Automatización de la extracción de características en tareas de análisis de sentimiento

Juan M. Rodríguez[1,2], Hernán D. Merlino[2], Patricia Pesado[1], Ramón García-Martínez[2]

[1] Programa de Doctorado en Ciencias Informáticas. Facultad de Informática. Universidad Nacional de La Plata. Argentina.
[2] Grupo de Investigación en Sistemas de Información. Departamento de Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús. Argentina.
jmrodriguez1982@gmail.com, hmerlino@gmail.com, ppesado@lidi.info.unlp.edu.ar, rgm1960@yahoo.com

Resumen. El siguiente artículo propone la utilización de un método de extracción de conocimiento para la Web (OIE), en particular ClausIE, para la obtención de características de películas de forma automática. En el marco de métodos de generación de resúmenes automáticos y de análisis de sentimiento, se compara este enfoque con otros dos en los cuales se utilizan pasos manuales para obtener las características de un servicio o producto. Los resultados obtenidos indican que ClausIE puede utilizarse para la extracción de características de forma semiautomática, requiere una intervención manual mínima que se explica en los resultados.

Palabras Clave. Análisis de sentimientos, extracción de características, extracción de conocimiento, extracción de relaciones semánticas, *open information extraction*, procesamiento de lenguaje natural.

1 Introducción

La tarea de realizar minería sobre críticas de cine y obtener un resumen de forma automática consiste en obtener primeramente el par: característica-opinión analizando una por una las oraciones constituyentes de la crítica para luego identificar la polaridad (positiva o negativa) de cada opinión. Y finalmente en construir una lista estructurada basada en las características y opiniones encontradas, promediando las polaridades de cada opinión para cada una de las características [Zhuang et al., 2006]. El presente trabajo se enfoca principalmente en mejorar la primera de las tareas mencionadas, es decir la identificación de características y palabras que expresan opiniones, pero principalmente en las características.

Las características, también llamadas aspectos, son elementos individuales que forman parte de una entidad mayor, siendo cada uno de ellos susceptibles de ser evaluado de forma independiente. Por ejemplo un restaurante tiene las siguientes características: comida, ambiente, servicio, precio. Incluso si se sabe que se está hablando de un restaurante en particular, que ofrece un plato particular como podría

ser: milanesas napolitana con papas fritas, este platillo puede constituir una característica.

La diferencia principal entre el análisis de sentimiento sobre críticas y el resumen automático de críticas con análisis de sentimientos radica en que en el primer caso solo se busca la polaridad global de un texto dado (la crítica) mientras que en el segundo se extraen de dicho texto la características principales y se evalúa la polaridad de cada una individualmente.

Los aspectos juegan un papel importante en el análisis de sentimiento ya que si bien es muy valioso contar con la opinión general sobre una cuestión, la revisión de aspectos o rasgos individuales juega un rol fundamental en la toma de decisiones, un ejemplo clásico es la revisión de un producto, en donde muchas veces un solo aspecto es decisivo para que el usuario decida, por ejemplo, comprarlo (típicamente el precio y/o la calidad).

En este trabajo se puso el foco en la extracción de características. Se buscó una solución automática basada en el uso de un método de extracción de conocimiento creado para la Web, o como se los llama en inglés: *Open Information Extraction* (OIE). En particular se planteó una solución basada en el método ClausIE [Del Corro&Gemulla, 2013].

1.1 Introducción a los métodos de extracción de conocimiento creados para la Web (OIE)

Extracción de conocimiento es cualquier técnica mediante la cual un proceso automatizable es capaz de analizar fuentes de información no estructurada, como por ejemplo textos escritos en lenguaje natural y extraer el conocimiento allí embebido para representarlo de una manera estructurada, manipulable en procesos de razonamiento automático, como por ejemplo: una regla de producción o un subgrafo en una red semántica. A la información obtenida como salida de este tipo de procesos se la llama: pieza de conocimiento [García-Martínez & Britos, 2004; Gómez et al., 1997]

En el año 2007 Michele Banko introduce un nuevo concepto en materia de extracción de conocimiento, al que llama en inglés: *Open Information Extraction* (OIE). Se trata de un paradigma de extracción de conocimiento en donde un sistema informático realiza una sola pasada sobre el total de las fuentes de información no estructurada en formato de lenguaje natural (llamado *corpus* de documentos), dadas como entrada y extrae un gran conjunto de tuplas relacionales sin requerir ningún tipo de participación humana. En el mismo trabajo Banko presenta un método llamado TEXT RUNNER, el cual es el primer método que trabaja dentro de este nuevo paradigma [Banko et al., 2007].

