IIT-TUDA at SemEval-2016 Task 5: Beyond Sentiment Lexicon: Combining Domain Dependency and Distributional Semantics Features for Aspect Based Sentiment Analysis

Ayush Kumar
IIT Patna, India
Dept. of Computer Science
ayush.cs12@iitp.ac.in

Sarah Kohail
TU Darmstadt, Germany
Computer Science Dept,
Language Technology Group
kohail@lt.informatik.tu-darmstadt.de

Amit Kumar
IIT Patna, India
Dept. of Computer Science
amit.mtmc14@iitp.ac.in

Asif Ekbal
IIT Patna, India
Dept. of Computer Science
asif@iitp.ac.in

Chris Biemann
TU Darmstadt, Germany
Computer Science Dept,
Language Technology Group
biem@cs.tu-darmstadt.de

Abstract

This paper reports the IIT-TUDA participation in the SemEval 2016 shared Task 5 of Aspect Based Sentiment Analysis (ABSA) for subtask 1. We describe our system incorporating domain dependency graph features, distributional thesaurus and unsupervised lexical induction using an unlabeled external corpus for aspect based sentiment analysis. Overall, we submitted 29 runs, covering 7 languages and 4 different domains. Our system is placed first in sentiment polarity classification for the English laptop domain, Spanish and Turkish restaurant reviews, and opinion target expression for Dutch and French in restaurant domain, and scores in medium ranks for aspect category identification and opinion target extraction.

1 Introduction

The advent of web technologies has made an unprecedented opportunity for online users to share and explain their views and opinions. The correlation between the views expressed by the users and the market strategies by the organizations strikes the importance of analyzing such reviews. But, valuable as they are, user-generated review texts are unstructured and very noisy. Major research studies adopted Natural Language Processing (NLP) and text mining techniques to better understand and process various types of information in user-generated reviews. Such efforts have come to be known as opinion mining, sentiment analysis or review mining (Pang and Lee, 2008).

Aspect level analysis performs a finer-grained sentiment analysis by addressing three subproblems: extracting aspects from the review text, identifying the entity that is referred to by the aspect, and finally classifying the opinion polarity towards the aspect (Liu, 2012). For example, a review of the "entity" laptop is likely to discuss distinct "aspects" like size, processing unit, and memory, and a single product can trigger a positive "opinion" about one feature, and a negative "opinion" about another.

In an attempt to support the efforts on Aspect Based Sentiment Analysis (ABSA), the SemEval 2016 shared Task 5 ABSA (Pontiki et al., 2016) offers the opportunity to experiment and evaluate on benchmark datasets (reviews) across various domains and languages through three subtasks. Subtask 1 covers the three sub-problems mentioned above, namely: aspect category identification (Slot 1), opinion target expression (OTE) (Slot 2) and sentiment polarity classification (Slot 3). We participated in Slot 1 and Slot 3 for English, Spanish, Dutch, French, Turkish, Russian and Arabic language for all available domains except telecoms. We also conducted experiments for Slot 2 for English, Spanish, Dutch and French. Overall, we submitted 29 runs, covering 7 languages and 4 different domains.

Proceedings of SemEval-2016, pages 311–317,
San Diego, California, June 16-17, 2016.

2 Method for Aspect Based Sentiment Analysis

In this section, we describe our data preprocessing and feature extraction. We also introduce an unsupervised approach for expanding the coverage of existing lexical resources based on the notion of distributional thesaurus. We use Support Vector Machine (SVM) (Cortes and Vapnik, 1995) as the baseline classifier for aspect category detection and sentiment polarity classification, and Conditional Random Fields (CRF) (Lafferty et al., 2001) for opinion target expression identification.

2.1 Preprocessing

We tokenize the data using Stanford tokenizer, normalize all digits to 'num' and remove stop words for tf-idf computation. For opinion target expression, we run Stanford CoreNLP[1] suite in order to extract information such as lemma, Part-of-Speech (PoS) and named entity (NE) in English language. For languages other than English, we use the universal parser[2] for tokenization and parsing. Since we deal with the OTE as a sequence labelling problem, it is necessary to identify the boundary of OT properly. We follow the standard BIO notation, where 'B-ASP', 'I-ASP' and 'O' represent the beginning, intermediate and outside tokens of a multi-word OTE respectively. e.g. In, '*Chow (B-ASP) fun (I-ASP) was (O) very (O) dry (O) . (O)*', 'Chow Fun' is the OTE.

