IHS R&D Belarus: Cross-domain Extraction of Product Features using Conditional Random Fields

Maryna Chernyshevich
IHS Inc. / IHS Global Belarus
131 Starovilenskaya St.
220123, Minsk, Belarus
Marina.Chernyshevich@ihs.com

Abstract

This paper describes the aspect extraction system submitted by IHS R&D Belarus team at the SemEval-2014 shared task related to Aspect-Based Sentiment Analysis. Our system is based on IHS Goldfire linguistic processor and uses a rich set of lexical, syntactic and statistical features in CRF model. We participated in two domain-specific tasks – restaurants and laptops – with the same system trained on a mixed corpus of reviews. Among submissions of constrained systems from 28 teams, our submission was ranked first in laptop domain and fourth in restaurant domain for the subtask A devoted to aspect extraction.

1 Introduction

With a rapid growth of the blogs, forums, review sites and social networks, more and more people express their personal views about products on the Internet in form of reviews, ratings, or recommendations. This is a great source of data used by many researchers and commercial applications that are focused on the sentiment analysis to determine customer opinions.

Sentiment analysis can be done on document, sentence, and phrase level (Jagtap, V. S., Karishma Pawar, 2013). Earlier works were focused mainly on the document (Turney, 2002; Pang, Lee and Vaithyanathan, 2002) and the sentence level (Kim and Hovy, 2004). However, this information can be insufficient for customers who are seeking opinions on specific product features (aspects) such as design, battery life, or screen. This fine-grained classification is a topic of aspect-based sentiment analysis (Moghaddam and Ester, 2012).

Traditional approaches to aspect extraction are based on frequently used nouns and noun phrases (Popescu and Etzioni, 2005; Blair-Goldensohn et al., 2008), exploiting opinions (Zhuang et al., 2006; Kobayashi, 2006), and supervised learning (Mukherjee and Liu, 2012).

In this paper, we describe a system (IHS_RD_Belarus in official results) developed to participate in the international shared task organized by the Conference on Semantic Evaluation Exercises (SemEval-2014) and focused on the phrase-level sentiment classification, namely aspect extraction (Pontiki et al., 2014). An *aspect term* means particular feature of a product or service used in opinion-bearing sentences (`My phone has amazing screen`), as well as in neutral sentences (`The screen brightness automatically adjusts`).

The organizers of SemEval-2014 task have provided a dataset of customer reviews with annotated aspects of the target entities from two domains: restaurants (3041 sentences) and laptops (3045 sentences). The results were evaluated separately in each domain. Table 1 shows the distribution of the provided data for each domain dataset, training and testing set, with number of sentences and aspects.

	Laptops	Restaurants
Training		
Sentences	3045	3041
Aspects	2358	3693
Testing		
Sentences	800	800
Aspects	654	1134

Table 1. Distribution of the provided data.

Many studies showed that sentiment analysis is very sensitive to the source domain (training corpus domain) and performs poorly on data from other domain (Jakob and Gurevych, 2010). This restriction limits the applicability of in-

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 309–313,
Dublin, Ireland, August 23-24, 2014.

domain models to a wide domain diversity of reviews. One of the common approaches to develop a cross-domain system is training on a mixture of labeled data from different domains (Aue and Gamon, 2005). Cross-domain approach has the advantage of better portability, but it suffers from lower accuracy compared to in-domain aspect extraction. Our cross-domain system is trained on mixed training data, and the same model was used unchanged for classification of both domain-specific test datasets.

2 System Description

Aspect extraction may be considered as a sequence labeling task because the product aspects occur at a sequence in a sentence (Liu, 2014). One of the state-of-the-art methods used for sequence labeling is Conditional Random Fields (CRF) (Lafferty, 2001). This method takes as an input a sequence of tokens, calculates the probabilities of the various possible labelings and chooses the one with the maximum probability.

We decided to deviate from Inside-Outside-Begin (IOB) scheme used by Jakob and Gurevych (Jakob and Gurevych, 2010) and Li (Li et al., 2010) and introduced the following labels: **FA** for the attribute word preceding head word of a noun group; **FH** for the head word of a noun group; **FPA** for attribute word after head word of a noun group (`Microsoft Office 2003`), and **O** for other non-aspect tokens. The following is an example of our suggested tagging: `I`**`/O`** `want`**`/O`** `to`**`/O`** `unplug`**`/O`** `the`**`/O`** *`external`*`/`**`FA`** *`keyboard`*`/`**`FH`**.

