Aspect-Based Sentiment Analysis of Russian Hotel Reviews

Valery Rybakov[0000-0003-4912-4816] and Alexey Malafeev[0000-0002-8962-7496]

National Research University Higher School of Economics
Nizhny Novgorod, Russia
valera210597@gmail.com; aumalafeev@hse.ru

Abstract. The paper presents an attempt to solve the task of aspect-based sentiment analysis in the domain of Russian-language hotel reviews, using distributed representation of words. The authors follow an approach similar to [Blinov, Kotelnikov, 2014], but applied to a different domain and using different parameters. The authors also present a new dataset that is made available to the community. To build the vector space of words with word2vec, a corpus comprising 50 329 hotel reviews was constructed. The next step was the compilation of aspect and sentiment lexicons in the vector space obtained. The lexicon construction approach was based on iteratively expanding a small set of initially specified terms. Finally, the sentiment of aspects in actual reviews was calculated given the aspect and sentiment terms found in the text and their weights, i.e. cosine similarity to the initial terms. The model was tested on a corpus of 6 876 texts from the same domain.

Keywords: Aspect-Based Sentiment Analysis, Distributed Representation of Words, Natural Language Processing, Machine Learning.

1 Introduction

Today, the opportunities of the Internet allow anyone to express their own opinion on any topic and in relation to any objects. This opinion can be presented in the form of a user reviews, usually of an informal style. The sentiment extracted from these reviews is of interest both for the potential customer who wants to purchase the best product on the market, and for enterprises engaged in the analysis of consumer preferences. The need for automatic sentiment extraction from texts has made widespread use of such a field of computer science and natural language processing as Sentiment Analysis.

The first attempts at opinion extraction were primarily focused on the document or sentence level. Now the sentiment analysis problem requires more complex consideration of opinion moving from the text and sentence to phrase level. Here the sentiment analysis problem boils down to the search for the author's attitude to certain aspects of the object, for example, the aspects food, service and price can be distinguished for the object restaurant. Thus, the sentiment expressed in the text is subjected to more detailed study, as it is considered at the level of significant aspects.

The task of aspect-based sentiment analysis [Liu, 2012; Pontiki et al., 2014; Pavlopoulos, 2014] is usually split into two subtasks: aspect terms extraction and aspect terms polarity estimation, which are concerned separately and often use different techniques.

A lot of research has been conducted in the field of aspect-based sentiment analysis. Traditional approaches are based on collecting the most frequent words and phrases which are contained in the manually constructed aspect or sentiment lexicon. State-of-the-art models make use of topic modeling methods, such as Latent Dirichlet Allocation (LDA), and Conditional Random Fields (CRF).

The model described in this paper is built using the distributed representation approach which serves as a tool for topic modeling. The topic itself is represented in the form of term lists (words and collocations), where all terms are semantically related to one another.

2 Related Work

Several evaluation initiatives have been undertaken to help promote the task of aspect-based sentiment analysis [Pontiki et al., 2014; Loukashevich et al., 2015], which is a very important step to solving it. There exist different methods for solving different aspect-based sentiment analysis subtasks [Liu, 2012; Pavlopoulos, 2014; Pontiki et al., 2014; Zhang and Liu, 2014].

Liu [2012] lists four main approaches to aspect extraction:

1. Using frequent nouns and noun phrases.
2. Using opinion and target relations.
3. Supervised learning.
4. Topic modeling.

The frequency-based approach was used in a number of studies: [Hu and Liu, 2004; Ku et al., 2006; Blair-Goldensohn et al., 2008]. The relation extraction approach (via a dependency parser) was notably used in [Zhuang et al., 2006], among others. As far as supervised learning is concerned, two main sequential labeling techniques dominate the task of aspect extraction: Hidden Markov Models [Rabiner, 1989] and Conditional Random Fields [Lafferty et al., 2001]. To give some prominent examples of applying these techniques to aspect term extraction, the first was used in [Jin et al., 2009], and the second in [Jakob and Gurevych, 2010]. Examples of using topic modeling for aspect extraction are [Mei et al., 2007; Titov and McDonald, 2008].

Another important note is that many methods often benefit from taking advantage of more data, i.e. additional reviews, even without annotated terms. This was well demonstrated by top performers in the SemEval-2014 aspect-based sentiment analysis task [Pontiki et al., 2014].

