo de los pesos de los términos se usa principalmente una medida TF*IDF. Al final de la construcción del grafo se aplica el algoritmo PageRank para identificar los nodos más importantes. Los nodos más importantes representan la centralidad de los documentos, es decir, los tópicos que tratan los documentos.

2.6.5 SemanticRank

SemanticRank (Tsatsaronis *et al*., 2010) es un método de ponderación para segmentos de textos (términos u oraciones) representados como nodos y las aristas representan la relación semántica entre los nodos; también considera información estadística. Al igual que otros métodos, con esta representación usan algoritmos estándar como los mencionados PageRank, HITS o alguna variante de estos. Lo novedoso del método es que para la ponderación de los nodos usa una métrica de relación semántica entre los segmentos de texto, la cual se basa en WordNet (Fellbaum, 1998) y Wikipedia[6]. La motivación principal es que los grafos semánticos generados podrían capturar la similitud respecto al significado entre los vértices del grafo.

[6] www.wikipedia.org

Para determinar la relación semántica entre las oraciones usan un modelo de bolsa de palabras en un espacio vectorial, y para el pesado de los términos se usa el procedimiento estándar TF*IDF. Adicionalmente, se aprovechan las rutas de las relaciones entre términos que se encuentran en la estructura de WordNet.

Aunado a esto, para mejorar la cobertura se combina con una métrica que usa artículos de Wikipedia y su estructura de enlaces como base de conocimiento. Se basa en la idea de la cantidad de artículos similares que apuntan a ambos términos, entre mayor sea el número de artículos entonces mayor será su similitud semántica.

Finalmente para determinar la importancia semántica de los términos u oraciones, se usan variantes de los algoritmos PageRank y HITS.

Capítulo 3

Marco teórico

En este capítulo se presentan los conceptos básicos necesarios para la comprensión de esta investigación.

Se presenta la importancia de la representación del conocimiento, es decir, la representación de los textos en una notación manipulable por las computadoras. Se da énfasis al formalismo en que se basa esta investigación, a saber, grafos conceptuales, la cual es la representación intermedia que se utiliza para describir los textos.

3.1 Estructuración del conocimiento

Un formalismo de representación del conocimiento debe proporcionar la forma de estructurar el conocimiento. Un aspecto importante para la estructuración del conocimiento es que la información semánticamente relacionada debe de estar reunida y junta. Esta consideración se aplicó a los primeros formalismos para representar el conocimiento: *Frames* y *Redes Semánticas*, pero que distaban de una semántica formal.

Los *frames* son estructuras de datos tipo registro que representan situaciones prototípicas y objetos. La idea clave fue agrupar la información relevante en la misma situación/objeto. En principio, se basan en que la memoria humana trabaja con estereotipos; éstos se adaptan a cada situación, y la información faltante se complementa para representar al mundo real. Los *frames* están puestos en una jerarquía de herencia donde los *frames* más específicos heredan las propiedades de los *frames* más generales, como en la programación orientada a objetos. Por medio de esta jerarquía se pueden hacer deducciones.

Por su parte, las redes semánticas se desarrollaron originalmente como modelos cognitivos y para el procesamiento semántico del lenguaje natural.

3.2 Redes semánticas

Una red semántica es un diagrama que representa relaciones entre objetos o conceptos para algún dominio de conocimiento específico. Los enlaces básicos son el enlace *IS-A* (es un) que relaciona un objeto y su concepto (*mesa* IS-A *mueble*, la mesa es un mueble), el enlace *A-Kind-Of* (es un tipo de) que relaciona dos conceptos (*gato A-Kind-Of mamífero*, un gato es un tipo de mamífero), uno es un tipo del otro, y el enlace *PROPERTY* que asigna una propiedad a un objeto o concepto (*visión* PROPERTY *gato*, la visión es una propiedad que poseen los gatos).

Una forma en que se heredan las propiedades a los objetos es por medio de las rutas que se forman con los enlaces *IS-A* y *A-Kind-Of* de la red que se forma.

Una de las críticas principales de las redes semánticas es su falta de semántica formal, ya que una misma red se puede interpretar de diferentes formas dependiendo de la compresión intuitiva del usuario o de su representación gráfica (Chein & Mugnier, 2009).

Las redes *IS-A* son muy flexibles, pero los investigadores del campo de la inteligencia artifical han señalado algunos problemas y desventajas importantes.

1) La elección de nodos y arcos es crucial en la fase de análisis. Una vez que se ha decidido una estructura determinada es muy complicado cambiarla.

2) Dificultad para expresar cuantificación. Por ejemplo, en expresiones tales como *algunos pájaros vuelan* o *todos los pájaros vuelan*.

Del mismo modo, las redes semánticas presentan grandes dificultades para representar la "intensionalidad". Por ejemplo, en proposiciones tales como *Juan cree que María desea casarse con un político* (véase la sección 3.3.5 Contextos). Esto llevó a idear otros esquemas de representación con una estructura más flexible que fuesen capaces de dar cabida a éstas y otras situaciones.

Una solución a los problemas antes mencionados son los grafos conceptuales propuestos por Sowa (Sowa J. F., 1984). Esta representación es más expresiva y flexible, por ejemplo, al definir contextos a diferencia de otras representaciones como KL-ONE o lógica de predicados (Sowa J. F., 1999; Chein & Mugnier, 2009).

Los grafos conceptuales son la culminación de varias direcciones en ciencia cognitiva, redes semánticas, procesamiento de lenguaje natural y lógica formal. Los grafos conceptuales son propuestos como una representación de conocimiento general para ayudar a resolver problemas de comprensión del lenguaje natural, toma de decisión humana, modelado conductual, y otras áreas afines.

3.3 Grafos conceptuales

En esta sección se presentan los conceptos básicos para la representación de oraciones en el formalismo de grafos conceptuales, estos temas fueron tomados principalmente de Sowa

(Sowa J. F., 1984; Sowa J. F., 1991; Sowa J. F., 1999), Chein y Mugnier, quienes han extendido la teoría de Sowa (Chein & Mugnier, 2009), y Montes-y-Gómez *et al.* (2000a; 200b; 2001), quienes aplicaron esta teoría al área de Recuperación de Información.

Los grafos conceptuales (GCs) son un esquema de representación del conocimiento muy poderoso y expresivo, el cual hereda los beneficios de la lógica y de los grafos de la matemática. Están basados en las redes semánticas y los grafos existenciales de C.S. Peirce.

Los GCs tienen dos propósitos principales: pueden usarse como representaciones canónicas de significado en un lenguaje natural; y como bloques de construcción para formar estructuras abstractas que sirvan como modelos.

Los grafos conceptuales propuestos por Sowa proporcionan los medios para representar semánticamente varios aspectos de los lenguajes naturales. Informalmente, un grafo conceptual es una estructura de conceptos y de relaciones entre éstos, donde un arco enlaza una relación conceptual ***r*** y un concepto ***c***. Los conceptos se representan mediante una forma cuadrada y las relaciones conceptuales por un óvalo.

En la figura 3.1 se muestra un ejemplo de un grafo conceptual, este grafo representa la oración: *El gato Tom está sobre el tapete azul.*

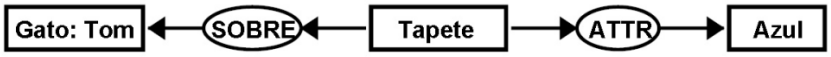


Figura 3.1. Ejemplo de grafo conceptual

La representación de la figura 3.1 se puede transformar a fórmulas del cálculo de predicados, siendo la siguiente forma su representación:

$$(\exists x{:}Tom)\ (\exists y{:}Tapete)(\exists z{:}Azul)\ Gato(x)\ sobre(x,y)\ Attr(y,z)$$

También pueden representarse en forma lineal:

$$[Gato{:}Tom] \rightarrow (sobre) \rightarrow [tapete] \rightarrow (Attr) \rightarrow [azul]$$

3.3.1 Definición

Formalmente un grafo conceptual (GC), ***g***, es un grafo bipartito que posee dos tipos de nodos: *conceptos* y *relaciones conceptuales* (Sowa J. F., 1984).

1. Cada arco ***a*** de ***g*** debe enlazar una relación conceptual ***r*** en ***g*;** o a un concepto ***c*** en ***g***
2. El GC ***g*** puede tener conceptos que no estén enlazados a ninguna relación conceptual, pero cada arco que pertenece a cualquier relación conceptual en ***g*** debe estar unida exactamente a un concepto en ***g***
3. Existen tres tipos de GCs:
    a. *Vacío* es un GC sin elementos: conceptos, arcos o relaciones conceptuales.
    b. *Singleton* en un GC que consiste de un solo concepto sin relaciones conceptuales ni arcos.
    c. *Estrella* es un GC que consiste de una sola relación conceptual y los conceptos que se unen por sus arcos.

En la figura 3.1, se mostró un ejemplo de un grafo conceptual.

3.3.2 Concepto

Los conceptos representan entidades, acciones y atributo. Cada concepto tiene un tipo ***t*** y un referente ***r***. El tipo conceptual indica la clase del elemento representado por el concepto, mientras que el referente indica el elemento específico (instancia de la clase) referido por éste. En la figura 3.1 el concepto definido como [*Gato:Tom*], el tipo es *Gato* y el referente es *Tom*.

De esta manera, se establece una noción de jerarquía de tipos, en la cual existe una *función tipo* que proyecta conceptos sobre un conjunto cuyos elementos son *etiquetas de tipo*, es decir, los referentes. En este contexto, todos los grafos conceptuales deben tener inherentemente una jerarquía de tipos conceptuales y de relaciones.

Referentes

Los referentes son de dos tipos genéricos e individuales. Los referentes genéricos denotan conceptos no específicos; y los referentes individuales denotan sustitutos de elementos específicos en un contexto real. Por ejemplo, el concepto [*Tapete*] denota un tapete, y el concepto [*Gato:Tom*] es un sustituto de *gato Tom,* el cual existe en algún sitio.

En la siguiente tabla, se muestran algunas notaciones estándar empleadas para los referentes, la descripción completa se puede consultar en (Sowa, 1984; 1999).

Tabla 3.1 Notación para conceptos

Notación	**Descripción**
[Gato]	Un gato
[Gato: *x]	Un gato x
[Gato: #]	El gato
[Gato: {*}]	Unos gatos
[Gato: {*}@3]	Tres gatos
[Gato: {Tom,Galore}]	Unos gatos llamados Tom y Galore
[Longitud:@15cm]	Longitud de 15 centímetros

Jerarquía de tipos conceptuales

Una jerarquía de tipos ***T*** es un conjunto parcialmente ordenado cuyos elementos son llamados ***etiquetas de tipo***. Cada etiqueta de tipo en ***T*** se especifica como *primitiva* o *definida.* En la jerarquía de tipos conceptuales se consideran los siguientes puntos.

1. La jerarquía ***T*** contiene dos tipos de etiquetas primitivas $\top$ (tipo universal) y $\bot$ (tipo absurdo). Esto es debido a que la jerarquía de tipos se considera un *lattice* (o retículo) y para cualquier par de elementos de la jerarquía existe un supremo (tipo universal) y existe un ínfimo (tipo absurdo), dado que es un conjunto parcialmente ordenado.
2. Una etiqueta de tipo definida y su definición son intercambiables.
3. El ordenamiento parcial sobre la jerarquía ***T*** está determinado por las relaciones:
    a. *subtipo,* $\leq$
    b. *subtipo propio,* $<$

c. *supertipo,* ≥

d. *supertipo propio,*>

Si ***t*** es una etiqueta de tipo, entonces $\top \geq t$ y $t \geq \bot$

Sea ***s*** y ***t*** etiquetas de tipo, si ***s*** ≤ ***t*** entonces ***s*** es el subtipo de ***t***.

En la figura 3.2, de acuerdo a la jerarquía, *Hombre* es un subtipo de *Persona*, es decir, *Hombre* ≤ *Persona.*

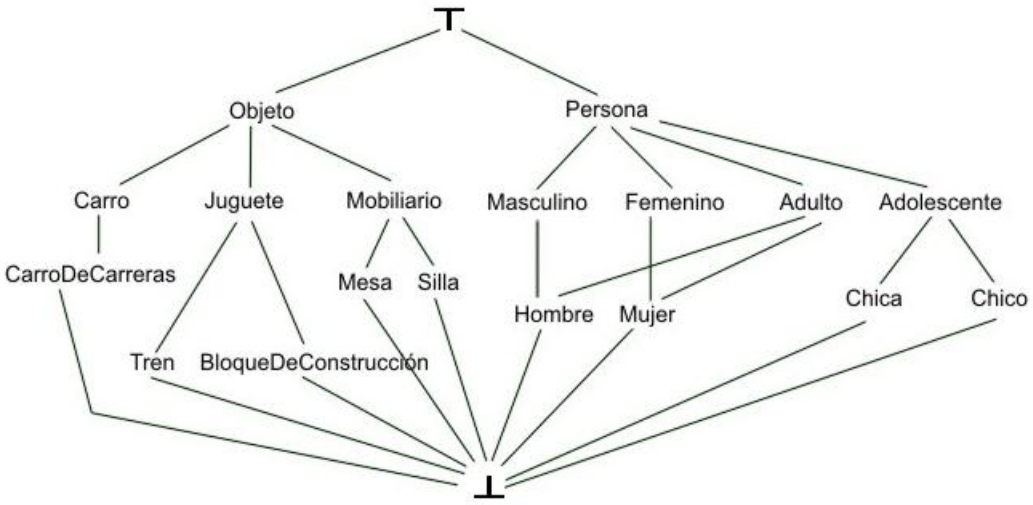


Figura 3.2. Jerarquía de tipos conceptuales

3.3.3 Relación conceptual

Cada relación conceptual ***r*** tiene un tipo ***t*** y un número no negativo ***n*** que representa la valencia. Esta valencia representa el número de arcos que puede tener dicha relación conceptual.

Al igual que los conceptos, se establece una noción de jerarquía de tipos, en la cual existe una función tipo que proyecta las relaciones sobre un conjunto cuyos elementos son etiquetas de tipo, es decir, referentes. En este contexto, todos los grafos conceptuales deben tener inherentemente una jerarquía de tipos conceptuales y de relaciones.

Jerarquía de tipos relacionales

Como se mencionó, toda relación conceptual *r* tiene un tipo relacional ***t*** y una valencia ***n*** (no negativa), en la cual el número de arcos que pertenecen a ***r*** es igual a la valencia ***n***; y toda relación conceptual del mismo tipo tiene la misma valencia.

De igual forma, existe la noción de jerarquía de tipos relacionales.

Una jerarquía de relación ***R*** es un conjunto parcialmente ordenado cuyos elementos son llamados ***etiquetas de relación***. Cada etiqueta de tipo en ***R*** se especifica como *primitiva* o *definida*.

1. Para cada etiqueta de relación ***r*** en ***R*** existe un entero positivo **n** que se llama *valencia.*
2. El ordenamiento parcial sobre la jerarquía ***R*** está determinado por las relaciones:
    a. *subtipo,* ≤
    b. *subtipo propio,* <
    c. *supertipo,* ≥
    d. *supertipo propio,*>

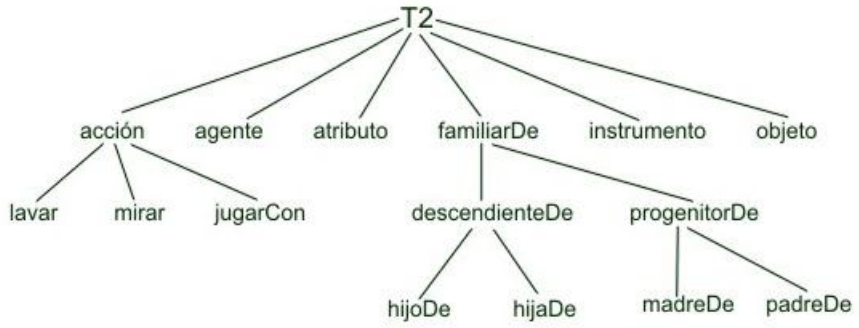


Figura 3.3. Jerarquía parcial de tipos relacionales

En la figura 3.3, *hijoDe* es un subtipo de *descendienteDe*, esto es, *hijoDe* ≤ *descendienteDe.*

3.3.4 Razonamiento con grafos

Todas las operaciones de los grafos conceptuales se basan en alguna combinación de las seis reglas canónicas de formación (núcleo de la teoría de grafos conceptuales). Cada una de estas reglas realiza una operación básica sobre los grafos conceptuales.

Por ejemplo, algunas de estas reglas los hacen más específicos, otras los generalizan, y otras únicamente cambian su forma pero los mantienen lógicamente equivalentes. El método para la generación de resúmenes propuesto en los siguientes capítulos se basa en cuatro operaciones sobre un conjunto de gafos conceptuales. En esta sección sólo se analizan las reglas canónicas.

