IITP:Supervised Machine Learning for Aspect based Sentiment Analysis

Deepak Kumar Gupta
Indian Institute of Technology Patna
Patna, India
deepak.mtmc13@iitp.ac.in

Asif Ekbal
Indian Institute of Technology Patna
Patna, India
asif@iitp.ac.in

Abstract

The shared task on *Aspect based Sentiment Analysis* primarily focuses on mining relevant information from the thousands of online reviews available for a popular product or service. In this paper we report our works on aspect term extraction and sentiment classification with respect to our participation in the SemEval-2014 shared task. The aspect term extraction method is based on supervised learning algorithm, where we use different classifiers, and finally combine their outputs using a majority voting technique. For sentiment classification we use Random Forest classifier. Our system for aspect term extraction shows the F-scores of 72.13% and 62.84% for the restaurants and laptops reviews, respectively. Due to some technical problems our submission on sentiment classification was not evaluated. However we evaluate the submitted system with the same evaluation metrics, and it shows the accuracies of 67.37% and 67.07% for the restaurants and laptops reviews, respectively.

1 Introduction

Nowadays user review is one of the means to drive the sales of products or services. There is a growing trend among the customers who look at the online reviews of products or services before taking a final decision. In sentiment analysis and opinion mining, aspect extraction aims to extract entity aspects or features on which opinions have been expressed (Hu and Liu, 2004; Liu, 2012). An aspect is an attribute or component of the product that has been commented on in a review. For example:*"Dell Laptop has very good battery life and click pads"*. Here aspect terms are *battery life* and *click pads*. Sentiment analysis is the task of identifying the polarity (*positive, negative* or *neutral*) of review. Aspect terms can influence sentiment polarity within a single domain. As an example, for the restaurant domain *cheap* is usually positive with respect to *food*, but it denotes a negative polarity when discussing the decor or ambiance (Brody and Elhadad, 2010).

A key task of aspect based sentiment analysis is to extract aspects of entities and determine the sentiment corresponding to aspect terms that have been commented in review document. In recent times there has been huge interest to identify aspects and sentiments simultaneously. The method proposed in (Hu and Liu, 2004) is based on information extraction (IE) approach that identifies frequently occurring noun phrases using association mining. Some other works include the methods, *viz* those that define aspect terms using a manually specified subset of the Wikipedia category (Fahrni and Klenner, 2008) hierarchy, unsupervised clustering technique (Popescu and Etzionir, 2005) and semantically motivated technique (Turney, 2002) etc. Our proposed approach for aspect term extraction is based on supervised machine learning, where we build many models based on different classifiers, and finally combine their outputs using majority voting. Before combining, the output of each classifier is post-processed with a set of heuristics. Each of these classifiers is trained with a moderate set of features, which are generated without using any domain-specific knowledge and/or resources. Our submitted system for the second task is based on Random Forest (Breiman, 2001).

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 319–323,
Dublin, Ireland, August 23-24, 2014.

2 Tasks

The SemEval-2014 shared task on Aspect based Sentiment Analysis [1] focuses on identifying the aspects of a given target entities and the sentiment expressed towards each aspect. A benchmark setup was provided with the datasets consisting of customer reviews with human-annotated annotations of the aspects and their polarity information. There were four subtasks, and we participated in the first two of them. These are defined as follows:

Subtask-1: The first task is related to aspect term extraction. Given a set of sentences with pre-identified entities, identify the aspect terms present in the sentence and return a list containing all the distinct aspect terms.

Substask-2: The second task addresses the aspect term polarity. For a given set of aspect terms within a sentence, determine whether the polarity of each aspect term is positive, negative, neutral or conflict (i.e. both positive and negative).

3 Methods

3.1 Pre-processing

Each review is in the XML form. At first we extract the reviews along with their identifiers. Each review is tokenized using the Stanford parser [2] and Part-of-Speech tagged using the Stanford PoS tagger [3]. At the various levels we need the chunk-level information. We extract these information using the OpenNLP chunker available at [4].

3.2 Aspect Term Extraction

The approach we adopted for aspect term extraction is based on the supervised machine learning algorithm. An aspect can be expressed by a noun, adjective, verb or adverb. But the recent research in (Liu, 2007) shows that 60-70% of the aspect terms are explicit nouns. The aspect terms could also consist of multiword entities such as *"battery life"* and *"spicy tuna rolls"* etc. As the classification algorithms we make use of Sequential minimal optimization (SMO), Multiclass classifier, Random forest and Random tree. For faster computation of Support Vector Machine, SMO (Platt, 1998) was proposed. Random tree (Breiman, 2001) is basically a decision tree, and in general used as a weak learner to be included in some ensemble learning method. Multiclass classifier is a meta learner based on binary SMO. This has been converted to multiclass classifier using the pairwise method. In order to reduce the errors caused by the incorrect boundary identification we define a set of heuristics, and apply on each output. At the end these models are combined together using a simple majority voting.

