RNPS No. 2143
ISSN 2072-6260
Versión Digital

SERIE GRIS

REPORTE TÉCNICO
Minería de Datos

Procesamiento de expresiones valorativas para la sumarización de opiniones

Lic. Gail García Delgado,
Dr. C. José E. Medina Pagola

RT_010 **marzo 2010**

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
Ciudad de La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2143
ISSN 2072-6260
Versión Digital

SERIE GRIS

REPORTE TÉCNICO
Minería de Datos

Procesamiento de expresiones valorativas para la sumarización de opiniones

Lic. Gail García Delgado,
Dr. C. José E. Medina Pagola

RT_010 **marzo 2010**

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
Ciudad de La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Índice

Procesamiento de expresiones valorativas para la sumarización de opiniones

Lic. Gail García Delgado, Dr. C. José E. Medina Pagola

Centro de Aplicaciones de Tecnología de Avanzada, 7a #21812 e/ 218 y 222, Siboney, Playa, Ciudad de La Habana, Cuba
ggarcia@cenatav.co.cu

RT_010 CENATAV
Fecha del camera-ready: 28 de enero de 2010

Resumen: La minería de opiniones es una disciplina reciente que deriva de la minería de texto, la recuperación de información y la lingüística computacional. Esta disciplina tiene como objetivo el estudio de la información subjetiva en los textos. A partir del análisis de la información subjetiva contenida en textos, se pueden realizar diversos procesamientos, entre los que se encuentra la generación automática de resúmenes valorativos. La generación automática de resúmenes valorativos deriva de una de las tareas de investigación de la minería de textos que concurre en la minería de opiniones, la generación automática de resúmenes de textos. Este reporte centra su atención en la generación automática de resúmenes valorativos, presentando previamente una serie de conocimientos relacionados tanto con opiniones, como con la generación de resúmenes de textos. En este trabajo se centra particular atención en dos pasos esenciales de la sumarización, la detección de tópicos y la selección de segmentos de textos a incluir en el resumen; analizando trabajos donde los textos a resumir son textos valorativos. Se muestran varias formas de realizar la detección de tópicos en opiniones, detectando problemas comunes en los trabajos de sumarización de opiniones consultados. Por otro lado, también se analizan distintas estrategias para la selección de los segmentos de textos valorativos, detectándose también problemas generales en estos trabajos.

Palabras clave: minería de opiniones, minería de texto, recuperación de información, generación automática de resúmenes de texto, generación automática de resúmenes de opiniones

Abstract: The opinion mining is a recent discipline that it derives of the text mining, the information retrieval and the computational linguistics. This discipline aims the study of the subjective information in texts. From the analysis of the subjective information in texts, various processings could be doing, such the automatic generation of opinion summaries. The automatic generation of opinion summaries derives of an one tasks of the texts mining that attends in the opinion mining, the automatic generation of texts summaries. This report centers its attention in the automatic generation of opinion summaries, previously presenting a series of knowledge related with opinions and with the generation of texts summaries. This report centers special attention in two essential steps of the sumarization, the topics detection and the selection of texts segments to include in the summary. We see a varied of ways to accomplish the topics detection in opinions, detecting common problems in the works of opinions sumarization of looked up. In addition, different strategies for the selection of the segments of opinion texts are explored, also detecting general problems in the works.

Keywords: Opinion Mining, Text Mining, Information Retrieval, Automatic Text Summarization, Automatic Opinion Summarization

1. Introducción

En los últimos años, debido al surgimiento y desarrollo de la Internet y las comunicaciones, se ha observado un interés creciente del hombre por extraer conocimiento útil de la información a la que tiene acceso. Gran parte de la información existente se encuentra en forma de textos o documentos escritos en algún lenguaje natural. La información expresada en textos escritos en lenguaje natural se divide principalmente en dos tipos: información subjetiva (también referida como opiniones) y hechos objetivos. Se entiende como opinión a

aquella unidad léxica (palabra, proposición, oración, párrafo) donde sea detectada información subjetiva, categorizándose como información subjetiva a: emociones, sentimientos, creencias, intenciones, especulaciones, propósitos, etc.

La minería de opiniones es la disciplina que tiene como objetivo el estudio de la información subjetiva en textos. La minería de opiniones es una sub-disciplina reciente que deriva de la minería de texto, la recuperación de información y la lingüística computacional. La motivación para el estudio de esta disciplina proviene del deseo de proporcionar herramientas para analistas que procesan información sobre diferentes contextos (político, comercial, gubernamental) con el objetivo de detectar o dar seguimiento automático a los propósitos y juicios insertados en la prensa y en diversos medios y fuentes.

A consecuencia del surgimiento y desarrollo de la Web 2.0[1] son cada vez más los sitios en los que los usuarios expresan sus opiniones sobre diversos temas. Para que la información emitida sea útil, se hace necesario el procesamiento de la misma y es aquí donde tareas relacionadas con la minería de opiniones desempeñan un papel fundamental. A partir del análisis de las opiniones contenidas en textos, se pueden realizar diversos procesamientos, entre los que pude mencionarse: la segmentación de textos valorativos[2], la recuperación de información que contenga opiniones, la generación automática de resúmenes valorativos, la indexación de opiniones, la recuperación de respuestas valorativas, etc.

Este reporte centra su atención en uno de los procesamientos antes mencionados, la generación automática de resúmenes valorativos. Este tipo de procesamiento deriva de una de las tareas de investigación de la minería de textos que concurre en la minería de opiniones, la generación automática de resúmenes de textos.

El presente trabajo se organiza como sigue: primeramente, para un mejor entendimiento de la generación de resúmenes valorativos, se introducen en este reporte una serie de conocimientos relacionados tanto con opiniones en la sección 2, como con la generación de resúmenes de textos en la sección 3. En la sección 4 se abordan dos pasos esenciales en la sumarización, analizándose cómo han sido realizados estos pasos en textos valorativos. El primero de estos pasos es la detección de tópicos, que en el caso de textos valorativos nos referimos a esta tarea como detección de tópicos valorativos. El segundo paso es la selección de segmentos de textos a incluir en el resumen, que será referido como selección de segmentos de textos valorativos a agregar en el resumen. La detección de tópicos valorativos se analiza en la sección 4.1 como un tipo de procesamiento de la información valorativa para la sumarización. Por otro lado, la selección de textos valorativos se analiza en la sección 4.2 como un tipo de información a presentar en el resumen, que para obtenerla se hace necesario realizar un procesamiento no superficial. En la sección 5 se presentan algunas técnicas de evaluación de la calidad de los resúmenes. En la sección 6 se presentan las conferencias en en las que se incluyen tareas relacionadas con la generación de resúmenes valorativos. En la sección 7 se hace una recapitulación del trabajo y finalmente, en la sección 8 se presentan las conclusiones de este reporte.

2. Análisis de opiniones

Como se ha mencionado anteriormente, la minería de opiniones es una disciplina que ha cobrado auge en la actualidad y a partir del análisis de estas se pueden realizar diversos procesamientos a los textos valorativos. En cada uno de dichos procesamientos juegan un papel fundamental tareas del análisis de opiniones, tales como: la extracción y catego-

[1] Un conjunto de aplicaciones donde el usuario tiene el control y se caracterizan por estar basadas en la inteligencia colectiva y el uso de servicios interactivos en red

[2] A partir de este momento en el resto del reporte se utilizarán análogamente las frases textos valorativos y textos evaluativos para referirnos a textos que contienen opiniones.

rización de expresiones que contengan opiniones y la clasificación de textos (documentos o segmentos de ellos) según los sentimientos y opiniones expresadas.

Para la ejecución de las tareas mencionadas anteriormente, es necesario identificar las unidades léxicas que contienen información subjetiva de aquellas que expresan hechos objetivos y posteriormente; definir modelos que permitan representar y procesar la información subjetiva con el propósito de determinar aspectos tales como: el tipo de opinión, su clasificación (positiva o negativa teniendo en cuenta la orientación semántica (o polaridad como también se llama) de la información subjetiva y afecto, apreciación y juicio de acuerdo al tipo de actitud que expresan), quién expresa la opinión (fuente), respecto a qué o a quién es emitida la opinión (tema), la intensidad (gradación como también se llama) de la opinión que indica un grado (puede ser un valor o una categoría) asignado a aspectos como: subjetividad, orientación semántica, etc.

En esta sección se presentan brevemente algunas definiciones y teorías abordadas en la literatura como tipos de opiniones. Además, se presentan los principales enfoques abordados por los métodos de detección y clasificación de opiniones, así como el nivel de profundidad al que se analizan las mismas.

2.1. Tipos de opiniones

Para identificar las unidades léxicas que contienen información subjetiva de aquellas que expresan hechos objetivos es necesario primero definir lo que va a ser contemplado como una opinión o expresión subjetiva. Desde que comenzó el interés en el estudio de las opiniones se han reportado en la literatura trabajos en los que se proponen diferentes definiciones de lo que se considera como opinión, identificando las opiniones como: adjetivos [1], [2], [3], sustantivos [4], términos [5], segmentos de oraciones [6], [7], [8], oraciones [6], [9], [7], [10], [8], [11], documentos [12], etc.

Las definiciones de opiniones son variadas, los investigadores en esta área no han asumido como consenso una definición o teoría general. Lo anterior se debe a que las opiniones se manifiestan de diversas formas en los diferentes textos valorativos, muchas veces dependiendo del dominio y estilo de los mismos.

En algunos trabajos se definen como opiniones a los adjetivos, basándose en el carácter subjetivo que expresan [1], [13], [3], [2], [5]. En algunos trabajos [8], [12], la evaluación y la especulación son reconocidas como tipos de opiniones. Los autores plantean que las oraciones subjetivas pueden ser identificadas debido a que contienen expresiones (evaluativas y especulativas) individuales de subjetividad. En [4] los autores se proponen el desarrollo de un sistema que pueda distinguir oraciones subjetivas de objetivas. Persiguiendo este propósito se clasifican las palabras en *Strong Subjective*, *Weak Subjective* y *Objective*. En [10] se basan en identificar las expresiones polares en oraciones individuales. Los autores identifica cuatro rasgos en los que se pueden clasificar las opiniones: lenguaje (explícito/implícito), mundo (real/posible), modalidad (temporalidad/condicionalidad) y atribución. La polaridad en el proceso de clasificación la obtienen mediante la relación con un léxico previamente seleccionado para el dominio en cuestión.

Algunos autores han definido las opiniones o sentimientos a partir de estados privados [9] , los cuales manifiestan estados de actitud de un sujeto hacia un objeto y están abiertos a la observación o verificación subjetiva. En el trabajo se identifican tres estados privados (intelectuales, emotivos y perceptivos) y se analizan dos tipos de oraciones subjetivas. Usando su definición de opinión, Wiebe *et al.* [6] proponen un amplio esquema de anotación para expresiones subjetivas y crean el corpus MPQA[3]. En su primera versión el MPQA (versión 1.2) contiene 535 documentos manualmente anotados; señalando las expresiones

[3] http://www.cs.pitt.edu/mpqa/databaserelease/

valorativas, sus fuentes, polaridades e intensidades. La segunda edición (versión 2.0) adiciona 157 documentos para un total de 692 y adiciona anotaciones referidas al tópico de las opiniones.

Bethard *et al.* [7] proponen como opinión una oración o segmento de oración que se ajuste como respuesta a la pregunta¨*What does X feel about Y*¨. Los autores definen los complementos del predicado de la oración como opinión proposicional y se proponen identificar tantos las opiniones como los emisores de la misma (fuente de la opinión). Bethard *et al.*, basados en su definición de opinión crean un corpus que contiene 5,139 oraciones anotadas para opiniones. A cada oración se le asigna una de las siguientes etiquetas: NON OPINION, OPINION-PROPOSITION y OPINION-SENTENCE; las cuales indican respectivamente que la oración no contiene opinión, que la oración contiene opinión proposicional y que la oración contiene opinión pero esta no es un complemento del predicado de la oración (no proposicional). Las anotaciones también contienen información acerca de las fuentes de algunas de las opiniones proposicionales.

