Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 157-162 recibido 09-05-12 revisado 04-06-12 aceptado 06-06-12

Learning a Statistical Model of Product Aspects for Sentiment Analysis[*]

Aprendizaje de un Modelo de Características de Productos para el Análisis de Opiniones

Lisette García-Moya
Universitat Jaume I
Castellon, Spain
lisette.garcia@uji.es

Rafael Berlanga Llavori
Universitat Jaume I
Castellon, Spain
berlanga@uji.es

Henry Anaya-Sánchez
Universitat Jaume I
Castellon, Spain
henry.anaya@uji.es

Resumen: En este artículo se introduce una nueva metodología para modelar características de productos a partir de una colección de opiniones de usuarios. La metodología propuesta se basa en modelos estadísticos de lenguajes y es aplicable a productos de dominio arbitrario. La metodología combina un kernel de palabras de opinión con un modelo de traducción de palabras para estimar el modelo de características. Se presenta además un método para modelar las opiniones vertidas sobre las características. Los experimentos realizados sobre diferentes colecciones de opiniones muestran resultados alentadores en el modelado tanto de características como de opiniones vertidas sobre éstas.

Palabras clave: Minería de Opiniones, Análisis de Opiniones, Modelado de Características de Productos

Abstract: In this paper, we introduce a new methodology for modeling product aspects from a collection of free-text customer reviews. The proposal relies on a language modeling framework and is domain independent. It combines both a kernel-based model of opinion words and a stochastic translation model between words to approach the aspect model of products. We also present a ranking-based methodology to model the sentiments expressed about the aspects. The experiments carried out over several collections of customer reviews show encouraging results in the modeling of product aspects and their sentiments even from individual customer reviews.

Keywords: Opinion Mining, Sentiment Analysis, Product Aspect Modeling

1. *Introduction*

With the increasing availability of user-generated contents, such as consumer opinion web sites, blogs, Internet forums and social networks, people have more opportunities to express their opinions and make them available to everyone. Publicly available opinions provide valuable information for decision-making processes based on a new collective intelligence paradigm designated as crowdsourcing. The consumer opinion web sites constitute an invaluable way of promotion in which satisfied customers tell other people how much they like a business, product, service, or event. It has became one of the most credible forms of advertising because people who do not stand to gain personally by promoting something put their reputations on the line every time they make a recommendation.

Therefore, in the last years the computational treatment of sentiment and opinions has been viewed as a challenging area of research that can serve different purposes.

One of the most relevant applications of sentiment analysis is the aspect-based summarization (Carenini, Ng, and Pauls, 2006; Yu et al., 2011). Broadly speaking, given a collection of opinion posts about a product or service, this task is aimed at obtaining the most relevant opined aspects (also called features) along with their most relevant

[*] This work has been partially funded by the “Ministerio de Economía y Competitividad” with contract number TIN2011-24147 and by the Fundació Caixa Castelló project P1-1B2010-49. Lisette García-Moya has been supported by the PhD Fellowship Program of the Universitat Jaume I (PREDOC/2009/12).

sentiment information expressed by customers (usually an opinion word and/or a polarity score). Aspect-based summarization is usually composed of three main tasks: *aspect identification*, *sentiment classification*, and *aspect rating.* Aspect identification is focused on extracting the set of aspects concerning the product from the reviews. The word *aspects* is intended to represent both components and attributes. For example, given the sentence, "The bed was comfortable", the review is about the "bed" aspect and the opinion is positively expressed by mean of the opinion word "comfortable". The sentiment classification task consists in determining the opinions about the product aspects and/or their polarities, whereas aspect rating leverages the relevance of aspects and their opinions to properly present them to the users.

In this paper, we address the aspect-based summarization task by introducing a domain independent methodology for modeling product aspects from a set of free-text customer reviews. The proposal relies on a language modeling framework, which combines both a probabilistic model of opinion words and a stochastic self-translation model between words to approach the aspect model of products. From the proposed method, we also derive a ranking-based method to approximate the sentiments expressed about the aspects.

Our work extends the preliminary approach introduced in García-Moya et al. (2012). Specifically, in the present work we propose a more general methodology that effectively allows –for example– the use of dependency relations between words in the modeling of product aspects. We carried out experiments on a wider set of review collections, including a taxonomy-based opinion dataset that entails a harder aspect modeling task. Finally, we also provide an evaluation of the proposed ranking-based method.

