UoM: Using Explicit Semantic Analysis for Classifying Sentiments

Sapna Negi
University of Malta
Msida, MSD2080, MALTA
sapna.negi13@gmail.com

Mike Rosner
University of Malta
Msida, MSD2080, MALTA
mike.rosner@um.edu.mt

Abstract

In this paper, we describe our system submitted for the Sentiment Analysis task at SemEval 2013 (Task 2). We implemented a combination of Explicit Semantic Analysis (ESA) with Naive Bayes classifier. ESA represents text as a high dimensional vector of explicitly defined topics, following the distributional semantic model. This approach is novel in the sense that ESA has not been used for Sentiment Analysis in the literature, to the best of our knowledge.

1 Introduction

Semantic relatedness measure gives the comparison of different terms or texts on the basis of their meaning or the content. For instance, it can be said that the word ”computer” is semantically more related to ”laptop” than ”flute”. Sentiment analysis refers to the task of determining the overall contextual polarity of the written text. In this paper, we propose the use of semantic relatedness models, specifically Explicit Semantic Analysis (ESA), to identify textual polarity. There are different approaches to model semantic relatedness like WordNet based models (Banerjee and Banerjee, 2002), distributional semantic models (DSMs) etc. DSMs follow the distributional hypothesis, which says that words occurring in the same contexts tend to have similar meanings (Harris, 1954). Therefore, considering sentiment classification problem, distributional hypothesis suggests that the words or phrases referring to positive polarity would tend to co-occur, and similar assumptions can be made for the negative terms.

DSMs generally utilize large textual corpora to extract the distributional information relying on the co-occurrence information and distribution of the terms. These models represent the text in the form of high-dimensional vectors highlighting the co-occurrence information. Semantic relatedness between two given texts is calculated by using these vectors, thus, following that the the semantic meaning of a text can be inferred from its usage in different contexts. There are several different computational models following distributional semantics hypothesis. Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) (Blei et. al., 2003), Explicit Semantic Analysis (ESA) are some examples of such models. However, in this work, we investigated the use of ESA for the given task of sentiment analysis (SA).

There are two sub-tasks defined in Task 2 at SemEval 2013 (SemEval, 2013). We participated in *Message Polarity Classification* sub-task, where we are required to automatically classify the sentiment of a given message into positive, negative, or neutral. The task deals with the short texts coming from Twitter and SMS (Short Message Service). We are provided with 8,000 - 12,000 twitter messages annotated with their sentiment label for the purpose of training the models. In this work, we present our approach for sentiment classification which uses a combination of ESA and Naive Bayes classifier. The rest of the paper is structured as follows : Section 2 discusses some related work in this context. Section

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 535–538, Atlanta, Georgia, June 14-15, 2013. ©2013 Association for Computational Linguistics

3 briefly explains ESA. Section 4 describes our approaches while Section 5 explains the submitted runs for our system to the task. Section 6 reports the results, and we conclude in section 7.

2 Related Work

The research in SA initiated with the classical machine learning algorithms like Naive Bayes, Maximum Entropy etc. using intuitive features like unigrams, bigrams, parts of speech information, position of words, adjectives etc. (Pang et. al., 2002). However, such approaches are heavily dependent upon the given training data, and therefore can be very limited for SA due to out of vocabulary words and phrases, and different meanings of words in different contexts (Pang and Lee, 2008). Due to these problems, several methods have been investigated to use some seed words for extracting more positive and negative terms with the help of lexical resources like WordNet etc., for instance, SentiWordNet, which defines the polarity of the word along with the intensity. In this paper, we model the sentiment classification using DSMs based on explicit topic models (Cimiano et. al., 2009), which incorporate correlation information from a corpus like Wikipedia, to generalize from a few known positive or negative terms. There have been some other attempts to utilize topic models in this regards, but they mainly focussed on latent topic models (Lin and He, 2009) (Maas et. al., 2011). Joint sentiment topic model introduced LDA based unsupervised topic models in sentiment analysis by pointing out that sentiments are often topic dependent because same word/phrase could represent different sentiments for different topics (Lin and He, 2009). The recent work by Maas et. al. (Maas et. al., 2011) on using latent concept models presented a mixture model of unsupervised and supervised techniques to learn word vectors capturing semantic term-document information along with the sentiment content.

