Subjetividad y Procesos Cognitivos, Vol. 14, Nº 2, 2010
Pág. 113-126, ISSN impreso: 1666-244X, ISSN electrónico: 1852-7310

Resúmenes de textos: nuevos retos en la Web 2.0*

Text summaries: new challenges in Web 2.0

Elena Lloret y Manuel Palomar**

Resumen

Este artículo presenta un estudio preliminar de los fenómenos presentes en la Web 2.0, concretamente en *blogs* y cómo se reflejan en los correspondientes resúmenes generados. El principal objetivo es cuantificar en qué medida dichos fenómenos están presentes tanto en los *blogs* como en los resúmenes. La presencia de estos fenómenos en los resúmenes tiene como consecuencia directa la disminución de la calidad de estos, en criterios como la corrección gramatical o la coherencia de los resúmenes. Los resultados preliminares obtenidos muestran que los nuevos géneros textuales derivados de la Web 2.0 contienen un alto número de rasgos lingüísticos típicos que es necesario tratar con métodos y herramientas adecuadas para que dichos rasgos no se propaguen a otras tareas del Procesamiento del Lenguaje Natural, en concreto, en este estudio, a los resúmenes de textos. Además, se proponen posibles soluciones para abordar el problema, con la finalidad de ayudar a que la calidad de los resúmenes no se vea afectada debido a la presencia de estos fenómenos.

Palabras clave: resúmenes de texto, corrección gramatical, rasgos lingüísticos.

Summary

This article presents a preliminary study of the phenomena present in Web 2.0, specifically in blogs and how they are reflected in the corresponding generated summaries. The main objective is to provide a measure of the occurrence of these phenomena in both blogs and summaries. The presence of these phenomena in the summaries has as a direct consequence in their diminishing quality in terms of grammar accuracy or coherence. Preliminary results obtained show that the new

* Esta investigación ha sido financiada a través de una beca FPI (BES-2007-16268) concedida por el Ministerio de Ciencia e Innovación del Gobierno de España, que a su vez está adscrita al proyecto TEXT-MESS (TIN2006-15265-C06-01) también financiado por el Gobierno de España. Además, ha sido parcialmente financiada por el proyecto PROMETEO "Desarrollo de Técnicas Inteligentes e Interactivas de Minería de Textos" (2009/119) de la Generalitat Valenciana.

** Grupo de Investigación "Procesamiento del Lenguaje Natural y Sistemas de Información", Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante. Dirección: Apartado de correos 99, E-03080, Alicante, España. E-mail: a elloret@dlsi.ua.es, mpalomar@dlsi.ua.es.

text genres derived from Web 2.0 contain a great quantity of linguistic typical traits which need to be tackled with appropriate tools for these traits not to propagate to other tasks of Natural Language Processing, in particular, in this study, to the text summaries. In addition, possible solutions to address the problem are proposed, in order that the quality of the summaries might remain unaffected by the occurrence of these phenomena.

Key words: text summaries, grammar accuracy, linguistic traits.

1. Introducción

Con el nacimiento de la Web 2.0 (o Web social) aparece una serie de aplicaciones en las que el usuario es el principal protagonista y cuyo papel es esencial. Dichas aplicaciones, como por ejemplo las redes sociales, los *blogs*, foros, o las wikis, fomentan la colaboración y el intercambio de información entre comunidades de usuarios.

Por otro lado, el Procesamiento del Lenguaje Natural (PLN) agrupa una serie de tareas (recuperación de información, búsqueda de respuestas, clasificación de textos, etc.) que ofrecen al usuario mecanismos computacionalmente efectivos para facilitar la interrelación hombre-máquina y que permiten una comunicación menos rígida que los lenguajes formales (Moreno Boronat et al., 1999). Una de estas tareas, Resúmenes de Textos (RT), tiene como principal objetivo identificar y presentar al usuario la información más relevante de uno o más documentos de forma breve y concisa (Spärck Jones, 2007). Además, debido al continuo aumento de información en formato digital, este tipo de herramientas son críticas en el marco de la sociedad de la información en la que vivimos.

