**PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ**

**ESCUELA DE POSGRADO**
**MAESTRÍA EN INFORMÁTICA**


# Generación automática de resúmenes abstractivos mono documento utilizando análisis semántico y del discurso

**TESIS PARA OPTAR EL GRADO ACADÉMICO DE MAGÍSTER EN INFORMÁTICA CON MENCIÓN EN CIENCIAS DE LA COMPUTACIÓN**

**AUTOR**

**Gregory Cesar Valderrama Vilca**

**ASESOR**

**MSc. Marco Antonio Sobrevilla Cabezudo**

**Septiembre de 2017**

# Abreviaturas

**PLN** Procesamiento del Lenguaje Natural

**NER** *Named Entity Recognition*

**RST** *Rhetorical Structure Theory*

**EDU** *Elementary Discourse Unit*

**DUC** *Document Understanding Conference*

**TAC** *Text Analysis Conference*

**AMR** *Abstract Meaning Representation*

**SRL** *Semantic Role Labeling*

**PAS** *Predicate Argument Structures*

**HAC** *Agglomerative Hierarchical Clustering*

**PSO** *Particle Swarm Optimization*

**ROUGE** *Recall-Oriented Understudy of Gisting Evaluation*

**SUMMAC** *TIPSTER Text Summarization Evaluation*

**BBN** *Bolt, Beranek and Newman*

**ACE** *Attempto Controlled English*

**AMRICA** *AMR Inspector for Cross-language Alignments*

# Agradecimientos

En primer lugar deseo agradecer a Dios por haberme guiado a lo largo de estos años de estudio.

Agradezco a mis padres por el apoyo brindado para forjarme como un profesional.

Agradezco a la universidad, mi *alma matter*, por haberme cobijado y brindado la formación que ahora me permitirá ayudar a construir una mejor sociedad.

Agradezco de forma muy especial a mi orientador MSc. Marco Antonio Sobrevilla Cabezudo por haberme guiado en esta tesis.

# Resumen

La web es un recurso gigantesco de datos e información sobre seguridad, salud, educación, y otros, que son de mucha utilidad para las personas, pero obtener una s´ıntesis o resumen de uno o varios documentos es una tarea costosa, que de manera manual sería imposible dados estos volúmenes de datos. La generación de resúmenes es una tarea desafiante debido a que involucra el análisis y comprensión del texto escrito en un lenguaje natural no estructurado altamente dependiente de un contexto y que debe describir dicha sintesis de eventos o conocimiento en una forma que resulte natural a las personas que lo leen. Existen distintos enfoques para resumir pudiendo categorizarse en extractivo o abstractivo. En la técnica extractiva, los resúmenes se generan a partir de la selección de oraciones consideradas sobresalientes en el texto origen. Los resúmenes abstractivos se crean regenerando el contenido extraído del texto fuente, por lo que se reformulan las frases por medio de procesos de fusión, compresión o supresión de términos, de esta manera se obtienen frases que en principio parafrasean o que no estaban en el texto original. Este tipo de resúmenes tienen una mayor probabilidad de alcanzar la coherencia y fluidez que tienen los resúmenes humanos. En el presente trabajo se implementa un método para la generación de resúmenes con un enfoque abstractivo, que permite integrar la información semántica (anotador AMR) y discursiva (RST) en un grafo conceptual que será sintetizado mediante el uso de una medida de similitud de conceptos en WordNet. Posteriormente, para encontrar los conceptos más importantes utilizamos PageRank considerando la información discursiva dada por la aplicación del método de O'Donell. Con los conceptos más importantes y la información de los roles semánticos obtenidos del PropBank (que están vinculados en AMR) se implementa un método de generación de lenguaje natural con la utilización de la herramienta SimpleNLG. En el trabajo podremos apreciar los resultados de aplicar este método al corpus de *Document Understanding Conference* 2002 y evaluados por la medida Rouge, ampliamente utilizada en la tarea de generación de resúmenes. El método propuesto alcanza una medida F1 de 24 % en la métrica Rouge-1 para la tarea de generación de resúmenes mono documento. Con esto se demuestra que es factible y más aún provechosa la utilización de estas técnicas, así como recomendamos configuraciones y herramientas útiles en esta tarea.

# Abstract

The web is a giant resource of data and information about security, health, education, and others, matters that have great utility for people, but to get a synthesis or abstract about one or many documents is an expensive labor, which with manual process might be impossible due to the huge amount of data. Abstract generation is a challenging task, due to that involves analysis and comprehension of the written text in non structural natural language dependent of a context and it must describe an events synthesis or knowledge in a simple form, becoming natural for any reader. There are diverse approaches to summarize. These categorized into extractive or abstractive. On abstractive technique, summaries are generated starting from selecting outstanding sentences on source text. Abstractive summaries are created by regenerating the content extracted from source text, through that phrases are reformulated by terms fusion, compression or suppression processes. In this manner, paraphrasing sentences are obtained or even sentences were not in the original text. This summarize type has a major probability to reach coherence and smoothness like one generated by human beings. The present work implements a method that allows to integrate syntactic, semantic (AMR annotator) and discursive (RST) information into a conceptual graph. This will be summarized through the use of a new measure of concept similarity on WordNet.To find the most relevant concepts we use PageRank, considering all discursive information given by the O"Donell method application. With the most important concepts and semantic roles information got from the PropBank, a natural language generation method was implemented with tool SimpleNLG.

In this work we can appreciated the results of applying this method to the corpus of *Document Understanding Conference* 2002 and tested by *Rouge* metric, widely used in the automatic summarization task. Our method reaches a measure F1 of 24 % in Rouge-1 metric for the mono-document abstract generation task. This shows that using these techniques are workable and even more profitable and recommended configurations and useful tools for this task.

# Índice general

# Índice de cuadros

# Índice de figuras

# Capítulo 1

# Introducción

## 1.1. Problemática

La Web es un recurso gigantesco de datos e información que, en las dos últimas décadas, ha experimentado un crecimiento exponencial. De acuerdo a un reporte elaborado por la empresa International Data Corporation (IDC), desde el 2005 hasta el 2020 el universo digital crecerá en un factor de 300 veces, desde los 130 hasta 40000 exabytes de datos, y contendrá datos importantes sobre distintos tópicos útiles a las personas como seguridad, salud, educación, economía, entre otros [Gantz and Reinsel, 2012]. En este contexto, el texto en lenguaje natural es la forma más abundante y natural de representar el conocimiento humano [Zhai and Massung, 2016]. Por ejemplo en la Web, las nuevas redes sociales como Facebook, Twitter, Google o Waze agregan millones de nuevos escritos cada día.

Leer, analizar y tomar decisiones en base a esta abundante información es imposible para una sola persona y económicamente inviable para la sociedad, por lo que es preciso utilizar nuevas tecnologías que permitan automatizar la extracción del contenido más importante y presentar dicha información al usuario en una manera que le resulte de utilidad.

El problema de convertir estos textos escritos en lenguaje natural a información estructurada es una tarea compleja y estudiada por el área del procesamiento del lenguaje natural (PLN), que busca hacer que los computadores realicen tareas útiles con el lenguaje humano, tareas como la comunicación humano-computador, mejorar la comunicación humano-humano o simplemente obteniendo resultados utiles del procesamiento del lenguaje o del habla [Jurafsky and Martin, 2009].

PLN es una tarea compleja, pues el lenguaje natural está diseñado para hacer lo más efectiva y eficiente la comunicación entre personas, por lo que omite gran cantidad de información asumiendo que tanto el emisor como receptor de la comunicación están inmersos en un mismo contexto de comunicación. Asimismo, el lenguaje natural contiene gran cantidad de ambigüedades que los seres humanos pueden resolver en función de este

tácito contexto de comunicación pero que resulta en una alta complejidad para su análisis computacional [Zhai and Massung, 2016].

A pesar de estos problemas, las técnicas de PLN han probado ser de mucha utilidad, por ejemplo en el contexto de las ciencias biomédicas donde han permitido manejar la creciente cantidad de publicaciones relacionadas al estudio de los genomas, construyendo bases de datos de manera automática, que hubieran tomado décadas en ser indexadas manualmente [Baumgartner et al., 2007]. Adicionalmente, en [Chieze et al., 2008] y [Farzindar and Lapalme, 2004] podemos ver como la generación automática de resúmenes ha sido utilizada para mejorar los procesos legales.

La generación de resúmenes es una tarea desafiante, debido a que involucra el análisis y comprensión de texto escrito en un lenguaje natural altamente dependiente de un contexto y que debe describir la s´ıntesis de eventos o conocimiento en una forma que resulte natural a las personas que lo leen.

Existen distintos enfoques que tener en cuenta al momento de generar resúmenes, como es la función del mismo, donde generamos un listado de los contenidos similar a un ´ındice o en otro caso una s´ıntesis del mismo. Otro enfoque esta dado por si estamos interesados en resumir uno o varios documentos. Otro enfoque es si será un resumen para una audiencia con un objetivo específico, o si este estará guiado por las consultas hechas por usuarios o consideraremos resúmenes anteriores [Torres-Moreno, 2014]. Por último otra clasificación está dada por la técnica utilizada para resumir, pudiendo ser extractiva o abstractiva [Mani, 2001].

En la técnica extractiva, los resúmenes se generan a partir de la selección de oraciones consideradas sobresalientes en el texto origen. Las palabras u oraciones se extraen literalmente y se presentan como resumen del texto. Usualmente se utilizan técnicas superficiales para el análisis de los textos, a nivel de sentencia o palabras, por lo que en general los resúmenes no tienen coherencia y solo dan una idea de lo que es sobresaliente en el texto. Pueden ser encontrados algunos ejemplos como la utilización de la frecuencia de términos en [Nenkova and Vanderwende, 2005]. También han sido utilizadas técnicas de análisis profundo para la identificación del contenido más relevante. Ha sido de especial interés el uso del Análisis del Discurso, el cual permite obtener información sobre la coherencia de un texto analizando las relaciones entre las sentencias del mismo. Trabajos originales como los realizados por [O' Donnell, 1997] y, más recientemente, [Uzeda et al., 2008] muestran cómo utilizar esta información para generar resúmenes extractivos.

Los resúmenes abstractivos se crean regenerando el contenido extraído del texto fuente, por lo que se reformulan las frases por medio de procesos de fusión, compresión o supresión de términos [Knight and Marcu, 2000], [Cohn and Lapata, 2009] y [Tanaka et al., 2009]. De esta manera se obtienen frases que en principio parafrasean o que no estaban en el texto original por lo que son originales al resumen.

Para entender mejor esta diferencia podemos hacer una referencia a un contexto cotidiano, por ejemplo cuando se nos pide resumir un texto podemos solo subrayar los términos o sentencias que en nuestro criterio contienen la información más relevante, por

otro lado, si fuera requerido crear un ensayo o artículo, será necesario que además de detectar la información más relevante nos aseguremos de que el texto resultante tenga una coherencia apropiada por lo que necesitaremos muchas veces abstraer los tópicos descritos y reformularlos con nuevas sentencias.

Los resúmenes extractivos han sido ampliamente estudiados como se muestra en las conferencias más importantes relacionadas, como son la *Document Understanding Conference* (DUC) y la *Text Analysis Conference* (TAC), pero estas conferencias también presentan el llamado a utilizar un enfoque abstractivo para mejorar la coherencia y alcanzar una fluidez similar a los resúmenes generados por los seres humanos [Genest and Lapalme, 2012].

Como podemos intuir el enfoque abstractivo necesita de una comprensión más profunda del contenido del texto y debe extraer los conceptos e identificar cómo estos se relacionan en un documento, para luego con esta información poder generar nuevas expresiones, por esta razón, requiere de técnicas no superficiales para el análisis del texto.

En los últimos años los esfuerzos por entender la semántica de las expresiones han dado como fruto recursos como WordNet y Propbank que son bases de conocimiento lingüísticos en la web. Podemos encontrar un ejemplo del uso de estos recurso en *Abstractive Meaning Representation* (AMR) que hace uso de Propbank para poder definir una representacion semántica simple y única a manera de grafo que es de utilidad para los procesos de abstracción [Banarescu et al., 2013]. En [Liu et al., 2015] podemos apreciar como estos grafos son utilizados para la generación de resúmenes abstractivos.

Como lo sugiere [Genest and Lapalme, 2011] para un enfoque abstractivo necesitaremos llevar el modelo de abstracción a un nivel superior, que utilice la información extraída del análisis semántico, pero que también nos permita fusionar los conceptos y manipularlos con el fin de resumir mejor la información. Podemos ver un ejemplo de este enfoque en [Miranda-Jiménez et al., 2014], donde se utiliza un modelo de grafos conceptuales y además se hace uso Wordnet para sintetizar los conceptos que tengan un uso significado común. En un sentido similar, el uso de ontologías para dominios específicos ha probado también ser de utilidad en la tarea de resumir textos con un enfoque abstractivo [Mohan et al., 2016].

Por otro lado, el uso del del Análisis del Discurso ha sido menos estudiado para el enfoque abstractivo, pero de igual manera, presenta aportes importantes en la necesidad de entender el documento como un todo como podemos apreciar en [Gerani et al., 2014]. Estos y muchos otros trabajos evidencian un interés, progreso y necesidad por la investigación del enfoque abstractivo para la generación de resúmenes.

Ante este contexto surge la pregunta, ¿ Es posible incorporar conocimiento semántico, a través de *Abstract Meaning Representation*, y del discurso en la generación automática de resúmenes abstractivos?

En el presente trabajo se muestra un modelo para la generación automática de resúmenes abstractivos para un solo documento utilizando métodos de análisis semántico y del discurso.

Esta información será condensada en un grafo conceptual por documento, que contendrá en sus nodos los conceptos y verbos unidos mediante aristas con información semántica. Además, este grafo considera la información proporcionada por un análisis del discurso entre las sentencias del documento. Esta información será asignada a los nodos en el grafo y posteriormente se utilizará el algoritmo PageRank para definir un ranking semántico-discursivo, para con esta información extraer los subgrafos con el contenido más relevante y que después explotamos para la generación de lenguaje natural y, por ende las sentencias del resumen.

## 1.2. Objetivos

La generación automática de resúmenes para un solo documento ha sido ampliamente estudiada mediante técnicas extractivas que a pesar de conseguir resultados en identificar los términos o sentencias más importantes están lejos de alcanzar la calidad de un resumen hecho por un ser humano, con el fin de poder mejorar la calidad de los resúmenes en cuanto a coherencia y cohesión es necesario continuar con la investigación en nuevas técnicas y enfoques.

### 1.2.1. Objetivo Principal

Implementar un método de generación automática de resúmenes mono documento con un enfoque abstractivo integrando información semántica y discursiva.

### 1.2.2. Objetivos Especificos

- Implementar un método de análisis sintáctico-semántico basado en *Abstractive Meaning Representation* (AMR), que nos permita generar un grafo conceptual por documento.

- Implementar un mecanismo que nos permita unificar conceptos del grafo conceptual generado mediante el uso de resolución de referencias y recursos de conocimiento como Propbank y Wordnet con la finalidad de resumir el texto original.

- Aplicar el algoritmo de PageRank sobre el grafo conceptual incorporando información a nivel de discurso, basado en *Rhetorical Structure Theory* (RST), para identificar los conceptos más importantes del texto a resumir.

- Implementar un método que permita construir oraciones con los conceptos más importantes con la finalidad de generar el resumen abstractivo.

## 1.3. Organización del Texto

El presente trabajo está organizado de la siguiente forma: en el Capítulo 2 se presenta el marco teórico de las técnicas y metodos empleados en los experimentos; en el Capítulo 3 se realiza una revisión sistemática orientada a la generación de resúmenes abstractivos. En el Capítulo 4 describiremos los experimentos y resultados de nuestro modelo de generación de resúmenes abstractivos. Finalmente en el Capítulo 5 serán mostradas las conclusiones y trabajos futuros de la presente tesis.

## 1.4. Publicaciones

El presente trabajo se realiza para la Maestría de Informática de la Pontificia Universidad Católica del Perú y como parte de la misma se han realizado los siguientes trabajos relacionados al presente trabajo.

- A Study of Abstractive Summarization using Semantic Representations and Discourse Level Information, Gregory Valderrama y Marco Sobrevilla, Text Speech and Dialogue International Conference (TSD 2017). Donde se presentaron los resultados obtenidos en el presente trabajo.
- Identificación del Nivel de Peligrosidad en Lima Mediante Minería de Datos en Contenidos de Noticias Web, Gregory Valderrama y Emilio Garcia 1er Workshop on Pattern Recognition and Applied Artificial Intelligence (WRPIAA 2014). Donde se utilizaron técnicas de análisis morfosintáctico para clasificar y asignar un grado de peligrosidad a documentos de noticias locales.
- Análisis de sentimientos en reseñas de películas mediante el uso de *Recursive Neural Tensor Networks*. 2nd Workshop on Pattern Recognition and Applied Artificial Intelligence (WRPIAA 2015)[1]. Donde se exploró el uso de modelos jerárquicos y redes neuronales para obtener una representac´ıon semantica de los comentarios sobre películas y clasificar su valoración positiva o negativa.

[1] http://grpiaa.inf.pucp.edu.pe/wrpiaa2015/wp-content/uploads/2015/10/Conference-agenda.pdf accesado en Febrero 2017

# Capítulo 2

# Marco Teórico

En este nuevo periodo Cámbrico de datos, son necesarias nuevas técnicas y tecnologías para el manejo de grandes volúmenes de información, para poder convertirlos en conocimiento útil para las personas. Ése es el objetivo de la llamada Minería de Datos, en particular cuando trabajamos sobre textos (*Text mining* ), y objetivo del presente trabajo en el contexto de la generación de resúmenes automáticos.

Es así, como distintos tipos de minería de datos utilizan distintos orígenes de datos, como pueden ser sensores de calor o geo-localización. La minería de texto tiene un carácter singular, pues el origen de sus datos es un ser humano, que podemos considerar como un "sensor subjetivo", tal como lo menciona [Zhai and Massung, 2016]. Las personas expresan su particular perspectiva sobre un evento o suceso de la realidad, en un lenguaje natural no estructurado, pudiendo poner mayor o menor atención a distintos aspectos del mismo evento en la realidad, desde esta perspectiva podemos decir que, el objetivo de la minería de texto es también revertir este proceso desde el texto descrito hasta la aproximación más cercana al conocimiento original.

Ahora bien el problema de convertir estos textos no estructurados en información estructurada es una tarea compleja y estudiada por el área del Procesamiento del Lenguaje Natural (PLN), a continuacion profundizaremos en los metodos y tecnicas que son de utilidad para el presente trabajo.

## 2.1. Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural (PLN) tiene como objetivo hacer que los computadores realicen tareas útiles con el lenguaje humano, tareas como la comunicación humano-computador, mejorar la comunicación humano-humano o simplemente obteniendo resultados utilices del procesamiento del lenguaje o del habla. Lo que distingue a PLN de otros sistemas de procesamiento de datos es que utiliza el conocimiento del lenguaje para por ejemplo identificar una palabra en una secuencia de caracteres

[Jurafsky and Martin, 2009].

Dependiendo del dominio de problema se pueden utilizar distintos niveles de conocimiento lingüístico, como [Jurafsky and Martin, 2009]:

- **Fonológico**, el estudio de los patrones de sonido en un lenguaje, el objetivo de este campo de estudio es determinar qué sonidos son significativos y tienen un significado.
- **Fonética**, el estudio de los sonidos del lenguaje humano, para detectar como estos son producidos y recibidos.
- **Morfológico**, el estudio de las unidades de significado en un lenguaje. Un morfema es la más pequeña unidad de un lenguaje que tiene un significado o función. Se incluyen palabras, prefijos, sufijos y otras estructuras que impactan en su significado.
- **Sintáctico**, el estudio de cómo las palabras son combinadas para formar sentencias.
- **Semántico**, el estudio del significado del lenguaje. La semántica examina las relaciones entre las palabras y qué están representando.
- **Discurso**, el estudio del intercambio de información, usualmente en la forma de conversaciones, y particularmente en el flujo de información entre las sentencias.
- **Pragmático**, el estudio de cómo el contexto afecta el significado de las expresiones y qué información es necesaria para inferir un conocimiento oculto o presupuesto.

En la Figura 2.1 se presentan los niveles de conocimiento lingüístico y el creciente grado de complejidad y abstracción de cada nivel. Es así como los análisis en niveles superiores manejan abstracciones más complejas, por esto, las aplicación de estos niveles son conocidos como aplicaciones de abordaje profundo y las que usan conocimientos de los niveles inferiores, son consideradas aplicaciones de enfoque superficial.

