USF: Chunking for Aspect Term Identification & Polarity Classification

Cindi Thompson
University of San Francisco
2130 Fulton St, HR 240 San Francisco, CA 94117
cathompson4@usfca.edu

Abstract

This paper describes the systems submitted by the University of San Francisco (USF) to Semeval-2014 Task 4, Aspect Based Sentiment Analysis (ABSA), which provides labeled data in two domains, laptops and restaurants. For the constrained condition of both the aspect term extraction and aspect term polarity tasks, we take a supervised machine learning approach using a combination of lexical, syntactic, and baseline sentiment features. Our extraction approach is inspired by a chunking approach, based on its strong past results on related tasks. Our system performed slightly below average compared to other submissions, possibly because we use a simpler classification model than prior work. Our polarity labeling approach uses two baseline hand-built sentiment classifiers as features in addition to lexical and syntactic features, and performed in the top ten of other constrained systems on both domains.

1 Introduction

As stated in the call for participation for this Semeval task, sentiment analysis focusing on overall polarity of a document, sentence, or similar context has been well studied in recent years (Liu, 2010; Pang and Lee, 2008; Tsytsarau and Palpanas, 2012). However, there is less prior work examining finer levels of granularity associated with individual entities and their characteristics or attributes, which the organizers for this task call *aspects*. The aspect based sentiment analysis task (ABSA) has the goal of identifying aspects of stated or implied target entities and the sentiment expressed towards each aspect. This problem has not been deeply studied in prior literature due to the lack, until now, of a large gold standard dataset. This Semeval task has provided two such datasets, in the domains of laptops and restaurants. A full description of the task and data is presented with this volume (Pontiki et al., 2014).

In this paper, we discuss our approach to the first two subtasks of the Semeval ABSA Task, those of aspect term extraction and aspect term polarity. In aspect term extraction the domain (restaurants or laptops) is known and the goal is to identify terms in a sentence that are features commonly associated with that domain, such as service and staff in the case of restaurants or size and speed in the case of laptops. In the polarity subtask, the aspect terms for a given sentence are already identified and the sentiment polarity (positive, negative, conflict, or neutral) must be assigned.

We approach both subtasks using supervised machine learning with background knowledge of sentiment lexicons and syntax included in our feature set. Our goal was to investigate whether techniques that have been successful in similar tasks would perform well on this newly created data set. We did not use additional corpus-based resources, so qualified for the constrained (versus unconstrained) version of the task. The remainder of the paper details related work, our approach, and experiments and the results we obtained.

2 Related Work

We divide related work into two areas: research related to aspect and aspect term identification, and research related to sentiment classification for aspect terms. We note that aspects have also been called topics and features in prior work. Until more recently, the community lacked a corpus

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 790–795,
Dublin, Ireland, August 23-24, 2014.

of gold-standard labeled data that focuses on aspect terms, rather than more general expressions of subjectivity or other private states. Thus, early work focused on learning or identifying aspects in an unsupervised (Hu and Liu, 2004) or semi-supervised setting (Moghaddam and Ester, 2010; Zhai et al., 2011). The earliest work on aspect detection focused on identifying frequently occurring noun phrases using information extraction (IE) techniques (Hu and Liu, 2004). Unsupervised techniques include clustering (Fahrni and Klenner, 2008; Popescu and Etzioni, 2005) and topic models (Titov and McDonald, 2008).

The benchmark corpus for sentiment analysis from Wiebe et al. (2005) inspired much work on learning subjective phrases in a supervised setting. The nature of the data and annotation differ from the data for this Semeval task, as it focuses on news articles and identifying an entire opinion phrase, including the source of the opinion, and only recently added aspect annotations. However, the techniques used by others to learn to extract this data and the associated polarity inspired our own approach. These include extraction-like approaches, usually using sequence modeling (Breck et al., 2007; Jin et al., 2009; Johansson and Moschitti, 2013; Li et al., 2010; Mitchell et al., 2013; Yang and Cardie, 2013) and semantic dependency or semantic parsing approaches (Kim and Hovy, 2006; Kobayashi et al., 2007; Wu et al., 2009) sometimes using background knowledge from sentiment lexicons (Zhang et al., 2009). The main differences between our approach and that of Breck et al. (2007) and Mitchell et al. (2013) are the classifier used and some of the features; they both use CRFs versus our Maximum entropy classifier, and they used a wider range of syntactic and dictionary-based features.

A second related corpus which includes more aspect information is that developed by Kim and Hovy (2006). This corpus also focuses on news articles rather than reviews, but does expand the types of aspects identified. The main focus of that work is on the identification, using FrameNet role labels, of the holder and target of an opinion, while the opinion itself is provided to the system.