Our experiments showed that the words used in aspect terms are easier to recognize when they are always tagged with the same tags. For example, let's consider the tagging of the word "camera" in the following cases: "camera" and "compact camera". We propose the **FH** tag for both examples, while the IOB scheme assumes the **FB** tag for the first example and the **FI** tag for the second.

2.1 Pre-processing

To facilitate feature generation for supervised CRF learning, sentences were pre-processed with IHS Goldfire linguistic processor that performs the following operations: slang and misspelling correction (*"excelent" → "excellent" , "amazin" → "amazing", "wouldnt" → "wouldn't"*), part-of-speech tagging, parsing, noun phrase extraction, semantic role labeling within expanded Subject-Action-Object (eSAO) relations (Todhunter et al., 2013), named entity recognition, labeling for predictive question-answering including rule-based sentiment analysis (Todhunter et al., 2014).

In addition, we designed some simple rules to detect entity boundaries that take precedence over CRF labeling. For example, in the sentence "`I run Final Cut Pro 7 and a few other applications`", our boundary detector recognizes "`Final Cut Pro 7`" as an entity represented by a single token (Tkachenko and Simanovsky, 2012).

2.2 Features

Below we will describe the features used in CRF model to represent the current token, two previous and two next tokens.

Word features:

- *Token feature* represents a base form of a token (word or entity) normalized by case folding. The vocabulary of terms is pretty compact within one domain, so this feature can have considerable impact on terms extraction performance.
- *Part of speech feature* represents the part-of-speech tag of the current token with slight generalization, for example, the NNS tag (plural noun) is mapped to NN (singular noun).
- *Named entity feature* labels named entities, e.g., people, organizations, locations, etc.
- *Semantic category* denotes the presence of the token in manually crafted domain-independent word-lists – sets of words having a common semantic meaning – such as parameter (characteristics of object, e.g., "durability"), process (e.g., "charging"), sentiment-bearing word (e.g., "problem"), person (e.g., "sister"), doer of an action (someone or something that performs an action, e.g., "organizer"), temporal word (date- or time-related words, e.g., "Monday"), nationality, word of reasoning (e.g., "decision", "reason"), etc.
- *Semantic orientation (SO) score of token* represents a low, mean or high SO score as separate feature values (the thresholds were determined experimentally). The SO of a word indicates the strength of its association with positive and negative reviews. We calculated SO of each word *w* using Pointwise Mutual Information (PMI) measures as

$$SO(w) = PMI(w, pr) - PMI(w, nr),$$

where PMI is the amount of information that we acquire about the presence of the word in positive *pr* or negative reviews *nr* (Turney, 2002). For the calculation of SO score, we used rated reviews from Epinions.com, Amazon.com and TripAdvisor.com. To make corpus more precise, we included only 5-star reviews in our positive corpus, and 1-star reviews in our negative corpus.

- *Frequency of token occurrence* is represented by five values ranging from very frequent to very rare words with an experimentally determined threshold. The frequency was obtained by dividing the number of reviews containing the token by the total number of reviews. The reason of using this as a feature is that people usually comment on the same product aspects and the vocabulary that they use usually converges (Liu, 2012).
- *Opinion target feature* is a binary feature that indicates whether a token is a part of an item which opinions are expressed on and comes from the rule-based sentiment analysis integrated in the predictive question-answering component of the IHS Goldfire linguistic processor. Opinion target can be a product feature as well as a product itself.

Noun phrase features:

- *Role of a token in a noun phrase*: head word or attribute word.
- *Noun phrase introduction feature* marks all tokens of noun phrase beginning with possessive pronoun, demonstrative pronoun, definite or indefinite article.
- *Number of attributes* with SO score higher than the experimentally chosen threshold. This feature labels all words in a noun group. Our research showed that people often use sentiment-bearing adjectives to describe an aspect, e.g., "`My phone has a great camera`".
- *List feature* was added to designate the availability of list indicators ("and" or comma) in the noun group, e.g., "`The leather carrying case, keyboard and mouse arrived in two days`".
- *Leaves-up feature* denotes the number of of-phrases in a noun phrase before the token under consideration. For example, the token "battery" has one preceding of-phrase in the phrase "durability of battery".
- *Leaves-down feature* denotes the number of of-phrases in a noun phrase after the token under consideration.