Kim y Hovy definen como opinión la tupla [Topic, Holder, Claim, Sentiment] [14], donde Topic se refiere a un tema sobre el cual una fuente (Holder) emite una argumentación valorativa (Claim) asociada a un sentimiento (Sentiment) que puede ser bueno o malo (positivo/negativo). Stoyano *et al.* [15], [16], [17] se basan en una definición de opinión muy parecida a la definida anteriormente, planteando que una opinión está compuesta por la tupla [Trigger, Source, Topic, Polarity] donde las componentes (en orden) son equivalentes en significado a Claim, Holder, Topic y Sentiment en la definición de Kim y Hovy.

Algunos autores se han basado en la teoría de la valoración para determinar la información subjetiva en los textos [18], [19]. Esta teoría se enfoca en particular en la expresión linguística de la actitud, así como en el conjunto de recursos que explícitamente posicionan de manera interpersonal las propuestas y las proposiciones textuales. La teoría de la valoración se divide en tres sub-sistemas en los que son clasificadas las opiniones: actitud (afecto, juicio y apreciación), gradación (fuerza y foco) y compromiso (apariencia, pronunciamientos, negaciones, etc.).

2.2. Detección y clasificación de opiniones

En la detección y clasificación de opiniones hay un aspecto que es muy importante tener en cuenta, se trata del dominio al que pertenecen los textos valorativos. Las expresiones que se utilizan para emitir opiniones pueden variar y tener diferentes significados en dependencia del dominio. Por ejemplo, en el segmento ¨El *objeto* es *bastante grande*¨, donde la expresión *bastante grande* es la opinión, si el *objeto* fuera un teléfono celular (lo deseable es que el teléfono sea pequeño), la opinión sería negativa respecto al objeto; sucediendo lo contrario si el objeto fuera un televisor (siendo deseable que el televisor sea grande). Desarrollar métodos de detección de opiniones independientes del dominio es todo un reto en este tema, porque aún cuando se intente el no ajuste a un dominio específico, existe un problema de estilo propio en cada tipo de texto valorativo que no se puede obviar.

Granularidad

De acuerdo a la granularidad o nivel al que se realiza el análisis para la detección y clasificación de opiniones, es posible separar los métodos en dos categorías [15]: granulado grueso (*coarse-grained* en inglés) y granulado fino (*fine-grained* en inglés).

En los métodos categorizados como granulado grueso, la detección de opiniones se realiza a nivel de documentos, donde se persigue como propósito clasificar los documentos en valorativos o no, pudiéndose además en el caso de los valorativos clasificarse de acuerdo a la orientación semántica. Muchos de los trabajos que detectan las opiniones a este nivel trabajan con textos cortos como son los comentarios sobre diversas entidades (productos comerciales, películas, editoriales, restaurantes, etc.) [20], [21], [22], encargándose

de detectar los comentarios valorativos y de clasificarlos en positivos o negativos. Estos trabajos generalmente se apoyan mucho en el dominio al que pertenecen los comentarios, identificando características y patrones regulares que cumplen las opiniones en el marco de los mismos. La complejidad y retos de estas aproximaciones no son tan grandes con respecto a otras que intentan divorciarse del dominio de los textos. La dependencia directa del dominio no permite aplicar las aproximaciones propuestas de manera ampliada a cualquier texto valorativo, lo cual es una deficiencia de estos métodos.

En los métodos de granulado fino la detección de opiniones se realiza al nivel de oraciones o a un menor nivel (expresiones o términos). Los trabajos clasificados en este nivel incluyen una amplia variedad de aproximaciones basadas en diferentes definiciones, características y dominios de las opiniones. Los trabajos clasificados bajo esta categoría realizan un análisis más profundo para la detección de opiniones. A diferencia del granulado grueso, en el granulado fino se detectan y clasifican las opiniones por las que están compuestos los textos valorativos, lo cual permite realizar luego de la detección procesamientos de mayor calidad con la información detectada, ya que es mucho más rica e informativa. En el granulado fino al igual que en el grueso la mayoría de los trabajos son dependientes del dominio (comentarios), aunque hay otros sobre noticias generales o editoriales que persiguen la independencia.

El granulado fino tiene diferentes niveles de complejidad que diferencian al granulado fino dependiente del dominio del granulado fino no dependiente del dominio, siendo este último el más complejo. En el caso de los comentarios sobre alguna entidad, donde se realice el análisis de opiniones mediante granulado fino dependiente del dominio, existen diferencias con respecto al análisis en otros textos valorativos en los que se aplique granulado fino independiente del dominio. Entre estas diferencias se puede mencionar que en el caso de los comentarios, los temas abordados están acotados, ya que la detección de opiniones en los comentarios es solo sobre aquellas relacionadas con la entidad y con los rasgos o aspectos de la misma. Los rasgos sobre una entidad son los aspectos que la caracterizan, por ejemplo, los rasgos de una cámara digital pueden ser: el enfoque del lente, el tamaño, la calidad de la fotos, etc. En el caso de los temas o tópicos abordados en textos valorativos en los que se aplique granulado fino independiente del dominio la detección de temas es más desafiante. En la sección 4.1 se le prestará atención diferenciada a la detección de temas, debido a la importancia que tiene este procesamiento en la generación de resúmenes valorativos.

Enfoques

Los métodos que persiguen la detección y clasificación de opiniones se han abordado fundamentalmente a partir de tres enfoques: realizando anotaciones manuales [23], apoyándose en corpus y basándose en diccionarios [24]. No se ha tomado como consenso la superioridad de alguno de los enfoques con respecto a los otros, cada uno tiene sus características, así como ventajas y desventajas.

El enfoque manual se basa en la observación de los textos para definir patrones de extracción a partir de sus características y peculiaridades. Luego, apoyados en estos patrones se procede anotando y etiquetando la información. Esta estrategia tiene como inconvenientes el alto costo en tiempo y esfuerzo humano.

Los enfoques que se basan en corpus se apoyan en patrones sintácticos [1], semánticos [13], [3] o de co-ocurrencia [25]. A partir de un conjunto inicial de términos semillas extraídos del corpus, se pueden aplicar métodos que identifican en el conjunto de datos, otros términos que se encuentran relacionados con los primeros. Estas relaciones han sido definidas mediante: conjunciones (adjetivos enlazados por conjunciones como ¨y, o¨ usualmente tienen orientación similar, mientras que ¨pero, ni¨ se utiliza con adjetivos de orientaciones opuestas) [1]; un grado significativo (depende de la definición de un umbral) de co-ocurrencia de términos en una vecindad determinada (términos con orientación sim-

ilar tienden a co-ocurrir en los mismos documentos) [25]; sinonimia y antonimia (términos que son sinónimos tienden a tener orientaciones similares, mientras que términos que son antónimos tienden a tener orientaciones opuestas) [3]; significados (términos con glosas similares tienden a tener orientación similar) [26]. La aproximación basada en corpus tiene como desventaja la fuerte dependencia del dominio del conjunto de datos utilizado. Por otra parte, de la definición inicial del conjunto de términos semilla depende la suficiencia del conjunto final de palabras identificadas (usualmente no es suficiente). Como ventaja cabe señalar, que en esta estrategia debido a que se basa en corpus, se tiene en cuenta en el análisis, el contexto en que se utilizan los términos, por lo que la eficacia de la clasificación es mayor que en el enfoque basado en diccionarios.

El enfoque basado en diccionarios consiste básicamente en un proceso similar al anterior, pero en este caso los términos semillas no son extraídos del corpus sino que se definen genéricos, por ejemplo, conjunto de adjetivos representativos por cada una de las orientaciones semánticas seleccionados de un diccionario [27]. Este enfoque apoyándose en una base de conocimientos como WordNet[4] o en un diccionario léxico como General Inquirer[5], identifican igualmente otros términos relacionados. Esta aproximación a diferencia de la anterior no es dependiente del dominio del corpus pero si es necesario la correcta definición del conjunto de términos semillas y tener disponible los lexicones que son recursos necesarios en este tipo de aproximación.

3. Generación automática de resúmenes de textos

La sumarización[6] es el proceso de abreviar un texto o un conjunto de ellos; o sea, brindar a partir de estos una versión más corta sin perder el contenido esencial de la información ofrecida. Los algoritmos tradicionales para la generación de resúmenes de textos se basan en los datos importantes presentes en los documentos, para eliminar información redundante y posteriormente presentar una vista sintetizada de los mismos. La generación automática de resúmenes es una tarea muy importante ya que ahorra tener que revisar manualmente grandes volúmenes de textos para encontrar información relevante a nuestras necesidades.

La sumarización se encuentra estrechamente relacionada con otras tareas de la minería de datos, entre las que se pueden mencionar: sistemas preguntas/respuestas, recuperación de pasajes, indexación, recuperación de información, etc.

Existen sistemas de sumarización que generan resúmenes que pueden verse como respuestas argumentadas [28], [29] y por ese motivo es que se dice que la sumarización está vinculada con la recuperación de respuestas (ver sección 6). Este tipo de sistema de sumarización, a partir de una consulta de usuario, produce un resumen enfocándose en el tipo de información requerida y solicitada mediante la consulta. También puede verse la sumarización relacionada con la recuperación de información y extracción de pasajes, ya que la tarea de resumir es precisamente la recuperación de información relevante a ciertas necesidades a partir de una fuente o conjunto de ellas, muchas veces extrayéndose segmentos textuales identificativos de las mismas.

En el proceso de sumarización, el conocimiento extraído del volumen de información a resumir o la propia información a resumir, de alguna manera tiene que almacenarse para que sea recuperada fácilmente y realizarle el tipo de procesamiento definido en cada

[4] Ontología del idioma inglés muy útil en diferentes tareas de procesamiento del lenguaje natural. WordNet es un sistema léxico de referencia donde los sustantivos, verbos, adjetivos y adverbios se agrupan en conjuntos de sinónimos.

[5] En GI se describen las orientaciones semánticas de las palabras de manera general. El GI es un sistema de análisis de texto que con el fin de llevar a cabo su tarea, usa un gran número de categorías (la definición de estas está disponible en http://www.webuse.umd.edu:9090/), cada una denotando la presencia de un trato específico en un término. Las dos categoría principales son Positivo (PRO)/Negativo (CON).

[6] En lo siguiente se utilizará también el término sumarización para referirse a la tarea de resumir

método. En cada uno de los sistemas de sumarización debe haber un módulo encargado del almacenamiento de la información; sin embargo, en muy pocos trabajos se aborda este tema, porque se ha centrado la atención a otro u otros aspectos. La información se almacena generalmente en bases de datos o se crean índices para luego realizar recuperaciones rápidas, en este caso se ha utilizado Lucene[7] como herramienta de indexación [30].

El interés en el desarrollo de sistemas que proporcionen de manera automática resúmenes de documentos o conjuntos de ellos data de varias décadas. La primera publicación registrada en la historia donde se describe la implementación de un sistema automático para la sumarización, es el método de Luhn publicado en 1958 [31]. El próximo trabajo notorio luego del propuesto por Luhn fue desarrollado diez años después por Edmundson [32]. Tras una disminución en el desarrollo de trabajos relevantes sobre el tema, en la década de los 90 se comienza a observar un renovado interés en esta línea, el cual ha ido creciendo notablemente.

En esta sección se presentan un conjunto de variables o criterios necesarios a tener en cuenta al hacer un resumen. Además, se argumentan diferentes clasificaciones de los tipos de resúmenes existentes.

3.1. Aspectos a tener en cuenta

Aunque todos los resúmenes de manera general persigan la sintetización de la información, mostrando el contenido relevante, no todos los resúmenes tienen las mismas necesidades. Existen un conjunto de variables o criterios que es necesario tener en cuenta a la hora de hacer un resumen automático de una fuente, como por ejemplo: la persona a la que va dirigido, la estructuración, el medio de donde se extrae (con sus tipos de datos, su idioma, la diversidad de fuentes textuales), etc. Esto provoca que en ocasiones sea necesario que el resumen se ajuste a las necesidades individuales del momento, y que por lo tanto se generen diferentes resúmenes para una misma fuente de entrada.