A partir de este trabajo se propusieron otros métodos de extracción de conocimiento bajo el paradigma que Banko llamó *Open Information Extraction* y que podríamos identificar de forma más concreta como métodos de extracción de conocimiento para la Web.

Los métodos de extracción de relaciones semánticas que trabajan de acuerdo con el paradigma anterior (OIE) devuelven una tupla para cada relación semántica

descubierta. La tupla tiene la forma (Entidad 1, Relación, Entidad 2), donde las entidades suelen ser objetos bien identificados, personas, lugares, empresas, fechas, etc., y la relación es la relación semántica entre las dos entidades, por lo general información fáctica del tipo: "Quién hizo qué a quién". Para ilustrar esto, considérese la siguiente oración en idioma inglés:

"Albert Einstein, que nació en Ulm, ha ganado el Premio Nobel".

Extrayendo las relaciones semánticas presentes en la oración y expresándolas como una tupla en la forma: "(Entidad 1, Relación, Entidad 2)" obtenemos lo siguiente:

- (Albert Einstein, ha ganado, el Premio Nobel)
- (Albert Einstein, nació en, Ulm)

1.2 Método escogido: ClausIE

En [Rodríguez et. al., 2015] se realizó una investigación documental sobre distintos métodos de extracción de relaciones semánticas para la Web y se encontró que ClausIE era, según sus autores [Del Corro&Gemulla, 2013] el método que lograba una mejor precisión. Esta aseveración fue puesta a prueba en [Rodríguez et. al., 2016] en donde se hizo una publicación parcial del resultado de una evaluación comparativa entre ClausIE y otros métodos de extracción de información similares (ReVerb y OLLIE). Una versión definitiva de los resultados se encuentra en proceso de publicación. Pero estos serían favorables a ClausIE, razón por la cual se escogió dicho método para este trabajo.

2 Trabajos relacionados

En [Blair-Goldensohn et al., 2008] utilizaron un método híbrido para extraer las características de las críticas, consistente en un método dinámico y uno estático de extracción. Buscaron sustantivos o sustantivos compuestos de hasta tres palabras y que aparecieran en ciertas frases que indicaban una carga de sentimientos (polaridad) y/o que respetaran ciertos patrones sintácticos que eran indicadores posibles de una opinión. Encontraron que los patrones eran más precisos que la ocurrencia de los sustantivos en frases con carga de sentimientos. El patrón más productivo que tuvieron buscaba secuencias de sustantivos que tuvieran inmediatamente antes un adjetivo, así encontraron por ejemplo frases como "…*great fish tacos*...", en críticas sobre restaurantes. Incluyeron "*fish tacos*" (tacos de pescado) como una característica, ya que este era un platillo característico de los restaurantes que habían sido evaluados en las críticas.

El segundo enfoque, el método estático para la extracción de características, consistió en lo siguiente: tomaron al azar 1500 oraciones de críticas sobre hoteles y restaurantes y las etiquetaron de forma manual indicando las características "de grano-grueso" que hallaron en ellas. Las llamaron características de grano-grueso

porque eran características generales que podían ser aplicadas a cualquier restaurante u hotel, no eran tan especificas como por ejemplo: "taco de pescado". Las características fueron las siguientes: comida, ambiente, servicio y precio para restaurantes. Para los hoteles usaron las características: habitaciones, ubicación, comedor, servicio y precio. También incluyeron una categoría *otras*, para etiquetar sentencias que no incluyeran ninguna de las anteriores. Luego etiquetaron un clasificador y lo entrenaron con el conjunto de casos etiquetados, finalmente utilizaron el clasificador ya entrenado para detectar aspectos en cualquier otra oración.

