ECNU: Extracting Effective Features from Multiple Sequential Sentences for Target-dependent Sentiment Analysis in Reviews

Zhihua Zhang, Man Lan*
Shanghai Key Laboratory of Multidimensional Information Processing
Department of Computer Science and Technology,
East China Normal University, Shanghai 200241, P. R. China
51131201039@ecnu.cn, mlan@cs.ecnu.edu.cn*

Abstract

This paper describes our systems submitted to the target-dependent sentiment polarity classification subtask in aspect based sentiment analysis (ABSA) task (i.e., Task 12) in SemEval 2015. To settle this problem, we extracted several effective features from three sequential sentences, including sentiment lexicon, linguistic and domain specific features. Then we employed these features to construct classifiers using supervised classification algorithm. In laptop domain, our systems ranked $2nd$ out of 6 constrained submissions and $2nd$ out of 7 unconstrained submissions. In restaurant domain, the rankings are $5th$ out of 6 and $2nd$ out of 8 respectively.

1 Introduction

Reviews express opinions of customers towards various aspects of a product or service. Mining customer reviews (i.e., opinion mining) has emerged as an interesting new research direction in recent years. Since sentiment expressed in reviews usually adheres to specific categories or target terms, it is much meaningful to identify the sentiment target and its orientation, which helps users gain precise sentiment insights on specific sentiment target.

Unlike most existing sentiment analysis methods which try to detect the polarity of a sentence or a review, the aspect based sentiment analysis task (ASBA) shared as task 12 in SemEval 2015 is aiming at addressing the category- or target- dependent sentiment analysis in reviews. There are two types of subtasks organized in ASBA. The first aspect detection subtask is to identify the sentiment adherent from reviews, i.e., the category (i.e., entity-attribute (E-A) pair) or opinion target expression (*OTE*) in reviews. In most cases, the customers may not explicitly indicate the entity and attribute words in reviews but the opinion target expression is a segment of review. For example, in a given review: "*The pizza is overpriced and soggy.*", *target="pizza", category="FOOD-QUALITY"*. Its category label *FOOD-QUALITY* does not exist in reviews, while its *OTE* word "*pizza*" is explicitly present in reviews. The second sentiment polarity classification subtask is to assign a polarity label (i.e., positive, negative or neutral) for every E-A pair or *OTE* identified from the given reviews. We participated the second type subtask, i.e., performing sentiment polarity classification on reviews. There are two domains in this sentiment analysis subtask, i.e, laptop and restaurant. In laptop domain, only E-A pairs are annotated and provided in reviews while in restaurant domain, both E-A pairs and *OTE* are provided. Comparing with laptop reviews, the restaurant reviews provide the annotated surface words adhering to sentiment. Therefore we speculate that the performance in restaurant domain would be much better than that of laptop domain.

The study of aspect based sentiment analysis focuses on discovering the opinions or sentiments expressed by a customer on different categories or aspects (Liu, 2012). In recent years, it has drawn a lot of attentions. For example, (Branavan et al., 2009; He et al., 2012; Mei et al., 2007) used topic or category information. (Lin and He, 2009; Jo and Oh, 2011) presented LDA-based models, which incorporate aspect and sentiment analysis together to model

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 736–741,
Denver, Colorado, June 4-5, 2015.

sentiments towards different aspects. (Hu and Liu, 2004; Ding et al., 2008) adopted lexicon-based approaches to detect the sentiment on different aspects. In addition, (Boiy and Moens, 2009; Jiang et al., 2011) explored the work to determine whether the reviews contain the aspect information. Unlike the above study, (Xiang et al., 2014) split the data into multiple subsets based on category distributions and then built seperate classifier for each category.

Following previous work (Brun et al., 2014; Brychcín et al., 2014; Castellucci et al., 2014; Kiritchenko et al., 2014), a rich set of features are adopted in this work: linguistic features (e.g., *n-grams*, grammatical relationship, *POS*, negations), sentiment lexicon features (e.g., MPQA, General Inquirer, SentiWordNet, etc) and domain specific features (e.g., in-domain word list, punctuation, etc). We also performed a series of experiments to compare supervised machine learning algorithms with different parameters and to choose effective feature subsets for performance of classification.

The rest of this paper is structured as follows. In Section 2, we describe our system in details, including preprocessing, feature engineering, evaluation metrics, etc. Section 3 reports data sets, experiments and result discussion. Finally, Section 4 concludes our work.

