NLANGP: Supervised Machine Learning System for Aspect Category Classification and Opinion Target Extraction

Zhiqiang Toh
Institute for Infocomm Research
1 Fusionopolis Way
Singapore 138632
ztoh@i2r.a-star.edu.sg

Jian Su
Institute for Infocomm Research
1 Fusionopolis Way
Singapore 138632
sujian@i2r.a-star.edu.sg

Abstract

This paper describes our system used in the Aspect Based Sentiment Analysis Task 12 of SemEval-2015. Our system is based on two supervised machine learning algorithms: sigmoidal feedforward network to train binary classifiers for aspect category classification (Slot 1), and Conditional Random Fields to train classifiers for opinion target extraction (Slot 2). We extract a variety of lexicon and syntactic features, as well as cluster features induced from unlabeled data. Our system achieves state-of-the-art performances, ranking 1st for three of the evaluations (Slot 1 for both restaurant and laptop domains, and Slot 1 & 2) and 2nd for Slot 2 evaluation.

1 Introduction

The amount of user-generated content on the web has grown rapidly in recent years, prompting increasing interests in the research area of sentiment analysis and opinion mining. Most previous work is concerned with detecting the overall polarity of a sentence or paragraph, regardless of the target entities (e.g. restaurants) and their aspects (e.g. food). By contrast, the Aspect Based Sentiment Analysis task of SemEval 2014 (SE-ABSA14) is concerned with identifying the aspects of given target entities and the sentiment expressed towards each aspect (Pontiki et al., 2014).

The SemEval-2015 Aspect Based Sentiment Analysis (SE-ABSA15) task is a continuation of SE-ABSA14 (Pontiki et al., 2015). The SE-ABSA15 task features a number of changes that address issues raised in SE-ABSA14 and also encourage further in-depth research. For example, (1) instead of isolated (potentially out of context) sentences, the input datasets will contain entire reviews; (2) information linking aspect terms and aspect categories are now provided; (3) besides in-domain ABSA (Subtask 1), SE-ABSA15 will include an out-of-domain ABSA subtask (Subtask 2).

We participate in Subtask 1 of SE-ABSA15, namely aspect category classification (Slot 1) and opinion target extraction (Slot 2). We also participate in the evaluation which assesses whether a system identifies both the aspect categories and opinion targets correctly (Slot 1 & 2).

For Slot 1, we model the problem as a multi-class classification problem where binary classifiers are trained to predict the aspect categories. We follow the one-vs-all strategy and train a binary classifier for each category in the training set. Each classifier is trained using sigmoidal feedforward network with 1 hidden layer. For Slot 2, we follow the approach of Toh and Wang (2014) by modeling the problem as a sequential labeling task, using Conditional Random Fields (CRF) as the training algorithm. For Slot 1 & 2, we perform a simple combination of Slot 1 predictions and Slot 2 predictions.

The remainder of this paper is structured as follows. In Section 2, we describe our system in detail, including the feature description and approaches. In Section 3, the official results are presented. Feature ablation results are shown in Section 4. Finally, Section 5 summarizes our work.

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 496–501,
Denver, Colorado, June 4-5, 2015.

2 System Description

In this section, we present the details of our sentiment analysis system. The training set consists of 254 English review documents containing 1315 sentences for the restaurant domain and 277 English review documents containing 1739 sentences for the laptop domain.

As a first step of our system, we perform basic data preprocessing. All sentences are tokenized and parsed using the Stanford Parser[1].

2.1 Features

This section briefly describes the features used in our system, where some of the features are useful across different slots. The features used are a subset of the features described in Toh and Wang (2014), which also provides a more detailed description of the features.

2.1.1 Word

The current word is used as a feature. For opinion target extraction, the previous word and next word are also used as features.

2.1.2 Bigram

All word bigrams found in a sentence are used as features.

2.1.3 Name List

For the restaurant domain, we extract two high precision name lists from the training set and use them for membership testing. For the first list, we collect and keep only high frequent opinion targets. For the second list, we consider the counts of individual words in the opinion targets and keep those words that frequently occur as part of an opinion target in the training set.

2.1.4 Head Word

From the sentence parse tree, we extract the head word of each word and use it as a feature.

2.1.5 Word Cluster

We induce Brown clusters and K-means clusters from two different sources of unlabeled dataset: the Multi-Domain Sentiment Dataset that contains Amazon product reviews (Blitzer et al., 2007)[2], and the Yelp Phoenix Academic Dataset that contains user reviews[3]. We also experiment using a third dataset that is created by combining the initial two datasets into one.