2.2 Features for Aspect Category Detection

- Domain Dependency Graph: We use the aspects list produced by Domain Dependency Graph (DDG) for each domain by (Kohail, 2015). The idea is to detect topics underlying a mixed-domain dataset, aggregate individual dependency relations between domain-specific content words, weigh them with tf-idf and produce a DDG by selecting the highest-ranked words and their dependency relations. Since the domains are already given, no topic modeling is required. However, only one domain was provided for French and Spanish, we used external reviews dataset to compute tf-idf. We use movies reviews[3] for Spanish; and books, music and DVD reviews[4] for French. The resulting graphs were filtered and only 'amod' (adjective modifying a noun) and 'nsubj' (nominal subjects of predicates) relations were selected. For each extracted aspect from the opinion-aspect pairs, we determine the existence or absence of this aspect using a binary feature.

- Distributional Thesaurus: A Distributional Thesaurus (DT) is an automatically computed lexical resource that ranks words according to the semantic similarity. We employ an open source implementation of DT computation as described in (Biemann and Riedl, 2013). For every top five significant words based on tf-idf score in each aspect category (for example: 'overpriced', '$', 'pricey', 'cheap', 'expensive' are the most significant terms in 'food#price' category), we find ten most similar words according to DT. The presence or absence of these words in the review is used as a feature for aspect category identification. Examples from the distributional thesaurus are presented in Table 1.

Token	DT Expansion
drinks	beers, wines, coffee, liquids, beverage
price	prices, pricing, cash, cost, pennies
fresh	fresher, new, refreshing, clean, frozen
laptop	pc, computer, notebook, tablet, imac
toshiba	samsung, sony, acer, asus, dell
touchpad	mouse, trackball, joystick, trackpad

Table 1: Example of DT expansions for frequent aspects.

- Tf-Idf Score: Each aspect category has some discriminative aspect terms. We extract a maximum of top five distinguishing words in each category based on tf-idf score. Presence or absence of each token in the review denotes the feature.

- Bag of Words: This feature denotes the number of occurrences of each word in the review.

[1] nlp.stanford.edu/software/corenlp.shtml

[2] http://www.undl.org/unlsys/uparser/UP.htm

[3] http://www.lsi.us.es/~fermin/index.php/Datasets

[4] http://www.uni-weimar.de/en/media/chairs/webis/corpora/corpus-webis-cls-10/

2.3 Features for Opinion Target Expression

- Word and Local Context: We use the current token, its lowercase form and the context tokens in a window of [-5..5] as features.
- Part-of-Speech (PoS) Information: We use PoS information of the current, preceding two and following two tokens as the features.
- Head Word and its PoS: We use the head word of the noun phrase and PoS information of the head word.
- Prefix and Suffix: We use prefix and suffix of length up to four characters.
- Frequent Aspect Term: We build a list of frequently occurring OTEs from the training set. An OTE is considered to be frequent if it appears at least four times in the training corpus. We define a binary feature for the presence or absence of extracted OTEs.
- Dependency Relations: In English language, features are defined in line with (Toh and Wang, 2014). For other languages, feature is defined by considering whether the current token is present in dependency relations 'nsubj', 'dep', 'amod', 'nmod' and 'dobj' or not.
- Character N-grams: We use all substrings up to length 5 of the current token as features.
- Orthographic feature: This feature checks whether the current token starts with the capitalized letter or not.
- DT features: We use the top 5 DT expansions of current token as the features.
- Expansion Score: OTEs have opinion around them. Opinions are regularly lexicalized with words found in sentiment lexicons. We calculate sentiment score based on SentiWordNet[5] (Esuli and Sebastiani, 2006) in English language. For Non-English language, we use our induced lexicons. We calculate sentiment score by considering the window size of 10 (preceding 5 and following 5 tokens of the target one).