Cada regla tiene uno de tres posibles efectos sobre las relaciones lógicas sobre un grafo inicial ***g*** y el grafo resultante ***h***

Equivalencia. Copiar y simplificar son reglas equivalentes, las cuales generan un grafo ***h*** que es lógicamente equivalente al original ***g***

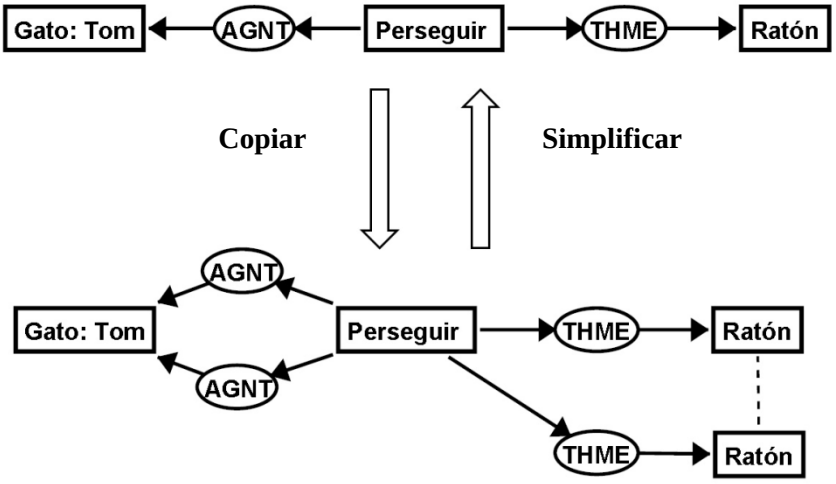


Figura 3.4. Reglas de equivalencia

En la figura 3.4 el grafo superior representa *El gato Tom está persiguiendo a un ratón*. La flecha hacia abajo representa dos aplicaciones de la regla copiar. Una copia el agente

(AGNT) y otra el Tema (THME). La flecha hacia arriba representa dos aplicaciones que simplifican el grafo por medio de la regla *simplicar*.

Especialización. Unir y Restringir son reglas para especializar, generan un grafo ***h*** que implica al original ***g*.**

En la figura 3.5, el grafo superior representa la oración: *Un gato está persiguiendo a un animal*. Después de realizar dos aplicaciones al grafo, éste se transforma en el grafo inferior, el cual puede leerse como *El gato Tom está persiguiendo a un ratón*. De estas reglas se deduce que un grafo más especializado implica a un grafo más general.

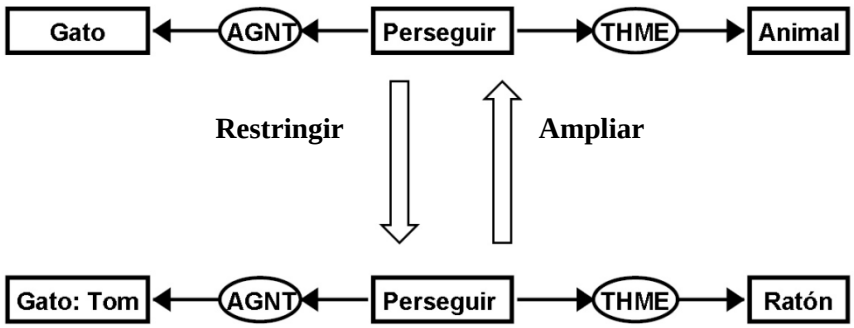


Figura 3.5. Reglas de especialización

Generalización. Unir y Separar son reglas de generalización. Un grafo generado ***h*** se implica por su original ***g***.

En la figura 3.6, hay dos oraciones: *Tom está persiguiendo a un ratón* (grafo superior del lado izquierdo) y *un ratón es café* (grafo superior del lado derecho). Las reglas de unión fusionan las dos copias del concepto [*Ratón*] y se forma un solo grafo conceptual, el grafo inferior, el cual puede leerse como *El gato Tom está persiguiendo a un ratón café*. Con la operación inversa (Separar) se obtienen nuevamente los grafos originales.

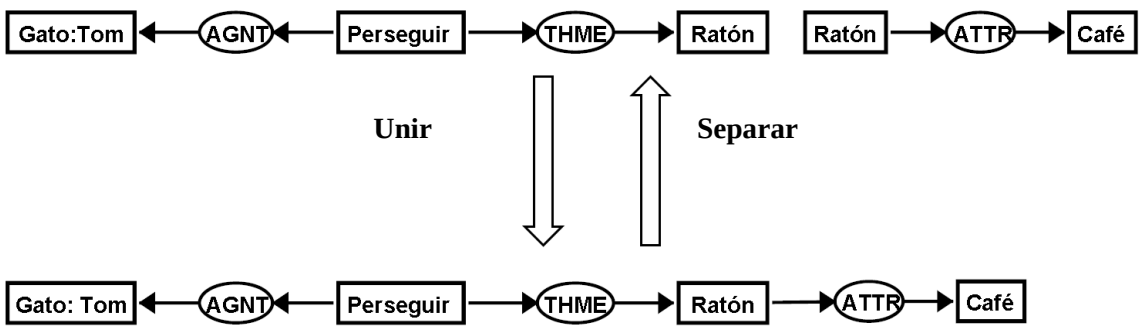


Figura 3.6. Reglas de generalización

La operación de generalización define un ordenamiento parcial de los grafos conceptuales conocido como *jerarquía de generalización*. Entonces si *u*, *v* y *w* son grafos conceptuales de esta jerarquía, las siguientes propiedades siempre son verdaderas:

- Reflexividad: $u \leq u$.
- Transitividad: si $u \leq v$ y $v \leq w$, entonces $u \leq w$.
- Antisimetría: si $u \leq v$ y $v \leq u$, entonces $u = v$.
- Subgrafo: Si *v* es un subgrafo de *u*, entonces $u \leq v$.

Además si *v* es una generalización de *u* ($u \leq v$), entonces debe de existir un subgrafo *u*′ inmerso en *u* que represente el grafo *v*. Este subgrafo *u*′ es llamado *proyección* de *v* en *u*.

Para dos grafos conceptuales cualesquiera *u* y *v*, siendo $u \leq v$, debe de existir un "mapeo" π: $v \rightarrow u$, donde πv es un subgrafo de *u* llamado proyección de *v* en *u*. Algunas propiedades de la proyección son:

- Para cada concepto *c* de *v*, πc es un concepto en πv, para el cual $type(\pi c) \leq type(c)$; y si *c* es un concepto individual, entonces también $referent(\pi c) = referent(c)$.

  Donde *type* representa a la función tipo, y *referent* representa a la función de etiquetas de tipo, es decir, a los referentes.

- Para cada relación conceptual r de v, πr es una relación conceptual en πv, para la cual $type(\pi r) = type(r)$. Esto implica que si el i-ésimo arco de r está conectado al concepto c, entonces el i-ésimo arco de πr debe de estar conectado a πc en πv.

Finalmente, si u_1, u_2 y v son grafos conceptuales, y $u_1 \leq v$ y $u_2 \leq v$, entonces v es una *generalización común* de u_1 y u_2. El grafo conceptual v es la *máxima generalización común* de u_1 y u_2, si y sólo si, no existe otra generalización común v' de u_1 y u_2 ($u_1 \leq v'$ y $u_2 \leq v'$), tal que $v' \leq v$.

Las propiedades anteriores son los mecanismos que proporciona el formalismo de grafos conceptuales para inferir/deducir nueva información a partir del conocimiento del mundo proporcionado por medio de las jerarquías de conceptos y relaciones conceptuales.

3.3.5 Semejanza entre grafos conceptuales

Un mecanismo importante en cualquier marco de trabajo es la forma en cómo se comparan los objetos que se definen, en nuestro caso, grafos conceptuales.

Se han propuesto algunas formas para comparar grafos conceptuales, las cuales se han usado en Recuperación de Información con resultados favorables en ese campo (Montes-y-Gómez *et al*., 2000a;2001).

Para medir la semejanza entre dos grafos conceptuales usamos las medidas propuestas por Montes-y-Gómez *et al.*,a saber, las medidas de semejanza conceptual y semejanza relacional. Tales medidas de semejanza se describen a continuación.

Dados dos grafos conceptuales G_1 y G_2, y uno de sus traslapes (elementos comunes), denotado por O, su semejanza S ($0 \leq S \leq 1$) es una combinación lineal de dos valores: una semejanza conceptual S_c y una semejanza relacional S_r.

Semejanza conceptual

La semejanza conceptual, $0 \leq Sc \leq 1$, depende de la cantidad de conceptos comunes de G_1 y G_2. Esta semejanza indica qué tan parecidas son las entidades, acciones y atributos mencionados en ambos grafos conceptuales.

La semejanza conceptual entre G_1 y G_2 se calcula con la siguiente ecuación.

$$s_c(G_1, G_2) = \frac{2\sum_{c \in O}\left(weight(c) \times \beta\left(\pi_{G_1} c, \pi_{G_2} c\right)\right)}{\sum_{c \in G_1} weight(c) + \sum_{c \in G_2} weight(c)} \qquad (3.1)$$

Donde *weight*(c) indica la importancia del concepto *c* dependiendo de su tipo, y la función $\beta(\pi_{G1}c, \pi_{G2}c)$ que expresa el nivel de generalización del concepto común $c \in O$ con respecto a sus proyecciones en los grafos originales, $\pi_{G1}c$ y $\pi_{G2}c$.

weight(c) toma diferentes valores dependiendo del tipo de concepto (entidad, acción, atributo). En este modelo son valores asignados por el usuario. Para no enfatizar ninguna preferencia por algún tipo de concepto se asigna el mismo valor para la función *weight*(c) para cualquier concepto. Nosotros no consideramos las preferencias del usuario dentro de este modelo de comparación, por lo que los asignamos el mismo valor para todos los conceptos.

Por su parte la función β expresa la semejanza semántica entre los conceptos originales en base a una jerarquía de conceptos establecida. La función β se define de la siguiente manera.

$$\beta(\pi_{G1}c, \pi_{G2}c) \begin{cases} 1 & \text{Si } type(\pi_{G1}c) = type(\pi_{G2}c) \text{ y } referent(\pi_{G1}c) = referent(\pi_{G2}c) \\ \dfrac{d}{d+1} & \text{Si } type(\pi_{G1}c) = type(\pi_{G2}c) \text{ y } referent(\pi_{G1}c) \neq referent(\pi_{G2}c) \\ \dfrac{2 \times d_c}{d_{\pi_{G_1} c} + d_{\pi_{G_2} c}} & \text{Si } type(\pi_{G1}c) \neq type(\pi_{G2}c) \end{cases}$$

Donde ***d***, para el segundo caso, indica el número de niveles de la jerarquía de conceptos proporcionada.

Donde d_i es la distancia, expresada como el número de nodos, desde el concepto *i* hasta la raíz de la jerarquía de conceptos.

Semejanza relacional

La semejanza relacional, $0 \leq Sr \leq 1$, indica qué tan similares son las relaciones entre los conceptos comunes en ambos grafos conceptuales G_1 y G_2. Es decir, la semejanza relacional indica que tan parecidos son los *vecindarios* de los conceptos del traslape en los grafos conceptuales originales.

Para calcular esta semejanza se usa la ecuación 3.2.

$$s_r(G_1, G_2) = \frac{2\sum_{r \in O} weight_O(r)}{\sum_{r \in N_O(G_1)} weight_{G_1}(r) + \sum_{r \in N_O(G_2)} weight_{G_2}(r)} \qquad (3.2)$$

El vecindario del traslape *O* en el grafo conceptual *G*, denotado como $N_O(G)$, es el conjunto de todas las relaciones conceptuales conectadas a los conceptos comunes en el grafo *G*; esto es:

$N_O(G)$: $\bigcup_{c \in O} N_G(c)$

Donde $N_G(c) = \{ r | r \text{ está conectada a } \pi_G c \text{ en } G\}$

En esta fórmula $weight_G(r)$ indica la importancia de la relación conceptual *r* en el grafo conceptual *G*. Este valor se calcula de acuerdo con el vecindario de *r* en *G*.

$$weight_G(r) = \frac{\sum_{c \in N_G(r)} weight(c)}{|N_G(r)|}$$

Donde $N_G(r) = \{ c | c \text{ está conectada a } r \text{ en } G\}$

Semejanza global

La semejanza global es la combinación de la semejanza conceptual *Sc* y la semejanza relacional *Sr*, de tal forma que la semejanza total es proporcional a ambos componentes. La semejanza total se define como:

$$S_g = S_c \times (a + bS_r) \qquad (3.3)$$

Donde 0 < a y b <1 ; y a + b = 1. Los valores de a y b se establecen en 0.5 para considerar iguales los valores de las semejanzas conceptual y relacional.

Contextos

Los grafos conceptuales son muy expresivos en relación con los contextos. Un contexto es un concepto con un grafo conceptual anidado, el cual describe al referente. En la figura 3.7, se muestra un ejemplo de un contexto de tipo Situación:

Juan cree que María desea casarse con un político

En este ejemplo, *Juan* es el que experimenta el concepto *creer*, el cual se enlaza al tema por la relación *THME* a la proposición *Juan cree*. El contenido del cuadro *Proposición* contiene otro grafo conceptual, en el cual *María* es la que experimenta el concepto *deseo*, el cual se enlaza con una situación en la que *María* espera que se cumpla. Tal situación se representa por otro grafo anidado en el cual indica que *María* se casa con un político, la línea punteada indica la correferencia del concepto.

Los grafos conceptuales que utilizan contextos se consideran grafos complejos; y los que no usan contextos se conocen como grafos conceptuales básicos según (Chein & Mugnier, 2009). Nuestro estudio lo dirigimos principalmente con grafos conceptuales básicos.

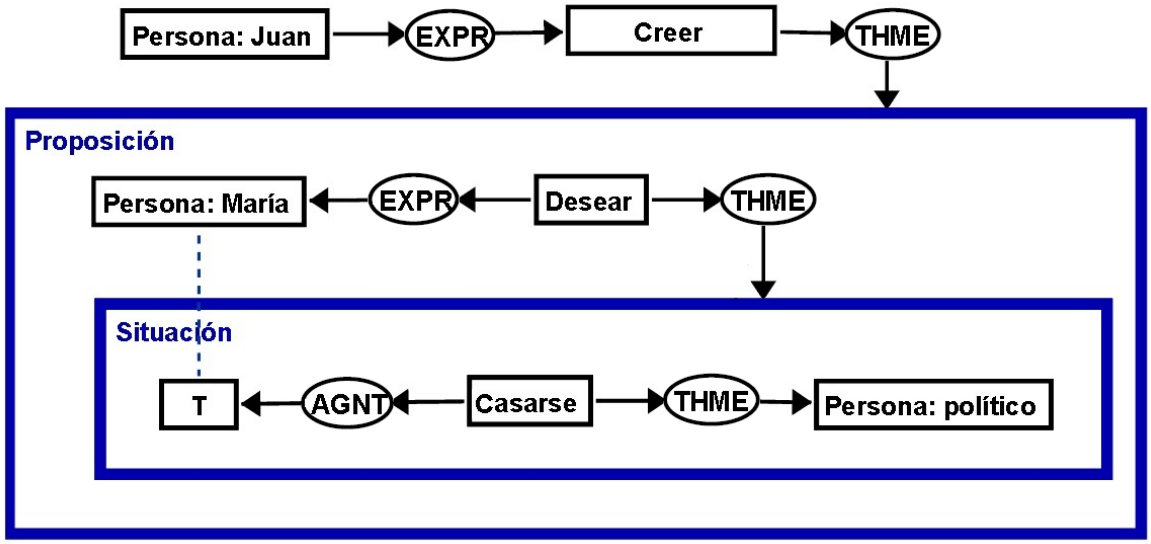


Figura 3.7. Grafo conceptual con dos contextos anidados

Capítulo 4

Método propuesto

En este capítulo se presenta el modelo propuesto para la generación de resúmenes abstractivos de un solo documento. La tarea principal es obtener un resumen del texto representado como grafos conceptuales. En este contexto, reducir las estructuras de los grafos conceptuales significa obtener el resumen del texto.

También se propone una nueva clase de grafos conceptuales: los grafos conceptuales ponderados. En nuestro contexto, los grafos conceptuales ponderados describen flujos semánticos, los cuales ayudan a identificar mejor a los actores principales en las estructuras gráficas. Estos grafos permiten a nuestro algoritmo de ponderación desempeñarse adecuadamente en nuestro marco de trabajo.

De acuerdo con Sowa (Sowa J. F., 1984), los grafos conceptuales son una manera natural y detallada de representar el lenguaje natural, es decir, los textos. Estas estructuras se han usado en otras áreas como la minería de textos para identificar los patrones del usuario (Montes-y-Gómez *et al*., 2000; 2001); sin embargo, para la generación de resúmenes, técnicas similares no han sido exploradas.