Cuando se intenta procesar la información que brinda la Web 2.0, la tarea de RT presenta una serie de retos originados por la naturaleza de estos nuevos géneros textuales, distintos de los que se tenía que hacer frente cuando el resumen se generaba a partir de géneros tradicionales, como es el caso de noticias periodísticas, textos legales o textos literarios. Estos retos incluyen, entre otros, la diversidad de fuentes, puntos de vista, opiniones, comentarios acerca de un tema, hecho o evento; los diversos idiomas en los que puede aparecer dicha información; la presencia de vocabulario informal o argot, así como también el uso frecuente de emoticonos para expresar emociones y sentimientos (Bandyopadhyay et al., 2008). Por tanto, es necesario investigar en aplicaciones de PLN, y en particular de RT, para lograr que sean capaces de tratar todos los fenómenos característicos de la Web 2.0, con el fin de proporcionar al usuario mecanismos que faciliten el manejo, de forma fiable, de la gran cantidad de información disponible, a la vez que le ayuden en el proceso de toma de decisiones. Estas peculiaridades no afectan solamente a la tarea de RT, sino que puede considerarse como un tema transversal a muchas otras tareas que procesan textos en lenguaje natural. Así, en el trabajo presentado en (Missen et al., 2009) se describen varios retos que la minería

de opiniones presenta a nivel de frase cuando se trabaja con *blogs*, como por ejemplo, determinar los límites de las oraciones o su lenguaje y se remarca que los métodos y herramientas tradicionales, tales como etiquetadores, analizadores sintácticos, etc. no son apropiados en muchas ocasiones ante el lenguaje característico de los nuevos géneros textuales de la Web 2.0. De esta manera, los fenómenos característicos de los nuevos géneros textuales pueden afectar al funcionamiento de herramientas tradicionales de PLN.

Sin embargo, si nos centramos en la tarea concreta de RT, la presencia de estos fenómenos en el resumen en sí puede ser transcendental para la calidad del mismo. Si lo que se pretende con el resumen generado es que sirva como sustituto del documento original y que, por tanto, los usuarios puedan usar dichos resúmenes para necesidades concretas, habrá que cuidar especialmente la calidad de los mismos, no pudiendo presentar el resumen por ejemplo lleno de faltas de ortografía. En cambio, cuando el resumen se utilizará para la consecución de otra tarea de PLN, como la búsqueda de respuestas o la recuperación de información, que el resumen arrastre ciertos rasgos propios del lenguaje utilizado en los nuevos géneros textuales, puede que no sea tan determinante como en el caso anteriormente comentado, aunque sí que influirá, seguramente, en el rendimiento de la tarea final que se pretende resolver.

El objetivo de esta investigación es, por tanto, realizar un estudio preliminar de la presencia de fenómenos que caracterizan los nuevos géneros textuales de la Web 2.0, en concreto los que aparecen en los *blogs*, y analizar en qué medida los resúmenes generados de esos *blogs* arrastran los mismos fenómenos, afectando de manera negativa la calidad de los mismos. La identificación de estos fenómenos será crucial para poder generar resúmenes de mayor calidad.

Este artículo está organizado de la siguiente manera. En primer lugar se realiza un breve repaso de los distintos enfoques que existen para generar resúmenes de opiniones (Sección 2). Posteriormente, en la Sección 3, se describe la tarea propuesta en el TAC 2008 y se detallan los fenómenos objeto del análisis. En la Sección 4 se cuantifican los fenómenos identificados en los *blogs*, así como se muestran los resultados preliminares de cómo estos fenómenos afectan a la calidad de los resúmenes generados. También se proponen posibles soluciones para disminuir los efectos negativos sobre la calidad final de los resúmenes. Finalmente, en la Sección 5 se exponen las conclusiones que se derivan de este estudio y se comentan las principales líneas de investigación para trabajos futuros.

2. Estado de la cuestión

Debido a la naturaleza subjetiva de la Web 2.0, una de las aplicaciones de la tarea de generación de resúmenes es su integración con la tarea de minería de opiniones para producir resúmenes que estén orientados a opiniones y, por tanto expresen sentimientos v

En lo que respecta a los resúmenes de opiniones generados a partir de *blogs*, ejemplos de enfoques propuestos se pueden encontrar en los sistemas que participaron en la tarea piloto de *Opinion Summarization* del TAC[1] 2008, cuyo objetivo era producir resúmenes orientados a un tema concreto a partir de un conjunto de *blogs* en los que se trataba dicho tema. Básicamente, los sistemas participantes primero identificaban las oraciones que hablan sobre el tema en cuestión y, una vez hecho esto, detectaban el carácter de las opiniones incluidas en esas frases para posteriormente agruparlas de acuerdo con su polaridad y generar finalmente el resumen (Bossard et al., 2008), (He et al., 2008), (Conroy and Schlesinger, 2008).