Casos como el mencionado anteriormente, se pueden representar en las necesidades de la audiencia de cierta película, donde la fuente de entrada sean las opiniones relacionadas con la misma y exista un grupo de usuarios que necesite un resumen de la trama de la película y otro grupo de usuarios necesite un resumen de las opiniones acerca de la misma. En cada caso la información de la fuente documental se procesa ajustándose a requerimientos.

En general, las propiedades que debe tener un resumen son diferentes para cada persona y situación, lo que hace que las posibilidades sean muy variadas. A continuación se expondrán brevemente algunas de estas variables.

- **Fuente**
  Un factor importante a tener en cuenta es si se trata de resumir un solo documento o de un conjunto de ellos. El procesamiento de varias fuentes puede hacer más compleja la tarea de búsqueda y la selección de información, pero también puede ayudar en la valoración de la redundancia (información repetitiva) presente entre las diferentes fuentes, para encontrar el contenido más identificativo.
- **Idioma**
  El idioma en el que se encuentra la fuente es otro factor importante, ya que es necesario en el tratamiento del lenguaje natural prestarle atención a las características específicas de cada idioma. Por otro lado las fuentes pueden estar escritas en un solo idioma (monolingüe) o en múltiples idiomas (multilingüe) lo cual dificulta aún más el procesamiento.

[7] Lucene es una librería de alto rendimiento que permite añadir funcionalidades de indexación y búsqueda a las aplicaciones facilitando tareas relacionadas con la recuperación de información. (http://lucene.apache.org)

- **Género**
  El género de la fuente es también una característica que las herramientas de resumen automático deben tener en cuenta para generar buenos resúmenes. Existen muchos tipos de géneros que también están muy relacionados con el estilo en que se presenta la información escrita (artículos científicos, noticias, novelas, poemas, editoriales, reportes, mails, blogs, etc.), los cuales tienen peculiaridades a considerar en el momento en que el procesamiento de su contenido se lleva a cabo para extraer los datos relevantes. Por ejemplo, cuando se trata de resumir una noticia, hay que darle especial importancia al titular y al subtitular de esta, además de a las referencias que puedan aparecer en el contenido.
- **Dominio**
  El dominio sobre el que se tiene que hacer un resumen es otra de las variables a considerar. Cuando se procesan documentos de un dominio particular, es preciso aplicar conocimientos y técnicas de tratamiento del lenguaje natural adaptadas y específicas para dicho dominio. Si el dominio utiliza términos médicos, es necesario que las técnicas utilizadas sean capaces de comprender correctamente lo que analizan, lo cual implica el uso de diccionarios de términos médicos y el conocimiento de expertos en la materia para que el tratamiento sea el adecuado.
- **Grado de compresión**
  En dependencia de las exigencias o necesidades de la persona, el sistema o el dispositivo que requiera el resumen, es necesario que este se ajuste a un volumen adecuado. Por ejemplo, un teléfono móvil de tamaño reducido necesita que el resumen sea pequeño, mientras que un resumen destinado a ser visualizado en una página web puede tener mayor tamaño.
- **Contenido a extraer**
  Cuando se trate de decidir la naturaleza de los datos que deben extraerse, es necesario tener en cuenta las necesidades de la persona o el sistema que solicite el resumen. Es posible que de todo el contenido, solo se deseen extraer datos relacionados con precios o con otro tipo de referencias.

3.2. Tipos de resúmenes

Los métodos para la generación automática de resúmenes de textos reportados en la literatura pueden clasificarse de acuerdo a varios criterios. También es posible tener en cuenta criterios de clasificación basados en las variables a considerar expuestas anteriormente, por ejemplo un resumen se puede clasificar atendiendo al lenguaje en: resumen multilingüe o resumen monolingüe. A continuación se expondrán brevemente algunos tipos de resúmenes.

Forma de la información mostrada

Los resúmenes se clasifican, según la forma en que se muestra la información, en estructurados y no estructurados.

- **Resumen estructurado**
  El resumen posee una estructura u organización fácil de entender a partir de algunos aspectos o etiquetas identificados en el contenido a resumir. Estas etiquetas pueden ser tomadas de un conjunto predefinido o extraerse de manera dinámica. Asociado a las etiquetas se puede ofrecer información relacionada, que en dependencia de lo que se quiera mostrar, puede ser muy diversa (ver sección 4.2).
  Como caso particular de lo antes mencionado, existen métodos que estructuran la información empaquetándola en tablas o plantillas. Estas tablas se llenan identificando en el documento o conjunto de documentos a resumir, un grupo de entidades y aspectos específicos predefinidos; así como, extrayendo conocimiento que permita generar algún

tipo de información específica requerida (contabilizar la ocurrencia de determinados rasgos o informaciones, por cientos de información, etc.).

- **Resumen no estructurado**
  La información mostrada en el resumen no se estructura a partir de etiquetas como en el caso anterior, sino que se muestran los resúmenes a partir de uno o varios segmentos de textos (palabras, expresiones, proposiciones, oraciones, etc.). Los resúmenes no estructurados pueden dividirse atendiendo a la forma de obtención de la información mostrada, en resúmenes de extracción y abstractos.
  - Resúmenes de extracción
    En este tipo de resumen, son procesados uno o más documentos con el fin de extraer de ellos información textual de interés. El resumen es conformado finalmente con información textual exacta (segmentos exactos de texto).
  - Resúmenes abstractos
    En este tipo de resumen, la información que se muestra no está constituida en su totalidad por información textual exacta extraída de los textos originales como en el caso anterior. En los resúmenes abstactos a diferencia del anterior, el resumen es generado a partir de la información de interés, haciendo uso de técnicas de procesamiento de la información para su generación en lenguaje natural.

Nivel de procesamiento de la información

Los resúmenes se clasifican, según el nivel al que se procesa la información, en resumen a nivel superficial y resumen a nivel profundo.

- **Resumen a nivel superficial**
  En los métodos superficiales no se utiliza análisis lingüístico complejo. En estos métodos la información es resumida mediante la selección de rasgos extraídos de los documentos. Ejemplos de estos rasgos pueden ser: términos frecuentes, posiciones de términos, oraciones o párrafos, términos contenidos en títulos y subtítulos, términos contenidos en lexicones, términos de dominios específicos o términos presentes en las consultas de usuarios.
- **Resumen a nivel profundo**
  En los métodos profundos es necesario hacer uso de técnicas complejas de procesamiento de información como análisis lingüístico para la generación en lenguaje natural. En estos métodos se hace uso de algún tipo de análisis semántico para encontrar los segmentos importantes en los textos.

Independientemente del nivel al que se procese la información para obtener el resumen, existen varios tipos de pre-procesamientos que se les realiza a los textos, ya sea para obtener un conjunto de términos más representativos de los documentos, reducir dimensionalidad, etc. Entre las operaciones más comunes realizadas en el pre-procesamiento inicial de los documentos se tienen:

- Eliminación de signos de puntuación, espaciados, acentos (en documentos en idioma español), reducción de mayúsculas, reconocimiento del formato del documento (eliminación de etiquetas si es una página HTML), reconocimiento de las palabras, etc. Cuando esta primera etapa es terminada se tiene el texto plano y las palabras identificadas en él.
- Eliminación de las palabras vacías (*stopwords*). Tras este paso se logra una reducción del documento entre un 30 y un 50 por ciento. Las *stopwords* son palabras que carecen de significado a la hora de representar un documento, ya que con seguridad aparecen en todos los documentos de la colección. Como ejemplos de palabras vacías se pueden mencionar: las preposiciones, las conjunciones, los artículos y los pronombres. Aunque algunas de estas pueden considerarse atendiendo al procesamiento lingüístico y semántico que se requiera.

- Extracción de raíces o de lemas (lematización). Los algoritmos de extracción de raíces o lemas obtienen un único término, a partir de diferentes palabras que constituyen variaciones morfológicas con un mismo significado. Por ejemplo, en el caso de los verbos un infinitivo puede ser obtenido a partir de sus conjugaciones verbales. Este proceso además comprende la eliminación de los plurales, de ciertos prefijos, de sufijos, etc. Existen herramientas libres que ofrecen entre otros procesamientos, la lematización de palabras, las cuales ofrecen de forma eficiente buenos resultados de eficacia, entre estas se puede mencionar el TreeTagger[8] y Freling[9].
- Etiquetado gramatical (*Part-of-speech tagging* en inglés y por sus siglas POS tagging o POST). Este etiquetado es el proceso de asignar a cada una de las palabras de un texto su categoría gramatical (sustantivo, adjetivo, artículo, verbo, etc.). Este proceso se puede realizar en base a la definición de la palabra o el contexto en que aparece, por ejemplo su relación con las palabras adyacentes en una frase, oración, o párrafo. Entre las herramientas libres que realizan este procesamiento se pueden mencionar Balie[10], TreeTagger y Freeling.
- Extracción de entidades nombradas. El reconocimiento de entidades nombradas es una tarea que persigue la identificación de elementos en el texto y su clasificación en distintas categorías predefinidas como: nombres de personas, organizaciones, direcciones de lugares, expresiones de tiempo (Fecha/Hora), cantidades numéricas (por ciento, cantidades monetarias, medidas de magnitud expresadas en distintas unidades de medidas: metros, litros, libras, km/h, etc.). Entre las herramientas libres que existen para la detección de entidades nombradas se puede mencionar Freeling y Balie.

No resulta necesario aplicar todas las operaciones antes comentadas, estas se emplean de acuerdo al tipo de procesamiento que vaya a realizar después. Por ejemplo, si se van a detectar entidades nombradas, no es conveniente eliminar previamente *stopwords* como la preposición ¨de¨, ya que dicha eliminación afectaría la recuperación de fechas en el formato ¨día de mes¨.

Distinción a partir de la audiencia

Los resúmenes se clasifican, según su distinción a partir de la audiencia, en genéricos, basados en consultas (*query-based* en inglés) y basados en tópicos (*topic-focused* o *user-focused* en inglés).

- **Resumen genérico** Este tipo de resumen brinda un resultado de interés para una amplia comunidad de lectores donde todos los asuntos abordados son valorados como importantes.
- **Resumen basado en consultas** En este tipo de resumen el resultado es basado en una consulta determinada, dando respuesta resumida sobre un interés particular expresado mediante la misma.
- **Resumen enfocado en tópicos** Este tipo de resumen es ajustado al interés particular de un tipo de usuario. Los sistemas enfocados en tópicos se especializan en un conjunto de temas particulares.

4. Generación automática de resúmenes valorativos

La tarea de sumarización de opiniones se encarga de la elaboración de un resumen a partir de procesar la información contenida en un documento o conjunto de documentos, teniendo en cuenta las opiniones insertadas. Debido a sus múltiples aplicaciones, la generación automática de resúmenes de opiniones es un tema que ha cobrado gran importancia en la actualidad. Métodos para generar de forma automática resúmenes de opiniones se han

[8] (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger)
[9] http://www.lsi.upc.edu/ñlp/freeling/
[10] http: //lordpimpington.com/codespeaks/drupal - 5,1/?q = node/5

aplicado en diferentes contextos: discursos, forums en línea, editoriales, actas de reuniones, comentarios sobre productos comerciales, en encuestas a grupos de personas que permitan a analistas de información crear un perfil de la opinión pública sobre determinados temas de interés, etc.

Los resúmenes de opiniones pueden verse como un tipo particular de resumen de texto, que persigue como propósito específico el análisis evaluativo o crítico de la información y brinda una vista sintetizada a partir de procesar las opiniones insertadas. Los resúmenes de comentarios u opiniones están fuera del ámbito de los actuales sistemas automáticos de resúmenes de textos. Como ha sido tratado anteriormente, los métodos de sumarización no se aplican indistintamente para cualquier situación, sino que dependen de ciertos parámetros y condiciones; por lo tanto, los resúmenes de texto no se pueden aplicar a opiniones directamente. Cuando se pretende hacer un resumen de opiniones es necesario realizar un procesamiento específico a la información para tener en cuenta el componente subjetivo de las opiniones y diversos aspectos relacionados con las mismas como pueden ser: la polaridad, el tipo de actitud, el emisor, el tema sobre el que se opina, etc.

