Semantic Spaces for Sentiment Analysis

Ivan Habernal[1,2] and Tomáš Brychcín[1,2]

[1] Department of Computer Science and Engineering
Faculty of Applied Sciences, University of West Bohemia
Univerzitní 8, 306 14 Plzeň, Czech Republic
nlp.kiv.zcu.cz
[2] NTIS – New Technologies for the Information Society
Faculty of Applied Sciences, University of West Bohemia
Univerzitní 22, 306 14 Plzeň, Czech Republic
{habernal,brychcin}@kiv.zcu.cz

Abstract. This article presents a new semi-supervised method for document-level sentiment analysis. We employ a supervised state-of-the-art classification approach and enrich the feature set by adding word cluster features. These features exploit clusters of words represented in semantic spaces computed on unlabeled data. We test our method on three large sentiment datasets (Czech movie and product reviews, and English movie reviews) and outperform the current state of the art. To the best of our knowledge, this article reports the first successful incorporation of semantic spaces based on local word co-occurrence in the sentiment analysis task.

Key words: document-level sentiment analysis, semantic spaces

1 Introduction

Supervised document-level sentiment analysis belongs to a very popular branch of opinion mining research [1]. Although the mainstream target language of this research area is English, other languages have been recently gaining attention. However, the lack of linguistic resources (e.g., sentiment lexicons) or publicly-available labeled datasets represents the main obstacle in research for many non-mainstream languages. The datasets for supervised sentiment analysis research usually deal with product or movie reviews, as the labeled data can be easily obtained from the web.

Many current approaches rely on bag-of-word (or bag-of-n-gram) document representation and features based on various weighting metrics of word frequencies [2]. However, for languages with high flection, such as Czech, the feature vectors are very sparse due to a large vocabulary; even after stemming or lemmatization [3].

To tackle the issue of data sparsity in sentiment analysis of Czech, we investigate the possibilities of clustering semantically similar words.[3] Words with similar meanings are clustered according to their distance in semantic spaces that are computed on unlabeled data. We enrich the baseline feature set by additional word cluster features.

[3] Recent research on language modeling of inflectional languages revealed that clustering semantically similar words can rapidly improve performance [4].

By employing Maximum Entropy classifier and the extended set of features, we can outperform the state of the art.

2 Related Work

Recent research in document-level sentiment analysis of English texts has shifted towards semi-supervised methods that exploit small labeled data or highly precise lexicons enriched with large unlabeled datasets [1].

Authors of [5] investigate learning word vectors that contain information about the word's semantics as well as sentiment. They use a graphical model for inferring word vectors and rely on words' global context. Document vectors are obtained as a vector product of words' vectors and document's bag-of-word vector (using TF-IDF weighing scheme). The SVM classifier is used do decide about sentiment of document.

Authors of [6] present a new probabilistic graphical model called JST (Joint Sentiment Topic) based on LDA, but extended with one latent node for word sentiment. The model trained by unsupervised manner performs poorly (accuracy about 60% on a binary classification task). By incorporating sentiment lexicons into the model, the accuracy increased to 82.8%.

Although the above-mentioned approaches give reasonable performance, authors of [7] conclude that it may be hard to beat the baseline word n-grams without additional linguistic processing, if a well-tuned classifier is used.

Sentiment analysis of Czech has gained attention very recently. Authors of [8] presents a pilot study on sentence-level sentiment analysis. They manually annotated 410 sentences from the news, generated a sentiment lexicon using this corpus, and reported only preliminary results using the Naive Bayes classifier.

An in-depth investigation of supervised methods for Czech sentiment analysis on three different domains was introduced in [3]. They tested various preprocessing methods, sets of features, and two classifiers (Maximum Entropy and SVM) on three large corpora (movie reviews, product reviews, and Facebook dataset)[4]. They achieved the best performance (F-measure 0.69 on Facebook data, 0.75 on product reviews, and 0.79 on movie reviews) using Maximum Entropy classifier and unigrams, bigrams, POS (Part-of-Speech) ratio, and emoticon features.

3 Semantic Spaces

The backbone principle of semantic spaces used in this paper is so-called Distributional Hypothesis, saying that *"a word is characterized by the company it keeps"* [9]. The word meaning is thus related to the context in which this word usually occurs, as confirmed in empirical tests carried out on human groups in [10]. In semantic spaces, words are represented as high-dimensional vectors. These vectors are usually derived from statistical properties of the words' contexts.

We briefly describe the most popular methods for building semantic spaces. All presented algorithms are available in an open source package *S-Space* [11].