As already shown in García-Moya et al. (2012), one strong point of our proposal is that it can effectively retrieve the product aspects even if we do not rely on NLP techniques. This is the main difference with respect to most of the approaches on sentiment analysis that consider the task of aspect identification (Wu et al., 2009; Qiu et al., 2009), as they strongly rely on dependency analysis.

2. Modeling Product Aspects

Given a collection of customer reviews about a specific product and a free-text document d –which can be either a subcollection of reviews or an individual review–, our goal is to obtain a probabilistic model for retrieving the product aspects from d.

Specifically, we consider modeling the set of aspects discussed in d as a statistical language model that assigns higher probability values to words defining aspects.

Let $V = \{w_1, \ldots, w_n\}$ represents the vocabulary of d. Let also $Q = \langle Q(w_1), \ldots, Q(w_n)\rangle^\top$ and $\mathcal{T} = \{\mathrm{p}(w_i|w_j)\}_{1\leq i\leq n, 1\leq j\leq n}$ be a vector-shaped model of opinion words and an n-by-n (column-wise) stochastic matrix representing an entailment-based self-translation model of words from d respectively.

Then, we propose to model unigram product aspects as follows:

$$P(w_i) \propto \left(\begin{pmatrix} \mathrm{p}(w_1|w_1) & \cdots & \mathrm{p}(w_1|w_n) \\ \vdots & \ddots & \vdots \\ \mathrm{p}(w_n|w_1) & \cdots & \mathrm{p}(w_n|w_n) \end{pmatrix}^k \cdot \begin{pmatrix} Q(w_1) \\ \vdots \\ Q(w_n) \end{pmatrix}\right)_i \tag{1}$$

$$= \left(\mathcal{T}^k \cdot Q\right)_i \tag{2}$$

where $k > 0$ is the number of times that the stochastic translation $\mathcal{T}$ is applied to the opinion model Q.

In the context of customer reviews, opinion words (e.g., "excellent", "terrible", etc.) are usually utilized to express sentiments about the different aspects of a product. This causes the review texts to reflect some entailment relationship from opinion words to aspect words. The idea behind the above model is that by applying successively the entailment model $\mathcal{T}$ to Q, we can capture such an entailment relationship between opinion and aspects, and define in this way a model of aspect words (García-Moya et al.,2012).

The unigram language model of aspects $P = \{P(w_i)\}_{1\leq i\leq n}$ can be extended to generate aspects of arbitrary length from the model:

$$P^*(s) = \prod_{t=1}^{r} P(w_{i_t})^{1/r} \tag{3}$$

where $s = w_{i_1} \ldots w_{i_r}$ $(r > 0)$.

In addition, we consider refining the unigram model P to avoid the assignment of high probability values to meaningless words (e.g., prepositions, conjunctions, etc.). The refined unigram language model P' is obtained by performing an *Expectation Maximization* process aimed at maximizing the cross entropy:

$$-\sum_{i=1}^{n} P(w_i)\log(\lambda P'(w_i)+(1-\lambda)P_{bg}(w_i)) \tag{4}$$

where P_{bg} is a background language model of the source language of the reviews (e.g. English). Currently, we estimate P_{bg} from the COCA corpus (Davies, 2011).

The estimation of both the self-translation model $\mathcal{T}$ and the opinion model Q are described in the next sections.

3. Self-translation Model $\mathcal{T}$

For all $i, j \in \{1,\ldots,n\}$, we define $\mathrm{p}(w_i|w_j)$ to be proportional to the number of times word w_i occurs in a local context of words from d containing an occurrence of w_j. In this way,

$$\mathrm{p}(w_i|w_j) = \frac{\mathrm{p}(w_i,w_j)}{\mathrm{p}(w_j)} \tag{5}$$

where:

$$\mathrm{p}(w_i,w_j) \propto \sum_{l\in\mathcal{L}} \mathrm{p}(w_i|l)\cdot\mathrm{p}(w_j|l)\cdot\mathrm{p}(l) \tag{6}$$

$$\mathrm{p}(w_j) = \sum_{w_i\in V} \mathrm{p}(w_i,w_j) \tag{7}$$

$\mathcal{L}$ is the set of all local contexts of words contained in d, $\mathrm{p}(w_i|l) = |l|_{w_i}/|l|$ and $\mathrm{p}(l) = |\mathcal{L}|^{-1}$ ($|l|_{w_i}$ is the number of times w_i occurs in l, and $|l|$ is the number of words contained in l).

