Procesamiento del Lenguaje Natural, Revista nº 52 marzo de 2014, pp 69-76 recibido 15-11-13 revisado 05-02-14 aceptado 06-02-14

Función de las secuencias narrativas en la clasificación de la polaridad de reviews

The function of narrative chains in the polarity classification of reviews

John Roberto
CLiC-UB
Gran Via 585, 08007 Bcn
roberto.john@ub.edu

Maria Salamó
Universidad de Barcelona
Gran Via 585, 08007 Bcn
maria.salamo@ub.edu

M. Antònia Martí
CLiC-UB
Gran Via 585, 08007 Bcn
amarti@ub.edu

Resumen: Los comentarios sobre productos o *reviews* son una fuente valiosa de información para entender las preferencias de los usuarios en los sistemas para la personalización de contenidos. En este artículo se analiza la función que desempeñan las secuencias narrativas en el cálculo de la polaridad de productos. Con esta finalidad hemos aplicado un algoritmo para extraer las oraciones que contienen eventos relacionados semánticamente y hemos realizado una serie de experimentos orientados a determinar el impacto que la omisión de dichas oraciones puede tener a nivel de la polaridad de los *reviews*. Los resultados obtenidos demuestran que las opiniones negativas de los productos se suelen expresar mediante secuencias narrativas mientras que las positivas son independientes de la narración.
Palabras clave: Análisis de la polaridad, perfiles de usuario, minería de opiniones

Abstract: Reviews are a powerful source of information about consumer preferences that can be used in personalization systems. In this paper we analyse the role played by narrative chains in determining the polarity of reviews. For this purpose, we applied an algorithm to remove sentences containing events semantically connected. We report experiments designed to evaluate the impact that the omission of those sentences has in determining the polarity of reviews. The results show that negative opinions are often expressed in terms of narrative chains while positive opinions are independent of narratives.
Keywords: Polarity analysis, user profiles, opinion mining

1 Introducción

Ante la enorme cantidad de información disponible en Internet, los sistemas para la personalización de contenidos (ej. Sistemas de Recomendación) se están convirtiendo en una herramienta indispensable para eliminar la sobrecarga de información. Los Perfiles de Usuario (PU) son un componente primordial de estos sistemas. Definimos el PU como una representación estructurada de los atributos de un usuario. Dichos atributos se pueden categorizar en dos clases (Vildjiounaite et al., 2007):

- Restricciones: información personal sobre el usuario como por ejemplo su edad, estado civil o su personalidad. Esta información se usa para limitar o restringir los productos a recomendar.
- Preferencias: información sobre los gustos, necesidades e intereses del usuario en relación con un determinado producto o servicio. Las preferencias pueden referirse al producto como totalidad ("me gusta el iPhone") o a algunas de sus características ("me gusta su diseño").

Las valoraciones expresadas como *ratings* y, más recientemente, los *reviews* suponen las formas habituales de obtener los PUs. En nuestro análisis sólo consideraremos los *reviews*, por el interés lingüístico que despiertan al estar constituidos íntegramente por texto en lenguaje natural.

En un trabajo previo (Roberto, Salamó, y Martí, 2014), determinamos que las restricciones y las preferencias se expresan en segmentos más o menos independientes de los *reviews*. Así, basándonos en (Ricci y Wietsma, 2006), definimos el *review* como un texto breve y subjetivo que: **1.** relata las *expe-*

ISSN 1135-5948

riencias personales de un usuario en relación con un producto (es decir, las restricciones), **2.** contiene una *descripción* más o menos detallada sobre las características del producto (es decir, las preferencias sobre las características) y **3.** hace una *valoración* general del mismo (preferencia a nivel de ítem). Adicionalmente, y según nuestro modelo, cada uno de estos tres tipos de segmentos se expresan mediante diferentes modalidades textuales:

- Las experiencias se expresan mediante *narraciones*: ej. "mi esposa y yo nos alojamos por 15 días".
- La descripción de las características del producto se expresan mediante *descripciones*: ej. "las habitaciones son enormes".
- La valoración del producto se expresa mediante *exhortaciones*: ej. "os recomiendo pasar unos días en este hotel".