We implement the following set of features for aspect terms extraction.

- **Local context:** Local contexts that span the preceding and following few tokens of the current word are used as the features. Here we use the previous two and next two tokens as the features.

- **Part-of-Speech information:** Part-of-Speech(PoS)information plays an important role in identifying the aspect terms. We use the PoS information of the current token as the feature.

- **Chunk Information:** Chunk information helps in identifying the boundaries of aspect terms. This is particularly more helpful to recognize multiword aspect terms.

- **Root word:** Roots of the surface forms are used as the features. We use the Porter Stemmer algorithm [5] to extract the root forms.

- **Stop word:** We use the list of stop words available at [6]. A feature is defined that takes the value equal to 1 or 0 depending upon whether it appears in the training/test set or not.

- **Length:** Length of token plays an important role in identifying the aspect terms. We assume an entity as the candidate aspect term if its length exceeds a predefined threshold value equal to five.

- **Prefix and Suffix**: Prefix and suffix of fixed length character sequences are stripped from each token and used as the features of classifier. Here we use the prefixes and suffixes of length upto three characters as the features.

[1] http://alt.qcri.org/semeval2014/task4/

[2] http://nlp.stanford.edu/software/lexparser.shtml

[3] http://nlp.stanford.edu/software/tagger.shtml

[4] http://opennlp.sourceforge.net/models-1.5/

[5] http://tartarus.org/martin/PorterStemmer/java.txt

[6] http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words

- **Frequent aspect term**: We extract the the aspect terms from the training data, and prepare a list by considering the most frequently occurring terms. We consider an aspect term to be frequent if it appears at least five times in the training data. A feature is then defined that fires if and only if the current token appears in this list.

The output of each classifiers is post-processed with a set of hand-crafted rules, defined as below:
Rule 1: If the PoS tag of the target token is noun, chunk tag is *I-NP* (denoting the intermediate token of a noun phrase) and the observed class of the previous token is *O* (other than aspect terms) then the current token should be assigned the class *B-Aspect* (denotes the beginning of an aspect term).
Rule 2: If the current token has PoS tag noun, chunk tag *I-NP* and the observed class of the immediately preceding token is *B-Aspect* then the current token should be assigned the class *I-Aspect* (denoting the intermediate token).

3.3 Polarity Identification

Polarity classification of aspect terms is the classical problem in sentiment analysis. The task is to classify the sentiments or opinions into semantic classes such as *positive*, *negative*, and *neutral*. We develop a Random Forest classifier for this task. In this particular task one more class *conflict* is introduced. It is assigned if the sentiment can either be positive or negative. For classification we make use of some of the features such as local context, PoS, Chunk, prefix and suffix etc., as defined in the previous Subsection. Some other problem-specific features that we implement for sentiment classification are defined as below:

- **MPQA feature**: We make use of MPQA subjectivity lexicon (Wiebe and Mihalcea, 2006) that contains sentiment bearing words as feature in our classifier. This list was prepared semi-automatically from the corpora of MPQA[7] and Movie Review dataset[8]. A feature is defined that takes the values as follows: 1 for positive; -1 for negative; 0 for neutral and 2 for those words that do not appear in the list.

- **Function words**: A list of function words is compiled from the web[9]. A binary-valued feature is defined that fires for those words that appear in this list.

4 Experiments and Analysis

We use the datasets and the evaluation scripts as provided by the SemEval-2014 shared task organizer.

4.1 Datasets

The datasets comprise of the domains of restaurants and laptop reviews. The training sets consist of 3,044 and 3,045 reviews. There are 3,699 and 2,358 aspect terms, respectively. The test set contains 800 reviews for each domain. There are 1,134 and 654 test instances in the respective domains.

4.2 Results and Analysis

At first we develop several machine learning models based on the different classification algorithms. All these classifiers were trained using the same set of features as mentioned in Section 3. We use the default implementations of these classifiers in Weka[10]. We post-process the outputs of all the models using some heuristics. Finally, all these classifiers are combined together using majority voting. It is to be noted that we determine the best configuration by carrying out different experiments on the development set, which is constructed by taking a part of the training set, and finally blind evaluation is performed on the respective test set. We use the evaluation script provided with the SemEval-2014 shared task. The training sets contain multiword aspect terms, and so we use the standard BIO notation[11] for proper boundary marking.