En [Zhuang et al., 2006] se llevó a cabo un experimento similar al propuesto en este artículo, se realizó un resumen automático de críticas cinematográficas de IMDB, focalizado en encontrar opiniones sobre las características de una película dada. Los autores definieron a una característica de película como un elemento (puesta en escena, música, etc.) o bien como personas (director, actor, etc.) mencionadas en una opinión. Los autores definieron de forma manual la lista de características principales (de tipo elemento) que son relevantes en una película y para las características asociadas a personas usaron el elenco completo de participantes tal y como está publicado en IMDB para una película dada.

Las características de tipo elemento escogidas de forma manual fueron las siguientes seis:

- OA: general
- ST: guión
- CH: diseño de personajes
- VP: efectos visuales
- MS: efectos de sonido y música
- SE: efectos especiales

Cada característica fue asociada a múltiples palabras claves, por ejemplo la característica "guion", fue asociada a las diferentes palabras claves en ingles: *story, plot, script, storyline, dialogue, screenplay, ending, line, scene, tale*. Para obtener las palabras claves, se trabajó con un conjunto de datos conformado por 1100 críticas cinematográficas de IMDB etiquetadas de forma manual. Luego las palabras claves asociadas a una característica las obtuvieron al quedarse solo con las más frecuentes.

3 Problemas encontrados

Los mismos autores de [Blair-Goldensohn et al., 2008] encontraron un problema fundamental con el primer enfoque, el del método dinámico y fue que los aspectos encontrados son de grano fino. No es trivial deducir que "sopa de pescado" y "sopa de langosta" forman parte de un aspecto mayor que podría ser: "sopas", "entradas" o "comida".

Respecto al segundo enfoque, el clasificador logró una precisión bastante alta, por ejemplo obtuvo un 86.9 % para la "servicio" y 90.3 % para "precio" en el caso de restaurantes. En el caso de hoteles logró un 83.9% de precisión para "servicio" y 83.3

% para “precio”. La exhaustividad (*recall* en inglés) fue un poco más baja, estuvo entre un 54.5 % y un 69.7 % para los casos mencionados. Sin embargo este método tiene la desventaja de necesitar un conjunto de casos etiquetados de forma manual.

El principal problema asociado al trabajo de [Zhuang et al., 2006] es la necesidad de conocer el conjunto de características relevantes de ante mano para poder generar el etiquetamiento manual.

4 Solución propuesta

Para la elaboración de las pruebas experimentales se utilizó un conjunto de datos de 2000 críticas cinematografías extraídas del sitio IMDB y etiquetadas a mano en dos conjuntos: un grupo de 1000 críticas positivas y otro de 1000 críticas negativas. El conjunto de datos fue creado originalmente por Pang y Lee en [Pang et al., 2002] para entrenar un clasificador de textos con el objetivo de realizar una tarea de análisis de sentimientos. Desde entonces el conjunto de datos ha estado disponible en la web y ha sido utilizado en otras publicaciones.

4.1 Obtención de características

Sobre el conjunto de datos se ejecutó el método de extracción de relaciones semánticas ClausIE. ClausIE devuelve por cada relación semántica una tupla de la forma: (Entidad 1, Relación, Entidad 2) en donde “Entidad” es cualquier elemento sintáctico que haga referencia a algo concreto: una persona, un lugar, una marca, etc. Aunque también puede ser una fecha u otro tipo de entidad más bien abstracta. ClausIE utiliza un algoritmo de detección de nombre de entidades para ello (NER por sus siglas en inglés). Se conjeturó que las características de una película tendrían que poder ser detectadas como entidades. Y en un corpus medianamente grande estas se repetirían con una frecuencia superior a otras entidades posibles. Por lo menos las características llamadas de grano-grueso según [Blair-Goldensohn et al., 2008].