La extracción automática de estos tres tipos de segmentos es una tarea útil para enriquecer los PUs y mejorar los procesos de recomendación basados en el análisis de *reviews*.

En este artículo analizamos la función que desempeñan las secuencias narrativas en el cálculo de la polaridad de productos. Para ello extraemos de los *reviews* las oraciones que relatan las *experiencias* del usuario adaptando el modelo de esquemas narrativos de Chambers y Jurafsky (2008b). Posteriormente, efectuamos varios experimentos orientados a determinar el impacto que la omisión de dichas oraciones puede tener en el cálculo de la polaridad de los *reviews*. Nuestro objetivo es demostrar que no todos los componentes de un *review* inciden de la misma manera en la detección de la polaridad.

La estructura del artículo es como sigue. En la Sección 2 presentamos una adaptación del algoritmo para la detección de eventos y secuencias narrativas de Chambers y Jurafsky. En la Sección 3 presentamos los experimentos y los resultados. En la Sección 4 hacemos un breve repaso de los trabajos relacionados con la detección de secuencias narrativas. Finalmente, en la Sección 5 presentamos las conclusiones y el trabajo futuro.

2 Detección de eventos y secuencias narrativas

Para efectos de nuestro trabajo, definiremos un *review* (R) como un texto compuesto por un conjunto de oraciones narrativas (O_N), descriptivas (O_D) y exhortativas (O_E), es decir, $R = \{O_N + O_D + O_E\} = \{o_1, o_2, o_3, ..., o_n\}$. Adicionalmente, una secuencia narrativa ($O_{\mathfrak{e}N}$) es un subconjunto de las oraciones que en O_N relatan eventos relacionados semántica y temporalmente: $O_{\mathfrak{e}N} \subseteq O_N$. Cada evento ($\mathfrak{e}$) es una tupla conformada por el verbo y sus argumentos: $\mathfrak{e} = \langle v, arg \rangle$ donde $arg \in \{suj, obj, prep\}$. Un caso concreto de $O_{\mathfrak{e}N}$ lo podemos ver en el Ejemplo (1):

(1) My dad bought me a Saturn for my graduation in 1992 before all the marketing hype (how embarrassing to be constantly asked if I went to Tennessee!). Shortly thereafter the problems started. ...
Mi papá me compró un Saturn para mi graduación en 1992 antes de todo el boom publicitario (qué vergüenza que te pregunten constantemente si fuiste a Tennessee!). Poco después empezaron los problemas. ...

Con el fin de evidenciar la relación semántica entre los eventos de un *review* recurriremos a la "presunción de la coherencia narrativa"[1] de Chambers y Jurafsky (2008b). Según estos autores, los verbos que comparten argumentos correferentes están relacionados semánticamente en virtud de la estructura narrativa del discurso. Por ejemplo, en el fragmento (1) los verbos *bought* y *went* tienen pronombres correferentes (*me* e *I*), por lo que se establece una relación semántica entre ambos verbos: $\langle bought, X_{objeto} \rangle$ y $\langle went, X_{sujeto} \rangle$. Siguiendo la notación gráfica utilizada por Chambers y Jurafsky:

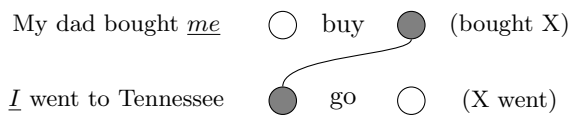

Figura 1: Relación semántica entre eventos (modelo básico).

En la Figura 1 los círculos sombreados representan el elemento correferente X (en un caso X es objeto y en otro sujeto oracional).

[1] *Narrative Coherence Assumption.*

Los círculos en blanco son las entidades que no tienen ningún nexo correferencial: *my dad* y $Tennessee$.

El problema de esta representación es que solo captura 2 de los 6 eventos subrayados en el Ejemplo (1). Para incluir los 4 eventos restantes hemos de considerar, además de los verbos, las nominalizaciones deverbales (*graduation* y *hype*), las expresiones temporales ($before$ y *shortly thereafter*) y los nexos coordinantes y subordinantes (if). El resultado del nuevo análisis se puede ver en la Figura 2.