En esta sección se abordan dos pasos esenciales en la sumarización, analizándose cómo han sido realizados estos pasos en textos valorativos. El primero de estos pasos es la detección de tópicos, que en el caso de textos valorativos nos referimos a esta tarea como detección de tópicos valorativos. El segundo paso es la selección de segmentos de textos a incluir en el resumen, que será referido como selección de segmentos de textos valorativos a agregar en el resumen. La detección de tópicos valorativos se analiza a continuación como un tipo de procesamiento de la información valorativa para la sumarización. Por otro lado, la selección de textos valorativos se analiza como un tipo de información a presentar en el resumen (sección 4.2).

4.1. Procesamiento de la información valorativa para la sumarización

Para la sumarización de opiniones, como ha sido mencionado anteriormente, es necesario hacer un análisis de las opiniones insertadas en los textos, ya sea para: detectarlas, clasificarlas, encontrar los temas abordados, detectar los emisores de las opiniones, etc. La calidad del resumen final depende en gran medida de la eficacia conque se realice cada uno de los procesamientos; sin embargo, hay trabajos de sumarización de opiniones que no se encargan, en el proceso de análisis de las opiniones, de la detección y clasificación de las mismas. Existen trabajos de sumarización de opiniones que asumen que las opiniones ya han sido detectadas y clasificadas previamente [29], [15], [33] y centran su atención y novedad en otros aspectos. Algunos trabajos han centrado la atención en la forma de agregar de manera eficiente la información en el resumen (ver sección 4.2) y otros se han concentrado en la detección de los tópicos sobre los que se opina.

La detección de los temas discutidos (tópicos de las opiniones) es un procesamiento de gran importancia en el análisis de las opiniones, el cual ha sido abordado en trabajos de sumarización de opiniones, sobre todo donde los textos evaluativos son comentarios. En la detección de los tópicos de las opiniones, al igual que en la detección y clasificación de las mismas, se pude tener en cuenta o no el dominio de los textos valorativos. La detección de tópicos se ha tratado en trabajos de sumarización de diversas maneras, aplicando técnicas de: agrupamiento [34], [35], [36], minería de asociación [27], patrones de extracción [29], secuencias frecuentes de palabras [37], etc. El problema de detección de tópicos en textos valorativos, exceptuando los comentarios, no ha sido muy explorado [15] y la detección no se ha realizado en profundidad aplicando técnicas de procesamiento de lenguaje natural y en particular de procesamiento de opiniones. A continuación se analizan brevemente varias formas de detectar tópicos en opiniones, haciéndose distinción entre dos tipos de textos valorativos: textos cortos (comentarios) y textos más largos (noticias, blogs, etc.).

La detección de los tópicos (rasgos) en comentarios se ha abordado de varias formas. En el sistema propuesto por Hu y Liu en el 2004 se aplica una técnica no dependiente del dominio para la detección de los rasgos [27]. Hu y Liu aplican un método de minería de asociación [38] para la detección de los rasgos más frecuentes, realizando luego dos tipos de podas para refinar el conjunto de rasgos seleccionados, en las que se eliminan los rasgos redundantes y los que no son compactos [13]. Una técnica dependiente del dominio aplicada para la detección de rasgos, es la de mapear rasgos detectados mediante el método propuesto en [13] a una taxonomía existente, donde se organizan los rasgos de la entidad, este mapeo se realiza a partir de la definición de medidas de semejanza [39], [40]. Entre otras estrategias dependientes del dominio también, se pueden mencionar: la definición de patrones de extracción definidos a partir de la observación de los rasgos en los textos a resumir [41], [29], la creación manual de listados de rasgos basándose en la observación de que los usuarios utilizan las mismas palabras o frases para referirse a los rasgos de ciertas entidades [42], la extracción de rasgos dinámicos y estáticos y su combinación [43], etc.

De manera general, la mayoría de los trabajos no tienen en cuenta en la detección de tópicos, a aquellos temas o rasgos que no estén expresados explícitamente en un segmento donde se emita una opinión sobre él. En todos estos trabajos, la detección de tópicos se realiza a partir de detectar ciertas entidades (rasgos) mencionados en los textos, por lo que la detección de tópicos en estos casos se basa en detección de entidades nombradas. Finalmente, los tópicos detectados en los comentarios están limitados; es por esto que se dijo anteriormente (ver sección 2.2) que los temas abordados en los comentarios son acotados. El conjunto de tópicos está definido por un conjunto de entidades mencionadas explícitamente, lo cual es una deficiencia que provoca pérdida de información, ya que hay opiniones referidas a rasgos implícitos que no se tienen en cuenta durante el análisis. Cabe destacar que en [42] se tratan dos casos de rasgos implícitos, aunque son dos casos bastantes simples, los autores reconocen su importancia y dieron un primer paso en su detección.

Otra deficiencia que está muy relacionada con la anterior, es que usualmente en los comentarios, en el momento de hacer referencia a un rasgo, en vez de mencionarlo se utilizan pronombres. La solución a este problema, denominado anaforismo, no se tiene en cuenta en muchos trabajos, lo cual también provoca pérdida de información relevante en la detección de los tópicos.

La detección de tópicos en textos valorativos de mayor extensión ha sido tratada en varios trabajos. Seki *et al.* en el 2005 [44], [45] aplican un método no dependiente del dominio para encontrar los temas principales. Mediante el algoritmo de agrupamiento Ward [46], se agrupan los párrafos que son previamente representados en el espacio vectorial y finalmente cada grupo representa un tópico. En el método de Chang y Tsai [34] igual que en el trabajo anterior, se utiliza un algoritmo de agrupamiento para detectar los tópicos. En este caso se agrupan por separado las razones positivas y negativas, aplicando el algoritmo de agrupamiento jerárquico FIHC basado en conjuntos de términos frecuentes [47]. En [34] las unidades léxicas en que se expresan las razones son los párrafos, los cuales pueden ser representados de dos maneras: por el conjunto de palabras que contienen o por el conjunto de palabras que contienen y que se encuentran relacionadas con el tema en mayor medida (conjunto previamente determinado). En el sistema propuesto por Zhan *et al.* en el 2008 [37], para la identificación de tópicos, los autores aplican una técnica no dependiente del dominio basada en la detección de secuencias frecuentes de palabras y en clases de equivalencia.

En ninguno de los trabajos anteriores se aprovecha el significado semántico de la información en la detección de tópicos. Al tener en cuenta solo frecuencias de términos se pierden las relaciones que tienen estos en base a su significado. Como solución a esto, en algunos trabajos donde también se agrupan las oraciones mediante la aplicación de algo-

ritmos de agrupamiento, se redefine la función de semejanza entre oraciones teniendo en cuenta la semántica de la información. Bossard *et al.* [48] aplican el algoritmo *fast global kmeans* para agrupar las oraciones, utilizando como función de semejanza una variante de Jaccard [49], donde si dos términos no son iguales se utilizan las relaciones de sinonimia/hiperonimia de Wordnet para tenerlos en cuenta en la función de semejanza. Gaurav y Roshan [50] redefinen la medida de semejanza entre oraciones teniendo en cuenta la semejanza entre palabras basándose en las relaciones que tiene las mismas en Wordnet; por ejemplo, fruta es ancestro de manzana en la jerarquía de conceptos de Wordnet.

Ku *et al.* [51] para la identificación de las oraciones relevantes a los tópicos abordados, se seleccionan términos (palabras) representativos que identifiquen los conceptos principales de un conjunto de documentos. En el caso la estrategia seguida sí depende del dominio y de las características del lenguaje también, ya para la selección de términos representativos se apoyan en la frecuencia de los caracteres que los componen, aprovechando las características especiales del lenguaje chino.

En los trabajos anteriores se tratan los textos valorativos, sin distinción alguna con respecto a textos no valorativos, de esta manera no se aprovechan las propiedades particulares de las opiniones. Las opiniones en ninguno de los trabajos anteriores se representan teniendo en cuenta características o aspectos identificativos de las mismas, sino que se representan a partir de modelos frecuentistas, tratando los textos valorativos como textos en general. Por ejemplo, en los trabajos donde se detectan los tópicos mediante algoritmos de agrupamiento, en la representación de las unidades léxicas a agrupar no se utilizan propiedades de las opiniones como pueden ser la polaridad o la intensidad. Para la correcta detección de los tópicos abordados en textos valorativos es importante definir modelos para representar las opiniones teniendo en cuenta aspectos o características de las mismas y proponer funciones que determinen grados de semejanza entre ellas.

Por otro lado, en algunos de los trabajos donde se detectan los tópicos mediante algoritmos de agrupamiento no se justifica la selección de los algoritmos de agrupamiento aplicados [34]; mientas que en otros simplemente se justifica a partir de la realización de pruebas con varios algoritmos, seleccionando finalmente aquel con el que se obtuvieron mejores resultados en la sumarización [44]. Otros trabajos utilizan como forma de justificación del algoritmo de agrupamiento, la exposición de las ventajas (respecto a algunos aspectos) del algoritmo aplicado con respecto a otros algoritmos de agrupamiento [35], [36]. Sin embargo, ninguna de las justificaciones mencionados anteriormente es válida. No es correcto decir que alguna forma o técnica de agrupamiento de las abordadas en la literatura es mejor que otra, pero sí, algunas son más apropiadas para ciertos problemas o aplicaciones. El conocimiento del dominio y la aplicación en específico, pueden en muchos casos ayudar a determinar qué tipo de grupos se van a formar y qué tipo de agrupamiento se va utilizar con el objetivo de obtener los mejores resultados [52].

4.2. Estructura de los resúmenes de opiniones

Los resúmenes de opiniones pueden ser al igual que los resúmenes de textos clasificados de acuerdo a la forma de la información mostrada en: resúmenes estructurados o no estructurados (sección 3.2).

Los resúmenes de opiniones estructurados poseen una organización fácil de entender a partir de etiquetas identificadas que pueden ser tomadas de un conjunto predefinido o extraerse de manera dinámica. En el caso de las opiniones, el conjunto predefinido puede ser una jerarquía predefinida de rasgos de ciertas entidades [40] y entre las etiquetas que se pueden extraer dinámicamente se puede mencionar: los principales temas comentados en textos valorativos, ya sean rasgos en comentarios [41], [29] o tópicos abordados en otros textos valorativos [34], [37], las fuentes emisoras de las opiniones, etc . Estas etiquetas pueden tenerse organizadas de acuerdo a diferentes criterios y visualizarse de alguna manera: espacio bidimensional [29], estructurados en listas [27], [42], grafos [15], etc. Los criterios de ordenación u organización de las etiquetas pueden ser dependientes o no de la información que se ofrece relacionada con estas, la cual en dependencia de lo que se quiera mostrar, puede ser muy diversa.

Como se dijo anteriormente (sección 3.2), la información que se muestra relacionada con las etiquetas puede ser muy variada y en cada caso relevante a necesidades específicas de información, como ejemplo se podría mencionar: vínculo a los documentos que contienen la etiqueta [37], opinión o conjunto de opiniones representativas de la etiqueta; quizás organizadas o categorizadas de acuerdo a algún criterio (de acuerdo a la polaridad [27], [42], mostrar opiniones más fuertes [40], etc.), resumen de las opiniones asociadas (como un resumen no estructurado de la etiqueta) [44], valores estadísticos [27], [15] (frecuencia, por ciento, polaridad, intensidad, promedio de la polaridad de las opiniones relacionadas con la etiqueta, etc.), etc.

La forma de obtención de la información relacionada con las etiquetas algunas veces es superficial, como es el caso de algunos de los valores estadísticos mencionadas en los ejemplos anteriores, ya que no es necesario realizar procesamiento profundo para determinarlos. Por otro lado, como es un resumen de opiniones, precisamente mucha de la información agregada es fácil de obtener una vez hecho el análisis de las opiniones, pues con este análisis se puede obtener información relacionada con las mismas (la polaridad, la intensidad, la fuente, etc.) y solo bastaría con agrupar las opiniones relacionadas con la etiqueta en positivas y negativas y mostrarlas todas o un subconjunto de ellas [27], determinar el promedio de la polaridad [15], mostrar la opinión de mayor intensidad [40], mostrar el número de opiniones positivas y negativas [42], ordenar las opiniones basados en su fortaleza o intensidad [53] , etc.

