AUEB-ABSA at SemEval-2016 Task 5: Ensembles of Classifiers and Embeddings for Aspect Based Sentiment Analysis

Dionysios Xenos, Panagiotis Theodorakakos, John Pavlopoulos,
Prodromos Malakasiotis and **Ion Androutsopoulos**
NLP Group, Department of Informatics
Athens University of Economics and Business
Patission 76, GR-104 34 Athens, Greece
http://nlp.cs.aueb.gr

Abstract

This paper describes our submissions to the Aspect Based Sentiment Analysis task of SemEval-2016. For Aspect Category Detection (Subtask1/Slot1), we used multiple ensembles, based on Support Vector Machine classifiers. For Opinion Target Expression extraction (Subtask1/Slot2), we used a sequence labeling approach with Conditional Random Fields. For Polarity Detection (Subtask1/Slot3), we used an ensemble of two supervised classifiers, one based on hand crafted features and one based on word embeddings. Our systems were ranked in the top 6 positions in all the tasks we participated. The source code of our systems is publicly available.

1 Introduction

The amount of user-generated content on the web has grown rapidly in recent years, leading to increased interest in sentiment analysis and, more generally, opinion mining. The task of Aspect Based Sentiment Analysis of SemEval-2014 (SE-ABSA14) and SemEval-2015 (SE-ABSA15) was concerned with identifying the aspects of given target entities and extracting the sentiment expressed towards each aspect (Pontiki et al., 2014; Pontiki et al., 2015). The task of Aspect Based Sentiment Analysis of SemEval-2016 (SE-ABSA16) is a continuation of those tasks (Pontiki et al., 2016). We participated in Aspect Category Detection (ACD, Subtask1/Slot1), Opinion Target Expression (OTE, Subtask1/Slot2), and Polarity Detection (PD, Subtask1/Slot3).

In ACD, we participated in the English language, for both Laptops and Restaurants, submitting both constrained and unconstrained systems. Our constrained system used only the provided training data for the corresponding domain. Features were extracted from lexicons created from the training data. One Support Vector Machine (SVM) classifier (Vapnik and Vapnik, 1998) was trained for each Entity and Attribute category (called E and A respectively). Our unconstrained system used word embeddings as additional resources (Mikolov et al., 2013). For each category (E or A), we used an ensemble of two systems: our constrained system and one new system, which was based on word embeddings.

In OTE, we participated with both a constrained and an unconstrained system.[1] The task is to identify aspects of given target entities. We addressed the problem as a sequential labeling task (Toh and Wang, 2014), assigning one label to each word in a sentence, indicating whether the word was an aspect term or not. In this task, we used Conditional Random Fields (Lafferty et al., 2001). Similarly to ACD, our unconstrained system differed in that it also used word embeddings as features.

In PD, we participated only with an unconstrained system, in both domains, in the English language. We used an ensemble of two classifiers. The first classifier used hand crafted features and sentiment lexicons with scores. The second classifier was based on IDF-weighted centroids of the word embeddings of each sentence (Kosmopoulos et al., 2015).

The remainder of this paper is structured as follows. In Section 2, we describe our systems in detail, including data preprocessing and feature de-

[1] Only the restaurants domain was available in OTE.

Proceedings of SemEval-2016, pages 324–329,
San Diego, California, June 16-17, 2016.

scriptions. In Sections 3 and 4, we present our official results and experiments, respectively. Finally, Section 5 summarizes our work and proposes future directions.

2 Systems

All our submissions used supervised learning. In Restaurants, the training data were 350 reviews (2,000 sentences), annotated with 2,499 aspects and their polarities. Each aspect consists of one Entity (E) and one Attribute of E (A), thus, forming an E#A pair. The sentiment polarity of any aspect may be positive, negative or neutral. Included, were also annotations for linguistic expressions, called Opinion Target Expressions (OTE), indicating the origin of each E. For example, in "The food was well prepared and the service impeccable." there were two annotations:

- 1st OTE: "food" offsets: 4 to 8
  - category: FOOD#QUALITY
  - polarity: positive
- 2nd OTE: "service" offsets: 35 to 42
  - category: SERVICE#GENERAL
  - polarity: positive

In Laptops, the training data were 450 reviews (2,500 sentences), annotated with 2,923 {E#A, Polarity} labels. In this domain, no OTE annotations were included.

