Extracción de opiniones sobre características adaptable al dominio *

Fermín L. Cruz, José A. Troyano, F. Javier Ortega and Fernando Enríquez

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla
Av. Reina Mercedes s/n 41012, Sevilla (Spain)
{fcruz,troyano,javierortega,fenros}@us.es

Resumen La extracción de opiniones sobre características es una tarea relacionada con la minería de opiniones, que consiste en extraer a partir de textos opiniones individuales acerca de las características de un objeto determinado. En los últimos años, esta tarea ha sido abordada desde una perspectiva no supervisada y sin concretar un dominio de aplicación específico. Nuestra propuesta, sin embargo, se centra en el desarrollo de un sistema de extracción que tenga en cuenta las particularidades de cada dominio de aplicación, y que se adapte con facilidad a los distintos dominios mediante la definición de una serie de recursos específicos. Los experimentos realizados muestran que el conocimiento aportado por estos recursos supone una valiosa ayuda para la construcción de sistemas precisos de extracción de opiniones.

1. Introducción

Durante los últimos años, la extracción de opiniones sobre características de productos ha sido estudiada en varios trabajos. La primera definición del problema se encuentra en [4]:

> Given a set of customer reviews of a particular product, the task involves three subtasks: (1) identifying features of the product that customers have expressed their opinions on (called product features); (2) for each feature, identifying review sentences that give positive or negative opinions; and (3) producing a summary using the discovered information.

Esta definición ha sido la base de distintos trabajos de dichos autores y otros investigadores [5],[8],[7],[10],[3]. En todos estos trabajos se aborda la tarea desde una perspectiva general, sin concretar ningún dominio de aplicación. Los sistemas tratan de identificar las menciones a características de los productos y las palabras que expresan la opinión de los usuarios sin tener en cuenta las particularidades del producto. Además, estos trabajos se apoyan en su mayoría

* Parcialmente financiado por el Ministerio de Educación y Ciencia (HUM2007-66607-C04-04).

en métodos automáticos, sin hacer apenas uso de ningún recurso manual que aporte conocimiento de calidad al sistema de extracción.

En contraposición a estos trabajos, nuestra propuesta se basa en un acercamiento más aplicado al problema: (1) antes de construir el sistema, se debe concretar un dominio de aplicación; (2) sólo se tendrán en cuenta las características incluidas en una taxonomía específica; y (3) el sistema de extracción se apoyará en un conjunto de recursos específicos del dominio de aplicación, generados automáticamente a partir de un conjunto de documentos anotados.

2. Nuestra propuesta

En nuestro trabajo, hemos partido de una definición distinta de la tarea de extracción de opiniones sobre características: dado un conjunto de documentos de opinión de un dominio concreto, se trata de reconocer las opiniones vertidas acerca de una serie de características opinables disponibles para el dominio, y clasificar las opiniones reconocidas según la polaridad (positivas/negativas). Llamamos a la primera tarea *reconocimiento de opiniones* y a la segunda *clasificación de opiniones*. En nuestro planteamiento es fundamental la participación como entrada al sistema de las características en las que estamos interesados. Dichas características son descritas en una taxonomía, en la que además se dispone de relaciones de especialización/generalización entre las mismas. Por ejemplo, la característica *sound quality* para el dominio *headphones* puede descomponerse a su vez en varias características, por ejemplo *bass*, *mids* y *highs*. La construcción de la taxonomía de características es un paso previo que es llevado a cabo de manera semiautomática, utilizando un algoritmo de *bootstrapping* a partir de un conjunto de documentos de opinión del dominio y con la participación de un experto.