[4] `http://liks.fav.zcu.cz/sentiment`

HAL — Hyperspace Analogue to Language [12] is a simple method for building a semantic space. HAL iterates over the corpus and records the co-occurring words (in some fixed window—typically 4 words) of each token, resulting into a co-occurrence matrix $\mathbb{M} = |W| \times |W|$, where $|W|$ is the vocabulary size. Finally, the row and column vectors of matrix $\mathbb{M}$ represent co-occurrence information of words appeared before and after, respectively.

COALS — Correlated Occurrence Analogue to Lexical Semantics [13] is an extension of HAL model as it starts with building a similar co-occurrence matrix $\mathbb{M}$. After this step, the raw counts are converted into the Pearson's correlations. Negative values are set to zero, other values are replaced by their square roots. The optional final step, inspired by LSA [14], reduces the dimension of matrix $\mathbb{M}$ using SVD (Singular Value Decomposition) and can also discover latent semantic relationship between words.

RI — Random Indexing [15] use a completely different approach from HAL and COALS. For each word in vocabulary, RI starts by creating random high-dimensional sparse vectors filled with few 1 and -1; the dimension is typically in order of thousands. Such vectors are very unlikely to overlap. Then the algorithm iterates over the corpus and for each token it sums up all the co-occurring words' vectors into the appropriate word vector in the final matrix $\mathbb{M}$.

RRI — Reflective Random Indexing [16] is an iterative extension of RI that focuses on modeling transitive relations between words. This approach is similar to the SVD reduction used in LSA and COALS, but it is less computationally expensive.

BEAGLE — Bound Encoding of the AggreGate Language Environment [17] shares some ideas with RI as it starts with generating high-dimensional vectors for each vocabulary word. The values are, however, drawn from a Gaussian distribution with 0 mean value and $1/D$ variance, where D is the vector dimension (in order of thousands). The final matrix $\mathbb{M}$ contains the co-occurrence information of vectors within a certain window (typically 5) as well as information about word order given by the convolution of n-gram vectors that contain the processed word.

4 Our model

Our method combines supervised machine learning, which employs a classifier on sentiment-labeled training data, and unsupervised feature space extension, which relies on clustering words represented in semantic spaces.

As a baseline, we adapt the supervised classification approach which has proven to be successful in the related work [3]. This approach requires a set of documents with their appropriate sentiment labels where each document is represented as a set of features. We rely on two kinds of binary features, namely the presence of word unigrams and bigrams. The feature space is pruned by filtering out unigrams and bigrams that occurred less than 10 times in the training data.

For classification we used the Maximum Entropy classifier [18] in the following form:

$$p(s|x) = \frac{1}{Z(x)} \prod_{i=1}^{n} e^{\lambda_i f_i(x,s)}, \tag{1}$$

where s is a sentiment label for a document, x is our knowledge about document, $f_i(x, s)$ is an i-th feature function, λ_i is corresponding weight and $Z(x)$ is a normalization factor. For estimating parameters of Maximum Entropy model we used limited memory BFGS (L-BFGS) method [19].

4.1 Word Clusters as Features

Since words in semantic space are represented as real-valued vectors, we can apply clustering methods. The main assumption is that words in the same cluster are semantically substitutable.

The selection of a suitable clustering algorithm is crucial for such a task. According to the study in [20] we selected Repeated Bisection algorithm because of its efficiency and acceptable computational requirements. We use the implementation from CLUTO software package.[5] As a similarity measure between two words we use cosine similarity of word vectors, calculated as the cosine of the angle between corresponding vectors: $S_{\cos}(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{a}\| \cdot \|\boldsymbol{b}\|}$.

We extend the baseline feature set (word unigram and bigram features) by adding binary features capturing the presence of cluster unigrams and bigrams in the document represented as a sequence of word clusters.[6] Again, we ignore word cluster unigrams and bigrams that occur less than 10 times in the corpus.

5 Datasets

We perform our experiments on two Czech datasets and one English dataset labeled with their sentiment.

The Czech datasets, provided by [3], contain a product review dataset (102,977 positive, 31,943 neutral, and 10,387 negative reviews) and a movie review dataset (30,897 positive, 30,768 neutral, and 29,716 negative reviews). We pre-process the data in the same way as in the original paper, namely using Ark-tweet tokenizer [21], Czech HPS stemmer,[7] and lowercasing.

The English movie review dataset, provided by [5], consists of 25,000 positive and 25,000 negative reviews for the training part, the same number of reviews is used for testing. There are additional 50,000 unlabeled reviews which are suitable for unsupervised methods.

