Revista Signos
ISSN: 0035-0451
revista.signos@ucv.cl
Pontificia Universidad Católica de Valparaíso
Chile

Miranda Jiménez, Sabino; Gelbukh, Alexander; Sidorov, Grigori
Generación de resúmenes por medio de síntesis de grafos conceptuales
Revista Signos, vol. 47, núm. 86, diciembre-, 2014, pp. 463-485
Pontificia Universidad Católica de Valparaíso
Valparaíso, Chile

Revista Signos. Estudios de Lingüística ISSN 0718-0934
© 2014 PUCV, Chile · DOI: 10.4067/S0718-09342014000300006 · 47(86) 463-485

# Generación de resúmenes por medio de síntesis de grafos conceptuales*

*Generating summaries by means of synthesis of conceptual graphs*

Sabino Miranda Jiménez
Instituto Politécnico Nacional
México
sabino_m@hotmail.com

Alexander Gelbukh
Instituto Politécnico Nacional
México
gelbukh@gelbukh.com

Grigori Sidorov
Instituto Politécnico Nacional
México
sidorov@cic.ipn.mx

**Recbido:** 08-IV-2013 / **Aceptado:** 27-XII-2013

## Resumen

En esta investigación, proponemos un modelo para la generación de resúmenes abstractivos de un solo documento, basado en la representación conceptual del texto. Aunque hay investigaciones que toman en cuenta la representación sintáctica o semántica parcial del texto, hasta ahora, una representación semántica completa de textos no se ha usado para la generación de resúmenes. Nuestro modelo usa una representación semántica completa del texto por medio de estructuras de grafos conceptuales. En este contexto, la tarea de la generación del resumen se reduce a resumir el conjunto de los grafos conceptuales correspondientes. Para realizar esto, se aplica un conjunto de operaciones sobre los grafos: generalización, unión o asociación, ponderación y poda. Además, se usan una jerarquía de conceptos (WordNet) y reglas heurísticas basadas en los patrones semánticos de VerbNet para apoyar a las operaciones. El conjunto resultante de grafos representa al resumen del texto a nivel conceptual. El método se evaluó con la colección de datos DUC 2003. Los resultados muestran que el método es efectivo para resumir textos cortos.

**Palabras Clave**: Resúmenes abstractivos, grafos conceptuales ponderados, algoritmos de ponderación basada en grafos, algoritmo HITS.

## Abstract

In this study, we propose a model for generating single-document abstractive summaries, based on the conceptual representation of the text. Although there are studies that take into account the partial syntactic or semantic representation of the text, so far, a complete semantic representation of texts has not been used for generating summaries. Our model uses a complete semantic representation of text by means of conceptual graph structures. In this context, the task of generating the summary is reduced to summarize the set of corresponding conceptual graphs. In order to do this, a set of operations on graphs is applied: generalization, join or association, ranking, and pruning. Furthermore, a hierarchy of concepts (WordNet) and heuristic rules based on the semantic patterns from VerbNet are used in order to support such operations. The resulting set of graphs depicts the text summary at the conceptual level. The method was evaluated on the DUC 2003 data collection. The results show that the method is effective for summarizing short texts.

**Key Words:** Abstractive summarization, weighted conceptual graphs, graph-based ranking algorithm, HITS algorithm.

## INTRODUCCIÓN

Actualmente, se cuenta con gran cantidad de información textual, pero las personas no tienen suficiente tiempo para leerla, analizarla y tomar decisiones importantes basándose en ella. De ahí que las tecnologías para la generación automática de resúmenes, capaces de presentar la información de manera concisa, estén adquiriendo gran importancia. La generación automática de resúmenes textuales de calidad es una tarea desafiante debido a que involucra análisis y comprensión del texto, el uso de información del dominio y generación de lenguaje natural.

Se han propuesto diferentes clasificaciones para la generación automática de resúmenes (Spärck Jones, 1999; Hovy & Chin-Yew, 1999; Nenkova & McKeown, 2011; Lloret & Palomar, 2012); sin embargo, los factores predominantes en tales clasificaciones son el medio de información (textos, imágenes, videos o voz), la entrada (un documento o múltiples documentos), la salida (extractivo o abstractivo), el propósito (genérico, personalizado, enfocado a una consulta, indicativo o informativo) y la cantidad de lenguas (monolingüe o multilingüe).

En este trabajo, se consideran especialmente dos factores: la cantidad de documentos en la entrada y el tipo de resumen que se genera a la salida: extractivo o abstractivo.

Los resúmenes extractivos se generan a partir de la selección de oraciones consideradas sobresalientes en el texto de origen. Las oraciones se extraen literalmente, se unen libremente, y se presentan como el resumen del texto. En este enfoque, que se ha estudiado ampliamente (Nenkova & McKeown, 2011; Lloret & Palomar, 2012),

se hace un análisis superficial de los textos, a nivel de palabras; por lo que, en general, los resúmenes no tienen coherencia y solo 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 fuente; esto es, se reformulan las frases por medio de procesos de fusión, combinación o supresión de términos (Spärck Jones, 1999; Hovy & Chin-Yew, 1999; Spärck Jones, 2007). 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. Para generar esta clase de resúmenes se requiere de una representación que emule la comprensión humana del texto. Se ha demostrado que para propósitos indicativos (líneas o párrafos relevantes del texto sin necesidad de estar coherentemente relacionados) los resúmenes extractivos han sido adecuados, pero para otros propósitos como informativos (descripción de los aspectos relevantes y la relación lógica del tema tratado) o que el resumen se adapte a los intereses del usuario es necesario generar resúmenes abstractivos (Hovy, 2005; Spärck Jones, 2007; Nenkova & McKeown, 2011).

En este trabajo, se presenta un modelo para la generación de resúmenes abstractivos de un solo documento, basado en grafos conceptuales (Sowa, 1984) como la representación textual subyacente; además, se describe el método para la síntesis de los grafos conceptuales. Nuestra investigación se centra en reducir las estructuras de los grafos (conceptos y relaciones) aplicando varias operaciones: generalización, unión, ponderación y poda. Las estructuras resultantes representan al resumen a nivel conceptual.