As a preprocessing step, we excluded sentences with no opinion tuples and sentences labeled as "Out of Scope".[2] All the features and hyper-parameter values used are described in a publicly available report, along with the source code of our systems.[3]

2.1 Aspect Category Detection

In Aspect Category Detection (ACD), each aspect E#A (e.g., FOOD#PRICE) in a sentence should be discovered. The possible E and A labels were predefined; thus, we considered this to be a classification task.

Constrained ACD system

One Support Vector Machine (SVM) classifier was trained for each predefined E and A, based on lexicons created from the training data.[4] [5] The lexicons assigned scores to unigrams (stemmed and unstemmed) and bigrams (stemmed, unstemmed or using only POS tag bigrams). For each unigram or bigram, we computed its Precision, Recall and F1 over the training data, following the work of Karampatsis et al. (2014). We used the average, median, maximum, and minimum values for each score (Precision, Recall, F1) and for each lexicon (stemmed unigrams, unstemmed unigrams, stemmed, unstemmed, POS tag bigrams), thus, yielding 12 features per lexicon and 60 features overall.

One Support Vector Machine (SVM) classifier was trained for each E and A label, yielding 11 classifiers for Restaurants and 31 for Laptops. Given a new text, the confidence scores of all the classifiers were examined and the Es and As of the classifiers whose confidence exceeded a threshold were used to form the E#A aspects.[6] All the possible E#A combinations are formed, provided that they appeared at least once in the training data. Thus, assuming that the classifiers of $|E|$ entities and $|A|$ attributes exceeded the threshold, we form at most $|E| \cdot |A|$ aspects.

Following the work of Hsu et al. (2010), we used a 5-fold cross validation on the training data for tuning and we performed a loose grid search, followed by a fine grid search. During the loose grid search, various kernels were examined (i.e., Linear, Sigmoid, Polynomial and RBF) and hyper-parameter values were searched with a big step and in a big range of values. During the fine grid search, the best kernel of the loose grid search was used and hyper-parameter values were searched with a smaller step in a much smaller range of values.

Unconstrained ACD system

Our unconstrained system was based on multiple ensembles, one for each E and A combination encountered in our training data. Each ensemble

[2] Sentences including opinions that can not be described by the SE-ABSA16 annotation schema, are "Out of Scope".

[3] https://github.com/nlpaueb/aueb-absa

[4] Lexicons were created only for tokens that appeared more than 2 times in the training dataset, for each E and A category.

[5] http://scikit-learn.org/stable/

[6] The threshold was manually fixed to 0.4.

returned the linear combination of the confidence scores of two systems.[7] The first system in each ensemble was the corresponding constrained system. The second system in each ensemble was a system based on word embeddings (Mikolov et al., 2013).

We used the Amazon product review data to produce word embeddings and Inverse Document Frequency (IDF) scores (McAuley et al., 2015).[8] Word embeddings were produced, with 200 dimensions and the skip-gram model, using Gensim.[9] Then, following the work of Kosmopoulos et al. (2015), for each sentence s_i in our data, we computed its centroid $c(s_i)$ as follows:

$$c(s_i) = \frac{\sum_{j=1}^{|V|} e_j \cdot TF(w_j, s_i) \cdot IDF(w_j)}{\sum_{i=1}^{|V|} TF(w_j, s_i) \cdot IDF(w_i)}$$

where $|\text{V}|$ is the size of the vocabulary, e_j is the embedding of word w_j, $\text{TF}(w_j, s_i)$ is the Boolean term frequency of w_j in sentence s_i, and $IDF(w_j)$ is the IDF score of w_j. Preliminary experiments in both domains, with IDF scores and term frequency (Boolean or not) scores, showed that the use of Boolean term frequency scores along with IDF scores was better than any other combination.