Actualmente, la mayoría de los métodos para la generación de resúmenes se han orientado principalmente al enfoque extractivo. No obstante, algunas investigaciones han utilizado algunas técnicas para la combinación de oraciones por medio del uso estructuras sintácticas y parafraseo (Barzilay *et al*., 2005); también se ha explorado el análisis de la estructura discursiva del texto (Marcu, 2000; da Cunha *et al*., 2007), así como la identificación de tripletas con la estructura sujeto-verbo-objeto (Tsatsaronis *et al*., 2010), y compresión de frases basado en LSA (Steinberger *et al*., 2006) o en la RST (Molina A. *et al*., 2013) . Sin embargo, todos esos enfoques no usan una representación conceptual y estructural completa de las oraciones del texto.

Nuestro modelo para la generación de resúmenes abstractivos se basa en la representación semántica completa del texto, a saber, estructuras de grafos conceptuales. De modo que el problema de la generación del resumen lo podemos simplificar en seleccionar los nodos más importantes de los grafos, y reducir las estructuras gráficas manteniendo su coherencia estructural. En este contexto, los grafos resultantes representarán al resumen a nivel conceptual.

En la siguiente sección presentamos la noción extendida de grafos conceptuales, esto es, los grafos conceptuales ponderados.

4.1 Grafos conceptuales ponderados

La noción de grafos conceptuales ponderados (GCP) se introdujo en (Miranda-Jiménez, Gelbukh, & Sidorov, 2013). Esta clase de grafos conceptuales surgen de la necesidad de ponderar los grafos conceptuales tradicionales, debido al uso de algoritmos de ponderación

como HITS o PageRank, en los cuales se puede aprovechar el peso que puede asignarse a las aristas que conectan a los nodos de los grafos conceptuales.

La estructura de los grafos conceptuales ya describe flujos de información semántica por medio de las conexiones entre las relaciones conceptuales y los conceptos que la conforman. Sin embargo, algunos de estos flujos son más importantes que otros, ya sea por las relaciones que interconectan o por la preferencia particular de algunos de ellos.

De lo anterior surge la necesidad de discriminar de cierta manera las conexiones entre los nodos que conforman a los grafos conceptuales, lo cual lleva a asignar un peso a las relaciones entre dos nodos cualesquiera. Con esto, se proporciona un mecanismo flexible para discriminar los flujos de interés en un contexto de ponderación.

Los flujos que se crean al pasar por las relaciones conceptuales, cuyas aristas entrantes y salientes tienen un peso asociado, se denominan 'flujos semánticos'. Un flujo semántico es básicamente el valor que acumulan los nodos y que se transmite hacia otros nodos aumentando o disminuyendo dicho valor al pasar por alguna relación conceptual que lo afecta directamente.

En este contexto, las relaciones conceptuales representan principalmente la semántica del texto, por lo que un peso con valor alto en sus aristas asociadas indican interés por el flujo que pasa a través de la relación en cuestión. De forma natural, los grafos conceptuales describen en su estructura las relaciones existentes de la semántica del texto, relaciones tales como agente, objeto, lugar, atributo, tema, etc. descritas por Sowa (Sowa J. F., 1984).

En la figura 4.1, se muestra la representación del grafo conceptual ponderado para la oración "*The cat Yojo is chasing a brown mouse*" (El gato Yojo está persiguiendo a un ratón café) tomada de (Sowa J. F., 1984).

En este grafo, el valor 2 sobre las aristas que conectan a la relación agentiva (AGNT), arista entrante y arista saliente, indican que esos flujos son más importantes que los otros flujos que tiene un valor de 1.

En el contexto de ponderación, se pueden asignar pesos a las aristas tanto positivos como negativos, incluyendo el valor cero. Aristas con valores positivos mayores a 1

recompensan el flujo que pasa a través de éstas; aristas con valores negativos y menores a 1 penalizan el flujo que pasa por ellas; el valor 1 es el elemento neutro; y valores iguales a cero cancelan el flujo que pasa por estas aristas.

Los pesos que se asignan a las aristas se determinan heurísticamente y de acuerdo a los intereses del usuario respecto a los actores (relaciones agentivas), cualidades (relaciones atributo), temas (relaciones de tema), lugares (relaciones locativas), etc.

Los flujos semánticos se usan durante el proceso de ponderación para elegir a los nodos más importantes, los cuales son los que transmiten mayor información a través de la red (véase la sección 4.3).

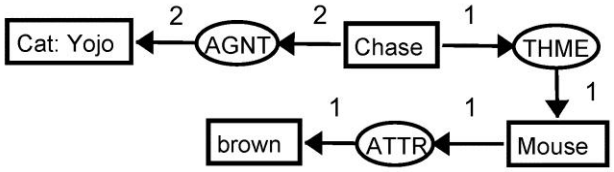


Figura 4.1. Grafo conceptual ponderado

Adicionalmente, los nodos concepto aceptan un peso que define la preferencia del nodo. Esto es, especifican el grado de interés en ciertos tópicos definidos por los conceptos. De manera similar a los pesos de las aristas, valores positivos recompensan el flujo que pasa por ellos y valores negativos penalizan los flujos que pasan por ellos. Los detalles y consideraciones tanto para los pesos de los grafos ponderados y las preferencias se detallan en la sección 5.2.

4.2 Esquema computacional

Para lograr nuestro objetivo, en la figura 4.2 se muestra el esquema general de nuestro modelo. Primeramente, se hace un preprocesamiento al texto para generar semiautomáticamente los grafos conceptuales a partir de un conjunto de documentos seleccionados de la competencia DUC. Durante el proceso de generación se agrega información semántica y sintáctica de fuentes externas.

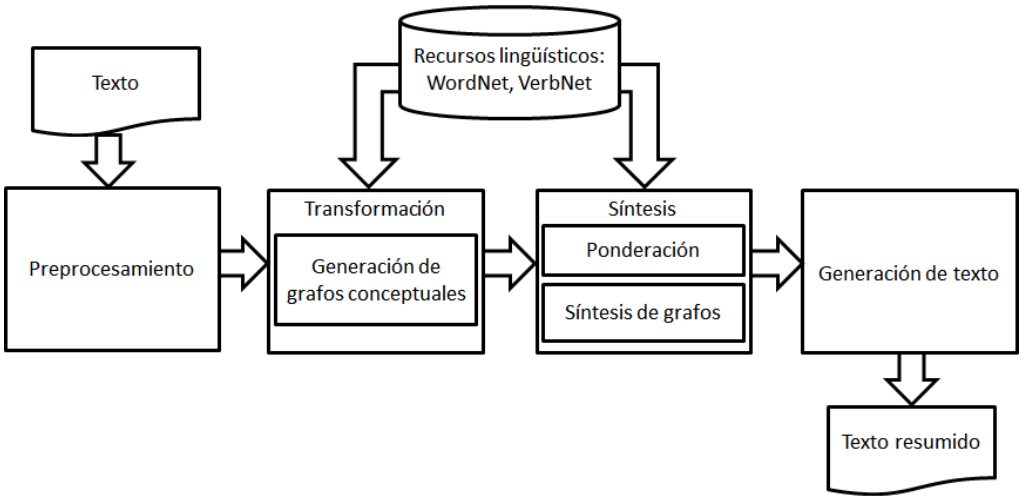


Figura 4.2. Modelo para la generación de resúmenes basado en grafos conceptuales

En la etapa de síntesis, los grafos se reducen de acuerdo a un conjunto de operaciones (poda, unión y generalización) que se les aplica. A partir de los grafos resultantes se genera el texto resumido.

En las siguientes secciones, describimos a detalle las etapas en que consiste la arquitectura.

4.2.1 Recursos lingüísticos

Nuestro enfoque se basa en el formalismo de grafos conceptuales, el cual requiere de información semántica adicional que se agrega a los grafos, a saber, una jerarquía de conceptos para propósitos de generalización; además de patrones verbales para mantener la coherencia estructural de los grafos.

Dichos recursos lingüísticos son específicos para el idioma de estudio y son necesarios para el funcionamiento adecuado de nuestro modelo.

En esta investigación, aplicamos el modelo presentado en la figura 4.2 al idioma inglés, debido a que este idioma cuenta con los recursos lingüísticos mencionados y son de libre acceso como WordNet (Fellbaum, 1998) y VerbNet (Kipper, Trang Dang, & Palmer, 2000); además se cuenta con documentos para poder evaluar nuestro método, dichos documentos los proporciona la competencia DUC (DUC, 2003). En la información proporcionada por DUC, se cuenta con los documentos originales, así como con los

resúmenes generados manualmente por humanos para distintas tareas que proporcionan la competencia.

Cabe resaltar que actualmente se ha dado un fuerte interés por el estudio de otros idiomas en el campo de la generación automática de resúmenes. Tal interés ha impulsado a otras iniciativas como el foro MultiLing[7] bajo el auspicio de la *Text Analysis Conference*[8] (TAC) para la generación de corpus para otros idiomas; así como la evaluación de los sistemas de generación de resúmenes tanto multidocumento como multilingüe.

MultiLing proporciona los corpora para diferentes lenguajes como el francés, español, checo, inglés, árabe, griego, hebreo, rumano, hindi y chino (Elhadad *et al*., 2013). Las colecciones de textos cuentan con los documentos originales y tres resúmenes hechos manualmente asociados a cada colección.

Aunque el autor de esta tesis participó en la iniciativa MultiLing para preparar la colección de datos para el español, todavía no se cuenta con todos los recursos lingüísticos para aplicar el modelo presentado en esta tesis al idioma español.

WordNet

WordNet es una base de datos léxica para el inglés, contiene cerca de 155,000 entradas entre sustantivos, verbos, adjetivos y adverbios. Está organizada jerárquicamente en grupos de sinónimos llamados '*synsets*', y está enlazada mediante relaciones semánticas de hiperonimia / hiponimia (clase / subclase), holonimia / meronimia (todo / parte), antonimia y algunas otras.

Cada *synset* expresa un concepto distinto, los *synsets* están enlazados por medio de relaciones semántico-conceptuales y léxicas, por lo que se forma una red de palabras relacionadas por su significado.

---

[7] *http://multiling.iit.demokritos.gr/pages/view/662/multiling-2013*

[8] *www.nist.gov/tac/*

Un ejemplo de una entrada de WordNet es la siguiente: *knowledge* (conocimiento). En esta entrada, todas las palabras del *synset* son equivalentes.

knowledge

- *cognition, knowledge, noesis* -- (*the psychological result of perception and learning and reasoning*).

La jerarquía de conceptos tipo se obtuvo por medio de las relaciones semánticas de hiperonimia / hiponimia. Por ejemplo, *atmospheric phenomenon* / *storm* (fenómeno atmosférico / tormenta), *residence* / *home* (residencia / casa).

Se obtuvieron las siguientes relaciones para los conceptos indexados en WordNet[9] 3.0. 32,028 relaciones directas entre el concepto y sus hipónimos representados por el *synset*; y se obtuvieron 130,717 relaciones directas entre el concepto y sus hiperónimos representados por el *synset*.

4.2.2 VerbNet

VerbNet[10] (Kipper, Trang Dang, & Palmer, 2000), otro recurso importante, es un diccionario computacional de verbos, se basa en la clasificación de verbos de Levin (Levin, 1993). Éste contiene información sintáctica y semántica de verbos para el inglés. . VerbNet asocia la semántica de un verbo con su marco sintáctico y combina la información semántica léxica tal como los roles semánticos con los marcos sintácticos y las restricciones de selección del verbo.

En VerbNet, la suposición fundamental es que los marcos sintácticos de un verbo (patrones verbales) como los elementos que toman sus argumentos son un reflejo directo de la semántica subyacente.

Cada clase definida en VerbNet contiene una lista de miembros, una lista de posibles roles temáticos, y una lista de *frames*, patrones donde se indica cómo los roles semánticos

[9] http://wordnet.princeton.edu/

[10] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

pueden realizarse en una oración. Los verbos que pertenecen la misma clase de VerbNet comparten los mismos marcos sintácticos; por eso se cree que tienen el mismo comportamiento sintáctico.

Por ejemplo, la clase *chase* (perseguir) describe un patrón principal definido como 'NP V NP' (frase sustantiva / verbo / frase sustantiva) y se indica que es verbo transitivo.

Los patrones verbales de VerbNet son un mecanismo para regir la coherencia de las oraciones. Con esta misma idea, los patrones ya descritos en VerbNet se usan directamente en las estructuras de los grafos conceptuales. Se extrajeron los patrones del diccionario con el fin de regir la coherencia estructural de los grafos.

Se extrajeron 1,420 patrones verbales primarios y un total de 3,343 formas verbales.

4.2.3 Colección de documentos de prueba

Los documentos que se utilizaron para probar el método propuesto son parte de la competencia *Document Understanding Conference* 2003 (DUC, 2003), en la cual se evalúan a los sistemas de generación de resúmenes tanto extractivos como abstractivos.

La colección preparada para evaluar nuestro enfoque consta de 30 documentos de noticias referentes, principalmente, a desastres naturales. Los textos seleccionados son breves, entre 50 y 100 palabras aproximadamente.

La colección de documentos que proporciona la competencia DUC consta de los textos fuente y en promedio tres o cuatro resúmenes (hechos por humanos) por cada documento. DUC preparó la información de esta forma con el fin de comparar los resultados de los sistemas que generan resúmenes.

Se utiliza esta versión de documentos, ya que los documentos creados para la competencia DUC 2003 eran resúmenes cortos, aproximadamente 100 palabras; también se incluyen los encabezados de las noticias que pueden considerarse un resumen muy breve de la noticia. Aquí se puede considerar, el resumen como documento y los encabezados como los resúmenes de este documento.

En la colección de documentos se cuenta con el texto original y 3 o 4 resúmenes construidos por asesores de la competencia; además de los encabezados de la noticia. Estos resúmenes hechos por humanos sirven como modelos para evaluar a los resúmenes construidos por los sistemas que los generan automáticamente, como se mencionó en la sección 2.5.

Por ejemplo, para la siguiente noticia, se cuenta con cuatro versiones de su texto resumido, a nivel de encabezado.

Texto original

Typhoon-Babs weakened into a severe tropical storm Sunday night after it triggered massive flooding and landslides in Taiwan and slammed Hong-Kong with strong winds. The storm earlier killed at least 156 people in the Philippines and left hundreds of thousands homeless.

Resúmenes

- *En route to China, Typhoon Babs weakens to severe tropical storm.*
- *Babs weakens after damaging Taiwan, Hong Kong, Philippines. 156 dead.*
- *Babs weakens to tropical storm; Philippines death toll at 156.*
- *Babs weakens to tropical storm after triggering flooding in Taiwan*

Del conjunto de prueba, sólo se seleccionaron textos cortos.

4.2.4 Preprocesamiento del texto

Nuestra colección de grafos conceptuales se creó a partir de los textos de la colección DUC. Para identificar las relaciones gramaticales entre las palabras de los textos, los documentos se procesaron con el analizador sintáctico (*parser*) de Stanford (de Marneffe, MacCartney, & Manning, 2006), es decir, se obtuvieron los árboles de dependencia (Mel'čuk, 1988) de los textos en cuestión, éstos se usan para la construcción de los grafos conceptuales.

Después de obtener las dependencias entre las palabras, se hace la creación de los grafos con un conjunto de reglas y considerando el tipo de relación obtenida por el parser. En la siguiente sección se detalla este procedimiento.

4.2.5 Generación de grafos conceptuales

Como lo mencionamos en la descripción del problema, no es nuestra tarea proporcionar los métodos para generar los grafos conceptuales, esta tarea tiene sus propios retos (Hensman & Dunnion, 2004; Ordoñez-Salinas & Gelbukh , 2010), aunque sí es necesario tener los textos representados en este formalismo. Por lo que la colección de datos se preparó semiautomáticamente.

Después de realizar el procesamiento de los textos, y a partir de la información gramatical de los árboles de dependencias, se generaron los grafos conceptuales por medio de un conjunto de reglas de transformación. Por ejemplo, las relaciones *nsubj* (sujeto nominal) y *agent* (agente) se convierten en AGNT (agente), la relación *amod* (modificador adjetivo) se convierte en ATTR (atributo), *dobj* (objeto directo) en THME (tema), etc.

El conjunto de relaciones gramaticales usadas por el parser para el inglés se definen en (de Marneffe & Manning, 2008). Las relaciones que se asignan incorrectamente o no se identifican por las reglas de transformación, se corrigen o añaden manualmente.

En la figura 4.3, se muestra un ejemplo de la construcción de un grafo conceptual. Se crean los nodos para la oración "*Bell distributes computers*" (Bell distribuye computadoras). Para la oración anterior, el parser de Stanford genera las relaciones *nsubj*(*distributes*, *Bell*) y *dobj*(*distributes*, *computers*). La tripleta está constituida por el nombre de la relación, la palabra gobernante y la palabra dependiente.

Los nodos que se generan a partir de la relación *nsubj* son Bell, AGNT (agente) y *distribute* (distribuir), y para la relación *dobj* son THME (tema) y *Computer*:{*} (número indeterminado de computadoras). Las características sintácticas del concepto, por ejemplo, en el caso del verbo, se mantienen codificadas en el nodo correspondiente tal como *distributes* (etiqueta *VBZ* generada por el parser que significa: verbo, tercera persona del

singular, tiempo presente) y sólo la palabra normalizada (verbo en infinitivo) se muestra en el grafo.