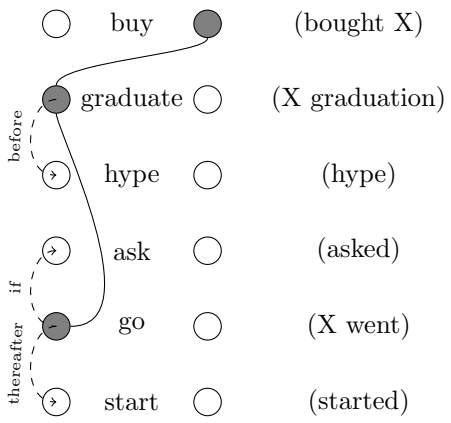

Figura 2: Relación semántica entre eventos (modelo extendido).

Como podemos observar, el sustantivo deverbal *graduation* se incorpora de forma directa (comparte el argumento X) a la narración gracias al pronombre posesivo *my*. Por su parte los eventos *hype* y *ask* son incorporados de forma indirecta (no comparten argumentos) por una expresión temporal ($before$) y por un nexo subordinante (if). Finalmente, el verbo *start* lo incluimos por la proximidad de ésta oración con las anteriores y porque está encabezada por una locución adverbial (*shortly thereafter*) que expresa un orden temporal.

El Algoritmo 1 describe el procedimiento que usamos para extraer de los *reviews* las oraciones que conformarán las secuencias narrativas según el modelo expuesto en la Figura 2. La entrada del algoritmo es el *review* (R), concretamente, las oraciones que lo componen. Mediante una llamada al parser de la Universidad de Stanford[2] obtenemos las dependencias sintácticas (línea 8) y resolvemos las correferencias de cada oración (línea 9). A continuación, identificamos las nominalizaciones basándonos en un listado de nombres deverbales[3] que hemos extraído de *NomBank* (línea 10). Aplicamos un procedimiento similar para identificar las expresiones temporales a partir de *TimeBank* (Pustejovsky et al., 2003) y de un listado de adverbios, adjetivos y preposiciones extraído del mismo corpus ($after$, $immediately$, $follows$, $meanwhile$, etc.) (línea 11). En las líneas 12 a la 16 seleccionamos los pares núcleo - modificador y, si el núcleo es un verbo o una nominalización y el modificador un argumento correferente, o una expresión temporal, agregamos el núcleo a la lista de eventos (línea 14). Dado que una secuencia narrativa se compone como mínimo de dos eventos, sólo la oración que contiene dos o más eventos (línea 17) pasa a formar parte de la secuencia narrativa (línea 18). Adicionalmente, siempre que sea posible, capturamos la referencia a las dos oraciones siguientes (líneas 19 a la 24) y si alguna de estas dos oraciones contiene una expresión temporal, también la incluimos en la secuencia narrativa (líneas 26 y 34). Finalmente, el algoritmo retorna el conjunto de oraciones seleccionadas ($O_{\in N}$).

En el Anexo A presentamos el Ejemplo (1) desarrollado en su totalidad y procesado de forma automática. Queremos aclarar que aunque en la Figura 2 incluimos el evento *ask* como parte del modelo, la implementación actual del algoritmo no captura dicho evento (ver Tabla 4). Una tarea pendiente como trabajo futuro es buscar alternativas al tratamiento de estos eventos "aislados".

3 Experimentos y resultados

Los experimentos que presentamos a continuación tienen como objetivo determinar la relación que existe entre el uso de las secuencias narrativas y la polaridad de los *reviews*.

3.1 Los datos

En este trabajo hemos usado el corpus de opiniones en inglés de (Cruz, 2012). El corpus se compone de 2547 documentos de los cuales 972 corresponden a opiniones sobre coches, 587 sobre auriculares y 988 sobre hoteles. Cada opinión lleva asociada, en un archivo independiente en formato XML, la puntuación del *review* (línea 2 en el siguiente fragmento) y las palabras que expresan opinión sobre

[2] http://nlp.stanford.edu/software/lex-parser.shtml

[3] Consideramos sólo los sustantivos que se derivan directamente de verbos.