We normalized each centroid using L2 normalization and we computed the cosine similarity between the centroid of the sentence and the word embedding of the label of each possible E or A. The final feature vector of each unconstrained E or A classifier (2nd classifier in each ensemble) is produced by concatenating all the cosine scores and the normalized centroid. We trained one SVM classifier per E and A, yielding 11 unconstrained classifiers for Restaurants and 31 for Laptops, as already noted. All the unconstrained classifiers were tuned similarly to our constrained classifiers.

In a final step, our unconstrained system returns one score per E and A, which is the linear combination of the confidence scores of the constrained and unconstrained classifiers in the corresponding ensemble. Similarly to our constrained system, E#A aspects are then returned.

[7] The weights of the two confidence scores were set to 0.5, i.e. we used the average of the two scores. Other weighs were also examined, but they did not lead to better performance.

[8] We used the files for individual product categories from the Amazon product data corpus, which had duplicate item reviews removed (`http://jmcauley.ucsd.edu/data/amazon/`).

[9] `https://radimrehurek.com/gensim/`

2.2 Opinion Target Expression

Opinion Target Expression is addressed as a sequential labeling problem. Each word in a sentence is assigned the label "B" to indicate the start of an aspect term, "I" to indicate the continuation of an aspect term, and "O" if the token is not an aspect term.

Both our constrained and unconstrained systems use Conditional Random Fields (CRF) (Lafferty et al., 2001).[10] However, our unconstrained system extends our constrained one, by incorporating more features. The features used in each system are discussed below.

Constrained OTE system

The features of our constrained system are the following:

- Morphological (Boolean), about the current token:
  - Capital first letter
  - All letters in capitals
  - Only digits
  - Existence of punctuation mark
  - Other
- Lexicon-based (one-hot)
  - POS tags (both of current token and context)
  - Word affixes (prefixes and suffixes) of the current token
  - Aspect terms

For features based on lexicons, we used the training data to form lists of POS tags, word affixes of various lengths (prefixes and suffixes of length 1, 2 and 3 characters) and aspect terms. Then, we formed a single one-hot vector (i.e., a vector of 0's and a single 1) per list, indicating which member of the list the token being examined corresponds to. We used 5 vectors for the POS based features (one for the current token and 4 for the context; two on the left and two on the right), 6 vectors for features based on word affixes (prefixes and affixes of one, two and

[10] `https://pystruct.github.io/index.html`

three letters) and 1 vector for aspect terms seen in the training data. Finally, we concatenated the vectors and the Boolean features to yield one overall feature vector.

Unconstrained OTE system

Our unconstrained system extends our constrained system by incorporating the following features:

- Word embedding of the current token
- Word embeddings of the context

For each word, we compute its embedding and incorporate it to our constrained system's feature vector. Word embeddings were calculated as in Section 2.1; each word corresponds to a 200-dimensional embedding. We also incorporate features for the word's context. We compute the embeddings of the words in a 5-context window (i.e., two words on the left and two on the right) and concatenate them to a 1000-dimensional vector.

If a word from the context had no embedding, we replaced it by the previous (for left context) or next (for right context) word that had an embedding. Also, in order to cope with missing words at the beginning and the end of a sentence, we introduced four special tokens, for the first two and the last two words of the sentence. The respective tokens were positioned before and after each sentence of the Amazon product reviews corpus, before producing word embeddings.

2.3 Polarity Detection

The objective of Polarity Detection (PD) was to detect the correct sentiment label for each aspect E#A in a sentence; possible sentiment labels were positive, negative and neutral (i.e., mildly positive or negative). Sentences could contain multiple aspects; e.g., the sentence "Excellent food, although the interior could use some help." contains two aspects; "FOOD#QUALITY" and "AMBIENCE#GENERAL", which should be labeled positive and negative respectively.

