Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2016”

Moscow, June 1–4, 2016

AUTOMATIC GENERATION OF THE DOMAIN-SPECIFIC SENTIMENT RUSSIAN DICTIONARIES

Dubatovka A. (alina.dubatovka@gmail.com)
Saint Petersburg State University, St. Petersburg, Russia

Kurochkin Yu. (yurakura@yandex-team.ru)
Yandex, St. Petersburg, Russia

Mikhailova E. (e.mikhaylova@spbu.ru)
Saint Petersburg State University, St. Petersburg, Russia

This paper presents an algorithm for generating the Domain-Specific Sentiment Russian dictionary using a graph model. It is important to emphasize that the described algorithm does not require any human-labeling, but just a sufficiently large corpus of Russian texts from the subject area, which can be generated automatically for most domains. Our algorithm is not strictly confined to the Russian language and, if necessary, can be generalized to develop dictionaries in other languages.
Dictionaries of positive and negative words are created using the analysis of the graph constructed on unlabeled corpus of the Domain-Specific Russian texts. The graph was built using the approach described in [6], pre-adapted to texts in Russian. The applicability of this method to create a graph for prediction of polarity of adjectives in reviews in Russian language is experimentally evaluated.
The original method of graph processing for splitting the vertex set of this graph into subsets of positive and negative words was proposed and implemented. The algorithm starts with gathering a small seed set of adjectives, polarity of which is unambiguous irrespective of a subject area (for example, “bad”, “good”, “terrible”, “excellent”).
Further, words are distributed iteratively: each time a vertex is added to the set, if the vertex is most strongly associated with the already existing vertices in the set. Several weighting functions on the edges were compared, as well as functions of attraction to the sets of positive and negative words with the aim of composing the most accurate dictionaries of positive and negative adjectives for a specific subject area.

Keywords: sentiment analysis, sentiment lexicon, opinion mining

1. Introduction

Public opinion and review studies have become an important part of decision-making processes. When a particular choice is associated with financial expenses (purchase of any goods, services, etc.), customers often rely on other people's experience. Information obtained from such studies is one of the most important factors in the final choice made.

Recently, the task of "opinion mining" has aroused significant interest among researchers of natural language texts as well as has become an important constituent of a decision-making process. Automatic sentiment analysis of a text finds a broad application in different fields of human activity—politics, business, film industry. Information obtained from review analysis affects the final choice that people make. Due to the broad variety of application areas, the actual goal is not just the review analysis itself, but also the opinion extraction from domain-specific texts. Since most of the algorithms for the review analysis are based on the use of sentiment dictionaries—dictionaries of positive and negative meaning of words, the automatic generation of such dictionaries becomes an important task. Solution of such kind of tasks strongly depends on the domain-specific area and language features, as the same words used in different situations can give diametrically opposite polarity. For example, the word "thin" is positive for laptop characteristics, while it would be uncomfortable to stay in a hotel with "thin walls". There are various situations when different meanings of one and the same word could be associated with alternative sentiments.

This paper describes an algorithm for generating the Russian sentiment dictionary for a domain-specific area using a graph model. We investigate the dependence of the dictionary's quality on graph building and analysis parameters. It is important to note that the described algorithm does not require any preliminary marking, but only a sufficiently large corpus of Russian texts from the domain area, which can be easily prepared by the user him-/herself. The algorithm is not bound to the Russian language and, if desired, can be generalized to create dictionaries of other languages.

Dictionaries of positive and negative words are created using the analysis of graphs constructed on a raw corpus of Russian texts from the domain area. Such corpus can be generated automatically for most domains. Graph nodes are adjectives, and edges connect the adjectives joined by a coordinating conjunction at least in one sentence.

The process of splitting graph nodes into positive and negative word subsets starts from small initial sets consisting of adjectives, the sentiment polarity of which is unambiguous irrespective of the subject area (*e.g*, "*bad*", "*good*", "*terrible*", "*excellent*"). Further, the remaining words are distributed iteratively: each time a node is added into the set most strongly associated with the already existing nodes. We compared efficiency of several weighting functions on the edges as well as distance functions to compile the most accurate dictionaries. This work has also demonstrated applicability of the approach described in [6] for sentiment analysis of adjectives in Russian-speaking reviews.