Algoritmo 1 Detección automática de secuencias narrativas

Require: $R = \{o_1, o_2, o_3, ..., o_n\}$
1: $\mathfrak{e}_{list} \leftarrow \varnothing$ // lista de eventos
2: $O_{\mathfrak{e}N} \leftarrow \varnothing$ // secuencia narrativa
3: $primero \leftarrow false$ // variable booleana auxiliar
4: $segundo \leftarrow false$ // variable booleana auxiliar
5: $o_x \leftarrow \varnothing$ // oración auxiliar
6: $o_y \leftarrow \varnothing$ // oración auxiliar
7: **while** $\forall o_i \in R$ **do**
8: $deps = \{par_k : 1 \leq k \leq n_k\}$ donde $par_k = \langle núcleo, modificador \rangle$ // se obtienen las dependencias
9: $crefs = \{c_m : 1 \leq m \leq n_m\}$ donde $c_m = \langle palabra \rangle$ // se resuelven las correferencias
10: $nomin = \{n_s : 1 \leq s \leq n_s\}$ donde $n_s = \langle palabra \rangle$ // se identifican las nominalizaciones
11: $time = \{t_t : 1 \leq t \leq n_t\}$ donde $t_t = \langle palabra \rangle$ // se identifican las expresiones temporales
12: **for all** $par = \langle núcleo, modificador \rangle \in deps$ **do**
13: **if** $((núcleo = verbo \| núcleo = nomin)$ **and** $(modificador = crefs \| modificador = time))$ **then**
14: $\mathfrak{e}_{list} \leftarrow \mathfrak{e}_{list} \cup \{núcleo\}$ // se guarda el evento
15: **end if**
16: **end for**
17: **if** $|\mathfrak{e}_{list}| \geq 2$ **then**
18: $O_{\mathfrak{e}N} \leftarrow O_{\mathfrak{e}N} \cup \{o_i\}$ // se agrega la oración a la secuencia narrativa
19: **if** $(i+1) \leq n$ **then**
20: $o_x \leftarrow o_{i+1}$
21: **end if**
22: **if** $(i+2) \leq n$ **then**
23: $o_y \leftarrow o_{i+2}$
24: **end if**
25: **end if**
26: **while** $\forall t_t \in time$ **and** $(o_x \neq \varnothing || o_y \neq \varnothing)$ **do**
27: **if** $t_t \subset o_x$ **then**
28: $O_{\mathfrak{e}N} \leftarrow O_{\mathfrak{e}N} \cup \{o_x\}$ // se agrega la oración a la secuencia narrativa
29: $o_x \leftarrow \varnothing$
30: **end if**
31: **if** $t_t \subset o_y$ **then**
32: $O_{\mathfrak{e}N} \leftarrow O_{\mathfrak{e}N} \cup \{o_y\}$ // se agrega la oración a la secuencia narrativa
33: $o_y \leftarrow \varnothing$
34: **end if**
35: **end while**
36: **end while**
Ensure: $O_{\mathfrak{e}N}$

el producto o alguna de sus características (líneas 7 a la 10):

```
1  ... <review id="5" item="Amerisuites Busch Gar-
2  dens" rating="2">...
3  <sentence id="1">
4  Stains(1) on(2) carpet(3) ,(4) dirty(5) pool(6) ,(7)
5  bad(8) elevator(9) /(10) housekeeper(11) setup(12)
6  .(13)
7  <opinion polarity="-" feature="swimming pool"
8  featWords="6" opWords="5"/>
9  <opinion polarity="-" feature="elevator" feat-
10 Words="9" opWords="8"/>
11 </sentence> ...
```

3.2 Configuraciones

El análisis de la polaridad de los *reviews* se ha realizado considerando seis configuraciones diferentes, cinco de ellas basadas en la supresión de diferentes fragmentos del texto:

R$_{ref}$ Es la configuración de referencia. Se utiliza el *review* en su totalidad para el cálculo de la polaridad.

ctr_{40} Análisis de la polaridad usando un 40 % del *review* seleccionado aleatoriamente.

ctr_{30} Análisis de la polaridad usando solo un 30 % del *review*. Las configuraciones ctr_{40} y ctr_{30} son de control (ctr) ya que se hace una fragmentación aleatoria del texto. Hemos seleccionado el 40 % y el 30 % de los textos como configuraciones de control puesto que, como veremos más adelante, estos porcentajes se aproximan al obtenido con la configuración que intentamos evaluar (des_{ext}).