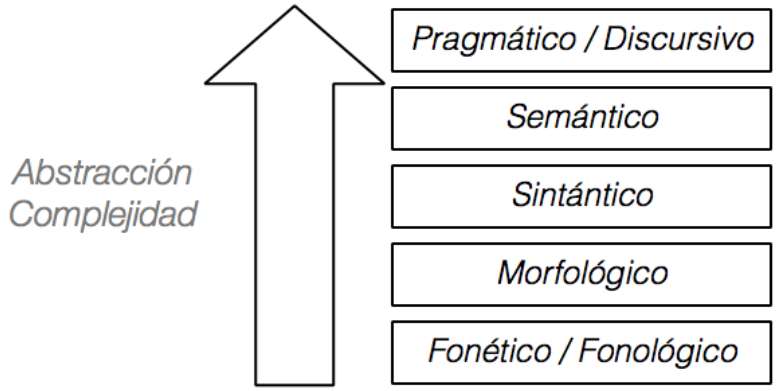


Figura 2.1: Abstracción y complejidad en los niveles de conocimiento linguístico. Extraído y adaptado de [Nóbrega et al., 2013]

## 2.2. Generación Automática de Resúmenes

Según el estándar ANSI[1], un resumen puede ser definido de manera general como:

"*Una representación abreviada y precisa de los contenidos de un documento, preferentemente preparado por sus autores para ser publicado con el mismo. Los resumenes son útiles para facilitar el acceso a publicaciones y generar bases de datos accesibles por computadores* "

La generación de resúmenes por seres humanos es un proceso costoso, que a profesionales en la materia toma entre 8 a 12 minutos [Cremmins, 1996]. Este factor se ve claramente influenciado por si el texto pertenece al campo de dominio del profesional, por esta razón, la posibilidad de generar dichos resúmenes utilizando computadores es una necesidad, dado el volumen de información actual.

En el contexto de las ciencias de la computación utilizaremos la siguiente definición [Hovy and Miktov, 2005]:

*"Un resumen automático es un texto generado por un software, que es coherente y contiene una significativa cantidad de información relevante de la fuente original y su ratio de compresión es menor a la tercera parte de la longitud original del documento".*

El concepto de ratio de compresión estará dado por la Fórmula 2.1. Como se dijo este ratio puede variar desde 10 % hasta 30 % de la longitud del texto original.

$$r = \frac{|Summary|}{|Source|} \tag{2.1}$$

Un resumen automático necesita considerar una etapa de selección del contenido más relevante, una vez identificado este contenido debe ser ordenado de una manera que siga un orden coherente y por último, las expresiones utilizadas en el nuevo texto o resumen deben seguir una fluidez apropiada en el idioma objetivo [Jurafsky and Martin, 2009].

En [Mani, 2001] se propone una arquitectura de tres etapas (Figura 2.2). En la etapa de análisis, los textos de entrada son interpretados y representados en un formato computacional, en la etapa de transformación dicha representación es procesada para identificar y seleccionar el contenido más relevante y como resultado tendremos una representación computacional condensada de los textos. En la etapa de síntesis es generado un texto en lenguaje natural.

Ahora bien un buen resumen debe mantener una cohesión y coherencia apropiada, como nos menciona [Barzilay and Elhadad, 1999].

La cohesión es un atributo lingüístico de la sentencia y es lograda por el uso apropiado de términos semánticamente relacionados, la correferencia, elipsis y conjunciones.

[1] http://www.ansi.org accesado en Febrero 2017

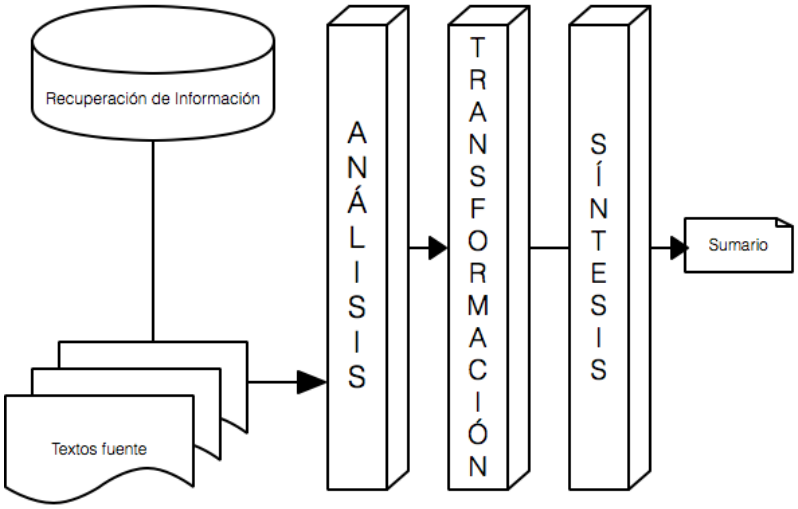


Figura 2.2: Arquitectura de un sistema de Sumarización Automática. Extraído y adaptado de [Mani, 2001]

La cohesión en las sentencias será disminuida si tenemos anáforas o referencias temporales no resueltas.

La coherencia es un atributo semántico localizado en un nivel superior de abstracción llamado Discurso, la aparición de contradicciones y redundancias afectarán negativamente la coherencia de un documento [Torres-Moreno, 2014].

## Categorías para la Generación de Resúmenes

La tarea de resumir documentos puede ser categorizada por diferentes criterios [Torres-Moreno, 2014]:

De acuerdo a su función:

- Resumen indicativo, el cual provee información sobre los tópicos discutidos en el documento, por ejemplo la generación de una tabla de contenidos.
- Resumen informativo, el cual busca reflejar el contenido del documento, como una versión corta del mismo.

De acuerdo al número de documentos para resumir:

- Resúmenes de un solo documento, es el resumen de un solo documento
- Resúmenes de varios documentos, es el resumen de varios documentos que usualmente pertenecen a un tópico común.

De acuerdo al tipo de técnica para resumir:

- Resumen extractivo, utiliza fragmentos del documento original
- Resumen abstractivo, para generar el resumen reescribe o parafrasea el documento original.

De acuerdo al contexto:

- Resumen genérico, un resumen que no está enfocado en un contexto.
- Resumen guiado por consultas, un resumen que toma en consideración preguntas específicas dadas por el usuario
- Resumen por actualización, es un resumen que toma en consideración conocimiento previo que pueda tener el usuario con respecto a documentos y resúmenes que previamente ha revisado, con el objetivo de evitar información repetida.

De acuerdo a la audiencia objetivo:

- Sin un tema, sera un resumen que no tomo en consideracion un tema específico del usuario.
- Con un tema objetivo, es un resumen que está enmarcado en un contexto particular como ideología, política, etc.

Como se ya se mencionó también para las tareas de sumarización existen enfoques tanto superficiales como profundos.

### 2.2.1. Métodos de Enfoque Superficial

Técnicas conocidas así por no profundizar en el conocimiento lingüístico sino que hacen uso de métodos estadísticos o empíricos con base en elementos superficiales del texto intentando descubrir patrones en el mismo.

Este enfoque no profundo hace uso de la presunción de independencia entre las palabras de uno o varios documentos, también llamado *bag-of-words*, que intencionalmente ignora la información de posición de las palabras por lo que un texto puede ser descrito simplemente por la existencia o no de ciertas palabras [Jurafsky and Martin, 2009].

Podemos utilizar esta presunción de independencia entre palabras por ejemplo, para la recuperación de la información más relevante, representando un documento o sentencia como un vector binario que indica la existencia o no de una palabra, es así cómo podemos llevar estos vectores a un espacio común de operación, en el cual usualmente utilizamos

una medida de distancia para calcular la similitud entre ellas y reducir su número para generar un resumen. [Zhai and Massung, 2016].

Tambien podemos utilizar modelos probabil´ısticos, donde se asume que las palabras son observaciones de una variable aleatoria por lo que podemos utilizar modelos probabilísticos para explicar la relación entre los ítems de un corpus, donde podemos detectar la probabilidad de pertenencia de un ítem a un tópico en particular y con esto generar un resumen.

De acuerdo con [Mani, 2001], en el contexto de la sumarización tenemos algunos métodos de enfoque superficial que utilizan palabras clave y su frecuencia en el texto, palabras clave en los títulos, localización de las sentencias y la utilización de ciertas palabras del diccionario como indicadores de importancia en determinadas sentencias.

### 2.2.2. Métodos de Enfoque Profundo

Estos métodos hacen uso de un conocimiento lingüístico para analizar y seleccionar el contenido de los resúmenes, estos comprenden el uso de reglas gramaticales, ontologías y otras informaciones semántico-discursivas, algunas de las cuales mostraremos a continuación.

#### Anotación Gramatical de Palabras *(Part-of-Speech Tagging)*

Este anotador se encarga de asignar una anotación con información relacionada a la clasificación gramatical que puede tener una palabra del corpus. Por ejemplo detectar que una palabra es un nombre propio, adjetivo, art´ıculo, verbo, entre otros.

#### Reconocimiento de entidades

El reconocimiento de entidades o *Named Entity Recognition* (NER) es el trabajo de identificar todas las entidades mencionadas en un texto que pueden ser nombres de personas, lugares, organizaciones, entre otras. Por ejemplo detectar todos los nombres de genes y prote´ınas en un corpus [Settles, 2005].

Esta tarea se enfrenta a dos tipo de ambigüedades, la primera relacionada a la posibilidad de que un nombre identifique dos entidades del mismo tipo, por ejemplo, padre e hijo pueden llevar el mismo primer nombre y apellido. Por otro lado, el nombre puede identificar dos entidades de distinto tipo como por ejemplo el nombre de un aeropuerto y de un héroe o figura nacional.

Existen en Internet listas de nombres de entidades que pueden ser consultadas para poder identificar lugares, personajes u organizaciones. Estas listas reciben el nombre de

*Gazetteers* [2].

### Detección y clasificacion de relaciones

La detección y clasificación de relaciones (*Relation Detection and Classification*), es la tarea de encontrar y clasificar las relaciones semánticas entre las entidades, por ejemplo relaciones de familia, empleador, parte-todo, pertenencia, geoespacial, entre otras. Se considera que existe una relación entre esta tarea y encontrar la relación semántica entre las palabras de un texto [Jurafsky and Martin, 2009].

### Reconocimiento de expresiones temporales y análisis temporal

El reconocimiento de expresiones temporales (*Temporal Expression Recognition*) es necesario para pasar al análisis temporal (*Temporal Analysis* ) que busca resolver cuando un evento ha sucedido y cómo está relacionado con otro en cuanto a la dimensión del tiempo se refiere [Jurafsky and Martin, 2009]. Por ejemplo: Lunes, Martes, Siguiente Feriado, 3.30 PM, medio dia, entre otros.

### Anotación de papeles semánticos

La tarea de la anotación de roles semánticos (*Semantic Role Labeling* (SRL) también llamado *Thematic Role Labeling* ) es la de asociar el significado de las palabras con el significado de la sentencia en conjunto. Por lo tanto, buscará encontrar automáticamente los roles semánticos para cada predicado en una sentencia. En específico esto significa determinar cuales constituyentes en una sentencia son argumentos semánticos para un predicado y entonces determinar el apropiado rol para cada uno de estos argumentos [Jurafsky and Martin, 2009].

Un ejemplo de los roles semánticos etiquedados puede encontrarse en [Cook, 1989]:

- *Agent*, es el participante del evento que provoca que el mismo suceda.
- *Theme/figure*, es el participante del evento que sufre un cambio en posición o estado.
- *Experiencer*, es el participante del evento que experimenta algo.
- *Source*, es la localización o lugar donde la acción empieza.
- *Goal*, es la localización o lugar a la que la acción se dirige o donde termina.
- *Recipient*, es la persona que está en posesión del theme.
- *Patient*, es el participante del evento que es afectado por el evento.

[2] Disponible en www.geonames.org accesado en Febrero 2017

- *Instrument*, es el participante del evento usado por el agente para hacer o causar el evento.
- *Location/ground*, La localización o el lugar asociado con el evento mismo.
- *Time*, Momento en el que un objeto o un evento está localizado.

Por ejemplo, en la sentencia adaptada de [Manchego, 2013], "*Juan rompio una ventana con una piedra*", al utilizar un anotador de papeles semánticos se producirá un resultado similar a este:

$$[Juan_{agent}][rompio_{verb}][una\ ventana_{patient}][con\ una\ piedra_{instrument}]$$

**Resolución de Referencias**

La resolución de referencias o *Reference Resolution (Coreference Resolution)* se da una vez detectadas las entidades mencionadas en un texto, y consiste en agrupar todas las referencias a la misma entidad. Por ejemplo, saber que en el primer párrafo se habla de una empresa aérea y en el párrafo tercero se menciona el mismo nombre con referencia a la misma instancia de línea aérea [Jurafsky and Martin, 2009].

**Análisis del discurso**

El análisis del discurso es una técnica de enfoque profundo que busca entender y explicar las relaciones entre las expresiones dentro de un mismo documentos, data la importancia de este tópico para nuestro trabajo se explicará en profundidad en la sección 2.4.

A continuación se describe la generación de resúmenes extractiva y abstractiva, que, como se mencionó puede ser visto como cuando un estudiante tiene que producir un resumen, donde en primera instancia puede solo identificar las sentencias más importantes o puede comprender el texto y con esto crear en sus propias palabras un resumen. Es claro que muchos avances se han hecho en cuanto a los resúmenes extractivos entre otras cosas porque requieren un enfoque menos profundo en cuanto a la comprensión del lenguaje pero es claro también que el enfoque abstractivo es el tópico de más interés para las investigaciones actuales. Por ejemplo, a continuación podemos realizar una comparación entre un resumen extractivo y abstractivo donde podemos notar una menor calidad en el resumen de un método extractivo en idioma Ingles [Mani, 2001].

> *"Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.*
>
> *Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here*

*gave their lives that that nation might live. It is altogether fitting and proper that we should do this.*

*But, in a larger sense, we can not dedicate – we can not consecrate – we can not hallow – this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us – that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion – that we here highly resolve that these dead shall not have died in vain – that this nation, under God, shall have a new birth of freedom – and that government of the people, by the people, for the people, shall not perish from the earth."*

Resumen Extractivo:

*Fourscore and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. The brave men, living and dead who struggled here, have consecrated it far above our power to add or detract.*

Resumen Abstractivo:

*This speech by Abraham Lincoln commemorates soldiers who laid down their lives in the Battle of Gettysburg. It offers an eloquent reminder to the troops that it is the future of freedom in America that they are fighting for.*

### 2.2.3. Generación Automática de Resúmenes Método Extractivo

La generación de resúmenes automáticos mediante un método extractivo consiste en identificar y seleccionar unidades de texto como sentencias, segmentos de sentencias o palabras que contienen la información más relevante y que posteriormente podamos utilizar para generar un resumen. [Das and Martins, 2007].

Los métodos extractivos pueden utilizar tanto enfoques superficiales como *bag-of-words*, métodos probabilísticos o la localización específica de palabras o sentencias en un documento. Y también hacer uso de enfoques profundos como algoritmos de resolución de dependencias o análisis del discurso, entre otros.

Pero debido a que no necesitan generar nuevas expresiones en lenguaje natural no les es necesario tener una comprensión profunda de los sucesos contenidos en el texto. Por lo que aún detectando las palabras con la información más valiosa, usualmente presentan limitaciones al momento de obtener un resumen conciso, coherente y que presente una fluidez apropiada tal como lo hace un ser humano [Carenini and Cheung, 2008].

### 2.2.4. Generación Automática de Resúmenes Método Abstractivo

En este tipo de método un resumen puede contener nuevas sentencias que no estaban presentes en el documento original, por lo que será necesario considerar la tarea de generación de lenguaje natural. Es así como [Genest and Lapalme, 2012] nos dicen que un método abstractivo que considere un proceso para el análisis del texto, la selección del contenido y la generación de nuevas sentencias tiene el mejor potencial para generar resúmenes comparables con los de un ser humano. Aunque es posible utilizar técnicas de enfoque superficial para obtener cierto conocimiento sobre el contenido de un texto y con esto generar nuevas expresiones, es usual que los métodos abstractivos recurran a métodos de enfoque profundo como reconocimiento de entidades, relaciones, resolución de anáforas entre otros y también bases de datos de conocimiento como ontologías que permitan aplicar cierta lógica y conocimiento para inferir nuevas sentencias.

Usualmente los métodos abstractivos recurren a técnicas de compresión y fusión de texto que buscarán eliminar las partes menos importante o combinarlas para mostrar la información más relevante [Radev et al., 2002].

A continuación presentaremos más información sobre las tareas de generación de resúmenes para uno y varios documentos.

### 2.2.5. Generación Automática de Resumen para un Documento

Por el número de documentos a resumir, la tarea de generar resúmenes automáticamente puede ser categorizada en mono documento, cuando buscamos resumir un solo documento, o multi documento cuando tenemos varios documentos que hablan de un tema en común.

La categoria mono documento está lejos de alcanzar la calidad del trabajo humano, entre otras razones debido a que al contar con un solo documento no podemos confiar en la redundancia de información que por ejemplo tenemos en la versión multi documento, esto requiere que utilicemos técnicas de enfoque profundo para extraer toda la información posible del texto original, adicionalmente la gran mayoría de trabajos son de carácter extractivo lo cual también dificulta conseguir la fluidez que tienen los resúmenes realizados por personas.

Uno de los primeros trabajos está dado por Hans Peter Luhn [Luhn, 1958] como parte del grupo de investigación de IBM donde propuso utilizar la frecuencia de palabras como un indicador de su importancia, después utilizó dicho valor para premiar las sentencias que mas de estas importantes palabras contuviese, con las mejor rankeadas sentencias se construiría el resumen. Posteriormente [Baxendale, 1958] en IBM, también y basando en el estudio de los párrafos, concluyó que el el 92 % de los párrafos contienen la información más relevante en las dos primeras líneas, por lo que utilizó la información de la posición de la sentencia en el documento como indicador de importancia. Posterior-

mente [Edmundson, 1969] propuso la utilización de ciertas palabras clave como indicador de importancia y la utilización de la información de estructura del documento por ejemplo si una sentencia era el título o estaba declarada como subtítulo. Posteriormente [Kupiec et al., 1995] adiciona la idea de tomar en consideración la longitud de las sentencias y la presencia de palabras en mayúscula, posteriormente se consideran características de un enfoque más profundo como el uso de atributos sintácticos o reconocimiento de entidades y se hace uso de algoritmos de aprendizaje del computador como Naive Bayes [Kupiec et al., 1995], Hidden Markov Models [Conroy and O'leary, 2001] y Modelos de regresión Lineal Logística (Osborne, 2002).

Durante la DUC 2001 y 2002 se introdujo la tarea de generar un resumen de 100 palabras desde documentos de noticias y lo sorprendente fue que ningun metodo pudo sobrepasar la línea base propuesta que utilizaba las primeras sentencias de cada artículo y por esta razón esta categoría de resumen no fue tomada en cuenta en las siguientes ediciones de la DUC. En [Nenkova and Vanderwende, 2005] se menciona que este resultado se debió a la particularidad que tiene los artículos de noticias de colocar siempre la parte más importante en las primeras líneas. Empero es necesario continuar con la investigación pues no toda la información está escrita en dicha manera y porque es posible obtener mejores resultados tal como lo presenta [Svore et al., 2007] donde propone un algoritmo basado en redes neurales que supera la línea base propuesta por la DUC. En un enfoque distinto con la utilización del análisis del discurso (RST) y de plantear la optimización del árbol discursivo mediante el algoritmo la programación lineal se alcanza una nueva línea base en [Hirao et al., 2013], así mismo nuevos enfoques en la utilización de teoría de grafos [Oliveira et al., 2016], resolución de anáforas [Durrett et al., 2016] y métodos abstractivos [Liu et al., 2015] demuestran que es factible crear mejores resúmenes tanto en su capacidad para ubicar la información más importante como en tener una calidad lingüística apropiada.

### 2.2.6. Generación Automática de Resumen para Varios Documentos

Este tipo de resumen consiste en extraer un resumen desde múltiples documentos que usualmente pertenecen a un dominio común de interés. Aunque no por esto están exentos de presentar información contradictoria por lo que esta tarea no es solo la de ubicar el contenido más importante entre varios documentos sino también la de asegurar que el resumen sea coherente [Das and Martins, 2007].

## 2.3. Métricas de Evaluación

La definición de un buen resumen sería, todo aquel texto que sea fácil de leer y dé una visión general del contenido del texto original o fuente [Saggion et al., 2016]. Dado que los resúmenes tienden a orientarse cada vez más hacia necesidades específicas, es necesario refinar los métodos de evaluación existentes.