We additionally extract the following features only for English language.

- Chunk information: To identify the boundaries of multi-word OTEs, we use chunk information of the current token as the features.
- Lemma: Lemmatization trims the inflectional forms and derivationally related forms of a token to a common base form.
- WordNet: We use top 4 noun synsets of current token from WordNet as the features.
- Named entity information: We extract named entity information of the current token with Stanford CoreNLP tool, and use the NER-sequence labels in BIO-scheme as features.

2.4 Features for Sentiment Polarity Classification

- Lexical Acquisition: We use lexical expansion for inducing sentiment words based on distributional hypothesis. We observe that for rare words, unseen instances and limited coverage of available lexicons, the distributional expansion can provide a useful backoff technique, also cf. (Govind et al., 2014).

  For all languages, we construct a polarity lexicon using an external corpus and seed sentiment lexicon. For seed lexicons, we use English (Salameh et al., 2015) and Arabic (Salameh et al., 2015) versions of Bing Liu's lexicon (Hu and Liu, 2004) for English and Arabic respectively, VU sentiment lexicon[6] for French, Dutch and Spanish, a lexicon by (Panchenko, 2014) for Russian, and SentiTurkNet (Dehkharghani et al., 2015) and NRC Emotion for Turkish[7]. For inducing a lexicon, we obtain the top 100 DT expansion of each word in the seed lexicon. Next we accept candidate terms that a) occur in the expansions of at least 10 seed terms, b) have a corpus frequency

[5] http://sentiwordnet.isti.cnr.it/

[6] https://github.com/opener-project/VU-sentiment-lexicon

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

Language	Seed Lexicon			Induced Lexicon	Common Entries
	Positive	Negative	Neutral		
English	2005	4789	-	12953	4120
Dutch	3314	5923	-	8496	2992
French	9338	10339	5993	18308	7636
Spanish	2175	1737	7869	12480	4306
Russian	3217	8849	-	7697	2945
Turkish	1900	2515	1382	6547	1838
Arabic	1916	4467	-	9077	1447

Table 2: Expansion statistics for induced lexicons. Common entries denote the number of words which are present both in the seed lexicon and the induced lexicon.

of more than 50 in the background corpus (English[8], French[9], Spanish[10], Dutch[11], Russian[12], Arabic[13]). Finally, we compute the normalized positive, negative and neutral score for each word similar to (Kumar et al., 2015), and inspired by (Hatzivassiloglou and McKeown, 1997). The core assumption is that words tend to be semantically more similar to words of same sentiment. Hence, words appearing more in the expansions of positive (negative/neutral) words get assigned a higher positive (negative/neutral) sentiment score, Here, in difference to (Kumar et al., 2015), we compute normalized positive, negative and neutral scores rather than assigning one of the polarity class to the words. It should be noted that the volume of induced lexicon depends on two factors: (i) number of words in the seed lexicon that have expansions and (ii) pruning threshold for obtaining the induced lexicon. The unavailability of expansions for all words in the seed lexicon and higher threshold on conditions for accepting candidate terms reduces the volume of induced lexicon. Expansion statistics for the induced lexicons are provided in Table 2.

We compute the sum of positive, negative and neutral scores of tokens using induced lexicon for that language as features. In addition, scores as given in the seed lexicon are also used as features. For English, we also computed these features from different lexicons: AFINN (Nielsen, 2011), NRC Hashtag, Sentiment 140 (Zhu et al., 2014), NRC Emotion (Mohammad and Turney, 2013) and Bing Liu (Hu and Liu, 2004).

- Word N-gram: All unigrams and bigrams tokens are extracted from the training set are used as a binary feature, where 1 and 0 indicates the presence and absence of n-grams in the review.
- Entity-Attribute Pair: We use E#A pair as a binary feature for sentiment classification.

