UO_UA: Using Latent Semantic Analysis to Build a Domain-Dependent Sentiment Resource

Reynier Ortega
Adrian Fonseca
Carlos Muñiz
CERPAMID
Ave Patricio Lumumba, S/N
Santigo de Cuba, Cuba
reynier.ortega@cerpamid.co.cu
adrian@cerpamid.co.cu

Yoan Gutiérrez
Andrés Montoyo
DLSI, University of Alicante
Carretera de San Vicente, S/N
Alicante, Spain
ygutierrez@dlsi.ua.es
montoyo@dlsi.ua.es

Abstract

In this paper we present our contribution to *SemEval-2014 Task 4: Aspect Based Sentiment Analysis (Pontiki et al., 2014), Subtask 2: Aspect Term Polarity* for Laptop domain. The most outstanding feature in this contribution is the automatic building of a domain-depended sentiment resource using Latent Semantic Analysis. We induce, for each term, two real scores that indicate its use in positive and negative contexts in the domain of interest. The aspect term polarity classification is carried out in two phases: opinion words extraction and polarity classification. The opinion words related with an aspect are obtained using dependency relations. These relations are provided by the Stanford Parser[1]. Finally, the polarity of the feature, in a given review, is determined from the positive and negative scores of each word related to it. The results obtained by our approach are encouraging if we consider that the construction of the polarity lexicon is performed fully automatically.

1 Introduction

Hundreds of millions of people and thousands of companies around the world, actively use Social Media[2]. Every day are more amazing websites and applications (Facebook, Twitter, MySpace, Amazon, etc.) that allow the easy sharing of information in near real time. For this reason, at present, the Web is flooded with subjective, personal and affective data. Mining this huge volume of information offer both interesting challenges and useful intelligent applications such as recommendation systems (Dong et al., 2013; Sun et al., 2009) and customer's reviews summarization (Bafna and Toshniwal, 2013; Balahur and Montoyo, 2008).

Nowadays, companies have redirected their marketing strategies toward the Web. Each one of them advertises that their products are the best, amazing, easy to use, long lasting and cheap. But are these advertisements really true? Obviously, not everything is true. The companies usually exaggerate the product's quality and in many cases tend not to advertise the limitations of their products. Therefore, taking a rational decision about which product is the best among the variety of existing options can be very stressful.

To avoid this situation, frequently we trust in the experiences gained by others who have purchased the product of our interest, or one similar. The existence of websites like Ciao[3], Epinions[4] and Cnet[5] make possible to the customers to interchange their experiences about a specific product, and to future clients avoid products advertising

However, the existence of a large volume of reviews entails that it is impossible to conduct an effective exploration before making a final decision. The most important benefit of having that amount of user-generated content on hand, specifically product's reviews, is that, these data can be explored by a computer system to obtain information about products and their features.

The task of aspect-based sentiment analysis (Liu, 2012) is a fine-grained level of Sentiment Analysis (Pang and Lee, 2008). This aim to identify the aspects (*e.g., battery, screen, food, service, size, weight, time-life*) of given target entities

---

[1]http://nlp.stanford.edu:8080/parser/
[2]http://en.wikipedia.org/wiki/Social_media

---

[3]www.ciao.com
[4]www.epinions.com
[5]www.cnet.com

(*e.g., laptops, restaurants, camera*) and the sentiment expressed towards each aspect (*e.g., positive, negative, neutral*). This are composed by two basic phases: feature extraction and feature polarity classification.

In this paper we present our contribution for *SemEval-2014 Task 4: Aspect Based Sentiment Analysis (Pontiki et al., 2014), Subtask 2: Aspect Term Polarity*. In this approach we only focus on the polarity classification problem. For this, we induce a domain-dependent sentiment lexicon applying Latent Semantic Analysis (LSA) on product reviews corpus, gathered from Ciao. The classification phase is carried out as follow: the opinion words related with the product aspect are draw out using the dependency relations provided by Stanford Parser, then the polarity of the extracted words are combined to obtain overall aspect polarity.

The paper is organized as follows. Section 2 describes our approach. Further on, in Section 3, we discuss the results obtained in the SemEval 2014 Task No. 4 subtask 2. Finally, section 4 provides concluding remarks.

2 UO_UA System

One of major challenge in sentiment analysis into product reviews, is dealing with a quite domain dependence. For instance, the word *"unpredictable"* can be considered as positive in Movie domain, however it is very negative in Airplane domain. For this reason, we propose to create a specific sentiment lexicon for addressing aspect based sentiment analysis in reviews.

Our proposal is divided in two main phases. The first one aims to build a domain-dependent sentiment resource for Laptop domain applying LSA. The second phase obtains the words related by means of some dependency relation with the aspect, and later, the polarity of these words are obtained from induced polarity lexicon and combined for computing overall aspect polarity.