Experiments show the precision, recall and F-score values 77.97%, 72.13% and 74.94%, respectively for the restaurant dataset. This is approximately 10 points below compared to the best system. But it shows the increments of 4.16 and 27.79 points over the average and baseline models, respectively. For the laptop dataset we obtain the precision, recall and F-score values of 70.74%, 62.84% and 66.55%, respectively. This is 8 points below the best one and 10.35 points

[7] http://cs.pitt.edu/mpqa/
[8] http://cs.cornell.edu/People/pabo/movie-review-data/
[9] http://www2.fs.u-bunkyo.ac.jp/ gilner/wordlists.html
[10] www.cs.waikato.ac.nz/ml/weka/
[11] B, I and O denote the beginning, intermediate and outside tokens

Model	precision	recall	F-score
Random Tree	65.21	59.63	62.29
Random Forest	70.93	62.69	66.55
SMO	71.18	64.22	67.52
Multiclass	73.44	68.50	70.88
Ensemble	77.97	72.13	74.94
Best system	85.35	82.71	84.01
Average	76.74	67.26	70.78
Baseline	-	-	47.15

Table 1: Result of Task-A for restaurants dataset with different classifiers (in %).

Model	precision	recall	F-score
Random Tree	56.52	56.17	56.34
Random Forest	58.38	58.02	58.19
SMO	63.62	63.22	63.39
Multiclass	65.30	64.90	65.09
Ensemble	70.74	62.84	66.55
Best system	84.80	66.51	74.55
Average	68.97	50.45	56.20
Baseline	-	-	35.64

Table 2: Results of aspect term extraction for laptops dataset with different classifiers (in %).

above the average system. Compared to the baseline it achieves more than 20 point increment. Detailed evaluation results for all the classifiers are reported in Table 1 and Table 2 for restaurant and laptop datasets, respectively. Results show that multiclass classifier achieves the highest performance with precision, recall and F-score values of 73.44%, 68.50% and 70.88%, respectively for the restaurant dataset. The same model shows the highest performance with precision, recall and F-score values of 65.30%, 64.90% and 65.09%, respectively for the laptop dataset. Because of majority ensemble we observe increments of 4.06% and 1.46% F-score points over the best individual model, respectively.

We also perform error analysis to understand the possible sources of errors. We show only the confusion matrix for Task-A in Table 3. It shows that in most cases I-ASP is misclassified as B-ASP. System also suffers because of the misclassification of aspect terms to others.

Experiments for classification are reported in Table 4. Evaluation shows that the system achieves the accuracies of 67.37% and 67.07% for

	B-ASP	I-ASP	Other
B-ASP	853	15	269
I-ASP	114	213	142
Other	123	35	11431

Table 3: Confusion matrix for Task-A on restaurants dataset.

Datasets	#Aspect Terms	#Correct Identification	Accuracy (in %)
Restaurants	1134	764	67.37
Laptops	654	438	67.07

Table 4: Results of aspect terms polarity (in %).

the restaurants and laptops datasets, respectively. Please note that our system for the second task was not officially evaluated because of the technical problems of the submitted zipped folder. However we evaluated the same system with the official evaluation script, and it shows the accuracies as reported in Table 4. We observe that the classifier performs reasonably well for the positive and negative classes, and suffers most for the conflict classes. This may be due to the number of instances present in the respective training set. Results show that our system achieves much lower classification accuracy (13.58 points below) compared to the best system for the restaurant datasets. However, for the laptop datasets the classification accuracy is quite encouraging (just 3.42 points below the best system). It is also to be noted that our classifier achieves quite comparable performance for both the datasets. Therefore it is more general and not biased to any particular domain.

5 Conclusion

In this paper we report our works on aspect term extraction and sentiment classification as part of our participation in the SemEval-2014 shared task. For aspect term extraction we develop an ensemble system. Our aspect term classification model is based on Random Forest classifier. Runs for both of our systems were constrained in nature, i.e. we did not make use of any external resources. Evaluation on the shared task dataset shows encouraging results that need further investigation.

Our analysis suggests that there are many ways to improve the performance of the system. In future we will identify more features to improve the performance of each of the tasks.

References

Leo Breiman. 2001. Random forests. 45(1):5–32.

S. Brody and N. Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Proceedings of NAACL*, pages 804–812, Los Angeles, CA.

Angela Fahrni and Manfred Klenner. 2008. Old wine or warm beer: Target-specic sentiment analysis of adjectives. In *Symsposium on Affective Language in Human and Machine*, pages 60–63. The Society for the Study of Artificial Intelligence and Simulation of Behavior (AISB).

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th KDD*, pages 168–177, Seattle, WAs.

B. Liu. 2007. *Exploring Hyperlinks, Contents, and Usage Data*. Springer.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

John C. Platt. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING.

Ana-Maria Popescu and Oren Etzionir. 2005. Extracting product features and opinions from reviews. In *Proceedings of the Conference on HLT/EMNLP*, pages 339–346.

P. D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th ACL*, pages 417–424.

Janyce Wiebe and Rada Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of the COLING/ACL*, pages 065–1072, Australia.