The model described in the paper is largely based on the approach presented in [Blinov and Kotelnikov, 2014]. In this work the authors suggest techniques for constructing the aspect and sentiment lexicons leveraging the distributed representation of words. For the vector space construction, the tool word2vec [Goldberg and Levy, 2014] was used. The parameters were the following: number of dimensions – 150, the size of a context window – 5 words, the minimal word frequency - 5. As the training data, they used a corpus of 47,301 reviews in the restaurant domain. The aspects Food, Interior and Service were selected. The values of F1-measure for each aspect are 0.664 (Food), 0.617 (Interior), and 0.667 (Service).

3 Dataset

At the moment there is no publicly available text corpus of Russian hotel reviews, marked for the sentiment of aspects. Thus, a new corpus of hotel reviews was assembled; the reviews were collected from the website TripAdvisor.com. To do this, an algorithm of site parsing was developed with the Python programming language using the BeautifulSoup framework.

The following information was collected from the site: the text of the review, the overall rating of the hotel (on a 5-point scale), an assessment of the hotel's characteristics, such as the price-quality ratio, location, room, cleanliness, service, quality of sleep. The site's infrastructure allows reviewers to choose from the proposed hotel characteristics only those he or she wants to evaluate or not rate any of them at all. For the sentiment identification stage of the algorithm, only three aspects were chosen: Room, Location and Service, since they are the most popular ones. The corpus includes reviews of hotels located in Barcelona, Berlin, Moscow, Istanbul, Phuket, and Helsinki. This choice was based on the ranking of countries visited by Russian tourists within 9 months in 2017, compiled by the agency TurStat, and also considering the need to reflect the culture and life diversity of selected regions to broaden the lexicon used by the algorithm. A snippet from the corpus is shown in Fig. 1

Общая оц	Цена/каче	Располож	Качество (	Номера	Чистота	Обслуживание	
4	4	4	5	4	4	5	Очень достойный отель с прекрасными номерами, хорошим интернетом, прекрасной
4		3			4	4	Остановились в Барселоне проездом, т.к. нужно было посетить консульство РФ. Распо
4	3	3	5	5	3	4	Типичная сетевая гостиница. Главный плюс-шикарные завтраки. Из минусов, это, коне
1				2	1	1	Начнем с того, что в этом отеле не берут деньги только за воздух. Звонок с телефона в
5							Отель находится в отдалении от центра,но пешком дойти можно.На входе имеется сто
5							Приехали с сестрой и её мужем на машине и с собакой. Парковка для постояльцев в п
5	5				5	5	Чистота и удобство в номерах – по высшему классу. Единственное, на что стоит пожал
5	4	4	5	5	5	5	В сети отелей NH Collection отдыхаем впервые. Все очень понравилось! Современный
5	5	4	5	5	5	5	У отеля неплохое расположение, рядом есть различные магазины. Обслуживающий пе
3			3		4	4	отель соответствует заявленным звездам.номера чистые и просторные,с явно прослеж

Fig. 1. The training corpus snippet.

In total, 50 329 reviews were collected for the training corpus.

The distribution of the training corpus reviews by aspects and sentiment marks is presented in Table. 1. Since users have the opportunity to assess aspects selectively, the table includes a column that stores the number of reviews that have not got a mark for a particular aspect. As can be seen, the share of such reviews is quite high. But since this markup is not taken into account when constructing the vector space, a significant number of unmarked reviews does not affect the quality of the algorithm.

Thus, the presented table with rating distribution only gives an approximate description of the corpus and is not used in training. The corpus was also not balanced for the number of positive and negative reviews, since that uneven distribution reflects the actual situation of the users' attitude to the hotel services, according to TripAdvisor.com.

Aspect	Mark						Total
	5	4	3	2	1	Not Marked	
Room	10155	6735	2708	753	358	29602	50329
Location	14018	4762	1689	368	163	29329	
Service	20487	9067	2960	885	719	16211	

Tab. 1. The distribution of the training corpus reviews by aspects and sentiment marks.

Additionally, based on the same algorithm, a test corpus was compiled comprising reviews about hotels in St. Petersburg (3650), Dubai (753) and Paris (2473). For the test corpus, only those texts were collected that contain a sentiment markup for the three studied aspects: Room, Location, and Service. Table 2 shows the number of test corpus reviews distributed by marks.