Como mismo hay información relacionada con las etiquetas que es fácil de obtener, hay información que para obtenerla se hace necesario realizar un procesamiento más profundo, como es el caso de la selección de segmentos de textos a incluir en los resúmenes de opiniones. En la literatura existen muchos métodos de selección propuestos en diversos trabajos de sumarización de textos basados en extracción [54], [55], [56]. En los métodos de selección en los trabajos de sumarización de textos no se tiene en cuenta en la definición de los criterios de selección, el trabajo específico con opiniones. Por lo tanto, como mismo no son aplicables los métodos para la sumarización de textos a la sumarización de opiniones (ver sección 4), no son aplicables ninguno de los criterios definidos para textos si lo que se quiere es resaltar segmentos valorativos.

En el caso de la sumarización de comentarios donde se detectan rasgos, la mayoría de los resúmenes tienen como objetivo ofrecer una panorámica general sobre las opiniones acerca de los mismos, mostrando los puntos de vista de las personas que ha interactuado con la entidad sobre la que se comenta. En la sumarización de comentarios no es de interés general definir criterios de selección de segmentos de textos valorativos, sino que la información que se agrega relacionada con los rasgos es superficial generalmente. La selección de segmentos de texto a agregar en el resumen no ha sido muy explorado en textos evaluativos, debido a que la mayoría de los trabajos de sumarización de opiniones son de sumarización de comentarios sobre ciertas entidades. Al igual que en la detección

de tópicos valorativos, en la selección de segmentos de textos valorativos a adicionar al resumen, no se han realizado trabajos en los que se aplique en profundidad técnicas de procesamiento de lenguaje natural y en particular de procesamiento de opiniones.

En todos los métodos de selección es necesario definir un criterio para asignar grados de importancia o representatividad a cada segmento (generalmente oraciones). Para determinar esta representatividad se han seguido estrategias estadísticas, en las que se tiene en cuenta valores de determinados aspectos para pesar las oraciones [44], [48] [57], [33]. Estos aspectos pueden ser criticables en cada caso, así como el peso o importancia que se asigna a cada uno de ellos, pero de manera general lo más cuestionable en cada trabajo es la no distinción entre segmentos valorativos y no valorativos en base a las propiedades de las opiniones insertadas en los textos. A continuación se describen de forma general, algunas estrategias de selección propuestas en métodos de sumarización de textos valorativos.

Meishan *et al.* en el 2007 [57] basados en que generalmente los lectores comentan los temas interesantes abordados en los blogs, extraen un conjunto de oraciones representativas de un blog a partir de información valiosa descubierta en el conjunto de comentarios asociados al mismo. La representatividad de las oraciones se determina a partir de la importancia o representatividad de las palabras que contiene. La importancia de una palabra es determinada a partir del conjunto de comentarios, teniendo en cuenta para ello tres estrategias estadísticas conocidas para el pesado de términos (Binario, CF y TF) y se propone una nueva estrategia (ReQuT) que se define en función de tres medidas: RM (*Reader*), QM (*Quotation*) y TM (*Topic*). El peso de las palabras mediante la aplicación de ReQuT se determina como:

$$Rep(w_k) = \alpha \cdot RM(w_k) + \beta \cdot QM(w_k) + \gamma \cdot TM(w_k). \tag{1}$$

En la función anterior se relacionan los comentarios a partir de tres aspectos mencionados en ellos: los lectores o usuarios que comentan (RM), las oraciones textuales (QM) y los principales tópicos comentados (TM), donde α, β, γ son grados de importancia para cada aspecto y cumplen que $\alpha + \beta + \gamma \leq 1$.

Fotis *et al.* en el 2008 [33], para la sumarización de comentarios sobre productos comerciales ofrecidos en la web, presentan una aproximación que selecciona las oraciones basándose en una estrategia que tiene en cuenta la información que ofrecen un conjunto de metadatos relacionados con los comentarios. Para la selección del conjunto de oraciones representativas, primeramente se determina la importancia (R_i) de cada una de ellas mediante un método estadístico que se apoya en un diccionario del dominio (D) creado previamente. Luego, esta importancia es ajustada (aumenta o disminuye) determinando un peso final para las oraciones. Para determinar el peso final (W_i), se realiza un análisis de la influencia que tienen los valores de un conjunto de metadatos relacionados con los comentarios: utilidad o importancia del comentario de un usuario para otros usuarios (*usefulness* en inglés), grado de familiaridad que tiene el usuario que comenta con el producto relacionado (*tech level* en inglés), el tiempo que hace que el usuario que comenta adquirió el producto (*ownership duration* en inglés) y el grado de confiabilidad que se tiene en el comentario de un usuario (*respectability* en inglés). Cabe destacar que la idea del uso de los metadatos es un aporte de este trabajo [33]. La importancia inicial R_i y el peso final (W_i) de las oraciones se determinan como:

$$R_i = \sum_{v_l \notin D} f_{v_l} + 2 \cdot \sum_{v_l \in D} f_{v_l}, \tag{2}$$

$$W_i = R_i + R_i \cdot \sum_{j=1}^{k} M_j, \tag{3}$$

donde v_l son las palabras contenidas en la oración i, f_{v_l} representa la frecuencia de aparición de la palabra v_l en la oración i que en caso de estar presente en el diccionario D se multiplica por 2 para aumentar su significado, k representa la cantidad total de metadatos y M_j define el valor que tiene el metadato j en el comentario donde está contenida la oración i.

En los trabajos anteriores se puede señalar la misma deficiencia que en la detección de tópicos en textos valorativos. En estas definiciones, no se tiene en cuenta la información valorativa que estos ofrecen y se contemplan como textos de manera general sin aprovechar las características de las opiniones en los textos valorativos.

En los métodos que previamente detectan los tópicos abordados es muy común que el proceso de selección de oraciones se apoye en la detección anterior. Zhan *et al.* [37] una vez que extraen los tópicos, por cada uno de ellos se detectan todas las oraciones relevantes iniciales. En caso de que los tópicos sean representados mediante secuencias frecuentes, las oraciones relevantes son aquellas que contienen la secuencia frecuente representativa. Si el tópico es representado mediante clases de equivalencia, las oraciones relevantes del tópico son aquellas que contengan alguna de las secuencias frecuentes que pertenezca a la clase. Luego del proceso inicial de selección de las oraciones, para reducir la redundancia de las mismas, se implementa el método Maximal Marginal Relevance para determinar el valor MMR de una oración y seleccionar finalmente las que van a pertenecer al resumen:

$$MM(S_i) = \lambda Sim(S_i, D) - (1 - \lambda) \max_{S_j \in S} Sim(S_i, S_j), \tag{4}$$

donde D es el conjunto de oraciones relevantes de un tema particular, S es el conjunto de oraciones ya incluidas en el resumen, λ es el parámetro de redundancia que toma valores entre 0 y 1. Como medida de semejanza Sim se adopta la medida del coseno y las oraciones son representadas a partir del modelo de espacio vectorial después de haber eliminado las palabras vacías y haber lematizado la colección de comentarios. La medida MMR de una oración determina la importancia o representatividad de la oración con respecto al tópico en un grado de redundancia (primera parte de la ecuación) y penaliza este valor por el grado de representatividad de la oración con respecto a las oraciones ya seleccionadas para visualizar en el resumen. Este trabajo presenta la misma deficiencia de los anteriores. La medida del coseno es una medida basada en la aparición de los términos, si se aplica esta medida sin ninguna variación para tener en cuenta la semántica de la información se pierde significado semántico, lo cual es una deficiencia de este método.

Carenini *et al.* [40] aprovechan en cierta medida el tratamiento de opiniones en la selección de las oraciones a presentar en el resumen. En el método se forman grupos (solapados) de oraciones por cada rasgo extraído cf, asignando a cada uno, todas las oraciones que contengan evaluaciones sobre el cf representativo del mismo. Luego, se selecciona de cada grupo (ordenados de forma descendente de acuerdo a la cantidad de oraciones que contienen) la oración de mayor valor CF_sum, teniendo en cuenta no seleccionar la misma oración dos veces. El valor de CF_sum de una oración S_k se basa en la cantidad y fortaleza de los cf_s presentes en cada oración:

$$CF\_sum(S_k) = \sum_{ps_i \in eval(S_k)} |ps_i|, \tag{5}$$

donde $eval(S_k)$ representa el conjunto de evaluaciones sobre los rasgos contenidas en la oración S_k. Cada evaluación ps_i es representada por un entero en el rango $[-3, -2, -1, 1, 2, 3]$ que refleja su orientación y fortaleza. El valor 3 representa la opinión más positiva posible y -3 la opinión más negativa posible. En este trabajo a diferencia de los anteriores, si se tiene en cuenta el tratamiento de las opiniones pero aun así cabe señalar que la polari-

dad es un aspecto característico muy importante de las opiniones y aún teniéndolo como información no lo aprovechan en el resumen.

Seki *et al.* [44] aplican un criterio para la selección de oraciones que radica en un peso que se asigna a cada una de ellas, teniendo en cuenta varios rasgos relacionados con la misma. El peso de una oración s es calculado como:

$$W(s) = L(s) \times (a_1 \times Q(s) + a_2 \times H(s) + a_3 \times T(s) + a_4 \times N(s) + a_5 \times S(s)), \qquad (6)$$

donde a los parámetros $a1, ..., a5$ se asignan valores de manera empírica ($a_1 = 0,4$, $a_3 = 1$, $a_4 = 0,4$, $a_5 = 20$, $a_2 = \frac{1}{Nc}$), $L(s)$ es el peso asociado a la oración s basado en la ubicación de la misma en el documento, $Q(s)$ es el número de palabras de contenido en narraciones y títulos que están presentes en la oración s, $H(s)$ es el número de palabras de cabecera que aparecen en s, $T(s)$ son los valores TF en el grupo. $N(s)$ y $S(s)$ son pesos opcionales basados en el análisis específico de opiniones, donde $N(s)$ es la frecuencia de entidades nombradas en s y $S(s) = 1$ si la oración es subjetiva y 0 en otro caso. En este trabajo también se tiene en cuenta el tratamiento con opiniones pero solo para distinguir las oraciones subjetivas de las objetivas.

5. Evaluación

Un aspecto de gran importancia a tener en cuenta en cualquier método, algoritmo o sistema, es la evaluación de los resultados, para contemplar aspectos como la validez, utilidad o correctitud de los mismos; por lo tanto, aunque no se considere que forme parte como tal del proceso de sumarización, es muy común incluir entre los pasos en los que se divide el proceso de resumir, un proceso de evaluación de la calidad de los resúmenes generados. La evaluación y validación de la calidad de los resúmenes es una tarea desafiante. Las medidas de evaluación de la sumarización pueden ser clasificadas en: manuales y automáticas, globales y locales, subjetivas y objetivas, internas y externas, etc.

Inicialmente los métodos de evaluación eran esencialmente manuales; es decir, las personas eran las encargadas de juzgar la calidad de los resúmenes generados. Los métodos de evaluación manuales requieren de personas que se dediquen a leer los resúmenes resultantes y las fuentes originales a partir de las cuales fue generado, para constatar de que en el resumen estén contenidas realmente las ideas principales expresadas en la fuente de información y también para comprobar aspectos como la coherencia y cohesión. Debido al alto costo en tiempo y esfuerzo humano característico de los métodos manuales; así como, de la gran probabilidad de la presencia de errores humanos, surgió la necesidad de desarrollar métodos de evaluación automática. La evaluación automática no se exonera de la necesidad de anotar manualmente algunos aspectos en la preparación de los datos a comparar (semi-automático), pero en esta se disminuye grandemente el alto costo en esfuerzo humano y tiempo.