2 System Description

2.1 Motivation

Unlike tweets with word length limitation, a review usually consists of several sentences and one single sentence may contain mixed opinions towards different targets. However, based on our observation and statistics on the data provided by SemEval 2015 Task 12, we find that most reviews (about 70%) have consistent opinion in their sentences, even though these sentences contain different category descriptions. Furthermore, although the E-A pair annotation is provided for each sentence, it is usually inferred by human being based on common knowledge from review rather than a single sentence. That is, the E-A pair information is supposed to be induced from contextual sentences rather than a single sentence alone. On the other hand, since one sentence may contain more than one category (i.e. E-A pair), this sentence alone may not provide enough information for every E-A pair. In consideration of above described reasons, we use multiple sentences rather than one single sentence to extract features for sentiment analysis. In this work, we used three sequential sentences, that is, for one given sentence, we combined its preceding and subsequent sentence with this current sentence together to perform sentiment analysis.

As we mentioned, one sentence may contain more than one E-A pair. As a result, for each E-A pair, not all words in this sentence or review are quite relevant and we need to select out only relevant words from three sequential sentences in terms of the corresponding E-A pair. Unlike the *OTE* words which already exist in reviews, most E-A pairs are not present in the sentence. Thus, for each E-A pair, we first extracted target words having top *tfidf* scores from three sequential sentences and then chose the relevant words from parse tree. Specifically, in laptop domain, the sentences contain only E-A pairs, so we selected two words having the highest *tfidf* scores from three sequential sentences in terms of corresponding E-A pair as its target words. Inspired by (Kiritchenko et al., 2014), for each target word in E-A pair, we selected the words from parse tree with distance $d =< 2$ as relevant words in terms of this E-A pair. After that, for all words in target words, we combined all their relevant words as pending words to extract features for sentiment analysis. While in restaurant domain since the sentences contain both E-A pair and opinion target expressions (*OTE*), we only combined the words in *OTE* with two words mentioned before as target words and chose their relevant words as pending words.

For each domain, each participant can submit two runs: (1) *constrained:* only the provided data can be used; (2) *unconstrained:* any additional resources can be used. In this task, we adopted 7 sentiment lexicons as external resources. Thus, the only difference of our two systems lies in the sentiment lexicon score features. For both systems, we extracted many traditional types of features to build classifiers for classification.

2.2 Data Preprocessing

Four preprocessing operations were performed. We first removed the XML tags from data and then transformed the abbreviations to their normal form-

s, i.e., "*don't*" to "*do not*". We used *Stanford Parser tools*[1] for tokenization, POS tagging and parsing. Finally, the WordNet-based Lemmatizer implemented in NLTK[2] was adopted to lemmatize words to their base forms with the aid of their POS tags.

2.3 Feature Engineering

In this work, we used three types of features: sentiment lexicon features, linguistic features and domain-specific features. All features were extracted from pending words as described above.

Sentiment Lexicon Features: Given pending words, we first converted them into lowercase and then calculated five feature values for each sentiment lexicon: (1) the ratio of positive words to pending words, (2) the ratio of negative words to pending words, (3) the maximum sentiment score, (4) the minimum sentiment score[3], (5) the sum of sentiment scores. If the pending word does not exist in one sentiment lexicon, its corresponding score is set to zero. The following 8 sentiment lexicons are used in our systems. Specifically, the first lexicon is employed to build constrained system and others 7 lexicons for unconstrained system.

- *Constrained PMI*: To build constrained system, we generated two domain-specific sentiment lexicons from the given training data respectively (i.e., laptop and restaurant). Given a term *w*, this PMI-based score is calculated from labeled reviews as below:

$$score(w) = PMI(w, pos) - PMI(w, neg)$$

  where *PMI* stands for pointwise mutual information.

- *Bing Liu opinion lexicon*[4]: This sentiment lexicon contains two annotated words lists: positive (about $2,000$) and negative(about $4,800$).

- *General Inquirer lexicon*[5]: The General Inquirer lexicon tries to classify English words along several dimensions, including sentiment polarity and we selected about $1,500$ positive words and $2,000$ negative words.

- *IMDB*[6]: This lexicon is generated from a large data set from IMDB which contains $25,000$ positive and $25,000$ negative movie reviews and the PMI-based sentiment score of each word is calculated as above.