Las extracciones semánticas obtenidas se ordenaron por la cantidad de veces que se repetía una “Entidad” inicial. Luego se filtraron los resultados para mostrar solo los que comienzan con el artículo en inglés “the”, de esta forma se evitó listar pronombres y otras palabras de uso frecuente. La lista obtenida se muestra en la tabla 1.

Tabla 1. Repeticiones de la primera entidad que comienzan con “the”

Entidad 1	Repeticiones
the film	3538
the movie	1637
the story	683
the plot	501
the audience	396
the script	387

the characters	320
the director	258
the two	234
the filmmakers	197
the acting	192
the actors	184
the camera	147
the world	143
the dialogue	140
the cast	128
the man	123
the ending	114
the music	112
the scenes	101
the result	100
the performances	99
the special effects	99

La lista que se muestra en la tabla 1 se corresponde muy bien con una lista de características, o según la nomenclatura de [Zhuang et al., 2006] pablaras claves que indican características. Sin embargo esta lista requirió de dos pasos manuales, por lo cual su generación no fue completamente automática. Estos pasos fueron los siguientes:

- Un corte de forma arbitraria en 99 repeticiones, no se tomaron más elementos que los que aprecian hasta 99 veces.
- La eliminación manual de algunas entidades que si bien se repetían no corresponden a características de una película: “the two”, “the camera”, “the world”, “the man” (marcadas en negrita)

Si se compara la lista encontrada con la lista de palabras claves de características que se presenta en [Zhuang et al., 2006], se observa que hay 12 palabras en común de un total de 38. Sin embargo en lista de la tabla 1 hay 8 palabras de uso frecuente que no fueron usadas en el trabajo de [Zhuang et al., 2006]. Por último hay que indicar que con las 12 palabras en común encontradas, se cubren todas las características de grano-grueso definidas en [Zhuang et al., 2006], aunque en algún conjunto solo quede una palabra clave. Esto se refleja en la tabla 2.

Tabla 2. Características de grano grueso y sus palabras clave asociadas en [Zhuang et al., 2006].

Características	Palabras clave
OA	**film, movie**
ST	**story, plot, script**, storyline, **dialogue**, screenplay, **ending**, line, **scene**, tale
CH	**character**, characterization, role

VP	**scene**, fight-scene, action-scene, action-sequence, set, battle-scene, picture, scenery, setting, visual-effects, color, background, image
MS	**music**, score, song, sound, soundtrack, theme
SE	**special-effects**, effect, CGI, SFX

En negrita se muestra las 12 palabras claves en común. Las otras palabras claves encontradas pertenecerían a las características de grano grueso OA y CH, según la siguiente lista:

- **CH:** acting, actors, cast, performances
- **OA:** director, audience, filmmakers, results

4.2 Análisis de sentimiento de cada característica

Para el siguiente análisis se tomó la lista de la tabla 1, sin contar las palabras filtradas, como una lista de características ya que el objetivo de este artículo es la obtención de características de forma automática a partir de un conjunto de críticas en lenguaje natural. Para cada característica se realizó una tarea de análisis de sentimiento utilizando el lexicón de sentimientos SentiWordnet 3.0 [Baccianella et al., 2010].

Se procedió del siguiente modo: se recuperaron todas las extracciones semánticas para una crítica dada, luego se unió cada extracción en una sola oración concatenando "Entidad 01" con "Relación" con "Entidad 02". Si en la oración resultante aparecía alguna de las características de la lista se evaluaba la misma utilizando el diccionario SentIWordNet 3.0. Luego según el resultado de la polaridad obtenido, positivo o negativo, se marcó dicha característica con un 1 o un -1 en una tabla de resultados final.

Finalmente con aquellas críticas para las cuales se encontraron características, se sumaron los valores de las polaridades de cada una de ellas para obtener un resultado o polaridad global. Este último paso se realizó con el fin de comparar el análisis de las características, el cual en conjunto debería ser idéntico al análisis global. Si esto no hubiese sido así, las características no habrían sido representativas de la película o bien el cálculo de su polaridad tendría que haber sido erróneo.

