NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif M. Mohammad
National Research Council Canada
1200 Montreal Rd., Ottawa, ON, Canada
{Svetlana.Kiritchenko, Xiaodan.Zhu, Colin.Cherry, Saif.Mohammad}
@nrc-cnrc.gc.ca

Abstract

Reviews depict sentiments of customers towards various *aspects* of a product or service. Some of these aspects can be grouped into coarser *aspect categories*. SemEval-2014 had a shared task (Task 4) on aspect-level sentiment analysis, with over 30 teams participated. In this paper, we describe our submissions, which stood first in detecting aspect categories, first in detecting sentiment towards aspect categories, third in detecting aspect terms, and first and second in detecting sentiment towards aspect terms in the laptop and restaurant domains, respectively.

1 Introduction

Automatically identifying sentiment expressed in text has a number of applications, including tracking sentiment towards products, movies, politicians, etc.; improving customer relation models; and detecting happiness and well-being. In many applications, it is important to associate sentiment with a particular entity or an aspect of an entity. For example, in reviews, customers might express different sentiment towards various aspects of a product or service they have availed. Consider:

> *The lasagna was great, but the service was a bit slow.*

The review is for a restaurant, and we can gather from it that the customer has a positive sentiment towards the lasagna they serve, but a negative sentiment towards the service.

The SemEval-2014 Task 4 (Aspect Based Sentiment Analysis) is a shared task where given a customer review, automatic systems are to determine *aspect terms, aspect categories*, and sentiment towards these aspect terms and categories. An aspect term is defined to be an explicit mention of a feature or component of the target product or service. The example sentence above has the aspect term *lasagna*. Similar aspect terms can be grouped into aspect categories. For example, *lasagna* and other food items can be grouped into the aspect category of 'food'. In Task 4, customer reviews are provided for two domains: restaurants and laptops. A fixed set of five aspect categories is defined for the restaurant domain: food, service, price, ambiance, and anecdotes. Automatic systems are to determine if any of those aspect categories are described in a review. The example sentence above describes the aspect categories of food (positive sentiment) and service (negative sentiment). For the laptop reviews, there is no aspect category detection subtask. Further details of the task and data can be found in the task description paper (Pontiki et al., 2014).

Restaurants				**Laptops**	
Term	T-Sent.	Cat.	C-Sent.	Term	T-Sent.
3	2	1	1	3	1

Table 1: Rank obtained by NRC-Canada in various subtasks of SemEval-2014 Task 4.

We present an in-house sequence tagger to detect aspect terms and supervised classifiers to detect aspect categories, sentiment towards aspect terms, and sentiment towards aspect categories. A summary of the ranks obtained by our submissions to the shared task is provided in Table 1.

2 Lexical Resources

2.1 Unlabeled Reviews Corpora

Apart from the training data provided for Task 4, we compiled large corpora of reviews for restaurants and laptops that were not labeled for aspect terms, aspect categories, or sentiment. We generated lexicons from these corpora and used them as a source of additional features in our machine learning systems.

Yelp restaurant reviews corpus: The Yelp Phoenix Academic Dataset[1] contains customer reviews posted on the Yelp website. The businesses for which the reviews are posted are classified into over 500 categories. Further, many of the businesses are assigned multiple business categories. We identified all food-related business categories (58 categories) that were grouped along with the category 'restaurant' and extracted all customer reviews for these categories. We will refer to this corpus of 183,935 reviews as the *Yelp restaurant reviews corpus*.

Amazon laptop reviews corpus: McAuley and Leskovec (2013) collected reviews posted on Amazon.com from June 1995 to March 2013. A subset of this corpus is marked as reviews for electronic products. We extracted from this subset all reviews that mention either *laptop* or *notebook*. We will refer to this collection of 124,712 reviews as the *Amazon laptop reviews corpus*.

Both the Yelp and the Amazon reviews have one to five star ratings associated with each review. We treated the one- and two-star reviews as negative reviews, and the four- and five-star reviews as positive reviews.