Lamentablemente, estas necesidades no dan una base clara para la evaluación y la definición de lo que és un buen resumen, sigue siendo en gran medida una cuestión abierta. Por lo tanto, la evaluación de resúmenes automáticos o realizados por humanos, se conoce como un tarea dif ́ıcil. Es dif ́ıcil para los seres humanos, lo que significa que la automatización de la tarea es aún más difícil de realizar y evaluar. Sin embargo, debido a la importancia del esfuerzo de investigación en el resumen automático, se han realizado una serie de propuestas para automatizar parcialmente o totalmente la evaluación [Galliers and Jones, 1993, Zajic et al., 2007]. También es útil señalar que en la mayoría de los casos las evaluaciones automáticas ya se correlacionan positivamente con las evaluaciones humanas.[Saggion et al., 2016]

### 2.3.1. Evaluación de Resúmenes Automáticos

En los Estados Unidos, desde finales de los 90s, se han organizado una serie de campañas de evaluación y discucsión que son, escencialmente *TIPSTER Text Summarization Evaluation* (SUMMAC) [Mani et al., 2002], *Document Understanding Conference* (DUC) [Ono et al., 1994] y, más recientemente *Text Analysis Conference* (TAC). La evaluación en estas conferencias se basa en la puntuación humana y automática de los resúmenes propuestos por los participantes. Por lo tanto, estas conferencias han desempeñado un papel importante en el diseño de las medidas de evaluación; También desempeñan un papel en la metaevaluación de los métodos de puntuación, ya que es posible comprobar hasta qué punto las puntuaciones obtenidas se correlacionan automáticamente con los juicios humanos. En términos generales, como se menciona en [Saggion et al., 2016], podríamos decir que tenemos tres dificultades principales al momento de la evaluación:

- Determinar cuáles son los datos más importantes que deberán guardarse del texto inicial.
- Los evaluadores deben ser capaces de reconocer automáticamente estas piezas de información en el resumen del candidato, ya que ésta información puede expresarse utilizando diversas expresiones.
- Evaluar la legibilidad (incluida la gramática y coherencia) del resumen.

Incluso para los resúmenes extractivos, los métodos de evaluación van desde enfoques puramente manuales a los puramente automáticos, y por supuesto hay muchas posibilidades en el medio. Los enfoques manuales se refieren a métodos en los que un ser humano evalúa un resumen de candidatos desde diferentes puntos de vista, por ejemplo, cobertura, gramática o estilo; Este tipo de evaluación es necesaria, pero se sabe que es altamente subjetiva. Los enfoques automáticos comparan segmentos de textos del resumen del candidato con uno o varios resúmenes de referencia; Este enfoque es fácil de reproducir pero no puede aplicarse cuando el sistema utiliza técnicas de reformulación. Los enfoques mixtos permiten analizar y anotar manualmente las informaciones más importantes y clasificar los resúmenes de los candidatos de acuerdo con estos (las informaciones más

importantes deben estar contenidas en el resumen del candidato, independientemente de su formulación lingüística).[Saggion et al., 2016]

### 2.3.2. Métodos Manuales

La forma más obvia y simple de evaluar un resumen, es la de tener 'asesores' o personas que validen la calidad del resumen obtenido. La evaluación manual puede proveer algunos indicadores sobre la cualidad y legibilidad de un texto. Un buen resumen debe de ser:

- Preciso sintácticamente
- Semánticamente coherente
- Lógicamente organizado
- No redundante

Estos puntos son muy complejos de ser calculados automáticamente, especialmente la coherencia semántica y la organizacion lógica. Con el fin de obtener una evaluación confiable de los puntos mencionados anteriormente, es necesario tener 'jueces' humanos. Para TAC 2009, resumenes escritos por expertos tubieron un promedio de 8.8/10. Por lo tanto, este valor puede ser visto como la puntuación límite alcanzable por resúmenes.[Saggion et al., 2016]

### 2.3.3. Métricas para la Generación de Resúmenes Automáticos

Desde principios de los años 2000, una serie de medidas han sido propuestas para automatizar la evaluacion de resúmenes. La mayoría de estas medidas estan basadas en una comparación directa con el resumen producido por un ser humano [Saggion et al., 2002, Radev et al., 2003].

La precisión, exhaustividad y la exactitud son medidas comúnmente usadas para medir la bondad de un método de clasificación. En el contexto de la generación de resúmenes podemos formular esta clasificación como la decisión de incluir o no ciertas palabras en un resumen de acuerdo a ciertas caracter´ısticas. Entonces si una palabra ha sido incluida en el resumen generado automáticamente y también está presente en el resumen generado manualmente diremos que es un verdadero positivo ($VP$ ), siguiendo este razonamiento podemos tener falsos positivos ($FP$), verdaderos negativos ($VN$) o falsos negativos ($FN$).

Entonces la precisión es la probabilidad de clasificar correctamente un elemento con respecto a todos los otros elementos que han sido escogidos correctos o incorrectos:

$$Precision_i = \frac{VP_i}{VP_i + FP_i} \tag{2.2}$$

Exhaustividad es la probabilidad de que un elemento se ha puesto en una la categor´ıa correcta dentro de todos los elementos que debieron estar en la misma.

$$Exhaustividad_i = \frac{VP_i}{VP_i + FN_i} \tag{2.3}$$

Se suele preferir el uso de la precisión y exhaustividad en una sola medida de bondad para un modelo, cuya combinación otorga la misma importancia para ambas medidas es conocida como medida *F1* que esta data por:

$$F = 2 \cdot \frac{Precision \cdot Exhaustividad}{Precision + Exhaustividad} \tag{2.4}$$

**ROUGE**

La medida *Recall-Oriented Understudy of Gisting Evaluation* (ROUGE) fue introducida por [Lin, 2004] y también hace uso de los conceptos de exhaustividad y de exactitud pero busca obtener una métrica más apropiada para el dominio de los resúmenes generados automaticamente.

Estas medidas están basadas en la comparación de n-gramas (ej. una secuencia de $n$ elementos) entre el resumen candidato (el resumen a ser evaluado) y una de va- rias referencias de resúmenes generados manualmente. ROUGE fue inspirado por BLEU [Papineni et al., 2002], una medida utilizada en la traducción automática, también basada en la comparación de $n$-gramas.

Existen varias variantes de ROUGE, las cuales son:

- **ROUGE-*n*** Basada en la comparación de n-gramas (una secuencia de 2 o 3 elementos, rara vez 4). Una serie de $n$-gramas, por lo tanto series de secuencias de $n$ palabras consecutivas, es extra´ıda de los resumenes referencia y el resumen candidato. La calificación es el radio entre el numero de n-gramas comunes, entre el resumen candidato y la referencia, y el numero de n-gramas extraidos desde solamente el resumen referencia.
- **ROUGE-*L*** Cubre las debilidades de *ROUGE-n*, es decir, el hecho de que la medida pudo estar basada en secuencias de texto muy pequeñas; *ROUGE-L* toma en consideración la secuencia común más larga entre dos secuencias de texto divididas por la longitud del texto. Incluso si este metodo es mas flexible de ROUGE-n, continua dependiendo de la continuidad de los n-gramas.

- **ROUGE-*SU*** *Skip-bi-gram* y *uni-gram* ROUGE toma en consideración bigramas tanto como unigramas. Sin embargo, los bi-gramas, en lugar de ser sólo secuencias continuas de palabras, permiten inserciones de palabras entre su primer y último elemento. La distancia máxima entre los dos elementos del bi-grama corresponde a un parámetro ($n$) de la medida (a menudo, la medida es instanciada con $n = 4$). Durante el TAC 2008, se ha demostrado que ROUGE-SU fue la medida más correlacionada con los juicios humanos.

## 2.4. Análisis del Discurso

Más allá del análisis de las palabras y su relación con las sentencias, este tipo de análisis está enfocado en obtener una estructura coherente entre sentencias que llamaremos discurso [Jurafsky and Martin, 2009]. Por ejemplo, si tomamos una serie de sentencias sintácticamente correctas y las colocamos en un documento, no necesariamente tendremos un discurso pues la coherencia está dada por la existencia de conexiones significativas (relaciones de coherencia) entre las sentencias, como la de resultado, explicación, paralelismo, elaboración, entre otros.

*Rhetorical Structure Theory* (RST) [Mann and Thompson, 1988] es una teoría ampliamente utilizada para este propósito. RST está basado en un grupo relaciones retóricas que pueden unir sentencias en un discurso y con esta información identificar las unidades elementales del mismo también llamadas *Elementary Discourse Unit* (EDU), que son las unidades mínimas de sentido lógico que comúnmente son expresadas mediante sentencias u oraciones.

Se identifican dos tipos de EDU en dichas relaciones, uno llamado nuclear y otro satelital. El nuclear representa la idea principal del escritor y puede ser interpretada independientemente y el satelital solo puede ser interpreta en relación al nuclear.

Por ejemplo en la expresión "Kevin debe estar aquí, su carro está parqueado afuera". Se define como nuclear la sentencia "Kevin deve estar aquí" y cómo satelital "su carro está parqueado afuera". Desde el punto de vista de la naturaleza de estas relaciones RST puede presentar tanto relaciones de naturaleza intencional-argumentativa cuanto de naturaleza semántica-informativa-factual.

Con base en este trabajo se han creado recursos para identificar estas relaciones como el presente en RST TreeBank [Carlson et al., 2003] que define 78 relaciones agrupadas en 16 clases. Por ejemplo en [Ono et al., 1994] y [Uzeda et al., 2008] se puede ver como esta información de información nuclear y satelital es utilizada para generar resúmenes automáticos.

## 2.5. Recursos lingüísticos

### 2.5.1. Corpus Anotado

Un corpus es una colección de textos, que pueden ser procesados por un computador [Jurafsky and Martin, 2009] y que sirven a un propósito de estudio en un dominio de problema específico.

La adición de metadata específica a este dominio es conocido como proceso de anotación. Un corpus que tiene estas anotaciones es conocido como un Corpus Anotado. Con un corpus anotado podemos utilizar distintos métodos de aprendizaje del computador supervisados, no supervisados y semi-supervisados con el fin de detectar patrones e inferencias, así como podemos establecer una medida de bondad en las distintas tareas computacionales que realizamos en un dominio específico.

### 2.5.2. WordNet

*WordNet* es un recurso que representa una gran base de datos léxica del idioma inglés. Los sustantivos, verbos, adjetivos y adverbios se agrupan en conjuntos de sinónimos sintéticos (*synsets* ), cada uno expresando un concepto distinto. Los *synsets* están interrelacionados por medio de relaciones conceptuales-semánticas y léxicas. Estas relaciones pueden ser de hiperonimia, hiponimia, coordinacion, holonimia, meronimia entre otras. [Miller, 1995].

*WordNet* agrupa las palabras en base al significado de las mismas. Tomando, sin embargo, algunas distinciones importantes:

- *WordNet* enlaza no solo las palábras similares, sino también los sentidos específicos de éstas. Resultando, en una desambiguación de las palabras que tienen una estrecha proximidad entre sí.
- *WordNet* etiqueta las relaciones semánticas entre palabras, a diferencia de la agrupación realizada en un tesauro, en donde no se sigue ningún patrón explícito que no sea la similitud del significado.

En el presente trabajo *WordNet* representa un recurso linguistico valioso para detectar si dos palabras se refieren al mismo concepto y con esto poder fusionarlas.

### 2.5.3. PropBank

El proyecto *PropBank* ha sido sumamente influyente en la última década para el procesamiento de lenguaje natural [Kingsbury and Palmer, 2003]. *PropBank* es un *corpus* o 'banco' de proposiciones verbales y sus argumentos en Inglés.[Kingsbury and Palmer, 2002]

Para el entrenamiento y generación de este *corpus* se utilizó como base un consenso desarrollado durante el año 2000, entre los grupos de *Bolt, Beranek and Newman* (BBN), MITRE, la Universidad de Nueva York y Penn. Tomando como punto de partida el *Penn Treebank II Wall Street Corpus* de un millón de palabras [Marcus et al., 1994].

Los argumentos esperados de cada sentido se numeran secuencialmente de Arg0 a Arg5. Según las directrices establecidas por la comunidad *Attempto Controlled English* (ACE) [Fuchs and Schwitter, 1996].

- Usos comunes para los argumentos:
  - Arg0: agent
    Arg1: patient
    Arg2: instrument/attribute
    Arg3: starting
    point/attribute Arg4: ending
    point
    ArgM: modifier

- Por ejemplo, en la sentencia "*Obama met him privately in the White House, on Thursday* ". Podemos apreciar el Arg0 que hace referencia al agente (*Agent* ), Arg1 que hace referencia al Paciente (*Patient* ), el argumento ArgM-MNR que hace referencia a la manera (*Manner*), el argumento ArgM-LOC que hace referencia a la ubicación (*Location*) y por Ãžltimo el ARGM-TMP que hace referencia al tiempo (*Time*) en que sucedió el evento. Como podemos apreciar FrameNet utiliza el identificador ArgM para identificar argumentos especiales.
  - Rel: met
    Arg0:
    Obama
    Arg1: him
    ArgM-MNR: privately
    ArgM-LOC: in the White House
    ArgM-TMP: on Thursday

PropBank no tiene como propósito hacer que los *tags* de los argumentos tengan el mismo "significado" de un sentido del verbo a otro, por lo que el "rol" desempeñado por Arg2 en un sentido de un predicado dado, puede ser interpretado por Arg3 en otro sentido [Kingsbury and Palmer, 2002].

Aunque no existe un acuerdo sobre el significado absoluto de un argumento en la base de datos de *PropBank* se encuentra una referencia al rol semántico que representa dicho argumento para cada verbo.

### *2.5.4. Abstractive Meaning Representation*

Como muchos autores intuyen y lo mencionan [Genest and Lapalme, 2012], se piensa

que un método completamente abstractivo requiere un proceso separado de análisis del

texto que sirva como un intermediario antes de la generación de nuevas sentencias.

Es asi como en 2013 múltiples autores propusieron un lenguaje de representación semántico común útil para los procesos de abstracción el cual llamaron *Abstract Meaning Representation* (AMR) [Banarescu et al., 2013]. En este lenguaje se puede describir la información semántica de las sentencias a manera de grafo, con el objetivo de proponer un simple y único modelo de representacion semántica de manera similar a los modelos sintácticos como Penn Treebank [Marcus et al., 1993], debido a que el hecho de tener múltiples formas de anotación para el reconocimiento de entidades, resolución de referencias, relaciones semánticas, reconocimiento de entidades temporales, etc. retrasa el desarrollo del área, como hubiera sido el caso del análisis sintáctico si hubiera recurrido a una distinta notación para el reconocimiento de sujeto, verbos, artículos, etc.

Los principios que rigen AMR son:

- AMR es un grafo que tiene un nodo raíz y nodos correctamente marcados con información semántica en base a los trabajos de [Shieber, 1986] y PENMAN [Mathiessen and Bateman, 1991] que debe ser fácilmente entendible por un ser humano y consultable por un programa.
- Trata de asignar una misma representación a sentencias que tengan el mismo significado como *"he described her as a genius", "his description of her: genius", "she was a genius, according to his description"* todas estas sentencias son asignadas a la misma representación AMR porque semánticamente es una persona masculina describiendo un adjetivo de una persona femenina en distintas voces pasiva y activa.
- AMR hace uso intensivo de PropBank [Kingsbury and Palmer, 2002, Palmer et al., 2005] para poder generar una correcta abstraction por ejemplo para el frameset "describe-01" Propbank nos da la información de que necesitamos 3 nodos de información (:arg0 la persona que describe, :arg1 la cosa que se describe, :arg2 que se describe de la cosa). AMR utiliza hasta 100 relaciones distintas que pueden estar basadas en PropBank aunque también se han adicionado relaciones especiales como de fecha, de cantidad y de usos comunes.
- AMR es agnóstico del mecanismo que utilizamos para llevar una sentencia a dicha representación o viceversa.
- Esta direccionado para el idioma Inglés por lo que no es una solución para problemas de traducción.

El lenguaje de AMR puede representar frames de manera similar a PropBank, relaciones semánticas, relaciones de referencia, relaciones inversas, expresiones de negación, sentencias de preguntas, verbos, nombre propios, adjetivos, preposiciones, entidades nombradas entre otras. Aunque es un modelo de representación bastante completo tiene limitaciones al no considerar el número y artículos de las sentencias, tampoco posee una representación para cuantificadores universales como *All* y tampoco puede diferenciar entre eventos reales y eventos hipotéticos por ejemplo en la sentencia *"the boy wants to go"*

las instancias de *"want-01"* y *"go-01"* tendrán el mismo estado a pesar de que *"go-01"* puede o no ocurrir.

Por ejemplo, para la expresión *"The dog wants to eat the bone"*, un anotador de AMR nos presentara el siguiente resultado en formato PENMAN:

```
(want-01
      :
├─ ARG0 (d / dog)
├─ :ARG1 (e / eat-01
   ├─ :ARG0 d
   ├─ :ARG1 (b / bone)))
```

Se puede representar la misma sentencia en un formato de grafo (Figura 2.3).:

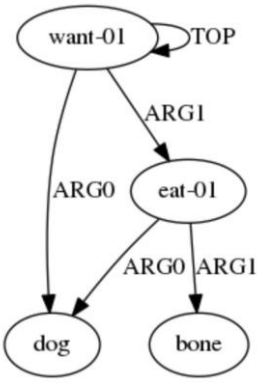


Figura 2.3: Representación AMR en forma de grafo de la oración *"The dog wants to eat the bone"*

AMR en este momento cuenta con un fuerte apoyo de la comunidad de investigacion y se han creado corpus y parsers para el idioma inglés [Knight et al., 2014] [Flanigan et al., 2014].

# Capítulo 3

# Estado del arte

Para poder entender el contexto actual de la generación automática de resúmenes y en particular su forma abstractiva, se realizó una revisión sistemática de la literatura relacionada.

## 3.1. Metodologia de Búsqueda

Una revisión sistemática es una forma de identificar las investigaciones relevantes para un problema de investigación específico, mediante un método debidamente detallado de los procedimientos y decisiones que tomaron los autores para llegar a sus conclusiones [Barbara and Charters, 2007].

Las preguntas que direccionaron la revisión fueron ¿Qué técnicas para generar resúmenes abstractivos han sido usadas? y como sub pregunta ¿Han usado Abstract Meaning Representation?. Para dar respuesta a estas preguntas se buscaron estudios desde el 2006 utilizando el motor de búsqueda *Google Scholar* [1]. La revisión fue realizada en el mes de agosto del 2016 usando los términos de búsqueda '*Abstractive Summarization*', '*Generative Summarization*'. A continuación la cadena de búsqueda (*"abstractive summarization" OR "generative summarization") AND (technique* OR approach* OR algorithm* OR method**). Se excluyeron los trabajos que no están relacionados al área de Ciencias de la Computación y trabajos no relacionados a la generación de resúmenes abstractivos. En total se seleccionaron 18 artículos, donde el 2 han sido publicados en el 2016, 6 en el 2015, 3 en el 2014, y los restantes 7 del 2006 al 2013. Esto nos muestra un creciente interés y actualidad del tema de estudio.

[1] Disponible en scholar.google.com accesado en Febrero 2017

## 3.2. Descripción de Trabajos Relacionados

En los trabajos de [Carenini et al., 2006] [Carenini and Cheung, 2008] se busca una solución al problema de la evaluación positiva o negativa de entidades en textos y abordan la generación automática de resúmenes para una entidad o producto. El corpus utilizado son textos de opiniones positivas y negativas obtenidos de [Hu and Liu, 2004] [Hahn and Mani, 2000].

Primero extraen las características más importantes en una sentencia, asignan una polaridad y peso a cada caracteristicas detectada. Después se procede a representar dicha información en una estructura jerárquica. Los autores desarrollan dos aplicaciones para la generación de resúmenes, una extractiva llamada MEAD* que es genérica e independiente del dominio del problema y que utiliza distintas técnicas para asignar una puntuación a cada sentencia aunque obtuvo una pobre coherencia en los resúmenes generados.

El otro método abstractivo propuesto por los autores se relaciona con la posibilidad de generar nuevas sentencias (*Natural Language Generation*) desde la estructura jerárquica previamente identificada, utilizan técnicas para agrupar las características y luego seleccionar las más representativas, se utilizan las relaciones a nivel del discurso y se define un set de templates que nos permitirá generar las nuevas sentencias por ejemplo para la siguientes caracter´ısticas [feature: *'digital zoom'* ; orientation: -3 unimodal; user: absolute-count=7, relative-coun=.3] utilizando el template adecuado podemos obtener *"several customers hated the digital zoom"*, estas sentencias generadas tendr´ıan mayor posibilidad de expresar correctamente la información al usuario.