We used an Ensemble of two Multi-class Logistic Regression (LR) classifiers, trained with different sets of features. Each classifier yields one confidence sentiment (for each E#A) per sentiment label. Then, for each sentiment label, our ensemble computes a linear combination of the corresponding two scores of the classifiers and the label with the highest combined score is returned.[11]

We describe below the feature sets of the two LR classifiers, LRI and LRII.

LR I PD classifier

In LRI, we used 50 hand crafted features, which could be categorized by nature:

- Morphological (Karampatsis et al., 2014)
  - frequency based (number of exclamation and question marks, etc.)
  - Boolean based (exclamation or question mark in the end of the sentence, etc.)
- POS based (number of nouns, adjectives, verbs and adverbs) (Karampatsis et al., 2014)
- E#A based (number of aspects and bags of entities and attributes)
- Sentiment lexicons
  - AFINN[12]
  - Hu & Liu[13]
  - NRC[14]

For features based on sentiment lexicons, besides the already given word scores, we compute new scores, which are relative to our data. Following the work of Karampatsis et al. (2014), we computed Precision and $F1$ scores per sentiment label, for each word in the lexicon and for each POS bigram (i.e., two sequential part of speech tags).

To take into consideration negation phenomena, we used the negation lexicon compiled by Zhang et al. (2014). If a negation word precedes a word in a lexicon, we reverse the word's score sign; i.e., positive becomes negative and vice versa. Also, we make special use of words only in upper case. If a word in upper case exists in a lexicon we multiply

[11] We used the average of the two scores in practice.

[12] http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010,AFINN-111

[13] http://github.com/woodrad/Twitter-Sentiment-Mining/tree/master

[14] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

Domain	Precision	Recall	F-measure	Rank
Restaurants (U)	67.75%	75.77%	71.54%	4th/30
Restaurants (C)	64.22%	70.79%	67.35%	4th/12
Laptops (U)	45.60%	53.19%	49.10%	2nd/22
Laptops (C)	40.69%	51.94%	45.63%	4th/9

Table 1: AUEB-ABSA's evaluation in Aspect Category Detection. The first column shows the domain and the run (C for constrained and U for unconstrained). The last column shows the rank of our system.

Domain	Precision	Recall	F-measure	Rank
Restaurants (U)	71.82%	69.12%	70.44%	2nd/19
Restaurants (C)	64.35%	58.99%	61.55%	6th/8

Table 2: AUEB-ABSA's evaluation in Opinion Target Expression.

its score, in order to make it more significant.[15] The resulting feature vector is normalized to [0,1] with the Euclidean norm.

LR II PD classifier

The second PD classifier uses the centroid of the word embeddings of each sentence as features. The centroids are compiled as in Section 2.1, but without the IDF scores in the denominator.[16] Words without embeddings or IDF scores are ignored when computing the centroids. The same applies to words with IDF scores below a given threshold.[17] Word embeddings are normalized with the Euclidean norm.

3 Results

Table 1 shows the results of our Aspect Category Detection (ACD) system. In Restaurants, our unconstrained ACD system was ranked 4th among 30 submissions. Our constrained ACD system was also ranked 4th, but amongst 12 submissions. In Laptops, our unconstrained ACD system was ranked 2nd among 22 submissions and our constrained one was ranked 4th among 9 submissions.

Table 2 shows the results of our Opinion Target Expression (OTE) system. Our unconstrained OTE system was ranked 2nd out of 19 submissions, while our constrained OTE system was ranked 6th out of 8 submissions.

[15] Here, we arbitrarily tripled the score

[16] Our experiments have shown that removing the IDF from the denominator improves the performance of LRII

[17] The threshold was set to 0.5, which led to best results on the validation data.

Domain	Accuracy	Rank
Restaurants (U)	83.24%	6th/28
Laptops (U)	76.90%	6th/21

Table 3: AUEB-ABSA's evaluation on Polarity Detection.

