BUTknot at SemEval-2016 Task 5: Supervised Machine Learning with Term Substitution Approach in Aspect Category Detection

Jakub Macháček
Brno University of Technology,
Faculty of Information Technology,
IT4Innovations Centre of Excellence
Božetěchova 2, 612 66 Brno, Czech Republic
qmachacek@stud.fit.vutbr.cz

Abstract

This paper describes an approach used to solve *Aspect Category Detection* (Subtask 1, Slot 1) of *SemEval 2016* Task 5. The core of the presented system is based on *Supervised machine learning* using bigram *bag-of-words* model. The performance is enhanced by several preprocessing methods, most importantly by a term substitution technique. The system has reached a very good performance in comparison with other submitted systems.

1 Introduction

As the Internet more and more becomes the means of expressing opinions about various subjects, the need to effectively process those opinions (subjective information) is becoming more and more important. Many companies around the world are now interested in gathering public opinion and performing strategic moves accordingly. Thus, *Sentiment Analysis* (also known as *Opinion mining*) has become an important area of interest.

Existing systems performing sentiment analysis usually only predict polarity of a given sentiment. While this can be sufficient in a lot of cases, sometimes we wish to analyze opinions about different aspects of the same entity. This task is known as *Aspect-based Sentiment Analysis*.

SemEval 2016 Task 5 (Pontiki et al., 2016) consists of three subtasks. The first subtask is about sentence-level sentiment analysis and is divided into three slots: *Aspect Category Detection*, *Opinion Target Expression* and *Sentiment Polarity Detection*.

There are multiple distinct domains in which participants were given the opportunity to test their systems. Each domain was available for one of the following languages: Arabic, Chinese, Dutch, English, French, Russian, Spanish and Turkish.

For all domains, there is a limited, known in advance, list of entities and aspects the system should recognize. Each entity can be associated with only certain aspect(s). The set of all possible entity-aspect pairs is also limited and known in advance.

Each submitted system could run in two modes: *Constrained* (using no external data sources like lexicons or additional training sets) and *Unconstrained* (no data source restriction).

The system I propose is focused on *Aspect Category Detection* only, i.e. it only predicts which aspects of a given entity a given sentiment has opinions about. I decided to participate in the English language, for which two domains were available: restaurant and laptop reviews. There were 12 entity-aspect pairs for the restaurants domain and over 80 for the laptops domain.

2 Approach

My approach was inspired by the NLANGP system (Toh and Su, 2015) which achieved excellent results in *SemEval 2015* (Pontiki et al., 2015) with a straightforward solution.

I model the task as a *multi-label* classification with *binary relevance* transformation, where labels correspond to the entity-aspect pairs. All train and test sentences are pre-processed (see Section 2.2). Words from each sentence are used as individual binary features of that sentence. For each entity-aspect

Proceedings of SemEval-2016, pages 306–310,
San Diego, California, June 16-17, 2016.

pair, all training sentences are used as positive or negative examples of that entity-aspect pair.

Vowpal Wabbit[1], a supervised machine learning tool, is used to train the resulting binary classifiers. More precisely, a variant of *online gradient descent* algorithm is used to perform *logistic regression* with *squared cost function*.

2.1 Model properties

It has been observed that the model offers higher accuracy if bigrams are allowed in VW. However, additional raising the n of VW's n-gram feature is counterproductive. Allowing skips inside bigrams also does not help. For the RESTAURANT#MISCELLANEOUS aspect category, however, setting $n = 5$ seemed to be a better choice. The system has generally unsatisfying score of predicting aspect categories associated with the MISCELLANEOUS attribute. Setting $n > 2$ did not improve the accuracy for any other aspect category.

I have also tuned VW's *learning rate* and a threshold T for comparing probabilities returned from VW's classifiers. When it predicts that a sentiment has opinion(s) about an aspect category C with the probability of P, the system drops the prediction iff $P < T$. Table 1 shows the tuned properties.

Domain	Restaurants	Laptops
Learning rate	0.41	0.38
Prediction threshold T	0.40	0.34

Table 1: Tuned learning rate and the prediction threshold

2.2 Text pre-processing

The initial text pre-processing step consists of removing the punctuation from each sentence and converting all characters to lower case. *Stanford CoreNLP*[2] is used to tokenize each sentence, *lemmatize* all words and also extract their *part of speech* (*POS*).

2.2.1 Filtering words

The POS tags are useful to estimate how much important given words are in terms of aspect category detection. The system does not introduce the tags to the machine learning algorithm, but it uses each tag to consider removing the corresponding word from its sentence. Also, some words are removed regardless of their POS.