Posteriormente, y con el fin de evaluar el desempeño de las técnicas extractivas y abstractas, [Carenini et al., 2006] realizaron una comparativa entre ellas, utilizando un grupo de personas donde se evaluó la exhaustividad, exactitud y precisión no solo de los métodos automáticos sino también de los resúmenes realizados por las personas. En dicho trabajo se concluye que en el aspecto lingüístico-gramatical y en cuanto a la calidad del contenido, las personas son aún superiores a ambos métodos de generación de resúmenes. Más aún en un trabajo posterior sobre un corpus que incluye opiniones controversiales, se demostró que el desempeño de los métodos abstractos, que generan nuevas sentencias, es muy superior a los extractivos, aunque aún se evidencian problemas al sintetizar la información y la generación de lenguaje natural.[Carenini and Cheung, 2008]

Así mismo en el contexto de la *Text Analysis Conference (TAC 2009)* se define una línea base de cuán bien los seres humanos realizan un resumen en comparación con los métodos automáticos [Genest et al., 2013], llegando nuevamente a la conclusión de que es aún muy superior el desempeño de un resumen realizado por un ser humano y que es necesario el desarrollo de nuevas técnicas que permitan extraer pequeños segmentos de las sentencias o comprimirlas o volverlas a formular con el fin de alcanzar mejores desempeños.

Otros esfuerzos pueden ser encontrados en la compresión de sentencias que busca mediante un mecanismo recursivo de simplificaciones generar un resumen que pre- serve las partes más importantes y su coherencia gramatical [Knight and Marcu, 2000]

[Cohn and Lapata, 2009]. En el trabajo de [Filippova and Strube, 2008b] se muestra la generación de una estructura jerárquica mediante el uso de un parser sintáctico, que será a su vez condensada o reducida mediante la optimización de una función objetivo que mide el aporte de cada palabra y la probabilidad de cada relación. Para obtener un resultado óptimo se recurre a la optimización de una función objetivo mediante el uso de programación lineal. Además los autores definen restricciones estructurales, sintácticas, semánticas para dicha función objetivo.

Posteriormente en [Filippova, 2010], el autor propone realizar dicha compresión mediante el uso de un grafo dirigido de palabras, donde las relaciones entre las mismas están dadas por su adyacencia en la sentencia origen, para complementar el grafo el autor adiciona un nodo de inicio y un nodo de fin, que representan el inicio (S) y fin (E) respectivo de cada sentencia. Una vez construido dicho grafo el autor formula el problema de la compresión como un problema de encontrar una ruta más corta en el grafo que vaya del nodo inicio al nodo fin y que pase por los nodos más importantes pero que no elija un mismo nodo varias veces. Posteriormente podemos ver en [Banerjee et al., 2015] como los autores utilizan primero las técnicas de clustering, teniendo en consideración las sentencias más importantes como puntos iniciales de cada cluster, para luego aplicar las técnicas de grafos en [Filippova, 2010].

Otra técnica relacionada a la generación de resúmenes automáticos es la fusión de sentencias que busca no sólo reducir, sino también complementar contenido, para esto primero genera grupos de sentencias en base a una medida de similaridad entre las mismas y luego selecciona qué temas son los más importantes para el resumen, apoyándose en los centroides de los clusters [Barzilay and McKeown, 2005], posteriormente en [Filippova and Strube, 2008a] se alcanza una calidad gramatical superior mediante la mejora en la generación de los árboles de dependencias, utilizando la información sintáctica en las sentencias y el valor de información aportado por cada palabra [Clarke and Lapata, 2008].

En [Ganesan et al., 2010] nos muestran la posibilidad de transformar el texto en una estructura de grafos y propone que el problema de la abstracción se transforme en un problema de encontrar un apropiado camino en dicho grafo. Este enfoque ha sido utilizado también en métodos extractivos como LexRank [Erkan and Radev, 2004] y TextRank [Mihalcea and Tarau, 2004] pero en dichos trabajos los grafos no eran direccionados y en Lexrank las sentencias eran tomadas como nodos, mientras que en este trabajo las palabras son los nodos. La desambiguación es resuelta por la aplicación de un *parser Part-Of-Speech (POS)* a las sentencias, entonces cada nodo contendrá la palabra más su anotación y la información sobre la sentencia a la que pertenecen y la posición donde fueron encontradas en el texto. Basados en esta información los autores buscan encontrar rutas válidas que evitan la redundancia. Aunque en la generación del resumen utilizan un método más cercano a un modelo extractivo, pues toman las palabras y expresiones del texto original.

En [Genest and Lapalme, 2011] se menciona que un método completamente abstractivo, requiere de un modelo intermedio entre las sentencias originales y las técnicas que utilizamos para generar las nuevas sentencias en el resumen. El autor propone que este

modelo intermedio está dado por los llamados ítems de información, que en su implementación serán tripletas sujeto-verbo-objeto (SVO) que están ubicadas en un tiempo y lugar. Los ítems de información son los elementos más pequeños de información coherente en un texto o sentencia, pudiendo referenciar a una entidad o predicado, para recuperar dichos ítems se utiliza un análisis sintáctico, aunque en su trabajo el autor aún usa las expresiones originales del documento, la selección de frases que conformarán el resumen dependen de un cálculo con base en la frecuencia de términos en los ítems de información identificados. Aunque los resultados en calidad lenguaje no fueron los mejores, se demostró la importancia de tener un modelo intermedio.

En un trabajo posterior [Genest and Lapalme, 2012], los autores proponen un método abstracto donde el modelo intermedio utiliza los aspectos informativos, propuestos en la TAC 2010, para la generación de resúmenes guiados por categorías. Para identificar los distintos aspectos necesarios, el modelo utiliza técnicas de extracción de la información y extracción de eventos, logrando superar el estado del arte en cuanto a densidad de información en el resumen generado. Como se mencionó en (TAC 2010), la categorización de los resúmenes en grupos específicos a un contexto pueden mejorar el rendimiento de las técnicas para la sumarización automática.

En [Lee et al., 2005] se presenta como las ontolog´ıas pueden ser utilizadas como fuentes de conocimiento específico a un dominio de problema, que nos ayuden a contextualizar el procesamiento de texto necesario para la generación de resúmenes, En este trabajo los expertos de un dominio de noticias definieron una ontología, después mediante el procesamiento de textos de noticias se extrajeron los términos más relevantes y se asociaron mediante una medida de pertenencia a los conceptos presentes en la ontolog´ıa, trabajos recientes como el de [Mohan et al., 2016] continuan con este enfoque.

En [Ramezani and Feizi-Derakhshi, 2015] se utiliza la ontolog´ıa FarsNet que es una versión de WordNet para el lenguaje Persa, donde en primera instancia los autores extraen los palabras (*Tokens* ), extraen la base morfológica de cada palabra (*Stemming* ), extraen las sentencias, resuelven las anáforas, aplican el reconocimiento de entidades con base en FarsNet y la extraen relaciones entre ellas utilizando la información sobre la relación semántica de los términos presentes en la ontología. Con toda esta información los autores proponen generar un grafo que servirá para extraer la información más relevante, para lo cual se evalúan las siguientes medidas, grado de centralidad que está dado por el número de relaciones que posee un nodo tanto de entrada como de salida, la idea principal es que un nodo con muchas relaciones tiene una alta importancia semántica; Centralidad agregada que engloba la idea de que la importancia de un nodo también depende de si los nodos relacionados también son importantes; y por último la medida de centralidad de baricentro donde evaluamos la distancia en número de nodos que debemos recorrer desde el nodo objetivo a todos los demás nodos, por lo tanto si este valor es pequeño asumimos que el nodo objetivo no tiene gran relación con las entidades identificadas. Los resultados del trabajo muestran que la medida de centralidad agregada es la que mejor resultados obtiene. La principal dificultad en las técnicas que utilizan un enfoque con base en ontologías está en la construcción de las mismas lo cual suele ser un trabajo manual y requerir de expertos en un dominio de problema.

Ya en [Miranda-Jiménez et al., 2014] podemos observar el uso de técnicas para la extracción de información, modelos de representación intermedia y el uso de fuentes de conocimiento presentes en la web. Los autores generan resúmenes de un solo documento, utilizando una representación semántica del texto mediante grafos conceptuales ponderados, en los cuales se asocian pesos a las aristas que conectan a los nodos concepto y los nodos relación creando un flujo denominado "flujo semántico". 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.

Las relaciones conceptuales representan principalmente la semántica del texto y están basadas en los roles semánticos [Jackendoff, 1972], relaciones como agente, objeto, lugar, atributo, etc…[Sowa, 1983].

Para la generación del grafo se utiliza un parser de Stanford [De Marneffe et al., 2006] y se hace uso de información sintáctica y semántica de fuentes externas como *WordNet* [Kilgarriff and Fellbaum, 2000] y *VerbNet* [Dang et al., 2000] que rigen la coherencia estructural de los grafos.

Una vez obtenidas las estructuras gramaticales en un modelo de árbol de dependencias se generan los grafos conceptuales en base a un conjunto de reglas de transformación, si algún nodo o relación es generada de manera incorrecta los autores las corrigen de manera manual.

En la etapa de síntesis los grafos se reducen de acuerdo a un conjunto de operaciones de generalización, unión, ponderación y poda mostradas en [Montes-y Gómez et al., 2001] y [Miranda-Jiménez et al., 2013]. La evaluación del método se realizó con documentos de noticias muy breves y se superó a la línea base con un promedio del 11 %, el set de datos corresponde a *DUC 2001* y *DUC 2002*. Los autores mencionan que una de las principales limitaciones está en la generación automática de los grafos conceptuales.

En [Gerani et al., 2014], los autores nos presentan la generación de resúmenes para el problema de minería de opiniones mediante un método abstractivo basado en la utilización del análisis de las estructuras y relaciones del discurso y también proponen un método para la generación de nuevas sentencias.

Partiendo de la idea de que todo texto coherente es estructurado para que la información que contiene pueda ser interpretada, el Análisis del Discurso (*Discourse Analysis*) nos permitirá identificar dichas estructuras, es así como el primer componente de [Gerani et al., 2014] es responsable de obtener un *Discourse Tree* (DT) como representación de cada texto, dicho árbol estará conformado de nodos de unidades básicas llamadas *Elementary Discourse Unit* (EDU) que estarán unidos usando las relaciones retóricas como e.j. Elaboración, Explicación, entre otras presentes en teoría de las estructuras retóricas (*Rhetorical Structure Theory, RST* ) [Mann and Thompson, 1988]. Este árbol será modificado para que cada nodo hoja solo contenga palabras que hagan referencia a "aspectos" , con esto obtendremos lo que los autores llaman un *Aspect-based Discourse Tree (ADT)* para cada texto de opinión. Como herramienta se utiliza un parser discursivo [Joty et al., 2013]

Como segundo componente los autores agregan todos los ADT y generan un grafo que llaman Aggregated Rhetorical Relation Graph (ARRG). Este grafo es dirigido y cada nodo puede tener varias aristas relacionando dos vértices. Cada uno de estos aspectos(nodos) tienen asociada una medida de fuerza de polaridad positiva o negativa de un aspecto. Las relaciones entre dichos nodos están dadas por las relaciones retóricas identificadas además poseen también un peso de confianza en la presencia de la relación entre dos aspectos.

El tercer componente se encarga de la selección del contenido, tomando como base el grafo ARRG y ejecutando un algoritmo Weighted PageRank (WPR) [Xing and Ghorbani, 2004]. Este algoritmo toma en cuenta la importancia tanto de los links que entran y salen de un aspecto (nodo) y asigna un ranking basado también en los pesos de las relaciones entre los aspectos. En este sentido los aspectos con mayor ranking que por ende tienen más relaciones o están en relación con los nodos con mayor ranking serán promovidos, con esta información se selecciona un subgrafo como representante de los aspectos más importantes.

Posteriormente, transforman el subgrafo en una estructura de árbol Aspect Hierarchy Tree (AHT), para esto se selecciona el nodo de mayor frecuencia y el más general (a menudo el producto) y como nodos hoja los nodos los menos frecuentes que usualmente representan características específicas de un producto.

Finalmente, tomando el AHT generado de la etapa anterior, se siguen las tareas propuestas por [Reiter et al., 2000] que consisten en una etapa de Microplanning que cubre el análisis léxica y una etapa de Sentence Realization que estará encargada de generar las nuevas sentencias con base en reglas previamente definidas.

En un trabajo similar de generación multi documento basado en la identificación de roles semánticos (Semantic Role Labeling) [Khan et al., 2016], los autores proponen utilizar dicha técnica para extraer de cada sentencia los llamados *Predicate Argument Structures* (PAS) que estarán conformados a su vez de argumentos semánticos que se subdividen en argumentos core (sujeto, objeto, objeto indirecto) y argumentos adjuntos (localización, tiempo, verbo). Una vez identificadas las estructuras PAS se procede a remover las palabras no importantes y aplicar un *parser* gramatical *Part-of-Speech* (POS), los autores proponen solo considerar las palabras etiquetadas como Sustantivos (*noun*), Verbos (*verb*), Localización (*Location*) y tiempo (*Time*). Una vez obtenidas estos atributos se comparara las sentencias utilizando la función de similaridad de Jiang [Jiang and Conrath, 1997] que hace uso de la WordNet para calcular la cantidad de información que los términos poseen y que los autores mencionan es la más cercana al juicio humano.

$$Jiang_{dist}(C1, C2) = IC(C1) + IC(C2) - 2 \times IC(Iso(C1, C2)) \quad (3.1)$$

Donde el contenido de información (IC) de cualquier concepto es estimado mediante el cálculo de la probabilidad de ocurrencia de un concepto en un corpus de texto:

$$IC(C) = -\log P(C) \quad (3.2)$$

Donde la probabilidad de que el concepto C ocurra es igual a:

$$P(C) = \frac{Freq(C)}{N} \quad (3.3)$$

Y donde la frecuencia del concepto C es la ocurrencia de C en una taxonom´ıa como la WordNet y N es el número total de sustantivos.

Basados en esta medida de similaridad los autores utilizan el algoritmo de clusterización *Agglomerative Hierarchical Clustering* (HAC) [Murtagh and Contreras, 2011] para agrupar las estructuras *Predicate Argument Structures* (PAS) identificadas. Posteriormente, para elegir la sentencia más representativa de cada cluster y que por ende debe ser utilizada en el resumen, se evalúa la utilización de la función similitud de Jiang, la información de aparición dentro del documento, el número de sustantivos y verbos que posee y por último la información dada por la frecuencia de los términos (TF-IDF). Para asignar la importancia de cada atributo los autores proponen la solución como un problema de optimización que utiliza la medida ROUGE-1 como una forma de medir la bondad de una solución y una metaheurística de optimización que ya ha sido utilizada para escenarios similares como es el algoritmo *Particle Swarm Optimization* (PSO) [Shi et al., 2001] y que ha sido utilizado en varias tareas relacionadas a la sumarización de textos [Van der Merwe and Engelbrecht, 2003] [Ziegler and Skubacz, 2007].

Una vez identificada la combinación óptima se procede a utilizar dicha función para obtener las estructuras PAS que tengan mejor calificación como parte del resumen. Finalmente se utiliza la herramienta *SimpleNLG* [Gatt and Reiter, 2009] para la generación de lenguaje natural en base a reglas que usan la información semántica extraída mediante la identificación de roles semánticos.

Como fue mencionado en el capítulo anterior, en 2014 múltiples autores proponen una representación común útil para los procesos de abstracción [Knight et al., 2014] la cual llamaron *Abstract Meaning Representation* (AMR), en esta representación se puede describir la información semántica de las sentencias a manera de grafo, con el objetivo de proponer un simple y único modelo de representacion semántica.

Con base en AMR [Liu et al., 2015] presentan un *framework* para la generación de resúmenes abstractivos para un solo documento. Los autores toman una a una cada sentencia del documento y con la ayuda del parser JAMR [Flanigan et al., 2014] se genera un grafo AMR. Posteriormente en la fase de construcción se fusionan los grafos en base a los conceptos que ocurren en dichos grafos. De esta manera tendremos un grafo único para el documento objetivo que reducirá su redundancia de conceptos, además porque la repetición de un concepto en el texto indica importancia se asignará el valor de frecuencia a cada concepto para ser utilizado en la síntesis posterior. Los autores aún no consideran el problema de la resolución de referencias de conceptos por ejemplo que ”Barack Obama” es igual a ”Obama”y que ”Say-01”es igual a ”Report-01”, aunque lo colocan como trabajos

futuros. Dado que por la fusión dos conceptos pueden ahora estar unidos por varias aristas, los autores agrupan todas en una solo arista que llevará como título las dos más comunes relaciones. Adicionalmente y para asegurar que el grafo es conectado se creará un nuevo nodo raíz y se conectará con cada concepto que fuera originalmente un nodo raíz a nivel de sentencia (Figura 3.1).

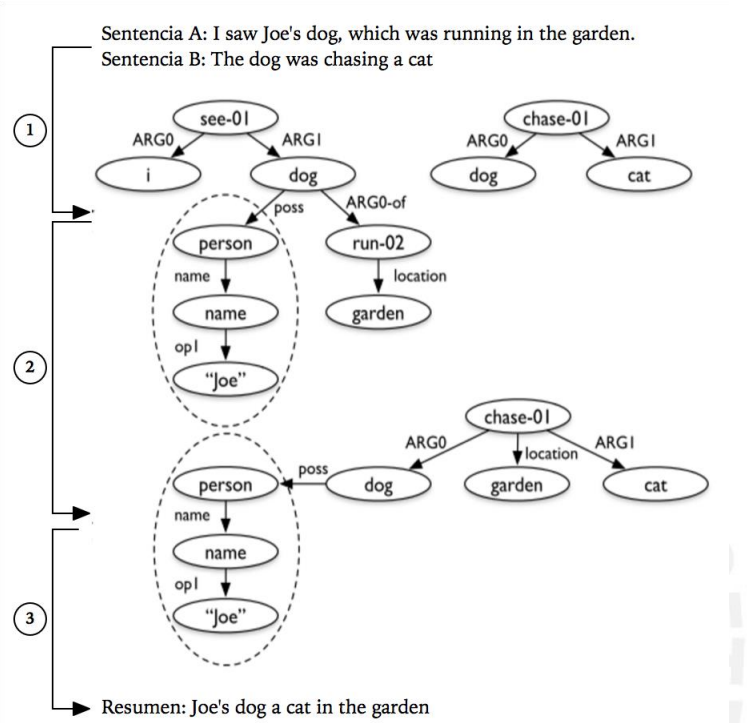


Figura 3.1: Proceso de sintesis entre varios grafos semánticos por sentencia. Extraido y adaptado de [Liu et al., 2015]

Una vez que se tiene un único grafo que representa todo el documento se procederá a la predicción del subgrafo resumen que debe incluir la información más importante sin alterar su significado. Para esto los autores formulan la selección de este subgrafo como un problema de programación lineal. Para la generación del lenguaje natural se utilizará el subgrafo identificado y los conceptos presentes en él, por lo que se buscará los términos más frecuentes alineados a dichos conceptos para colocarlos en el resumen generado sin un orden específico por lo que la única medida que se podrá aplicar es la que utiliza la existencia o no de un término como ROUGE-1.

## 3.3. Conclusiones

Con base en los trabajos analizados podemos concluir que los métodos abstractivos para la generación de resúmenes alcanzan mejores resultados tanto lingüísticos como semánticos en los textos generados, así como pueden alcanzar mejor desempeño en la síntesis de la información.

Los métodos abstractivos también requieren de capacidades de análisis profundo sobre el texto, pues a diferencia de los métodos extractivos, requieren entender la semántica de las expresiones por lo que requerirá del uso de técnicas para el procesamiento del lenguaje natural (PLN) como extracción de información, roles semánticos, extracción de eventos, aspectos informativos, análisis del discurso entre otros. Asimismo los avances en estos campos repercuten positivamente en generación automática de resúmenes.

También se ha podido identificar que los métodos abstractivos que tienen un mejor desempeño recurren a un modelo intermedio entre la información textual original y el texto generado, que usualmente es construido mediante alguna o muchas técnicas combinadas de PLN. Dicho modelo intermedio permite a los autores aplicar distintas técnicas para mejorar la capacidad de compresión, fusión o síntesis de la información presente en el texto original. Usualmente dicho modelo ha sido de carácter jerárquico o basado en grafos, siendo AMR un modelo semántico de los últimos propuestos y que busca desarrollar un ecosistema similar al de los modelos sintácticos.

Asimismo, el integrar la información discursiva de un método RST complementa la información semántica obtenida a nivel de sentencia y en ambos casos es representada en una forma de grafo, lo que nos alienta a utilizar algoritmos de grafos como PageRank o Hits para encontrar métodos eficientes para navegar e identificar los elementos más importantes.

Una vez que hemos alcanzado una s´ıntesis en el modelo intermedio, todo modelo abstractivo debe generar nuevas sentencias en lenguaje natural que permitan mostrar la información más relevante identificada de una manera correcta y coherente a los usuarios.

Estas tres tareas, la generación de un modelo intermedio desde el texto, la síntesis de dicho modelo intermedio y la generación de lenguaje natural son actividades complejas que aún se encuentran lejos de tener una solución apropiada a las necesidades de las personas.