2.2 Lexicons

Sentiment Lexicons: From the Yelp restaurant reviews corpus, we automatically created an in-domain sentiment lexicon for restaurants. Following Turney and Littman (2003) and Mohammad et al. (2013), we calculated a sentiment score for each term w in the corpus:

$$score\,(w) = PMI\,(w, pos) - PMI\,(w, neg) \quad (1)$$

where *pos* denotes positive reviews and *neg* denotes negative reviews. PMI stands for pointwise mutual information:

$$PMI\,(w, pos) = log_2 \frac{freq\,(w, pos) * N}{freq\,(w) * freq\,(pos)} \quad (2)$$

where $freq\,(w, pos)$ is the number of times a term w occurs in positive reviews, $freq\,(w)$ is the total frequency of term w in the corpus, $freq\,(pos)$ is the total number of tokens in positive reviews, and N is the total number of tokens in the corpus. $PMI\,(w, neg)$ was calculated in a similar way. Since PMI is known to be a poor estimator of association for low-frequency events, we ignored terms that occurred less than five times in each (positive and negative) groups of reviews.

A positive sentiment score indicates a greater overall association with positive sentiment, whereas a negative score indicates a greater association with negative sentiment. The magnitude is indicative of the degree of association.

Negation words (e.g., *not*, *never*) can significantly affect the sentiment of an expression (Zhu et al., 2014). Therefore, when generating the sentiment lexicons we distinguished terms appearing in negated contexts (defined as text spans between a negation word and a punctuation mark) and affirmative (non-negated) contexts. The sentiment scores were then calculated separately for the two types of contexts. For example, the term *good* in affirmative contexts has a sentiment score of 1.2 whereas the same term in negated contexts has a score of -1.4. We built two lexicons, *Yelp Restaurant Sentiment AffLex* and *Yelp Restaurant Sentiment NegLex*, as described in (Kiritchenko et al., 2014).

Similarly, we generated in-domain sentiment lexicons from the Amazon laptop reviews corpus.

In addition, we employed existing out-of-domain sentiment lexicons: (1) large-coverage automatic tweet sentiment lexicons, Hashtag Sentiment lexicons and Sentiment140 lexicons (Kiritchenko et al., 2014), and (2) three manually created sentiment lexicons, NRC Emotion Lexicon (Mohammad and Turney, 2010), Bing Liu's Lexicon (Hu and Liu, 2004), and the MPQA Subjectivity Lexicon (Wilson et al., 2005).

Yelp Restaurant Word–Aspect Association Lexicon: The Yelp restaurant reviews corpus was also used to generate a lexicon of terms associated with the aspect categories of food, price, service, ambiance, and anecdotes. Each sentence of the corpus was labeled with zero, one, or more of the five aspect categories by our aspect category classification system (described in Section 5). Then, for each term w and each category c an association score was calculated as follows:

$$score\,(w, c) = PMI\,(w, c) - PMI\,(w, \neg c) \quad (3)$$

2.3 Word Clusters

Word clusters can provide an alternative representation of text, significantly reducing the sparsity of the token space. Using Brown clustering algorithm (Brown et al., 1992), we generated 1,000 word clusters from the Yelp restaurant reviews corpus. Additionally, we used publicly available

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

word clusters generated from 56 million English-language tweets (Owoputi et al., 2013).

3 Subtask 1: Aspect Term Extraction

The objective of this subtask is to detect aspect terms in sentences. We approached this problem using in-house entity-recognition software, very similar to the system used by de Bruijn et al. (2011) to detect medical concepts. First, sentences were tokenized to split away punctuation, and then the token sequence was tagged using a semi-Markov tagger (Sarawagi and Cohen, 2004). The tagger had two possible tags: O for *outside*, and T for *aspect term*, where an aspect term could tag a phrase of up to 5 consecutive tokens. The tagger was trained using the structured Passive-Aggressive (PA) algorithm with a maximum step-size of $C = 1$ (Crammer et al., 2006).