El artículo está organizado de la siguiente forma. En la sección 1, se describe el trabajo relacionado y el formalismo de los grafos conceptuales. En la sección 2, se presenta la metodología y se detalla el método para la síntesis de grafos. En la sección 3, se discuten los resultados obtenidos. Por último, se presentan las conclusiones.

## 1. Marco teórico

### *1.1. Trabajo relacionado*

Como se mencionó, los resúmenes extractivos se crean a partir de la selección de oraciones o párrafos más representativos del texto; generalmente, las oraciones seleccionadas se presentan en el mismo orden en que aparecen en los documentos originales.En los métodos que implementan este enfoque, se decide automáticamente si una oración se incluirá en el resumen. Por lo que, el resumen resultante es usualmente incoherente debido a las técnicas usadas.

Los métodos extractivos emplean frecuentemente un modelo de ponderación lineal para determinar la importancia de las unidades textuales (palabras, oraciones o

párrafos) de acuerdo a diversas características en el documento tales como la posición, la frecuencia de ocurrencia, la aparición de ciertas palabras de entrada ('en conclusión', 'en resumen', etc.), entre otras (Baxendale, 1958; Luhn, 1958). Este enfoque ha sido estudiado ampliamente y sus limitaciones son conocidas (Spärck Jones, 1999; Hovy & Chin-Yew, 1999; Spärck Jones, 2007; Nenkova & McKeown, 2011).

Por otra parte, en los últimos años, se ha explorado la generación de resúmenes abstractivos. Entre ellos destacan los métodos que se enfocan en la compresión de frases donde se busca reducir la frase a su forma esencial por medio del uso de modelos probabilísticos (Knight & Marcu, 2000; Cohn & Lapata, 2009), de la estructura retórica del documento (Molina, Torres-Moreno, SanJuan, da Cunha & Sierra, 2013); o la fusión de frases (Barzilay & McKeown, 2005), donde se hace uso de técnicas de reescritura con base en el análisis de las estructuras sintácticas del texto.

También, se han implementado esquemas de abstracción con sus reglas de extracción definidas para un dominio específico (Genest & Lapalme, 2012). La representación se orienta a responder consultas que cubran una necesidad de información específica. Se usan elementos de información sintáctica, los cuales se definen como tripletas sujeto–verbo–objeto, por ejemplo, persona–mata–mujer.

Otro método es el *Semantic Graph*, el cual se basa en la información semántica del documento (Leskovec, Grobelnik & Milic-Frayling, 2004; Tsatsaronis, Varlamis & Nørvåg, 2010). En este método, se implementa un análisis sintáctico profundo para extraer las relaciones sintácticas del texto, y se extraen tripletas de la forma lógica sujeto-predicado-objeto para cada oración. Cada tripleta se caracteriza por un conjunto de atributos basados en el grafo, estadísticos y lingüísticos. Se usa un clasificador supervisado, a saber, *Support Vector Machine* (Vapnik, 1998) para identificar las tripletas importantes que conformarán al resumen. Otro método basado en la información semántica del texto hace uso de grafos de conceptos (Plaza, 2010; Plaza, Díaz & Gervás, 2011). En este enfoque, se usa una representación conceptual del texto basándose en una red semántica y un tesauro. Se agrupan los conceptos similares del documento formando un mapa de grupos conceptuales, se determinan los grupos más relevantes y se asocian las oraciones a los grupos, seleccionando finalmente las oraciones más relevantes según a los grupos que pertenezcan.

En los enfoques que consideran al texto como un grafo conectado, por ejemplo, TextRank (Mihalcea & Tarau, 2004) y LexRank (Erkan & Radev, 2004), los nodos representan oraciones, palabras u otro tipo de unidad, y las aristas se crean de acuerdo al traslape del contenido entre las unidades; otros autores usan las relaciones sintácticas como aristas (Litvak & Last, 2008). En este contexto, para identificar la importancia de un nodo se usan algoritmos de ponderación como HITS (Kleinberg, 1999) o PageRank (Page & Brin, 1998), donde los nodos con mejor ponderación se consideran los más relevantes y formarán parte del resumen.

Los enfoques mencionados usan principalmente técnicas de reescritura basándose en un análisis sintáctico parcial o análisis semántico parcial como el uso de algunas relaciones sintácticas o patrones semánticos. Sin embargo, no se ha usado una representación semántica de los textos de gránulo fino como la que se puede realizar con una representación más expresiva como los grafos conceptuales (Sowa, 1984), donde se aprovechan los roles temáticos como agente, paciente, tema, etc. (Jackendoff, 1972; Fillmore & Atkins, 1992) que forman parte de la estructura semántica del texto representado. La representación semántica detallada del texto es la principal característica que diferencia esta investigación de los métodos mencionados. En este contexto, el problema de la generación de resúmenes se reduce a simplificar los grafos conceptuales que representan al texto y mantener la información importante y esencial, así como la coherencia en tales estructuras.

### *1.2. Grafos conceptuales*

Los grafos conceptuales (GCs) son estructuras para la representación del conocimiento basados en lógica (Sowa, 1984). Con este formalismo, los textos se representan de forma natural, simple y con una representación semántica detallada. En estas estructuras, existen dos tipos de nodos: las relaciones conceptuales (representadas por óvalos) y los conceptos (representados por rectángulos). Un concepto se conecta a otro concepto a través de una relación conceptual, y una relación conceptual debe conectarse estrictamente con algún concepto.

En la Figura 1, tomado de Sowa (1984), se representa la oración "*Joe buys a necktie from Hal for $10*" (Joe compra una corbata a Hal en $10). En este grafo, se detalla la semántica de la oración: quién compró (AGNT), quién vendió (SRCE), qué se vendió (OBJ) y por medio de qué se vendió (INST).