# Capítulo 4

# Generación de Resúmenes Abstractivos

Nuestro objetivo es immplementar un método de generación automática de resúmenes con un enfoque abstractivo integrando información semántica y discursiva, esto requerirá de una comprensión de los conceptos presentes en un texto, utilizamos técnicas de análisis profundo con el fin de identificar dichos conceptos y las relaciones entre ellos, presentes tanto a nivel sentencial como a nivel del documento.

En la sección 4.1 se presenta el corpus utilizado, en la sección 4.2 se explica el método propuesto, en la sección 4.3 se mostrarán los resultados de aplicar el método sobre el corpus el cual ha sido dividido en dos grupos de documentos conocidos como datos de entrenamiento y datos de prueba. Primero en la data de entrenamiento, buscamos calibrar e identificar los mejores parámetros para nuestro método y después aplicamos dicha configuración al segundo grupo de documentos de prueba para validar los resultados obtenidos.

## 4.1. Corpus

En los experimentos utilizamos el Corpus provisto por la *Document Understanding Conference* (DUC) que contiene artículos y resúmenes abstractivos escritos por personas. Estos resúmenes son de aproximadamente 100 palabras que corresponden en promedio a una tasa de compresión del 20 %. Dicho corpus ha sido ampliamente utilizado para la tarea de generación de resúmenes extractivos para un documento [1].

En los experimentos hemos seleccionado 275 documentos de entrenamiento para la construcción de nuestro modelo. Adicionalmente 307 documentos, distintos a los anteriores, para la validación del mismo. La metrica utilizada fue ROUGE. [2]. Los documentos

[1] Disponible en http://duc.nist.gov/data.html accesado en Febrero 2017

[2] Disponible en http://www.isi.edu/~cyl/ROUGE/ accesado en Febrero de 2017

del corpus han sido extra´ıdos de distintas fuentes de noticias que contemplan distintos formatos y representaciones, por este motivo el primer trabajo realizado fue remover los marcadores web y las secciones usualmente XML referentes a la representación web, para el presente trabajo no se ha considerado las secciones de t´ıtulo, ni las palabras clave contenidas en muchos de estos art´ıculos. Este proceso fue realizado de manera manual con el fin de evitar errores en este nivel.

## 4.2. Descripción del Método Propuesto

El método propuesto sigue la arquitectura propuesta por [Mani, 2001] que consta de tres etapas, en la etapa de análisis, los textos de entrada son interpretados y representados en un formato computacional, en la etapa de transformación dicha representación es procesada para identificar y seleccionar el contenido más relevante y como resultado se obtiene una representación computacional condensada de los textos. En la etapa de s´ıntesis es generado un texto en lenguaje natural. En la figura 4.1. podemos apreciar una vista en general de las etapas y técnicas utilizadas.

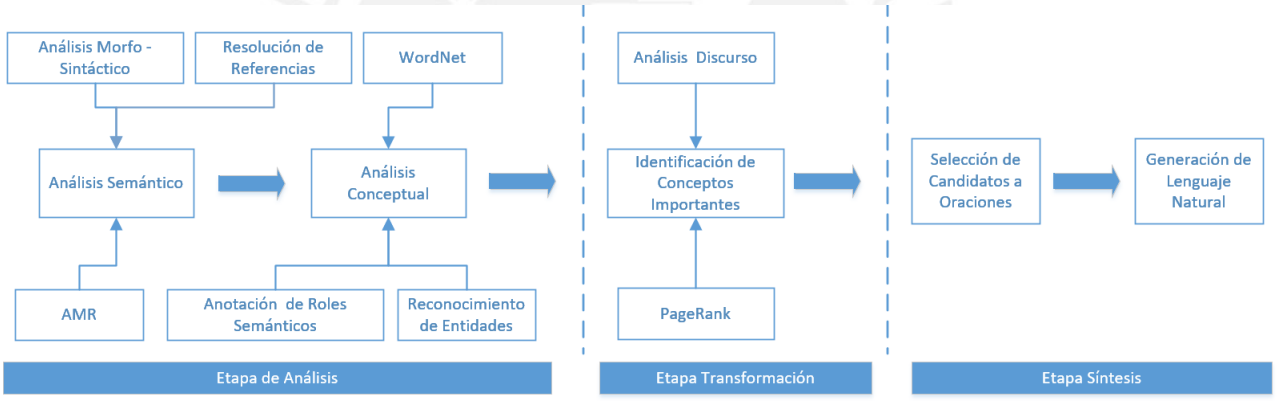


Figura 4.1: Diagrama del Método Propuesto

### 4.2.1. Fase de Análisis

#### Análisis Morfo-Sintáctico

El análisis Morfo-Sintáctico nos permite asignar una estructura sintáctica dada por una gramática a una sentencia [Jurafsky and Martin, 2009]. En el trabajo este análisis es base para el funcionamiento del anotador semántico y anotador discursivo.

Fue utilizada la implementación de la Universidad de *Stanford* [3] [Manning et al., 2014], que incluye además de la generación un árbol sintáctico, la anotación part-of-speech, delimitación de sentencias y resolución de referencias entre otras tareas.

[3]Disponible en http://stanfordnlp.github.io/CoreNLP/ accesado en Febrero 2017

Por ejemplo, para la sentencia a continuación podemos obtener una visualización en la pagina web del proyecto [4], que incluye la anotación *Part-of-Speech* en la Figura 4.2.

*"The United Nations Food and Agriculture organization said hot and dry conditions in January and February were expected to reduce the total cereal harvest in 11 southern African countries to 16m tonnes, 25 per cent down on the average.*

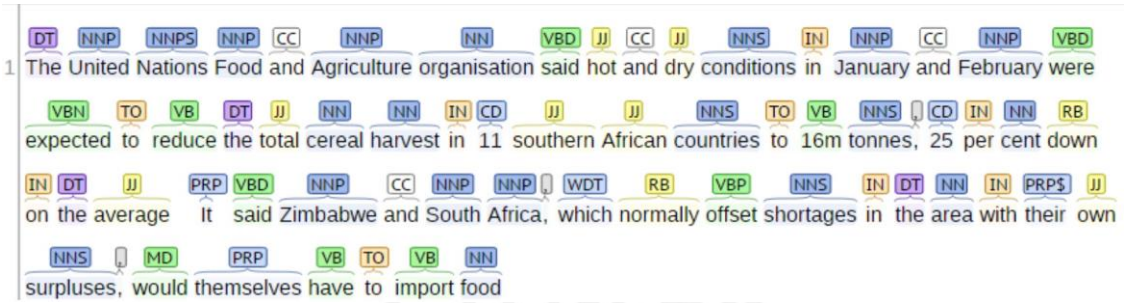


Figura 4.2: Ejemplo de *Part-of-Speech* Stanford Online Parser

La delimitación de sentencias es una tarea necesaria pues los documentos contenidos en el corpus no tienen una especificación al respecto. Utilizaremos la capacidad del anotador sintáctico para identificar las sentencias en los documentos, esto se dará mediante el uso de los signos de puntuación que indican nuevas sentencias en el idioma inglés, este paso es requisito para el anotador semántico y discursivo elegidos en el trabajo.

**Resolución de Referencias**

Dado el enfoque abstractivo, podemos modificar el texto original utilizando las técnicas de resolución de referencias para expandir el mismo y así incrementar la cantidad de información en cada sentencia, este proceso ayudara posteriormente al análisis conceptual como también lo mencionan en [Liu et al., 2015] aunque no llegan implementarlo.

En los experimentos y por la complejidad de la tarea solo se consideró explotar las referencias de pronombres hacia entidades reconocidas con las anotaciones (***NN*** , ***NNS*** , ***NNP***, ***NNPS*** ) en el anotador *Part-of-Speech*.

Por ejemplo, para las siguientes sentencias se puede apreciar cómo este proceso incrementa la información contenida al reemplazar el pronombre It por el texto completo de la organización que referencia.

*"**The United Nations Food and Agriculture** organization said hot and dry conditions in January and February were expected to reduce the total cereal harvest in 11 southern African countries to 16m tonnes, 25 per cent down on the average.*

[ ***It (PRP) | The United Nations Food and Agriculture (NNP)*** ] *said*

[4]Disponible en http://nlp.stanford.edu:8080/corenlp/process accesado en Febrero 2017

*Zimbabwe and South Africa , which normally offset shortages in the area with their own surpluses , would themselves have to import food"*

**Análisis Semántico**

El análisis semántico busca encontrar significado de las palabras más allá de su rol sintáctico. En este punto se decidió utilizar *Abstract Meaning Representation* [Banarescu et al., 2013].

Al igual que en [Liu et al., 2015] recurrimos a un parser de AMR, entre otras razones porque AMR engloba no sólo el análisis semántico desde una perspectiva sintáctica, sino que además hace uso de recursos de conocimiento como es *Propbank*, que como pudimos apreciar en trabajos como [Ramezani and Feizi-Derakhshi, 2015] [Mohan et al., 2016], demuestran ser de mucha utilidad para los resúmenes abstractivos.

AMR tiene una característica llamada de reentrada que resulta de sumo interés para los resúmenes, pues nos permite fusionar expresiones, como podemos apreciar en el siguiente ejemplo de grafo AMR para la expresión *"The dog wants to eat the bone"* que en la representación AMR se escribiría de la siguiente forma:

```
(want-01 :
├─ ARG0 (d / dog)
└─ :ARG1 (e / eat-01
   ├─ :ARG0 d
   └─ :ARG1 (b / bone)))
```

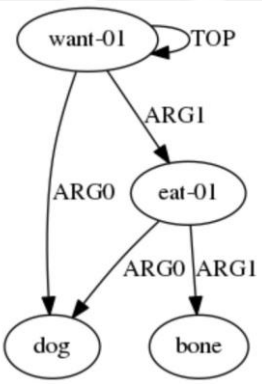


Figura 4.3: Representación AMR en forma de grafo de la oración *"The dog wants to eat the bone"*

Y podemos apreciar su representación gráfica en la figura 4.3, donde se puede apreciar con mayor claridad como la variable *d* que representa el concepto *dog* posee dos aristas debido a que participa en varios elementos de la sentencia. Para la generación de los diagramas de grafos AMR utilizamos el proyecto *AMR Inspector for Cross-language Alignments* (AMRICA) presente en el enlace a internet[5].

[5]Disponible en https://github.com/nsaphra/AMRICA accesado en Febrero 2017

AMR es un tópico actual de interés, como lo demuestra al ser la tarea numero ocho en la edición 2016 del *International Workshop on Semantic Evaluation 2016* [6], donde se llegaron a proponer hasta 6 nuevos *parsers*, uno de estos es CAMR parser [Wang et al., 2016], el cual llega a alcanzar una media F1 de 66.5 % sobre el corpus de la competencia. Para el trabajo solo se tuvo acceso a la versión entrenada sobre el corpus original de AMR (LDC2013E117) [Banarescu et al., 2013], donde los autores mencionan un F1 del 61 % [Wang et al., 2015].

Se aplicó el anotador CAMR al corpus. A diferencia del trabajo en [Liu et al., 2015], en el presente trabajo se utilizó un corpus que no tiene una versión de AMR generada manualmente, entonces también se tuvó un aporte empírico en el hecho de aplicar este anotador al corpus de la DUC, que hasta donde se investigó no ha sido aplicado anteriormente.

Por ejemplo para la sentencia:

*"**The United Nations Food and Agriculture** organization said hot and dry conditions in January and February were expected to reduce the total cereal harvest in 11 southern African countries to 16m tonnes, 25 per cent down on the average.*

Utilizamos el parser AMR para obtener su representación semántica:

```
(x8 / say-01
   :ARG0 (x4 / food
      :null_edge (x2 / null_tag)
      :null_edge (x3 / null_tag)
      :null_edge (x5 / null_tag
         :op2 (x7 / organization
            :null_edge (x6 / null_tag))))
   :ARG1 (x18 / expect-01
      :ARG0 (x10 / and
         :op1 (x9 / hot)
         :op3 (x11 / dry-02)
         :op2 (x12 / condition)
         :location (x15 / and
            :op1 (x14 / date-entity)
            :op2 (x16 / date-entity)))
      :ARG1 (x20 / reduce-01
         :ARG1 (x24 / harvest-01
            :ARG2-of (x22 / total-01)
            :ARG1 (x23 / cereal)
            :location (x29 / country
               :quant 11
               :mod (x27 / south)
               :name (x28 / name
                  :op1 "African")))
         :ARG4 (xap0 / multiple
            :op1 (x32 / mass-quantity
               :unit (t / tonne)
               :null_edge (x31 / null_tag)
```

[6] Disponible en http://alt.qcri.org/semeval2016/ accesado en Febrero 2017

```
:ARG3 (x36 / monetary-quantity
   :unit (c / cent)
   :mod (x37 / down
      :prep-on (x40 / average))))))))
```

Ahora, esta representación también puede ser visualizada en un grafo, para este propósito utilizaremos el proyecto AMRICA[7] el cual genera un grafo a partir del formato AMR. (Figura 4.4)

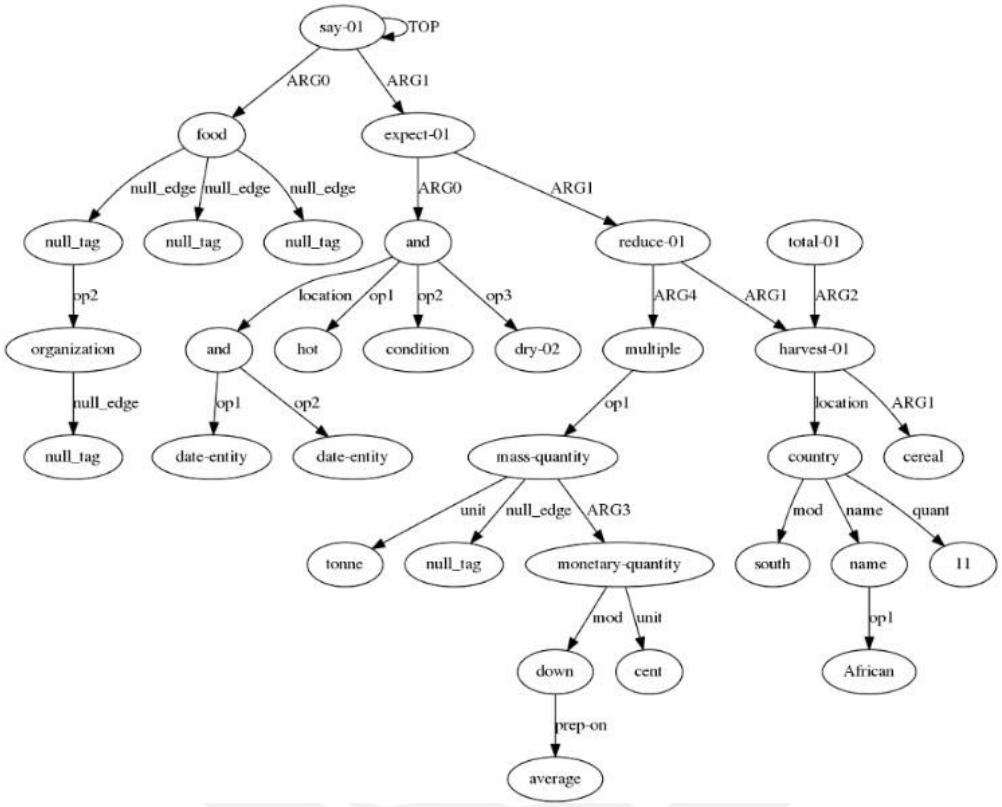


Figura 4.4: Visualización gráfica del resultado de aplicar el anotador CAMR en una sentencia del corpus

### Fusión de grafos semánticos en un grafo conceptual por documento

En este punto se ha creado un grafo con información semántica por sentencia en un documento, que fue fusionado para obtener un único modelo conceptual por documento apoyándonos en los roles semánticos y el reconocimiento de entidades.

### Análisis Conceptual

En la misma línea del trabajo de [Miranda-Jiménez et al., 2013], se creyó necesario un modelo que esté basado en el análisis sintáctico y semántico, pero que lleve el análisis

[7]Disponible en https://github.com/nsaphra/AMRICA accesado en Febrero 2017

a un nivel superior el cual llamaremos igualmente Conceptual, debido a que necesitamos abstraer los conceptos para poder fusionarlos y generar nuevas sentencias.

A diferencia del trabajo de [Miranda-Jiménez et al., 2014], donde utilizan (VerbNet [Kipper et al., 2000]) para, mediante un trabajo en parte manual, alinear los conceptos y las relaciones semánticas entre ellos. Nosotros generamos los gráfos conceptuales automáticamente en base a la salida AMR que ya está alineada a un recurso lingüístico como es Propbank.

### Roles Semánticos

En AMR las relaciones entre conceptos tienen un identificador Arg0...Arg5 que suelen estar asociados a un rol semántico como es agente, paciente, etc.. En el trabajo hemos decidido expandir este conocimiento y utilizar la relación que existe entre Prop- bank y Verbnet para identificar de manera más exacta las relaciones semánticas y roles semánticos de cada concepto.

De esta forma si revisamos los *frames* en Propbank[8], por ejemplo para el verbo offset-01 encontraremos que para el Arg1 el rol semántico asignado es Goal y no Patient. Por esta razón en nuestro trabajo y siempre que exista la información de un frame en Propbank colocamos el rol semántico asociado de la VerbNet, ahora si esta información no existiese se utiliza la convención por defecto de AMR descrita en la tabla a continuación.

| Relación AMR | Rol Semántico |
|---|---|
| Arg0 | agent |
| Arg1 | patient |
| Arg2 | goal |
| Arg3 | start |
| Arg4 | end |

Cuadro 4.1: Relación por defecto entre AMR y roles semánticos

Nuestro grafo AMR lucirá ahora los roles semánticos entre los conceptos.

### Reconocimiento de Entidades

En la experimentación también se hizo uso de la capacidad de AMR de reconocer entidades agrupadas en 8 tipos principales (*Person, Organization, Location, Facility, Event, Product, Publication, Natural object, Other*) que pueden a su vez contener varios subtipos como es el caso de la categoría *Organization* que puede contener a *company, government, military, criminal organization,* entre otras. Aunque en el futuro se espera que estas entidades tengan una referencia a recursos externo, como puede ser wikipedia,

[8]Disponible en https://github.com/propbank/propbank-frames accesado en Febrero 2017

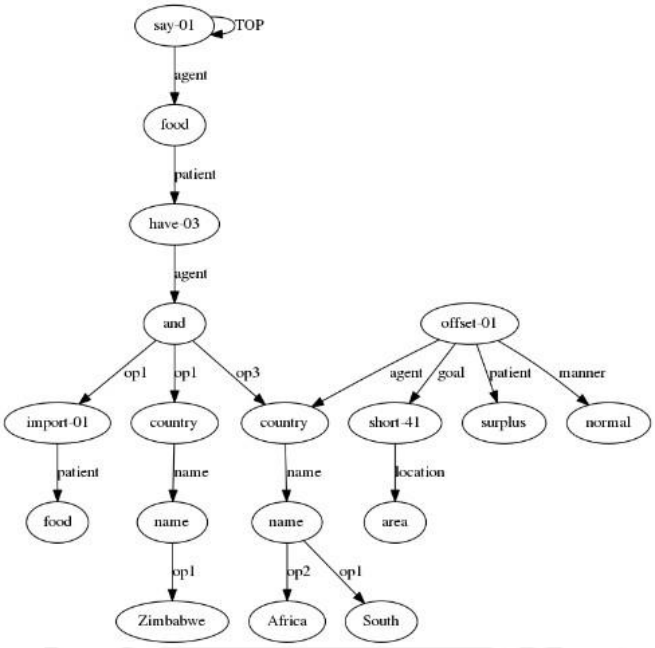


Figura 4.5: Grafo conceptual anotado con roles semánticos como aristas entre los nodos

de manera que sirvan para expandir la información sobre su significado, esta capacidad aún no esta presente en el *parser* y no ha sido explorada por el presente trabajo.

El formato AMR representa las entidades como un sub grafo cuyo nodo ra´ız es un nodo con el identificador Name, en nuestro trabajo estos subgrupos fueron fusionados en un solo nodo.

Al igual que en [Liu et al., 2015] fusionamos los nodos por los conceptos similares entre las sentencias pero utilizamos la información extraída de los roles semánticos de manera que solo fusionamos los conceptos que son *Agents*, *Patients*, *Goals* y *Themes* evitando la fusión de los grafos por verbos debido a que esto genera confusión y ambiguedad en el grafo.

### Fusion de Conceptos con WordNet

Adicionamos al proceso de fusion tambien la idea de utilizar un recurso lingüístico como es la Wordnet[9] [Miranda-Jiménez et al., 2014] para fusionar conceptos que esten relacionados a pesar de no tener la misma representacion textual.