Por otro lado, fuera del ámbito de dicha competición, también se pueden encontrar enfoques interesantes para abordar esta tarea. Por ejemplo, en (Beineke et al., 2004) se propone un sistema en el que una vez identificados los fragmentos de texto que expresan opiniones, se utilizan técnicas de aprendizaje automático para seleccionar las frases que pertenecerán al resumen. De manera similar, en (Zhuang et al., 2006) se identifican las palabras que expresan opinión y el tipo de las mismas (bien expresando una opinión positiva o negativa), para componer posteriormente el resumen, pero orientado solamente a reseñas de películas de cine o de productos electrónicos, como en el caso presentado en (Meng and Wang, 2009). En (Hu and Liu, 2004) se propone un tipo particular de resúmenes de opinión que se centra en agrupar y resumir las opiniones a favor o en contra según distintas características de un producto. Por ejemplo, de una cámara fotográfica, el resumen podría girar en torno al tamaño, la calidad de las fotos, el diseño, etc. En (Zhou and Hovy, 2006) se realiza un estudio para abordar el problema de la generación de resúmenes tanto en *blogs* como en conversaciones mantenidas de forma online, y se compara con los tipos de documentos usados tradicionalmente para esta tarea (noticias periodísticas). Los autores también hacen una reflexión muy interesante acerca de la dificultad de evaluar la calidad de un resumen. En el enfoque propuesto en (Lloret et al., 2009), se utiliza un sistema de resúmenes genérico para decidir qué frases son las más relevantes, una vez que las oraciones subjetivas han sido identificadas y clasificadas. Además, se emplean métodos para minimizar el efecto de la redundancia por un lado, y para corregir los errores ortográficos y gramaticales que puedan aparecer en el resumen.

En muchas ocasiones, las opiniones que se encuentran en la Web vienen acompañadas de índices numéricos o símbolos que representan el grado de satisfacción de un usuario ante un determinado producto o servicio. En (Titov and McDonald, 2008) se propone un modelo (*Multi-Aspect Sentiment Model*) para identificar el carácter de las opiniones expresadas en dicho formato y poder extraer los fragmentos de textos correspondientes para generar un resumen. Recientemente, no solo encontramos en la literatura enfoques centrados en la generación de resúmenes de *blogs*, sino que otros

[1] Text Analysis Conference: http://www.nist.gov/tac/

se orientan a producir resúmenes enfatizando las diferencias sobre la misma entidad (Lerman and McDonald, 2009), (Carenini et al., 2006), o bien intentan resumir la información más importante a partir de los comentarios que la gente expresa en una entrada de un *blog* (Balahur et al., 2009), (Hu et al., 2007). En este último, el enfoque propuesto consiste en extraer las frases más significativas de los comentarios que sirvan como representación del tema que se trata. Para determinar las frases más relevantes, además de utilizar la frecuencia de las palabras, consideran tres elementos adicionales: el autor del comentario, el objeto del comentario y el tema del *blog*.

Todos los sistemas presentados anteriormente se comportan de manera similar: primero se detectan qué frases expresan opiniones y cuál es la polaridad de las mismas, y en segundo lugar se emplean técnicas para seleccionar las frases más relevantes y presentarlas de forma conjunta para conseguir que el resumen sea lo más coherente posible. Sin embargo, el paradigma más común a la hora de abordar la tarea de RT (paradigma extractivo) no es el más adecuado para generar resúmenes cuando las frases pueden provenir de diferentes comentarios, fuentes, etc. y por tanto, la tarea de generar resúmenes de estos nuevos géneros textuales tiene la dificultad añadida de que es necesario seguir un paradigma abstractivo para que el resumen final sea de calidad. Como se demuestra en (Carenini and Cheung, 2008) para este tipo de resúmenes el paradigma abstractivo ofrece, en general, mejores resultados que el extractivo. Además, tampoco se tienen en cuenta los fenómenos presentes en la Web 2.0, como la presencia de errores gramaticales y ortográficos, vocabulario informal, información contradictoria y redundante, derivados del hecho de que distintas fuentes exponen diversos puntos de vista sobre un mismo tema. Todos estos fenómenos son muy frecuentes en *blogs*, foros, etc. y, por tanto, hay que tenerlos en cuenta a la hora de generar un resumen, para que la calidad del mismo no se vea afectada de forma negativa. Por desgracia, la dificultad y alto coste asociado a la evaluación de resúmenes, así como la escasez de herramientas automáticas, hacen que la calidad de los resúmenes generados por sí sola sea difícil de evaluar.