El conjunto completo de las etiquetas para las categorías gramaticales usadas por el parser de Stanford está definido en el marco del proyecto Penn TreeBank (Santorini, 1990).

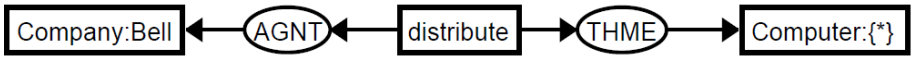


Figura 4.3. Construcción de los nodos a partir de relación *nsubj* y *obj*

Posteriormente, se adjuntó manualmente el 'concepto tipo' (hiperónimo) a cada concepto del grafo, existente en la jerarquía de WordNet. Por ejemplo, en la figura 4.3, se asignó *Company* (Compañía) al concepto *Bell* y *Computer* (Computadora) al concepto *computers* (computadoras) –{*} indica un conjunto indeterminado de elementos de la clase *Computer*–, véase Sowa (1984).

Por último, también se añadió la clase verbal asociada a VerbNet que define al patrón verbal para los nodos concepto de tipo verbo, lo cual permite mantener la coherencia de la estructura gráfica.

Por ejemplo, la clase *contribute* (contribuir) de VerbNet contiene al concepto verbo *distribute* (distribuir). En esta clase, se define al patrón verbal básico como 'NP V NP.Theme' (frase sustantiva / verbo / frase sustantiva como tema). Lo que indica que este grafo debe tener un complemento definido como su tema.

Un ejemplo más de un patrón verbal, en la figura 4.4, la clase *chase* (perseguir) define el patrón básico 'NP V NP' (frase sustantiva / verbo / frase sustantiva), y ésta contiene a verbos como *chase* (perseguir), *pursue* (perseguir, seguir), *follow* (seguir a), etc. que usan este mismo patrón, por lo que una estructura similar debe existir en el grafo, en este caso, la estructura se define por *Cat–Chase–Mouse* (Gato–perseguir–Ratón).

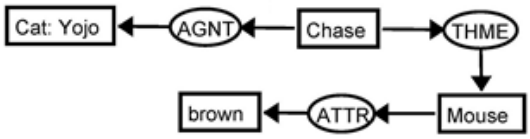


Figura 4.4. Ejemplo de patrón verbal

Los grafos generados son grafos simples con el fin de facilitar y demostrar nuestro modelo. Es decir, son grafos sin negaciones, ni las llamadas situaciones, ni contextos (Sowa J. F., 1984).

En el grafo de la figura 4.5, representa el texto de noticia que se encuentra abajo. En el grafo las líneas punteadas representan la correferencia de los conceptos asociados para facilitar la lectura. Esta misma correferencia está definida por los números referenciados, por ejemplo , '?1' hace referencia al concepto etiquetado como '*1'.

"*Typhoon Babs weakened into a severe tropical storm Sunday night after it triggered massive flooding and landslides in Taiwan and slammed Hong Kong with strong winds. The storm earlier killed at least 156 people in the Philippines and left hundreds of thousands homeless."*

"El tifón Babs se debilitó a tormenta tropical severa la noche del domingo después de haber provocado graves inundaciones y deslizamientos de tierra en Taiwán y golpeó a Hong Kong con fuertes vientos. La tormenta mató al menos a 156 personas en Filipinas y dejó a cientos de miles sin hogar."

En nuestro enfoque, consideramos como conceptos a palabras de contenido, es decir, todas las palabras excepto palabras auxiliares (como artículos, preposiciones, etc.), y como relaciones conceptuales consideramos los roles semánticos (Jackendoff, 1972): agente, iniciador, instrumento, experimentante, paciente, lugar, tiempo, objeto, fuente, y meta; así como otras relaciones tales como atributo, cantidad, medida, etc., aproximadamente 30 relaciones usadas en Sowa (1984).

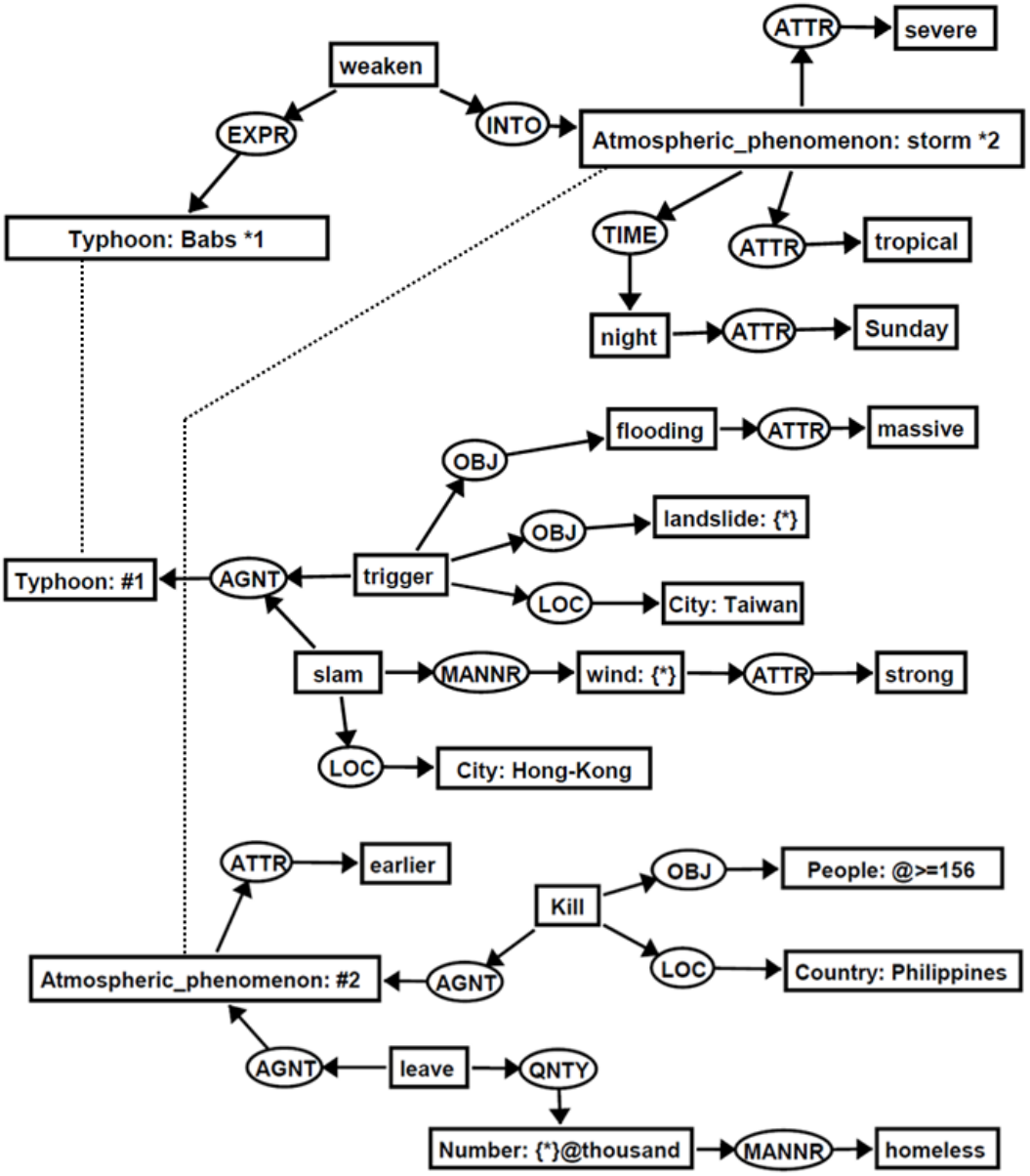


Figura 4.5. Ejemplo de noticia como grafo conceptual

4.2.6 Síntesis

En nuestro marco de trabajo, los textos están representados como grafos conceptuales. Por lo que la tarea para la generación del resumen se reduce a resumir los grafos conceptuales.

En general, el método consiste en los procesos de ponderación y poda de los nodos de los grafos conceptuales (síntesis de grafos), apoyado con las operaciones de generalización y unión que para relacionan a los grafos conceptuales.

El proceso de ponderación se basa en la estructura y los flujos semánticos de los grafos conceptuales ponderados (Miranda-Jiménez *et al*., 2013), así como en el algoritmo HITS de Kleinberg (Kleinberg, 1999; Mihalcea *et al*., 2004) para determinar la importancia de los nodos.

El proceso de poda considera la información de ponderación del algoritmo HITS y utiliza los patrones verbales de VerbNet para mantener la coherencia en las estructuras durante el proceso de eliminación de los nodos. Los grafos resultantes después de aplicar las operaciones se consideran la representación del resumen a nivel conceptual.

El procedimiento completo se detalla en la sección 4.3.

4.2.7 Generación de texto

El desarrollo de esta etapa queda fuera del alcance de esta investigación. Sin embargo, hay algunos trabajos que se han propuesto para la generación del texto a partir de los grafos conceptuales. Por ejemplo, utilizar rutas de expresión y el uso de reglas gramaticales (Sowa, 1983; 1984; 1999), o la asociación de estructuras semánticas a las estructuras sintácticas por medio de una gramática de adjunción de árboles (Nicolov, Mellish, & Ritchie, 1995).

4.3 Síntesis de grafos conceptuales

Como mencionamos, la entrada de datos a nuestro modelo son grafos conceptuales, lo cuales representan al texto que queremos resumir.

En nuestro enfoque, la forma que proponemos para obtener un resumen es por medio de varias operaciones sobre los grafos conceptuales. Las operaciones básicas que realizamos son cuatro: generalización, unión o asociación, ponderación y poda.

Un mecanismo importante para que soporte las operaciones de generalización y unión es la comparación entre grafos conceptuales. Este mecanismo de comparación permite identificar las generalizaciones y asociaciones entre los grafos conceptuales

Comparación de grafos

El procedimiento que usamos para la comparación de los grafos conceptuales es el propuesto por Montes-y-Gómez (Montes-y-Gómez *et al*., 2000; 2001), el cual consta de dos etapas: el apareamiento de los grafos y la medición de la semejanza.

En la primera etapa se identifican los conceptos y relaciones conceptuales comunes, a partir de esto, se construye la semejanza (similitud), por medio de los traslapes. En la segunda etapa, se calcula la semejanza por medio de la importancia relativa del traslape con respecto a los grafos originales.

Para llevar a cabo ambas etapas, se requiere de conocimiento del domino, el cual es expresado por medio de las jerarquías de conceptos, en este caso, WordNet. Básicamente, estas jerarquías de conceptos permiten determinar las semejanzas entre los grafos a diferentes niveles de generalización. Para llevar a cabo la comparación se realizan los procedimientos C-I y C-II.

Procedimiento C-I

1. Dados los grafos conceptuales G_1 y G_2.
2. Para cada concepto c_1 perteneciente a G_1 y cada concepto c_2 perteneciente a G_2.
   a. Se extraen las generalidades comunes de c_1 y c_2, se guardan en el conjunto P: $P \leftarrow c_{12}$.
3. Para cada relación r_1 perteneciente a G_1 y cada relación r_2 perteneciente a G_2.
   a. Se extraen las generalidades comunes de r_1 y r_2, se guardan en el conjunto P: $P \leftarrow r_{12}$.

El conjunto resultante *P* representa el conjunto de semejanzas entre los dos grafos conceptuales.

Procedimiento C-II

1. A partir del conjunto P de la etapa C-I, $Traslapes_1$ representa todos los conceptos de P.

2. Mientras haya traslapes en $Traslapes_{k-1}$ para $k > 1$ (k inicia en el nivel 2).
   a. Construir $Traslapes_k$ a partir de $Traslapes_{k-1}$.
   b. Eliminar $Traslapes_{k-1}$ cubiertos por $Traslapes_k$.
   c. Pasar al siguiente nivel de k.
3. *MaximoTraslapes* ← Unión de todos los $Traslapes_k$ desde k=1.
4. Para cada relación r perteneciente a P y para cada traslape O_i perteneciente a *MaximoTraslapes, se prueba:*
   a. Si todos los conceptos vecinos a r existen en O_i; entonces agregar la relación O_i ←r.

Inicialmente el procedimiento C-I, considera que cada uno de los conceptos del conjunto *P* es por sí solo un traslape. Para cada nuevo nivel de *k*, se generan los traslapes basándose en el nivel anterior *k*-1. En cada nivel anterior se eliminan los traslapes que no fueron máximos para el nivel en proceso, es decir, los elementos que formaron traslape. El proceso se detiene hasta cuando ya no se encuentra ningún nuevo traslape. Cada uno de los traslapes de tamaño *k* se genera de la unión de dos traslapes compatibles de tamaño *k*-1. Al final se agregan las relaciones de *P* en los traslapes correspondientes, deben existir todas las conexiones de la relación para que pueda insertarse.

Medición de la semejanza

Una vez construidos los traslapes de los grafos conceptuales. Se procede a medir la semejanza entre esos traslapes.

La medición de la semejanza es otra etapa de la comparación de los grafos conceptuales. En esta etapa se recibe como entrada los dos grafos que se compararán y el conjunto de todos sus posibles traslapes, el conjunto *P*.

Para cada traslape se calcula una medida de semejanza. Se obtiene la mayor medida y el traslape que la produce (que es la descripción final de la semejanza).

Para el cálculo de la semejanza global se usa la ecuación (3.3) que combina la semejanza conceptual, ecuación (3.1), y la semejanza relacional, ecuación (3.2).

4.3.1 Generalización

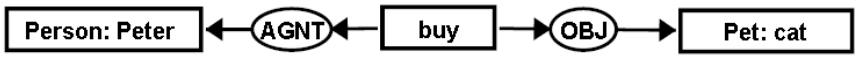
La operación de generalización consiste en "abstraer" los conceptos candidatos por otro concepto relacionado en un jerarquía de tipos donde existen relaciones semánticas (hipónimos / hiperónimos), para este propósito usamos la jerarquía de WorNet.

En la figura 4.6 se muestra un ejemplo de síntesis por generalización, los conceptos *cat* (gato), y *bird* (pájaro) tienen asociados el tipo *Pet* (Mascota); de forma similar, los conceptos *Peter* y *Mary* que tiene asociado el concepto tipo *Person* (Persona).

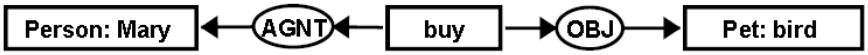
Los grafos pueden leerse como sigue

G_1: *Peter buys a cat* (Peter compra un gato)

G_2: *Mary buys a bird* (Mary compra un pájaro)



(G_1)



(G_2)

Figura 4.6. Grafos conceptuales candidatos para generalización

Para realizar esta operación la semejanza entre los grafos debe ser máxima, es decir, todos los nodos y las relaciones deben ser semejantes para aplicar la reducción de los nodos.

Al realizar la generalización por medio del traslape de conceptos obtenemos el grafo G_3 de la figura 4.7, el cual puede leerse: *Peter and Mary buy a pet each one* (Peter y Mary compran una mascota cada uno).

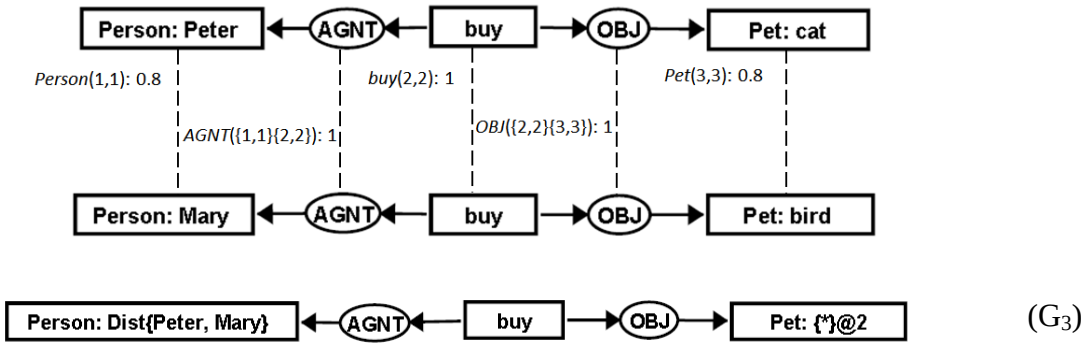
En la figura 4.7, se muestran los traslapes que se obtiene al comparar los grafos G_1y G_2. Las condiciones para que se lleve a cabo esta operación son las siguientes:

1. El número de conceptos y relaciones deben ser semejantes para ambos grafos.
2. Las entidades se generalizan al concepto tipo mínimo común de acuerdo a la jerarquía de conceptos.
3. Se aplican reglas heurísticas para agrupar entidades, de acuerdo al grado de generalización requerido.

Ejemplo de reglas:

a. Se aplica el operador *Dist* (distribución) a los referentes de los conceptos asociados a las relaciones agente (conceptos {1,1} de la relación ***AGNT***), bajo el concepto tipo mínimo común.
b. Se contabilizan las entidades asociadas como objeto (conceptos {3,3} de la relación ***OBJ***).