In this paper, we consider two alternatives for defining local contexts. The first one defines local contexts as the N-grams occurring in the sentences of d. Given a bag D of dependency relations observed among word occurrences in d, the second alternative defines local contexts as the word tuples of D.

4. Modeling Opinion Words

We rely on a kernel-based density estimation approach to define Q from a predefined set of (general-domain) opinion words $\{u_1,\ldots,u_m\}$. Thus, we define:

$$Q(w) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{K}(w,u_i) \tag{8}$$

where $w \in V$ and $\mathrm{K}(w,u_i)$ is the gaussian kernel:

$$\mathrm{K}(w,u_i) = \exp\left(-0{,}5\cdot h(g(w),g(u_i))^2/\sigma^2\right) \tag{9}$$

such that h represents the geodesic distance between distributions (Dillon et al., 2007), $g(v)$ is the posterior distribution of words $\{\mathrm{p}(w_i|v)\}_{1\leq i\leq n}$, and σ is a predetermined distribution width. In our experiments, we set $\sigma = 0{,}3$.

Relying on $\mathcal{T}$, we also propose to rank sentiment words with respect to the sequence $s = w_{i_1}\ldots w_{i_r}$ by regarding the score:

$$R(w) = \prod_{t=1}^{r} p(w_{i_t}|w)^{1/r} Q(w). \tag{10}$$

5. Evaluation

To evaluate our approach, we firstly rely on four collections of customer reviews each one corresponding to a product (Apex AD2600, Canon G3, Nokia 6610 and Norton).[1] These collections of reviews are manually annotated at the sentence level with the relevant product aspects referred to in the text (Hu and Liu, 2004; Ding, Liu, and Yu, 2008).

We compare several aspect language models obtained from our approach (by varying the value of k, and using either N-grams of different sizes or dependency relations to estimate the translation model) to the baseline language model obtained by replacing the model $\{P(w_i)\}_{1\leq i\leq n}$ by the MLE model of words from each product collection.

Since the goal is to measure the effectiveness of the statistical language models on the generation of products aspects, the performance of each language model is measured from the *log-likelihood* of generating the bag of aspects $S = \{s_1,\ldots,s_k\}$ that have been manually annotated in the collection. Specifically, we have considered the shifted likelihood:

$$\ell_{shift}(P^*,S) = \ell(P^*,S) - k\log n \tag{11}$$

where the likelihood $\ell(P^*,S)$ is defined as:

$$\ell(P^*,S) = \sum_{i=1}^{k} \log P^*(s_i) \tag{12}$$

[1] `http://www.cs.uic.edu/~liub/FBS`

Table 1: Language model performance for generating the bag of product aspects.

k	**Model**	**Apex AD2600**		**Canon G3**		**Nokia 6610**		**Norton**	
		review	**product**	**review**	**product**	**review**	**product**	**review**	**product**
	baseline	0.0028	6.2978	0.0065	4.327	0.0087	4.5503	0.0033	1.2247
5	**w2**	0.0139	6.0104	0.0146	4.1133	0.0322	4.0846	0.009	1.4627
	w3	0.0141	7.1017	0.0156	4.572	0.0333	4.7469	0.0094	1.6359
	w4	0.0141	7.3233	0.0157	4.6456	0.0335	4.8683	0.0096	1.6923
	w5	0.014	7.3554	0.0157	4.6598	0.0334	4.8742	0.0096	1.7083
	w6	0.014	7.3191	0.0157	4.6645	0.0333	4.8392	0.0096	1.7071
	drAll	0.2324	-2.6241	0.3045	1.2349	0.3744	-2.3262	0.2435	-2.2707
	drSelected	**0.8955**	**17.0883**	**1.1059**	**15.2788**	**1.5542**	**15.2324**	**0.997**	**13.3926**
10	**w2**	0.0141	7.125	0.0154	4.6545	0.0332	4.7512	0.0094	1.6571
	w3	0.0142	7.428	0.0158	4.7432	0.0335	4.9571	0.0096	1.7134
	w4	0.0141	7.448	0.0158	4.7323	0.0334	4.97	0.0097	1.7294
	w5	0.014	7.4237	0.0158	4.717	0.0332	4.9368	0.0096	1.7263
	w6	0.014	7.3736	0.0157	4.7037	0.0331	4.885	0.0096	1.7182
	drAll	0.2321	-2.5711	0.3038	1.3244	0.3732	-2.346	0.2434	-2.2561
	drSelected	**0.8951**	**17.2074**	**1.1047**	**15.3229**	**1.5518**	**15.0994**	**0.9969**	**13.3938**
15	**w2**	0.0142	7.3454	0.0157	4.7474	0.0335	4.8923	0.0096	1.7059
	w3	0.0142	7.4502	0.0158	4.7613	0.0335	4.9797	0.0097	1.7254
	w4	0.0141	7.4523	0.0158	4.7381	0.0334	4.9837	0.0097	1.7322
	w5	0.014	7.4254	0.0158	4.7203	0.0332	4.9522	0.0096	1.7274
	w6	0.014	7.3752	0.0157	4.7059	0.033	4.9004	0.0096	1.7188
	drAll	0.2323	-2.5589	0.3042	1.3386	0.3738	-2.3192	0.2435	-2.2467
	drSelected	**0.8932**	**17.2486**	**1.1057**	**15.3953**	**1.5537**	**15.2319**	**0.9971**	**13.4158**
20	**w2**	0.0142	7.3952	0.0158	4.7737	0.0336	4.9356	0.0097	1.7236
	w3	0.0142	7.4518	0.0159	4.7642	0.0335	4.9837	0.0097	1.7281
	w4	0.0141	7.4524	0.0158	4.7386	0.0333	4.9873	0.0097	1.7324
	w5	0.014	7.4254	0.0157	4.7206	0.0332	4.9588	0.0096	1.7274
	w6	0.014	7.3753	0.0157	4.706	0.033	4.9079	0.0096	1.7189
	drAll	0.2322	-2.5593	0.304	1.3445	0.3733	-2.3267	0.2434	-2.2461
	drSelected	**0.8952**	**17.2414**	**1.105**	**15.3912**	**1.5522**	**15.2217**	**0.9969**	**13.4179**