SAO features:

- *Semantic label feature* represents the role of the token in eSAO relation: subject, action, adjective, object, preposition, indirect object or adverb.
- *SAO feature* labels all words presented in an eSAO relation. We used a set of eSAO patterns to determine basic relations between words. To form a SAO pattern, each non-empty component of an eSAO relation was mapped to an abstract value, e.g., proper noun phrases to "PNP", common noun phrases to "CNP", predicates are left in their canonical form. For example, the sentence "`The restaurant Tal offers authentic chongqing hotpot.`" is represented by the SAO pattern "PNP offer CNP". All words from eSAO are marked with the same SAO feature.

2.3 Results and Experiments

Our CRF model was trained on the mixed set of 6086 sentences with annotated aspect terms (3045 from the laptop domain and 3041 from the restaurant domain). The same model was applied unchanged to the test dataset from laptop domain (800 sentences) and restaurant domain (800 sentences). We evaluated our system using 5-fold cross-validation: in each of the five iterations of the cross-validation, we used 80% of the provided training data for learning, and 20% for testing.

	laptops	restaurants
training set	0.707	0.7784
development set	0.7214	0.7865
test set	0.7455	0.7962
baseline	0.3564	0.4715

Table 2. Performance on different datasets (F_1-score).

The Table 2 shows the model performance (F_1-score) obtained on the training set (using 5-fold cross validation), on the development set (we used a part of the training set as development set), on the final test set and the baseline provided by the task organizers.

To evaluate the individual contribution of different feature sets, we performed ablation experiment, presented in Table 3. This test involves removing one of the following feature sets at a time: current token and its POS tag (TOK), combinations with two previous and two next tokens

and their POS tags (CONT), named entity (NE), semantic category (SC), semantic orientation (SO), word frequency (WF), opinion target (OT), noun phrase related features (NP_F), and SAO pattern and semantic label (SAO_F). Some features complement each other, so that despite small individual contribution, a cumulative improvement is generally achieved by using them in a set.

	Dev set		Test set	
	lap	rest	lap	rest
overall	**0.7214**	**0.7865**	**0.7455**	**0.7962**
-TOK	0.6642 (-7.9%)	0.7244 (-7.9%)	0.692 (-7.2%)	0.7445 (-6.4%)
-CONT	0.7101 (-1.6%)	0.77 (-2.1%)	0.7323 (-1.8%)	0.7811 (-1.9%)
-SC	0.6982 (-3.3%)	0.7854 (-0.1%)	0.7048 (-5.8%)	0.7864 (-1.2%)
-SO	0.709 (-1.7%)	0.7815 (-0.6%)	0.7442 (-0.2%)	0.7937 (-0.3%)
-OT	0.7026 (-2.6%)	0.7812 (-0.7%)	0.7381 (-1%)	0.7973 (0.1%)
-NP_F	0.717 (-0.6%)	0.777 (-1.2%)	0.7303 (-2%)	0.7801 (-2%)
-WF	0.716 (-0.8%)	0.788 (0.2%)	0.7399 (-0.7%)	0.7937 (-0.3%)
-SAO_F	0.7198 (-0.2%)	0.7854 (-0.1%)	0.7297 (-2.1%)	0.7981 (0.2%)
-NE	0.7191 (-0.3%)	0.7836 (-0.4%)	0.7444 (-0.1%)	0.7961 (0)

Table 3. Ablation experiment (F_1-score).

The importance of a feature set is measured by F_1-score on development and testing datasets for both domains separately.

Feature sets are listed in descending order of their impact on overall performance. The analysis shows that the most important feature set is the combination of Token and POS features. Other features contribute to the performance to a smaller degree.

As can be seen, the relative influence of features on F_1-score is similar on test and development sets, showing that our model effectively overcomes the overfitting problem.