2.1 Domain-Dependent Polarity Lexicon

The use of sentiment resource has been proven to be useful to build, train, and evaluate systems for sentiment analysis (Gutiérrez et al., 2013; Balahur, 2011). In order to build sentiment resource, several approach has been presented. In one of the first works, presented by (Hatzivassiloglou and McKeown, 1997), was proposed to take into ac-

count if adjectives are linked by adversative or copulative conjunctions for detecting its polarity. In (Turney and Littman, 2003) the authors exposed a method for inferring the semantic orientation of a word from its statistical association with a set of positive and negative paradigm words, measured by point-wise mutual information (PMI). In (2004), Hu and Liu suggested a technique to expand the lexicon using the relations of synonymy and antonym provided by WordNet (Fellbaum, 1998). In (2009), Cruz et al., created a sentiment resource based on a graph, constructed from conjunctive expressions between pairs of adjectives, observed in a review corpus. PageRank algorithm (Page et al., 1999) was adapted to be used on graphs with positive and negative edges, in order to obtain the semantic orientation of words.

Despite the wide range of existing proposals for resources construction, the results achieved with them are far from expected. As we have already seen, in aspect based sentiment analysis, the polarity of a word is heavily dependent on the domain; and general propose sentiment resource such as General Inquirer (Stone et al., 1966), WordNet-Affect(Strapparava and Valitutti, 2004), SentiWordNet(Baccianella et al., 2010) or HowNet (Dong et al., 2010) do not capture this dependency. On the other hand, the human annotators can not create specific sentiment resources for each new product launched to market. Therefore, propose methods to create these resources is a challenging task.

In this paper we address this task, presenting a framework for building domain-dependent sentiment resource. Our proposal is compounded of four phases. (See figure 1).

Firstly, review pages about the product of interest can be retrieved from different websites, for instance, Ciao, Epinions and Cnet (in this work we only use reviews from Ciao). This reviews are parsed and cleaned (this time we use Python XML Parser[6]). For each page we extract: pros, cons, title, full review and rating. In this work we have only focus on the pros and cons attributes because they are usually very brief, consist of short phrases or sentence segments and give a positive and negative evaluation about the product aspects. Each pros and cons in remainder paper will be considered as positive and negative samples, respectively.
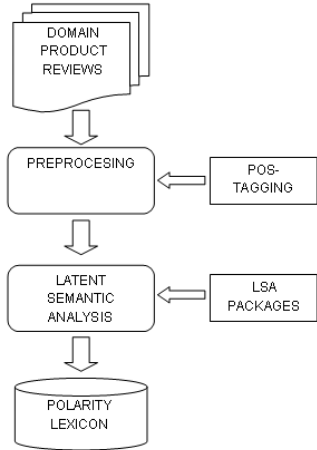
---

[6]https://docs.python.org/2/library/xml.html

Figure 1: Building domain-dependent sentiment resource.

Subsequently, the samples are preprocessed, applying a POS-Tagging tool (Padró and Stanilovsky, 2012) to convert all words in lemmas. After that, the stopwords are removed from text. Afterward each sample is represented using the classic vector space model (Salton et al., 1975). Intending to measure the association between term and class we add a special term to the vectors. In positive samples the term t_{pos} is added whereas in the negative samples the term t_{neg} is aggregated.

Later, we apply a Latent Semantic Analysis (this time we use, Gensim python package) to calculate the strength of the semantic association between words and classes. LSA uses the Singular Value Decomposition (SVD) to analyze the statistical relationships among words in a corpus.

The first step is construct a matrix $M_{n \times m}$, in which the row vectors v_i represent lemmas and the column vectors s_i the positive and negative sample (pros and cons). In each cell t_{ij}, we have the TF score (Term Frequency) of the ith lemma in j^{th} sample. The next step is apply Singular Value Decomposition to matrix $M_{n \times m}$ to decompose it into a product of three matrices $U\Sigma V^T$, then, we select the k largest singular values, and their corresponding singular vectors from U and V, obtained an approximation $\hat{M} = U_k \Sigma_k V_k^T$ of rank k to original matrix $M_{n \times m}$. After LSA is performed, we use the new matrix $\hat{M}$ to measure the association between lemmas l_i and l_j computing the cosine measure between vectors v_i and v_j, with the equation 1.

$$LSA_{score}(l_i, l_j) = \frac{< v_i, v_j >}{\| v_i \| \cdot \| v_j \|} \qquad (1)$$

Finally, the polarity lexicon contains lemmas l_i and its positive and negative scores. This values are computed using $LSA_{score}(l_i, t_{pos})$ and $LSA_{score}(l_i, t_{neg})$ respectively. The table 1 show some top positive and negative words computed with this strategy.