2. Related work

Over the last five years, there has been a tremendous increase in demand for sentiment analysis tools by companies for monitoring people's opinions on company's products and services and by social science researchers. All sentiment analysis tools rely on dictionaries of words and phrases with positive and negative connotations. Such dictionaries are necessary for different languages and different domain-specific areas affecting the polarity of words. Thus, the task of building sentiment dictionaries for a particular language and the domain-specific area is highly relevant, because even if there is a large amount of publicly available labeled data, most of such dictionaries are composed for the English language and examine either movie reviews or reviews on equipment.

For example, [1] describes an algorithm for compiling one of few publicly available Russian-language dictionaries of opinion words for the product meta-domain with the help of learning several classifiers on one domain and then migrating the resulting model to other domain areas. Further, in [7], authors try to clarify the dictionary obtained by analysis of the corresponding subgraph of the RuThes thesaurus.

The task of creating the sentiment vocabularies is relevant not only for Russian language but also for many other languages. Thus, in [10] authors propose a method for automatic dictionary creation for new languages (German, Russian, Italian, French, Arabic and Czech) using manually compiled dictionaries in English and Spanish.

In [2], the original manually compiled dictionary for the German language was enlarged by the construction of the graph based on the untagged German corpus, as described in [6], and by its further analysis using the classification method of maximum entropy.

Different graph models are widely used for subtasks such as adapting the model to a new domain area, highlighting sentiment sentences from the text or ranking words according to the opinion polarity. The authors of [11], having the corpus from marked up documents in one domain and unlabeled corpus from a different domain, determine the polarity by building and analyzing a weighted graph composed with the feedback as nodes and the cosine measure of similarity between documents as weight of the edge. In [8] it is proposed to build a graph using sentences and relationships between them. The problem of automatic detection of sentiment sentences from the text is solved by searching the minimum cut in the graph.

Since graph models play a key role in the social network analysis, various algorithms have been developed for their analysis. Some of them could be applied to the problem considered in this paper. For example, in [3, 4] the various algorithms of random walks (in particular, PageRank) are adapted to the graph constructed on the basis of eXtended WordNet [5] to rank the sentiment polarity of words.

3. Methodology

As shown in [6], coordinating conjunctions, connecting coordinate adjectives and adverbs convey the relation of the polarity of the connected words. As a rule, copulative conjunctions are placed between words that have the same polarity («*Tasty*

and healthy Breakfast"), and the adversative conjunctions are placed between words with nearly opposite sentiment ("*Cheap but nice hotel*").

These relations between the word sentiments allow to build a weighted graph whose nodes are the adjectives and edges are connections between them, labeled with the number of sentences with the words connected either by the copulative or adversative conjunction.

Analysis of the obtained graph permits to evaluate the "positivity" or "negativity" of words, which are its nodes: the better the node is connected with other "positive" nodes and the worse with the "negative", the more positive it is.

Thus, the algorithm for constructing the sentiment dictionary consists of two main stages, described below: building the graph of connections between words and its processing.

3.1. Constructing the graph

As described above, we construct the graph using adjectives as nodes. The edges are copulative and adversative relations between them. To build such edges, we extract coordinate adjectives from the texts and consider connections between them. The adjectives are coordinate if they are consistent in their gender, number and case, and satisfy the template

$$(\text{ADV} \mid \text{NEG}) * \text{ADJ}(, ? (\text{AND} \mid \text{BUT})? (\text{ADV} \mid \text{NEG}) * \text{ADJ}) +,$$

where AND is the conjunction "and", BUT is one of adversative conjunctions ("but", "instead", "however", "nevertheless "), NEG is a negation, ADV is an adverb of measure and degree ("very", "quite", "too", "completely") and ADJ is an adjective.