Aspect	Mark						Total
	5	4	3	2	1	Not Marked	
Rom	3199	2205	1025	287	160	0	6876
Location	4533	1555	611	134	43	0	
Service	4031	1821	656	218	150	0	

Tab. 2. The distribution of the test corpus reviews by aspects and sentiment marks.

The proposed training and test corpuses are publicly available at `https://goo.gl/DTEpxs`

4 System Description

4.1 Normalization

Before entering the program, all reviews in the training corpus are pre-processed. The review marks are deleted, the texts are lemmatized (mystem is used) and segmented by sentences. Each segment is tokenized, the punctuation marks are deleted.

Also, the negation problem is dealt with at this stage. It is important that in the text, the word to which the particle *не* (not) belongs gets the opposite meaning, so it becomes necessary to designate the given word differently. Due to the fact that it is rather difficult to automatically determine which word the particle belongs to, it was decided to add the prefix *not_* to the first adjective, adverb or verb following the particle, and thus to regard the construction *not* + *word* as a separate term. The part-of-speech identification was carried out using the library pymorphy2. The collocations with the adverb *очень* (very) were processed in the same way. The normalization stage also includes the removal of stop words.

4.2 Terms Extraction

To extract aspect and sentiment terms from the training corpus, the method of vector representation of words was used. For this purpose, the tool word2vec with skip-gram model was applied using the Gensim library for Python.

All texts from the training set (50329) were used to construct a vector space of words with dimension 300. The context window size of 7 words was chosen. The words whose frequency is less than 5 in the corpus were not selected for training.

The method of extracting aspect and sentiment terms consists in automatically expanding a predefined set of five terms for each aspect. For the aspect Room the initial terms *номер* (room), ванная (bathroom), *телевизор* (TV), *свет* (light), *кровать* (bed) are selected. For the aspect Service the initial terms are *сервис* (service), *персонал* (staff), *администратор* (administrator), *сотрудник* (staff member), *консьерж* (concierge). For the aspect Location the words *местоположение* (location), *достопримечательность* (attraction), *центр* (center), *транспорт* (transport), *месторасположение* (location) were chosen. In such a set, for each term other terms close to it were sought, using the vector representation of words. To find the distance between the vectors, the cosine similarity measure was used.

Thus, for each term a list of 10 new terms closest to the original one was found. These lists were combined, with duplicate terms removed. This process continues and the resulting list again generates a new one according to the same principle. Repeating this procedure for new term lists is an iterative process that generates aspect terms.

To remove noise words which appear during term generation, an additional restriction was used: each newly generated term was stored in the resulting list of aspect terms only if the similarity value with at least three the five terms in the initial list exceeded 0.3 for each aspect. For each term, the cosine similarity with initial terms is calculated and the maximum is assigned to it as the weight. The weight value will be used at the sentiment assignment step.

As a result, each of the three aspects has its own list of terms. The number of terms for each aspect is the following: 2550 for Room, 1317 for Location, and 1740 for Service. The t-SNE algorithm allows one to visualize word vectors. Figure 2 shows the vectors of the three aspects of the first 300 generated terms. It can be seen from the graph that, in accordance with the three aspects, three separate clusters are distinguished.

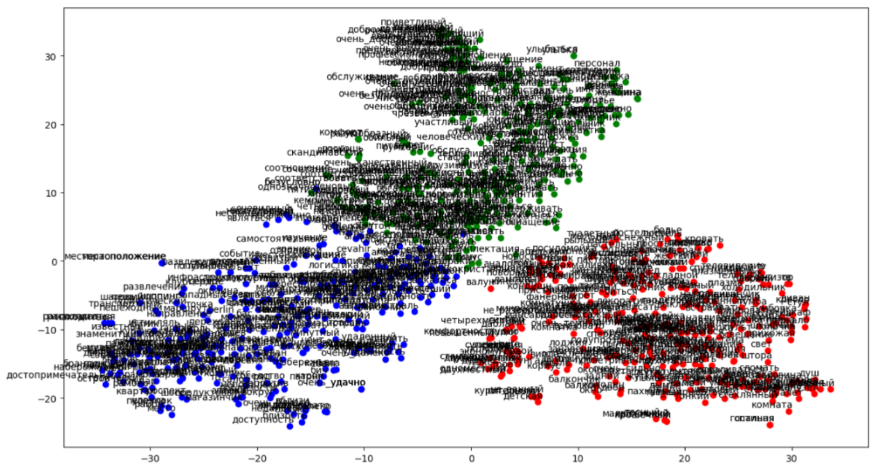

Fig. 2. Visualization of the aspect vectors.