Table 3 shows the evaluation of our PD system (unconstrained only). We were ranked 6th out of 28 submissions in Restaurants and 6th out of 21 submissions in Laptops.

4 Experiments & Discussion

As described in Section 2.1, our unconstrained ACD system used an ensemble of two systems, one based on word embeddings and one based on features calculated only on the training data. Preliminary experiments showed that the ensemble is better than each system alone and better than a single system combining all the features of the two systems.

In the same task, experiments in the domain of Restaurants showed that there is no important difference between having one classifier for each possible E#A label, and having separate classifiers for each E and each A label. However, in Laptops, the latter approach led to slightly better results and, hence, it was preferred.

In the task of OTE, our constrained system is below the median participant (6th out of 8 submissions). However, when extended with features based on word embeddings (our unconstrained system), its performance is outstanding (2nd out of 19 submissions).

It is also worth noting that, in preliminary experiments for OTE, the use of a context vector (i.e., concatenated embeddings of words in a 5-context window of the word in question) gave far better results than using the centroid of these word embeddings.

System	Restaurants	Laptops
LRI (hand crafted)	80.71% (12th)	73.92% (9th)
LRII (embeddings)	74.94% (22th)	73.12% (11th)
Ensemble	**83.24% (6th)**	**76.90% (6th)**

Table 4: Accuracy of our submitted PD ensemble and its two subsystems, LRI (feature based) and LRII (based on word embeddings). In parentheses are estimated ranks of the two subsystems and the official rank of our ensemble.

For PD, we performed some post-experiments on the gold test data. As can be seen in Table 4,

our ensemble, which is our final PD system, outperforms its two subsystems, both in Restaurants (2.5% - 8.3%) and Laptops (3%-3.8%). Also, post-experiments showed that the negation lexicon improved the Accuracy of our ensemble by 0.5% in Restaurants and 1% in Laptops.

5 Conclusions and future work

We presented our approach to sentence level Aspect Based Sentiment Analysis (SE-ABSA16), which includes the subtasks of Aspect Category Detection (ACD), Opinion Target Expression (OTE) and Polarity Detection (PD). We observed and showed in post-experiments the benefits of using word embeddings and ensembles. The source code of our systems is publicly available. Future work includes the incorporation of neural networks in our ensembles.

Acknowledgments

This work was carried out during the BSc projects of the first two authors, which were co-supervised by the other three authors.

References

C. Hsu and C. Chang. 2010. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University.

R. M. Karampatsis, J. Pavlopoulos, and P. Malakasiotis. 2014. Aueb: Two stage sentiment analysis of social network messages. In *Proceedings of SemEval 2014, at COLING 2014*, pages 114–118, Dublin, Ireland.

A. Kosmopoulos, I. Androutsopoulos, and G. Paliouras. 2015. Biomedical semantic indexing using dense word vectors in bioasq. *Journal of Biomedical Semantics, supplement on Semantics-Enabled Biomedical Information Retrieval*, pages 5–7.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th ICML*, pages 282–289, San Francisco, CA, USA.

J. McAuley, R. Pandey, and J.Leskovec. 2015. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD*, pages 785–794.

T. Mikolov, I. Sutskever, G. S. Corrado K. Chen, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS 2013*, pages 3111–3119.

M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of SemEval 2014*, pages 27–35, Dublin, Ireland.

M. Pontiki, D. Galanis, H. Papageogiou, S. Manandhar, and I. Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of SemEval 2015*, pages 27–35, 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 SemEval 2016*, San Diego, California.

Z. Toh and W. Wang. 2014. Dlirec: Aspect term extraction and term polarity classification system. In *Proceedings of SemEval 2014*, pages 235–240, Dublin, Ireland.

V. N. Vapnik and V. Vapnik. 1998. *Statistical learning theory*, volume 1. Wiley New York.

F. Zhang, Z. Zhang, and M. Lan. 2014. Ecnu: A combination method and multiple features for aspect extraction and sentiment polarity classification. In *Proceedings of SemEval 2014*, pages 252–258, Dublin, Ireland.