- *MPQA*[7]: MPQA contains about $8,000$ subjective words with 6 types of label: strong/weak positive, strong/weak negative, both (having positive and negative sentiment) and neutral. Then we transformed these above nominal labels to $1, 0.5, -1, -0.5, 0, 0$ respectively.

- *SentiWordNet*[8]: The sentiment scores of each item in SentiWordNet is represented as a tuple i.e., positivity and negativity.We use the difference between positive and negative score as its sentiment score. When locating the corresponding item, we retrieved the word lemma and selected the first term in searched results according to its POS tag.

- *NRC Hashtag Sentiment Lexicon*[9]: (Mohammad et al., 2013) collected two tweet sets containing hashtags and used the sentiment of its hashtags as the sentiment label for each tweet. In this experiment, we used both unigrams and bigrams sentiment lexicons.

- *NRC Sentiment140 Lexicon*[10]: This lexicon is generated from a collection of 1.6 million tweets with positive or negative emoticons and contains about $62,000$ unigrams, $677,000$ bigrams and $480,000$ non-contiguous pairs. We used unigrams and bigrams.

Linguistic Features

- *Word n-grams:* We converted all pending words into lowercase and removed low frequency terms (≤ 5). After that, we extracted word-level unigram and bigrams.

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

[2] http://nltk.org

[3] We convert the sentiment scores in all sentiment lexicons to the range of $[-1, 1]$, where "-" denotes negative sentiment.

[4] http://www.cs.uic.edu/liub/FBS/sentiment-analysis.html#lexicon

[5] http://www.wjh.harvard.edu/inquirer/homecat.htm

[6] http://anthology.aclweb.org//S/S13/S13-2.pdf#page=444

[7] http://mpqa.cs.pitt.edu/

[8] http://sentiwordnet.isti.cnr.it/

[9] http://www.umiacs.umd.edu/saif/WebDocs/NRC-Hashtag-Sentiment-Lexicon-v0.1.zip

[10] http://help.sentiment140.com/for-students/

- *POS Features:* (Pak and Paroubek, 2010) found that subjective texts often contain more adjectives or adverbs and less nouns than objective texts. Therefore, the POS tags are important features for sentiment analysis. We recorded the number of nouns (the corresponding POS tags are *NN, NNP, NNS and NNPS*), verbs (*VB, VBD, VBG, VBN, VBP and VBZ*), adjectives (*JJ, JJR and JJS*) and adverbs (*RB, RBR and RBS*) in pending words.

- *Grammatical Relationship:* The grammatical relationship usually expresses the role of words in phrase and contains certain semantic information (Zhao et al., 2014). We obtained dependency information from parse tree and the grammatical information is denoted as a tuple, e.g., *amod*(*surprises*, *great*), where *amod* represents the dependency relationship between *surprises* and *great* (here *great* is a modifier). We presented two types of features: the relationship with the first word in tuple as *Rel1* and with the second word as *Rel2*. The size of each feature set is approximately 150.

- *Negation Features:* We collected 29 negations from Internet and designed this binary feature to record if there is negation in pending words.

Domain Specifical Features

- *In-domain word list*: For different domains, the words indicative of viewpoints are quite different. For example, *useful, fast, excellent* represent positive opinion in laptop domain and *delicious, cheap, beautiful* stand for positive opinion in restaurant domain. Therefore, we manually built two in-domain word lists from training instances indicative of positive and negative for both domains respectively. This feature records the number of in-domain words in pending words.

- *Punctuation*: Exclamation (!) and question (?) signs often indicate emotions (i.e., surprise, shock, interrogative, etc.) of users. Thus this feature counts the number of exclamations and questions in pending words.

- *All-caps*: This feature is the number of uppercase words in pending words.

2.4 Evaluation Measures

To evaluate the performance of different systems, the official evaluation measure *accuracy* is adopted.