In the same way, sentiment terms were obtained. As the initial terms that set the overall sentiment, the words *отличный* (excellent) for the positive class and *ужасный* (terrible) for the negative class were chosen. For each newly generated term, the cosine similarity value with the initial term was found and was assigned to the term as the weight. As a result, for the positive sentiment, 342 terms were found (with the threshold of 0.2) and 1203 terms for the negative sentiment (with the threshold of 0.25).

Similarly to aspect terms, the sentiment term vectors can be visualized using the t-SNE method. Figure 3 illustrates the distribution of the vectors of the 300 most positive and most negative terms.

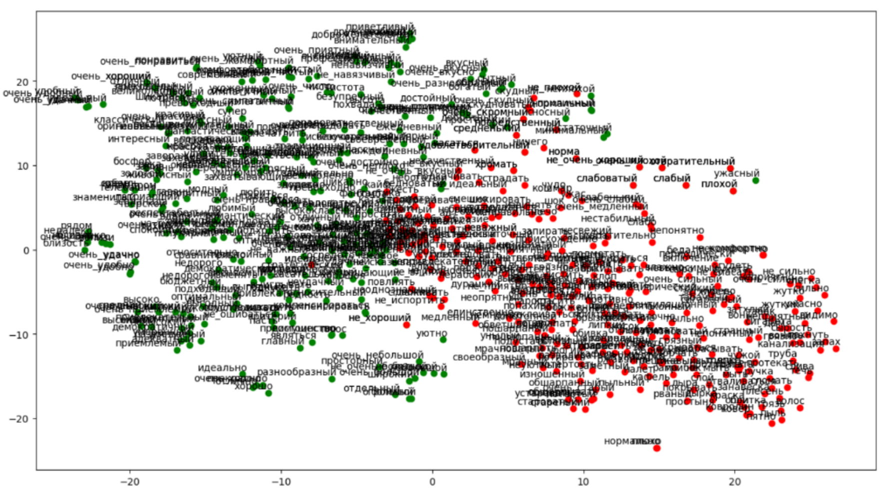

Fig. 3. Visualization of the sentiment vectors.

4.3 The Aspect Score Calculation

The final stage of the system consists in assigning each aspect a sentiment value (positive or negative). The input text of the review is segmented by the following punctuation marks: {*? ! , . : ;*}.

For each segment, the aspect and sentiment terms and their weights are identified from the corresponding lists prepared at the previous stages. Then, the weights of the sentiment terms from the current, previous and subsequent segments are added together. As a result, the final sentiment value for a given aspect is equal to the product of its weight and the sum of the sentiment term weights from the three segments. The value of the aspect's sentiment can take either a positive or a negative value depending on the sentiment class (positive or negative). To determine the sentiment of the aspect for the whole review, the sum of the sentiment values of all aspect terms in the review is calculated.

5 Results and Discussion

The program was tested based on the common metrics for assessing the quality of classifiers: precision, recall, F-score, and classification accuracy (micro).

Initially, the assessments were presented on a five-point scale. The conversion to the binary scale was performed according to the following scheme: {1, 2} → negative, {4, 5} → positive. Reviews that have a score of 3 on an aspect were not considered for this aspect when assessing the quality of the algorithm.

The number of correct and incorrect decisions of the algorithm, as well as the precision, recall and F1-measure metrics for each aspect are presented below.

Room

Category		Actual class	
		Positive	Negative
Predicted class	Positive	3228	60
	Negative	2176	387

Location

Category		Actual class	
		Positive	Negative
Predicted class	Positive	4787	119
	Negative	1301	58

Service

Category	Actual class

		Positive	Negative
Predicted class	Positive	4090	145
	Negative	1762	223

Performance

	F "+"	**F "-"**	**F mean**	**Accuracy**
Room	0.743	0.257	0.5	0.618
Location	0.871	0.076	0.473	0.773
Service	0.811	0.19	0.501	0.693

In the last table, we show the F-scores for both the positive and the negative classes in all three aspects, as well as the mean F-score and classification accuracy (micro). It can be seen that the model fails on the negative class in all three aspects, but since the negative classes are much smaller than the positive ones, classification accuracy ranges from 0.618 to 0.773. Unfortunately, we cannot directly compare these with any results obtained by other researchers, since, as far as we know, no performance scores are reported in the literature for the same task, domain and language. In [Blinov and Kotelnikov, 2014], whose approach was the basis for ours, the task and language are the same, but not the domain (restaurants, not hotels); as already mentioned in the Related Work section, the F-scores for each aspect were 0.664 (Food), 0.617 (Interior), and 0.667 (Service), considerably higher than ours.