La WordNet es una base de datos léxica del idioma inglés que contiene sustantivos, verbos, adjetivos y adverbios; organizada jerárquicamente en grupos de sinónimos llamados *synsets* y está enlazada mediante relaciones semánticas de hiperonimia, hiponimia meronimia, antonimia y más.

[9]Disponible en https://github.com/wordnet/wordnet accesado en Febrero 2017

Entonces para fusionar nuestros grafos utilizaremos la medida de similaridad propuesta por [Wu and Palmer, 1994] que hace uso de la medida profundidad de dos *SynSets* en la *Wordnet* con respecto a un concepto común, descrita por la siguiente fórmula:

$$score = 2 * depth(lcs)/(depth(s1) + depth(s2))$$

Donde depth(LCS) es la medida de profundidad con respecto al nodo raíz de la Wordnet para el término común entre S1 y S2. En caso de que múltiples *SynSets* sean compartidos se tomará el más común.

En el trabajo la medida de similitud debio ser mayor al 90 % para fusionar dos conceptos.

En la Figura (4.6) se presenta un ejemplo del método de fusión utilizando la información de los roles semánticos donde podemos apreciar que se pueden fusionar conceptos que han sido identificados como *Agent* o *Patient* en sentencias distintas, entidades reconocidas como países o personas y WordNet para fusionar conceptos similares como *Past* y *History*.

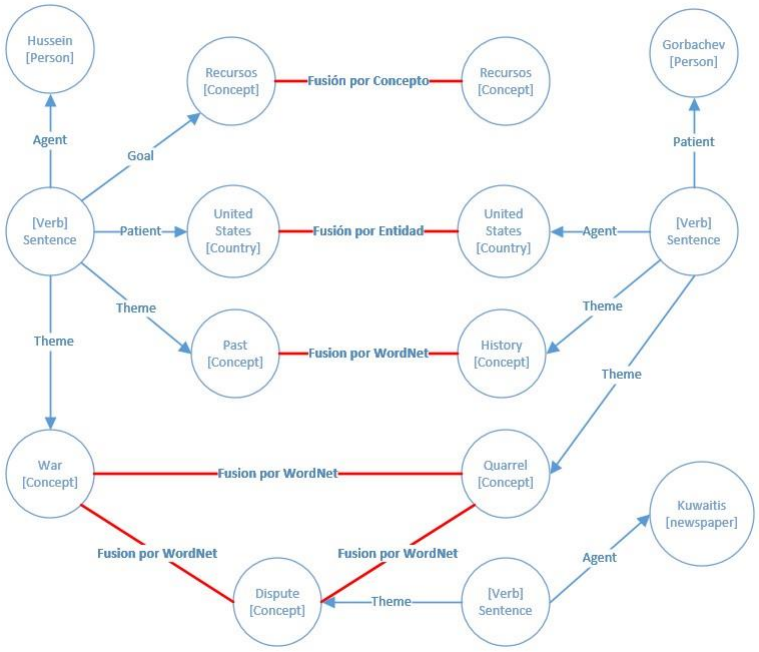


Figura 4.6: Fusión de grafos semánticos

A diferencia de [Liu et al., 2015], en este punto no tenemos garantía de tener un único grafo conceptual, debido a que solo estamos fusionando conceptos que estén relacionados y no hemos generado una relación ficticia entre los grafos por sentencia. Aunque la utilización de WordNet representa un método de fusión más apropiado [Miranda-Jiménez et al., 2014].

## 4.2.2. Fase de Transformación

### Análisis del Discurso

Como se mencionó el estudio de la coherencia y semántica entre las sentencias de un texto es objetivo del Análisis del discurso. En particular, RST (*Rhetorical Structure Theory* ) ha dado resultados positivos en la generación de resúmenes tanto para un documento [O' Donnell, 1997] como para múltiples documentos [Uzêda et al., 2010], aunque mayormente ha sido utilizado con un enfoque extractivo debido a que su énfasis está a nivel de sentencia y la forma en que estas se relacionan.

El establecer una medida de importancia para cada EDU en un árbol RST ha tenido distintos enfoques como es la utilización de la información nuclear o de satélite [Ono et al., 1994] [Marcu et al., 2000] y en un enfoque más orientado a la importancia de cada relación [O' Donnell, 1997] donde recorremos el árbol retórico de la raíz a las hojas y cuando encontremos un nodo satélite multiplicaremos el valor del nodo padre por un factor asociado al tipo de relación. La asignación de dichos pesos es un trabajo manual y empírico, pero contamos con valores óptimos propuestos en [de Uzêda et al., 2007]. En el presente trabajo se escogió el método de O'Donnell debido a que se optó por técnicas que hagan uso de la toda información semántica, como es el caso de este método que asigna un valor de importancia de acuerdo al tipo de relación discursiva, podemos ver un ejemplo de la asignación de puntajes en (Figura 4.7).

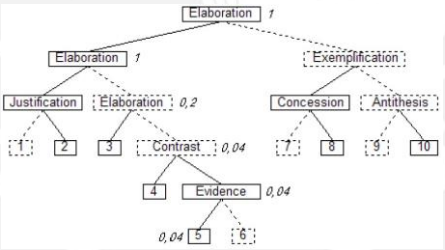


Figura 4.7: Recorrido de puntuación según O'Donnell

En [Cardoso, 2014] se prueba la relevancia de este análisis en la generación automática de resúmenes multidocumento mayormente bajo un enfoque extractivo. También podemos encontrar esfuerzos por aplicar este método a enfoques abstractivos como se presentó en el capítulo anterior en [Gerani et al., 2014].

En este punto, contamos con una clasificación de importancia de cada EDU dada por la aplicación del método de O'Donnell como podemos apreciar en (Figura 4.8) donde aparece entre paréntesis al costado del identificador del EDU.

Entonces asignamos a cada palabra de cada sentencia el valor del peso del EDU que le corresponde. Ahora bien una palabra puede estar repetida en distintos EDU de una sentencia, en este caso se tomo el peso de mayor valor.

Entonces, cada concepto de nuestro modelo ha sido fusionado mediante la información de WordNet y cada concepto tiene un valor de importancia en el documento dado

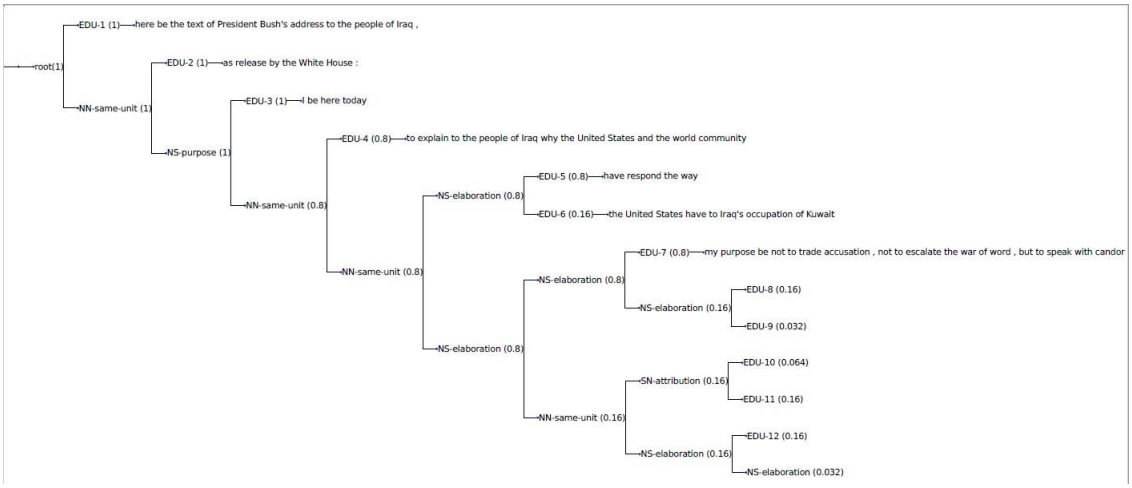


Figura 4.8: Recorrido de puntuación utilizando O'Donnell en un documento del corpus

por el uso del método de O'Donnell que hace uso del análisis Discursivo. Lo que nos da un grafo conceptual con pesos de importancia discursiva en sus nodos.

Para poder identificar los nodos más importantes considerando la cantidad de relaciones y los pesos de los mismos recurriremos al método de PageRank [Brin and Page, 1998] el cual genera un ranking de importancia de los nodos de acuerdo a la importancia estructural de los mismos, no requiriendo que exista un camino entre ellos pues establece un salto aleatorio manejado por una probabilidad de utilizar este salto aleatorio desde un nodo específico.

Entonces diremos que para un grafo G con N vertices $V_1 \ldots V_N$ y di como el grado de aristas que salen del nodo $i$; diremos que M es una matriz probabilidades de transición de NxN, donde $M_{ij} = 1/d_i$ si un camino existe desde el nodo $i$ al $j$ de otra manera el valor será 0.

Entonces para calcular el vector PageRank se deberá resolver la siguiente ecuación:

$$P_r = cMP_r + (1 - c)v \tag{4.1}$$

En la ecuación v es un vector de $Nx1$ y sus elementos tienen el valor de 1/N y c es el llamado *damping* factor, un valor escalar entre 0 y 1. El primer término de la ecuación describe la información sobre las relaciones entre los nodos, el segundo define la probabilidad de moverse aleatoriamente desde un nodo a otro sin tener ninguna relación entre uno y otro. El valor c indica el grado de importancia que le daremos al segundo factor.

En el modelo tradicional de PageRank el vector v es un vector normalizado cuyos valores son 1/N lo que asigna igual probabilidad para moverse desde un nodo a otro. Sin embargo como lo indica [Haveliwala, 2002], el vector v no tiene porque ser uniforme y puede asignar mejores probabilidades a ciertos tipos de nodos modificando la tendencia del algoritmo ha preferir ciertos nodos. Por lo tanto, si asignamos una alta probabilidad a un

nodo entonces dicho nodo tendrá un ranking más alto así como su vecindad. Utilizaremos esta capacidad para incluir la información discursiva recolectada por la aplicación del método de O'Donell, por lo que los nodos con un valor alto en el análisis discursivo transmitirán dicha importancia a su vecindad.

PageRank se presenta como un método muy útil pues a diferencia del método usado en [Miranda-Jiménez et al., 2014] de HITS [Kleinberg, 1999], PageRank nos permite incluir la información discursiva recolectada.

Entonces después de aplicar PageRank nuestro modelo conceptual estará dado por un grafo que contiene nodos que representan conceptos en el documento y relaciones entre ellos que son dadas por la información semántica contenida en Propbank y VerbNet. Estos nodos tienen un grado de importancia dado por el algoritmo PageRank que ha explotado no solo las relaciones semánticas sino también la información discursiva.

### 4.2.3. Fase de Síntesis

Una vez construido nuestro grafo necesitamos recorrerlo para extraer de él información sobre las acciones que se produjeron (*Verbs*), los agentes (*Agents*), hacia quien se realizaron estas acciones (*Patients*), sobre que tema fueron (*Themes*) y con qué objetivo (*Goals*).

Entonces nuestro algoritmo se posicionará en los nodos verbo y a partir de él intentará extraer el o los nodos que estén unidos a él con la relación semántica de Agent y así para los roles semánticos de *Patient*, *Theme*, *Goal*. Una vez identificado este subgrafo será la base de una nueva sentencia cuya importancia estará dada por:

$$Importancia\ de\ la\ sentencia = Sumatoria(P(Agents)+P(Verbs)+P(Themes)+P(Goals))$$

Con base en nuestros experimentos se obtuvo una ganancia significativa del 1 % si sólo se consideran los subgrafos que contengan como mínimo un nodo *Agent* y por lo menos un nodo *Patient*, *Theme* o *Goal*.

| **Rol Semántico** | **Concepto** | **Valor** |
|---|---|---|
| **Agent** | And | 0.066985066741923949 |
| **Agent** | Peaple | 0.029765304343543915 |
| **Agent** | United States | 0.010885067942962847 |
| **Agent** | World Community | 0.025989257063427702 |
| **Verb** | Respond | 0.039531389611232572 |
| **Patient** | Way | 0.026002720054770188 |
| | Total | 0.19915880575786116 |

Cuadro 4.2: Valores dados por PageRank para una sentencia extra´ıda desde el grafo conceptual

En la tabla 4.2, se muestran los valores obtenidos para cada concepto elegido por nuestro método de navegación, la suma total de estos valores es el peso total de una expresión.

Si ordenamos las sentencias de mayor a menor importancia podemos construir un resumen que contemple las expresiones mas importantes.

**Generación del lenguaje natural**

Ahora bien en este punto hemos logrado extraer información sintáctica, semántica y llevarla a un modelo conceptual, pero es necesario generar nuevas sentencias que puedan presentar la información en una forma similar a la que un ser humano la utiliza. Esta tarea es abordada por la generación de lenguaje natural o producción de lenguaje que utiliza un una forma de conocimiento que ha sido inferida en este caso por un computador.

Este problema es un tópico de interés para el procesamiento del lenguaje natural y actualmente está en desarrollo e investigación y es una tarea compleja pero clave para materializar las ventajas de la técnica de resúmenes abstractivos. Usualmente este proceso necesitará de una fase para determinar los contenidos, una fase de planeamiento donde se decidirá qué léxico, como las sentencias serán combinadas y el uso de referencias entre sentencias, y por último una fase de realización del texto donde se generan las sentencias dependiendo del lenguaje objetivo.[Reiter et al., 2000]

Entonces, los contenidos estan dados por las sentencias ordenadas por su importancia en la evaluación conceptual. Para apoyo a la fase de planeamiento contamos con una clasificación por roles semánticos de los verbos objetivo como son *Agent*, *Patient*, *Goal*, *Theme* obtenidos por AMR.

Para la fase de realización utilizaremos el proyecto SimpleNLG [Gatt and Reiter, 2009][10] que es un motor de generación de lenguaje natural para el idioma Inglés.

Este proyecto nos permite definir las partes de una sentencia y por ejemplo, el tiempo en cual deseamos que se genera la misma y la sentencia generada será coherente en cuanto a tiempo y persona.

Por ejemplo el siguiente código:

```
SPhraseSpec p = nlgFactory.createClause();
p.setSubject('Mary');
p.setVerb('chase');
p.setObject('the monkey');
```

[10]Disponible en https://github.com/simplenlg/simplenlg accesado en Febrero 2017

Genera el texto "*Mary chases the monkey*" apropiado en tiempo presente y tercera persona. Como se puede apreciar, hay una buena relación entre los elementos requeridos por el proyecto y nuestra abstracción desde el grafo conceptual, más aún la generación del texto tiene una conjugación correcta a pesar de que nosotros cambiemos el tiempo de la narración como es común al momento de generar resúmenes dado que la narración suele ser en tiempo pasado.

Más aún en la generación de lenguaje natural buscamos emular la capacidad de un ser humano para generar y agrupar varias sentencias en un párrafo que tenga contenga una síntesis de dicha información, por esto también utilizaremos la capacidad del proyecto SimpleNLG para generar texto desde múltiples sentencias, por ejemplo:

```
SPhraseSpec s1 = nlgFactory.createClause('my cat', 'like', 'fish'); SPhraseSpec s2 = nlgFactory.createClause('my dog', 'like', 'big bones'); SPhraseSpec s3 = nlgFactory.createClause('my horse', 'like', 'grass');
```

Buscamos generar una sola sentencia que englobe estas ideas de la misma forma que lo haría un ser humano para lo cual SimpleNLG nos ofrece la posibilidad de coordinar expresiones de varios sujetos, objetos y objetos indirectos pudiendo obtener por ejemplo la siguiente expresión.

*My cat likes fish, my dog likes big bones and my horse likes grass.*

En el experimento esta posibilidad se vio reflejada, por ejemplo, en la generación de la siguiente expresión:

*"We agreed with objective of possible international peaceful order devour large state and Gorbachev neighbor".*

## 4.3. Experimentación

### 4.3.1. Corpus de Entrenamiento

#### Técnica Extractiva

Las técnicas extractivas han probado tener resultados importantes en la identificación de los componentes relevantes en un texto y con esta información generar resúmenes. Una de estas técnicas, que también forma parte de nuestro modelo, es la propuesta por [O' Donnell, 1997].

Nuestro primer experimento consistió en utilizar un parser RST para extraer dicha representación del corpus de entrenamiento, el parser elegido es DPLP [Ji and Eisenstein, 2014] que obtiene un 71.3 % de éxito en detectar las sentencias nucleares y un 61.63 % en de- tectar las relaciones entre las sentencias evaluado en el corpus RST Discourse TreeBank

[Carlson et al., 2003]. Con dicho parser obtenemos una representación en formato Tree-Bank donde se evidencian los núcleos y relaciones entre las sentencias.

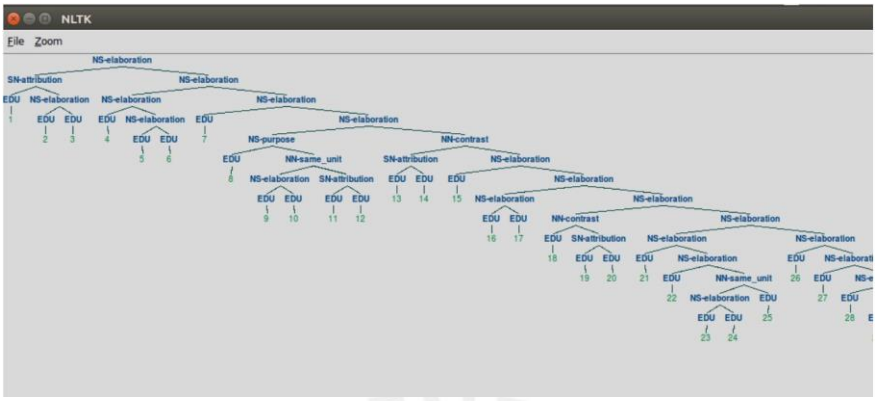


Figura 4.9: Ejemplo de gráfico del resultado del parser DPLP para un documento

Se utilizó el método de O'Donnell para la generación de resúmenes extractivos para los textos expandidos con la resolución de referencias, obteniendo un valor F-1 de Rouge-1 de alrededor de 40 % como se puede ver en la tabla.

La valoración de la importancia de las relaciones entre EDU fue tomada del estudio de [de Uzêda et al., 2007] y se puede encontrar en el Anexo A.

| | **Rouge-1** | **Rouge-L** | **Rouge-SU4** |
|---|---|---|---|
| **RST+RR** | 0.42 | 0.4 | 0.148 |

Cuadro 4.3: Resultados Rouge mediante el método O'Donnell en nuestro Corpus

### Tecnica Abstractiva

A continuación presentamos los resultados sobre el Corpus de entrenamiento que ha sido expandido por la aplicación del de la resolución de referencias.

En primera instancia, obtenemos el resultado de aplicar nuestro modelo de navegación sobre el grafo conceptual que contiene la información semántica y que ha sido fusionado por los conceptos comunes mediante el uso de WordNet. Se extrae las sentencias con mayor importancia calculada por la aplicación de PageRank que en este primer experimento solo considera las relaciones entre los nodos, llamaremos a este resultado (Conceptual + RR).

Posteriormente en (Conceptual + RR + RST), incluimos la información dada por el análisis del discurso, a manera de pesos a cada nodo como se explicó en la fase de análisis. El algoritmo PageRank considera tanto las relaciones como los pesos en cada nodo, en nuestros experimentos el mejor resultado se obtuvo en 30 iteraciones con un damping factor de 65 %, lo que significa que le damos un 65 % de importancia a las relaciones en el grafo conceptual y 35 % de importancia a las relaciones discursivas, en nuestros

experimentos el incrementar aún más la importancia discursiva no significó una mejora en los resultados.