The sentiment link was formed for each pair of the selected coordinate adjectives (either positive or negative depending on the conjunction). This link is necessary to calculate the weight of an edge. For example, for the phrase "*Tasty, plentiful but not very varied and expensive breakfast*" three positive links are produced: (*tasty, plentiful*), (*tasty, varied*), (*plentiful and varied*); and three negative links are: (*tasty, expensive*), (*plentiful, expensive*), (*varied, expensive*).

To determine the part of speech and word forms we use the morphological analyzer of the Russian language Mystem[1] [9]. Mystem works on the basis of the dictionary and normalizes words into the primary form, and also processes their grammatical information. For words missing in the dictionary, Mystem performs a hypothetical analysis of words.

In case the negation comes before the adjective, an orientation of connection between words reversed. For example, in the sentence "The pool is large, but not very deep", the adjectives "large" and "deep" will have a positive connection, meaning the sentiment coincidence, although they are connected by the adversative conjunction «but». Similarly, by processing the phrase "Delicious and not expensive food", "delicious" and "expensive" will have a negative connection and therefore obtain opposite polarity.

[1] https://tech.yandex.ru/mystem/

Since we analyzed the feedback of Internet users, not literary texts, it is necessary to consider not only the punctuation rules of the Russian language, but also the most common (though erroneous) forms. So, for example, people sometimes miss a comma, even if grammar rules require to use it: a comma between coordinate adjectives, a comma before "but" or between repeated conjunctions "and", so the template should not be very strict.

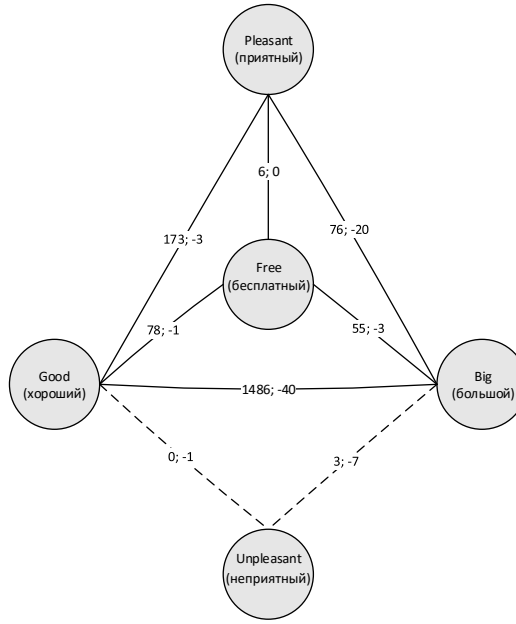


Fig. 1. A fragment of the graph without removal the "un-" prefixes

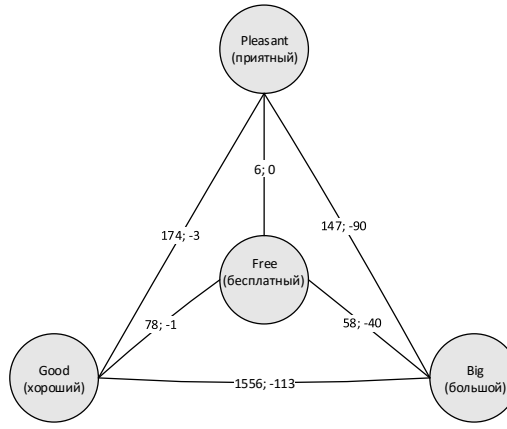


Fig. 2. A fragment of the graph with removal the "un-" prefixes

3.1.1. Particle “not” and the prefix “un-”

Apart from the fact that negation can be expressed by a free-standing particle “not”, it can appear in the prefix form “*un-*” with adjective. For example, “*not very pleasant*”, “*unpleasant*” and even “*not pleasant*”, by and large, represent the same negation of the adjective “*pleasant*”. So we analyzed how the negative prefix “*un-*” affects the sentiment adjective. For this purpose, we have implemented and compared two approaches. The first approach is to consider such adjectives (for example “*unpleasant*” and “*pleasant*”) as two different nodes of the graph. The second one is to separate the prefix “*un-*” if it is possible (i.e. the word without the prefix is identifiable by Mystem) and consider it as a negative particle “not”, that is, the adjective “*unpleasant*” is equated with the phrase “*not pleasant*” and, hence, we do not have to create a special node for the word “*unpleasant*”. Fig. 1 and 2 present fragments of the graphs built without and with removal the “un-“ prefix respectively.