3. Resúmenes de *blogs*

La tarea piloto *Opinion Summarization* propuesta en el marco de conferencias TAC 2008 tenía como principal objetivo la generación de un resumen tomando como documentos de entrada distintos *blogs* que contenían opiniones de la gente sobre diferentes temas. La idea era que los resúmenes generados contuvieran información acerca de las razones por la que la gente estaba a favor o en contra de algo, o que reflejara la opinión de las personas respecto a un tema. Por tanto, dado un tema, unas preguntas asociadas a dicho tema y una colección de *blogs* donde se encontraban las respuestas a las preguntas, los resúmenes generados debían tener en cuenta las frases pertinentes, así como también el tipo de sentimiento expresado en dichas oraciones (positivo, negativo o neutro). Al tratarse en su totalidad de preguntas subjetivas, como por ejemplo *"What motivated negative opinions regarding purchasing a car from CAR-MAX?"*

para el tema *Carmax* o *"Why do people like Starbucks better than Dunkin Donuts?"* para el tema *Starbucks coffee shops*, fue necesario combinar técnicas de minería de opiniones con técnicas de generación de resúmenes para resolver con éxito la tarea propuesta.

Para llevar a cabo este análisis preliminar, se parte de un subconjunto de 53 *blogs* y de 144 resúmenes generados por los sistemas participantes en la tarea piloto *Opinion Summarization* del TAC 2008 para cuatro temas concretos (1001, 1003, 1004, 1019). Es importante remarcar que todos los resúmenes analizados fueron participantes reales en dicha tarea, y por tanto, siguieron las pautas y las instrucciones que la organización daba a la hora de generar los resúmenes, como es el caso de la longitud de los mismos que no podía superar los 14.000 caracteres (sin considerar los espacios en blanco). El cuadro 1 muestra las preguntas asociadas a cada uno de los temas seleccionados para el análisis y el número de *blogs* relacionados con cada uno. Una de las evaluaciones realizadas en el seno de la conferencia TAC fue determinar si los resúmenes generados cumplían una serie de criterios lingüísticos que indicaban la calidad de los mismos. En esta evaluación, que se realizó de forma manual, se consideraron los siguientes criterios: corrección gramatical, cuyo objetivo es comprobar que un resumen carezca de errores ortográficos y sintácticos; ausencia de redundancia para ver si el resumen tiene información repetida; estructura y coherencia, para ver si la información recogida en el resumen está interrelacionada y organizada de manera correcta y adecuada; legibilidad, para ver si el resumen es fácil de leer y se entiende; y nivel de información que mide el grado en el que el resumen contiene información que responde a las preguntas formuladas para un tema concreto. Cada resumen se evaluaba para cada uno de estos criterios, según la siguiente escala de puntos: 1-2 (muy pobre); 3-4 (pobre); 5-6 (aceptable); 7-8 (bueno); 9-10 (muy bueno).

De todos los criterios anteriormente descritos, para este trabajo hemos seleccionado solamente dos: i) corrección gramatical y ii) la estructura y coherencia, que serán los que utilicemos para toda la experimentación y evaluación llevada a cabo. En base a estos criterios evaluados, analizamos con qué frecuencia están presentes estos fenómenos en la colección de *blogs* y cómo repercuten en los resúmenes finales generados. Por tanto, de forma más detallada, los fenómenos considerados en este estudio son:

- **Corrección gramatical:** cualquier persona puede escribir en un *blog*, y generalmente, distintas personas, dependiendo de su edad o nivel de formación, emplean distintos tipos de lenguaje. Generalmente, el lenguaje utilizado se caracteriza por ser informal (e.g. *You know*, *Yeah*, *It's gonna*), utilizar abreviaturas conocidas (e.g. *LOL*[2] o *Xmas* para referirse a *Christmas*) y emoticonos (:-D, para expresar alegría),

[2] LOL puede tener diferentes significados: *Laugh Out Loud*, *Lots of Love*, o *Lots of Luck*, aunque el más utilizado en los foros es el que se refiere a *Laugh Out Loud*, que significa *riéndose a carcajadas*.

vocabulario inapropiado (*shit*), o incluso contener errores de ortografía (e.g. *becasue*, *didnt*) o de gramática (frases incorrectas gramaticalmente como *"they still have that national proxy, no?"*). Concretamente para este estudio se analizó la presencia de los siguientes fenómenos: errores ortográficos y gramaticales; vocabulario informal, abreviaturas; emoticonos y vocabulario inapropiado.

- **Estructura y coherencia:** en un mismo *blog*, pueden existir un gran número de comentarios o *threads*, a través de los cuales diferentes usuarios pueden dar su opinión en relación a un tema. Por tanto, a la hora de generar un resumen es imprescindible saber estructurar la información y agruparla correctamente. Cuando el resumen está dirigido a un perfil de usuario determinado o a una pregunta en concreto, esto es más sencillo puesto que tenemos indicaciones de qué información debe aparecer en el resumen. Sin embargo, cuando queremos realizar un resumen genérico[3] este criterio es crucial para poder generar un resumen que tenga sentido y que no sea un conjunto de frases o de informaciones independientes que se han agrupado. En este análisis identificamos, por un lado, si el contenido de los *blogs* ha sido elaborado por varias personas, y por otro, analizamos si las frases de un resumen provienen de varias fuentes y si guardan relación entre ellas.