3 Explicit Semantic Analysis

Explicit Semantic Analysis (ESA) is a technique for computing semantic relatedness between texts using distributional information (Gabrilovich and Markovitch, 2007). ESA represents text as vectors of concepts explicitly defined by humans, like Wikipedia articles. This provides an intuitive and easily understandable topic space for humans, in contrast to the latent topic space in latent models.Input texts are represented as multidimensional vectors of weighted concepts. The procedure of computing semantic relatedness involves comparing the vectors corresponding to the given texts e.g. using cosine product. The magnitude of each dimension in the vector is the associativity weight of the text to that explicit concept/dimension. To quantify this associativity, the textual content related to the explicit concept/dimension is utilized. This weight can be calculated by considering different methods, for instance, tf-idf score. ESA has been proved to be a generalized vector space model (Gottron et. al., 2011).

4 Methodology

We implemented a combination of traditional machine learning based approach for SA using Naive Bayes algorithm, and ESA based sentiment identification. To perform sentiment classification solely using ESA, we asses the similarity of a new text against the text whose sentiment is already known, using ESA. More similar is a text to a particular sentiment annotated text, better are its chances to belong to the same sentiment class. On the other hand, we followed a standard classification approach by learning Naive Bayes over the given training data. Finally, we consult both ESA and Naive Bayes for classifying the text. The overall probability of a text belonging to a particular sentiment class was determined by weighted sum of ESA similarity score, and the scores given by Naive Bayes classifier. The sentiment class with the highest total score was accepted as the sentiment of the input text. The individual weights of ESA and Naive Bayes were determined by linear regression for our experiments.

5 System Description

We created three bags of words (BOW) corresponding to the different sentiment classes (positive, negative, and neutral) annotated in the training data. These BOWs were used as the definition of the particular sentiment class for making the ESA comparisons, and for learning Naive Bayes. We used unigrams and bigrams as features for the Naive

Task	Approach	F score	Highest F score	Rank
Twitter, with constrained data	ESA with Naive Bayes	.5182	.6902	24/35
SMS, with constrained data	ESA with Naive Bayes	.422	.6846	24/28
Twitter, with unconstrained data	ESA with Naive Bayes	.4507	.6486	16/16
SMS, with unconstrained data	ESA with Naive Bayes	.3522	.4947	15/15
Twitter, with constrained data	ESA	.35	.6902	NA

Table 1: Results

Bayes algorithm. The ESA implementation was replicated from the version available on Github[1], replacing the Wikipedia dump by the version released in February 2013.

We submitted two runs each for Twitter and SMS test data. The first run (constrained) used only the provided training data for learning while the second run (unconstrained) used a combination of external training data coming from the popular movie review dataset (Pang et. al., 2002), and the data provided with the task.

6 Results and discussion

The first four entries provided in the table 1 correspond to the four runs submitted in SemEval-2013 Task 2. The fifth entry corresponds to the results of a separate experiment performed by us, to estimate the influence of ESA on SA. According to the F-scores, ESA is unable to identify the sentiment in the texts following the mentioned approach. The results suggest that combining Naive Bayes to the system improved the overall scores. However, even the combined system could not perform well. Also, the mixing of external data lowered the scores indicating incompatibility of the external training data with the provided data.

7 Conclusion

We presented an approach of using ESA for sentiment classification. The submitted system follow a combination of standard Naive Bayes model and ESA based classification. The results of the task suggests that the approach we used for ESA based classification is unable to identify the sentiment accurately. As a future step, we plan to investigate more on the usability of ESA for sentiment classification, for instance, by using suitable features in the concept definitions, and weighing them according to the different sentiment classes.

[1] `https://github.com/kasooja/clesa`

References

Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. pp. 142–150. HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2002472.2002491

Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 375–384. CIKM '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1645953.1646003

Cimiano, P., Schultz, A., Sizov, S., Sorg, P., Staab, S.: Explicit versus latent concept models for cross-language information retrieval. In: Proceedings of the 21st international jont conference on Artifical intelligence. pp. 1513–1518. IJCAI'09, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2009), http://dl.acm.org/citation.cfm?id=1661445.1661688

Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. 2(1-2), 1–135 (Jan 2008), http://dx.doi.org/10.1561/1500000011

Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10. pp. 79–86. EMNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002), http://dx.doi.org/10.3115/1118693.1118704

Wilson, T., Kozareva, Z., Nakov, P., Rosenthal, S., Stoyanov, V., Ritter, A.: SemEval-2013 task 2: Sentiment analysis in twitter. In: Proceedings of the International Workshop on Semantic Evaluation. SemEval '13 (June 2013)

Banerjee, S., Banerjee, S.: An adapted lesk algorithm for word sense disambiguation using wordnet. In: In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics. pp. 136–145 (2002)

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

P. W. Foltz, W. Kintsch, and T. K. Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25:285–307, 1998.

Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, 2007.

Thomas Gottron, Maik Anderka, and Benno Stein. Insights into explicit semantic analysis. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 1961–1964, New York, NY, USA, 2011. ACM.

Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.