We conducted several experiments on the training data to prove the domain portability of our CRF model. The results are shown in Table 4. As can be seen, the training on single-domain data improves the performance of in-domain classification by about 2%, but lowers the performance of cross-domain classification by about 40%. The training on the mixed dataset demonstrates acceptable accuracy on both domain-specific test sets.

Training dataset	Results on laptops dataset	Results on restaurants dataset
laptops	0.7667	0.3778
restaurants	0.2961	0.8223
mixed	0.7455	0.7962

Table 4. Results of classification with different training datasets (F_1-score).

2.4 Error Analysis and Further Work

The error analysis showed three main error types: not recognized, excessively recognized and partially recognized aspect terms (head word is recognized correctly, e.g., *"*`separate RAM memory`*"* instead of *"*`RAM memory`*"*). While first types are recall and precision errors respectively, partial aspect extraction yields both recall and precision errors. A summary of the errors on test dataset is presented in Table 5.

	laptops	restaurants
not recognized	68%	58%
partially recognized	18%	30%
excessively recognized	14%	12%

Table 5. Error types distribution.

From Table 5, we can see that a major source of errors is related to not recognized aspect terms. In the future, we would like to experiment with additional techniques to overcome recall problem, e.g., using dictionaries or concept taxonomies and employ skip-chain CRF, proposed by Li et al. (2010). Further improvements can also be made by tuning parameters of CRF learning.

To verify the cross-domain portability of the system, we are going to test it on a third domain test dataset without including additional instances in the training corpus, as proposed by Aue and Gamon (2005).

3 Conclusion

In this paper, we have presented a CRF-based learning technique applied to the aspect extraction task. We implemented rich set of lexical, syntactic and statistical features and showed that our approach has good domain portability and performance ranked first out of 28 participating teams in the laptop domain and fourth in restaurant domain.

References

Anthony Aue and Michael Gamon. 2005. Customizing sentiment classifiers to new domains: a case study. In *Proceedings of Recent Advances in Natural Language Processing*, RANLP-2005.

Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A. Reis, and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews. In *Proceedings of WWW-2008 workshop on NLP in the Information Explosion Era*.

V. S. Jagtap and Karishma Pawar. 2012. Analysis of different approaches to Sentence-Level Sentiment Classification. *International Journal of Scientific Engineering and Technology, Volume 2, Issue 3.*

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*, EMNLP'10.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of Interntional Conference on Computational Linguistics*, COLING'04.

Nozomi Kobayashi, Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2006. Opinion mining on the Web by extracting subject-attribute-value relations. In *Proceedings of AAAI-CAAW '06*.

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*, ICML'01.

Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, and Hao Yu. 2010. Structure-aware review mining and summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING'10.

Bing Liu. 2012. Sentiment Analysis and Opinion Mining.

Samaneh Moghaddam and Martin Ester. 2012. Aspect-based opinion mining from online reviews. Tutorial at SIGIR Conference.

Arjun Mukherjee and Bing Liu. 2012. Aspect Extraction through Semi-Supervised Modeling. In *Proceedings of 50th Anunal Meeting of Association for Computational Linguistics*, ACL'12.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, EMNLP'02.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, EMNLP'05.

Lawrence R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, 77(2): p. 257-286.

Maksim Tkachenko and Andrey Simanovsky. 2012. Named Entity Recognition: Exploring Features. In *Proceedings of* KONVENS'12.

James Todhunter, Igor Sovpel and Dzianis Pastanohau. System and method for automatic semantic labeling of natural language texts. *U.S. Patent 8 583 422, November 12, 2013*.

James Todhunter, Igor Sovpel and Dzianis Pastanohau. Question-answering system and method based on semantic labeling of text documents and user questions. *U.S. Patent 8 666 730, September 16, 2014*.

Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, ACL'02.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*, HLT/EMNLP'05.

Lei Zhang and Bing Liu. 2014. Aspect and Entity Extraction for Opinion Mining. Data Mining and Knowledge Discovery for Big Data.

Li Zhuang, Feng Jing, and Xiaoyan Zhu. 2006. Movie review mining and summarization. In *Proceedings of ACM International Conference on Information and Knowledge Management*, CIKM'06.

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), Dublin, Ireland.