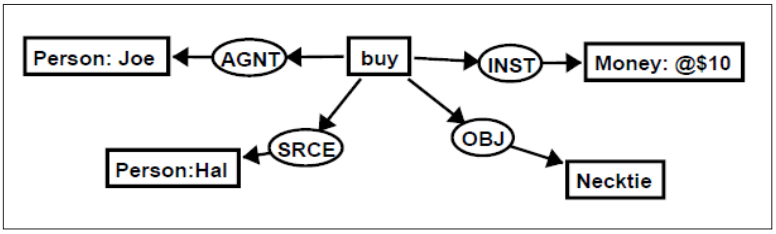


**Figura 1**. Grafo conceptual.

Otro elemento clave de los GCs es el 'concepto tipo'. Los conceptos tipo representan clases de entidades como *Person* (Persona), *Money* (Dinero), etc. Esto se conoce como una relación jerárquica de conceptos tipo que representa a una jerarquía 'ES-UN'. La jerarquía se usa para vincular conceptos para propósitos de inferencia (Sowa, 1984; Sowa, 1999; Chein & Mugnier, 2009). Por ejemplo, en la Figura 1, *Person: Joe* denota el concepto tipo *Person* y su referente *Joe* que es una instancia de *Person*.

Adicionalmente, el formalismo de GCs permite operaciones basadas en grafos para razonamiento. Un conjunto de operaciones como restricción, simplificación, unión, indexamiento (Sowa, 1999; Chein & Mugnier, 2009), y empatamiento de grafos (Montes-y-Gómez, Gelbukh, López-López & Baeza-Yates, 2001).

### *1.2.1. Grafos conceptuales ponderados*

Los grafos conceptuales ponderados (Miranda-Jiménez, Gelbukh & Sidorov, 2013) son GCs, en lo cuales se asocian pesos a las aristas que conectan a los nodos concepto y los nodos relación creando un flujo denominado 'flujo semántico' (véase la Figura 2). Un flujo semántico es básicamente el peso que acumulan los nodos y que se transmite hacia otros nodos aumentando o disminuyendo su valor al pasar por alguna relación conceptual.

En este contexto, las relaciones conceptuales representan principalmente la semántica del texto, relaciones tales como agente, objeto, lugar, atributo, tema, etc. (Sowa, 1984). De ahí que un peso con valor alto en la arista indica interés por el flujo de la relación en cuestión. Adicionalmente, en este tipo de grafos, los nodos concepto aceptan un peso que define la preferencia del nodo, es decir, el grado de interés de ciertos tópicos definidos por los conceptos.

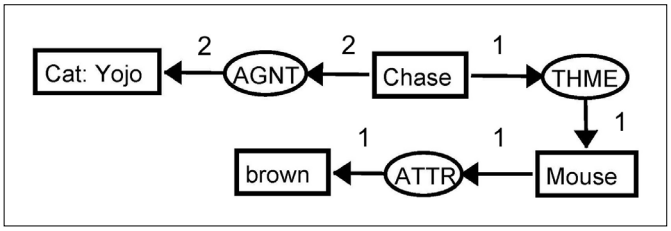


**Figura 2.** Grafo conceptual ponderado.

En la Figura 2, se muestra un grafo conceptual ponderado para la oración "*The cat Yojo is chasing a brown mouse*" (El gato Yojo está persiguiendo a un ratón café) (Sowa, 1984). En este grafo, se asigna mayor importancia a la relación agentiva (AGNT) asignando un peso mayor a su arista entrante y a su arista saliente.

Los flujos semánticos se usan durante el proceso de ponderación para elegir a los nodos más importantes (véase la sección 2.2.2.4. Ponderación).

## 2. Metodología

Nuestro modelo para la generación de resúmenes abstractivos se basa en la representación semántica completa del texto por medio de GCs. De modo que el problema de la generación del resumen se simplifica en seleccionar los nodos más importantes de los grafos y reducir las estructuras gráficas manteniendo su coherencia estructural.

En la Figura 3, se muestra el esquema general. Primeramente, se hace un preprocesamiento al texto para generar semiautomáticamente los GCs a partir de un conjunto de documentos seleccionados. Durante el proceso de generación, se agrega información semántica y sintáctica de fuentes externas.

En la etapa de síntesis, los grafos se reducen de acuerdo a un conjunto de operaciones (generalización, unión, ponderación y poda) que se les aplica. A partir de los grafos resultantes, se genera el texto resumido. La generación del texto queda fuera del alcance de esta investigación, ya que en este estudio nos interesa obtener las estructuras conceptuales resumidas.

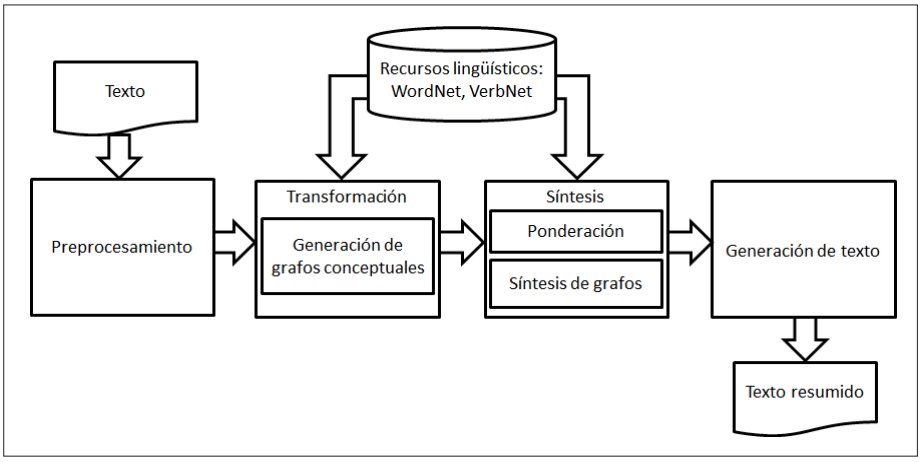


**Figura 3.** Generación de resúmenes basado en grafos conceptuales (GCs).

## *2.1. Herramientas*

Nuestro enfoque se basa en el formalismo de GCs, el cual requiere de información semántica adicional que se agrega a los grafos durante su creación: una jerarquía de conceptos para propósitos de generalización y 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 3 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).

**WordNet** es una base de datos léxica para el inglés agrupada en 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. De esta base de datos léxica se obtuvo la jerarquía de 'conceptos tipo' por medio de las relaciones semánticas de hiperonimia / hiponimia. Por ejemplo, *atmospheric phenomenon* / *storm* (fenómeno atmosférico / tormenta), *residence* / *home* (residencia / casa).