Cuadro 1. Ejemplo de temas, preguntas asociadas y número de *blogs* para cada tema

ID	Tema	Preguntas	Num. *blogs*
1001	Carmax	What motivated negative opinions regarding purchasing a car from CARMAX? What motivated positive opinions of CARMAX from car buyers?	10
1003	Jiffy Lube	What reasons are given for liking the services provided by Jiffy Lube? What reasons are given for not liking the services provided by Jiffy Lube?	19
1004	Starbucks coffee shops	Why do people like Starbucks better than Dunkin Donuts? Why do people like Dunkin Donuts better than Starbucks?	9
1019	China one-child per family law	What complaints are made about China's one-child per family law? What reason is given for approval of China's one-child per family law?	15

[3] En este contexto, genérico se refiere a un resumen que intenta captar la información más relevante de uno o varios documentos, sin estar centrado en ningún tema concreto.

4. Evaluación preliminar

Los resultados preliminares obtenidos del análisis realizado se muestran en el cuadro 2. Los resultados muestran el porcentaje de *blogs* en los que un determinado fenómeno está presente. La primera fila (*Blogs*-VC) representa el 60% de los *blogs*, ya que su contenido está formado por distintas fuentes que provienen de comentarios realizados por varias personas (en la columna *Núm. fuentes* vemos que el número medio de fuentes distintas para estos *blogs* ha sido de 20), mientras que la segunda (*Blogs*-UC) representa el 40% restante de los *blogs* analizados cuyo contenido solo está formado por la entrada principal (y como vemos, el número medio de fuentes distintas es 1). Como podemos observar, los *blogs* formados por distintas fuentes tienen una mayor presencia de fenómenos característicos de la Web 2.0, debido a la diversidad de personas que pueden participar en ellos. Los usuarios que participan en los *blogs* pueden provenir de diferentes países, nivel cultural, edades diferentes, etc., y por tanto, cada usuario escribirá de forma distinta. De todos estos fenómenos, el que más predomina es el de la utilización de vocabulario informal, seguido de la presencia de errores de gramática y de ortografía, así como también el uso de abreviaturas o emoticonos. Sin embargo, en los *blogs* que contienen solamente una entrada, estos fenómenos ya no se dan con tanta frecuencia, como se deduce de los resultados. En este caso, vemos cómo ahora solo el 47,4% de los *blogs* emplean un lenguaje informal, o solo el 5,3% frente al 65,5% anterior utilizan emoticonos. Por otro lado, en lo que se refiere a los resúmenes generados a partir de los *blogs* analizados anteriormente, se observa (cuadro 3) que, al ser generados seleccionando frases literales de los *blogs* originales, se cumple la hipótesis inicial de que los resúmenes arrastran muchos de los fenómenos que se daban en los *blogs*, como en el caso del uso de vocabulario informal o de errores ortográficos y gramaticales. Es importante destacar que el resto de fenómenos no se da con tanta frecuencia en los resúmenes (a excepción de la presencia de vocabulario inapropiado, que para el tema 1003 que presenta una alta concentración en los resúmenes). Esto es debido a la variabilidad de los resúmenes, y a que las frases relevantes con respecto a la pregunta puede que no sean las que en el *blog* original contienen emoticonos o abreviaturas. También, como se ha comentado anteriormente, el 40% de los *blogs* analizados contienen una única entrada, lo que hace que para estos casos, la presencia de fenómenos característicos de la Web 2.0 sea menor y por tanto, todos los resúmenes que se han generado a partir de estos *blogs* tampoco contienen dichos fenómenos. Cabe destacar, sin embargo, que en todos los resúmenes analizados, las frases que los forman presentan poca relación entre ellas y no están conectadas, de forma que la coherencia de los mismos se ve afectada negativamente. Esto es lo que se ha querido reflejar en la última columna de esta tabla (*Distintas fuentes*), que refleja que, aunque las frases se hayan extraído de distintos *blogs* o de distintos comentarios de un mismo *blog*, han sido simplemente agrupadas una a continuación de otra, sin mostrar conexión alguna. Teniendo en cuenta la evaluación manual realizada en el TAC, en la que se asignaba una puntuación de 1 a 10 (1=muy pobre, 10=muy bueno) a cada resumen y para cada criterio explicado anteriormente, las puntuaciones obtenidas para estos resúmenes con

respecto a los dos criterios analizados en este trabajo (corrección gramatical y estructura y coherencia) son bastantes pobres. El resultado medio para el primer criterio es 4,7 sobre un máximo de 10 (calidad pobre-aceptable), mientras que para el criterio que evalúa la estructura y la coherencia, es todavía más bajo, 2,7 (calidad muy pobre), tal y como se muestra en el Cuadro 4. Esto concuerda con los fenómenos analizados en este estudio, ya que se observa que están influyendo negativamente en la calidad de los resúmenes y que, por tanto, hay que tenerlos muy en cuenta cuando se pretende generar un resumen de forma automática, especialmente en este tipo de géneros textuales.