For Brown clusters[4], we experiment with different datasets, cluster sizes ($\{100, 200, 500, 1000\}$), minimum occurrences ($\{1, 2, 3\}$) and binary prefix lengths. The best settings to use are determined using 5-fold cross validation.

K-means clusters are induced using the `word2vec` tool (Mikolov et al., 2013)[5]. Similarly, among different datasets, word vector sizes ($\{50, 100, 200, 500, 1000\}$), cluster sizes ($\{50, 100, 200, 500, 1000\}$), and sub-sampling thresholds ($\{0.00001, 0.001\}$), we use 5-fold cross validation to select the best settings.

2.1.6 Name List Generated using Double Propagation

For the restaurant domain, we generate a name list of possible opinion targets using the Double Propagation (DP) algorithm (Qiu et al., 2011). The propagation rules are modified from the logic rules presented in Liu et al. (2013), where we write our rules in Prolog and use SWI-Prolog[6] as the solver. As the rules can only identify single-word targets, to consider multi-word targets, we extend the left boundary of the identified target to include any consective noun words right before the target.

2.2 Approaches

We developed our system to return results for Slot 1 (restaurant and laptop domains), Slot 2 (restaurant domain) and Slot 1 & 2 (restaurant domain). This section describes our machine learning approaches used to generate the predictions for each slot.

2.2.1 Aspect Category Classification (Slot 1)

Aspect category classification is based on a set of one-vs-all binary classifiers, one classifier for each

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

[2] We used the unprocessed.tar.gz archive found at http://www.cs.jhu.edu/ mdredze/datasets/sentiment/

[3] http://www.yelp.com/dataset_challenge/

[4] Brown clusters are induced using the implementation by Percy Liang found at https://github.com/percyliang/brown-cluster/

[5] https://code.google.com/p/word2vec/

[6] http://www.swi-prolog.org/

Parameter	Restaurant	Laptop
learning rate	0.9	0.7
hidden units	4	4
threshold	0.2	0.2

Table 1: Tuned parameter values for Slot 1 on the restaurant and laptop domain.

category found in the training set. For each sentence in the training set, we extract features from all words in the sentence to create a training example. The label of the example depends on which category *C* we are training: 1 if the sentence contains *C* as one of its categories, -1 otherwise. The number of binary classifiers is 13 for the restaurant domain and 79 for the laptop domain, which equals to the number of categories annotated in the training set for the respective domain.

We use the Vowpal Wabbit tool[7] to train the binary classifiers. Each classifier is trained using sigmoidal feedforward network with 1 hidden layer (`--nn`), with `--ngram` enabled to generate word bigrams. The learning rate (`-l`) and number of hidden units are tuned using 5-fold cross validation.

We also tuned the probability threshold where we regard the classifier output as positive outcome. Any classifier that returns a probability score greater than the threshold will be added to the output set of categories. The tuned parameter values used are shown in Table 1.

Table 2 shows the features used for the restaurant and laptop domain, as well as the 5-fold cross-validation performances after adding each feature group.

2.2.2 Opinion Target Extraction (Slot 2)

Opinion target extraction is modeled as a sequential labeling task, where each word in the sentence is assigned a label using the IOB2 scheme (Sang and Veenstra, 1999). The classifier is trained using Conditional Random Fields (CRF), which has shown to achieve state-of-the-art performances in previous work (Toh and Wang, 2014; Chernyshevich, 2014). We use the CRFsuite tool (Okazaki, 2007) for CRF training and enable negative state and transition features (`-p feature.possible_states=1 -p feature.possible_transitions=1`).

Restaurant	
Feature	F1
Word	0.6245
+ Bigram	0.6423
+ Name List	0.6608
+ Head Word	0.6660
+ Word Cluster	0.7038
Laptop	
Feature	F1
Word	0.4520
+ Bigram	0.4611
+ Head Word	0.4721
+ Word Cluster	0.4841

Table 2: 5-fold cross-validation performances for Slot 1 on the restaurant and laptop domain. Each row uses all features added in the previous rows.

We experiment with two different methods of returning predicted opinion targets, one suitable for Slot 1 & 2 evaluation (Method-1), the other suitable for Slot 2 evaluation (Method-2).