Las medidas globales describen la calidad del resultado completo de un resumen usando un único valor real [58], [59], [60], mientras que las locales evalúan por separado cada uno de los pasos realizados para la obtención del resumen final [53], [42], [13]. Para la evaluación de la eficacia con que se realiza cada uno de los pasos donde se extrae información (extracción de rasgos, extracción y clasificación de opiniones, detección de tópicos, etc.) se realizan comparaciones con datos anotados de forma manual por un conjunto de expertos. También se pueden efectuar comparaciones con los resultados obtenidos por otros métodos que se tracen los mismo objetivos y que extraigan de manera automática la misma información, teniendo como base, la información extraída de forma manual.

La eficacia se determina generalmente haciendo uso dos métricas ampliamente utilizadas en tareas de recuperación de información, la precisión (*Precision* en inglés), el

recobrado (*Recall* en inglés). El Recall representa la cantidad de información correcta obtenida en los resultados dividido por el total de información obtenida manualmente ($\frac{nij}{ni}$) y la Precision representa la cantidad de información correcta obtenida en los resultados dividido por el número total de información obtenida por el sistema ($\frac{nij}{nj}$). Además de la Precision y el Recall, también se puede utilizar alguna medida que combine las dos anteriores como el F-measure [61]. Las medidas objetivas miden propiedades de los resultados de los resúmenes, por ejemplo, la coherencia, nivel de redundancia, precisión, etc. La presencia de tales propiedades no garantiza que los resultados sean interesantes para el usuario, estas medidas carecen del enlace con los usuarios, aunque su principal atractivo es que son independientes del dominio. Las medidas subjetivas evalúan considerando la utilidad que tienen los resúmenes para los usuarios.

Otra clasificación divide la validación del resumen en medidas internas y externas. Normalmente, sea cual sea la tarea a evaluar donde se clasifican las medidas de evaluación en externas e internas; las medidas externas se basan en un criterio externo que es impuesto sobre los datos, por ejemplo, una estructura previamente especificada que refleje la intuición que se tenga de los resultados ideales para la tarea (*gold standard*). No es posible aplicar estas medidas a situaciones del mundo real donde usualmente no está disponible una clasificación de referencia. Por otra parte, normalmente las medidas internas evalúan considerando solamente los resultados de la tarea evaluada en términos de las propiedades que deben cumplir los resultados ideales, sin tener en cuenta para esto muestras de control.

En el caso de las sumarización también se pueden clasificar las medidas en internas y externas basados en lo explicado anteriormente, donde en la evaluación externa se compara el resumen generado con sumarios referenciales y en la interna no. A diferencia de la evaluación externa, la interna no requiere de sumarios de referencia para comparar el resumen generado, lo cual se debe principalmente al grado de dificultad que tiene en algunos casos la definición de resúmenes de referencias ideales para la comparación. Cuando el resumen persigue un propósito general, donde se brinda igual importancia a todos los asuntos abordados, entonces es posible realizar comparaciones con sumarios de referencia. Sin embargo, como muchos sistemas de sumarización son orientados a satisfacer objetivos específicos, siendo estos objetivos muy variados y la información extraída para conformar el resumen muy ser diversa; no es factible construir resúmenes de referencia para comparar con los obtenidos automáticamente y se procede a evaluar de forma interna los resultados de un sistema generador de resúmenes.

A pesar de lo explicado anteriormente, en la sumarización algunos autores realizan la clasificación en interna y externa teniendo en cuenta la tarea a la que esté orientada el resumen [62] y no a la dependencia de resúmenes referenciales. En la evaluación interna no se tiene en cuenta la audiencia a la que va dirigida; a diferencia de la externa que tiene en cuenta la audiencia [62]. Es conveniente diferenciar de alguna manera estas dos formas de entender lo interno y externo en la evaluación de resúmenes, debido a que los resúmenes dependen en gran medida de a quién van dirigidos y de los objetivos de los mismos. Esta diferenciación permite determinar de forma más sencilla las características a evaluar y la manera de proceder en dicha evaluación.

Para que no exista confusión entre ambas categorizaciones de interna y externa (teniendo en cuenta o no a la audiencia y la dependencia o no de resúmenes referenciales), en este trabajo nos referiremos como: objetivas (evaluación sin tener en cuenta la audiencia) y subjetivas (evaluación teniendo en cuenta la audiencia), a las medidas internas y externas respectivamente, definidas en [62].

Cualquiera de las clasificaciones mencionadas anteriormente son válidas para estructurar los métodos de evaluación de la sumarización, pero algunas son muy generales, por ejemplo: la clasificación en métodos manuales y automáticos [63] y la división en métodos globales y locales. Es más común clasificar los métodos de evaluación teniendo en cuenta la

dependencia o no de resúmenes referenciales o la tarea a la que esté orientada al resumen. Esta última categorización es la que se sigue en este trabajo para estructurar las distintas aproximaciones a la evaluación de resúmenes.

5.1. Evaluación objetiva

Generalmente, este tipo de evaluación se aplica a resúmenes genéricos y en la evaluación se puede brindar importancia a aspectos generales como: la precisión, la coherencia, el nivel de redundancia, la similitud con resúmenes de referencia, lo informativo que pueda resultar el resumen, etc. En estos métodos, generalmente la evaluación se lleva a cabo comparando de forma automática el resumen generado con sumarios de referencia creados por jueces o por sistemas automáticos reconocidos. Estos métodos son automáticos o semi-automáticos y su principal deficiencia es que generalmente en ellos no se tiene en cuenta la semántica de la información, sino que se basan en la co-ocurrencia de elementos léxicos entre el resumen a evaluar y el de referencia. Por ejemplo, si se aplica una medida para comparar dos resúmenes, donde en uno (resumen a evaluar) apareciera el segmento *José llamó a Juan* y en otro (resumen referencial) *Juan llamó a José*, la medida consideraría adecuada la aparición del segmento ejemplificado en el resumen a evaluar; no siendo así, si en el resumen a evaluar apareciera el segmento *Le dió un regalo* y en el referencial *Le ofreció un obsequio*. A continuación se describen un conjunto de métodos donde la evaluación se realiza sin tener en cuenta a la audiencia.

Uno de los métodos de evaluación es el *Basic Elements*[11] [60], [64], usado en las conferencias DUC 2006, 2007 y TAC 2008 (ver sección 6). En este método se divide cada oración del resumen a evaluar en un grupo de unidades semánticas mínimas denominadas *Basic Elements* (BEs). Luego de la obtención de los BEs en el resumen a evaluar, se asigna a cada uno de ellos una puntuación teniendo en cuenta su presencia en los resúmenes de referencia y se determina la similitud entre BEs. Finalmente, el método ofrece una puntuación global a partir de una lista evaluada de BEs.

Otro de los métodos de evaluación más conocidos es el *Pyramid Method* [58], incluido en el 2005 y 2006 en las conferencias DUC. Este método se basa en la detección de *Summarization Content Units* (SCUs) en el resúmen a evaluar, las cuales se dividen en filas horizontales en el interior de una pirámide. La pirámide tendrá tantas filas como resúmenes de referencia se estén tomando para la comparación y en cada fila estarán ubicadas las SCUs con el mismo peso que el número de la fila que ocupan, comenzando desde la fila 1. El grado de relevancia (peso de las SCUs) asignado a cada SCU se determina como la cantidad de resúmenes de referencia que la contienen. Por ejemplo, en la fila 1 que es la de la base de la pirámide están las SCUs de peso 1 y así sucesivamente en asenso, quedando en la cima las SCUs de peso igual a la cantidad de resúmenes referenciales. De esta manera en las filas superiores estarán las SCUs más relevantes y la relevancia gobal de un resumen se determina como la suma de las relevancias de todas sus SCUs.

Como uno de los métodos de evaluación más utilizados en la actualidad se puede citar un conjunto de medidas propuestas en [59] denominadas ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*). Estas medidas han sido utilizadas en la serie de conferencias DUC (2002-2007) y TAC 2008. Estas métricas se basan en la co-ocurrencia de n-gramas entre los resúmenes a evaluar y los de referencia. A continuación serán descritas varias de las variantes de *ROUGE*[12]:

La medida *ROUGE-N* es útil cuando se tiene más de un resumen como base para la comparación con el resumen obtenido automáticamente. ROUGE-N toma valores entre 0 y 1 y mientras mayor sea su valor, mejor es la calidad del resumen evaluado. Esta medida

[11] http://www.isi.edu/ cyl/BE/
[12] http://berouge.com/

se calcula como:

$$ROUGE - N(s) = \frac{\sum_{r \in R} \langle \Phi_n(r), \Phi_n(s) \rangle}{\sum_{r \in R} \langle \Phi_n(r), \Phi_n(r) \rangle}, \tag{7}$$

donde:

- s es el resumen generado automáticamente por algún sistema y $R = \{r_1, r_2, ..., r_m\}$ representa el conjunto de resúmenes de referencia.
- $\Phi_n(r)$ es un vector binario cuyas componentes representan los *n-gramas* contenidos en el resumen r, la componente i-ésima toma como valor 1 si el *n-grama* i-ésimo está contenido en r y en otro caso toma valor 0.
- $\langle i, j \rangle$ denota el producto del vector i por la transpuesta de j.
- $\langle \Phi_n(r), \Phi_n(s) \rangle$ representa el máximo número de *n-gramas* que co-ocurren entre los resúmenes s y r.

ROUGE-N_{multi} es una medida alternativa a la anterior donde se toma del conjunto de resúmenes por referencia el resumen que más se asemeja al obtenido automáticamente. ROUGE-N_{multi} se calcula como:

$$ROUGE - N_{multi}(s) = \max_{r \in R} \frac{\langle \Phi_n(r), \Phi_n(s) \rangle}{\langle \Phi_n(r), \Phi_n(r) \rangle} \tag{8}$$

En *ROUGE-L* se aplica el concepto de subsecuencia común más larga (*LCS* por sus siglas en inglés). Una subsecuencia de una cadena $c = c_1, c_2, ..., c_n$ es una cadena de forma $c_{i_1}...c_{i_n}$, donde $1 \leq i_1 < ...i_n \leq n$. Esta medida se basa en que mientras mayor sea la longitud de la mayor subsecuencia común entre las oraciones de dos resúmenes, estos serán más similares. ROUGE-L queda definida con base en la medida F-measure, donde la precisión (R_{LCS}) y el recobrado (P_{LCS}) se determinan teniendo en cuenta el concepto de subsecuencia común más larga. ROUGE-L se calcula como:

$$ROUGE - L(s) = \frac{(1 + \beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}}, \tag{9}$$

donde R_{LCS} y P_{LCS} se determinan como: $\frac{\sum_{i=1}^{u} LCS(r_i, s)}{\sum_{i=1}^{u} |r_i|}$ y $\frac{\sum_{i=1}^{u} LCS(r_i, s)}{|s|}$ respectivante, siendo:

- $r_1, r_2, ..., r_n$ el conjunto de oraciones en los resúmenes de referencia contenidos en R y s el resumen obtenido de forma automática considerado como una concatenación de oraciones.
- $|r_i|$ el tamaño de la oración r_i.
- $|s|$ la cantidad de oraciones que contiene s.
- $LCS(x, y)$ el tamaño de la LCS entre las oraciones x y y.
- β el parámetro de balance entre la precisión y el recobrado tomado usualmente grande.

La medida *ROUGE-S* también está definida con base en F-measure pero en este caso la precisión y el recobrado se determinan inspirados en la idea de la medida ROUGE-N definida con anterioridad, tomando 2 como valor para n. La medida ROUGE-S de define como:

$$ROUGE - S(s) = \frac{(1 + \beta^2) R_s P_s}{R_s + \beta^2 P_s} \tag{10}$$

donde R_s y P_s se determinan como: $\frac{\sum_{i=1}^{u} \langle \Psi_2(r_i), \Psi_2(s) \rangle}{\sum_{i=1}^{u} \langle \Psi_2(r_i), \Psi_2(r_i) \rangle}$ y $\frac{\sum_{i=1}^{u} \langle \Psi_2(r_i), \Psi_2(s) \rangle}{\langle \Psi_2(s), \Psi_2(s) \rangle}$ respectivamente, siendo:

- $\Phi_2(r)$ un vector binario cuyas componente representan parejas de palabras, la componente i-ésima toma como valor 1 si la subsecuencia i-ésima está contenida en r y en otro caso toma valor 0.