**VerbNet** es un diccionario computacional que 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.

Cada clase definida en VerbNet contiene una lista de miembros, una lista de roles temáticos y una lista de *frames* (patrones donde se indica cómo los roles semánticos pueden realizarse en una oración). Los verbos que pertenecen a la misma clase comparten los mismos marcos sintácticos. 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, en nuestro modelo, rigen la coherencia estructural de los grafos.

**Colección de textos.** Nuestra colección de textos para la evaluación del enfoque consta de 30 documentos de noticias referentes a desastres naturales que se tomaron de la colección de datos de la competencia *Document Understanding Conference 2003* (DUC, 2003). Se eligió esta versión de documentos, ya que en este año se consideró la tarea de la generación de resúmenes muy breves, a nivel de encabezado; además, se cuenta con documentos (resúmenes) cortos creados para esta competencia. De ahí que nosotros consideramos, al resumen como el documento de origen y a los encabezados como los resúmenes de este documento. Los textos seleccionados son breves, de entre 50 y 100 palabras.

## *2.2. Procedimientos metodológicos*

Los procedimientos metodológicos de esta investigación se realizan principalmente en dos etapas: la generación de GCs y la síntesis de GCs, es decir, la aplicación de las operaciones sobre los GCs para obtener el resumen a nivel conceptual.

### *2.2.1. Generación de grafos conceptuales*

La construcción de los GCs a partir del texto no es directa. Se han propuesto métodos para la generación automática de GCs (Hensman, 2005; Ordoñez-Salinas & Gelbukh , 2010); sin embargo, las herramientas no están disponibles y los resultados que se reportan no son apropiados para nuestro objetivo. Por lo cual, generamos semiautomáticamente una colección de GCs basados en los documentos de noticias que usa la competencia DUC 2003.

**Preprocesamiento.** El conjunto de GCs se creó a partir de los textos de la colección DUC. Para identificar las relaciones entre las palabras de los textos, los documentos se preprocesaron con el analizador sintáctico (*parser*) de Stanford (de Marneffe, MacCartney & Manning, 2006), es decir, se obtuvieron los árboles de dependencia (Mel'cuk, 1988) y se implementó un conjunto de reglas considerando el tipo de relación obtenida por el *parser* para establecer la relación conceptual.

**Construcción de GCs.** Después de realizar el preprocesamiento de los textos y a partir de la información gramatical de los árboles de dependencias, se generaron los GCs 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 y Manning (2008). Las relaciones que se asignan incorrectamente o que no se identifican por las reglas de transformación se corrigen o añaden manualmente. Además, se asoció manualmente el patrón verbal a todos los conceptos verbales y se asignó el 'concepto tipo' para cada concepto del grafo según la jerarquía de WordNet.

En la Figura 4, 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 solo la palabra normalizada (verbo en infinitivo) se muestra en el grafo. El conjunto de etiquetas para las categorías gramaticales usadas por el *parser* de Stanford es el definido en el marco del proyecto *Penn TreeBank* (Santorini, 1990).

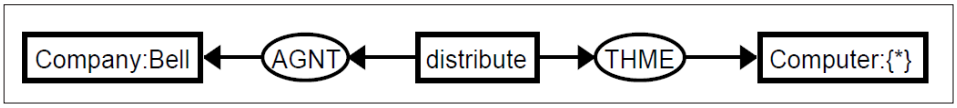


**Figura 4.** Construcción de los nodos a partir de relación *nsubj* y *obj*.

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, 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) para detalles de la sintaxis usada en los GCs.

Por último, también se añadió manualmente 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.

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, 1984).

En el grafo de la Figura 5, se representa el siguiente texto.

> *"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 cerró en Hong Kong con fuertes vientos. La tormenta mató al menos a 156 personas en Filipinas y dejó a cientos de miles sin hogar."

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.

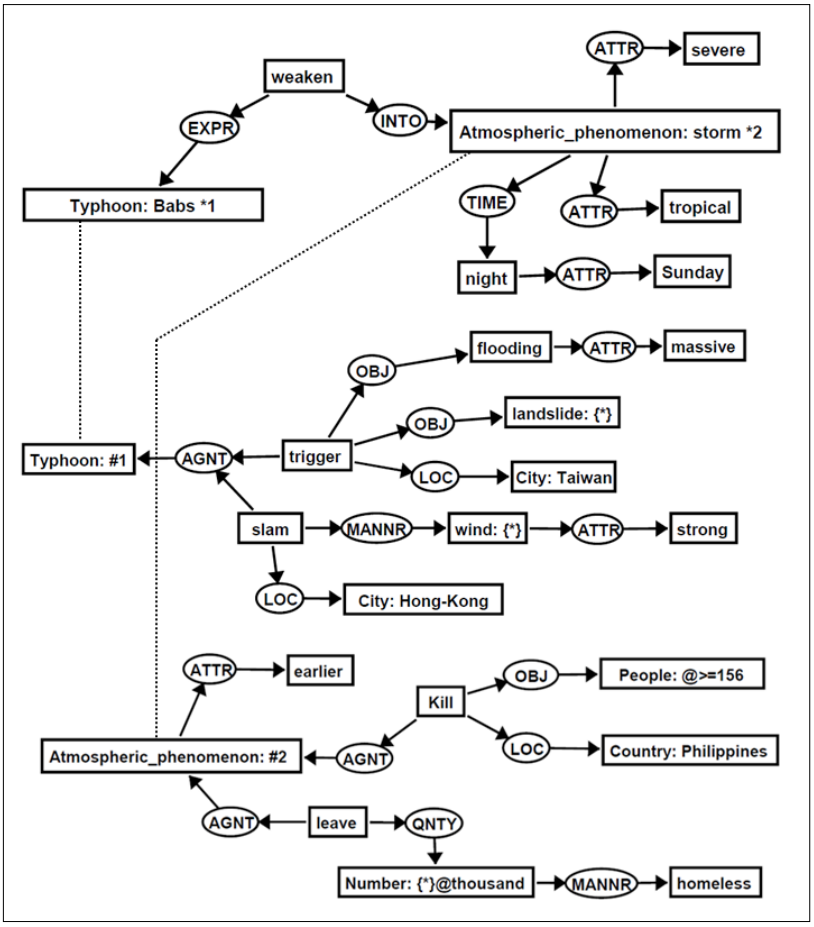


**Figura 5**. Ejemplo de noticia como grafo conceptual.

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 a 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).