3.2. Graph processing

The next stage is to split the previously obtained graph into two clusters: the positive set and the negative one. To do this, we initialize the positive and the negative sets and iteratively add one by one the non-assigned nodes to each of these sets. The candidate node is added into the nearest set (“positive” or “negative”). The candidates are the nodes with at least one edge connected with these sets. After adding a node into the set, all its neighbors that have not yet been assigned are added to the set of candidates and the distances between the candidate node and the final sets are recalculated.

3.3. Initialization

It is possible to initialize the positive and negative sets according to the fact that the sentiment polarity of certain words is obvious regardless of the context and domain area.

Therefore, the initial set of “*positive*” words contains such apparently positive adjectives like “*good*”, “*excellent*”, “*wonderful*”, “*lovely*”, “*best*”, but “*negative*” words set consist of negative adjectives “*bad*”, “*awful*”, “*disgusting*”, “*worst*”, “*poor*”.

3.3.1. Weight of the graph edges

In the course of the graph construction, a pair of numbers is assigned to every edge—the number of positive and negative connections. The question is how to calculate the weight of edges using these numbers. The weight of edges can be calculated using these numbers as a basic difference, as arbitrary linear combination or a nonlinear function. In our experiments, we calculate the edge weight according to the formula

$$weight(word_1, word_2) = \#(word_1 AND\ word_2) - K * \#(word_1 BUT\ word_2),$$

where K is the coefficient of the adversative conjunctions relevance (the number of negative connections is much less than the number of positive connections, so we give greater weight to the first ones).

3.3.2. Distance to the final set

A similar problem arises—when calculating the distance between the candidate node and each of the final sets. Multiple edges of different weights can connect the node with the set. Is one "heavy" edge better than a lot of "light" edges? The node may have edges connected with the opposite set. We compared the following most common intuitive techniques for distance calculation.

3.3.3. The heaviest edge

The distance between the node and the set is the weight of the heaviest edge connecting each other.

3.3.4. The sum of the weights of edges

The edge weight is the sum of the edge weights connecting with the considered set, subtracted by the sum of weights of the edges connected with the opposite set.

4. Description of experiments

The algorithms' input was supplied with 259,023 depersonalized unlabeled hotel reviews. The size of the dataset was 660 Mb. The reviews were about different hotels over the world. As long as the texts were written by real users, they contained a lot of misspellings and grammatical errors and informal words. As a rule, these reviews described hotel location, rooms, staff, meal and beaches, but a lot of texts contained unrelated information concerning flight, excursions, places of interest *etc*.

In order to evaluate precision of the proposed algorithms on all the data available we have manually labeled all the adjectives extracted by the algorithms into three classes: positive, negative, and neutral. So we obtained "large" dictionaries of positive, negative, and neutral words consisting of 970, 1,000 and 2,591 words respectively. These dictionaries were used for the precision evaluation, because after processing by the algorithm each word resulted in being placed into one of the "large" dictionaries. In this case we calculated not only classical precision, but also a precision of separation positive words from negative ones. For this propose, we discarded all words contained in the "large" neutral dictionary from the result, since the detection of neutral words is actually a separate challenge [1, 8], and then calculated a classical precision. Table 1 contains sizes of "large" dictionaries as well as the resulting dictionaries for the algorithm with and without removing the "*un-*" prefixes.