The restaurant reviews used in this Semeval task are a 3000-plus sentence subset of those harvested by Ganu et al. (2009), plus newly annotated sentences used for test data. The original corpus contains over 50,000 structured restaurant reviews including restaurant information and a star rating. The original star rating was not made available for the Semeval tasks, and the aspect term annotations and their associated sentiment were added for this task; the original sentence-level sentiment annotations were not provided. Most of the work exploring this corpus to date uses unsupervised (Brody and Elhadad, 2010) or semi-supervised (Mukherjee and Liu, 2012) approaches.

As there has been an explosion of research in sentiment classification, it is impossible to review all of the related work. See Tsytsarau and Palpanas (2012) for a recent survey. We will note that our approach follows a somewhat standard machine learning approach inspired by that of Wilson et al. (2005), but with a different feature set. We did not thoroughly explore as many classifiers as this work and others have done. Finally, we note that some work has investigated the joint task of identifying opinion phrases or targets simultaneously with polarity (Choi and Cardie, 2009; Johansson and Moschitti, 2013; Mitchell et al., 2013).

3 Approach

For both subtasks, we take a supervised machine learning approach, examining several classifiers and their variants, and converging on feature sets which performed best in small-scale cross-validation experiments. After the official competition ended, we continued to examine different variants and discuss alternative approaches and their accuracy in the experimental results section. For all tasks we use the Maximum Entropy classifier, "iib" variant from the Natural Language Toolkit (NLTK) in Python (Bird et al., 2009). We experimented with several other classifiers from NLTK and found that Maximum Entropy performed best on a hold out set of data. We had originally planned to use a Conditional Random Field (CRF) model (Lafferty et al., 2001) because of its strong performance on similar tasks, but met with time limitations when converting the data to the appropriate format (there is no CRF provided with NLTK at this time). We had also planned to try classifiers from the scikit-learn toolkit (Pedregosa et al., 2011), but again met with time constraints due to the necessity to manually convert the features to a binary representation.

We first preprocess the data using NLTK's tokenization and part-of-speech tagging modules and align the results with the aspect terms in the data,

as detailed further below. The sentiment lexicon we use as the basis of all sentiment features discussed below combines two standard lexicons (Liu et al., 2005; Wilson et al., 2005).

3.1 Aspect Term Extraction

While it is difficult to give a precise definition of aspect, it can be roughly thought of as a characteristic of a target concept such as a restaurant or laptop. Examples include the italicized terms in the following:

- I liked the *service* and the *staff*, but not the *food*.
- The *hard disk* is very noisy.

We use a sequence labeling approach, which can also be thought of as a tagging or chunking approach, to identify the aspect terms in each sentence. Specifically, and similar to Breck et al. (2007) and Mitchell et al. (2013), as the target class for each token, we use the IOB2 sequence labeling scheme (Tjong et al., 2000), where the aspect terms are considered as the chunks to be labeled. Using this approach, each token is tagged as either Beginning an aspect term, being In an aspect term, or being Outside an aspect term. We also experimented with an IO labeling scheme as discussed in the experimental results section, in which each token is tagged as being either In or Outside an aspect term. Here is an example of a sentence with its IOB tags:

- The-O pizza-B is-O the-O best-O if-O you-O like-O thin-B crusted-I pizza-I .-O

Of course, unlike an HMM or CRF, a standard classifier such as Maximum entropy does not label entire sequences. Therefore, each example presented to our classifier represents a single token from the sentence being labeled, and the target label is the IOB tag of that token. Further, we present the tokens of a given sentence in order from the first word in the sentence to the last.

The features used for each token are derived from the token, the prior token, and the next token in the sentence (thus using a three-token window). In addition, we include the IOB tag of the prior token, using the gold standard at training time and the classifier's output at testing time, even if it is incorrect. For each token we extract the word, its stem, its part-of-speech (POS) tag, its polarity from the sentiment dictionary, and whether the word is objective or subjective, from the same sentiment dictionary. We use dummy values for the prior and next words of the first and last token in a sentence, respectively. All feature-value pairs are converted to binary features automatically by NLTK.

Because we believed that the data would prove to be sparse and that new words would appear in the testing data, we also include an unknown word feature, replacing the 50% least frequent words in the training data with the "UNK" token, and doing the same for both these words and unseen words in the test set. However, we later found that we should have used cross-validation to support our hypothesis, and that using the full vocabulary would have improved our results, as shown in the experimental results section.