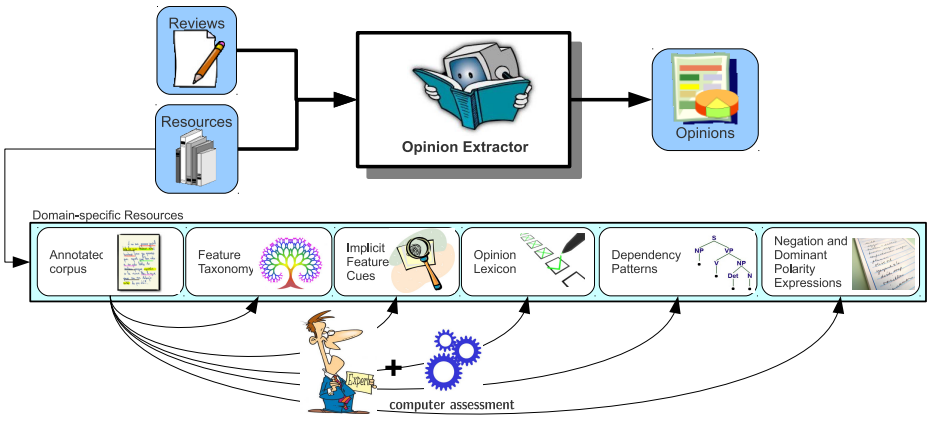

Figura 1. Esquema conceptual de nuestra propuesta

Uno de los pilares de nuestra propuesta es la generación de recursos dependientes del dominio, que facilitan la tarea de extracción de las opiniones. Dichos recursos incluyen, entre otras cosas, patrones de dependencias sintácticas que permiten conectar las *palabras de característica* (menciones a las características de la taxonomía) con las *palabras de opinión* (palabras a partir de las cuáles se decide la polaridad de una opinión), lexicones de palabras de opinión del dominio (incluyendo estimaciones de la orientación semántica de las mismas y de la probabilidad de ser utilizadas en una opinión) y listas de términos indicativos de características implícitas (una característica implícita es aquella que no aparece mencionada en el texto, sino que se deduce de las palabras de opinión utilizadas, como ocurre por ejemplo con la palabra de opinión *expensive*, que asociaremos a la característica *price*). Los recursos son inducidos a partir de un conjunto de documentos anotados. Dado que el proceso de anotación de las opiniones puede ser costoso, hemos investigado métodos para facilitar dicho proceso, así como algoritmos de ampliación automática que nos permiten construir los recursos a partir de unos pocos documentos anotados y un conjunto mayor de documentos sin anotar. Además, en la construcción del sistema extractor hemos incluido implementaciones independientes del dominio (y que por tanto no hacen uso de los recuros) de algunas de las subtareas identificadas. De esta forma pretendemos, por un lado, permitir la construcción rápida de sistemas para nuevos dominios y, por otro lado, evaluar la aportación real de los recursos a la resolución de la tarea. En [2] se describen con mayor detalle los recursos, los métodos utilizados para su generación y la arquitectura del sistema extractor.

3. Experimentación

En este apartado incluimos algunos resultados experimentales realizados sobre un dominio concreto (*headphones*), con una taxonomía de 31 características. Para su realización, dispusimos de un corpus de 587 documentos anotados de análisis (*reviews*) de auriculares, en inglés, extraídos de Epinions.com. El corpus utilizado, incluyendo también los dominios *hotels* y *cars*, está disponible para uso público[1]. Los experimentos se realizaron usando validación cruzada sobre diez particiones, utilizando en cada una de las ejecuciones una parte para evaluación y el resto para la generación de los recursos.

En la tabla 1 mostramos los resultados obtenidos por cinco aproximaciones distintas. Las tres primeras no hacen uso de los recursos del dominio; en su lugar, usan un método léxico basado en ventanas para enlazar las palabras de característica y de opinión, y clasifican la polaridad de las opiniones utilizando distintos algoritmos de la literatura : información mutua entre términos de opinión y semillas (algoritmo PMI-IR) [9], cálculo de distancias en WordNet [6], y el recurso léxico SentiWordNet [1]. El cuarto sistema se basa en los recursos del dominio para llevar a cabo la extracción. Finalmente, el quinto sistema representa un acercamiento híbrido en el que se utilizan componentes basados en recursos y algunos independientes del dominio. Como se puede observar, los

[1] http://www.lsi.us.es/ fermin/index.php/Datasets

sistemas que hacen uso de los recursos mejoran significativamente los resultados de los sistemas que no hacen uso de ellos.

	Opinion Recognition			Opinion Classification	Opinion Recognition + Classification		
Experimento	p	r	$F_{\frac{1}{2}}$	accuracy	p	r	$F_{\frac{1}{2}}$
PMI-IR	0,6092	0,3039	0,5073	0,8706	0,5512	0,2754	0,4593
WordNet	0,6756	0,3002	0,5405	0,8940	0,6111	0,2720	0,4892
SentiWordnet	0,6744	0,3643	0,5763	0,8688	0,5972	0,3230	0,5105
Resource-based	0,7869	0,5662	0,7300	0,9503	0,7557	0,5436	0,7010
Hybrid	0,7836	0,5736	**0,7301**	**0,9572**	0,7573	0,5543	**0,7056**

Cuadro 1. Resultados obtenidos por los *pipelines* basados en recursos

Referencias

1. S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).
2. F. L. Cruz, J. A. Troyano, F. Enríquez, J. Ortega, and C. G.Vallejo. A knowledge-rich approach to feature-based opinion extraction from product reviews. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, pages 13–20. ACM, 2010.
3. X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 231–240, New York, NY, USA, 2008. ACM.
4. M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, New York, NY, USA, 2004. ACM.
5. M. Hu and B. Liu. Mining opinion features in customer reviews. In *Proceedings of AAAI*, pages 755–760, 2004.
6. J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke. Using wordnet to measure semantic orientation of adjectives. In *National Institute for*, volume 26, pages 1115–1118, 2004.
7. B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of WWW*, 2005.
8. A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 2005.
9. P. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 417–424, 2002.
10. L. Zhuang, F. Jing, X. Zhu, and L. Zhang. Movie review mining and summarization. In *Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM)*, 2006.