Table 1. Sizes of the "large" dictionaries and result dictionaries

	Positive	Negative	Neutral	Total
Algorithm without removing the "un-" prefix	5,252	2,815	—	8,067
Algorithm after removing the "un-" prefix	4,936	2,695	—	7,631
"Large" dictionary	1,948	1,946	4,951	8,845

Because of the large amount of input data, a human assessment is impossible, so recall was estimated by using "manual" dictionaries. For this purpose, we manually labeled 500 random reviews from the input data. Every occurring adjective was labeled as positive or negative depending on its sentiment polarity in the review. Thus we compiled "manual" dictionaries of positive and negative words, consisting of 173 and 127 words respectively, for recall estimation. Since these 500 reviews were selected randomly, and adjective distribution over the reviews is considered to be uniform, we can assume the sample as unbiased, and therefore, the recall calculated for these 500 reviews is a good approximation for the recall of all data. To estimate recall for positive dictionary we calculated what part of words from "manual" positive dictionary occurs in the positive dictionary generated by the algorithm, the same was done for the negative dictionaries. Table 2 contains sizes of "manual" dictionaries and sizes of the intersections of "manual" and result dictionaries for both algorithms.

Table 2. Sizes of the "manual" dictionaries and "small" result dictionaries obtained as the intersection of the result dictionaries and corresponding "manual" dictionaries

	Positive dictionary	Negative dictionary	Total
"Manual" dictionary	173	127	300
Algorithm without "un-" prefix removing	164	74	238
Algorithm with "un-" prefix removing	163	83	246

To study the rate of dictionary degradation and relationship between the result quality and the stop point we built a plot of Precision@n, where Precision@n is a precision of the top n words from each dictionary.

In addition, to explore the dependence of the dictionary quality on the importance coefficient of negative edges K we built a plot of the F_1-measure for different values of parameter K.

5. Results

Tables 3 and 4 contain the results of the algorithms without and with removing the «un-» prefix respectively.

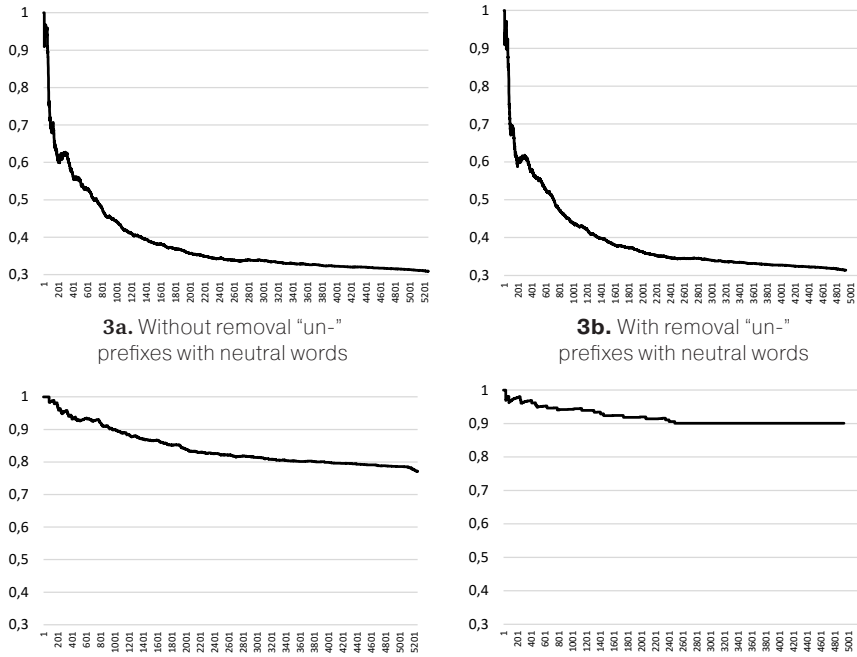
Table 3. The result of the algorithm without removing the "un-" prefix

Metric	Positive dictionary	Negative dictionary	Total dictionary
Recall	0.806	0.684	0.754
Precision	0.309	0.521	0.381
Precision without neutral words	0.770	0.827	0.796
F_1-measure	0.447	0.591	0.506
F_1-measure without neutral words	0.788	0.749	0.774