The greater the value of this measure, the better the performance of the language model P^*.

In Table 1, we show the performance of each language model. For each product, we include two columns: one measuring the average performance obtained by applying the method to each individual review (column labeled as *review*), and the other one measuring the performance obtained by applying the method to each (entire) product review collection. The label "wN" represents the model obtained by using N-grams as local contexts of words to build the translation model $\mathcal{T}$; whereas the labels "drAll" and "drSelected" refer to models that defines the local contexts from dependency relations. The model "drAll" uses all the dependency relations obtained with the Stanford dependency parser (De Marneffe, MacCartney, and Manning, 2006), while "drSelected" considers only the set of relations {"nn", "acomp", "advmod", "amod", "det", "dobj", "infmod", "iobj", "measure", "nsubj", "nsubjpass", "partmod", "prep", "rcmod", "xcomp", "xsubj"}.

Several observations can be made by analyzing Table 1. Firstly, it can be appreciated that models obtained by instantiating our approach clearly outperform the baseline (except in the case of drAll when applied to the entire collection, and in the case of $w2$ when $k = 5$ for some products). This shows that only relying on the frequency of words is not enough for modeling product aspects from a collection of customer reviews. Secondly, it can be seen that the models based on dependency relations outperforms those models based on N-grams when applied to individual customer reviews; being *drSelected* the best model. However, *drAll* surprisingly performs the worst at collection level. It seems that using arbitrary dependency relations increases the uncertainty associated to the transla-

Table 2: Performance obtained on the Taxonomy-Based Opinion Dataset.

Measure	**Model**	**cars**		**headphones**		**hotels**	
		review	**category**	**review**	**category**	**review**	**category**
$\ell_{\textbf{shift}}(\mathbf{P}^*, \mathbf{S})$	**baseline**	0.0046	0.1419	0.002	0.0344	0.0068	0.0621
	w2	0.012	0.1766	0.012	0.0616	0.0172	0.0888
	w3	0.0117	0.1757	0.0115	0.059	0.0168	0.0875
	w4	0.0116	0.1747	0.0112	0.0573	0.0165	0.0859
	w5	0.0114	0.1739	0.0111	0.056	0.0163	0.085
	w6	0.0113	0.1728	0.0109	0.0551	0.0161	0.0841
	drAll	0.0948	-0.3437	0.2619	0.2727	0.2965	0.2596
	drSelected	**0.8578**	**1.8708**	**0.9015**	**1.6247**	**1.2073**	**1.9802**
MAP	**w2**	0.5769	0.4436	0.6148	0.498	0.5432	0.416
	w3	0.591	0.4506	0.5841	0.4524	0.5331	0.4013
	w4	0.5523	0.4253	0.5631	0.4275	0.5093	0.3802
	w5	0.5231	0.3927	0.5382	0.4031	0.496	0.3763
	w6	0.501	0.3567	0.5261	0.3837	0.503	0.4002
	drAll	0.6994	0.5623	0.6609	0.5455	0.6574	0.5580
	drSelected	**0.7109**	**0.5988**	**0.6903**	**0.5889**	**0.6832**	**0.5942**

tion model $\mathcal{T}$ as d becomes larger. Finally, we can notice that using N-grams of size $3 \leq N \leq 5$ produce overall similar results.