3.2 Polarity

In the polarity subtask, the aspect terms are provided, and the goal is to classify them as positive, negative, conflict, or neutral. In this case, we use a simple classification approach that includes features of the aspect term and surrounding tokens (again in a three-token window), and also some simple baseline sentiment classification features. First, we use similar features as for the aspect term extraction task, with changes to incorporate the fact that aspect terms are occasionally *phrases*, not single words. In fact, we hypothesize that features of the words before and after an aspect phrase could be more useful than the words prior to and after a particular *word* in the phrase.

Thus, instead of using features from the three-token window including the current token, we use features from the words on each side of the aspect phrase, and use the head of the aspect phrase and its features as the middle of the window. This approach is similar to that of Johansson and Moschitti (2013), who use features from the words before and after opinion expressions. In our case, these features are again the word, its POS tag, its sentiment polarity and objectivity, and its IOB tag. Note that in this case we use the IOB tag from all terms in the window, since the aspect term extraction task is treated as a prerequisite to the polarity classification task.

In addition to these word-based features, we add four higher-level features. The first is an indicator of the number of aspect terms in the entire sentence, since this might indicate a more de-

tailed sentence, and we believe that more specific sentences might correlate with positive sentiment. The other three features are baselines connected to the estimated sentiment of the sentence or phrase. First, we apply a hand-built sentence level sentiment classifier that follows a now standard baseline approach (Zhu et al., 2009): using a sentiment lexicon (Liu's), it counts the number of positive and negative sentiment words in the sentence, flipping polarity when negation words are encountered, and discontinuing the polarity flip when punctuation is encountered. This results in a "high level sentiment" feature consisting of the number of positive sentiment words minus the number of negative sentiment words.

Our other two sentiment features provide finer granularity information, based on the sentiment of the "chunks" in which an aspect term appears. First, we use the punctuation within the sentence to divide it into punctuation-separated chunks. Then, we calculate the number of positive and negative sentiment words within each chunk, again flipping polarity after the presence of a negation word. The positive and negative counts associated with the chunk within which an aspect phrase appears are then used as features when classifying the phrase. We also experimented with using conjunctions (and, or, but, etc.) as chunk boundaries, but preliminary results indicated that this resulted in reduced accuracy.

4 Experimental Results & Analysis

In this section we report our results and some additional analysis for the ABSA subtasks 1 and 2. Please refer to Pontiki et al. (2014) for details on the tasks, corpora, and evaluation criteria. We chose the constrained condition, which allows the use of sentiment lexicons in addition to the training data provided, but no additional data such as other reviews.

Aspect term extraction is evaluated using Precision, Recall, and F-measure on an unseen set of sentences. Table 1 shows our results[1] on both domains, the top results,[2] and the mean score of all constrained submissions (21 entries). Note that for Restaurants, COMMIT-P1WP3 had the best Precision, at 0.909, but XRCE had the best F-measure, so we show their three scores. Our results were close to the mean for both corpora and quite a bit above the lowest scoring submissions and the baseline provided by the organizers; the latter is also shown in the Table.

	System	P	R	F1	Rank
Lap	Belarus	0.848	0.665	0.746	1
	mean	*0.760*	*0.503*	*0.562*	*11*
	USF	0.754	0.404	0.526	14.7
	baseline	*0.443*	*0.298*	*0.356*	
Rest	XRCE	0.862	0.818	0.840	1
	mean	*0.770*	*0.649*	*0.693*	*11*
	USF	0.783	0.645	0.707	14.3
	baseline	*0.525*	*0.428*	*0.472*	

Table 1: Aspect Term Extraction Results, Constrained.

	Approach	P	R	F1
Lap	FV-No-Snt	0.724	**0.622**	**0.669**
	Full Voc.	**0.733**	0.601	0.660
	Original	0.715	0.493	0.583
	IO	0.696	0.501	0.582
Rest	Full Voc.	**0.792**	0.704	**0.746**
	FV-No-Snt	0.784	**0.710**	0.745
	Original	0.777	0.657	0.711
	IO	0.769	0.660	0.710

Table 2: Aspect Term Extraction Cross-Validation Results.

After the submission deadline, we continued to experiment with alternative approaches using 5-fold cross validation on the training set, shown in Table 2. We found that using the full vocabulary was better than our original approach of only using the top 50% occurring words, even with 28% unseen words in the restaurant test set and 21% in laptops. We also found that leaving out the polarity feature while using all vocabulary words (FV-No-Snt) improved our F-measure score to 0.669 for laptops but reduced it slightly to 0.745 for restaurants. Finally, using IO versus IOB tagging did not influence the F-measure significantly. About 25% of the aspect terms in the restaurant training set have length greater than one, and 37% of the laptop terms.