6 Results

The experiments on the Czech corpora were performed using 10-fold cross-validation. The English corpus was already separated into training and test data. The vocabulary size $|V|$ for building semantic spaces and clustering was limited to the 10,000 most

[5] `http://www.cs.umn.edu/~karypis/cluto`

[6] According to our experiments, incorporating only the cluster features (without words unigrams and bigrams), in order to reduce the feature space, leads to worse performance than a baseline.

[7] `http://liks.fav.zcu.cz/HPS/`

frequent words (excluding stopwords). Semantic spaces were constructed using training data (Czech) or training+unlabeled data (English). We used default settings for all semantic spaces according to their original papers.

For each semantic space (HAL, COALS, RI, RRI, and BEAGLE) we conducted the experiments with seven different numbers of clusters (50, 100, 200, 500, 1,000, 2,000, and 5,000 clusters, respectively).

State of the art [3]					0.75
Number of clusters	Semantic space				
	COALS	HAL	RI	RRI	BEAGLE
50	0.72	0.72	0.71	0.70	0.75
100	0.75	0.77	0.75	0.70	0.71
200	0.75	0.73	0.73	0.74	0.74
500	0.74	0.72	0.71	0.70	0.69
1000	0.75	0.75	0.72	0.73	0.71
2000	0.73	0.73	0.73	0.73	0.73
5000	0.72	0.73	0.76	**0.78**	0.76

95% confidence interval = ± 0.002.

Table 1. F-measure results for the Czech product reviews dataset (best results are bold-faced).

cluster 1	**cluster 2**	**cluster 3**	**cluster 4**
pěkný (pretty)	složitější (more complex)	šmejd (junk)	silnější (stronger)
líbivý (pleasing)	komplikovaný (complicated)	spokojennost (satisfaction)	širší (wider)
moderní (modern)	nepřehledný (confusing)	geniální (ingenious)	užší (tighter)
nadčasový (timeless)	nepraktický (unpractical)	úchvatný (fascinating)	kratší (shorter)
milý (nice)	nešikovný (inept)	spokojenost (satisfaction)	menší (smaller)
atraktivní (attractive)	problematický (problematic)	vyhovující (suitable)	větší (bigger)
hezký (handsome)	zmatený (chaotic)		méně (less)

Table 2. Example of clusters obtained from the RRI model on the Czech product review dataset. To make this example easy to read, we transformed the stemmed words back into their nominative cases. Note that the Czech word *spokojennost* contains a typo (double *n*), however, is was clustered into the same cluster together with the correct word *spokojenost*.

Czech product reviews — The best result (F-measure: 0.78) was achieved using RRI and 5,000 clusters, as shown in Table 1. We assume that clustering into 5,000 clusters produces small and precise clusters, where the words tend to have the same meaning as well as the same sentiment. See Table 2 for examples.

Czech movie reviews — The best result (0.80) was obtained from COALS with a small number of clusters (50 and 100), as shown in Table 3. Also BEAGLE with 1,000 clusters performs well.

State of the art [3]					0.79
Number of clusters	Semantic space COALS	HAL	RI	RRI	BEAGLE
50	**0.80**	0.78	0.79	0.79	0.78
100	**0.80**	0.78	0.78	0.77	0.79
200	0.79	0.78	0.79	0.78	0.79
500	0.79	0.79	0.78	0.79	0.79
1000	0.78	0.77	0.78	0.78	**0.80**
2000	0.79	0.79	0.78	0.79	0.77
5000	0.79	0.79	0.79	0.79	0.79

95% confidence interval = ± 0.003.

Table 3. F-measure results for the Czech movie reviews dataset (best results are bold-faced).

English movie reviews — Table 4 shows that the best results (0.895) on the English movie review corpus were achieved using small clusters—50 for RRI and 100 for COALS. However, both semantic spaces yield similar results for larger clusters as well.

State of the art [5]					0.889
No. of clusters	Semantic space COALS	HAL	RI	RRI	BEAGLE
50	0.891	0.890	0.890	**0.895**	0.892
100	**0.895**	0.892	0.891	0.893	0.892
200	0.893	0.891	0.894	0.894	0.892
500	0.893	0.890	0.892	0.893	0.889
1000	0.894	0.889	0.889	0.893	0.889
2000	0.893	0.890	0.889	0.893	0.889
5000	0.893	0.893	0.890	0.894	0.892

95% confidence interval = ± 0.004,
90% confidence interval = ± 0.003,

Table 4. F-measure results for the English movie reviews dataset; best results are bold-faced. Note that in a two-class balanced-data scenario, F-measure equals accuracy.

In our experiments, RRI and COALS tend to perform consistently best for both languages. On the other hand, HAL and RI give no satisfactory results, regardless of the cluster size. Surprisingly, COALS with 100 clusters gives the best results in the movie review domain in both languages. On both Czech datasets, the current state of the art was significantly outperformed by our method (95% confidence interval). On the English dataset, our results are significantly better than the state of the art on 90% confidence interval.