Cuadro 2. Fenómenos presentes en los *blogs*

	Err. ort./gram.	Voc. informal	Abreviaturas	Emoticonos	Voc. inapropiado	Núm. fuentes
Blogs-VC	72,4%	100%	72,4%	65,5%	45,0%	20
Blogs-UC	10,5%	47,4%	10,5%	5,3%	10,5%	1

Cuadro 3. Fenómenos presentes en los resúmenes

ID Tema	Err. ort./ gram.	Voc. informal	Abreviaturas	Emoticonos	Voc. inapropiado	Distintas fuentes
1001	94,4%	100%	25,0%	0%	50,0%	100%
1003	77,8%	100%	47,2%	13,9%	72,2%	100%
1004	66,6%	100%	58,3%	25%	22,2%	100%
1019	75,0%	72,2%	11,1%	22,2%	5,6%	100%
Promedio	78,5%	93,1%	35,4%	15,3%	37,5%	100%

Cuadro 4. Resultados obtenidos en la evaluación manual del TAC para el mismo conjunto de resúmenes

Criterio	Resultado medio
Corrección gramatical	4,7
Estructura y coherencia	2,7

Incrementar la calidad de los resúmenes generados

A partir de los resultados obtenidos donde se observa que existe una abundante presencia de fenómenos tales como vocabulario informal o inapropiado y errores ortográficos o gramaticales, tanto en los *blogs*, como en los resúmenes generados a partir de estos, es necesario plantear una serie de posibles soluciones para abordar el problema y con seguir así, que la calidad de los resúmenes no se vea afectada negativamente por culpa del lenguaje utilizado en los documentos originales que se pretenden resumir. De esta manera, en esta sección solamente se plantean algunas técnicas que pueden ser útiles para que este tipo de fenómenos no se propague a los resúmenes generados. En un futuro, se pretende estudiar la efectividad de cada una de las medidas propuestas.

- **Diccionarios:** mediante el uso de diferentes diccionarios o listados de términos se podría identificar por un lado los acrónimos y abreviaturas presentes en los *blogs*, de tal forma que en el resumen aparecería la palabra o frase completa en lugar de la abreviatura. También las listas de emoticonos[4] podrían aprovecharse para saber el sentimiento asociado a una oración y poder así clasificar la emoción de la oración, o bien emplearse para generar nuevo lenguaje para ser incorporado al resumen. Finalmente, listas de palabras inapropiadas como insultos nos permitirían identificar dichas palabras en los documentos originales y eliminarlas para que no aparezcan en los resúmenes.

- **Correctores ortográficos o de estilo:** estas herramientas son de gran utilidad para identificar los errores ortográficos de un texto. Productos como *STILUS*[5], *Style Checker*[6] , o *Language Tool*[7] podrían emplearse sobre los documentos o bien sobre el resumen una vez generado (en la etapa de post-procesamiento) para corregir los errores existentes.

- **Analizadores sintácticos o de dependencias:** si realizáramos un análisis de dependencias sobre los documentos fuente, como etapa previa a generar el resumen podríamos, por un lado, detectar las frases que no estuvieran completas y por otro lado, podríamos identificar la información que realmente valdría la pena extraer de una oración para que formara parte del resumen en caso de contener información relevante. De esta manera evitaríamos incorporar en el resumen frases incompletas o con ruido. También nos permitiría identificar frases que incluyeran ruido, puesto que en el caso de los documentos Web existen muchas oraciones con publicidad o con información adicional que no son relevantes para la generación del resumen (e.g. *contact us*). Herramientas como *Freeling*[8] o *Minipar*[9] podrían utilizarse para este fin. Para generar resúmenes de opiniones a partir de *blogs*, foros, wikis, etc. sería necesario investigar y centrar los esfuerzos en técnicas de abstracción que fueran capaces de identificar la información relevante y fusionarla para formar frases con sentido e interrelacionadas. De esta manera, se disminuiría el efecto que estos fenómenos tienen sobre el resumen. Sin embargo, aunque cada vez son más los sistemas que apuestan por seguir un paradigma abstractivo, la realidad es que los recursos y herramientas actuales para tal fin (por ejemplo, de generación de lenguaje) hacen que esta tarea sea todavía un reto.

