SAP-RI: A Constrained and Supervised Approach for Aspect-Based Sentiment Analysis

Nishtha Malhotra[1,2*], **Akriti Vij**[1,2,*], **Naveen Nandan**[1] and **Daniel Dahlmeier**[1]
[1]Research & Innovation, SAP Asia, Singapore
[2]Nanyang Technological University, Singapore
{nishtha.malhotra,akriti.vij,naveen.nandan,d.dahlmeier}@sap.com

Abstract

We describe the submission of the SAP Research & Innovation team to the SemEval 2014 Task 4: Aspect-Based Sentiment Analysis (ABSA). Our system follows a constrained and supervised approach for aspect term extraction, categorization and sentiment classification of online reviews and the details are included in this paper.

1 Introduction

The increasing popularity of the internet as a source of information, and e-commerce as a way of life, has led to a major surge in the number of reviews that can be found online, for a wide range of products and services. Consequently, more and more consumers have taken to consulting these online reviews as part of their pre-purchase research before deciding on availing services from a local business or investing in a product from a particular brand. This calls for innovative techniques for the sentiment analysis of online reviews so as to generate accurate and relevant recommendations.

Sentiment analysis has been extensively studied and applied in different domains. Predicting the sentiment polarity (*positive, negative, neutral*) of user opinions by mining user reviews (Hu and Liu, 2004; Liu, 2012; Pang and Lee, 2008; Liu, 2010) has been of high commercial and research interest. In these studies, sentiment analysis is often conducted at one of the three levels: *document level*, *sentence level* or *attribute level*.

Through the SemEval 2014 Task 4 on *Aspect Based Sentiment Analysis* (Pontiki et al., 2014), we explore sentiment analysis at the aspect level. The task consists of four subtasks: in subtask 1 *aspect term extraction*, participants need to identify the aspect terms present in a sentence and return a list containing all distinct aspect terms, in subtask 2 *aspect term polarity*, participants were to determine the polarity of each aspect term in a sentence, in subtask 3 *aspect category detection*, participants had to identify the aspect categories discussed in a given sentence, and in subtask 4 *aspect category polarity*, participants were to determine the polarity of each aspect category. The polarity classification subtasks consider sentiment analysis to be a three-way classification problem between positive, negative and neutral sentiment. On the other hand, the aspect category detection subtask is a multi-label classification problem where one sentence can be labelled with more than one aspect category.

In this paper, we describe the submission of the SAP-RI team to the SemEval 2014 Task 4. We make use of supervised techniques to extract the aspects of interest (Jakob and Gurevych, 2010), categorize them (Lu et al., 2011) and predict the sentiment of customer online reviews on *Laptops* and *Restaurants*. We developed a constrained system for aspect-based sentiment analysis of these online reviews. The system is constrained in the sense that we only use the training data that was provided by the challenge organizers and no other external data sources. Our system performed reasonably well, especially with a F_1 score of 75.61% for the aspect category polarity subtask, 79.04% F_1 score on the aspect category detection task and 66.61% F_1 score on the aspect term extraction task.

2 Subtask 1: Aspect Term Extraction

Given a review with annotated entities in the training set, the task was to extract the aspect terms for reviews in the test set. For this subtask, training, development and testing were conducted for both

[*] The work was done during an internship at SAP.

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 517–521,
Dublin, Ireland, August 23-24, 2014.

the laptop and the restaurant domain.

2.1 Features

Each review was represented as a feature vector made up of the following features:

- **Word N-grams:** all unigrams, bigrams and trigrams from the review text
- **Casing:** presence or absence of capital case/ title case words
- **POS tags:** POS tags of a word and its neighbours
- **Parse dependencies and relations:** parse dependency relations of the aspects, i.e., presence/absence of adjectives and adverbs in the dependency parse tree
- **Punctuation Marks:** presence/absence of punctuation marks, such as *?*, *!*

2.2 Method

We approach the task by casting it as a sequence tagging task where each token in a candidate sentence is labelled as either *Beginning*, *Inside* or *Outside* (BIO). We then employ conditional random fields (CRF), which is a discriminative, probabilistic model for sequence data with state-of-the-art performance (Lafferty et al., 2001). A linear-chain CRF tries to estimate the conditional probability of a label sequence $\mathbf{y}$ given the observed features $\mathbf{x}$, where each label y_t is conditioned on the previous label y_{t-1}. In our case, we use BIO CoNLL-style tags (Sang and De Meulder, 2003).