Para saber si la diferencia entre los valores medios obtenidos en nuestras dos mues- tras pareadas son estadísticamente significativos necesitamos de un análisis estadístico inferencial. Este análisis puede ser paramétrico o no paramétrico dependiendo de si nuestros resultados siguen o no una distribución normal. Dados los datos continuos y mayores a 50 elementos el test de normalidad utilizado fue el de Kolmogorov-Smirnov con la corrección de Lillefors y un nivel de significancia del 5 %. Esta prueba será formulada de la siguiente manera:

$H_0$ = Nuestra muestra NO ES significativamente diferente de una población normal

$H_a$ = Nuestra muestra ES significativamente diferente de una población normal

De esta manera, si nuestras muestras siguen una distribución normal utilizaremos el test paramétrico T-Student y si no es así aplicaremos la prueba de rangos con signo de Wilcoxon [Hernández Sampieri et al., 2003], todo esto con el fin de saber si existe una diferencia significativa al 5 % entre las medias de nuestros resultados. Las hipótesis para ambas pruebas serán formuladas de la siguiente manera:

$H_0$ = Los grupos de muestras no difieren significativamente

$H_a$ = Los grupos de muestras difieren significativamente

En la Tabla 4.4, se comparan los resultados del método Conceptual + RR y Conceptual +RR + RST. En nuestro análisis los resultados no superaron el test de normalidad por lo que debemos aceptar la hipótesis alternativa la cual indica que los resultados no siguen una distribución normal y debemos utilizar la prueba no paramétrica. En la tabla 4.4, podemos visualizar como el incluir la información discursiva mejora el desempeño significativamente, como lo indica el nivel p-value de 0.043 para Rouge-1 y de 0.03 en la medida Rouge-L, lo que nos lleva a aceptar la hipótesis alternativa en el test de Wilconxon, la cual indica que si hay una diferencia significativa en nuestros resultados. Lo mismo no se cumple para la medida Rouge-SU4 que alcanza un p-value de 0.057 lo que nos obliga a aceptar la hipótesis nula que indica que no hay una diferencia significativa entre los resultados.

| **F1/Rouge** | **Rouge-1** | **Rouge-L** | **Rouge-SU4** |
|---|---|---|---|
| **Conceptual + RR** | 0.199 | 0.187 | 0.024 |
| **Conceptual + RR + RST** | 0.212 | 0.2 | 0.027 |
| **Wilcoxon Test (P-Valor)** | 0.04314 | 0.03689 | 0.05701 |

Cuadro 4.4: Comparación resumen conceptual y resumen conceptual apoyado por datos discursivos

Este incremento entre la versión puramente conceptual y la que utiliza la información discursiva, hemos notado se debe a que en el modelo conceptual solo se utiliza la información de las relaciones entre los nodos lo que suele favorecer a un número limitado de Agentes o Pacientes que puede estar presentes en la mayor´ıa de expresiones generadas. Por ejemplo, en el siguiente resultado de aplicar el metodo Conceptual + RR, de las 6 sentencias generadas 4 de ellas hablan sobre el mismo agente.

*have nominee judgeship judge belly president United States United States prerogative chairman committee use any circumstance view personal*.

*lie nominating power desire have nominee judgeship judge belly president United States United States prerogative chairman committee*.

*have nominee judgeship judge belly president United States United States prerogative chairman committee nominate someone*.

*Simon call have nominee judgeship judge belly president United States United States prerogative chairman committee*.

*have nominee judgeship judge belly president United States United States prerogative chairman committee want nominate*.

*have nominee judgeship judge belly president United States United States prerogative chairman committee consider conservative*.

Por otro lado al utilizar la información discursiva los pesos permiten encontrar otras expresiones que pueden tener más valor semántico-discursivo y ser más concisas para un resumen. En el mismo documento pero mediante la técnica Conceptual y RST se pudieron obtener ahora 9 sentencias y donde solo 2 de ellas refieren al mismo Agente e incrementaron el valor F1 Rouge-1 en 15 % para dicho documento.

*example name David Souter Thurgood Marshall Edward Kennedy Roman Hruska make plain*.

*Marshall explain other case and have*.

*Ervin point exasperated mean come have*.

*name David Souter Thurgood Marshall Edward Kennedy Roman Hruska express this think*.

*Clarence Thomas answer question explicit believe avoid Supreme*.

*Marshall reveal justice Supreme Court*.

*Howard Metzenbaum say entitle categorical answer direct*.

*have nominee judgeship judge belly president United States United States prerogative chairman committee use any circumstance view personal*.

*lie nominating power desire have nominee judgeship judge belly president United States United States prerogative chairman committee*.

**Generación de Lenguaje**

En este punto tomaremos las sentencias más importantes de nuestro grafo conceptual hasta obtener un total aproximado de 100 palabras pues no tomamos sentencias parciales y obtendremos los resúmenes utilizando la Generación de Lenguaje Natural. Es importante notar que producto de la fusión de nodos se pueden generar expresiones compuestas o nuevas sentencias no presentes en el documento original.

En la siguiente tabla (Tabla 4.5), se muestran los resultados obtenidos que en su mayor´ıa muestran una ligera mejora en Rouge 1, Rouge L y Rouge SU. En nuestros experimentos la mejor combinación se dio cuando adicionamos la frase *"with objective of"* cuando detectamos una relación semántica "*goal* ", llevando por ejemplo una frase generada que no exist´ıa en el documento original:

*"We agree possible international peaceful order devour large state and Gorbachev neighbor"*

Transformandola en:

*"We agreed with objective of possible international peaceful order devour large state and Gorbachev neighbor"*.

| **F1/Rouge** | **Rouge-1** | **Rouge-L** | **Rouge-SU4** |
|---|---|---|---|
| **Conceptual + RR + RST** | 0.212 | 0.2 | 0.027 |
| **Conceptual + RR + RST + NLG** | 0.23 | 0.216 | 0.031 |
| **Wilcoxon Test (P-Valor)** | 0.002 | 0.003223 | - |
| **T-Student Test (P-Valor)** | - | - | 2.2e-16 |

Cuadro 4.5: Comparación resumen conceptual apoyado por datos discursivos y el incluir *SimpleNLG* para la generación de lenguaje natural

Como podemos apreciar en la sentencia generada se identifica una correcta utilización del pronombre *We* con la capitalización adecuada, se identifica también claramente el verbo y el objetivo del mismo y la combinación de dos subexpresiones mediante el conector *And*.

En la tabla (Tabla 4.5) podemos apreciar como los resultados para Rouge-1 y Rouge-L no superaron el test de normalidad por lo que la prueba aplicada fue la de Wilconxon, pero en el caso de Rouge-SU4 si obtuvimos resultados que superaron el test de normalidad por lo que la prueba utilizada fue T-Student.

Adicionalmente podemos apreciar como existe una mejora significativa tanto en Rouge-1, como en Rouge-L y principalmente en la medida Rouge-SU4, lo que indica que el texto es más coherente en relación a la versión provista por las personas. En particular, la utilización de los conectores como *And* y por ende la adecuada conjugación de las frases tanto en genero y numero mejoran claramente los resultados.

### 4.3.2. Validación en Corpus de Prueba

Por último para validar los resultados se ejecuto la misma operación sobre los documentos seleccionados para prueba del corpus de la DUC, esto quiere decir que se utilizo la resolución de referencias para expandir los documentos, el análisis sintáctico y semántico

presente en AMR, para después construir un grafo conceptual aprovechando las relaciones semánticas presentes en Propbank, en el cual utilizo entre otras técnicas la similitud de conceptos dada por Wordnet.

Por otro lado se extrajo la información discursiva mediante la utilización del análisis RST, para utilizarla con el grafo conceptual. Aplicamos el algoritmo PageRank aprovechando los pesos dados por RST para detectar los nodos más importantes y con esto extraer las sentencias más relevantes incluyendo su información semántica con la cual alimentar un generador de lenguaje natural para mejorar la coherencia de las sentencias.

A continuación los resultados obtenidos en cuanto a precisión (Tabla 4.6) y exhaustividad (Tabla 4.7), donde podemos apreciar que un menor resultado en la exhaustividad del modelo lo que indica la deficiencia en ubicar los términos relevantes. Podemos apreciar también como el uso de NLG mejora ligeramente este factor.

| **Precisión/ Rouge** | **Rouge-1** | **Rouge-L** | **Rouge-SU4** |
|---|---|---|---|
| **RST** | 0.409 | 0.389 | 0.127 |
| **Conceptual+ RR** | 0.241 | 0.227 | 0.031 |
| **Conceptual+ RR+RST** | 0.244 | 0.232 | 0.032 |
| **Conceptual+ RR+RST+ NLG** | 0.243 | 0.23 | 0.033 |

Cuadro 4.6: Tabla de precisión

| **Exhaustividad/ Rouge** | **Rouge-1** | **Rouge-L** | **Rouge-SU4** |
|---|---|---|---|
| **RST** | 0.42 | 0.399 | 0.131 |
| **Conceptual+ RR** | 0.21 | 0.198 | 0.027 |
| **Conceptual+ RR+RST** | 0.215 | 0.204 | 0.028 |
| **Conceptual+ RR+RST+ NLG** | 0.247 | 0.233 | 0.033 |

Cuadro 4.7: Tabla de exhaustividad

En los resultados de la tabla (4.8), podemos ver la comparativa entre el método conceptual y el método conceptual con información discursiva (RST), en el caso de Rouge-1 y Rouge-L no se cumplio con el supuesto de normalidad por lo que tuvimos que recurrir a un test no paramétrico, en el caso de de Rouge-SU4 se validó el supuesto de normalidad por lo que se utilizó una prueba paramétrica. Ninguno de los resultados en los datos de prueba mostró una diferencia significativa, sin embargo debemos indicar, que en nuestros experimentos no se ha utilizado ningún algoritmo que presente un sobre ajuste hacia la data de entrenamiento, lo que significa que la bondad de esta técnica depende solamente y de manera individual del texto contenido en cada set de documentos. Adémas es importante notar que el método nunca perjudica los resultados.

| F1/Rouge | Rouge-1 | Rouge-L | Rouge-SU4 |
|---|---|---|---|
| **Conceptual+ RR** | 0.224 | 0.211 | 0.029 |
| **Conceptual+ RR+RST** | 0.228 | 0.217 | 0.029 |
| **Wilcoxon Test (P-Valor)** | 0.46 | 0.3494 | - |
| **T-StudentTest (P-Valor)** | - | - | 0.2059 |

Cuadro 4.8: Tabla de la relación F1 entre el metodo Conceptual y Conceptual con RST

En los resultados de la tabla (4.9), de igual manera el supuesto de normalidad solo se pudo probar para Rouge-SU4, en esta ocasión los resultados mostraron una mejora significativa en todas las metricas.

| F1/Rouge | Rouge-1 | Rouge-L | Rouge-SU4 |
|---|---|---|---|
| **Conceptual+ RR+RST** | 0.228 | 0.217 | 0.029 |
| **Conceptual+ RR+RST+ NLG** | 0.244 | 0.231 | 0.033 |
| **Wilcoxon Test (P-Valor)** | 0.002755 | 0.005253 | - |
| **T-StudentTest (P-Valor)** | - | - | 2.2e-16 |

Cuadro 4.9: Tabla de la relación F1 entre el metodo Conceptual con RST y Conceptual con RST y NLG

En los resultados de la tabla (Tabla 4.10) podemos apreciar una mejora constante al aplicar cada técnica, aunque no tan buena y estadísticamente significativa como en el set de entrenamiento. Contamos también con la información del método extractivo con base en RST que nos muestra que aún existe una gran diferencia entre el éxito alcanzado al utilizar segmentos de las sentencias y la capacidad de sintetizar el contenido más importante en expresiones originales.

| F1/Rouge | Rouge-1 | Rouge-L | Rouge-SU4 |
|---|---|---|---|
| **RST** | 0.413 | 0.393 | 0.129 |
| **Conceptual+ RR** | 0.224 | 0.211 | 0.029 |
| **Conceptual+ RR+RST** | 0.228 | 0.217 | 0.029 |
| **Conceptual+ RR+RST+ NLG** | 0.244 | 0.231 | 0.033 |

Cuadro 4.10: Tabla resumen de la relación F1 entre precisión y exhaustividad de los experimentos en el corpus de prueba

De igual manera vemos como los datos son consistentes entre el corpus de entrenamiento y el corpus de prueba por lo que podemos decir que el modelo propuesto tiene un desempeño estable alrededor del 24 % como medida F1 entre la relación de precisión y exhaustividad en la medida Rouge-1.

# Capítulo 5

# Conclusiones y Trabajos Futuros

En este capítulo se presentan las conclusiones, limitaciones, contribuciones y trabajos futuros. De esta manera el capitulo esta dividido en 4 secciones; en la sección 5.1, presentamos las conclusiones del trabajo, en la sección 5.2 se presentan las contribuciones, en la sección 5.3 las limitaciones encontradas y las sugerencias para trabajos futuros.

## 5.1. Conclusiones

El objetivo central del presente trabajo fue la investigación y desarrollo de un método automático para la generación de resúmenes utilizando un enfoque abstractivo que utilice información semántica y discursiva. Dicho método fue implementado siguiendo la arquitectura propuesta por [Mani, 2001], donde se definen 3 etapas, la primera es la de análisis donde interpretamos y representamos en un formato computacional la información, la segunda es la de transformación donde identificamos y seleccionamos el contenido más relevante y como resultado tendremos una representación computacional condensada de los texto y la tercera es la etapa de síntesis donde es generado un texto en lenguaje natural.

En la etapa de análisis, se abordó nuestro primer objetivo específico, que está asociado con la hipótesis de utilizar un anotador semántico *Abstract Meaning Representation* (AMR) [Banarescu et al., 2013] para generar automáticamente los grafos conceptuales. En la propuesta podemos apreciar cómo este objetivo es alcanzado, en primera instancia mediante la utilización de la resolución de referencias para expandir y resolver mejor la información contenida en los documentos. Posteriormente, fue utilizado el anotador de AMR (CAMR) [Wang et al., 2016] para extraer automáticamente grafos semánticos por sentencia.

Estos grafos iniciales son expandidos al utilizar la información sobre los roles semánticos contenida en PropBank, de esta forma tenemos grafos cuyos nodos son conceptos y sus aristas representan relaciones semánticas como *Agent*, *Patient*, *Goal*, *Theme* entre

otras.

Estos grafos por sentencia deben ser fusionados en un solo grafo conceptual mediante la unión de los conceptos comunes, esta unión se dará sólo entre conceptos y no verbos. Además, para expandir la capacidad de síntesis del método se utiliza una comparación de conceptos basada en la medida de distancia propuesta por [Wu and Palmer, 1994] que utiliza el conocimiento de la WordNet. De esta manera se prueba la utilidad de la resolución de referencias y recursos de conocimiento como Propbank y Wordnet en la síntesis de conceptos, que es el segundo objetivo específico de nuestro trabajo.

En la etapa de transformación es necesario identificar el contenido más importante por esto, nuestro método adiciona la información discursiva al grafo conceptual que ya contiene información semántica con el fin de tener un enfoque que englobe tanto el contexto sentencial como a nivel del documento. En nuestro modelo la información discursiva es obtenida mediante el uso de *Rhetorical Structure Theory* (RST) [Mann and Thompson, 1988], que nos permite identificar partes nucleares y satelitales en los llamados *Elementary Discourse Unit* (EDU), asi como también se identifican relaciones discursivas entre EDUs.

Utilizamos el método propuesto por [O' Donnell, 1997], el cual toma en consideración tanto la clasificación nuclear como también las relaciones discursivas, obtenemos así valores de importancia discursiva para los EDU de cada sentencia que asociamos a los conceptos en nuestro grafo.

Dado que necesitamos identificar los contenidos más importantes, utilizamos el algoritmo PageRank dado que contamos con un grafo de conceptos, donde a mayor número de relaciones se intuye una mayor frecuencia de utilización en el texto y por ende una más alta importancia; además, se conto con la información discursiva asignada como pesos a cada concepto.

PageRank nos permitió incluir en un solo método de calificación, la información semántica y discursiva. En nuestro trabajo se demuestra como esta simbiosis es siempre provechosa y cumple con nuestro tercer objetivo.

La etapa de síntesis necesito de la implementación de un método para navegar el grafo conceptual que ahora en cada nodo contiene una clasificación semántica y discursiva y con esto generar nuevas expresiones.

En el presente trabajo, se utilizaron los roles semánticos, asociando el rol *Agent* como sujeto de la expresión, el verbo por ser el eje del análisis semántico en AMR es fácilmente identificable y los roles semánticos *Patient*, *Theme* y *Goal* cómo constituyentes del predicado. Entonces por cada verbo presente en el grafo se construyeron las expresiones que contengan uno o varios sujetos y uno o varios predicados. Con esta regla se extrajeron varias expresiones cuyo peso total fue la suma del peso de todos los conceptos que contienen, posteriormente fueron ordenadas de manera decreciente y las principales fueron tomadas para la generación de los resúmenes que en nuestros experimentos tienen una tasa de compresión del 20 %, lo que significa al rededor de 100 palabras.

Por último, estas tripletas sujeto, verbo y predicado fueron utilizadas en conjunto

con la herramienta SimpleNLG para la generación de lenguaje natural. En nuestro trabajo configuramos la herramienta para generar los resúmenes en tiempo pasado y de esta manera se generaron expresiónes que fueron coherentes en tiempo y número.

En nuestro trabajo se valida la viabilidad de utilizar los grafos conceptuales como base de conocimiento para la posterior generación de lenguaje natural. Nuestros resultados muestran una medida F1 del 24 % en la métrica Rouge-1, con esto queda demostrada nuestro objetivo específico final. Es importante notar que en nuestros experimentos no se pudo superar los resultados obtenidos por la técnica extractiva, que llegó a alcanzar una medida Rouge-1 de 41 %. Esto no significa que el método planteado no es útil para la generación de resúmenes, pues como se ha descrito anteriormente, las técnicas abstractivas tienen una mayor dificultad pero también representan el camino para superar la cohesión y coherencia de los resúmenes generados por técnicas extractivas.

El método propuesto fue evaluado sobre el Corpus DUC 2002 el cual es específico para el trabajo en la generación de resúmenes mono documento, que consta de artículos de noticias de diferentes fuentes y resúmenes generados manualmente.

Nuestro objetivo principal fue cubierto al presentar un modelo que ejemplifica cómo podemos integrar la información semántica y discursiva en un grafo conceptual que pueda ser utilizado para generar sentencias con la información más importante.

## 5.2. Contribuciones

Las contribuciones realizadas en el presente trabajo son descritas a continuación:

- Un método que ejemplifica la posibilidad de generar resúmenes con un enfoque abstractivo que utiliza información semántica y discursiva en un grafo conceptual ponderado con un algoritmo PageRank para luego generar lenguaje natural.
- Un modelo de cómo aplicar la resolución de referencias para expandir la información que puede ser obtenida por el anotador AMR, siendo esta una capacidad de expandir el texto propia de los métodos abstractivos y una recomendación en [Liu et al., 2015] que no fue implementada en dicho trabajo.
- Un modelo de cómo aplicar un anotador semántico AMR y la información en Prop-Bank asociada a esta representación para generar automáticamente grafos conceptuales, tal como se menciona en [Miranda-Jiménez et al., 2014] los grafos conceptuales son útiles para la generación de resúmenes pero es necesario encontrar métodos para poder generarlos automáticamente, por lo que el presente trabajo representa una forma viable de hacerlo.
- Un modelo de s´ıntesis para nuestra propuesta de grafo conceptual que utiliza el conocimiento en WordNet.

- Un modelo que utiliza la información semántica y discursiva mediante la aplicación de un algoritmo PageRank con pesos obtenidos de la utilización del método en [O' Donnell, 1997], que en conocimiento del autor es la primera vez que se aplica sobre un grafo conceptual generado con un anotador AMR. Se aporta también un número de iteraciones y valor de *Dumping* recomendados. PageRank en nuestra opinión es superior al método HITS pues permite ponderar información adicional en forma de pesos asociados a cada nodo del grafo conceptual, por lo que su aplicacion a este tipo de grafos es un aporte a la teoria presentada en [Miranda-Jiménez et al., 2014].

- Un metodo para navegar un grafo conceptual que utiliza la información de los roles semánticos obtenidos mediante AMR y PropBank para alimentar la herramienta SimpleNLG y generar lenguaje natural.

- Nuestro experimento muestra un estado del arte en cuanto a la utilización de las herramientas existentes con el fin de materializar en un solo modelo las ideas propuestas por [Miranda-Jiménez et al., 2014], [Liu et al., 2015] y [Gerani et al., 2014], todo esto sobre el corpus para la generación de resúmenes DUC 2002.

## 5.3. Limitaciones y Trabajos Futuros

- Aunque los anotadores morfo sintácticos tienen una certeza importante al momento actual, no es así el caso de los anotadores discursivos y semánticos. Este hecho ha sido una limitante en los experimentos, porque si bien podemos obtener una representación de los documentos aún tenemos pérdidas importantes de información principalmente en el parser semántico AMR. Aunque esto no retira la importancia que tendrán en el futuro. Pensamos en el mismo sentido de [Miranda-Jiménez et al., 2014] [Liu et al., 2015] que la utilización de estas bases de conocimiento como es Propbank y Wordnet son una respuesta a un abordaje más profundo sobre el significado de las sentencias, pues nos permite desambiguar y establecer mejor los roles semánticos.

- Un trabajo futuro nace de la necesidad de un mejor modelo de abstracción, aunque AMR será un elemento importante en el análisis semántico, en su forma actual resulta insuficiente para abstraer los conceptos principales pues aún está muy influenciado por la sintáxis, lo cual se evidencia en la distinta representación que se obtiene de una misma idea dependiendo de si esta se encuentra escrita en una sentencia en voz pasiva o activa. Pensamos al igual que [Miranda-Jiménez et al., 2014], que es necesario llevar la abstracción un nivel conceptual.