3 Experiment

3.1 Datasets

The organizers provided two XML format documents regarding *laptop* and *restaurant* domain. In laptop, the {*E-A*, *P*} (i.e., {*EntityAttribute*, *Polarity*}) annotations are assigned at the sentence level taking the context of the whole review into account. In restaurant, it is a quadruple, i.e., {*E-A*, *OTE*, *P*}, where *OTE* stands for opinion target expression. In laptop, 22 entities (e.g., *LAPTOP, DISPLAY, CPU*, etc.) and 9 attributes (e.g., *PORTABILITY, PRICE, CONNECTIVITY,* etc.) are tagged while the restaurant data contains 6 entities (e.g., *SERVICE, RESTAURANT, FOOD,* etc.) and 5 attributes (i.e., *PRICES, QUALITY, STYLE_OPTIONS,* etc.). Table 1 shows the statistics of the data sets used in our experiments. Specifically, in restaurant, the opinions are adhered to *OTEs* and if the target does not exist explicitly, the *OTE* is tagged as *NULL*.

Dataset	Reviews	Sentences	Positive	Negative	Neutral	All
Laptop:						
train	277	1,739	1,103	765	106	1,974
test	173	725	541	329	79	949
Restaurant:						
train	254	1,315	1,198	403	53	1,654
test	96	663	454	346	45	845

Table 1: Statistics of training and test dataset in laptop and restaurant domains. *Positive, Negative, Neural* and *All* stand for the number of corresponding instances.

3.2 Experiments on Training data

To address this task, we adopted similar methods for both laptop and restaurant domains, i.e, employing rich features to build classifiers and adopting *Constrained PMI* features as sentiment lexicon feature for constrained systems while other sentiment lexicons for unconstrained systems. The 5-fold cross validation was performed for system development.

Table 2 shows the results of feature selection experiments for unconstrained and constrained systems in restaurant and laptop domains.

From Table 2, it is interesting to find: (1) SentiLexi features are the most effective feature type-

Restaurant				Laptop			
Constrained		Unconstrained		Constrained		Unconstrained	
Feature	**Accuracy**	**Feature**	**Accuracy**	**Feature**	**Accuracy**	**Feature**	**Accuracy**
ConPMI	79.80	SentiLexi	82.82	ConPMI	80.09	SentiLexi	81.21
+bigram	80.77(+0.97)	+Domain	83.49(+0.67)	+Domain	80.49(+0.40)	+bigram	82.02(+0.81)
+Negation	81.07(+0.30)	+Negation	84.28(+0.77)	+Negation	81.30(+0.81)	+rel2	83.54(+1.52)
+rel2	81.25(+0.18)	+rel1	**84.52**(+0.24)	+rel2	**81.71**(+0.41)	+rel1	83.94(+0.40)
+Domain	**81.43**(+0.18)	+rel2	84.34(-0.18)	+POS	81.25(-0.46)	+Negation	**84.19**(+0.25)
+rel1	81.07(-0.36)	+bigram	84.03(-0.31)	+unigram	81.00(-0.25)	+unigram	84.04(-0.15)
+POS	80.89(-0.18)	+POS	83.67(-0.36)	+rel1	79.53(-0.47)	+Domain	83.99(-0.05)
+unigram	78.71(-2.18)	+unigram	81.63(-2.04)	+bigram	79.38(-0.15)	+POS	82.77(-0.78)

Table 2: Results of feature selection experiments for restaurant and laptop domains on training datasets. The numbers in the brackets are the performance increments compared with the previous results. *ConPMI* stands for *Constrained PMI* features while *SentiLexi* is other external sentiment lexicons features.

s to detect the polarity regardless of constrained or unconstrained. (2) *POS* features are not quite effective in all systems. The possible reasons may be that *POS* aims at identifying the subjective instances from objective ones and it has no discriminating power for the type of sentiment polarity. (3) The *unigram* features are not as effective as expected because most words are already present in *rel1* or *rel2* feature. (4) The performances in laptop and restaurant domain are comparable, which is inconsistent with our previous speculation (i.e., the result of restaurant domain performs better than that of laptop domain since both A-E pair and *OTE* are provided in restaurant). We do a deep analysis and find that the top two words with *tfidf* score usually include the *OTE* words in restaurant domain. This also confirmed that this target words selection method is effective for laptop domain.

Besides, in our preliminary experiments for both domains, we examined the SVM classifiers with various parameters implemented in scikit-learn tools[11]. Finally we employed the configurations listed in Table 3 for test data.

Domain	Constrained	Unconstrained
Restaurant	SVM,kernel=linear,c=0.1	SVM,kernel=linear,c=0.5
Laptop	SVM,kernel=linear,c=0.1	SVM,kernle=linear,c=1

Table 3: System configurations for the constrained and unconstrained runs in two domains.