Positive	Score	Negative	Score
sturdy	0.8249	*prone*	0.8322
superb	0.7293	*weak*	0.8189
durable	0.7074	*disaster*	0.8120
sexy	0.6893	*erm*	0.8118
powerfull	0.6700	*ill*	0.8107
robust	0.6686	*uncomfortable*	0.8084
affordable	0.6630	*noisy*	0.7917
suuupeerrr	0.6550	*overwhelm*	0.7514
lighweight	0.6550	*unsturdy*	0.7491
unbreakable	0.6542	*lousy*	0.7143

Table 1: Examples of positive and negative words.

With aim to do our contribution to *SemEval-2014, Task 4: Aspect Based Sentiment Analysis (Pontiki et al., 2014), Subtask 2: Aspect Term Polarity*, we gathered 3010 Laptop Reviews, from Ciao and create a corpus with 6020 samples, 3010 positives (Pros) and 3010 negatives (Cons). This corpus was used as input in the developed framework (See figure 1). In this time we utilize Freeling[7] as POS-Tagging tool and Gensim Python Packages[8] to perform LSA (only the most 100 most significant eigenvalue are used). After that, a domain-dependent sentiment resource (DLSR) with 4482 term was created for Laptop reviews.

2.2 Aspect Polarity Classification

In order to exploit our domain-dependent sentiment resource building for Laptop domain, we develop an unsupervised method based on language rule to classify the product aspect. The basic rules are used to find dependency relation between aspect and their attributes. The figure 2 show the architecture of our proposal.

The proposed method receive as input a tuple $(P_{feature}, R)$, where $P_{feature}$ represent the aspect to evaluate, and R is the context (review) in it appears.

---

[7]http://nlp.lsi.upc.edu/freeling/
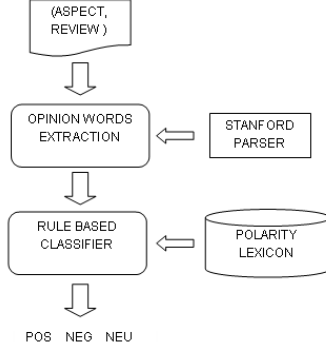[8]https://pypi.python.org/pypi/gensim

Figure 2: Apect polarity classification.

The dependency parsed is applied to review R, using Stanford Parser. Following that, we extract a set of tuples W, each tuple is represented as a pair (Att, Mod) where Att is a word related with the aspect $P_{feature}$ through some dependency relations shown in Table 2, and Mod is a integer value indicating if Att is modified by a valence shifter (Polanyi and Zaenen, 2004), (we only consider negation words, e.g., *never, no, not, don't, nothing,* etc.) , and default value of 0 is assign, in case that, the Att is modified by a valence shifter, we assign value of -1.

Dependency relations		
mod	subj	nsubj
amod	csub	csubpass
advmod	obj	dobj
vmod	iobj	pobj
rcmod	npadvmod	nn
subj	xcomp	advcl

Table 2: Stanford Parser dependency relations.

Once, the set of pairs W was obtained, the polarity of the feature $P_{feature}$ is determined from the scores of the attributes (related words) that describe it. To sum up, for each pair $(Att, Mod) \in W$, the positive $Pos((Att, Mod))$ and negative $Neg((Att, Mod))$ scores are calculated as:

$$Neg((Att, Mod)) = \begin{cases} -N(Att) \, if \, Mod < 0 \\ N(Att) \, otherwise \end{cases}$$
(2)

$$Pos((Att, Mod)) = \begin{cases} -P(Att) \, if \, Mod < 0 \\ P(Att) \, otherwise \end{cases}$$
(3)

Where $P(Att)$ and $N(Att)$ are the positive and negative score for Att in domain-dependent sentiment resource DLSR.

Finally, the global positive and negative scores (SO_{pos}, SO_{neg}) are calculated as:

$$SO_{pos}(P_{feature}) = \sum_{w \in W} Pos(w)$$
(4)

$$SO_{neg}(P_{feature}) = \sum_{w \in W} Neg(w)$$
(5)

If SO_{pos} is greater than SO_{neg} then the aspect is considered as positive. On the contrary, if SO_{pos} is less than SO_{neg} the aspect is negative. Finally, if SO_{pos} is equal to SO_{neg} the aspect is considered as neutral.

3 Results

In this section we present the evaluation of our system in the context of *SemEval-2014, Task 4: Aspect Based Sentiment Analysis (Pontiki et al., 2014), Subtask 2: Aspect Term Polarity*. For evaluating the participant's system two unlabeled domain-specific datasets for laptops and restaurants were distributed. For each dataset two runs can be submitted, the first (constrained), the system can only be used the provided training data and other resources such as lexicons. In the second (unconstrained), the system can use additional data for training. We send one run for laptop dataset and it only use external data retrieved from Ciao website (the training data was not used) (unconstrained).