Table 4. The results of the algorithm after removing the "un-" prefix

Metric	Positive dictionary	Negative dictionary	Total dictionary
Recall	0.793	0.683	0.746
Precision	0.314	0.502	0.380
Precision without neutral words	0.779	0.820	0.799
F_1-measure	0.450	0.579	0.504
F_1-measure without neutral words	0.786	0.745	0.772

Fig. 3 and 4 present plots of Precision@n for positive and negative dictionaries, obtained as a result of processing all reviews, respectively. It is easy to see that the dictionaries start degrading very quickly due to the inclusion of neutral words, however, degradation of the filtered dictionaries, containing sentiment words only, is much slower. We should notice that removing the "un-" prefix does not affect the plot behavior much, if neutral words are taken into account, while for the filtered dictionaries it gives a significant increase in precision especially for the positive dictionary.



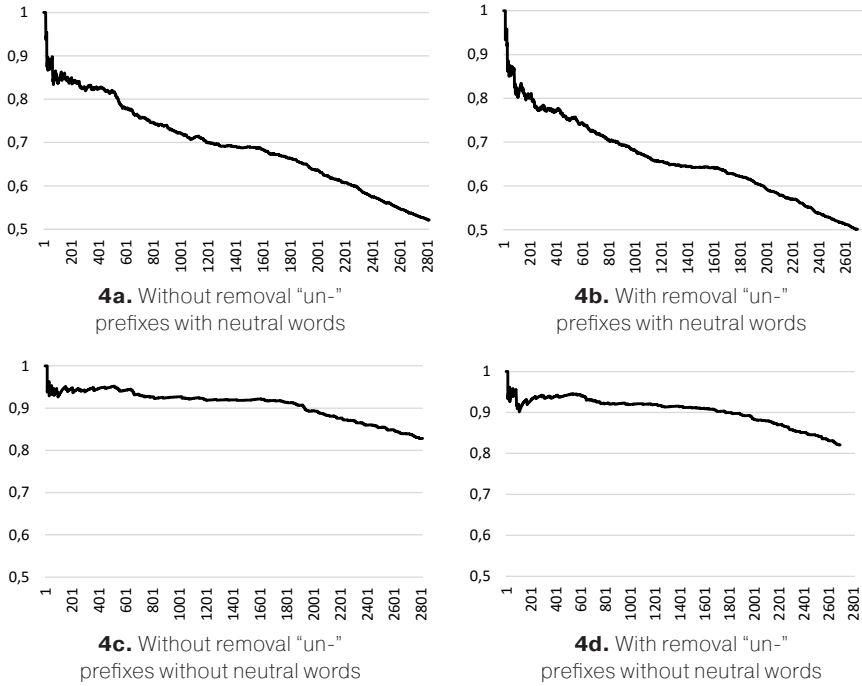
3a. Without removal "un-" prefixes with neutral words

3b. With removal "un-" prefixes with neutral words

3c. Without removal "un-" prefixes without neutral words

3d. With removal "un-" prefixes without neutral words

Fig. 3. Plot of Precision@n for the positive dictionary



4a. Without removal "un-" prefixes with neutral words

4b. With removal "un-" prefixes with neutral words

4c. Without removal "un-" prefixes without neutral words

4d. With removal "un-" prefixes without neutral words

Fig. 4. Plot of Precision@n for negative dictionary

Fig. 5 and 6 present the plots of dependence of F_1-measure with and without neutral words on the parameter K for positive and negative dictionaries. Fig. 7 and 8 contain a scatter plot of recall (the x-axis) and precision (y-axis) with different values of K from 1 to 10 (the points are marked with corresponding parameter values).