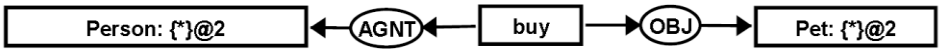
Un ejemplo de la regla se muestra en la en grafo G_3 de la figura 4.7. Donde las entidades [*Person: Peter*] y [*Person: Mary*] difieren en su referente, esto es, en el concepto *Peter* y *Mary*. Al realizar la comparación entre los traslapes se obtuvo *Person*(1,1): **08**; lo que nos indica que hay discrepancia entre el concepto 1 de G_1 y el concepto 1 de G_2. Lo mismo sucede con el concepto *Pet*, la discrepancia es en el referente del concepto 3 de G_1 y concepto 3 de G_2. Las reglas *a* y b se aplican a los nodos y el resultado se refleja en el grafo G_3.



(G_3)

Figura 4.7 Síntesis de grafos por generalización

En la figura 4.8, se muestra el grafo sin aplicar la regla *a* a la generalización de los grafos G_1 y G_2 y sólo se aplican reglas para contabilizar como la regla *b*. El grafo G_4 puede leerse como *Two persons buy two pets* (dos personas compran dos mascotas).

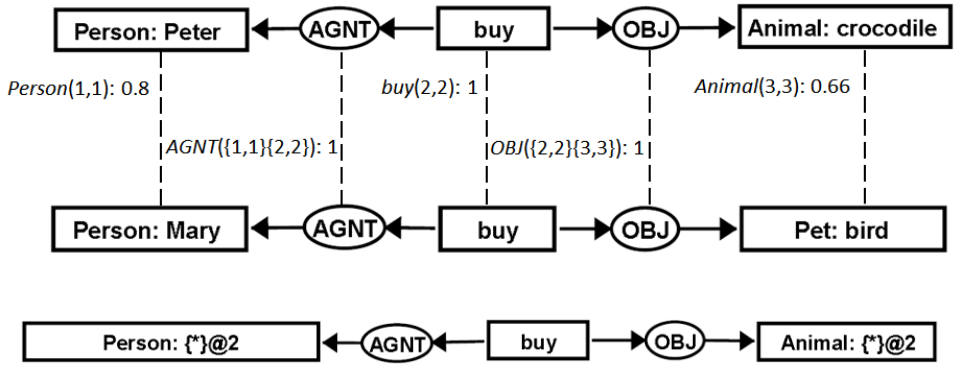


(G_4)

Figura 4.8 Síntesis de grafos por generalización: abstracción

Un ejemplo más sobre la operación de generalización. Si se tienen los siguientes grafos, y se cuenta con la jerarquía (a) de la figura 4.10, donde se describen las asociaciones de hiponimia / hiperonimia entre los conceptos. El grafo G_5 que se quiere comparar puede leerse como sigue.

G_5: *Peter buys a crocodile* (Peter compra un cocodrilo)



(G_5)

(G_2)

(G_6)

Figura 4.9 Síntesis de grafos por generalización: abstracción (2)

En la figura 4.9, se muestran los traslapes que se identificaron al comparar los grafos G_5 y G_2. El grafo G_6 muestra la operación de generalización para conceptos de distinto nivel en la jerarquía de conceptos. Al realizar esta operación el traslape entre los conceptos [*Animal: crocodrile*] y [*Pet: bird*] se generalizan al concepto tipo mínimo común, es decir al tipo *Animal*. Como se muestra en el grafo G_6, el cual puede leerse como *Two persons buy two animals* (dos personas compraron dos animales).

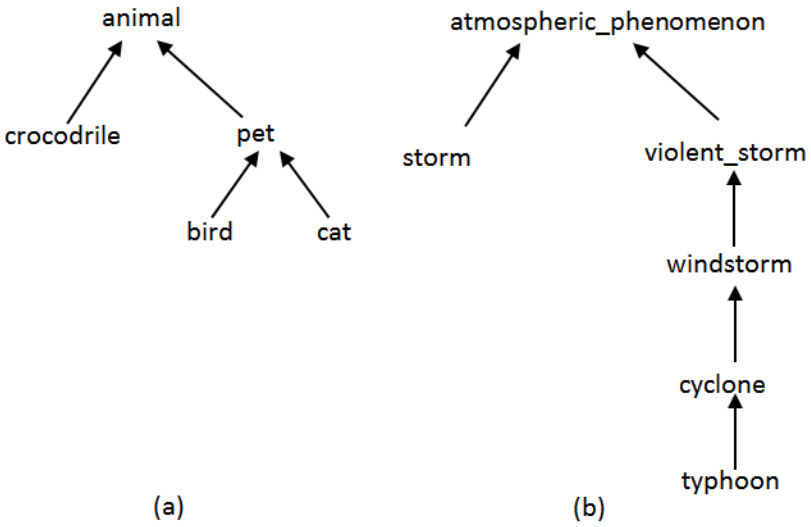


Figura 4.10. Fragmento de jerarquía de conceptos tipo

4.3.2 Unión o asociación

Esta operación se encarga de asociar nodos semejantes de dos grafos conceptuales. Para realizar esta operación asumimos que el texto de origen es cohesivo y coherente, el cual las oraciones probablemente refieren a conceptos que fueron previamente mencionados u otros conceptos relacionados (Halliday & Hasan, 1976; Budanitsky & Hirst, 2006).

Con esta operación se identifican los conceptos relacionados entre los grafos conceptuales para ayudar a mejorar los resultados del proceso de ponderación.

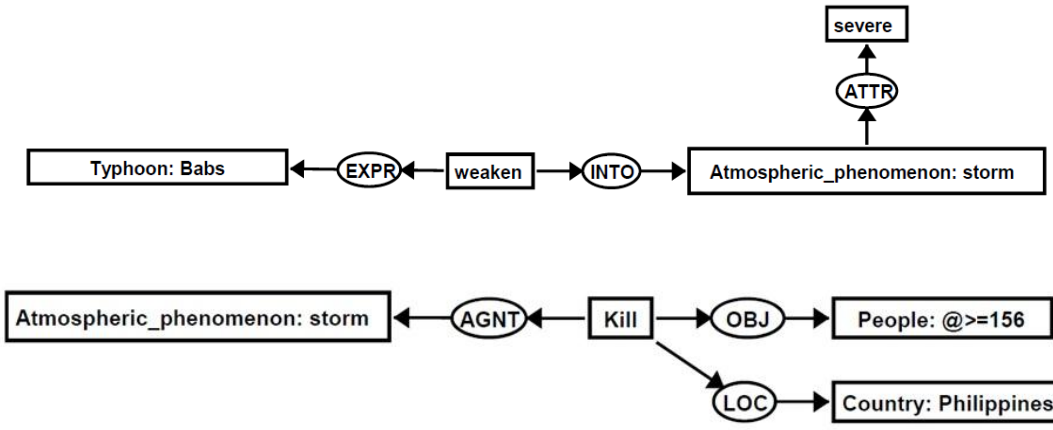


Figura 4.11. Unión de grafos conceptuales

En la figura 4.11, se tiene dos grafos G_7 y G_8, los cuales pueden leerse como sigue.

G_7: *Typhoon Babs weakened into severe storm*

(El tifón Babs se debilitó a tormenta severa)

G_8: S*torm killed at least 156 people in Philippines*

(La tormenta mató al menos a 156 personas en Filipinas)

La detección de asociaciones para unificar nodos entre los grafos conceptuales consiste en el siguiente procedimiento:

1. Comparar todos los grafos conceptuales por pares.
2. Para cada par de grafos G_x y G_y se calculan los traslapes para los conceptos y las relaciones.
3. Si sólo hay conceptos, asociar sólo los conceptos que representan sustantivos que tengan la similitud más alta de acuerdo a sus traslapes.
4. Si hay conceptos y relaciones, asociar la vecindad de conceptos y relaciones (las relaciones deben tener similitud máxima de 1 para asociarse).

En la figura 4.12, se ejemplifica el proceso de unión/asociación para el par de grafos G_7 y G_8 de la figura 4.11. Aquí, se detectan dos traslapes en los grafos: traslape de [*Typhooon: Babs*] y [*Atmospheric_phenomenon*:], conceptos relacionados identificados por (1,1). La semejanza de similitud entre estos conceptos es de 0.333, usando la jerarquía de conceptos (b) de la figura 4.10. Sin embargo, se identifica otra asociación [*Atmospheric_phenomenon*: *storm*] y [*Atmospheric_phenomenon*: *storm*] que representan a la misma instancia, por lo que la similitud es la máxima. Esta última asociación se prefiere por obtener el máximo valor. Los conceptos están identificados por (3,1) el concepto tres del gafo G_7 y el concepto uno del grafo G_8 y son los nodos que forma la asociación.

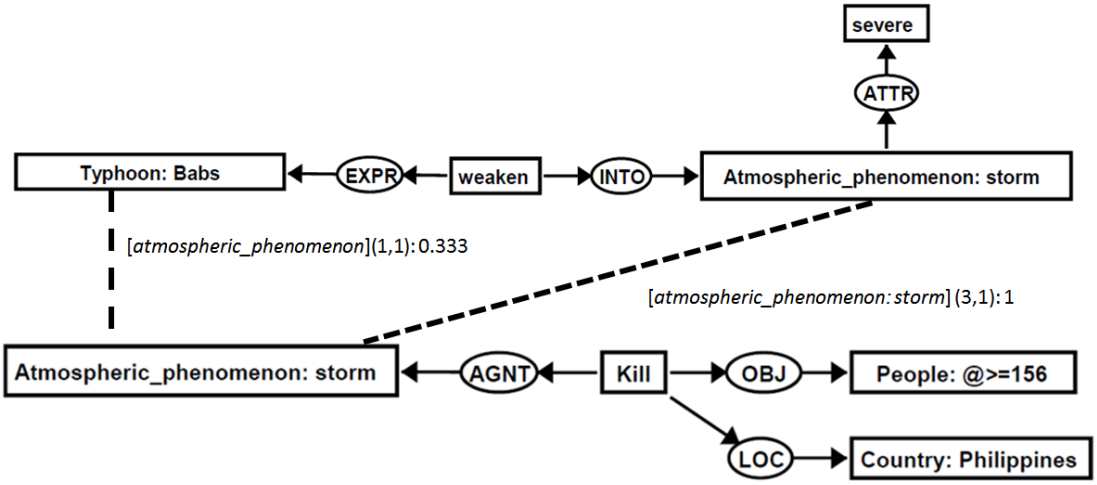


Figura 4.12. Unión de grafos: dos posibles asociaciones

En la figura 4.13, se presenta el caso cuando se deben asociar más de un nodo, es decir, una vecindad. En este caso, las relaciones deben ser idénticas para poder hacer la asociación. Si los referentes no coinciden, en este caso *Babs*, se pueden asociar bajo el concepto tipo.

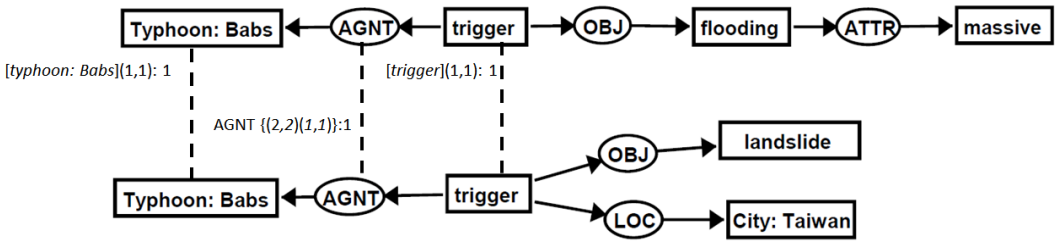


Figura 4.13. Unión de grafos: asociación de una vecindad

Posteriormente a la generalización e identificación de asociaciones, los grafos resultantes deben ponderarse para identificar los nodos más importantes de las estructuras. El proceso de poda es el último en aplicarse.

4.3.3 Ponderación y poda

Como mencionamos, nuestro enfoque considera a los textos como representaciones de grafos conceptuales; por lo que el problema de resumir el texto (sus grafos conceptuales) se simplifica en seleccionar los nodos más importantes de las estructuras y quitar aquellos nodos que no se consideran importantes, manteniendo la coherencia estructural de los

grafos. Después de aplicar los procesos de selección y poda, los grafos resultantes se consideran como el resumen del texto a nivel conceptual.

Para determinar la importancia de los nodos proponemos utilizar el algoritmo HITS, el cual ha sido usado para la extracción de palabras clave para la generación de resúmenes extractivos (Mihalcea *et al*., 2005; Litvak *et al*., 2008). HITS es un algoritmo iterativo que toma en cuenta el grado de entrada y el grado de salida del nodo para determinar su importancia (véase la sección 2.6).

El método HITS proporciona dos métricas: AUTH y HUB (autoridad y concentración). Nuestro enfoque usa la métrica AUTH para determinar la importancia, ya que nos indica qué nodo es bueno como fuente de información; no obstante, se calculan ambas métricas ya que son dependientes. De manera que, un nodo con un valor alto en la métrica AUTH formará parte del grafo resumido, mientras que un nodo con valor AUTH por debajo de un umbral predefinido se excluirá del resumen.

El umbral que se preestablece es un parámetro que se define por el usuario, el cual indica el porcentaje de compresión del resumen, esto es, la compresión de las estructuras conceptuales.

Nosotros utilizamos una versión modificada del método HITS a la propuesta por Mihalcea *et al*. (2004). Para el cálculo de las métricas AUTH y HUB se usan las ecuaciones (4.1) y (4.2), las cuales consideran los flujos semánticos y la preferencia sobre algunos tópicos de interés predefinidos por el usuario.

$$AUTH(V_i) = \sum_{V_k \in I(V_i)} W_{ki} \cdot HUB(V_k) \cdot PREF(V_k) \qquad (4.1)$$

$$HUB(V_i) = \sum_{V_k \in O(V_i)} W_{ik} \cdot AUTH(V_k) \cdot PREF(V_k) \qquad (4.2)$$

Donde, I es el conjunto de los enlaces entrantes al nodo V_i. O es el conjunto de enlaces salientes del nodo V_i. W_{ki} es el peso del flujo que parte del nodo V_k hacia el nodo V_i. *HUB* y

AUTH indican el valor de las métricas autoridad y concentración para el nodo indicado. *PREF* indica la preferencia del usuario por el nodo V_k, lo cual representa un tópico.

La idea general del algoritmo se presenta en la figura 4.14. Consiste en calcular las métricas *AUTH* y *HUB* para el nodo *x*. La métrica *AUTH*(*x*) se calcula por medio de la suma del valor de *HUB* de todos los nodos a los que apuntan a *x*. Esto es, los nodos *n1*, *n2* y *n3* transmiten su peso de qué tan buenos son ellos para obtener información de la estructura gráfica. Si *x* obtiene un valor alto para *AUTH* indica que es un buen nodo como fuente de información.

El cálculo de *HUB*(*x*) se realiza por medio de la suma de *AUTH* de todos los nodos a los que apunta *x* (n1, n2, n3). Es decir, *x* contabiliza el peso de los nodos a los que conecta para identificar qué tan buenos son como fuente de información. Si *x* obtiene un valor alto para *HUB* indica que es un buen nodo para obtenerse información a través de él.

Ambas medidas son complementarias y se calculan para todos los nodos del grafo. Además, los valores que se transmiten a los nodos se aumentan o disminuyen de acuerdo al peso entre cada par de nodos, por ejemplo, *w*(*n*,*x*).

En las ecuaciones (1) y (2), se incorporan los pesos que definen los flujos entre los nodos, *w*(*n*,*x*). Además de la preferencia del nodo *x* como un posible tópico de interés.

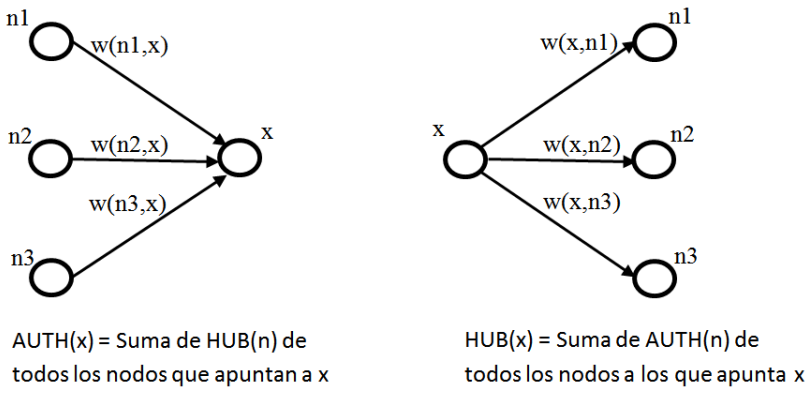


Figura 4.14 Operaciones para cálculo de métricas HITS

4.3.4 Reducción de grafos conceptuales

Para identificar los nodos más importantes en los grafos se aplican las ecuaciones (4.1) y (4.2) iterativamente hasta que converja el algoritmo o hasta un determinado número de

iteraciones previamente definidas. Mihalcea *et al*. (2004) usan de 20 a 30 iteraciones para convergir el algoritmo, otros autores usan una sola iteración (Litvak *et al*., 2008). Nosotros hemos identificado que 15 iteraciones son suficientes para nuestra colección de grafos, más iteraciones no mejoran la selección de los nodos.