An automated experiment over the oficial training set has been implemented to produce a POS filter and a list of stop words, for each domain separately. Both features are included in the constrained mode.

The experiment showed that for the restaurants domain, it was not demonstrably beneficial to remove any word based on its POS. On the other hand, it generated a list of tags for the laptops domain.

The lists of stop words that the experiment produced were surprisingly small. It seemed there were just few high frequency words which were irrelevant in the learning process.

Food	Service	Opinion	Laptop
pizza	staff	good	computer
sauce	clerk	average	machine
entree	she	terrific	notebook
hot dog	friendly	disappointing	netbook
croquette	attentive	poor	desktop
...	...	...	...

Table 2: Example term lists compilation

2.2.2 Term groups

As a part of the pre-processing phase, a simple substitution system has been implemented to support machine learning. When multiple n-grams have roughly the same meaning, or they are related in a certain way, it is often beneficial if classifiers do not distinguish between them. A set of n-gram (term) lists in the fashion depicted in Table 2 has been manually compiled. Presence of the listed terms is then checked in each sentence and if found, each occurrence is replaced by its representative. Terms are always compared by their lemmas. The lists have been compiled by applying the following:

1. For each entity and attribute independently, only those sentences from the train dataset that contain the entity or attribute were selected. Then all unigrams from the sentences were sorted by number of occurrences. The most frequent words were manually checked, one by one, and the ones closely related to a particular entity or attribute were added to the respective lists.

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

[2] Homepage: http://stanfordnlp.github.io/CoreNLP/

2. In case of the restaurants domain, opinion targets from the train datasets were also extracted. This resulted in much shorter lists of high precision terms. All terms could be therefore individually checked in a reasonable time. Again, terms closely related to an entity or attribute were added to the respective lists. Some n-grams were split into multiple pieces, e.g. *lava cake dessert* $\rightarrow$ {*lava cake*, *dessert*}.

3. While performing the preceding two methods, I also noticed that some terms played a certain role in aspect category detection even if they were not associated with just a single entity or attribute but rather a set of them. This, for example, included opinion words (indicating attributes GENERAL and QUALITY) and words describing problems (e.g. *fail*, *problematic*, *blue screen* – indicating attributes OPERATION_PERFORMANCE, QUALITY and possibly some other).

The following methods for extending the lists were also used but they are not applied in the constrained mode:

4. For some specific words (e.g. adjectives expressing food taste), an online dictionary[3] has been used to search for their synonyms and the term lists have been manually appended with suitable words.

5. Several lists of words publicly available on the Internet have also been included:

   - Food: http://eatingatoz.com/food-list/ and https://www.atkins.com/how-it-works/atkins-20/phase-1/low-carb-foods; both lists have been manually checked and some misleading items removed (mostly words related to drinks)
   - Drinks: http://cocktails.lovetoknow.com/List_of_Popular_Cocktails and others compiled manually
   - Laptop manufacturers: https://en.wikipedia.org/wiki/List_of_laptop_brands_and_manufacturers#Major_brands
   - Laptop series: manually extracted names of laptop series available at https://en.wikipedia.org/wiki/Asus#Laptops, https://en.wikipedia.org/wiki/List_of_Hewlett-Packard_products#Business_notebooks and http://www.acer.com/ac/en/US/content/models/laptops
   - Processors: manually extracted names of CPU series from Wikipedia pages https://en.wikipedia.org/wiki/List_of_AMD_microprocessors and https://en.wikipedia.org/wiki/List_of_Intel_microprocessors
   - Operation systems: manually extracted names of mainstream Linux distributions from the list published at https://en.wikipedia.org/wiki/List_of_Linux_distributions
   - Screen resolution names: https://en.wikipedia.org/wiki/Display_resolution#Common_display_resolutions

[3] Search engine available at http://www.thesaurus.com/

2.2.3 Other pre-processing steps

The system also tries to improve its accuracy by replacing all numbers by the word *number*. Numbers preceded by a currency symbol (\$, □, £) are replaced with the word *price* (which then indicates the PRICES attribute).

Words containing both alpha and numeric characters are replaced with the word *model* as it was observed that in most cases, such words represent particular model names (e.g. *i7*, *G73JH-x3*, *d620*). This seemed to be helpful notably in the laptops domain.

The system removes all words shorter than two characters. It is important to note that this step comes after removing all non-alphanumeric characters. This means words like *w/* are eventually removed.

The system also neutralizes consecutive letters which effectively replaces words like *waay* or *waaaaaay* by the word *way*.