3.3 Results and Discussion

Using the optimum feature set shown in Table 2 and configurations described in Table 3, we trained separate models for each domain and evaluated them against the SemEval-2015 Task 12 test set.

Table 4 presents the results of our systems and top-ranked systems on test data provided by organizer for laptop and restaurant domain. In laptop domain, our systems ranked 2nd out of 6 constrained submissions and 2nd out of 7 unconstrained submissions while in restaurant domain, the rankings are 5th/6 and 2nd/8 respectively.

The results in Table 4 shows that in both domains our unconstrained systems performed comparable to the best results. It indicated that using the external sentiment lexicons as additional resources makes great contribution although the majority of these external sentiment lexicons are out of domain, e.g., NRC lexicons are generated from tweets and IMDB is about movie reviews. On the other hand, the constrained system which calculated the PMI score for each word from training data only, would involve a lot of noise due to (1) no sufficient training instances and (2) without consideration of the relationship between word sentiment and its opinion adherent.

TeamID	Restaurant		Laptop	
	Con	Uncon	Con	Uncon
ECNU	69.82(5)	78.11(2)	74.50(2)	78.29(2)
lsislif	75.50(1)	-	77.87(1)	-
sentiue	-	78.70(1)	-	79.35(1)

Table 4: Performance of our systems and the top-ranked system for laptop and restaurant domains in terms of *Accuracy(%)* on test datasets. *Con* stands for *constrained* and *Uncon* represents *unconstrained*. The numbers in the brackets are the rankings on corresponding submissions.

[11] http://scikit-learn.org/stable/

4 Conclusion

In this paper, we examined several feature types, i.e., surface text, syntax feature, sentiment lexicon feature, etc, to detect sentiment polarity towards category or opinion target expression in reviews. Moreover, we extracted features from three sequential sentences in consideration of the characteristic of review. Our systems perform better than majority of submissions (e.g., rank 2nd out of 7 and 2nd out of 8 on unconstrained submissions in laptop and restaurant domains respectively). For the future work, we would like to construct domain-specific sentiment lexicons and present more effective in-domain features to settle this problem.

5 Acknowledgments

This research is supported by grants from Science and Technology Commission of Shanghai Municipality under research grant no. (14DZ2260800 and 15ZR1410700) and Shanghai Collaborative Innovation Center of Trustworthy Software for Internet of Things (ZF1213).

References

Erik Boiy and Marie-Francine Moens. 2009. A machine learning approach to sentiment analysis in multilingual web texts. *Information retrieval*, 12(5):526–558.

SRK Branavan, Harr Chen, Jacob Eisenstein, and Regina Barzilay. 2009. Learning document-level semantic properties from free-text annotations. *Journal of Artificial Intelligence Research*, 34(2):569.

Caroline Brun, Diana Nicoleta Popa, and Claude Roux. 2014. XRCE: Hybrid classification for aspect-based sentiment analysis. *SemEval 2014*, page 838.

Tomáš Brychcín, Michal Konkol, and Josef Steinberger. 2014. UWB: Machine learning approach to aspect-based sentiment analysis. *SemEval 2014*, page 817.

Giuseppe Castellucci, Simone Filice, Danilo Croce, and Roberto Basili. 2014. UNITOR: Aspect based sentiment analysis with structured learning. *SemEval 2014*, page 761.

Xiaowen Ding, Bing Liu, and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 231–240.

Yulan He, Chenghua Lin, Wei Gao, and Kam-Fai Wong. 2012. Tracking sentiment and topic dynamics from social media. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent Twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies-Volume 1*, pages 151–160.

Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the 4th ACM international conference on Web search and data mining*, pages 815–824.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif M Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. *SemEval 2014*, page 437.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384.

Bing Liu. 2012. Sentiment analysis and opinion mining: synthesis lectures on human language technologies. *Morgan & Claypool Publishers*.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on WWW*, pages 171–180.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327, Atlanta, Georgia, USA, June.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, pages 1320–1326.

Bing Xiang, Liang Zhou, and Thomson Reuters. 2014. Improving Twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In *Proceedings of the 52nd Annual Meeting of the ACL (Short Papers)*, pages 434–439.

Jiang Zhao, Tian Tian Zhu, and Man Lan. 2014. ECNU: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment. *SemEval 2014*, page 271.