nar_{ext} Análisis de la polaridad usando las secuencias narrativas ($O_{\mathfrak{e}N}$). Esta configuración utiliza el modelo extendido para detectar las secuencias narrativas (ver Figura 2).

des_{bas} Análisis de la polaridad usando los segmentos NO narrativos, es decir, la

Dominio	**pol.**	$\mathbf{R}_{ref}$	ctr_{40}	ctr_{30}	nar_{ext}	des_{bas}	des_{ext}
Coches	+	89.4	71.4	72.2	77.8	83.5	84.8
	ren.	0.8	1.7	2.4	1.2	1.8	2.1
	–	77.3	75.5	72.6	67.2	68.2	66.7
	ren.	0.7	1.8	2.4	1.1	1.5	1.6
Hoteles	+	95.7	87.2	86.7	78	94.5	95.5
	ren.	0.9	2.1	2.8	1.2	2.1	2.4
	–	77.5	67	67.5	66.7	69	67.2
	ren.	0.7	1.6	2.2	1.1	1.5	1.7
Auriculares	+	82.9	65.8	66.6	69.7	82.5	81.7
	ren.	0.8	1.6	2.2	1.1	1.8	2.0
	–	72.4	68.1	68.1	60.3	65.7	62.4
	ren.	0.7	1.7	2.2	0.9	1.4	1.5
Longitud	%	100	40	30	60.6	44.2	39.4
	pals.	1.184.496	473.798	355.348	717.805	524.505	466.691

Tabla 1: Resultados del cálculo de la polaridad en los *reviews*.

parte del *review* que queda tras eliminar las secuencias narrativas ($O_{\not\in N}$). Por comodidad nos referiremos a los segmentos no narrativos como segmentos descriptivos ($des = O_D \cup O_E$). En esta configuración, la eliminación de las secuencias narrativas se realizó según el modelo básico (ver Figura 1).

des_{ext} Es la configuración objetivo, es decir, la que nos informa directamente sobre el impacto que tiene la omisión de las secuencias narrativas en el cálculo de la polaridad. des_{ext} está basada en el modelo extendido de la Figura 2. Recalcamos que tanto des_{bas} como des_{ext} evalúan la polaridad de los *reviews* suprimiendo sus segmentos narrativos.

Para el análisis de la polaridad hemos usado el *Semantic Orientation CALculator* (SO-CAL)(Taboada et al., 2011). El SO-CAL utiliza diccionarios de palabras anotadas con su orientación semántica en una escala que va del 5 para los términos más positivos (*exquisite*) a -5 para los más negativos ($horrific$). El SO-CAL también incorpora modificadores de la polaridad como son los intensificadores (*most excellent*) y la negación (*not good*).

3.3 Resultados

Con el fin de determinar el impacto que tiene la omisión de las secuencias narrativas en el cálculo de la polaridad de los *reviews* hemos realizado 36 evaluaciones del Algoritmo 1. Las 36 evaluaciones se efectúan considerando 3 dominios (coches, hoteles y auriculares), 2 polaridades (positiva + y negativa –) y 6 configuraciones (ver Sección 3.2). Los resultados se pueden ver en la Tabla 1. En esta tabla tenemos los niveles de precisión que se obtienen al predecir la polaridad de los *reviews*. También evaluamos el rendimiento (ren.), es decir, la relación entre la precisión y la cantidad de texto necesaria para obtener dicha precisión ($ren = precisión/longitud\_texto$). La cantidad o longitud de texto usada en cada configuración está descrita en la última fila de la Tabla 1, tanto en términos de porcentajes (%) como de número de palabras (pals.).

Las principales observaciones que podemos extraer de este análisis son las siguientes:

- R_{ref} obtuvo las mejores precisiones en el cálculo de la polaridad pero su rendimiento es muy bajo puesto que requiere de todo el texto de los *reviews* (más de 1 millón de palabras) para alcanzar tales valores. La precisión más alta bajo esta configuración es de 95.7 % (hoteles) sin embargo en ningún momento el rendimiento supera el valor del 1.0.