For Slot 1 & 2 evaluation, the explicit opinion targets may have more than one categories. Thus, we use the following method (Method-1): we train a separate CRF model for each category found in the training set. That is, for each category *C*, we assign the label "B-*C*" to indicate the start of an opinion target, "I-*C*" to indicate the continuation of an opinion target, and "O" if the opinion target does not have *C* as one of its categories.

Using FOOD#PRICES category as an example, for the training set that is used to train the FOOD#PRICES CRF model, we assign the label "B-FOOD#PRICES" to indicate the start of a FOOD#PRICES opinion target, "I-FOOD#PRICES" to indicate the continuation of a FOOD#PRICES opinion target, and "O" if the opinion target does not have FOOD#PRICES as one of its categories.

However, our initial experiments suggest that Method-1 does not achieve optimum performance for Slot 2 evaluation. The reason is that the number of positive training examples for most of the categories is small.

[7] https://github.com/JohnLangford/vowpal_wabbit/wiki

	Slot 1									
	Restaurant					Laptop				
System	Type	Rank	P	R	F1	Type	Rank	P	R	F1
NLANGP (U)	U	1	0.6386	**0.6155**	**0.6268**	U	1	**0.6425**	0.4209	**0.5086**
NLANGP (C)	C	2	**0.6637**	0.5806	0.6194	C	4	0.5743	0.4283	0.4906
1st	U	1	0.6386	**0.6155**	**0.6268**	U	1	**0.6425**	0.4209	**0.5086**
2nd	C	2	**0.6637**	0.5806	0.6194	U	2	0.5773	0.4409	0.5000
3rd	C	3	0.5698	0.5742	0.5720	C	3	0.5548	**0.4483**	0.4959
Baseline	–	–	–	–	0.5133	–	–	–	–	0.4631
	Slot 2					Slot 1 & 2				
	Restaurant									
System	Type	Rank	P	R	F1	Type	Rank	P	R	F1
NLANGP (U)	U	2	0.7053	0.6402	0.6712	U	1	0.4463	**0.4130**	**0.4290**
NLANGP (C)	C	7	**0.7129**	0.5406	0.6149	C	4	0.4387	0.3645	0.3982
1st	U	1	0.6893	**0.7122**	**0.7005**	U	1	0.4463	**0.4130**	**0.4290**
2nd	U	2	0.7053	0.6402	0.6712	C	2	**0.5937**	0.3337	0.4273
3rd	C	3	0.6723	0.6661	0.6691	U	3	0.5832	0.3278	0.4197
Baseline	–	–	–	–	0.4807	–	–	–	–	0.3444

Table 4: Comparison of our unconstrained (U) and constrained (C) systems with the top three participating systems and official baselines for Slot 1, Slot 2 and Slot 1 & 2. P, R, and F1 denote the precision, recall and F1 measure respectively.

Restaurant	
Feature	F1
Word	0.6225
+ Name List	0.6796
+ Head Word	0.6840
+ Word Cluster	0.7224
+ DP Name List	0.7237

Table 3: 5-fold cross-validation performances of Slot 2 on the restaurant domain. Each row uses all features added in the previous rows. The cross-validation experiments use Method-1 to train the models.

Since Slot 2 evaluation only requires the identified text span to be returned and does not require any category information, we can increase the number of positive training examples by collapsing all categories into a single category (e.g. "TERM"). Thus, for Slot 2 evaluation, the following method (Method-2) is used: we train a single CRF model where all opinion targets in the training set are assigned the labels "B-TERM", "I-TERM" and "O" accordingly.

Table 3 shows the features used for the restaurant domain as well as the 5-fold cross-validation performances after adding each feature group.

Due to time constraints, all cross-validation experiments for Slot 2 use Method-1 to train the models. The same settings will then be used to train the final models using both Method-1 (for Slot 1 & 2 evaluation) and Method-2 (for Slot 2 evaluation).

2.2.3 Slot 1 & 2

To create the predictions for Slot 1 & 2 evaluation, we perform a simple combination of Slot 1 predictions and Slot 2 predictions. First, we use all Slot 2 predictions. Next, for each sentence, we add categories that are found in Slot 1 predictions but not Slot 2 predictions of the same sentence. Those additional categories are assumed to be NULL targets.