Our features can be divided into two categories: emission and transition features. Emission features couple the tag sequence y to the input w. Most of these work on the token level, and conjoin features of each token with the tag covering that token. If a token is the first or last token covered by a tag, then we produce a second copy of each of its features to indicate its special position. Let w_i be the token being tagged; its token feature templates are: token-identity within a window $(w_{i-2} \dots w_{i+2})$, lower-cased token-identity within a window $(lc(w_{i-2}) \dots lc(w_{i+2}))$, and prefixes and suffixes of w_i (up to 3 characters in length). There are only two phrase-level emission feature templates: the cased and uncased identity of the entire phrase covered by a tag, which allow the system to memorize complete terms such as, "getting a table" or "fish and chips." Transition features couple tags with tags. Let the current tag be y_j. Its transition feature templates are short n-grams of tag identities: y_j; y_j, y_{j-1}; and y_j, y_{j-1}, y_{j-2}.

During development, we experimented with the training algorithm, trying both PA and the simpler structured perceptron (Collins, 2002). We also added the lowercased back-off features. In Table 2, we re-test these design decisions on the test set, revealing that lower-cased back-off features made a strong contribution, while PA training was perhaps not as important. Our complete system achieved an F1-score of 80.19 on the restaurant domain and 68.57 on the laptop domain, ranking third among 24 teams in both.

	Restaurants		
System	**P**	**R**	**F1**
NRC-Canada (All)	**84.41**	**76.37**	**80.19**
All − lower-casing	83.68	75.49	79.37
All − PA + percep	83.37	76.45	79.76

	Laptops		
System	**P**	**R**	**F1**
NRC-Canada (All)	**78.77**	**60.70**	**68.57**
All − lower-casing	78.11	60.55	68.22
All − PA + percep	77.76	61.47	68.66

Table 2: Test set ablation experiments for Subtask 1: Aspect Term Detection.

4 Subtask 2: Aspect Term Polarity

In this subtask, the goal is to detect sentiment expressed towards a given aspect term. For example, in sentence "The asian salad is barely eatable." the aspect term *asian salad* is referred to with negative sentiment. There were defined four categories of sentiment: positive, negative, neutral, or conflict. The conflict category is assigned to cases where an aspect term is mentioned with both positive and negative sentiment.

To address this multi-class classification problem, we trained a linear SVM classifier using the LibSVM software (Chang and Lin, 2011). Sentences were first tokenized and parsed with the Stanford CoreNLP toolkits[2] to obtain part-of-speech (POS) tags and (collapsed) typed dependency parse trees (de Marneffe et al., 2006). Then, features were extracted from (1) the target term itself; (2) its *surface context*, i.e., a window of n words surrounding the term; (3) the *parse context*, i.e., the nodes in the parse tree that are connected to the target term by at most three edges.

Surface features: (1) unigrams (single words) and bigrams (2-word sequences) extracted from a term and its surface context; (2) context-target bigrams (i.e., bigrams formed by a word from the surface context and a word from the term itself).

Lexicon features: (1) the number of positive/negative tokens; (2) the sum of the tokens' sentiment scores; (3) the maximal sentiment score. The lexicon features were calculated for each manually and automatically created sentiment lexicons described in Section 2.2.

Parse features: (1) word- and POS-ngrams in

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

System	Laptops Acc.	Rest. Acc.
NRC-Canada (All)	**70.49**	**80.16**
All − sentiment lexicons	63.61	77.13
All − Yelp lexicons	68.65	77.85
All − Amazon lex.	68.13	80.11
All − manual lexicons	67.43	78.66
All − tweet lexicons	69.11	78.57
All − parse features	69.42	78.40

Table 3: Test set ablation experiments for Subtask 2: Aspect Term Polarity.