The results achieve by our method are illustrate in Table 3. As may be observed, the accuracy

Label	Pr	Rc	F1
conflict	0.0	0.0	0.0
negative	**0,5234**	0,3764	0,4379
neutral	0,4556	0,4074	0,4302
positive	**0,6364**	**0,7561**	**0,6911**
Accuracy		0.55198777	

Table 3: Results in aspect polarity classification for laptop dataset.

achieve by UA_OU was 0.55, and F1 measure for negative, neutral and positive were 0,4379, 0,4302 and 0,6911 respectively. In case of conflict polarity we reached a 0.0 F1 value because our system not handle this situation. For this subtask (Laptop domain) a total of 32 runs was submitted by all

systems participant's and our run was ranked as 25th. The results despite not achieving expected, are encouraging. These evidence the feasibility of building resources from data available on the web, for aspect-based sentiment analysis.

4 Conclusions

In this article, we presented and evaluated the approach considered for our participation in *SemEval-2014 Task 4: Aspect Based Sentiment Analysis (Pontiki et al., 2014), Subtask 2: Aspect Term Polarity*, specifically for Laptop Domain. We present a framework for building domain-dependent sentiment resources applying Latent Semantic Analysis and build a special resource for polarity classification in Laptop domain. This resource was combined into unsupervised method to compute the polarity associated to different aspect in reviews. The results obtained by our approach are encouraging if we consider that the construction of the polarity lexicon is performed fully automatically.

Acknowledgements

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, LREC '10, Valletta, Malta, May.

Kushal Bafna and Durga Toshniwal. 2013. Feature based summarization of customers' reviews of online products. *Procedia Computer Science*, 22(0):142 – 151. 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - KES2013.

Alexandra Balahur and Andrés Montoyo. 2008. Multilingual feature-driven opinion extraction and summarization from customer reviews. In Epaminondas Kapetanios, Vijayan Sugumaran, and Myra Spiliopoulou, editors, *Natural Language and Information Systems*, volume 5039 of *Lecture Notes in Computer Science*, pages 345–346. Springer Berlin Heidelberg.

Alexandra Balahur. 2011. *Methods and Resources for Sentiment Analysis in Multilingual Documents of Different Text Types*. Ph.D. thesis, Department of Software and Computing Systems. Alcalant, Alcalant University.

Fermín Cruz, José Antonio Troyano, Francisco Javier Ortega, and Carlos García Vallejo. 2009. Inducción de un lexicón de opinión orientado al dominio. *Procesamiento del Lenguaje Natural*, 43:5–12.

Zhendong Dong, Qiang Dong, and Changling Hao. 2010. HowNet and its computation of meaning. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, COLING'10, pages 53–56, Stroudsburg, PA, USA.

Ruihai Dong, Markus Schaal, Michael P. O'Mahony, Kevin McCarthy, and Barry Smyth. 2013. Opinionated product recommendation. In Sarah Jane Delany and Santiago Ontañón, editors, *Case-Based Reasoning Research and Development*, volume 7969 of *Lecture Notes in Computer Science*, pages 44–58. Springer Berlin Heidelberg.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Yoan Gutiérrez, Andy González, Roger Pérez, José Abreu, Antonio Fernández Orquín, Alejandro Mosquera, Andrés Montoyo, Rafael Muñoz, and Franc Camara. 2013. UMCC_DLSI-(SA): Using a ranking algorithm and informal features to solve sentiment analysis in Twitter. *Atlanta, Georgia, USA*, page 443.

Vasileios Hatzivassiloglou and Kathleen McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the Joint ACL/EACL Conference*, pages 174–181.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *Proceedings of AAAI*, pages 755–760.

Bing Liu. 2012. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank Citation Ranking: Bringing Order to the Web.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis.

Livia Polanyi and Annie Zaenen. 2004. Contextual lexical valence shifters. In Yan Qu, James Shanahan, and Janyce Wiebe, editors, *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*. AAAI Press. AAAI technical report SS-04-07.

Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect Based Sentiment Analysis. In *International Workshop on Semantic Evaluation (SemEval), 2014*.

Gerard Salton, Andrew Wong, and Chung Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Philip James Stone, Dexter Colboyd Dunphy, Marshall Smith, and Daniel Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: an affective extension of WordNet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, LREC, pages 1083–1086, Lisbon, Portugal, May 26-28. European Language Resources Association.

Jianshu Sun, Chong Long, Xiaoyan Zhu, and Minlie Huang. 2009. Mining reviews for product comparison and recommendation. *Polibits*, pages 33 – 40, 06.

Peter Turney and Michael Lederman Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.