Para seleccionar los nodos que consideramos importantes, llevamos a cabo los siguiente pasos.

1. Establecer las medidas AUTH y HUB asociadas a cada nodo al valor de 1.
2. Aplicar las operaciones AUTH y HUB a cada nodo.
3. Normalizar los valores AUTH y HUB por medio de la norma euclidiana.
4. Repetir los pasos 2 y 3 hasta convergir el algoritmo o alcanzar el número máximo de iteraciones.
5. Ordenar los nodos por la métrica AUTH en orden descendente.
6. Expandir los conceptos conectados para cada relación conceptual seleccionada.
7. Expandir los nodos asociados para conceptos verbales de acuerdo con su patrón verbal.
8. Seleccionar los conceptos sobresalientes de acuerdo al porcentaje de compresión establecido para la poda del grafo.
9. Los nodos seleccionados forman parte del resumen así como las relaciones que los unen. Si algún concepto no están relacionado a la estructura o a otros nodos no se considera al final.

En los pasos 1 a 5 se calculan los valores para el algoritmo HITS basado en las ecuaciones (4.1) y (4.2). El paso 6 aplica las reglas para expandir los conceptos que conectan a una relación conceptual, si es que fue seleccionada. Por ejemplo, la relación OBJ(*trigger*,*flooding*) (véase la tabla 5.3) se expande en dos conceptos *trigger* (provocar) y *flooding* (inundación).

El paso 7 aplica los patrones verbales asociados a los conceptos de tipo verbo para mantener la coherencia de la estructura. Por ejemplo, en la figura 4.1, el patrón verbal para el concepto verbal *chase* (perseguir) es 'NP V NP' (frase sustantiva / verbo / frase sustantiva), y el verbo es transitivo. El papel del primer NP es el agente, y el segundo NP es el tema. Ambas partes se requieren para que el concepto *chase* tenga semántica y estructura completa debido a que es definido como verbo transitivo. Por lo que, ambos, agente y tema deben formar parte del resumen (véase la sección 4.2.2).

En el paso 8, se aplica la operación de poda de acuerdo con un porcentaje de compresión establecido por el usuario. Aquí se seleccionan los nodos, considerados importantes, sin duplicarlos de acuerdo al porcentaje de compresión.

Por último, el paso 9 considera si algún nodo no tiene alguna relación con otro nodo, es decir, está aislado a la estructura, se descarta.

Los grafos resultantes después de aplicar todo el proceso son considerados como el resumen del texto a nivel conceptual (véase la tabla 5.2).

Capítulo 5
Resultados

En este capítulo, presentamos los resultados obtenidos con el modelo propuesto. Los experimentos se realizaron sobre una selección de documentos de la colección de datos DUC 2003. Los experimentos muestran que el modelo es viable para la generación automática de resúmenes.

En esta investigación se presenta un modelo para la generación automática de resúmenes con enfoque abstractivo basado en la representación semántica del texto por medio de grafos conceptuales.

Dado que la mayoría de los métodos actuales para la generación automática de resúmenes se centran en el enfoque extractivo (no se requiere de representaciones complejas del texto), por lo que hacen poco uso o ninguno de la semántica del documento para la generación del resumen.

Los métodos que hacen uso de alguna manera la semántica del texto no la implementan a un nivel granular fino y se pierden los detalles de las estructuras textuales, los cuales son importantes para la generación de un resumen abstractivo (Marcu, 2000; Barzilay, 2003; Barzilay *et al*., 2005; Litvak *et al*. 2008; da Cunha, 2008; Tsatsaronis *et al*. 2010).

No obstante, representar los textos semánticamente es complejo, debido a que no existen herramientas totalmente automatizadas, precisas y disponibles para trasformar un texto a otra representación intermedia tal como grafos conceptuales; pero tales representaciones con niveles de granularidad fina y flexibles de manipular son necesarias para lograr resúmenes abstractivos.

5.1 Evaluación del modelo

Uno de los componentes importantes para todo sistema es la forma en que se evalúa. En áreas como la generación automática de resúmenes el problema es complejo debido a la subjetividad de la tarea misma y de la propia evaluación. Esto es debido a que no existe un solo resumen correcto, ya que depende de los propósitos, el enfoque, así como de las necesidades particulares del lector (Spärck Jones, 2007; Nenkova *et al*., 2011, Torres-Moreno, 2011).

Como se mencionó en la sección 2.5, la forma estandarizada para evaluar a un sistema de generación automática de resúmenes es a través de una evaluación intrínseca automática. Es decir, el resumen generado por el sistema se compara contra los resúmenes hechos

manualmente por un humano. Aquí entra un factor importante de subjetividad que es la variabilidad del juicio humano.

La métricas comunes para la evaluación de los resúmenes en la competencia DUC es ROUGE (Chin-Yew, 2004), el cual encuentra la correlación entre los dos resúmenes el generado automáticamente y el creado manualmente. Se calcula básicamente la proporción de los *n-gramas* (traslapes de palabras) que comparten ambos resúmenes.

También, actualmente, se ha usado el método AutoSummENG (Giannakopoulos *et al*., 2011) para evaluar los resúmenes en la competencia MultiLing del TAC 2011 y 2013, el cual se basa en una estructura en forma de grafos.

5.1.1 Métricas utilizadas

Los dos métodos antes mencionados requieren que el resumen generado sea texto. Por lo que ninguno de los dos métodos podemos usar para evaluar los resultados de nuestro enfoque; ya que como se mencionó al inicio de esta investigación, la generación del texto a partir de los grafos conceptuales queda fuera del alcance de este trabajo.

No obstante, hemos realizado mediciones de los resultados obtenidos por nuestro enfoque. De la misma manera que las métricas como ROUGE y sus variantes, requerimos de los resúmenes modelo para poder comparar los resultados. Usamos métricas conocidas como Precisión, *Recall* y Medida-F (Nenkova A. , 2006) para la evaluación de los sistemas de generación de resúmenes.

La precisión la definimos en la ecuación 5.1 como la fracción de los conceptos elegidos por nuestro método que fueron correctos, esto es, se identifican los conceptos que el humano eligió en su resumen (resumen modelo) y que el método también los eligió.

Recall se define en la ecuación 5.2 como la fracción de conceptos elegidos por el humano que fueron correctamente identificados por el método.

Por último, la *Medida-F* se define en la ecuación 5.3 que es la media harmónica de Precisión y *Recall.*

$$\text{Precisión} = \frac{\text{Traslape de num. conceptos elegidos por el método y el humano}}{\text{Num. conceptos elegidos por el método}} \quad (5.1)$$

$$\textit{Recall} = \frac{\text{Traslape de num. conceptos elegidos por el método y el humano}}{\text{Num. conceptos elegidos por el humano}} \quad (5.2)$$

$$\text{Medida} - \text{F} = 2 \text{ x} \frac{\text{Precisión x } \textit{Recall}}{\text{Precisión} + \textit{Recall}} \quad (5.3)$$

En nuestro enfoque de evaluación, la forma en que consideramos un concepto correctamente seleccionado es comprobando que el concepto o su "concepto tipo se encuentra en el resumen modelo. Por ejemplo, para el concepto *storm* (tormenta) definido como [*Atmospheric_phenomenon*: *storm*] (Fenómeno_atmosférico: tormenta), si se encuentra *storm* o *Atmospheric_phenomenon* en el resumen modelo se contabiliza como un concepto correctamente seleccionado.

Se usa la jerarquía de conceptos tipo para realizar esta comprobación, ya que el concepto seleccionado puede ser cualquiera de sus abstracciones y alcanzable por medio de sus conceptos tipo, se establece un nivel límite para la comprobación, por ejemplo, un máximo de 5 niveles en la jerarquía.

5.1.2 Línea base

Nuestro modelo lo comparamos contra otro método simple que se construyó y es denominado línea base o *baseline*. La línea base que usamos consiste en extraer las palabras de los primeros párrafos hasta alcanzar el porcentaje de conceptos que se desean comparar.

En nuestro caso, se seleccionaron palabras de contenido como conceptos y se excluyeron las palabras denominadas vacías o *stopwords* tales como artículos y preposiciones.

Es importan enfatizar que la línea base como se considera en nuestra evaluación ha demostrado ser una línea base difícil de superar, ya que la mayoría de los sistemas de generación de resúmenes de un único documento evaluados en las competencias DUC para

esta tarea, no superaron a la línea base, de ahí que esta clase de resúmenes se descontinúo de la competencia y sigue siendo un problema abierto para la investigación (Nenkova *et al*., 2011).

El problema antes mencionado es en gran medida a la naturaleza de los documentos que se usan en la competencia. Los documentos son noticias y dada su naturaleza, en este género de documentos, las ideas principales siempre están en los primeros párrafos del documento. Por lo que la línea base capta adecuadamente las ideas principales del documento.

También se resalta en los resultados de la evaluación de la competencia DUC, que las personas que hicieron los resúmenes manualmente superaron significativamente a la línea base; lo que indica que esta clase de resúmenes es un reto importante y que mejores métodos que la línea base son posibles.

5.1.3 Definición del umbral de compresión

El propósito de un resumen es representar al texto original con una longitud mucho menor, pero manteniendo la mayor cantidad de las ideas principales del documento original. Con esto, es necesario un parámetro que indique el tamaño de reducción que se desea del documento original.

El parámetro que indica la cantidad de reducción del texto original es el umbral de compresión o porcentaje de compresión que el usuario establece. Este parámetro se define como la proporción de conceptos que queremos extraer del documento de texto original para conformar el resumen.

El umbral de compresión, como se mencionó, lo define el usuario, y se aplica tanto para la compresión de los grafos como para la extracción de los conceptos por el método considerado como la línea base.

5.2 Configuración de los experimentos

Los experimentos se realizaron con documentos seleccionados de la colección de artículos de noticias de la competencia DUC. Se eligió la colección del 2003 (DUC, 2003) ya que en este año se cuenta con resúmenes muy cortos a nivel de encabezados.

Los documentos elegidos son breves con longitud de entre 50 y 100 palabras aproximadamente. Se definieron tres grupos de documentos: 3 oraciones (grupo I), 4 oraciones (grupo II), y si superaban las 4 oraciones (grupo III). Cada grupo consta de 10 documentos representados como grafos conceptuales.

5.2.1 Descripción de los parámetros

Para llevar a cabo los experimentos de nuestro método, se configuración los siguientes parámetros.

Umbral de compresión

El parámetro del umbral de compresión se estableció al 20%, esto es, la reducción de los grafos a nivel conceptual debe estar en este rango, 20% de la longitud del texto original.

Sin embargo, en el caso de nuestro modelo, para algunas circunstancias, puede ser que se necesiten más conceptos excediendo el umbral establecido. Esto debido a que si un concepto verbal está en el límite del umbral y requiere de algunos conceptos adicionales que mantenga su estructura coherente (de acuerdo a su patrón verbal), entonces será necesario agregar los conceptos; de esta manera, se rebasará el umbral establecido. Este hecho se verá reflejado directamente en las métricas con una penalización.

Configuración HITS

Como mencionamos en la sección 4.3.3, para ponderar los nodos de los grafos conceptuales se usan las ecuaciones (4.2) y (4.3) que representan el modelo del método HITS. En el modelo HITS propuesto se integran dos factores: los pesos de los grafos conceptuales y las preferencias sobre un tópico de interés.

Pesos de los grafos conceptuales

En la sección 4.1, se presentó el esquema para el uso de los flujos semánticos en el contexto de los grafos conceptuales. Estos flujos se aumentan o disminuyen (valores en las métricas AUTH y HUB) a través de la estructura del grafo gracias a los pesos de las aristas que conectan a las relaciones conceptuales del grafo.

En la tabla 5.1, se describen el significado de los posibles valores para las aristas en el contexto de los flujos semánticos.

Tabla 5.1 Interpretación de los pesos asociados a las aristas con respecto al flujo semántico

Peso de la arista	Descripción
mayor que 1	El flujo semántico que pasa por la arista es de interés, aumentando su valor en la proporción indicada por el peso.
igual a 1	No se afecta el flujo semántico que se transmite, representa el valor neutro.
menor que 1	El flujo semántico que pasa por la arista no es de interés, y se penaliza reduciendo su valor en la proporción indicada.
igual a 0	Cancela el flujo semántico que pasa por la arista.

En este contexto, asumimos que los actores que realizan las acciones (los agentes) poseen un papel importante en el desarrollo del texto, en especial en el género de noticias. Por lo que se configuró el peso para todos los nodos que conectan relaciones agentivas (**AGNT**) con el valor de 2 tanto para los flujos de entrada como para los flujos de salida, es decir, duplicamos el valor del flujo que pasa por cada arista que asocia a un agente. El valor se determinó heurísticamente.

Para las demás relaciones como atributo (ATTR), objeto (OBJ), etc., se usó el valor neutro, valor de 1.

Preferencias del usuario

El otro factor que afecta a las métricas HITS son las preferencias del usuario por algún tópico de interés. Estas preferencias se representan con un valor asociado al tópico, nodo concepto, al cual le interesa al usuario. Este valor afecta directamente al flujo de información que se transmite a lo largo de la red y que pasa a través del tópico. En la tabla 5.2, se describen el significado de los posibles valores para el tópico.

Tabla 5.2 Interpretación de los valores asociados a los nodos (tópico) con respecto al flujo semántico

Peso del nodo (tópico)	Descripción
mayor que 1	El tópico es de interés y el flujo semántico que pasa por el nodo incrementa su valor en la proporción indicada por el peso.
igual a 1	No se afecta el flujo semántico que se transmite, representa el valor neutro.
menor que 1	El tópico no es de interés y el flujo semántico que pasa por el nodo se penaliza reduciendo su valor en la proporción indicada por el peso.
igual a 0	Cancela el flujo semántico que pasa por la arista.

Cabe mencionar que tanto para los pesos de las aristas como para las preferencias sobre los tópicos, nosotros consideramos sólo valores positivos; sin embargo, el modelo permite el uso de valores negativos. Usar valores negativos como pesos asociados a las aristas resultaría en una penalización para el flujo que pase por esa arista. El mismo efecto ocurriría para los tópicos de interés. Pero la combinación de valores negativos para ambos factores afectaría al modelo debido a su estructura. Por lo que se recomienda exclusión mutua para estos factores para los valores negativos.

Iteraciones del algoritmo

El método HITS es un método de reforzamiento mutuo por lo que requiere repetirse el proceso iterativamente hasta que converja, o repetir el proceso hasta un determinado número de iteraciones previamente definidas.

Como ya mencionamos algunos autores usan entre 20 a 30 iteraciones para convergir el algoritmo. Nosotros identificamos que 15 iteraciones son suficientes para nuestra colección de grafos, más iteraciones no mejoran la selección de los nodos.

5.3 Discusión de los resultados

Después de aplicar las operaciones propuestas en la sección 4.3 a los grafos conceptuales que representan los documentos fuente, se obtienen resultados similares, para cada documento fuente, a los que se presentan en las tablas siguientes.

En la tabla 5.3, se muestra un ejemplo de los nodos seleccionados por el método. En esta tabla se enfatizan los resultados obtenidos por el método de ponderación. Los resultados presentados se calcularon para el grafo presentado en la figura 4.5 que repetimos nuevamente aquí para facilitar la lectura (figura 5.1).

La primera columna de la tabla 5.3 representa al nodo seleccionado; la segunda columna representa la expansión de los conceptos que asocia un nodo de tipo relación. Es decir, que los nodos relación no son elegibles por sí mismos, y se deben considerar los conceptos que asocia la relación como nodos elegibles. Por ejemplo, la relación OBJ(*trigger-flooding*) los nodos que expande son los conceptos *trigger* y *flooding*; tales conceptos son elegibles para formar parte del resumen.

La tercera columna de la tabla 5.3 representa los valores para la métrica autoridad (AUTH), esta es la métrica que se usa para elegir a los nodos más importantes. Los nodos están ordenados descendentemente de acuerdo a la métrica AUTH. La última columna representa los valores para la métrica concentración (HUB).

En esta misma tabla aparece una marca (**req**) que indica que el concepto es requerido. Esta marca indica que el concepto fue agregado como concepto necesario para completar la coherencia de la estructura, en este caso se agregó para el concepto *kill* y *slam*.

El primer concepto (*kill*) es un concepto verbal y tiene un patrón asociado (NP)–(V)–(NP) donde empatan los conceptos. (*storm* / NP)–(*kill* / V)–(*People:@>=156* / NP) **(req)**.

El último NP es el concepto que hacía falta para completar la coherencia de esta estructura, por lo que se agregó a los conceptos elegibles.