### *2.2.2. Síntesis de grafos conceptuales*

El método de síntesis se basa en un conjunto de operaciones sobre los GCs: generalización, unión o asociación, ponderación y poda. Esta investigación extiende el trabajo presentado en Miranda-Jiménez et al. (2013) precisando las operaciones que se aplican a los GCs para resumirlos. También se usa una operación de comparación entre los grafos para fines de asociación y generalización (Montes-y-Gómez et al., 2001).

Para identificar las estructuras más importantes en los GCs, se realizan los siguientes pasos.

1. Identificar todas las generalizaciones en los grafos por medio de la operación de generalización.
2. Identificar todas las asociaciones entre los grafos por medio de la operación de unión.
3. Establecer las medidas AUTH y HUB asociadas a cada nodo al valor de 1.
4. Ponderar los nodos calculando los valores AUTH y HUB para cada nodo.
5. Normalizar los valores AUTH y HUB por medio de la norma euclidiana.
6. Repetir los pasos 4 y 5 hasta alcanzar el número máximo predefinido de iteraciones.
7. Ordenar descendentemente los nodos de acuerdo al valor de la métrica AUTH.
8. Expandir los conceptos conectados para cada relación conceptual seleccionada.
9. Expandir los nodos asociados a conceptos verbales de acuerdo con su patrón verbal.
10. Seleccionar los conceptos sobresalientes de acuerdo al porcentaje de compresión establecido para la poda del grafo.

En los pasos 1 y 2, se establecen las relaciones entre todos los grafos para mejorar la ponderación de los nodos, identificando la relación que tiene cada nodo respecto a otros grafos. En los pasos 3 a 6, se calculan los valores para el algoritmo HITS basado en las ecuaciones (1) y (2) de ponderación de la sección 2.2.2.4. En el paso 8, se aplican las reglas para expandir los conceptos que conectan a una relación conceptual, si es que fue seleccionada como nodo relevante en el paso 7. Por ejemplo, la relación OBJ(*trigger*, *flooding*) (véase la Tabla 1, sección 3) se expande en dos conceptos *trigger* (provocar) y *flooding* (inundación).

En el paso 9, se aplican los patrones verbales asociados a los conceptos verbales para mantener la coherencia de la estructura. Por ejemplo, en la Figura 2, el patrón verbal para el concepto verbal *chase* (perseguir) es 'NP V NP' (frase sustantiva / verbo / frase sustantiva), y el verbo es transitivo. El 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.

Por último, en el paso 10, se aplica la operación de poda por medio de un porcentaje de compresión establecido previamente por el usuario. Aquí, se seleccionan los nodos sin duplicados de acuerdo a la tabla de ponderación ordenada descendentemente según la métrica AUTH. Los nodos seleccionados, los cuales forman parte de los grafos, se consideran como el resumen final, a nivel conceptual (véase la Tabla 2, sección 3).

En los siguientes puntos se detallan las operaciones propuestas para la síntesis de GCs.

### *2.2.2.1. Comparación entre GCs*

Esta operación consiste en la comparación de dos GCs (Montes-y-Gómez et al., 2001). Primero, los dos GCs se traslapan y se identifican sus elementos comunes (conceptos y relaciones). La medida de similitud se calcula como un tamaño relativo de sus elementos comunes, se obtiene un valor entre 0 y 1 (0 significa ninguna similitud; 1 significa la máxima similitud). Esta operación requiere de una jerarquía de conceptos para determinar la similitud. Por ejemplo, en la Figura 6, para determinar la similitud entre *cocodrile* y *bird*, se usa la jerarquía de conceptos (a) de la Figura 7. En el ejemplo, *Animal*(3,3): 0.66 significa que el concepto '*Animal*' es el concepto común mínimo para *cocodrile* y *bird* con una similitud de 0.66, de esta manera, se puede establecer una asociación.

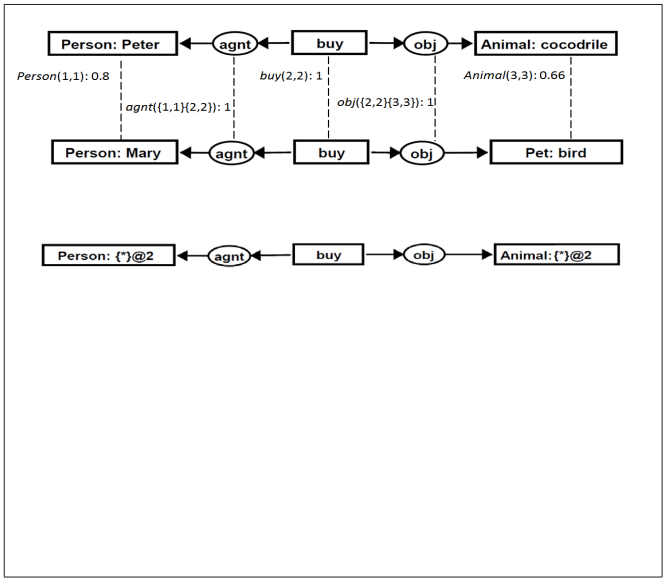


**Figura 6.** Generalización.

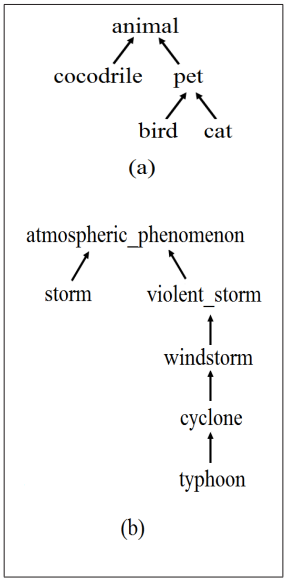


**Figura 7.** Jerarquía de conceptos.

### *2.2.2.2. Generalización*

La operación de generalización combina dos GCs de acuerdo a sus elementos comunes. Por ejemplo, los siguientes grafos se pueden leer como $G_1$: *Peter buys a crocodile* (Peter compra un cocodrilo) y $G_2$: *Mary buys a bird* (María compra un pájaro).