Aspect term polarity is evaluated on accuracy over all labels: positive, negative, neutral, or conflict. Table 3 shows our results on both domains, the top results, the mean score of all constrained submissions (24 entries for laptops, 28 for restaurants), and the baseline accuracy. In this case our scores are above average in all cases.

[1] Rank averaged over P, R, and F for USF

[2] We abbreviate IHS_RD_Belarus as Belarus.

	System	Acc	Rank
Lap	NRC-Canada	0.705	1
	USF	0.645	6
	mean	*0.604*	*12.5*
	baseline	*0.514*	
Rest	DCU	0.810	1
	USF	0.732	9
	mean	*0.702*	*14.5*
	baseline	*0.643*	

Table 3: Aspect Term Polarity Results, Constrained.

5 Conclusions

In conclusion, we show that a chunking approach to supervised learning works fairly well in the aspect term extraction task, and that local sentence features and a baseline sentiment classifier work well for aspect term polarity classification. Our systems for both tasks performed reasonably well considering the relatively simple classification techniques and features incorporated. In future work, we plan to apply more sophisticated classifiers which have shown to be accurate on related tasks, including CRFs and Support Vector Machines. We also would like to experiment with variants of the features used here, such as the exploration of smaller or larger context windows, or the usefulness of stemming compared to the original tokens. We also believe that more sophisticated syntactic or semantic features, or topic models, could improve results on one or both tasks.

We thank the organizers for the provision of this interesting dataset.

Acknowledgements

The author thanks her spring 2014 research assistant, Hao Chen, for his help in preparing some of the code used in the experiments.

References

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Eric Breck, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *Twentieth International Joint Conference on Artificial Intelligence*, pages 2683–2688.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812.

Yejin Choi and Claire Cardie. 2009. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 590–598.

Angela Fahrni and Manfred Klenner. 2008. Old wine or warm beer: Target-specific sentiment analysis of adjectives. In *Proc. of the Symposium on Affective Language in Human and Machine, AISB*, pages 60–63.

Gayatree Ganu, Noemie Elhadad, and Amelie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *12th International Workshop on the Web and Databases*, pages 1–6.

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. ACM.

Wei Jin, Hung Hay Ho, and Rohini K Srihari. 2009. A novel lexicalized HMM-based learning framework for web opinion mining. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 465–472.

Richard Johansson and Alessandro Moschitti. 2013. Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39(3):473–509.

Soo-min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Workshop on Sentiment and Subjectivity in Text*, pages 1–8.

Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. 2007. Extracting aspect-evaluation and aspect-of relations in opinion mining. In *EMNLP-CoNLL*, pages 1065–1074.

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

Fangtao Li, Chao Han, Minlie Huang, and Xiaoyan Zhu. 2010. Structure-aware review mining and summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 653–661.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International World Wide Web conference*, pages 342–351. ACM.

Bing Liu. 2010. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing 2*, pages 627–666. CRC Press.

Margaret Mitchell, Jacqueline Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *EMNLP*, pages 1643–1654.

Samaneh Moghaddam and Martin Ester. 2010. Opinion digger: An unsupervised opinion miner from unstructured product reviews. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1825–1828. ACM.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 339–348.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Maria Pontiki, Dimitrios 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)*.

Ana-maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 339–346.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceeding of the 17th international conference on World Wide Web*, pages 111–120, New York, New York, USA. ACM.

Erik F Tjong, Kim Sang, Walter Daelemans, Rob Koeling, Yuval Krymolowski, Vasin Punyakanok, Dan Roth, Millers Yard, Mill Lane, and Ramat Gan. 2000. Applying system combination to base noun phrase identification. In *Proceedings of the 18th conference on Computational linguistics*, pages 857–863.

Mikalai Tsytsarau and Themis Palpanas. 2012. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478–514.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354.

Yuanbin Wu, Qi Zhang, Xuanjing Huang, and Lide Wu. 2009. Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1541.

Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *Proceedings of ACL*, pages 16550–1649.

Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. 2011. Clustering product features for opinion mining. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 347–354, New York, New York, USA. ACM.

Qi Zhang, Yuanbin Wu, Tao Li, Mitsunori Ogihara, Joseph Johnson, and Xuanjing Huang. 2009. Mining product reviews based on shallow dependency parsing. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 726–727. ACM.

Jingbo Zhu, Muhua Zhu, Huizhen Wang, and Benjamin Tsou. 2009. Aspect-based sentence segmentation for sentiment summarization. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 65–72. ACM.