3 Datasets, Experimental Results and Discussions

For feature selection and hyperparameter tuning, we perform five-fold cross-validation on the training set. For Slot 1 and Slot 3, we use supervised classification using Support Vector Machine (SVM)[14]. Based on cross-validation results, we set the probability threshold of 0.185, 0.13 and 0.145 for restaurants, laptops and phones, respectively, for predicting aspect categories in the review. All E#A pairs having predicted probability greater than the threshold are enlisted as aspect categories. For Slot 2, we use CRFSuite[15] with default parameters. CRF-

[8] https://snap.stanford.edu/data/web-Amazon.html

[9] http://wacky.sslmit.unibo.it/doku.php?id=corpora

[10] http://corporafromtheweb.org/escow14/

[11] http://corporafromtheweb.org/nlcow14/

[12] lib.ruc.ecebooks

[13] http://corpora2.informatik.uni-leipzig.de

[14] https://github.com/bwaldvogel/liblinear-java

[15] http://www.chokkan.org/software/crfsuite/

Language	Domain	Slot 1: F1/#Entries	Slot 2: F1/#Entries	Slot 3: Acc./#Entries
English	Restaurants	63.0 (U, 17), 61.2 (C, 20) / 30	42.6 (U, 18) / 19	86.7 (U, 2) / 29
Dutch	Restaurants	55.2 (U, 3), 54.9 (C, 4) / 6	56.9 (U, 1) / 3	76.9 (U, 2) / 4
Spanish	Restaurants	59.8 (U, 6), 59.0 (C, 7) / 9	64.3 (U, 3) / 5	83.5 (U, 1) / 5
French	Restaurants	57.8 (U, 2), 57.0 (C, 3) / 6	66.6 (U, 1) / 3	72.2 (U, 5) / 6
Russian	Restaurants	62.6 (C, 3), 58.1 (C, 4) / 7	-	73.6 (U, 3) / 6
Turkish	Restaurants	56.6 (U, 3), 55.7 (C, 4) / 5	-	84.2 (U, 1) / 3
Dutch	Phones	45.4 (U, 2), 45.0 (C, 3) / 4	-	82.5 (U, 2) / 3
English	Laptops	43.9 (U, 12), 42.6 (C, 14) / 22	-	82.7 (U, 1) / 22
Arabic	Hotels	-	-	81.7 (U, 2) / 3

Table 3: Evaluation result for Subtask 1. Mode of submission (C-constrained, U-unconstrained) and rank is given in the parenthesis. F1 and Acc. denote F1-Score and Accuracy. #Entries is the total number of submissions for respective slot and domains.

Dataset	All Features	All w/o E#A Pair	All w/o Induced Lexicon
English Restaurants	86.729	86.224	86.390
English Laptops	82.772	82.310	82.457
Dutch Restaurants	76.998	76.250	74.228
Dutch Phones	82.576	82.058	80.896
Russian Restaurants	73.615	73.158	70.657
French Restaurants	72.222	71.898	70.154
Spanish Restaurants	83.582	82.920	79.589
Turkish Restaurants	84.277	83.650	80.788
Arabic Hotels	81.720	80.650	78.680

Table 4: Feature Ablation Experiment for Sentiment Polarity Classification

Suite is a fast implementation of Conditional Random Field (CRFs) for segmenting and labelling sequential data.

Teams were allowed to submit their systems in two modes: constrained and unconstrained modes. In constrained mode, the participants are allowed to use only the resources and dataset provided by the organizers whereas in unconstrained submission, participants can use any external resource. For Slot 2 and Slot 3, we only sent unconstrained submission, while for Slot 1 we sent constrained as well as unconstrained submissions except for Russian restaurants.

Our system achieves the best results in sentiment polarity classification for reviews about English laptops, Spanish restaurants and Turkish restaurants. We score second for English restaurants. We also produce the maximum F1-score value for opinion target expression for French and Dutch restaurants. Our evaluation results across all domains and languages are given in Table 3.

The results show that our system performs comparably well for sentiment polarity classification and opinion target expression. A feature ablation experiment given in Table 4 shows the effectiveness of induced lexicon for Slot 3 task. We get a significant improvement on adding information from the induced lexicons in each language. This holds especially for languages other than English, where existing sentiment lexicons are less comprehensive. We also note that entity-attribute pairs also help in resolving conflicting sentiments (for example: *cheap food (positive)* to *cheap service (negative)*).