5 Resultados y Conclusiones

La precisión global para el análisis de sentimientos (más específicamente la obtención de la polaridad), usando SentiWordNet 3.0 sobre las 2000 críticas cinematográficas es 0,662; son 1324 críticas categorizadas correctamente. Este es el piso, sobre el cual se cimentan los análisis de características, un piso bajo, sobre todo al comprar los resultados obtenidos con métodos de clasificación supervisada como los que se usaron en [Pang et al., 2002].

Solo en 1187 críticas se encontró al menos una característica para poder analizar, lo que equivale al 59% de las mismas.

La suma de las polaridades positivas y negativas de cada una de las características, para obtener la polaridad global de la crítica arrojó una precisión de 0.619, es decir 735 clasificadas correctamente de las 1187 que tenían al menos una característica. Si bien es un número bajo, es una precisión cercana a la precisión global de SentiWordNet 3.0. Sobre ese mismo segmento de críticas (las 1187 que tienen al menos una característica) SentiWordNet 3.0 obtuvo por su cuenta una precisión de 0.666, es decir un total de 790 correctamente clasificadas.

Sin embargo la precisión promedio obtenida fue mayor que la calculada en [Zhuang et al., 2006], en donde se calculó la precisión promedio de diferentes pares de características-opiniones para distintas películas siendo el promedio total de la precisión 0.483. Sin embargo, al ser diferente el conjunto de críticas utilizado (el de los autores no está disponible) y diferente la forma de analizar la polaridad, las precisiones no son directamente comparables. Se la cita solo a modo referencia.

Por último, el principal resultado positivo es la extracción de forma casi automática (con una intervención manual mínima) de las características de un producto o servicio (en este caso películas). Las características pueden no ser exhaustivas, al comprarlas con las utilizadas en el trabajo de [Zhuang et al., 2006] pero son representativas y sin duda utilizadas con mayor frecuencia en el conjunto de datos analizado. El análisis de sentimiento sobre las características individuales, no mejora el rendimiento global del método utilizado (en este caso el lexicón de sentimientos SentiWordNet) pero se mantiene coherente con la precisión del mismo.

6 Futuras líneas de investigación

Queda como trabajo futuro la revisión y comparación de este enfoque con otros métodos de extracción de características automáticos como SABRE [Caputo et al., 2017].

Referencias

Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 2200-2204).

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007, January). Open information extraction for the web. In IJCAI (Vol. 7, pp. 2670-2676).

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A., & Reynar, J. (2008, April). Building a sentiment summarizer for local service reviews. In WWW workshop on NLP in the information explosion era (Vol. 14, pp. 339-348).

Caputo, A., Basile, P., de Gemmis, M., Lops, P., Semeraro, G., & Rossiello, G. (2017). SABRE: A Sentiment Aspect-Based Retrieval Engine. In Information Filtering and Retrieval (pp. 63-78). Springer International Publishing.

Del Corro, L., & Gemulla, R. (2013, May). ClausIE: clause-based open information extraction. In Proceedings of the 22nd international conference on World Wide Web (pp. 355-366). International World Wide Web Conferences Steering Committee.

García-Martínez, R. & Britos, P. V. (2004). Ingeniería de sistemas expertos. Nueva Librería. ISBN 987-1104-15

Gómez, A., Juristo, N., Montes, C., & Pazos, J. (1997). Ingeniería del conocimiento. Editorial Centro de Estudios Ramón Areces. ISBN 84-8004-269-9.

Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

Rodríguez, J. M., Merlino, H., García-Martínez, R. (2015). Revisión Sistemática Comparativa de Evolución de Métodos de Extracción de Conocimiento para la Web. XXI Congreso Argentino de Ciencias de la Computación (CACIC 2015). Buenos Aires, Argentina.

Rodríguez, J. M., Merlino, H. D., Pesado, P., & García-Martínez, R. (2016, August). Performance Evaluation of Knowledge Extraction Methods. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 16-22). Springer International Publishing.

Zhuang, L., Jing, F., & Zhu, X. Y. (2006, November). Movie review mining and summarization. In Proceedings of the 15th ACM international conference on Information and knowledge management (pp. 43-50). ACM.