During development, we split the training data in the ratio of 60:20:20 as training, development (dev) and testing (dev-test). We train the CRF model on the training set of the data, perform feature selection based on the dev set, and test the resulting model on the dev-test. In all experiments, we use the CRF++[1] implementation of conditional random fields with the parameter c=4.0. This value was chosen based on manual observation. We perform a feature ablation study and the results are reported in Table 1. Features listed in section 2.1 were those that were retained for the final run.

3 Subtask 2: Aspect Term Polarity Estimation

For this subtask, the training, development and testing was done using reviews on laptops and restaurants. Given the aspect terms in a sentence, the task was to predict their sentiment polarities.

3.1 Features

For each review, we used the following features:

- **Word N-grams:** all lowercased unigrams, bigrams and trigrams from the review text
- **Polarity of neighbouring adjectives:** extracted word sentiment from SentiWordNet lexicon (Baccianella et al., 2010)
- **Neighbouring POS tags:** the POS tags of up to neighbouring 3 words
- **Parse dependencies and relations:** parse dependency relations of the aspects, i.e., presence/absence of adjectives and adverbs in the dependency parse tree

3.2 Method

For each aspect term of a sentence, the aforementioned features were extracted. For example, for the term *Sushi* in the sentence *Sushi was delicious.*, the following feature vector is constructed, {*aspect: 'sushi', advmod:'null', amod:'delicious', uni_sushi: 1, uni_was: 1, uni_delicious, uni_the: 0, ..* }.

We then treat the aspect sentiment polarity estimation as a multi-class classification task where each instance would be labelled as either *positive*, *negative* or *neutral*. For the classification task, we experimented with Naive Bayes and Support Vector Machines (SVM) – both linear and RBF kernels – and it was observed that linear SVM performed best. Hence, we use linear SVM for the classification task. Table 2 summarizes the results obtained from our experiments for various feature combinations. The classifiers used are implementations from *scikit-learn*[2], which is also used for the remaining tasks.

4 Subtask3: Aspect Category Detection

Given a review with annotated entities or aspect terms, the task was to predict the aspect categories.

[1] `code.google.com/p/crfpp/`

[2] `scikit-learn.org/stable/`

Features	Precision	Recall	F1-Score
N-grams, POS tags	0.7655	0.4283	0.5496
N-grams, Parse relations, POS tags	0.8192	0.6641	0.7336
N-Grams, Parse relations, POS tags, casing	0.8101	0.6641	0.7299
N-grams, Parse relations, POS tags, !	0.8116	0.6641	0.7305
N-grams, Parse relations, POS tags,!, ?	0.8123	0.6672	0.7326

Table 1: Training-phase experimental results for Subtask 1 on Restaurant reviews.

Features	Laptops	Restaurants
Neighbouring words, 2,3 POS grams, bigrams, trigrams, Sentiment,1,2 ngram lower	0.4196	0.5997
Parse Relations, 2,3 POS grams, bigrams, trigrams, Sentiment, 1,2 ngram lower	0.5869	0.6375
Parse Relations, Neighbouring words, bigram, trigrams, Sentiment, 1,2 ngram lower	0.5848	0.6380
Parse Relations, 2,3 POS grams, Neighbouring words, Sentiment, 1,2 ngram lower	0.5890	0.6240
Parse Relations, 2,3 POS grams , Neighbouring words, bigram, trigrams, 1,2 ngram lower	0.5626	0.6239
Parse Relations, 2,3 POS grams , Neighbouring words, bigram, trigrams, Sentiment	0.5922	0.6409

Table 2: Training-phase experimental results (Accuracy) for Subtask 2.

As one sentence in a review could belong to multiple aspect categories, we model the task as a multi-label classification problem, i.e., given an instance, predict all labels that the instance fits to.