El segundo concepto (*slam*) es también un concepto verbal y tiene un patrón asociado similar al anterior (NP)–(V)–(NP) donde empatan los conceptos. (*Typhoon:Babs* / NP)–(*slam* / V)–(*City:Hong Kong* / NP) **(req)**. Similarmente, como en el caso anterior, el último NP es el concepto que hacía falta para completar la coherencia de esta estructura, por lo que se agregó a los conceptos elegibles.

En la tabla 5.4, se muestran los conceptos finales seleccionados que formarán parte del resumen. Las relaciones que conectan a estos conceptos también se consideran como parte del resume, de hecho es a partir de ellas que se expanden los nodos concepto por lo que ya estaban consideradas (véase la tabla 5.3).

Tabla 5.3 Relaciones y conceptos seleccionados por el método de ponderación con expansión de relaciones conceptuales

Nodo	Expansión de relación	AUTH	HUB
Typhoon:Babs	-	0.729	0.3e–16
Atmospheric_phenomenon:storm	-	0.680	0.70e–03
AGNT(trigger-Typhoon:Babs)	trigger / Babs	0.054	0.147
OBJ(trigger-flooding)	trigger / flooding	0.027	0.10e–04
OBJ(trigger-landslide)	trigger / landslide	0.027	0.67e–05
LOC(trigger-City:Taiwan)	trigger / Taiwan	0.027	0.137
AGNT(kill-Atmospheric_phenomenon:storm)	kill / storm / People:@>=156 (**req**)	0.022	0.147
AGNT(slam-Typhoon:Babs)	slam / Typhoon:Babs / City:Hong Kong (**req**)	0.022	0.67e–05
LOC(kill-Country:Philippines)	kill / Philippines	0.011	0.38e–16

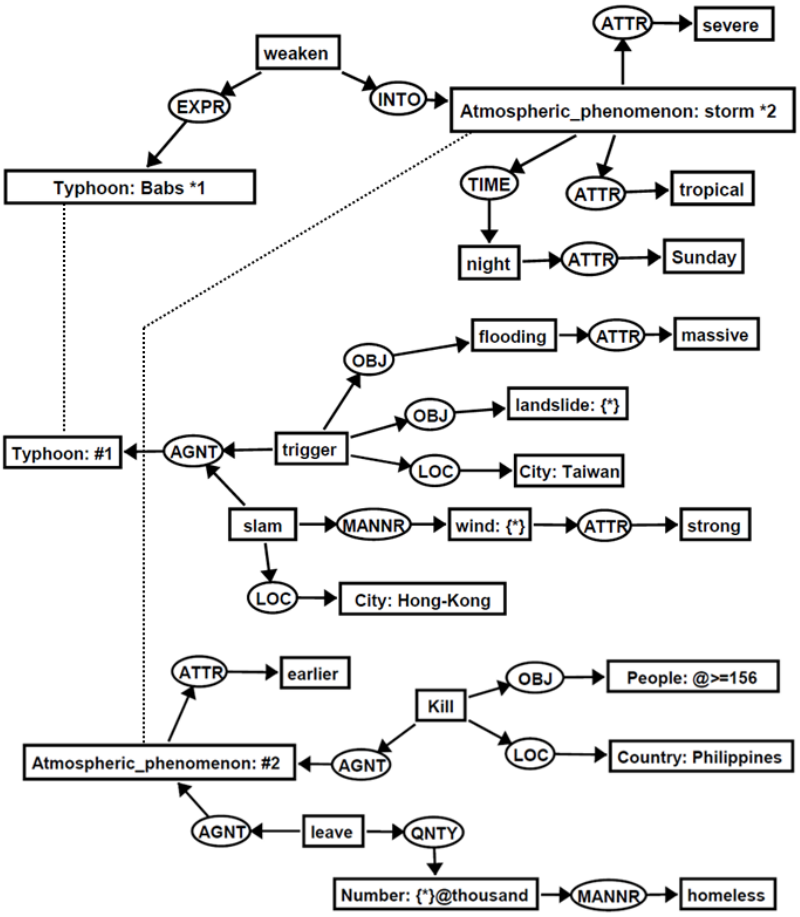


Figura 5.1. Ejemplo de noticia como grafo conceptual

Tabla 5.4 Conceptos finales seleccionados por el método de ponderación

Nodo	AUTH	HUB
Typhoon:Babs	0.729	0.3e-16
Atmospheric_phenomenon:storm	0.680	0.70e-03
trigger	0.054	0.147
flooding	0.027	0.10e-04
landslide	0.027	0.67e-05
City:Taiwan	0.027	0.137
kill	0.022	0.147
slam	0.022	0.67e-05
City:Hong Kong	0.022	0.147
People:@>=156	0.022	0.70e-03

Finalmente, los conceptos seleccionados en la tabla 5.4 representan al resumen, de acuerdo con el grafo de la figura 5.1, el cual puede leerse como el siguiente texto:

> "*Typhoon Babs triggered flooding and landslides in Taiwan. Typhoon Babs slammed Hong Kong. The storm killed at least 156 people.*"
>
> "El tifón Babs provocó inundaciones y deslizamientos de tierra en Taiwán. El tifón Babs golpeó a Hong Kong. La tormenta mató al menos a 156 personas."

En la tabla 5.5, se agrupan los resultados obtenidos por nuestro enfoque. Se muestra el promedio de la evaluación para la colección de grafos. Nuestro método supera en promedio 11% a la línea base.

Tabla 5.5 Evaluación del método

	Precisión		***Recall***		**Medida-F**	
	Línea base	**Método**	**Línea base**	**Método**	**Línea base**	**Método**
Grupo I	0.45	0.52	0.44	0.67	0.45	0.58
Grupo II	0.53	0.68	0.53	0.74	0.53	0.71
Grupo III	0.56	0.66	0.56	0.69	0.56	0.67
Promedio	0.51	0.62	0.51	0.70	0.51	0.65

El grupo I que consta de 3 oraciones tenemos una precisión promedio que supera en 7% a la línea base; para el grupo II que consta de 4 oraciones, la precisión promedio la supera en 15%; y para el grupo III que consta de más de 4 oraciones, la precisión promedio la supera en 10%.

Los datos presentados en el grupo I nos indican que el método para textos muy breves casi se desempeña igual que un método más simple de implementar y computacionalmente menos costos y sin usar tantos recursos lingüísticos como el método que presentamos.

No obstante, para los datos mostrados para el grupo II y III, se puede inferir que para textos a nivel de párrafos, nuestro método puede identificar aceptablemente (68% en promedio) los conceptos relevantes del texto analizado. De lo anterior, podemos deducir que el método se desempeña mejor debido a que, a nivel de párrafo, se tiene un texto más estructurado y cohesionado, por lo que el método aprovecha esa característica del texto al representarse como grafo conceptual.

También podemos observar que la línea base va mejorando mientras aumentamos la cantidad de oraciones. Sin embargo, el método propuesto se mantiene por arriba de la línea base.

Se esperaba que la línea base mejorara ya que hay estudios que han demostrado que las primeras y las últimas oraciones en los párrafos son buenos indicadores para identificar la información relevante (Luhn, 1958; Baxendale, 1958; Hovy *et al*. 1999).

Aunado a lo anterior, la competencia DUC reportó los resultados para la tarea de generación de resúmenes de un solo documento para los años 2001 y 2002, donde ningún sistema participante superó a la línea base, la cual consistía en seleccionar las primeras 100 palabras del documento (Nenkova A. , 2005). Los documentos usados en esos años eran aproximadamente entre 350–800 palabras.

El problema del bajo desempeño de los sistemas que generan resúmenes de un solo documento en comparación con la línea base surge debido a la naturaleza de los documentos. Los documentos son textos de noticias, debido a las características de escritura de este género de documentos, las partes más importantes del documento se colocan al inicio de los párrafos.

Capítulo 6
Conclusiones

En este capítulo se presentan las aportaciones de esta investigación. Así como las limitaciones de nuestro enfoque. También describimos las líneas de trabajo futuras; y por último se listas las publicaciones y trabajos alrededor de esta investigación.

La mayoría de los trabajos realizados en el área de generación automática de resúmenes se orienta al enfoque extractivo; y los métodos que se orientan al enfoque abstractivo usan parte de la semántica de los documentos o no la usan definitivamente. Aunado a esto, se ha relegado el estudio para la generación de resúmenes de un solo documento y se han enfocado más al estudio de los resúmenes multidocumento. En parte porque los foros de competencia como DUC/TAC promueven otros retos para construir resúmenes como resúmenes de opiniones, resúmenes multilingüe, etc.

Una idea para mejorar la generación de resúmenes de un solo documento es el uso de representaciones expresivas que detallen la semántica del texto. Con ello, se espera que se "entienda" mejor lo que se tiene representado.

De acuerdo a lo anterior, en esta investigación propusimos un nuevo modelo para la generación de resúmenes abstractivos monodocumento basado en una semántica detallada. Este modelo usa los grafos conceptuales como representación intermedia de los textos, la cual se aprovecha para "entender" las relaciones importantes del texto que conllevan la semántica subyacente.

El enfoque se basó en manipular a los grafos conceptuales por medio de las operaciones de generalización, unión, ponderación y poda para reducir las estructuras conceptuales.

Una de las etapas principales para simplificar las estructuras fue la etapa de selección de nodos, la cual se propuso resolver con una variación del método HITS, y apoyada por los grafos conceptuales ponderados como representación del texto.

Adicionalmente, se propuso el uso de recursos lingüísticos como VerbNet para mantener la coherencia de las estructuras resumidas por medio de sus patrones verbales.

Así también, se mostró cómo los flujos semánticos vinculados a los grafos conceptuales ponderados proporcionan un esquema flexible para la creación de resúmenes orientados a los intereses del usuario ya sean por la preferencia de los tópicos, o por el interés de ciertos actores inherentes a los grafos tales como los agentes, los lugares, los temas, etc.

Finalmente, evaluamos nuestro método con documentos breves de la colección de datos DUC. A pesar de realizar la evaluación con noticias cortas, con lo que la información más importante se encuentra al inicio de éstas, se superó a la línea base con un promedio del 11%.

Lo anterior demuestra que el método funciona para textos cortos y un indicio de que puede funcionar para textos más largos.

6.1 Aportaciones

Las contribuciones científicas de este trabajo son las siguientes:

- Modelo para la generación de resúmenes abstractivos de un solo documento basado en un conjunto de operaciones que operan sobre los grafos conceptuales que representan al texto.
- Conjunto de operaciones para manipular a los grafos conceptuales.
- Método para la selección de nodos importantes en un contexto de grafos conceptuales a partir la conectividad y de la información sintáctica y semántica de la estructura.
- Conceptualización de una nueva clase de grafos conceptuales, a saber, grafos conceptuales ponderados, los cuales expresan flujos semánticos por medio de las relaciones conceptuales.

Las contribuciones técnicas de este trabajo son las siguientes:

- Implementación de los algoritmos de las operaciones para la manipulación de los grafos conceptuales.
- Conjunto de grafos para la evaluación del modelo.
- Implementación de algoritmo para la extracción de la jerarquía de conceptos de WordNet
- Implementación de algoritmo para la extracción de patrones de VerbNet

6.2 Limitaciones del método

El modelo para la generación de resúmenes propuesto en este trabajo tiene las siguientes limitaciones:

- **Es dependiente del lenguaje**

  La principal desventajas de nuestro modelo es que requiere de recursos lingüísticos externos como WordNet y VerbNet, lo cual lo hacen dependiente del idioma. Por lo que, no podemos aplicar el modelo a lenguajes que carezcan de estos recursos.

- **Generación de los grafos conceptuales a partir del texto**

  Crear automáticamente los grafos conceptuales tiene sus retos como se mencionó en el capítulo 4 y no hay herramientas disponibles para realizar esta tarea. Por lo que, se tienen que crear semiautomáticamente y realizar los ajustes necesarios manualmente. De ahí que para probar el modelo a gran escala será necesario construir los grafos que representen a los textos largos lo cual consume tiempo y se torna compleja su manipulación para realizar los ajustes y agregar la información semántica.

- **Generación del texto a partir de los grafos conceptuales**

  Se requiere generar el texto a partir de los grafos obtenidos para evaluar los resultados de nuestro modelo con métricas como ROUGE y sea comparable con otros enfoques.

  Recordamos que la métrica ROUGE requiere del resumen textual generado por el sistema para compararlo contra un resumen modelo que es generado por un humano; entre más similar sea el resumen generado por el sistema al resumen generado por el humano, se considera un mejor sistema.

6.3 Trabajo futuro

En esta investigación nos centramos en proporcionar un modelo para la generación de resúmenes de un solo documento dentro del marco de trabajo de los grafos conceptuales. Asumimos que se cuenta con los grafos conceptuales y a partir de ahí se proyectó el modelo.

Los siguientes puntos presentan el trabajo futuro que puede realizarse alrededor de esta investigación.

- Probar el método con documentos grandes (al menos 1000 palabras), por lo que es necesario construir los grafos conceptuales, cuyo tamaño complica administrar su estructura.

- Probar otros algoritmos de ponderación en nuestro modelo de trabajo como PageRank y SemanticRank.

- Extender el modelo para contemplar resúmenes multidocumento, puesto que la necesidad de resumir es mayor, y tiene requerimientos diferentes a los que tienen los resúmenes de un solo documento.

- Desarrollar un método que transforme los textos a grafos conceptuales para automatizar el proceso.

- Desarrollar un método que genere el texto a partir de los grafos conceptuales, con el fin de contrastar nuestra propuesta con otros enfoques y utilizar métricas como ROUGE que normalmente se usan para evaluar a los sistemas de generación de resúmenes.

6.4 Publicaciones

- Elhadad, M., Miranda-Jiménez, S., Steinberger, J., & Giannakopoulos, G. (2013). **Multi-document multilingual summarization corpus preparation, Part 2: Czech, Hebrew and Spanish**. The 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria.
- Miranda-Jiménez, S., Gelbukh, A., & G. Sidorov. **Abstractive Summarization of Short Texts Using Conceptual Graphs.** Revista Polibits. Manuscrito aceptado: 24 de junio de 2013.
- Miranda-Jiménez, S., Gelbukh, A., & G. Sidorov. **Generación de resúmenes por medio de síntesis de grafos conceptuales.** Revista Signos. Manuscrito enviado: abril de 2013, (en revisión).
- Miranda-Jiménez, S., Gelbukh, A., & Sidorov, G. S**ummarizing Conceptual Graphs for Automatic Summarization Task**. In: Pfeiffer, H.D., Ignatov, D.,Poelmans, J. (eds.). 20th International Conference on Conceptual Structures: Conceptual Structures for Knowledge Representation for STEM Research and Education (ICCS 2013), Mumbai, India, January 10-12, 2013. LNCS (LNAI), vol. 7735, pp. 245-253. Springer, Heidelberg (2013).
- Miranda-Jiménez, S., Gelbukh, A., & Sidorov, G. **Model for Automatic Text Summarization based on Conceptual** Graphs. Miguel González Mendoza, Oscar Herrera Alcántara (Eds.). In: Proceedings of the Doctoral Consortium at the 10th Mexican International Conference on Artificial Intelligence, MICAI-2011, November 28, 2011, Puebla, Mexico, ISBN 978-607-95367-4-9
- Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Treviño, A., & Gordon, J. **Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets**. LNAI 7630, MICAI 2012, November 2012, pp. 1-14.
- Viveros-Jiménez, F., Sidorov, G., Castillo Velásquez, F., Castro- Sánchez, N., Miranda-Jiménez, S., Treviño, A., & Gordon, J. **Sondeos automatizados en**

las redes sociales a través de la minería de opinión. Komputer Sapiens, Año IV Vol. II, Julio-Diciembre, 2012.

6.4.1 Ponencias sin publicación

- **Ponderación ordenada de grafos conceptuales para la generación de resúmenes**. 5to. Coloquio de Lingüística Computacional (UNAM), agosto 2011, Ciudad de México, México.
- Poster: **Automatic Text Summarization based on Conceptual Graphs**. 8vo Taller de Tecnologías del Lenguaje Humano (INAOE), noviembre 2011, Tonantzintla, Puebla, México.
- Poster: **Generación de resúmenes con métodos simbólicos**. 6to Taller de Tecnologías del Lenguaje Humano (INAOE), octubre 2009, Tonantzintla, Puebla, México.

6.5 Recursos generados

Recursos desarrollados asociados al Laboratorio de Procesamiento del Lenguaje Natural durante el desarrollo de esta tesis.

- **Corpus en español para la evaluación de sistemas de generación de resúmenes multidocumento y multilingüe**
  Disponible a través de la competencia MultiLing 2013
  http://multiling.iit.demokritos.gr/pages/view/662/multiling-2013
- **Corpus en inglés para evaluar los sistemas de detección de plagio**
  Corpus basado en ofuscación de resúmenes basado en la colección de documentos DUC 2003 y DUC 2006.
  Disponible a través de la competencia PAN-2013
  http://pan.webis.de/

Referencias

Barzilay, R. (2003). *Information Fusion for MultiDocument Summarization: Paraphrasing and Generation.* Thesis (Ph.D.), Columbia University, New York, USA.