Se realiza una comparación entre los GCs, posteriormente, se determinan los conceptos comunes mínimos para unirlos. De acuerdo con la jerarquía (a) de la Figura

7 para *cocodrile* y *bird*, *Animal* es el concepto común mínimo entre ambos conceptos; y *Person* es el concepto mínimo común para *Peter* y *Mary*. Por lo que $G_3$ es el grafo resultante después de combinar los dos grafos. $G_3$ puede leerse como "*Two persons buy two animals"* (Dos personas compran dos animales).

### *2.2.2.3. Unión o asociación*

Esta operación une dos conceptos relacionados de dos GCs. Nuestro supuesto es que un texto fuente es coherente y cohesivo, donde las oraciones probablemente refieren a conceptos que fueron previamente mencionados u otros conceptos relacionados (Halliday & Hasan, 1976).

Esta operación apoya y mejora los resultados del proceso de ponderación. En la Figura 8, los GCs pueden leerse como $G_4$: "*Typhoon Babs weakened into severe storm*" y $G_5$: "*Storm killed at least 156 people in Philippines*".

En estos grafos hay dos traslapes de conceptos relacionados: la primera asociación se identifica por **(1,1)**, *Typhoon: Babs* (concepto 1, $G_4$) y *atmospheric_phenomenon: storm* (concepto 1, $G_5$); y la otra asociación identificada por **(3,1)**, *atmospheric_phenomenon*: *storm* (concepto 3, $G_4$) y *atmospheric_phenomenon*: *storm* (concepto 1, $G_5$), esta última tiene la similitud máxima usando la jerarquía (b) de la Figura 8. Ambas asociaciones son válidas, pero se usa la asociación con la medida de similitud máxima para establecer la asociación entre los nodos. Por lo tanto, consideramos como el mismo nodo ambos conceptos: concepto 3 de $G_4$ y concepto 1 de $G_5$.

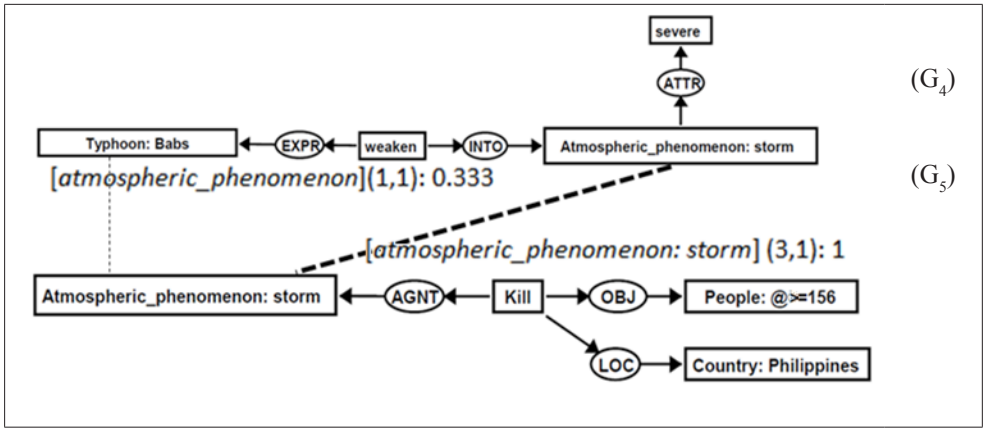


**Figura 8**. Unión o asociación.

### *2.2.2.4. Ponderación*

La operación de ponderación filtra los nodos más importantes del grafo. Para determinar la importancia de los nodos, se utiliza el algoritmo HITS (Kleinberg,

1999). HITS es un algoritmo iterativo que toma en cuenta el grado de entrada y el grado de salida del nodo para determinar su importancia.

El método HITS proporciona dos métricas AUTH y HUB (autoridad y concentración). Usamos la métrica AUTH para determinar la importancia del nodo, ya que nos indica qué nodo es bueno como fuente de información; no obstante, se calculan ambas métricas ya que son dependientes, ambos valores se calculan para cada nodo del grafo. 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.

Utilizamos una versión modificada del método HITS a la propuesta por Mihalcea y Tarau (2004). Para el cálculo de las métricas AUTH y HUB se usan las ecuaciones (1) y (2), las cuales consideran los flujos semánticos entre dos nodos ($W_{ki}$) y la preferencia de los tópicos (PREF). 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 nodo $V_k$.

$$AUTH(V_i) = \sum_{V_k \in I(V_i)} W_{ki} \times HUB(V_k) \times PREF(V_k) \qquad (1)$$

$$HUB(V_i) = \sum_{V_k \in O(V_i)} W_{ik} \times AUTH(V_k) \times PREF(V_k) \qquad (2)$$

Las ecuaciones (1) y (2) se calculan iterativamente para cada nodo hasta que converja el algoritmo (no haya cambios en las métricas AUTH y HUB) o hasta un determinado número de iteraciones previamente definidas. Mihalcea y Tarau (2004) usan de 20 a 30 iteraciones en el algoritmo, otros autores usan una sola iteración (Litvak & Last, 2008). Nosotros hemos identificado que 15 iteraciones son suficientes para la colección de grafos, más iteraciones no mejoran la selección de los nodos.

#### *2.2.2.5. Poda*

La operación de poda se aplica para reducir los grafos. Esta operación toma en cuenta los resultados de la ponderación, los patrones verbales para remover los nodos irrelevantes y la tasa de compresión o umbral para establecer cuántos nodos deben incluirse en el resumen resultante. Esta operación selecciona los nodos según su valor de la métrica AUTH y que se encuentren dentro del porcentaje de compresión.

De acuerdo con Hovy (2005) un resumen es útil si se encuentra entre el 15% y el 35% de longitud con respecto al documento original. Nosotros usamos el 20% de compresión, ya que al analizar los resúmenes con los cuales se compara nuestro método, la mayoría de los resúmenes hechos manualmente estaban dentro de este porcentaje.