Como un ejemplo de evaluación objetiva manual, en algunos trabajos [45], [65], [44], [36], se definen un conjunto de aspectos o criterios a ser evaluados por un conjunto de usuarios y varios valores para categorizar cada uno de estos aspectos. Los valores pueden ser, por ejemplo, enteros entre 1 y 5, siendo 1 la calidad mínima y 5 la máxima asignada a cada categoría. Como ejemplo de los criterios de evaluación definidos se puede mencionar la calidad lingüística, valorando: la calidad de la gramática en el lenguaje, la no redundancia de los datos, la claridad referencial, la estructura y coherencia, etc.

5.2. Evaluación subjetiva

En esta evaluación de los resúmenes, se tiene en cuenta a los usuarios que harán uso del mismo, contemplando: la utilidad que tiene para ese tipo de usuarios en específico, los objetivos del resumen y el tipo de información que se muestra. Esta aproximación se aplica generalmente a resúmenes específicos orientados a resumir aspectos determinados. En estos métodos, debido al grado de dificultad que tiene en algunos casos la definición de resúmenes de referencias ideales se crean estrategias para no depender de resúmenes referenciales para la evaluación. Pueden haber muchas propuestas y formas de este tipo de evaluación, se da libertad a los autores que evalúen sus sistemas proponiendo nuevas estrategias y métodos ajustándolos al tipo de resumen particular. Entre este tipo de estrategias de evaluación hay métodos manuales y automáticos. A continuación se comentan algunos de estos métodos.

Las estrategias basadas en cuestionarios (test de preguntas), son métodos de evaluación manual bastante conocidos, en algunos trabajos se realizan comparaciones entre respuestas a un conjunto de interrogantes ofrecidas por un grupo de jueces [66]. Los evaluadores deben responder las preguntas de acuerdo a la información que tengan luego de leer con anterioridad el resumen generado y la fuente original a partir de la que se obtuvo el mismo. Luego se comparan las respuestas para medir su nivel de concordancia en el resumen y en la fuente original.

En [67] se ofrece a los jueces la fuente original marcada, resaltando un conjunto de frases que se refieren al asunto que debe abordar el resumen, de esta manera los jueces tienen una noción de la información de la fuente original que debe contener el resumen. En otra aproximación [68], en vez de resaltar frases en los textos, se ofrece un lista de conceptos claves a los que se supone que se deba hacer alusión en los resúmenes. Para la evaluación, los jueces deben relacionar conceptos ofrecidos y observar su presencia o no en los resúmenes.

Además de estrategias para la evaluación subjetiva manuales como las comentadas anteriormente, también existen métodos automatizados. Un ejemplo de estrategia de evaluación subjetiva automática utilizada en varios trabajos [42], [27], consiste en evaluar por separado los resultados de cada una de las sub-tareas en las que es dividido el problema. Los autores han evaluado sus sistemas de resúmenes teniendo en cuenta: la eficacia de la extracción de rasgos, la eficacia de la extracción de las opiniones y la exactitud en la predicción de las orientaciones semánticas de las mismas. Para la evaluación, los autores crean corpus; por ejemplo, en el trabajo de Hu y Liu [13] la experimentación fue conducida utilizando comentarios de compradores de cinco productos electrónicos adquiridos de Amazon.com y C—net.com. Los comentarios fueron revisados y etiquetados manualmente, detectando los rasgos sobre los cuales los clientes expresan sus opiniones e identificando la orientación de las oraciones que contienen opiniones. Zhuang *et al.* [42] como conjunto de datos seleccionan los 100 primeros comentarios acerca de 10 películas de IMDB (*Internet Movie Data Base*) que cubren diferentes géneros. Los comentarios son procesados manualmente para etiquetarlos, asignando las clases a los rasgos y a las opiniones respectivamente.

En los trabajos anteriores se determinan la *Precision* y *Recall* de los sistemas pero no se comparan con trabajos previos porque no hay otros trabajos que se puedan tomar como

líneas bases para comparaciones. Popescu and Etzioni [53] utilizan para sus experimentos un total de 1621 comentarios sobre 7 clases de productos, en los que están contenidos el conjunto definido por Hu y Liu. En este trabajo los autores se comparan con el sistema propuesto por Hu y Liu ya que es el más adecuado por realizar un procesamiento similar. El sistema propuesto por Popescu and Etzioni demostró ser mejor que el de Hu y Liu en cuanto a Precision en todas las tareas evaluadas, 22 por ciento más preciso en la identificación de rasgos, 0.08 por ciento más preciso en la detección de opiniones y 0.06 más preciso en la determinación de la polaridad.

6. Foros de evaluación

El NIST (National Institute of Standards and Technology) ha organizado en esta década una serie (2001-2007) de conferencias o evaluaciones denominadas DUC (Document Understanding Conference), seguidas por TAC (Text Analysis Conference) que comenzó en el 2008[13]. Los organizadores de estas conferencias cada año anuncian tareas de sumarización como: sumarización orientada a tópicos, sumarización actualizada, etc. Los participantes producen sus resúmenes que son evaluados manual y automáticamente, para lo cual los organizadores crean corpus de validación. Para la evaluación también se da libertad a los participantes a que evalúen sus sistemas proponiendo nuevas estrategias y métodos de validación. Después de la evaluación de los sistemas, se muestran los resultados y se debaten las técnicas aplicadas.

La evaluación de las opiniones comenzó en el 2006 en TREC (Text REtrieval Conference), donde se incluyó como tarea la detección de opiniones y ha continuado con algunas variaciones en TREC 2007 y 2008. En TREC[14] se han presentado varios trabajos de sumarización de opiniones [69], [70], [71].

El objetivo de la detección de opiniones en TREC es descubrir lo que piensan los *bloggers* acerca de un tópico dado. A los participantes en la tarea se les provee de un corpus de blogs, así como de un conjunto de preguntas del tipo ¿Qué piensan los *bloggers* acerca de X?. Los participantes por su parte deben ofrecer un conjunto de blogs como respuesta a esta interrogante. En TREC a partir del 2007, la tarea de detección de opiniones incluye la polaridad, proponiéndose encontrar las opiniones positivas/negativas acerca de X. Para la detección de opiniones en TREC no es necesario realizar un procesamiento profundo, por lo que esta tarea en estas conferencias es de granulado grueso. La detección de opiniones, aunque sea de granulado grueso en TREC es considerada como la predecesora de tareas relacionadas con opiniones en TAC 2008.

En TAC 2008 se presentan dos tareas acerca de opiniones muy relacionadas entre si: la recuperación de respuestas valorativas y la summarización de textos valorativos, en las cuales se aplican técnicas de granulado fino para detectar las opiniones. En esta conferencia se utilizó para la evaluación la colección Blog06 de TREC 2006 que contiene alrededor de 3.2 millones de *blogs post* referentes a 100,000 blogs.

La tarea de sumarización de opiniones en TAC 2008 es una extensión natural de la tarea de recuperación de respuestas valorativas. A los sistemas de sumarización de opiniones participantes se les provee de un tema X y de una o dos preguntas argumentativas sobre X. Las preguntas argumentativas son aquellas cuyas respuestas no son entidades nombradas, sino que es necesario ofrecer argumentos en la respuesta, por ejemplo: ¿Por qué las personas se sienten atraídas por X?, ¿Por qué las personas no sienten atracción por X?, ¿Qué piensan los personas acerca de X?, etc. Por su parte, los sistemas participantes deben producir un resumen del tema dado que sintetice las respuestas a las preguntas provistas. La tarea

[13] http://duc.nist.gov/
[14] http://trec.nist.gov/

de sumarización de opiniones en TAC 2008 asigna importancia tanto al contenido del resumen como a su facilidad de palabra o fluidez y a su legibilidad. Los resúmenes son evaluados automáticamente utilizando el *Pyramid Method* y manualmente a partir de 5 aspectos: gramática (grammaticality en inglés), poco redundancia (non-redundancy en inglés), estructura/coherencia (structure/coherence en inglés), legibilidad global (overall readability en inglés) y sensibilidad global (contenido y legibilidad).

7. Recapitulación

En este trabajo se ha presentado un estudio sobre la generación automática de resúmenes de opiniones como tarea derivada de la generación automática de resúmenes de textos y de la minería de opiniones. Se presentaron una serie de conocimientos relacionados tanto con opiniones, como con la generación de resúmenes de textos. Se centró particular atención en dos pasos esenciales de la sumarización de opiniones, la detección de tópicos valorativos y la selección de segmentos de textos valorativos a incluir en el resumen.

La detección de tópicos valorativos se analiza en este trabajo como un tipo de procesamiento de la información valorativa para la sumarización de opiniones, haciéndose distinción entre dos tipos de textos valorativos: textos cortos y textos más largos. La selección de segmentos de textos valorativos a incluir en el resumen como se abordó en la sección 4.2 es un tipo de información a agregar que para obtenerla es necesario realizar un procesamiento más profundo, a diferencia de otros casos donde la obtención se realiza de forma superficial. Tanto en la detección de tópicos valorativos, como en la selección de textos valorativos a mostrar en el resumen se señalaron problemas detectados en los distintos trabajos analizados que se resumirán de forma breve a continuación.

Los tópicos detectados en los comentarios están limitados y la detección de tópicos se realiza a partir de detectar ciertas entidades (rasgos) mencionados en los textos, por lo que la detección de tópicos en estos casos se basa en la detección de entidades nombradas. La mayoría de los trabajos no tienen en cuenta en la detección de tópicos, a aquellos temas o rasgos que no estén expresados explícitamente en un segmento donde se emita una opinión sobre él. El conjunto de tópicos está definido por un conjunto de entidades mencionadas explícitamente, lo cual es una deficiencia que provoca pérdida de información, ya que hay opiniones referidas a rasgos implícitos que no se tienen en cuenta durante el análisis. Cabe destacar que en [42] se tratan dos casos de rasgos implícitos, aunque son dos casos bastantes simples, los autores reconocen su importancia y dieron un primer paso en su detección.

En la detección de tópicos en textos valorativos largos se señaló como deficiencia que se tratan los textos valorativos sin distinción alguna con respecto a textos no valorativos, de esta manera no se aprovechan las propiedades particulares de las opiniones como pueden ser la polaridad y la intensidad. Las opiniones en ninguno de los trabajos analizados se representan teniendo en cuenta características o aspectos identificativos de las mismas.

En la detección de tópicos a partir de la aplicación de algoritmos de agrupamiento se señaló que ninguna de las justificaciones en las que se basan para seleccionar los algoritmos son correctas. En caso de utilizar agrupamiento en la detección de tópicos para la sumarización, es necesario tener en cuenta los objetivos de este tipo de aplicación (sumarización) y en base a esto utilizar algún algoritmo que se ajuste o definir nuevos métodos de agrupamiento. Estos nuevos métodos de agrupamiento para la detección de tópicos para la sumarización, en el caso de las opiniones (sumarización de opiniones) debe aprovechar las características que tienen las mismas (deficiencia anterior) para representar las opiniones teniendo en cuenta dichas características y proponer funciones que determinen grados de semejanza entre ellas.

Al igual que en la detección de tópicos valorativos, en la selección de segmentos de textos valorativos a adicionar al resumen, no se han realizado trabajos en los que se aplique en profundidad técnicas de procesamiento de lenguaje natural y en particular de procesamiento de opiniones. De manera general lo más cuestionable en cada trabajo es la no distinción entre segmentos valorativos y no valorativos en base a las propiedades de las opiniones insertadas en los textos.

Finalmente, tanto los problemas detectados en la detección de tópicos valorativos en textos largos, como en la selección de segmentos de textos valorativos subyacen en el trabajo con opiniones y en el poco aprovechamiento de las características particulares de los textos valorativos. La identificación de estas características es importante, tanto para la correcta detección de los tópicos abordados en textos valorativos, como para la selección de segmentos de textos a incluir en el resumen, ya que se podrían definir modelos que representen las opiniones teniendo en cuenta estos aspectos y luego proponer funciones que determinen grados de semejanza entre las opiniones. Aún queda mucho que hacer en el futuro, la minería de opiniones es una disciplina que ofrece muchos desafíos, al igual que cada uno de los procesamientos que se pueden realizan a partir del análisis de las opiniones, entre ellos la generación automática de resúmenes valorativos.