5.1. Ranking Sentiment Words for Product Aspects

A second experiment was focused on evaluating the score function R proposed for ranking sentiment words for each product aspect.

In this case, we consider the Taxonomy-Based Opinion Dataset from Cruz et al. (2010). This dataset consists of three review collections, each one corresponding to a product category (namely, cars, headphones and hotels). For each collection, there is a set of customer reviews about different products in the category. The customer reviews have been manually annotated at the sentence level with the following elements: (i) the product aspects (explicitly or implicitly) referred in the sentence, (ii) the category of each aspect (based on a given taxonomy), and (iii) the sentiment or opinion word associated to each aspect.

To measure the quality of the ranking of sentiment words obtained for each aspect (in a review/category), we consider calculating the *Mean Average Precision* (MAP) of the obtained ranking with respect to the set of expressed sentiments about the aspect.

In Table 2, we show the average performance of drAll and drSelect (using $k = 15$) in both the generation of product aspects and the retrieval of sentiment words for each manually labeled aspect in each category of the Taxonomy-Based Opinion Dataset.

Similar to the previous experiments, the model drSelect performs the best in each case. The values of ℓ_{shift} (measured according to the entire categories) are relatively smaller in this dataset, since it entails a harder task (there is more than one product in each collection/category). This also justify that MAP values in the case of individual reviews are larger than the values obtained in the case of the entire categories.

6. *Conclusion*

In this paper, a new methodology for modeling product aspects from a collection of customer reviews has been presented. The proposed method is based on the language modeling framework and is both domain and language independent. We have also presented a ranking-based methodology to model the sentiments expressed about the aspects. The experiments carried out over several collections of customer reviews (with different degree of difficulty) have shown the usefulness of the proposal for properly modeling the product aspects and retrieving their sentiments even from individual reviews. As future work, we plan to develop a generative routine to produce a set of (multi-word) product aspects likely to be generated from both the language model of aspects and the language model of noun phrases from a set of customer reviews. We also consider to extend our methodology to include modeling the polarity of sentiment words.

References

Carenini, G., R. Ng, and A. Pauls. 2006. Multi-document summarization of evaluative text. In *Proc. of EACL 2006*, pages 305–312.

Cruz, F.L., J.A. Troyano, F. Enríquez, F.J. Ortega, and C.G. Vallejo. 2010. 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.

Davies, Mark. 2011. Word frequency data from the Corpus of Contemporary American English (COCA). Downloaded from http://www.wordfrequency.info on June 01, 2011.

De Marneffe, M.C., B. MacCartney, and C.D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

Dillon, J., Y. Mao, G. Lebanon, and J. Zhang. 2007. Statistical Translation, Heat Kernels, and Expected Distance. In *Proc. of the 23rd Conference on Uncertainty in Artificial Intelligence.*

Ding, Xiaowen, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the international conference on Web search and web data mining*, pages 231–240.

García-Moya, Lisette, Henry Anaya-Sánchez, and Rafael Berlanga-Llavori. 2012. Combining Probabilistic Language Models for Aspect-Based Sentiment Retrieval. In *Proceedings of the 34th European Conference on Information Retrieval*, volume 7224 of *Lecture Notes in Computer Science*, pages 561–564. Springer-Verlag.

Hu, M and B Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177. ACM Press, New York, NY.

Qiu, Guang, Bing Liu, Jiajun Bu, and Chun Chen. 2009. Expanding domain sentiment lexicon through double propagation. In *Proceedings of the 21st international jont conference on Artifical intelligence*, IJCAI'09, pages 1199–1204.

Wu, Yuanbin, Qi Zhang, Xuanjing Huang, and Lide Wu. 2009. Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, pages 1533–1541.

Yu, J., Z.J. Zha, M. Wang, and T.S. Chua. 2011. Aspect ranking: identifying important product aspects from online consumer reviews. In *Proc. of ACL 2011*, pages 1496–1505.