Barzilay, R., & Elhadad, M. (1997). Using Lexical Chains for Text Summarization. *Proceedings of the ACL EACL'97 Workshop on Intelligent Scalable Text Summarization*, (págs. 10–17). Madrid.

Barzilay, R., & McKeown, K. R. (2005). Sentence Fusion for Multidocument News Summarization. *Computational Linguistics, 31*(3), 297-328.

Baxendale, P. (1958). Machine-made index for technical literature: an experiment. *IBM Journal of Research and Development, 2*(4), 354-361.

Boudin, F., & Torres-Moreno, J. M. (2007). NEO-CORTEX: a performant user-oriented multi-document summarization system. *Computational Linguistics and Intelligent Text Processing* (págs. 551-562). Springer Berlin Heidelberg.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics, 32*(1), 13-47.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, (págs. 335 - 336). Melbourne, Australia.

Chein, M., & Mugnier, M.-L. (2009). *Graph-based Knowledge Representation: Computational Foundations of Conceptual Graphs.* London: Springer-Verlag.

Chin-Yew, L. (2004). ROUGE: a Package for Automatic Evaluation of Summaries. *Proceedings of the Workshop on Text Summarization Branches Out.* Barcelona, Spain.

Chin-Yew, L., & Hovy, E. (1997). Identifying Topics by Position. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, (págs. 283–290). Washington.

Cohn, T., & Lapata, M. (2009). Sentence compression as tree transduction. *Journal of Artificial Intelligence Research, 34*, 637-674.

da Cunha, I. (2008). Hacia un modelo lingüístico de resumen automático de artículos médicos en español. Ph.D. thesis, IULA, Barcelona, España.

da Cunha, I., Torres-Moreno, J. M., Velázquez-Morales, P., & Vivaldi, J. (2009). Un algoritmo lingüístico-estadístico para resumen automático de textos especializados. *Linguamática*, 67-79.

da Cunha, I., Torres-Moreno, J.-M., Sierra, G., Cabrera-Diego, L. A., Castro Rolón, B. G., & Rolland Bartilotti, J. M. (2011). The RST Spanish Treebank On-line Interface. *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, (págs. 698-703). Hissar, Bulgaria.

da Cunha, I., Wanner, L., & Cabré, T. (2007). Summarization of specialized discourse: The case of medical articles in Spanish. (J. B. Company, Ed.) *Terminology, 13*(2), 249-286.

Dang, H. T. (2005). Overview of DUC 2005. *Proceedings of the Document Understanding Conf. Wksp. 2005 (DUC 2005) at the Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP*.

de Marneffe, M.-C., & Manning, C. D. (11 de 2008). *Stanford Parser Manual*. Recuperado el 15 de 01 de 2013, de Stanford Parser:

http://nlp.stanford.edu/software/dependencies_manual.pdf

de Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. *5th International Conference on Language Resources and Evaluation (LREC 2006)*, (págs. 449-454). GENOA, ITALY.

DUC. (2003). *Document Understanding Conference*. Recuperado el 15 de 09 de 2010, de http://duc.nist.gov/pubs.html#2003

Edmundson, H. P. (1969). New Methods in Automatic Extracting. *Journal of the Association for Computing, 16*(2), 264-285.

Elhadad, M., Miranda-Jiménez, S., Steinberger, J., & Giannakopoulos, G. (2013). Multi-document multilingual summarization corpus preparation, Part 2: Czech, Hebrew and Spanish. *The 51st Annual Meeting of the Association for Computational Linguistics.* Sofia, Bulgaria.

Erkan, G., & Radev, D. (2004b). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research, 22*(1), 457-479.

Erkan, G., & Radev, D. R. (2004a). LexPageRank: Prestige in Multi-Document Text Summarization. *EMNLP*, (págs. 365-371).

Fellbaum, C. (Ed.). (1998). *WordNet: An Electronic Lexical Database.* Cambridge MA: MIT Press.

Fernández, S., SanJuan, E., & Torres-Moreno, J. (2007). Textual Energy of Associative Memories: Performant Applications of Enertex Algorithm in Text Summarization and Topic Segmentation. En A. Gelbukh, & Á. Kuri Morales (Ed.), *MICAI 2007: Advances in Artificial Intelligence* (págs. 861-871). Springer Berlin Heidelberg.

Fillmore, C., & Atkins, B. T. (1992). Towards a Frame-Based Lexicon: the Case of RISK. En A. Lehrer, & E. Kittay, *Frames and Fields* (págs. 75-102). Erlbaum Publ.

Ganesan, K., Zhai, C., & Han, J. (2010). Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, (págs. 340–348). Beijin, China.

Genest, P. E., & Lapalme, G. (2011). Framework for abstractive summarization using text-to-text generation. (págs. 64-73). Association for Computational Linguistics.

Genest, P.-E., & Lapalme, G. (2012). Fully Abstractive Approach to Guided Summarization. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, (págs. 354–358). Jeju, Korea.

Giannakopoulos, G., & Karkaletsis, V. (2010). Summarization system evaluation. *TAC 2010 Workshop.* Maryland, USA.

Giannakopoulos, G., & Karkaletsis, V. (2011). AutoSummENG and MeMoG in Evaluating Guided Summaries. *TAC 2011 Workshop NIST Gaithersburg, Maryland, USA*.

Hahn, U., & Mani, I. (2000). The Challenges of Automatic Summarization. *IEEE Computer, 33*(11), 29-36.

Halliday, M. (1985). *An introduction to Funtional Grammar.* (E. Arnold, Ed.) London.

Halliday, M., & Hasan, R. (1976). *Cohesion in English.* Longman.

Hensman, S. (2005). Constructing conceptual graphs using linguistic resources. *Proceedings of the 4th WSEAS International Conference on Telecommunications and Informatics*, (págs. 1-6). Prague, Czech Republic.

Hensman, S., & Dunnion, J. (2004). Automatically Building Conceptual Graphs Using VerbNet and WordNet. *Proceedings of the 3rd International Symposium on Information and Communication Technologies*, (págs. 115-120). Las Vegas, USA.

Hirst, G., & St-Onge, D. (1998). *Lexical chains as representations of context for the detection and correction of malapropisms.* Cambridge, MA: The MIT Press,.

Hovy, E. (2005). Automated text summarization. En R. MitKov, *The Oxford Handbook of Computational* (págs. 583-598). Oxford: Oxford University Press.

Hovy, E., & Chin-Yew, L. (1999). Automating Text Summarization in SUMMARIST. En I. Mani, & M. T. Maybury (Edits.), *Advances in Automatic Text Summarization* (págs. 81-94). Cambridge MA: MIT Press.

Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar*. Cambridge : MIT Press.

Jing, H. (2001). *Cut-and-Paste Text Summarization.* Thesis (Ph.D.) Columbia University, Columbia.

Kipper, K., Trang Dang, H., & Palmer, M. (2000). Class-Based Construction of a Verb Lexicon. *Proceedings of Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, (págs. 691–696). Austin, Texas, USA.

Kleinberg, J. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM, 46*(5), 604–632.

Knight, K., & Marcu, D. (2000). Statistics-Based Summarization - Step One: Sentence Compression. *The 17th National Conference of the American Association for Artificial Intelligence AAAI'2000*, (págs. 703-710). Austin, Texas,USA.

Knight, K., & Marcu, D. (2002). Summarization beyond sentence extraction: A Probabilistic Approach to Sentence Compression. *Artificial Intelligence, 139*(1), 91–107.

Kupiec, J., Pedersen, J., & Chen, F. (1995). A Trainable Document Summarizer. *Proceedings of the 18th Annual International ACM, 1995 SIGIR Conference on Research and Development in Information Retrieval*, (págs. 68-73). Seattle, USA.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Levin, B. (1993). *English Verb Classes And Alternations: A Preliminary Investigation.* The University of Chicago Press.

Litvak, M., & Last, M. (2008). Graph-based keyword extraction for single-document summarization. *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, (págs. 17-24). Manchester, United Kingdom.

Luhn, H. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development, 2*(2), 159-165.

Mani, I. (2001). Summarization Evaluation: An Overview. *Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL).*

Mani, I., Klein, G., House, D., & Hirschmans, L. (2002). SUMMAC: a text summarization evaluation. *Natural Language Engineering, 1*(8), 43-68.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text 8, 3,* 243-281.

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

Marcu, D. (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts.* Thesis (Ph.D), University of Toronto, Toronto, Canada.

Marcu, D. (1999). Discourse Trees Are Good Indicators of Importance in Text. En I. Mani, & M. T. Maybury (Edits.), *Advances in AutomaticText Summarization* (págs. 123-136). Cambridge MA: MIT Press.

Marcu, D. (2000). *The Theory and Practice of Discourse and Summarization.* Cambridge MA: MIT Press.

McKeown, K. R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J., Nenkova, A., y otros. (2002). Tracking and summarising news on a daily basis with Columbia's Newsblaster. *Proceedings of the human.*

McKeown, K., & Radev, D. (1995). Generating Summaries of Multiple News Articles. *SIGIR '95 Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, 74-82.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice.* Albany: State University Press of New York.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Texts. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, (págs. 404-411). Barcelona, Spain.

Mihalcea, R., & Tarau, P. (2005). An Algorithm for Language Independent Single and Multiple Document Summarization. *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, (págs. 602–607). Jeju, Korea.

Miranda-Jiménez, S., Gelbukh, A., & Sidorov, G. (2013). Summarizing Conceptual Graphs for Automatic Summarization Task. *Conceptual Structures for STEM Research and Education. 20th International Conference on Conceptual Structures, ICCS 2013*, (págs. 245-253). Mumbai, India.

Molina, A., da Cunha, I., Torres-Moreno, J., & Velázquez-Morales, P. (2011). La compresión de frases : un recurso para la optimización de resumen automático de documentos. *Linguamática, 2*(3), 13–27.

Molina, A., Torres-Moreno, J. M., da Cunha, I., SanJuan, E., & Sierra, G. (2012). Sentence Compression in Spanish driven by Discourse Segmentation and Language Models. *Cornell University ArXiv :1212.3493.* Computation and Language (cs.CL), Information Retrieval (cs.IR) Vol.1212.

Molina, A., Torres-Moreno, J. M., SanJuan, E., da Cunha, I., & Sierra, G. (2013). Discursive Sentence Compression. *Computational Linguistics and Intelligent Text Processing* (págs. 394–407). Springer Berlin Heidelberg.

Montes-y-Gómez, M., Gelbukh, A. F., López-López, A., & Baeza-Yates, R. A. (2001). Flexible Comparison of Conceptual Graphs. *Proc. DEXA-2001, 12th International Conference and Workshop on Database and Expert Systems Applications*, (págs. 102-111). Munich, Germany.

Montes-y-Gómez, M., Gelbukh, A., & López-López, A. (2000b). Comparison of Conceptual Graphs. *MICAI 2000: Advances in Artificial Intelligence, Mexican*

International Conference on Artificial Intelligence, (págs. 548-556). Acapulco, Mexico.

Montes-y-Gómez, M., López-López, A., & Gelbukh, A. (2000a). Information Retrieval with Conceptual Graph Matching. *Proceedings of 11th International Conference on Database and Expert Systems Applications DEXA-2000.* London, UK.

Nenkova, A. (2005). Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference. *Proceedings of the 20th national conference on Artificial intelligence* (págs. 1436-1441). Pittsburgh, Pennsylvania: AAAI Press.

Nenkova, A. (2006). Summarization Evaluation for Text and Speech: Issues and Approaches. *INTERSPEECH-2006, paper 2079-Wed1WeS.1*.

Nenkova, A., & McKeown, K. (2011). Automatic Summarization. *Foundations and Trends in Information Retrieval*, 103-233.

Nenkova, A., & Passonneau, R. (2004). Evaluating Content Selection in Summarization: The Pyramid Method. *HLT/NAACL*.

Nicolov, N., Mellish, C., & Ritchie, G. (1995). Sentence Generation from Conceptual Graphs. *Conceptual Structures: Applications, Implementation and Theory*, (págs. 74-88). Santa Cruz, CA, USA.

Ordoñez-Salinas, S., & Gelbukh , A. F. (2010). Generación de Grafos Conceptuales. En M. González Mendoza, & M. Herrera Alcántara (Ed.), *Avances en sistemas inteligentes en México. Sociedad Mexicana de Inteligencia Artificial*, (págs. 139–150). Ciudad de México, México.

Page, L., & Brin, S. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*, 1-7.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web.

Pardo, T. A., Rino, L. H., & Nunes, M. D. (2003). GistSumm: A summarization tool based on a new extractive method. *Computational Processing of the Portuguese Language* (págs. 210-218). Springer Berlin Heidelberg.

Radev, D. (1999). *Generating Natural Language Summaries from Multiple On-Line Sources: Language Reuse and Regeneration.* Thesis (Ph.D.), Columbia University, New York, USA.

Radev, D. A., Blair-Goldensohn, S., Blitzer, J., Celebi, A., Dimitrov, S., Drabek, E., y otros. (2004). MEAD-a platform for multidocument multilingual text summarization. *Proceedings of the 4th International Conference on Language Resources and Evaluation*.

Saggion, H., Torres-Moreno, J. M., da Cunha, I., & SanJuan, E. (2010). Multilingual summarization evaluation without human models. *Proceedings of the 23rd International Conference on Computational Linguistics* (págs. 1059-1067). Association for Computational Linguistics.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval.* New York: McGraw-Hill.

Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic Text Structuring and Summarization. *Information Processing & Management, 33*(2), 193-207.

Santorini, B. (1990). *Part-of-speech Tagging Guidelines for the Penn Treebank Project.* Reporte técnico MS-CIS-90-47, University of Pennsylvania, Department of Computer and Information Science.

Soricut, R., & Marcu, D. (2003). Sentence Level Discourse Parsing using Syntactic and Lexical Information. *Proceedings of HLT-NAACL 2003*, (págs. 149-156). Edmonton.

Sowa, J. F. (1984). *Conceptual Structures: Information Processing in Mind and Machine.* Reading, MA: Addison-Wesley.

Sowa, J. F. (Ed.). (1991). *Principles of Semantic Networks: Explorations in the Representation of Knowledge.* San Mateo, CA: Morgan Kaufmann Publishers.

Sowa, J. F. (1999). *Knowledge Representation: Logical, Philosophical, and Computational Foundations.* Pacific Grove, CA: Brooks Cole Publishing Co.

Spärck Jones, K. (1999). Automatic Summarising: Factors and Directions. En I. Mani, & M. T. Maybury (Edits.), *Advances in Automatic Text Summarization* (págs. 1-12). Cambridge MA: MIT Press.

Spärck Jones, K. (2007). Automatic summarising: The state of the art. *Information Processing & Management, 43*(6), 1449-1481.

Sparck-Jones, K., & Galliers, J. R. (1996). Evaluating Natural Language Processing Systems: An Analysis and Review. *Lecture Notes in Artificial Intelligence 1083.* Springer-Verlag.

Steinberger, J., & Jezek, K. (2006). Sentence compression for the lsa-based summarizer. *Proceedings of the 7th International Conference on Information Systems Implementation and Modelling*, (págs. 141–148).

Thakkar, K., Dharaskar, R., & Chandak, M. (2010). Graph-Based Algorithms for Text Summarization. *3rd International Conference on Emerging Trends in Engineering and Technology (ICETET)*, (págs. 516 - 519).

Torres-Moreno, J. M. (2011). *Résumé Automatique de Documents.* Lavoisier.

Torres-Moreno, J. M. (2012). Artex is AnotheR TEXt summarizer. *CoRR abs/1210.3312*.

Torres-Moreno, J. M., Saggion, H., da Cunha, I., SanJuan, E., & Velázquez-Morales, P. (2010). Summary Evaluation with and without References. *Polibits*(42), 13-20.

Torres-Moreno, J. M., Velázquez-Morales, P., & Meunier, J. (2001). Cortex: un algorithme pour la condensation automatique de textes. *ARCo*, (págs. 65–75). Lyon, France.

Tsatsaronis, G., Varlamis, I., & Nørvåg, K. (2010). SemanticRank: ranking keywords and sentences using semantic graphs. *Proceedings of the 23rd International Conference on Computational Linguistics*, (págs. 1074-1082). Beijing, China.

Vandeghinste, V., & Pan, Y. (2004). Sentence Compression for Automated Subtitling: A Hybrid Approach. *Proc. of ACL Workshop on Summarization*, (págs. 89–95).

Vivaldi, J., & Rodríguez, H. (2001). Improving term extraction by combining different techniques. *Terminology, 7*(1), 31–47.

Vivaldi, J., da Cunha, I., Torres-Moreno, J. M., & Velázquez-Morales, P. (2010). Automatic Summarization Using Terminological and Semantic Resources. *LREC.*

Wang, W., Wei, F., Li, W., & Li, S. (2009). HyperSum: Hypergraph Based Semi-Supervised Sentence Ranking for Query-Oriented Summarization. *CIKM '09 Proceeding of the 18th ACM conference on Information and knowledge management* , (págs. 1855-1858). Hong Kong, China.