	Restaurants		
System	P	R	F1
NRC-Canada (All)	**91.04**	**86.24**	**88.58**
All − lex. resources	86.53	78.34	82.23
All − W–A lexicon	88.47	80.10	84.08
All − word clusters	90.84	86.15	88.43
All − post-processing	91.47	84.78	88.00

Table 4: Test set ablation experiments for Subtask 3: Aspect Category Detection. 'W–A lexicon' stands for Yelp Restaurant Word–Aspect Association Lexicon.

the parse context; (2) context-target bigrams, i.e., bigrams composed of a word from the parse context and a word from the term; (3) all paths that start or end with the root of the target terms. The idea behind the use of the parse features is that sometimes an aspect term is separated from its modifying sentiment phrase and the surface context is insufficient or even misleading for detecting sentiment expressed towards the aspect. For example, in sentence "The food, though different from what we had last time, is actually great" the word *great* is much closer to the word *food* in the parse tree than in the surface form. Furthermore, the features derived from the parse context can help resolve local syntactic ambiguity (e.g., the word *bad* in the phrase "a bad sushi lover" modifies *lover* and not *sushi*).

Table 3 presents the results of our official submission on the test sets for the laptop and restaurant domains. On the laptop dataset, our system achieved the accuracy of 70.49 and was ranked first among 32 submissions from 29 teams. From the ablation experiments we see that the most significant gains come from the use of the sentiment lexicons; without the lexicon features the performance of the system drops by 6.88 percentage points. Observe that the features derived from the out-of-domain Yelp Restaurant Sentiment lexicon are very helpful on the laptop domain. The parse features proved to be useful as well; they contribute 1.07 percentage points to the final performance. On the restaurant data, our system obtained the accuracy of 80.16 and was ranked second among 36 submissions from 29 teams.

5 Subtask 3: Aspect Category Detection

The objective of this subtask is to detect aspect categories discussed in a given sentence. There are 5 pre-defined categories for the restaurant domain: food, price, service, ambience, and anecdotes/miscellaneous. Each sentence can be labeled with one or more categories from the predefined set. No aspect categories were defined for the laptop domain.

We addressed the subtask as a multi-class multi-label text classification problem. Five binary one-vs-all Support Vector Machine (SVM) classifiers were built, one for each category. The parameter C was optimized through cross-validation separately for each classifier. Sentences were tokenized and stemmed with Porter stemmer (Porter, 1980). Then, the following sets of features were generated for each sentence: ngrams, stemmed ngrams, character ngrams, non-contiguous ngrams, word cluster ngrams, and lexicon features. For the lexicon features, we used the Yelp Restaurant Word–Aspect Association Lexicon and calculated the cumulative scores of all terms appeared in the sentence for each aspect category. Separate scores were calculated for unigram and bigram entries. Sentences with no category assigned by any of the five classifiers went through the post-processing step. For each such sentence, a category c with the maximal posterior probability $P(c|d)$ was identified and the sentence was labeled with the category c if $P(c|d) \geq 0.4$.

Table 4 presents the results on the restaurant test set. Our system obtained the F1-score of 88.58 and was ranked first among 21 submissions from 18 teams. Among the lexical resources (lexicons and word clusters) employed in the system, the Word–Aspect Association Lexicon provided the most gains: an increase in F1-score of 4.5 points. The post-processing step also proved to be beneficial: the recall improved by 1.46 points increasing the overall F1-score by 0.58 points.

6 Subtask 4: Aspect Category Polarity

In the Aspect Category Polarity subtask, the goal is to detect the sentiment expressed towards a given aspect category in a given sentence. For each input pair (sentence, aspect category), the output is a single sentiment label: positive, negative, neutral, or conflict.

We trained one multi-class SVM classifier (Crammer and Singer, 2002) for all aspect categories. The feature set was extended to incorporate the information about a given aspect category c using a domain adaptation technique (Daumé III, 2007) as follows: each feature f had two copies, $f\_general$ (for all the aspect categories) and f_c (for the specific category of the instance). For example, for the input pair ("The bread is top notch as well.", 'food') two copies of the unigram top would be used: $top\_general$ and top_food. With this setup the classifier can take advantage of the whole training dataset to learn common sentiment features (e.g., the word *good* is associated with positive sentiment for all aspect categories). At the same time, aspect-specific sentiment features can be learned from the training instances pertaining to a specific aspect category (e.g., the word *delicious* is associated with positive sentiment for the category 'food').