2.3 Prediction post-processing

The presence of previously detected opinion words indicates the QUALITY and GENERAL attributes. The corresponding aspect categories are correctly

Domain	Rest	Lapt
Basic text pre-processing	63.9	49.6
Bigrams	65.3	50.1
Minimal word length	65.7	50.7
Lemmatization	65.8	50.7
Consecutive letters neutralization	65.8	
Prices recognition	66.1	50.9
Numbers+models recognition	66.2	51.0
POS filter		52.4
Stop words	66.4	52.4
Prediction post-processing	68.0	
Term groups (constrained)	71.6	53.5
Term groups (unconstrained)	72.1	53.8

Table 3: The F-measure in percentage using 4-fold cross validation over the training sets. Each row represents the system's accuracy when the corresponding technique and all those from previous rows are enabled. The prediction threshold, as well as VW's learning rate, is always optimized. Missing values represent unused techniques.

predicted by VW only in cases in which the opinion words are directly preceded or followed by words indicating entities (i.e. they make up bigrams). When these words are separated by one or more other words, no aspect category is predicted. For this reason, the system always looks for an entity-related word which is closest to the opinion word and still not farther than four skips. If such word is found, the corresponding aspect category is additionally predicted. The same process is repeated with the PRICES attribute with the difference that when no suitable entity is found, RESTAURANT#PRICES is predicted. The system does no post-processing in the laptops domain.

3 Results

All techniques and features described in Section 2 have been tuned separately for each domain using *4-fold cross* validation. Table 3 displays the reached accuracy.

The tuned system has been trained using the official training sets containing 2000 and 2500 sentences in the restaurants and laptops domains respectively. The test sets consisted of 676 sentences in the restaurants domain and 808 sentences in the laptops domain.

The system achieved very good results, especially in the restaurants domain where it ranked *third* in the unconstrained mode, falling behind the winner only by 0.635%, and *first* in the constrained mode. Table 4 shows the F-score of the top five systems in each domain and mode. The best accuracy in the unconstrained mode has been reached in both domains by the NLANGP team which has ranked first also in SemEval 2015.

Domain	Restaurants		Laptops	
Mode	C	U	C	U
1st place	**71.494**	73.031	47.891	51.937
2nd place	68.701	72.886	**47.527**	49.105
3rd place	67.817	**72.396**	46.728	49.076
4th place	67.350	71.537	45.629	**48.396**
5th place	65.563	**71.494**	43.754	47.891

Table 4: Rankings in ACD (slot 1) of the Subtask 1. The F-score of the submitted systems is represented in percentage. Results of my system are highlighted in bold-faced font. C stands for the constrained and U for the unconstrained mode.

4 Conclusion

This paper described my approach to aspect category detection. The presented system has ranked as one of the most accurate in this task. As I have never contributed to the field of aspect-based sentiment analysis (and SA generally) before, I find my results more than satisfactory.

The system has an advantage of its relatively high adaptability to work with previously unseen domains. All techniques except the term groups and the prediction post-processing can be tuned automatically with no additional manual help needed. Multilinguality is limited by the used lemmatizer(s) but since lemmatization offers just a mild increase in accuracy, it is possible to omit it when working with unsupported languages.

The system does not take review context of input sentences into consideration. In my future work I would like to remove this flaw. The term groups described in Section 2.2.2 will be significantly improved by extending them and creating new groups for other important terms.

Acknowledgments

This work was supported by the H2020 project *MixedEmotions*, grant agreement No. 644632.

References

Everton Alvares Cherman, Maria Carolina Monard, and Jean Metz. 2011. Multi-label problem transformation methods: a case study. *CLEI Electronic Journal*, 14(1).

José Saias. 2015. Sentiue: Target and Aspect based Sentiment Analysis in SemEval-2015 Task 12. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado.

Satarupa Guha, Aditya Joshi, and Vasudeva Varma. 2015. SIEL: Aspect Based Sentiment Analysis in Reviews. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado.

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*. Seattle, Washington, USA.

John Pavlopoulos. 2014. *Aspect based sentiment analysis*. PhD thesis, Dept. of Informatics, Athens University of Economics and Business, Greece.

Maria Pontiki, Dimitrios Galanis, Haris 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)*, Denver, Colorado.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeny Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June. Association for Computational Linguistics.

Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. 2014. On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter. *Ninth International Conference on Language Resources and Evaluation*, pages 810–817.

Zhiqiang Toh and Jian Su. 2015. NLANGP: Supervised Machine Learning System for Aspect Category Classification and Opinion Target Extraction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado.

Bo Wang and Min Liu. 2015. Deep Learning for Aspect-Based Sentiment Analysis.

Lei Zhang and Bing Liu. 2014. Aspect and Entity Extraction for Opinion Mining. In *Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenges, and Opportunities*. Springer.