[4] http://en.wikipedia.org/wiki/List of emoticons

[5] http://stilus.daedalus.es/stilusint.php

[6] http://www.whitesmoke.com/style-checker

[7] http://community.languagetool.org/

[8] http://www.lsi.upc.edu/n˜ lp/freeling/

[9] http://ai.stanford.edu/ rion/parsing/minipar viz.html

5. Conclusión y trabajo futuro

En este artículo se ha presentado un análisis preliminar de diferentes aspectos que pueden influir de forma negativa en la calidad de los resúmenes generados a partir de los nuevos géneros textuales que brinda la Web 2.0; en particular, se han analizado *blogs* de diferentes temáticas. De este estudio preliminar se puede concluir que los *blogs* contienen fenómenos característicos respecto al lenguaje empleado, el estilo de escritura, etc. que aparecen con mucha frecuencia y son aceptados por los usuarios, siendo el uso de vocabulario informal o la presencia de errores de ortografía y gramática los que más abundan. Del análisis realizado se observa que estos fenómenos también se reflejan en los resúmenes generados en porcentajes altos (aproximadamente, un 93% de los resúmenes incorporar lenguaje informal, o un 79% contienen errores ortográficos). Por tanto, para generar resúmenes de calidad y en especial, si queremos que estos resúmenes no arrastren los fenómenos que influyen directamente en su calidad, es importante analizar y tener en cuenta dichos fenómenos. Si no tienen en cuenta, por mucho que se investigue en técnicas para identificar la información relevante de uno o varios documentos, los resúmenes generados seguirán careciendo de la calidad suficiente. Por consiguiente, además de tener presentes estos fenómenos, una solución para disminuir el efecto negativo de los mismos sería investigar en técnicas de generación de resúmenes para quedarnos solamente con los fragmentos de información relevantes y añadir lenguaje nuevo, dando lugar a la generación de abstractos y, por tanto, lograr que el resumen generado sea un fragmento de texto coherente y gramaticalmente correcto. De cara a trabajos futuros, sería conveniente extender el análisis a un mayor número de *blogs* y una mayor gama de géneros textuales de la Web 2.0, como reseñas, wikis o foros para ver en qué medida se cumple la hipótesis en un corpus de mayor tamaño. Es interesante también replicar el análisis sobre un corpus de *blogs* especializados, por ejemplo en temas financieros y estudiar la calidad de los resúmenes generados a partir de este tipo de textos, comparándola con la obtenida con *blogs* de carácter más informal. Sería de gran utilidad realizar un análisis paralelo para ver si existe alguna relación entre número de errores que contienen los *blogs* y la cantidad de información que el resumen recoge, es decir, si los *blogs* con mayor cantidad de errores generan resúmenes menos fieles al contenido. También sería muy interesante investigar en técnicas de resúmenes que siguieran un paradigma abstractivo, y no se centraran solamente en la extracción de frases relevantes. De esta manera, muchos de estos problemas quedarían resueltos, al no basarnos únicamente en extraer frases literales de los documentos fuente. Además, a corto plazo se pretende investigar y analizar con mayor detalle, la efectividad de las soluciones propuestas, como por ejemplo, el uso de diccionario especializados para filtrar las palabras inapropiadas o traducir las abreviaturas por su correspondiente palabra completa, o también la utilización de analizadores sintácticos para poder identificar frases incompletas o incorrectas desde el punto de vista gramatical. El objetivo final que se pretende al generar resúmenes orientados a opiniones es que faciliten al usuario el manejo de grandes cantidades de información, ayudándoles en el proceso de toma de decisiones.

Referencias bibliográficas

Balahur, Alexandra; Lloret, Elena; Boldrini, Ester; Montoyo, Andrés; Palomar, Manuel and Martínez-Barco, Patricio (2009). "Summarizing Threads in Blogs Using Opinion Polarity". In *Proceedings of the International Workshop on Events in Emerging Text Types* (eETTs), pages 5-13.

Bandyopadhyay, Sivaji; Poibeau, Thierry; Saggion, Horacio and Yangarber, Roman (editors) (2008). Coling 2008: *Proceedings of the Workshop Multi-source Multilingual Information Extraction and Summarization*.