Sentences were tokenized and part-of-speech tagged with CMU Twitter NLP tool (Gimpel et al., 2011). Then, each sentence was represented as a feature vector with the following groups of features: ngrams, character ngrams, non-contiguous ngrams, POS tags, cluster ngrams, and lexicon features. The lexicon features were calculated as described in Section 4.

A sentence can refer to more than one aspect category with different sentiment. For example, in the sentence "The pizza was delicious, but the waiter was rude.", food is described with positive sentiment while service with negative. If the words *delicious* and *rude* occur in the training set, the classifier can learn that *delicious* usually refers to food (with positive sentiment) and *rude* to service (with negative sentiment). If these terms do not appear in the training set, their polarities can still be inferred from sentiment lexicons. However, sentiment lexicons do not distinguish among aspect categories and would treat both words, *delicious* and *rude*, as equally applicable to both categories, 'food' and 'service'. To (partially) overcome this problem, we applied the Yelp Restaurant Word–Aspect Association Lexicon to collect all the terms having a high or moderate association with the given aspect category (e.g., *pizza*, *delicious* for the category 'food' and *waiter*, *rude* for the category 'service'). Then, the feature set described above was augmented with the same groups of features generated just for the terms associated with the given category. We call these features *aspect term features*.

System	Restaurants Accuracy
NRC-Canada (All)	**82.93**
All − lexical resources	74.15
All − lexicons	75.32
All − Yelp lexicons	79.22
All − manual lexicons	82.44
All − tweet lexicons	84.10
All − word clusters	82.93
All − aspect term features	82.54

Table 5: Test set ablation experiments for Subtask 4: Aspect Category Polarity.

Table 5 presents the results on the test set for the restaurant domain. Our system achieved the accuracy of 82.93 and was ranked first among 23 submissions from 20 teams. The ablation experiments demonstrate the significant impact of the lexical resources employed in the system: 8.78 percentage point gain in accuracy. The major advantage comes from the sentiment lexicons, and specifically from the in-domain Yelp Restaurant Sentiment lexicons. The out-of-domain tweet sentiment lexicons did not prove useful on this subtask. Also, word clusters did not offer additional benefits on top of those provided by the lexicons. The use of aspect term features resulted in gains of 0.39.

7 Conclusion

The paper describes supervised machine-learning approaches to detect aspect terms and aspect categories and to detect sentiment expressed towards aspect terms and aspect categories in customer reviews. Apart from common surface-form features such as ngrams, our approaches benefit from the use of existing and newly created lexical resources such as word–aspect association lexicons and sentiment lexicons. Our submissions stood first on 3 out of 4 subtasks, and within the top 3 best results on all 6 task-domain evaluations.

References

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Koby Crammer and Yoram Singer. 2002. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL '07, pages 256 – 263.

Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. 2011. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18(5):557–562.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the International Conference on Language Resources and Evaluation*, LREC '06.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL '11.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research (to appear)*.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, LA, California.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '13, Atlanta, Georgia, USA, June.

Olutobi Owoputi, Brendan OConnor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, pages 380–390.

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 International Workshop on Semantic Evaluation*, SemEval '14, Dublin, Ireland, August.

M.F. Porter. 1980. An algorithm for suffix stripping. *Program*, 3:130–137.

Sunita Sarawagi and William W Cohen. 2004. Semi-markov conditional random fields for information extraction. In *Advances in Neural Information Processing Systems*, volume 17, pages 1185–1192.

Peter Turney and Michael L Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4).

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*, HLT '05, pages 347–354, Stroudsburg, PA, USA.

Xiaodan Zhu, Hongyu Guo, Saif M. Mohammad, and Svetlana Kiritchenko. 2014. An empirical study on the effect of negation words on sentiment. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL '14.