- ctr_{40} y ctr_{30} representan el caso contrario a R_{ref}: si bien su rendimiento llega a superar el 2.0, sus niveles de precisión son muy bajos en comparación con las otras configuraciones (alrededor del 72 %).

- nar_{ext} es la configuración menos eficaz: tiene los niveles de precisión más bajos de todas las configuraciones y su rendimiento es modesto. Por ejemplo, la precisión para el dominio de los auriculares con polaridad negativa es de tan solo el 60.3 % con un rendimiento del 0.9.
- des_{bas} y des_{ext}, de otra parte, presentan los niveles de precisión más próximos a R_{ref} en cuanto al cálculo de la polaridad positiva utilizando tan solo el 39.4 % del *review* (466.691 palabras) (ver celdas sombreadas en la Tabla 1). Estos resultados contrastan con su bajo rendimiento en el cálculo de la polaridad negativa. El contraste al que hacemos referencia es más evidente en des_{ext} que en des_{bas}. Por ejemplo, en el dominio de los auriculares la precisión de aciertos positivos es del 95.5 % frente a un 67.2 % de precisión en aciertos negativos. De la misma forma, si atendemos al promedio de las diferencias en el rendimiento (ren) entre des_{ext} y las configuraciones de control, que trabajan con un número de palabras similar al de des_{ext}, veremos una divergencia notable en el rendimiento debida al tipo de polaridad: $des_{ext} = 0.5$[4], $ctr_{30} = 0.2$ y $ctr_{40} = 0.2$.

Por tanto, los resultados indican que al omitir las secuencias narrativas de los *reviews* estamos descartando información relevante para entender las opiniones negativas. Desde el punto de vista lingüístico, este hallazgo nos revela que los usuarios suelen recurrir a la narración para describir aspectos negativos de los productos mientras que las valoraciones positivas son independientes de la narración: "*The management staff is the worst I've ever encountered ... The employees at this hotel were the rudest bunch of people I have ran into in a while.*".

4 Trabajos relacionados

En esta sección describimos algunos de los trabajos más relevantes relacionados con la detección de secuencias narrativas en lenguaje natural.

Chambers y Jurafsky han estudiado la forma de inferir el orden de los eventos que aparecen en textos narrativos. En (Chambers y Jurafsky, 2008b; Chambers y Jurafsky, 2008a) los autores presentan las bases de su modelo. En trabajos posteriores, (Chambers y Jurafsky, 2009; Chambers y Jurafsky, 2010) amplian el modelo para incluir los diferentes roles que puede desempeñar el "protagonista" de un evento (ej. *criminal*, *sospechoso*) y para ampliar la lista de argumentos que hacen servir (*sujeto*, *objeto*, *prep*). También introducen el concepto de "esquema narrativo" con el que buscan relacionar en un mismo "escenario" todos los "protagonistas" de cada una de las cadenas de eventos narrativos.

De otro lado, Regneri, Koller y Pinkal (2010) aplican aprendizaje no supervisado para detectar las frases que describen un mismo evento ("sentarse a la mesa", "tomar asiento", etc.)[5] y el orden en que suelen presentarse en una narración (*script*). Su procedimiento se basa en el uso de una matriz de frases semánticamente relacionadas donde aplica el alineamiento múltiple de secuencias (*Multiple Sequence Alignment*, MSA). Sobre esta representación matricial se construye un "grafo temporal" que, mediante un algoritmo de agrupamiento, determina el orden en que se han de presentar los eventos.

Hajishirzi et al. (2011) y Hajishirzi y Mueller (2012) analizan la forma de interpretar oraciones narrativas mediante la representación simbólica de los eventos, estados y entidades que ellas contienen. Su aproximación se vale de dos tipos de conocimiento: la descripción de los eventos y las entidades más importantes del dominio. Para la descripción temporal de los eventos (y estados) utilizan un lenguaje simbólico caracterizado por la presencia de condiciones y consecuencias: $\langle evento(\vec{x}), condición(\tilde{x}), consecuencia(\tilde{x})\rangle$.