6 Conclusion and Future Work

In this paper, an aspect-based sentiment analysis system was described, which employs the distributed representation of word vectors to compile the aspect and sentiment lexicons. A training set amounting to 50,379 reviews in the hotel domain was compiled and marked according to the sentiment of the selected aspects (Room, Location and Service). A test set of 6876 reviews was also compiled. Our code and the dataset are made available to the expert community (see the link at the end of Section 3). Based the constructed training corpus, sentiment and aspect terms were obtained; these can also be used for further research in the field.

The developed algorithm shows average classification performance. The F-score and accuracy values for the aspects are: 0.473, 0.773 (Location); 0.501, 0.693 (Service); 0.5, 0.618 (Room). We believe that, despite the not so impressive results, the present paper is still a contribution in the field. We have compiled and made available for the community a sufficiently large dataset for the aspect-based sentiment analysis task in the hotel review domain. Additionally, we report some first results that can be improved on in further research. Admittedly, our approach is not entirely novel, yet it differs from what has previously been done by other researchers in some aspects, such as the domain and the new dataset for this domain, parameter values and other specifics of our attempt at this task.

For future work, it is possible to experiment with other distributed representations for the purposes of solving the same task, such as doc2vec or fasttext, that might be helpful in identifying the aspect sentiment of entire reviews. Additionally, it might be interesting to adjust the parameters of word2vec for a better model. Lastly, systems based on support vector machines show excellent results in the field, so it might be beneficial to combine our approach with some of the proven machine learning methods for classification tasks.

References

1. Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G.A., and Reynar, J. (2008). Building a sentiment summarizer for local service reviews. In WWW Workshop on NLP in the Information Explosion Era,
2. Blinov, P. D., & Kotelnikov, E. V. (2014). Using distributed representations for aspect-based sentiment analysis. In Proceedings of International Conference Dialog (No. 13, p. 20).
3. Goldberg Y., Levy O. (2014) Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method //arXiv preprint arXiv:1402.3722.
4. Hu, M., and Liu, B. (2004). Mining opinion features in customer reviews. In AAAI, pp. 755–760.
5. Jakob, N., and Gurevych, I. (2010). Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, (Association for Computational Linguistics), pp. 1035–1045.
6. Jin, W., Ho, H.H., and Srihari, R.K. (2009). OpinionMiner: a novel machine learning system for web opinion mining and extraction. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (ACM), pp. 1195–1204.
7. Ku, L.-W., Liang, Y.-T., and Chen, H.-H. (2006). Opinion Extraction, Summarization and Tracking in News and Blog Corpora. In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs,
8. Lafferty, J., McCallum, A., and Pereira, F.C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
9. Liu, B. (2012). Sentiment Analysis and Opinion Mining (Morgan & Claypool Publishers)
10. Loukashevich, N.V., Blinov, P.D., Kotelnikov, E.V., Rubtsova, Y.V., Ivanov, V.V., and Tutubalina, E. (2015). SentiRuEval: Testing Object-Oriented Sentiment Analysis Systems in Russian. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2015,."
11. Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. (2007). Topic sentiment mixture: modeling facets and opinions in weblogs. In Proceedings of the 16th International Conference on World Wide Web, (ACM), pp. 171–180.
12. Pavlopoulos, I. (2014). Aspect based sentiment analysis. Athens University of Economics and Business.
13. Pontiki, M., Papageorgiou, H., Galanis, D., Androutsopoulos, I., Pavlopoulos, J., and Manandhar, S. (2014). Semeval-2014 task 4: Aspect based sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 27–35.

14. Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77, 257–286.
15. Titov, I., and McDonald, R.T. (2008). A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In ACL, (Citeseer), pp. 308–316.
16. Zhuang, L., Jing, F., and Zhu, X.-Y. (2006). Movie review mining and summarization. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management, (ACM), pp. 43–50.