We score in medium ranks for Slot 1 task. Distributional thesaurus based expansion for discriminative terms and aspects obtained through domain dependency graph results in marginal increments. This could be attributed to conflict in very fine grained aspect categories (for example: Restaurant#Prices, Food#Prices, Drink#Prices)

Language	F1-measure
English	68.45
Dutch	64.37
Spanish	69.73
French	69.64

Table 5: Result on Slot 2 task after modification

which may not have been captured explicitly by the external features. For the Slot 2 task, we have found some inconsistencies in our extraction pipeline, unfortunately were not able to correct them in time for the submission.

After the evaluation period, we revised our feature representation to ensure that it matches the correct input format for CRF. We also added two new features including unsupervised PoS tags (Biemann, 2009) as the feature for all the languages and SentiWordNet score for English language. For the current token, we use PoS tag, distributional thesaurus, lexical expansion score, unsupervised PoS tag, SentiWordNet score of context tokens [-2, -1, 0, 1, 2], prefix (upto 3-character), suffix (upto 3-character) and chunk information of context tokens [-1, 0, 1]. The updated results of after modification are shown in Table 5. If we had incorporated these changes earlier, we would have scored third for English and first for the other three languages.

4 Conclusions and Future Work

In this paper, we report our work on the task of Aspect Based Sentiment Analysis, which covers three slots: aspect identification, opinion target extraction and sentiment polarity classification. By leveraging a distributional thesaurus, we expand the existing domain specific aspect list and sentiment lexicons for different languages to reach a higher coverage on sentiment words. Our system was ranked first in five out of 29 submitted runs. While performance is satisfactory for Slot 3 and Slot 2 (after correction), our setup compares infavorably to others for Slot 1. We will continually improve our system in future work.

References

Chris Biemann and Martin Riedl. 2013. Text: Now in 2D! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95.

Chris Biemann. 2009. Unsupervised Part-of-Speech Tagging in the Large. *Research on Language and Computation*, 7(2):101–135.

Corinna Cortes and Vladimir Vapnik. 1995. Support vector machine. *Machine learning*, 20(3):273–297.

Rahim Dehkharghani, Yucel Saygin, Berrin Yanikoglu, and Kemal Oflazer. 2015. SentiTurkNet: a Turkish polarity lexicon for sentiment analysis. *Language Resources and Evaluation*, pages 1–19.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 417–422, Genoa, Italy.

Govind, Asif Ekbal, and Chris Biemann. 2014. Multi-objective optimization and unsupervised lexical acquisition for named entity recognition and classification. In *Proceedings the 11th International Conference on Natural Language Processing (ICON)*, Goa, India.

Vasileios Hatzivassiloglou and Kathleen R McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181, Madrid, Spain.

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*, pages 168–177, Seattle, WA.

Sarah Kohail. 2015. Unsupervised topic-specific domain dependency graphs for aspect identification in sentiment analysis. In *Student Research Workshop Associated with the International Conference Recent Advances in Natural Language Processing (RANLP 2015)*, pages 16–23, Hissar, Bulgaria.

Ayush Kumar, Sarah Kohail, Asif Ekbal, and Chris Biemann. 2015. IIT-TUDA: System for sentiment analysis in indian languages using lexical acquisition. In *Mining Intelligence and Knowledge Exploration*, pages 684–693. Hyderabad, India.

John D Lafferty, Andrew McCallum, and Fernando C N 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, pages 282–289, Williamstown, MA, USA.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, pages 93–98, Heraklion, Greece.

Alexander Panchenko. 2014. Sentiment index of the Russian speaking Facebook. In *In Proceedings of International Conference on Computational Linguistics. Dialogue 2014*, pages 506–517, Moscow, Russia.

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

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. Association for Computational Linguistics.

Mohammad Salameh, Saif M Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, Denver, Colorado.

Zhiqiang Toh and Wenting Wang. 2014. Dlirec: Aspect term extraction and term polarity classification system. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 235–240, Dublin, Ireland.

Xiaodan Zhu, Svetlana Kiritchenko, and Saif M Mohammad. 2014. NRC-Canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 443–447, Dublin, Ireland.