## 3. Resultados de la investigación

Los experimentos se realizaron con la colección de noticias de la competencia DUC 2003. Se seleccionaron documentos con longitud de entre 50 y 100 palabras. Se definieron tres grupos de documentos según su longitud: 3 oraciones (Grupo I), 4 oraciones (Grupo II) y más de 4 oraciones (Grupo III). Cada grupo consta de 10 documentos representados como GCs. Además, se estableció un porcentaje de compresión del 20%, y se fijó el flujo semántico con el valor de 2 para las relaciones agentivas tanto para el flujo de entrada como para el flujo de salida. Los demás flujos son neutrales, el valor se estableció en 1 (debido a las ecuaciones), para no cancelar el flujo que transmiten los nodos. Estos valores se establecieron basados en el supuesto de que los agentes (actores) son más importantes en el contexto de un documento referente a una noticia, y se determinaron heurísticamente.

Para comparar el método, se definió un algoritmo básico (una 'línea base'), el cual consiste en seleccionar los primeros conceptos de la noticia hasta llegar al porcentaje de compresión establecido (excepto palabras auxiliares: artículos, preposiciones, etc.). También, usamos métricas estándares como precisión y *recall* (términos comúnmente usados en recuperación de información). *Recall* es la fracción de conceptos elegidos por el humano que fueron correctamente identificados por el método, ecuación (3). La precisión está definida como la fracción de conceptos elegidos por el método que fueron correctos, ecuación (4). La *Medida-F* es la media harmónica de precisión y r*ecall*, ecuación (5).

$$Recall = \frac{\text{Traslape de num. conceptos elegidos por el método y el humano}}{\text{Num. conceptos elegidos por el humano}} \quad (3)$$

$$\text{Precisión} = \frac{\text{Traslape de num. conceptos elegidos por el método y el humano}}{\text{Num. conceptos elegidos por el método}} \quad (4)$$

$$\text{Medida–F} = 2 \times \frac{\text{Precisión} \times Recall}{\text{Precisión} + Recall} \quad (5)$$

En la Tabla 1, se muestra un ejemplo de los nodos seleccionados por el método de ponderación incluyendo las relaciones conceptuales para el grafo de la Figura 5, así como la expansión de las relaciones conceptuales tales como las relaciones objeto (OBJ). La tabla está ordenada descendentemente por la métrica AUTH. 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 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.

**Tabla 1.** Relaciones conceptuales y conceptos seleccionados por el método de ponderación con expansión de relaciones conceptuales.

| Nodo | Expansión de relación | AUTH | HUB |
|---|---|---|---|
| Cyclone:Typhoon-Babs | - | 0.729 | 0.3e–16 |
| Atmospheric_phenomenon:storm | - | 0.680 | 0.70e–03 |
| AGNT(trigger-Cyclone:Typhoon-Babs) | trigger / Typhoon-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-Cyclone: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 |

En la Tabla 2, se muestran los conceptos finales seleccionados que formarán parte del resumen. Las relaciones que conectan a estos conceptos también se consideran como parte del resumen, de hecho, es a partir de ellas que se expanden los nodos concepto por lo que ya estaban consideradas (véase la Tabla 1).

Finalmente, los conceptos seleccionados en la Tabla 2 representan al resumen, de acuerdo con el grafo de la Figura 5, el cual puede leerse como el siguiente texto:

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

**Tabla 2.** Conceptos finales seleccionados por el método de ponderación.

| Nodo | AUTH | HUB |
|---|---|---|
| Cyclone: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 |

En la Tabla 3, se agrupan los resultados obtenidos según los resúmenes generados a nivel conceptual. Nuestro método supera en promedio 11% a la línea base. Para el Grupo I que consta de 3 oraciones, obtenemos 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%.

**Tabla 3.** 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** |

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

No obstante, según los datos mostrados en los grupos II y III, se puede inferir que para textos a nivel de párrafo, 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 al representarse como GCs.

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 existen estudios que han demostrado que, en general, las primeras y últimas oraciones en los párrafos son buenos indicadores para identificar la información relevante, y, en el caso particular de documentos de noticias, las primeras líneas contienen la información más relevante (Baxendale, 1958; Luhn, 1958; Hovy & Chin-Yew, 1999).

Aunado a lo anterior, la competencia DUC reportó los resultados para la tarea de generación de resúmenes de un solo documento de noticias para los años 2001 y 2002, donde ningún sistema participante superó a la 'línea base', la cual consistía en recuperar el inicio de los párrafos de los artículos de noticia, similar a la usada en esta investigación. 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 automática de resúmenes monodocumento ya no se continuó realizando en los años siguientes de la competencia DUC, sigue siendo un problema abierto (Nenkova, 2005; Nenkova & McKeown, 2011).

## CONCLUSIONES

En esta investigación propusimos un nuevo modelo para la generación de resúmenes abstractivos de un solo documento basado en GCs como la representación intermedia del texto. El modelo proporciona una combinación del contenido del texto con los roles semánticos dentro de un contexto de ponderación basado en la conectividad de los conceptos por medio de su semántica.

El modelo presentando, también, incorpora los flujos semánticos vinculados a los GCs que 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. Esta característica detallada de la semántica del texto no se había usado para la generación de resúmenes.

También, cabe mencionar las limitaciones del método presentado ya que se requiere de recursos lingüísticos externos (WordNet y VerbNet), los cuales lo hacen dependiente del idioma; adicionalmente, la obtención de los GCs a partir del texto es otra limitante ya que por el momento no hay herramientas que los generen automáticamente.

Finalmente, la evaluación del método se realizó con documentos de noticias muy breves y se superó a la 'línea base' con un promedio del 11%, similar a la usada en los años de la competencia DUC 2001 y DUC 2002, donde ningún sistema superó a la 'línea base'. Un reto futuro es aplicar nuestro modelo a textos más largos y a la generación de resúmenes multidocumento, siendo la principal limitante la construcción automática de los GCs.

## REFERENCIAS BIBLIOGRÁFICAS


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.

Chein, M. & Mugnier, M. L. (2009). *Graph-based knowledge representation: Computational foundations of conceptual graphs*. Londres: Springer-Verlag.