4.1 Features

We experimented with different features, for example unigrams, dependency tree relations, bigrams, POS tags and sentiment of the words (SentiWordNet), but using just the unigrams alone happened to yield the best result. The feature vector was merely a bag-of-words vector indicating the presence or absence of a word in an instance.

4.2 Method

The training instances were divided into 5 sets based on the aspect categories and thereby, we treated the multi-label classification task as 5 different binary classification tasks. Hence, we used an ensemble of binary classifiers for the multi-label classification. An SVM model was trained using one classifier per class to distinguish it from all other classes. For the binary classification tasks, directly estimating a linear separating function (such as linear SVM) gave better results, as shown in Table 3. Finally, the results of the 5 binary classifiers were combined to label the test instance.

The category *Miscellaneous* was observed to have the lowest accuracy, probably due to the fact that miscellaneous captures all those aspects terms that do not have a clearly defined category.

5 Subtask4 Aspect Category Polarity Detection

For each review with pre-labelled aspect categories, the task was to produce a model which predicts the sentiment polarity of each aspect category.

5.1 Features

The training data contains reviews with the polarity for the corresponding aspect category. The models performed best on using just unigram and bigram features.

5.2 Method

The training instances were split into 5 sets based on the aspect categories. We make use of the sentiment polarity classifier, as described in section 3.2, thereby, training one sentiment polarity classifier for each aspect category. Table 4 indicates the performance of different classifiers for this task, using features as discussed in section 5.1.

6 Results

Table 5 gives an overview of the performance of our system in this year's task based on the official scores from the organizers. We see that our system performs relatively well for subtasks 1, 3 and 4, while for subtask 2 the F_1 scores are behind the best system by about 12%. As observed, a sentence could have more than one aspect and each of these aspects could have different polarities expressed. Including features that preserve the context of the aspect could probably improve the performance in the subtask 2. In most cases, a simple set of features was enough to result in a

Restaurants Category	Naive Bayes	AdaBoost	LinearSVC
Food	0.7130	0.8000	0.8470
Service	0.6064	0.9137	0.8997
Miscellaneous	0.6710	0.7490	0.7890
Ambience	0.6770	0.9063	0.8940
Price	0.7608	0.8548	0.9590

Table 3: Training-phase experimental results (F_1 score) for Subtask 3.

Restaurants Category	Naive Bayes	AdaBoost	LinearSVC
Food	0.7136	0.6711	0.7417
Service	0.6733	0.5244	0.6688
Miscellaneous	0.4756	0.3170	0.4756
Ambience	0.6574	0.7232	0.6885
Price	0.7477	0.7752	0.6651

Table 4: Training-phase experimental results (F_1 score) for Subtask 4.

high F_1 score, for example, in subtask 3 a bag-of-words feature set proved to yield a relatively high F_1 score. In general, for the classification tasks, we observe that the linear SVM performs best.

Subtask	Dataset	Best score	Our score	Rank
1	Laptops	74.55	66.61	8/27
1	Restaurants	84.01	77.88	12/29
2	Laptops	70.48	58.56	18/32
2	Restaurants	80.95	69.92	22/36
3	Restaurants	88.57	79.04	7/21
4	Restaurants	82.92	75.61	5/25

Table 5: Results (F_1 score and ranking) for the Semeval-2014 test set.

7 Conclusion

In this paper, we have described the submission of the SAP-RI team to the SemEval 2014 Task 4. We model the classification tasks using linear SVM and the term extraction task using CRF in order to develop an aspect-based sentiment analysis system that performs reasonably well.

Acknowledgement

The research is partially funded by the Economic Development Board and the National Research Foundation of Singapore.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. 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)*, volume 10, pages 2200–2204.

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.

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1035–1045.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Bing Liu. 2010. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing*, pages 627–666. Chapman & Hall, 2 edition.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Bin Lu, Myle Ott, Claire Cardie, and Benjamin Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *Proceedings of Sentiment Elicitation from Natural Text for Information Retrieval and Extraction*, pages 81–88.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*.

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In

Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, pages 142–147.