Since RRI extends RI by modeling transitive relations between words, authors of RRI claim that RRI surpasses RI [16]. Similarly, the COALS model extends HAL (see section 3). We suppose that this is the reason why these two models perform best in

our task. We also think that the improvement on Czech movie review dataset using BEAGLE with 1000 clusters is caused only by chance, because BEAGLE performed consistently worst in all other experiments.

7 Conclusion and Future Work

This article presented a promising semi-supervised method for document-level sentiment analysis using semantic spaces and word cluster features. We outperformed the current state of the art on two datasets in Czech and one dataset in English. Our method benefits from its independence of any additional labeled data as it improves the existing methods in a fully-unsupervised manner. We also prove that the method is language independent. Furthermore, it can significantly help in morphologically rich languages.

In the future work, we plan to investigate a combined model that incorporates the best-performing semantic space for each cluster size.

Acknowledgments

This work was supported by grant no. SGS-2013-029 Advanced computing and information systems, by the European Regional Development Fund (ERDF) and by project "NTIS - New Technologies for Information Society", European Centre of Excellence, CZ.1.05/1.1.00/02.0090. Access to the MetaCentrum computing facilities provided under the program "Projects of Large Infrastructure for Research, Development, and Innovations" LM2010005, funded by the Ministry of Education, Youth, and Sports of the Czech Republic, is highly appreciated. The access to the CERIT-SC computing and storage facilities provided under the programme Center CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, reg. no. CZ. 1.05/3.2.00/08.0144 is acknowledged.

References

1. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Mining Text Data. Springer (2012) 415–463
2. Martineau, J., Finin, T.: Delta TFIDF: An improved feature space for sentiment analysis. In: Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA, The AAAI Press (2009)
3. Habernal, I., Ptáček, T., Steinberger, J.: Sentiment analysis in czech social media using supervised machine learning. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Atlanta, Georgia, Association for Computational Linguistics (June 2013) 65–74
4. Brychcín, T., Konopík, M.: Semantic spaces for improving language modeling. Computer Speech and Language (2013) DOI 10.1016/j.csl.2013.05.001.
5. 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, Portland, Oregon, USA, Association for Computational Linguistics (June 2011) 142–150

6. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM conference on Information and knowledge management. CIKM '09, New York, NY, USA, ACM (2009) 375–384
7. Wang, S., Manning, C.D.: Baselines and bigrams: simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2. ACL '12, Stroudsburg, PA, USA, Association for Computational Linguistics (2012) 90–94
8. Veselovská, K., Hajič Jr., J., Šindlerová, J.: Creating annotated resources for polarity classification in Czech. In: Proceedings of KONVENS 2012, ÖGAI (September 2012) 296–304 PATHOS 2012 workshop.
9. Firth, J.R.: A Synopsis of Linguistic Theory, 1930-1955. Studies in Linguistic Analysis (1957) 1–32
10. Charles, W.G.: Contextual correlates of meaning. Applied Psycholinguistics **21**(04) (2000) 505–524
11. Jurgens, D., Stevens, K.: The s-space package: An open source package for word space models. In System Papers of the Association of Computational Linguistics (2010)
12. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods Instruments and Computers **28**(2) (1996) 203–208
13. Rohde, D.L.T., Gonnerman, L.M., Plaut, D.C.: An improved method for deriving word meaning from lexical co-occurrence. Cognitive Psychology **7** (2004) 573–605
14. Landauer, T.K., Foltz, P., Laham, D.: An Introduction to Latent Semantic Analysis. Discourse Processes (25) (1998) 259–284
15. Sahlgren, M.: An Introduction to Random Indexing. Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005 (2005)
16. Cohen, T., Schvaneveldt, R., Widdows, D.: Reflective random indexing and indirect inference: a scalable method for discovery of implicit connections. Journal of Biomedical Informatics **43**(2) (2010) 240–56
17. Jones, M.N., Mewhort, D.J.K.: Representing word meaning and order information in a composite holographic lexicon. Psychological Review **114** (2007) 1–37
18. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Computational Linguistics **22** (March 1996) 39–71
19. Nocedal, J.: Updating Quasi-Newton Matrices with Limited Storage. Mathematics of Computation **35**(151) (1980) 773–782
20. Zhao, Y., Karypis, G.: Criterion functions for document clustering: Experiments and analysis. Technical report, Department of Computer Science, University of Minnesota, Minneapolis (2002)
21. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for twitter: annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2. HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 42–47