8. Conclusiones

La minería de opiniones es una disciplina que ha cobrado auge en la actualidad, provocado por un notable incremento en el interés del hombre en procesar información sobre diferentes contextos donde se expresan opiniones: discursos, forums en linea, editoriales, actas de reuniones, comentarios sobre productos comerciales, encuestas, etc. Junto con la minería de opiniones también han cobrado importancia diversos procesamientos que se pueden realizar a los textos teniendo en cuenta la información valorativa que contienen, entre los cuales se encuentra la generación automática de resúmenes de opiniones.

En este trabajo se presentó un estudio sobre la generación automática de resúmenes de opiniones como tarea derivada de la generación automática de resúmenes de textos y de la minería de opiniones. Como se pudo apreciar, se presentaron una serie de conocimientos relacionados tanto con opiniones, como con la generación de resúmenes de textos. Se centró particular atención en dos pasos esenciales de la sumarización de opiniones, la detección de tópicos valorativos y la selección de segmentos de textos valorativos a incluir en el resumen; señalándose problemas detectados en los distintos trabajos analizados. Por otro lado, se presentaron un conjunto de medidas de evaluación de la calidad de los resúmenes y se identificaron los principales foros de evaluación que contemplan procesamientos que concurren en el análisis de las opiniones.

Referencias bibliográficas

1. Hatzivassiloglou, V., Mckeown, K.R.: Predicting the semantic orientation of adjectives. In: Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Association for Computational Linguistics (1997) 174–181
2. Hatzivassiloglou, V., Wiebe, J.: Effects of adjective orientation and gradability on sentence subjectivity. In: 18th International Conference on Computational Linguistics (COLING-2000). (2000)
3. Kamps, J., Marx, M., Mokken, R., de Rijke, M.: Using wordnet to measure semantic orientation of adjectives. (2004) 1115–1118
4. Riloff, E., Wiebe, J., Wilson, T.: Learning subjective nouns using extraction pattern bootstrapping. (2003) 25–32
5. Takamura, H., Inui, T., Okumura, M.: Extracting semantic orientations of words using spin model. In: Proceedings of ACL-05, 43rd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (2005)

6. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language Resources and Evaluation (formerly Computers and the Humanities) **39**(2/3) (2005) 164–210
7. Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., Jurafsky, D.: Automatic extraction of opinion propositions and their holders. In: Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications. (2004)
8. Wiebe, J., Bruce, R., Bell, M., Martin, M., Wilson, T.: A corpus study of evaluative and speculative language. In: Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue (SIGdial-2001). (2001) 186–195
9. Wiebe, J.: Tracking point of view in narrative. Computational Linguistics **20**(2) (1994) 233–287
10. Nigam, K., Hurst, M.: Towards a robust metric of polarity. In Shanahan, J.G., Qu, Y., Wiebe, J., eds.: Computing Attitude and Affect in Text: Theories and Applications. Number 20 in the Information Retrieval Series, Springer (2006)
11. Wiebe, J., Riloff, E.: Creating subjective and objective sentence classifiers from unannotated texts. In: Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005). (2005) 486–497
12. Wiebe, J., Wilson, T., Bell, M.: Identifying collocations for recognizing opinions. In: Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation. (2001) 24–31
13. Hu, M., Liu, B.: Mining opinion features in customer reviews. (2004)
14. Kim, S.M., Hovy, E.: Determining the sentiment of opinions. (2004) 1267–1373
15. Stoyanov, V.S.: Opinion summarization: Automatically creating useful representations of the opinions expressed in text. (2009)
16. Stoyanov, Cardie, C.: Toward opinion summarization: Linking the sources. (2006)
17. Stoyanov, Cardie, C.: Topic identification for fine-grained opinion analysis. (2008)
18. Kaplan, N.: Nuevos desarrollos en el estudio de la evaluación en el lenguaje: La teoría de la valoración. (2004) 52–78
19. Kaplan, N.: La construcción discursiva del evento conflictivo en las noticias por televisión. (2007)
20. Turney, P.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. (2002) 417–424
21. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. (2003) Available at http://www2003.org.
22. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL. (2004) 271–278
23. Bruce, R., Wiebe, J.: Recognizing subjectivity: A case study of manual tagging. Natural Language Engineering (1999) 187–205
24. Liu, B.: Chapter 11. opinion mining. In: UIC ACL-07. (2007)
25. Turney, P., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS) (2003) 315–346
26. Esuli, A., Sebastiani, F.: Determining the semantic orientation of terms through gloss analysis (2005)
27. Hu, M., Liu, B.: Mining and summarizing customer reviews. (2004) 168–177
28. Feiguina, O., Lapalme, G.: Query-based summarization of customer reviews. In: Canadian Conference on AI. (2007) 452–463
29. Fujii, A., Ishikawa, T.: A system for summarizing and visualizing arguments in subjective documents: Toward supporting decision making. In: Proceedings of the Workshop on Sentiment and Subjectivity in Text, Association for Computational Linguistics (2006) 15–22
30. Cruz, F., T.J.O.J.E.F.: The italica system at tac 2008 opinion summarization task. In: In Proceedings of the Text Analysis Conference (TAC). (2008)
31. Luhn, H.P.: The automatic creation of literature abstracts. IBM Journal of Research and Development (1958) 155–164
32. Edmundson, H.P.: New methods in automatic extracting. Journal of the Association for Computing Machinery (1969) 264 – 285
33. Kokkoras, F., Lampridou, E., Ntonas, K., Vlahavas, I.: Mopis: A multiple opinion summarizer. In: SETN '08: Proceedings of the 5th Hellenic conference on Artificial Intelligence. (2008) 110–122
34. Chang, C.H., Tsai, K.C.: Aspect summarization from blogsphere for social study. In: ICDM Workshops. (2007) 9–14
35. Esaú Villatoro-Tello, Luis Villaseñor-Pineda, M.M.y.G.y.D.P.A.: Multi-document summarization based on locally relevant sentences. 8th Mexican International Conference on Artificial Intelligence (2009) 227–238
36. Pons-Porrata, A., Ruiz-Shulcloper, J., Llavori, R.B.: A method for the automatic summarization of topic-based clusters of documents. In: CIARP. (2003) 596–603
37. Zhan, J., Loh, H.T., Liu, Y.: Summarizing online customer reviews automatically based on topical structure. (2007) 245–256

38. Liu, B., Hsu, W., Ma, Y. In: Knowledge Discovery and Data Mining. (1998) 80–86
39. Carenini, G., Ng, R.T., Zwart, E.: Extracting knowledge from evaluative text. In: K-CAP '05: Proceedings of the 3rd international conference on Knowledge capture. (2005) 11–18
40. Carenini, G., Ng, R., Pauls, A.: Multi-document summarization of evaluative text. In: Proceedings of the European Chapter of the Association for Computational Linguistics (EACL). (2006) 305–312
41. Puspesh, K.: Multi-document update and opinion summarization. (2008)
42. Zhuang, L., Jing, F., Zhu, X., Zhang, L.: Movie review mining and summarization. In: Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM). (2006)
43. S. Blair-Goldensohn, K. Hannan, R.M.T.N.G.A.R., Reynar, J.: Building a sentiment summarizer for local service reviews. (2008)
44. Seki, Y., Eguchi, K., Kando, N., Aono, M.: Multi-document summarization with subjectivity analysis at DUC 2005. In: Proceedings 2005 Document Understanding Conference (DUC-2005). (2005)
45. Seki, Y., Eguchi, K., Kando, N., Aono, M.: Multi-document summarization reflecting information needs on subjectivity. In: The 5th NTCIR Workshop. (2005)
46. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association (1963) 236–244
47. Fung, B.C.M., Wang, K., Ester, M.: Hierarchical document clustering using frequent itemsets. In: Proc. of the 3rd SIAM International Conference on Data Mining (SDM). (2003) 59–70
48. Bossard, A., G.M., Poibeau, T.: Description of the lipn systems at tac 2008: Summarizing information and opinions. In: In Proceedings of the Text Analysis Conference (TAC). (2008)
49. Wan, X., Peng, Y.: A measure based on optimal matching in graph theory for document similarity. In LNCS 3411 (2005) 227–238
50. Gaurav Aggarwal, Roshan Sumbaly, S.S.: Updatesummarization. In: ClassProject,NaturalLanguageProcessing,Stanford. (2009)
51. Ku, L.W., Li, L.Y., Wu, T.H., Chen, H.H.: Major topic detection and its application to opinion summarization. In: Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR). (2005) 627–628
52. García, L.A.: Agrupamiento basado en el concepto de intermediación diferencial y la aplicación de la teoría de los conjuntos aproximados para valorar resultados de agrupamientos. In: Universidad central Marta Abreu de las villas. (2008)
53. Popescu, A.M., E.O.: Extracting product features and opinions from reviews. (2005) 339–346
54. Wang, D., Zhu, S., Li, T., Gong, Y.: Multi-document summarization using sentence-based topic models. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. (2009) 297–300
55. Aliguliyev, R.M.: A new sentence similarity measure and sentence based extractive technique for automatic text summarization. Expert Syst. Appl. (2009) 7764–7772
56. Wan, X., Yang, J., Xiao, J.: Manifold-ranking based topic-focused multi-document summarization. In: IJCAI. (2007) 2903–2908
57. Hu, M., Sun, A., Lim, E.P.: Comments-oriented blog summarization by sentence extraction. In: CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. (2007) 901–904
58. Nenkova, A., Passonneau, B.: Evaluating content selection in summarization: The pyramid method. In: In Proceedings of the HLT-NAACL Conference. (2004) 145–152
59. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: In Proceedings of the ACL Text Summarization Workshop. (2004) 2903–2908
60. Hovy, E.; Lin, C.Y.Z.L.: Evaluating duc 2005 using basic elements. In: In Proceedings of the Document Understanding Conferences (DUC). Vancouver. (2005) 1–6
61. Peñas, A., l.R.S.V.V.F.: Overview of the answer validation exercise 2006. (2006) 257–264
62. Amigó, E.: Síntesis de información: desarrollo y evaluación de un modelo interactivo. In: Madrid, Universidad Nacional de Educación a Distancia. [Tesis doctoral]. (2006)
63. da Cunha Fanego, I.: Hacia un modelo lingüístico de resumen automático de artículos médicos en español (2008)
64. Hovy, Eduard, C.Y.L.L.Z., Fukumoto, J.: Automated summarization evaluation with basic elements. In: In Proceedings of LREC. (2006)
65. Carenini G, C.J.: Extractive vs. nlg-based abstractive summarization of evaluative text: The effect of corpus controversiality. In: International Conference on Natural Language Generation. (2008)
66. Morris, A. H.; Kasper, G.M.A.D.A.: The effects and limitations of automated text condensing on reading comprensión performance. In: Information Systems Research. (1992) 17–35
67. Mani, I.; House, D.K.G.H.L.O.L.F.T.C.M.B.: The tipster summac text summarization evaluation: Final report. technical report. In: DARPA. (1998)
68. Saggion, H.; Lapalme, G.: Concept identification and presentation in the context of technical text summarization. In: In Proceedings of the ANLP/NAACL Workshop on Automatic Summarization. Seattle. (2000)
69. I. Ounis, M. de Rijke, C.M.G.M., Soboroff, I.: Overview of trec-2006 blog track
70. C. Macdonald, I.O., Soboroff, I.: Overview of trec-2007 blog track
71. I. Ounis, C.M., Soboroff, I.: Overview of trec-2008 blog track

RT_010, marzo 2010
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: DCG Matilde Galindo Sánchez
RNPS No. 2143
ISSN 2072-6260
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. No. 21812 e/218 y 222, Rpto. Siboney, Playa;
Ciudad de La Habana. Cuba. C.P. 12200
Impreso en Cuba