3 Results

We have submitted results for unconstrained and constrained (using only the provided training set of the corresponding domain) systems. The constrained system only uses Word, Bigram (for Slot 1) and Name List (for the restaurant domain) features. Table 4 presents the official results of our submissions. We also include the results of the top three

Restaurant		
System	Method-1	Method-2
NLANGP (U)	0.6099	0.6712
NLANGP (C)	0.5489	0.6149

Table 5: Comparison of F1 performances for Slot 2 evaluation. Our official submissions for Slot 2 evaluation use Method-2, which is better than Method-1 used for Slot 1 & 2 evaluation.

participating systems and official baselines for comparison (Pontiki et al., 2015).

As shown from the table, our system performed well for all four evaluations. Our system is ranked 1st for three of the evaluations (Slot 1 for both restaurant and laptop domains, and Slot 1 & 2) and 2nd for Slot 2 evaluation. In addition, our constrained system also achieves competitive results, ranking 2nd in Slot 1 Restaurant and 4th in Slot 1 Laptop and Slot 1 & 2. Another observation is that our unconstrained systems achieved better performances than the corresponding constrained systems for all evaluations, indicating the use of external resources are beneficial.

We are interested to know whether the Slot 2 predictions that help to achieve best results in Slot 1 & 2 evaluation are also useful for Slot 2 evaluation. Table 5 shows the F1 performances of Slot 2 evaluation if we have used Method-1 (Section 2.2.2) to generate the Slot 2 predictions. As shown from the table, using the same Slot 2 predictions for both Slot 2 evaluation and Slot 1 & 2 evaluation are detrimental to Slot 2 performances, with performance difference greater than 6.0%. Our approach of using a different method to generate Slot 2 predictions for Slot 2 evaluation helps to overcome the data sparseness problem and improves the performances of target extraction.

4 Feature Ablation

Table 6 and Table 7 show the (unconstrained) F1 measure and loss on the test set resulting from training with each group of feature removed for Slot 1 and Slot 2 respectively. The ablation experiments indicate that each feature is helpful in improving the performance, with performance gains in the range of $1.0\% - 6.0\%$. The only exception is the use of bigram feature in Slot 1 evaluation on the laptop domain, where a slight decrease of 0.27% is observed. Among the external resources used, the Word Cluster feature consistently provides the most gain: an increase in F1 measure greater than 3.0% for both slots on the restaurant domain.

Restaurant		
Feature	F1	Loss
Word	0.5914	0.0354
Bigram	0.6031	0.0237
Name List	0.6123	0.0145
Head Word	0.6136	0.0132
Word Cluster	0.5910	0.0358
Laptop		
Feature	F1	Loss
Word	0.4483	0.0603
Bigram	0.5114	-0.0027
Head Word	0.4978	0.0108
Word Cluster	0.4940	0.0146

Table 6: Test set ablation experiments for Slot 1 on the restaurant and laptop domain. The quantity is the (unconstrained) F1 measure and loss resulted from the removal of a single feature group.

Restaurant		
Feature	F1	Loss
Word	0.6280	0.0432
Name List	0.6540	0.0172
Head Word	0.6602	0.0110
Word Cluster	0.6387	0.0325
DP Name List	0.6608	0.0104

Table 7: Test set ablation experiments for Slot 2 on the restaurant domain. The quantity is the (unconstrained) F1 measure and loss resulted from the removal of a single feature group.

5 Conclusion

In this paper, we report our work on aspect category classification and opinion target extraction using supervised machine learning approaches. By leveraging on external resources, careful feature selection and performance tuning, our system achieves top performances in all four evaluations, ranking 1st for three of the evaluations, and second for the re-

maining evaluation. In future, we hope to improve our opinion target extraction system by taking into account surrounding sentence context and incorporating sentiment lexicon features to better classify aspect categories and detect opinion expressions.

Acknowledgments

This work is a study conducted at Baidu-I^{2}R Research Centre.

We thank the anonymous reviewers for their helpful comments and suggestions.

References

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic, June.

Maryna Chernyshevich. 2014. IHS R&D Belarus: Cross-domain Extraction of Product Features using Conditional Random Fields. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 309–313.

Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. 2013. A Logic Programming Approach to Aspect Extraction in Opinion Mining. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 01*, WI-IAT '13, pages 276–283, Washington, DC, USA, November.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris 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)*, pages 27–35.

Maria Pontiki, Dimitris Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion Word Expansion and Target Extraction through Double Propagation. *Computational Linguistics*, 37(1):9–27.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing Text Chunks. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, pages 173–179, Stroudsburg, PA, USA.

Zhiqiang Toh and Wenting Wang. 2014. DLIREC: Aspect Term Extraction and Term Polarity Classification System. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 235–240.