Li, Lee-Urban, y Riedl (2012) y Li et al. (2012) exponen una técnica para identificar los eventos característicos de una determinada situación y su disposición temporal más habitual. Primero, los autores obtienen ejemplos reales de secuencias narrativas en diferentes ámbitos. Posteriormente, agrupan las oraciones que están semánticamente relacionadas (ej. "la policía detiene al criminal", "los agentes arrestan al ladrón") aplicando la similitud de cosenos. Por último, establecen restricciones entre eventos (precedencia, opcionalidad, exclusión) mediante un análisis

[4] des_{ext}: coches(2.1 - 1.6 = 0.5); hoteles(2.4 - 1.7 = 0.7); auriculares(2.0 - 1.5 = 0.5) $\rightarrow \frac{0.5+0.7+0.5}{3} = \mathbf{0.5}$

[5] Aunque parte de su trabajo consiste en detectar las diferentes realizaciones lingüísticas de un mismo evento, Regneri, Koller, y Pinkal (2010) aclaran que no se trata de un problema de paráfrasis.

de su frecuencia y probabilidad de aparición.

5 Conclusiones y trabajo futuro

En este artículo se analiza la función que desempeñan las secuencias narrativas en el cálculo de la polaridad de productos. Con este propósito adaptamos el modelo de esquemas narrativos de Chambers y Jurafsky (2008b) para extraer de los *reviews* las oraciones que relatan eventos y efectuamos un análisis de la polaridad de los textos bajo diferentes configuraciones.

Los resultados obtenidos indican que la modalidad textual narrativa se suele emplear para valorar negativamente los productos mientras que las valoraciones positivas son independientes de la narración. Concluimos, por tanto, que la omisión de las secuencias narrativas sólo afecta a los *reviews* con polaridad negativa. Este conocimiento es útil para comprender la forma en que los usuarios evalúan productos en lenguaje natural.

El trabajo futuro está enfocado a mejorar el rendimiento del algoritmo para detectar las secuencias narrativas mediante la incorporación de una fase de preprocesamiento de los *reviews* que corrija errores tipográficos, ortográficos y, especialmente, de segmentación de las oraciones. Adicionalmente, pensamos evaluar diferentes recursos para resolver la correferencia puesto que es un elemento importante para obtener buenos resultados en la detección de secuencias narrativas. Por último, creemos que sería productivo restringir el tipo de nominalizaciones que se hacen servir para la identificación de eventos ya que la función eventiva de algunas de ellas en el texto es cuestionable.

Agradecimentos

Esta investigación ha sido posible gracias a la financiación de los proyectos TIN2012-38603-C02 y TIN2009-14404-CO2 del Ministerio de Ciencia e Innovación así como a la Generalitat de Catalunya mediante una beca predoctoral FI (2010FI_B 00521).

Bibliografía

Chambers, N. y D. Jurafsky. 2008a. Jointly combining implicit constraints improves temporal ordering. En *Proc. of the Conference on Empirical Methods in NLP*, páginas 698–706, Stroudsburg, USA.

Chambers, N. y D. Jurafsky. 2008b. Unsupervised learning of narrative event chains. En *Proc. of ACL-08: HLT*, páginas 789–797, Columbus, Ohio.

Chambers, N. y D. Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. En *Proc. of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on NNLP of the AFNLP*, volumen 2, páginas 602–610, Stroudsburg, USA.

Chambers, N. y D. Jurafsky. 2010. A database of narrative schemas.

Cruz, F. 2012. *Extracción de opiniones sobre características: un enfoque práctico adaptable al dominio.* Colección de monografías de la Sociedad Española para el Procesamiento del Lenguaje Natural. SEPLN.

Hajishirzi, H., J. Hockenmaier, T. Mueller, y E. Amir. 2011. Reasoning about robocup soccer narratives. En Fabio Gagliardi Cozman y Avi Pfeffer, editores, *Proc. of the Conference on Uncertainty in AI*, páginas 291–300.

Hajishirzi, H. y T. Mueller. 2012. Question answering in natural language narratives using symbolic probabilistic reasoning. En G. Michael Youngblood y Philip M. McCarthy, editores, *Proc. of the 25th International Florida Artificial Intelligence Research Society Conference*, páginas 38–43.

Li, B., D. Appling, S. Lee-Urban, y M. Riedl. 2012. Learning sociocultural knowledge via crowdsourced examples.