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

de Marneffe, M. C. & Manning, C. D. (2008). *Stanford parser manual* [en línea]. Disponible en: 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*. Ponencia presentada en el 5th International Conference on Language Resources and Evaluation, Genova, Italia.

DUC (2003). *Document Understanding Conference* [en línea]. Disponible en: http://duc.nist.gov/pubs.html#2003

Erkan, G. & Radev, D. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*(1), 457-479.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Fillmore, C. & Atkins, B. T. (1992). Towards a frame-based lexicon: The case of RISK. En A. Lehrer & E. Kittay (Eds.), *Frames and Fields* (pp. 75-102). Hillsdale, NJ: Erlbaum.

Genest, P. E. & Lapalme, G. (2012). *Fully abstractive approach to guided summarization*. Ponencia presentada en el 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Corea.

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

Hensman, S. (2005). Constructing conceptual graphs using linguistic resources. En M. Husak & Mastorakis (Eds.), *The 4th WSEAS International Conference on Telecommunications and Informatics* (pp. 1-6). Wisconsin, USA:WEAS Press.

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

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

Jackendoff, R. (1972). *Semantic interpretation in generative grammar*. Cambridge, MA: MIT Press.

Kipper, K., Trang Dang, H. & Palmer, M. (2000). Class-based construction of a verb lexicon. En R. Engelmore & H. Hirsh (Eds.), *The 17th National Conference on Artificial Intelligence* (pp. 691-696). Menlo Park, CA: AAAI Press.

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. En R. Engelmore & H. Hirsh (Eds.), *The 17th National Conference on Artificial Intelligence* (pp. 703-710). Menlo Park, CA: AAAI Press.

Leskovec, J., Grobelnik, M. & Milic-Frayling, N. (2004). *Learning semantic graph mapping for document summarization*. Ponencia presentada en ECML/PKDD-2004 Workshop on Knowledge Discovery and Ontologies, Pisa, Italia.

Litvak, M. & Last, M. (2008). Graph-based keyword extraction for single-document summarization. En S. Bandyopadhyay, T. Poibeau, H. Saggion & R. Yangarber (Eds.), *Workshop on Multi-source Multilingual Information Extraction and Summarization* (pp. 17-24). Manchester, UK: Coling.

Lloret, E. & Palomar, M. (2012). Text summarisation in progress: A literature review. *Artificial Intelligence Review*, *37*(1), 1-41.

Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, *2*(2), 159-165.

Mel'cuk, I. (1988). *Dependency syntax: Theory and practice*. Albany, NY: State University Press of New York.

Mihalcea, R. & Tarau, P. (2004). *TextRank: Bringing order into texts*. Ponencia presentada en Conference on Empirical Methods in Natural Language Processing, Barcelona, España.

Miranda-Jiménez, S., Gelbukh, A. & Sidorov, G. (2013). Summarizing conceptual graphs for automatic summarization task. En H. Pfeiffer, D. Ignatov, J. Poelmans & N. Gadiragu (Eds.), *The 20th International Conference on Conceptual Structures* (pp. 245-253). Berlin: Springer-Verlag.

Molina, A., Torres-Moreno, J. M., SanJuan, E., da Cunha, I. & Sierra, G. (2013). Discursive sentence compression. En A. Gelbukh (Ed.), *The 14th International Conference CICLing 2013* (pp. 394-407). Berlin: Springer-Verlag.

Montes-y-Gómez, M., Gelbukh, A., López-López, A. & Baeza-Yates, R. A. (2001). Flexible comparison of conceptual graphs. En H. Mayr, J. Lazansky, G. Quirchmayr & P. Vogel (Eds.), *The 12th International Conference and Workshop on Database and Expert Systems Applications* (pp. 102-111). Berlin: Springer-Verlag.

Nenkova, A. (2005). Automatic text summarization of newswire: Lessons learned from the document understanding conference. En M. Veloso & S. Kambhampati (Eds.), *The 20th National Conference on Artificial Intelligence* (pp. 1436-1441). Menlo Park, CA: AAAI Press..

Nenkova, A. & McKeown, K. (2011). Automatic summarization. *Foundations and Trends in Information Retrieval*, *5*(2-3), 103-233.

Ordoñez-Salinas, S. & Gelbukh, A. (2010). Generación de grafos conceptuales. En M. González Mendoza & M. Herrera Alcántara (Eds.), *Avances en sistemas inteligentes en México* (pp. 139-150). Ciudad de México: SMIA.

Page, L. & Brin, S. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1-7), 107-117.

Plaza, L. (2010). *The use of semantic graphs in automatic summarization: Comparative case studies in Biomedicine, Journalism and Tourism*. Tesis doctoral, Universidad Complutense de Madrid, Madrid, España.

Plaza, L., Díaz, A. & Gervás, P. (2011). A semantic graph-based approach to biomedical summarisation. *Artificial Intelligence in Medicine*, *53*(1), 1-14.

Santorini, B. (1990). *Part-of-speech tagging guidelines for the penn treebank project*. Reporte técnico MS-CIS-90-47, Universidad de Pennsylvania, Pennsylvania, USA.

Sowa, J. (1984). *Conceptual structures: Information processing in mind and machine*. Reading, MA: Addison-Wesley.

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

Spärck Jones, K. (1999). Automatic summarising: Factors and directions. En I. Mani & M. T. Maybury (Eds.), *Advances in Automatic Text Summarization* (pp. 1-12). Cambridge MA: MIT Press.

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

Tsatsaronis, G., Varlamis, I. & Nørvåg, K. (2010). *SemanticRank: Ranking keywords and sentences using semantic graphs*. Ponencia presentada en el 23rd International Conference on Computational Linguistics, Beijing, China.

Vapnik, V. N. (1998). *Statistical learning theory*. Nueva York, USA: John Wiley & Sons.

## * AGRADECIMIENTOS

Este trabajo fue realizado con el apoyo parcial del gobierno de México (CONACYT, SNI), del IPN, México (proyectos SIP 20120418, 20131441, 20131702; COFAA, PIFI), del proyecto 122030 CONACYT-DST India "Answer Validation through Textual Entailment", y del proyecto 269180 de la comunidad europea: FP7-PEOPLE-2010-IRSES "*Web Information Quality - Evaluation Initiative* (WIQ-EI)".