- Encontrar mejores técnicas para recorrer el grafo conceptual ponderado es un desafío a futuro pues de esta habilidad también depende la capacidad de generar mejores sentencias para el resumen.

- Será necesario encontrar una manera de generar lenguaje a partir del mismo, pensamos que la generación de lenguaje natural es también una tarea de la cual dependerá el éxito de los modelos de generación de resúmenes abstractivos.

# Appendices

# Apéndice A

# Valores de importancia en las relaciones semánticas

| antithesis | List | purpose | reason-e |
|---|---|---|---|
| antithesis-e | manner | purpose-e | Reason |
| cause | manner-e | question-answer | result |
| cause-e | otherwise | question-answer-e | result-e |
| Cause-Result | otherwise-e | question-answer-n | Result |
| concession | Otherwise | question-answer-s | Same-Unit |
| concession-e | problem-solution | question-answer-n-e | Same-Unit-NS |
| condition | problem-solution-e | question-answer-s-e | Same-Unit-SN |
| condition-e | problem-solution-n | Question-Answer | Sequence |
| Contrast | problem-solution-s | statement-response-n | topic-drift |
| Disjunction | problem-solution-n-e | statement-response-s | topic-shift |
| Inverted-Sequence | problem-solution-s-e | Statement-Response | Topic-Drift |
| Joint | Problem-Solution | reason | Topic-Shift |

Cuadro A.1: Relaciones pertenecientes a la categor´ıa ++ Importantes (factor de importancia = 0.8)

| comparison | Enablement | evaluation-n-e | nonrestrictive-relative-e |
|---|---|---|---|
| comparison-e | evaluation | evaluation-s-e | preference |
| Comparison | evaluation-e | Evaluation | preference-e |
| enablement | evaluation-n | means | relative-e |
| enablement-e | evaluation-s | means-e | restrictive-rel-e |

Cuadro A.2: Relaciones pertenecientes a la categor´ıa + Importantes (factor de importancia = 0.6)

| | | | |
|---|---|---|---|
| Abstract | consequence-n-e | interpretation-n | summary |
| analogy | consequence-s-e | interpretation-s | summary-e |
| analogy-e | Consequence | interpretation-n-e | summary-n |
| Analogy | contingency | interpretation-s-e | summary-s |
| Attribution | contingency-e | Interpretation | summary-n-e |
| Author | evidence | justify | summary-s-e |
| Column-Title | evidence-e | justify-e | Summary |
| comment | explanation-argumentative | Parallel | Text |
| comment-e | explanation-argumentative-e | Proportion | TextualOrganization |
| Comment-Topic | Heading | restatement | Title |
| conclusion | hypothetical | restatement-e | Topic |
| conclusion-e | hypothetical-e | rhetorical-question | Topic-Comment |
| consequence-n | interpretation | SectionText | Topic-WA-Comment |
| consequence-s | interpretation-e | SectionTitle | |

Cuadro A.3: Relaciones pertenecientes a la categor´ıa - Importantes (factor de importancia = 0.4)

| | | |
|---|---|---|
| attribution | elaboration-part-whole | OTHERrel |
| attribution-e | elaboration-process-step | OTHERrel-e |
| attribution-n | elaboration-project-attribute | OTHERmultinuc |
| background | elaboration-general-specific | parenthetical |
| background-e | elaboration-additional-e | temporal-after |
| circumstance | elaboration-set-member-e | temporal-before |
| circumstance-e | elaboration-part-whole-e | temporal-sametime |
| definition | elaboration-process-step-e | temporal-after-e |
| definition-e | elaboration-object-attribute-e | temporal-before-e |
| elaboration | elaboration-general-specific-e | temporal-sametime-e |
| elaboration-e | example | TemporalSameTime |
| elaboration-additional | example-e | |
| elaboration-set-member | motivation | |

Cuadro A.4: Relaciones pertenecientes a la categor´ıa - Importantes (factor de importancia = 0.4)

# Bibliografía


[Banarescu et al., 2013]Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract meaning representation for sembanking. *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

[Banerjee et al., 2015]Banerjee, S., Mitra, P., and Sugiyama, K. (2015). Multi-document abstractive summarization using ilp based multi-sentence compression. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1208–1214. AAAI Press.

[Barbara and Charters, 2007]Barbara, K. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. *EBSE Technical Report EBSE-2007-01.2007*.

[Barzilay and Elhadad, 1999]Barzilay, R. and Elhadad, M. (1999). Using lexical chains for text summarization. *Advances in automatic text summarization*, pages 111–121.

[Barzilay and McKeown, 2005]Barzilay, R. and McKeown, K. R. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

[Baumgartner et al., 2007]Baumgartner, W. A., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G., and Hunter, L. (2007). Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, 23(13):i41–i48.

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

[Brin and Page, 1998]Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117.

[Cardoso, 2014]Cardoso, P. C. F. (2014). *Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo*. PhD thesis, Universidade de São Paulo.

[Carenini and Cheung, 2008]Carenini, G. and Cheung, J. C. K. (2008). Extractive vs. nlg-based abstractive summarization of evaluative text: The effect of corpus controver- siality. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 33–41. Association for Computational Linguistics.

[Carenini et al., 2006]Carenini, G., Ng, R., and Pauls, A. (2006). Multi-document summarization of evaluative text. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.

[Carlson et al., 2003]Carlson, L., Marcu, D., and Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

[Chieze et al., 2008]Chieze, E., Farzindar, A., and Lapalme, G. (2008). Automatic summarization and information extraction from canadian immigration decisions. In *Proceedings of the Semantic Processing of Legal Texts Workshop*, pages 51–57.

[Clarke and Lapata, 2008]Clarke, J. and Lapata, M. (2008). Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:399–429.

[Cohn and Lapata, 2009]Cohn, T. A. and Lapata, M. (2009). Sentence compression as tree transduction. *Journal of Artificial Intelligence Research*, 34:637–674.

[Conroy and O'leary, 2001]Conroy, J. M. and O'leary, D. P. (2001). Text summarization via hidden markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 406–407. ACM.

[Cook, 1989]Cook, W. A. (1989). *Case grammar theory*. Georgetown University Press.

[Cremmins, 1996]Cremmins, E. T. (1996). The art of abstracting.

[Dang et al., 2000]Dang, H. T., Kipper, K., and Palmer, M. (2000). Integrating compositional semantics into a verb lexicon. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 1011–1015. Association for Computational Linguistics.

[Das and Martins, 2007]Das, D. and Martins, A. F. (2007). A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, 4:192–195.

[De Marneffe et al., 2006]De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

[de Uzêda et al., 2007]de Uzêda, V. R., Pardo, T. A. S., and Nunes, M. d. G. V. (2007). *Estudo e avaliação de métodos de sumarização automática de textos baseados na RST*. ICMC-USP.

[Durrett et al., 2016]Durrett, G., Berg-Kirkpatrick, T., and Klein, D. (2016). Learning-based single-document summarization with compression and anaphoricity constraints. *arXiv preprint arXiv:1603.08887*.

[Edmundson, 1969]Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285.

[Erkan and Radev, 2004]Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

[Farzindar and Lapalme, 2004]Farzindar, A. and Lapalme, G. (2004). Legal text summarization by exploration of the thematic structures and argumentative roles. In *Text Summarization Branches Out Workshop held in conjunction with ACL*, pages 27–34.

[Filippova, 2010]Filippova, K. (2010). Multi-Sentence Compression: Finding Shortest Paths in Word Graphs. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING10)*, (August):322–330.

[Filippova and Strube, 2008a]Filippova, K. and Strube, M. (2008a). Dependency tree based sentence compression. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 25–32. Association for Computational Linguistics.

[Filippova and Strube, 2008b]Filippova, K. and Strube, M. (2008b). Sentence fusion via dependency graph compression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 177–185. Association for Computational Linguistics.

[Flanigan et al., 2014]Flanigan, J., Thomson, S., Carbonell, J., Dyer, C., and Smith, N. a. (2014). A Discriminative Graph-Based Parser for the Abstract Meaning Representation. *Acl*, pages 1426–1436.

[Fuchs and Schwitter, 1996]Fuchs, N. E. and Schwitter, R. (1996). Attempto controlled english (ace). *arXiv preprint cmp-lg/9603003*.

[Galliers and Jones, 1993]Galliers, J. R. and Jones, K. S. (1993). Evaluating natural language processing systems.

[Ganesan et al., 2010]Ganesan, K., Zhai, C., and Han, J. (2010). Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd international conference on computational linguistics*, pages 340–348. Association for Computational Linguistics.

[Gantz and Reinsel, 2012]Gantz, J. and Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, 2007:1–16.

[Gatt and Reiter, 2009]Gatt, A. and Reiter, E. (2009). Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 90–93. Association for Computational Linguistics.

[Genest and Lapalme, 2011]Genest, P.-E. and Lapalme, G. (2011). Framework for abstractive summarization using text-to-text generation. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 64–73. Association for Computational Linguistics.

[Genest and Lapalme, 2012]Genest, P.-E. and Lapalme, G. (2012). Fully abstractive approach to guided summarization. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 354–358. Association for Computational Linguistics.

[Genest et al., 2013]Genest, P.-E., Lapalme, G., and Yousfi-Monod, M. (2013). Hextac: the creation of a manual extractive run. *Génération de résumés par abstraction*, page 7.

[Gerani et al., 2014]Gerani, S., Mehdad, Y., Carenini, G., Ng, R. T., and Nejat, B. (2014). Abstractive summarization of product reviews using discourse structure. In *EMNLP*, pages 1602–1613.

[Hahn and Mani, 2000]Hahn, U. and Mani, I. (2000). The challenges of automatic summarization. *Computer*, 33(11):29–36.

[Haveliwala, 2002]Haveliwala, T. H. (2002). Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526. ACM.

[Hernández Sampieri et al., 2003]Hernández Sampieri, R., Lucio, B., Collado, M. d. P. F., Sampieri, C. H., Collado, C. F., and Lucio, P. B. (2003). *Metodología de la investigación*. Number 303.1. McGraw-Hill,.

[Hirao et al., 2013]Hirao, T., Yoshida, Y., Nishino, M., Yasuda, N., and Nagata, M. (2013). Single-document summarization as a tree knapsack problem. In *EMNLP*, volume 13, pages 1515–1520.

[Hovy and Miktov, 2005]Hovy, E. and Miktov, R. (2005). Automated text summarization. In *The Oxford Handbook of Computational Linguistics*, pages 583–598. Oxford University Press.

[Hu and Liu, 2004]Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

[Jackendoff, 1972]Jackendoff, R. S. (1972). Semantic interpretation in generative grammar.

[Ji and Eisenstein, 2014]Ji, Y. and Eisenstein, J. (2014). Representation learning for text-level discourse parsing. In *ACL (1)*, pages 13–24.

[Jiang and Conrath, 1997]Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

[Joty et al., 2013]Joty, S. R., Carenini, G., Ng, R. T., and Mehdad, Y. (2013). Combining intra-and multi-sentential rhetorical parsing for document-level discourse analysis. In *ACL (1)*, pages 486–496.

[Jurafsky and Martin, 2009]Jurafsky, D. and Martin, J. h. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

[Khan et al., 2016]Khan, A., Salim, N., and Isiaka obasa, A. (2016). An Optimized Semantic Technique for Multi-Document Abstractive Summarization. *Indian Journal of Science and Technology*, 8(32).

[Kilgarriff and Fellbaum, 2000]Kilgarriff, A. and Fellbaum, C. (2000). Wordnet: An electronic lexical database.

[Kingsbury and Palmer, 2002]Kingsbury, P. and Palmer, M. (2002). From treebank to propbank. In *LREC*. Citeseer.

[Kingsbury and Palmer, 2003]Kingsbury, P. and Palmer, M. (2003). Propbank: the next level of treebank. In *Proceedings of Treebanks and lexical Theories*, volume 3. Citeseer.

[Kipper et al., 2000]Kipper, K., Dang, H. T., Palmer, M., et al. (2000). Class-based construction of a verb lexicon. In *AAAI/IAAI*, pages 691–696.

[Kleinberg, 1999]Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.

[Knight et al., 2014]Knight, K., Baranescu, L., Bonial, C., Georgescu, M., Griffitt, K., Hermjakob, U., Marcu, D., Palmer, M., and Schneifer, N. (2014). Abstract meaning representation (amr) annotation release 1.0. *Web download*.

[Knight and Marcu, 2000]Knight, K. and Marcu, D. (2000). Statistics-based summarization-step one: Sentence compression. *AAAI/IAAI*, 2000:703–710.

[Kupiec et al., 1995]Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68–73. ACM.

[Lee et al., 2005]Lee, C.-S., Jian, Z.-W., and Huang, L.-K. (2005). A fuzzy ontology and its application to news summarization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(5):859–880.

[Lin, 2004]Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.

[Liu et al., 2015]Liu, F., Flanigan, J., Thomson, S., Sadeh, N., and Smith, N. A. (2015). Toward abstractive summarization using semantic representations.

[Luhn, 1958]Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

[Manchego, 2013]Manchego, F. E. A. (2013). *Anotação automática semissupervisionada de papéis semânticos para o português do Brasil*. PhD thesis, Universidade de São Paulo.

[Mani, 2001]Mani, I. (2001). *Automatic summarization*, volume 3. John Benjamins Publishing.

[Mani et al., 2002]Mani, I., Klein, G., House, D., Hirschman, L., Firmin, T., and Sundheim, B. (2002). Summac: a text summarization evaluation. *Natural Language Engineering*, 8(01):43–68.

[Mann and Thompson, 1988]Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

[Manning et al., 2014]Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

[Marcu et al., 2000]Marcu, D., Carlson, L., and Watanabe, M. (2000). The automatic translation of discourse structures. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 9–17. Association for Computational Linguistics.

[Marcus et al., 1994]Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The penn treebank: A revised corpus design for extracting predicate argument structure. In *Proceedings of the ARPA Human Language Technology Workshop*.

[Marcus et al., 1993]Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

[Mathiessen and Bateman, 1991]Mathiessen, C. M. and Bateman, J. (1991). Text generation and systemic-functional linguistics. *London: Pinter*.

[Mihalcea and Tarau, 2004]Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing order into texts. *Proceedings of EMNLP*, 85:404–411.

[Miller, 1995]Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

[Miranda-Jiménez et al., 2013]Miranda-Jiménez, S., Gelbukh, A., and Sidorov, G. (2013). Summarizing conceptual graphs for automatic summarization task. In *International Conference on Conceptual Structures*, pages 245–253. Springer.

[Miranda-Jiménez et al., 2014]Miranda-Jiménez, S., Gelbukh, A., and Sidorov, G. (2014). Conceptual graphs as framework for summarizing short texts. *International Journal of Conceptual Structures and Smart Applications (IJCSSA)*, 2(2):55–75.

[Mohan et al., 2016]Mohan, M. J., Sunitha, C., Ganesh, A., and Jaya, A. (2016). A study on ontology based abstractive summarization. *Procedia Computer Science*, 87:32–37.

[Montes-y Gómez et al., 2001]Montes-y Gómez, M., Gelbukh, A., Lopez-Lopez, A., and Baeza-Yates, R. (2001). Flexible comparison of conceptual graphs. In *International Conference on Database and Expert Systems Applications*, pages 102–111. Springer.

[Murtagh and Contreras, 2011]Murtagh, F. and Contreras, P. (2011). Methods of hierarchical clustering. *arXiv preprint arXiv:1105.0121*.

[Nenkova and Vanderwende, 2005]Nenkova, A. and Vanderwende, L. (2005). The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101*.

[Nóbrega et al., 2013]Nóbrega, F. A. A., Pardo, T. A. S., and de Linguística Computacional, N. I. (2013). Desambiguação lexical de sentido com uso de informação multidocumento por meio de redes de co-ocorrência. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Techonology*, pages 138–147.

[Oliveira et al., 2016]Oliveira, H., Lima, R., Lins, R. D., Freitas, F., Riss, M., and Simske, S. J. (2016). Assessing concept weighting in integer linear programming based single-document summarization. In *Proceedings of the 2016 ACM Symposium on Document Engineering*, pages 205–208. ACM.

[Ono et al., 1994]Ono, K., Sumita, K., and Miike, S. (1994). Abstract generation based on rhetorical structure extraction. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 344–348. Association for Computational Linguistics.

[O' Donnell, 1997]O ' Donnell, M. (1997). Variable-length on-line document generation. In *the Proceedings of the 6th European Workshop on Natural Language Generation, Gerhard-Mercator University, Duisburg, Germany*.

[Palmer et al., 2005]Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

[Papineni et al., 2002]Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

[Radev et al., 2002]Radev, D., Winkel, A., and Topper, M. (2002). Multi document centroid-based text summarization. In *ACL 2002*.

[Radev et al., 2003]Radev, D. R., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Qi, H., Celebi, A., Liu, D., and Drabek, E. (2003). Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 375–382. Association for Computational Linguistics.

[Ramezani and Feizi-Derakhshi, 2015]Ramezani, M. and Feizi-Derakhshi, M.-R. (2015). Ontology-based automatic text summarization using farsnet. *Advances in Computer Science: an International Journal*, 4(2):88–96.

[Reiter et al., 2000]Reiter, E., Dale, R., and Feng, Z. (2000). *Building natural language generation systems*, volume 33. MIT Press.

[Saggion et al., 2016]Saggion, H., Poibeau, T., Saggion, H., Poibeau, T., Text, A., and Past, S. (2016). Automatic Text Summarization : Past , Present and Future.

[Saggion et al., 2002]Saggion, H., Teufel, S., Radev, D., and Lam, W. (2002). Meta-evaluation of summaries in a cross-lingual environment using content-based metrics. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

[Settles, 2005]Settles, B. (2005). Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192.

[Shi et al., 2001]Shi, Y. et al. (2001). Particle swarm optimization: developments, applications and resources. In *evolutionary computation, 2001. Proceedings of the 2001 Congress on*, volume 1, pages 81–86. IEEE.

[Shieber, 1986]Shieber, S. M. (1986). An introduction to unification-based approaches to grammar.

[Sowa, 1983]Sowa, J. F. (1983). Conceptual structures: information processing in mind and machine.

[Svore et al., 2007]Svore, K. M., Vanderwende, L., and Burges, C. J. (2007). Enhancing single-document summarization by combining ranknet and third-party sources. In *EMNLP-CoNLL*, pages 448–457.

[Tanaka et al., 2009]Tanaka, H., Kinoshita, A., Kobayakawa, T., Kumano, T., and Kato, N. (2009). Syntax-driven sentence revision for broadcast news summarization. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, pages 39–47. Association for Computational Linguistics.

[Torres-Moreno, 2014]Torres-Moreno, J.-M. (2014). *Automatic text summarization*. John Wiley & Sons.

[Uzeda et al., 2008]Uzeda, V. R., Pardo, T. A. S., and Nunes, M. D. G. V. (2008). Evaluation of automatic text summarization methods based on rhetorical structure theory. *Intelligent Systems Design and*.

[Uzêda et al., 2010]Uzêda, V. R., Pardo, T. A. S., and Nunes, M. D. G. V. (2010). A comprehensive comparative evaluation of rst-based summarization methods. *ACM Transactions on Speech and Language Processing (TSLP)*, 6(4):4.

[Van der Merwe and Engelbrecht, 2003]Van der Merwe, D. and Engelbrecht, A. P. (2003). Data clustering using particle swarm optimization. In *Evolutionary Computation, 2003. CEC'03. The 2003 Congress on*, volume 1, pages 215–220. IEEE.

[Wang et al., 2016]Wang, C., Pradhan, S., Pan, X., Ji, H., and Xue, N. (2016). Camr at semeval-2016 task 8: An extended transition-based amr parser. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1173–1178, San Diego, California. Association for Computational Linguistics.

[Wang et al., 2015]Wang, C., Xue, N., and Pradhan, S. (2015). A transition-based algorithm for amr parsing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–375, Denver, Colorado. Association for Computational Linguistics.

[Wu and Palmer, 1994]Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.

[Xing and Ghorbani, 2004]Xing, W. and Ghorbani, A. (2004). Weighted pagerank algorithm. In *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*, pages 305–314. IEEE.

[Zajic et al., 2007]Zajic, D., Dorr, B. J., Lin, J., and Schwartz, R. (2007). Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Information Processing & Management*, 43(6):1549–1570.

[Zhai and Massung, 2016]Zhai, C. and Massung, S. (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Association for Computing Machinery and Morgan & Claypool, New York, NY, USA.

[Ziegler and Skubacz, 2007]Ziegler, C.-N. and Skubacz, M. (2007). Content extraction from news pages using particle swarm optimization on linguistic and structural features. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 242–249. IEEE Computer Society.