Li, B., S. Lee-Urban, y M. Riedl. 2012. Toward autonomous crowd-powered creation of interactive narratives. En *Intelligent Narrative Technologies 5, Papers from the 2012 AIIDE Workshop*, páginas 20–25.

Pustejovsky, J., P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, y M. Lazo. 2003. The timebank corpus. En *Proc. of Corpus Linguistics*, páginas 647–656.

Regneri, M., A. Koller, y M. Pinkal. 2010. Learning script knowledge with web experiments. En *Proc. of the 48th Annual Meeting of the Association for Comput. Linguist.*, páginas 979–988, Stroudsburg, USA.

Ricci, F. y R. Wietsma. 2006. Product reviews in travel decision making. *Proceeding of Information and Communication Technologies in Tourism*, páginas 296–307.

Roberto, J., M. Salamó, y M. Martí. 2014. Genre-based stages classification for polarity analysis. *Dialogue and Discourse*, (en proceso de revisión).

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, y M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307, Junio.

Vildjiounaite, E., O. Kocsis, V. Kyllönen, y B. Kladis. 2007. Context-dependent user modelling for smart homes. En C. Conati K. McCoy, y G. Paliouras, editores, *User Modeling*, volumen 4511 de *Lecture Notes in Computer Science*, páginas 345–349. Springer.

A Anexo 1: Identificación de las secuencias narrativas del Ejemplo (1)

o_1: My dad bought me a Saturn for my graduation in 1992 before all the marketing hype (how embarrassing to be constantly asked if I went to Tennessee!).
o_2: Shortly thereafter the problems started.
o_3: With a little research online you can find plenty of evidence that Saturns have a history of excessive oil consumption.
o_4: My 92 SL2 is a total lemon - blown head gasket, six alternators and a plethora of other problems with less than 60,000.
o_5: The dealer has not been helpful - saying that a blown head gasket at 57,000 miles is not uncommon (well for Saturns maybe!).

Figura 3: La entrada al Algoritmo 1 lo constituye las oraciones del *review* (R), es decir, $R = \{o_1, o_2, o_3, o_4, o_5\}$.

root(ROOT-0, started-5)
advmod(thereafter-2, Shortly-1)
advmod(started-5, thereafter-2)
det(problems-4, the-3)
nsubj(started-5, problems-4)

Figura 4: Mediante el parser de Stanford se obtienen las dependencias con los pares núcleo-modificador (el ejemplo corresponde a o_2). (Ver Algoritmo 1 línea 8).

o	**p**	**crefs**$_{id}$	**crefs**$_{ent}$
1	1	1	my
1	4	1	I
1	8	1	my
1	25	1	I
4	1	1	my
3	13	2	saturn
5	23	2	saturn
4	3	3	sl2
4	7	3	lemon
4	11	4	gasket
4	17	4	plethora

Tabla 2: Listado de las entidades (crefs$_{ent}$) que comparten un nexo correferencial (crefs$_{id}$) en el *review*. Se especifica la oración (o) a la que pertenecen y la posición que ocupa según el número de palabras en o. (Ver Algoritmo 1 línea 9).

o	**p**	**nominalización**	**verbo**
1	9	graduation	graduate
1	15	marketing	market
1	16	hype	exaggerate
3	4	research	research
3	11	evidence	evidence
3	16	history	record
3	20	consumption	consume
4	10	head	head
5	2	dealer	deal
5	12	head	head

Tabla 3: Listado de las nominalizaciones presentes en el *review*. En la tercera columna tenemos las nominalizaciones y en la cuarta los verbos de los que proceden según el $NomBank$. (Ver Algoritmo 1 línea 10).

o	**p**	**crefs**$_{id}$	*time*	**eventos**
1	3	1	$\emptyset$	buy
1	9	1	$\emptyset$	graduation
1	16	$\emptyset$	before	hype
1	26	1	$\emptyset$	go
2	5	$\emptyset$	thereafter	start

Tabla 4: Secuencia narrativa ($O_{\mathbb{e}N}$) obtenida mediante la aplicación del Algoritmo 1. En la primera y quinta columnas tenemos las oraciones y los eventos seleccionados por el algoritmo. En la tercera y cuarta columna tenemos la correferencia (crefs$_{id}$) y/o la expresión temporal (*time*) que selecciona cada evento.