Bautin, Mikhail; Vijayarenu, Lohit and Skiena, Steven (2008). "International Sentiment Analysis for News and Blogs". In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Beineke, Philip; Hastie, Trevor; Manning, Christopher and Vaithyanathan; Shivakumar (2004). "An Exploration of Sentiment Summarization". In Shanahan, James G.; Wiebe, Janyce and Qu, Yan (editors). *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

Bossard, A.; Généreux, M. and Poibeau, T. (2008). "Description of the LIPN Systems at TAC 2008: Summarizing Information and Opinions". In *Proceedings of the Text Analysis Conference (TAC)*.

Carenini, Giuseppe and Cheung, Jackie C.K. (2008). "Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality". In *Proceedings of the Fifth International Natural Language Generation Conference, ACL 2008*. Ohio, pages 33-40.

Carenini, Giuseppe; Ng, Raymond and Pauls, Adam (2006). "Multi-Document Summarization of Evaluative Text". In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*.

Conroy, J.M. and Schlesinger, J.D. (2008). "CLASSY at TAC 2008 metrics". In *Proceedings of the Text Analysis Conference (TAC)*.

He, T.; Chen, J.; Gui, Z. and Li, F. (2008). "CCNU at TAC 2008: Proceeding on Using Semantic Method for Automated Summarization Yield". In *Proceedings of the Text Analysis Conference (TAC)*.

Hu, Minqing and Liu, Bing (2004). "Mining and Summarizing Customer Reviews". In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 168-177.

Hu, Meishan; Sun, Aixin and Lim, Ee-Peng (2007). "Comments-oriented Blog Summarization by Sentence Extraction". In *Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM),* pages 901-904.

Kim, Jungi; Li, Jin-Ji and Lee, Jong-Hyeok (2009). "Discovering the Discriminative Views: Measuring Term Weights for Sentiment Analysis". In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP,* pages 253-261.

Lerman, Kevin and McDonald, Ryan (2009). "Contrastive Summarization: An Experiment with Consumer Reviews". In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 113-116.

Lloret Elena and Palomar, Manuel (2009). "A Gradual Combination of Features for Building Automatic Summarisation Systems". In *Proceedings of the 12th International Conference on Text, Speech and Dialogue*, pages 16-23.

Lloret, Elena; Balahur, Alexandra; Montoyo, Andrés and Palomar, Manuel (2009). "Towards Building a Competitive Opinion Summarization System: Challenges and Keys". In *Proceedings of the NAACL Student Research Workshop*, pages 72-77.

Marcu, Daniel (1999). "Discourse Trees Are Good Indicators of Importance in Text". In *Inderjeet Mani and Mark Maybury, editors, Advances in Automatic Text Summarization*, pages 123-136.

Meng, Xinfan and Wang, Houfeng (2009). "Mining User Reviews: from Specification to Summarization". In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers,* pages 177-180.

Mihalcea, Rada (2004). "Graph-based ranking algorithms for sentence extraction, applied to text summarization". In *Proceedings of the Association for Computational Lingusitics*, pages 170-173.

Muhammad, Malik; Missen, Saad and Boughanem; Mohand (2009). "Sentence-Level Opinion-Topic Association for Opinion Detection in Blogs". In *WAINA '09: Proceedings of the 2009 International Conference on Advanced Information Networking and Applications Workshops*, pages 733-737.

Muhammad, Malik; Missen, Saad; Boughanem, Mohand and Cabanac, Guillaume (2009). "Challenges for Sentence Level Opinion Detection in Blogs". In *ICIS '09:*

Proceedings of the 2009 Eight IEEE/ACIS International Conference on Computer and Information Science, pages 347-351.

Moreno Boronat, Lidia; Palomar Sanz, Manuel; Molina Marco, Antonio and Ferrández Rodríguez, Antonio (1999). *Introducción al procesamiento del lenguaje natural*. Alicante: Universidad de Alicante.

Pang, Bo and Lee, Lillian (2008). "Opinion Mining and Sentiment Analysis". *Foundations and Trends in Information Retrieval*, *2*(1-2):1-135.

Saggion, Horacio (2008). "Automatic Summarization: An Overview". *Revue française de linguistique appliquée, XIII*(1).

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

Titov, Ivan and McDonald, Ryan (2008). "A Joint Model of Text and Aspect Ratings for Sentiment Summarization". In *Proceedings of ACL-08: HLT*, pages 308-316.

Zhou, Liang and Hovy, Eduard (2006). "On the Summarization of Dynamically Introduced Information: Online Discussions and Blogs". In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW),* pages 237-242.

Zhuang, Li; Jing, Feng and Zhu, Xiao-Yan (2006). "Movie Review Mining and Summarization". In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43-50.

Fecha de recepción: 15/12/09
Fecha de aceptación: 10/05/10