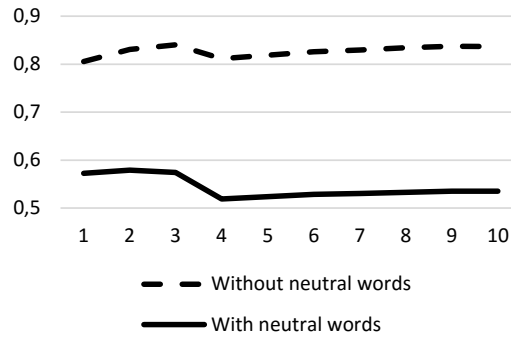


Fig. 5. The dependence of the F_1-measures on the parameter K without removal of "un-" prefix

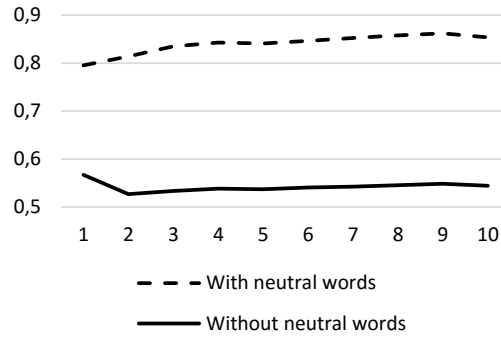


Fig. 6. The dependence of the F1-measures on the parameter K after removing the "un-" prefix

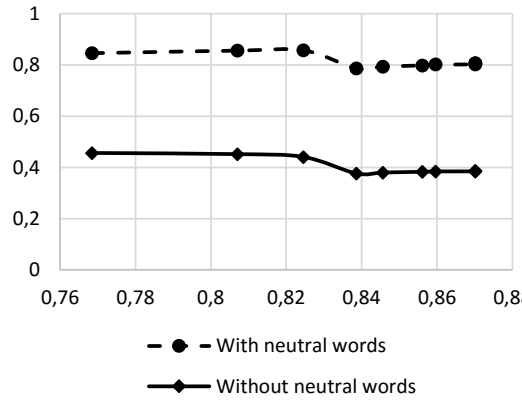


Fig. 7. The values of precision/recall for different values of the parameter K without removal of "un-" prefixes

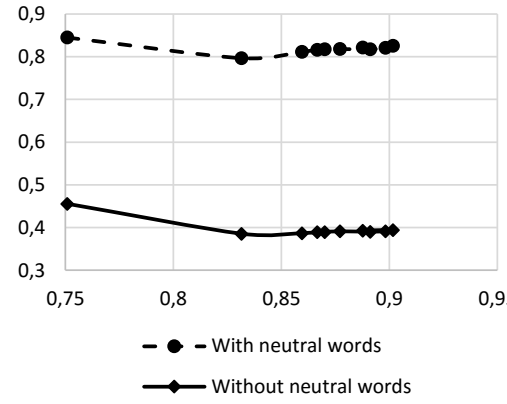


Fig. 8. The values of precision/recall for different values of parameter K after removing the "un-" prefixes

6. Conclusion

This work describes an unsupervised method for building dictionaries of sentiment adjectives based on the analysis of unlabeled reviews from chosen domain. For this propose, we consider and analyze a graph, built using adjectives from the text as the nodes and the syntactic relations between them as edges. To separate the adjectives into positive and negative sets, this graph is split into two clusters using the initial set of «universal» adjectives, whose sentiment polarity does not depend on the chosen domain or context. The paper provides a comparison of several implementations of algorithms for graph constructing and analyzing. These algorithms ensure the construction of dictionaries with 79.9% precision and 75.4% recall.

We considered hotel reviews in Russian language as data for our experiments. The described method is applied to unlabeled texts, and thus input corpus for the algorithm can be formed automatically (e.g., using a crawler), without human assessment. This allows to use this algorithm for an arbitrary domain. Furthermore, this approach can be applied to texts in other languages, as its implementation requires only a morphological analyzer, a list of copulative and adversative conjunctions, and initial set of “universal” sentiment words.

References

1. *Chetviorkin I. I. and Loukachevitch N. V.* (2012), Extraction of Russian sentiment lexicon for product meta-domain, Proceedings of COLING 2012: Technical Papers, Mumbai, pp. 593–610.
2. *Clematide S., Klenner M.* (2010), Evaluation and extension of a polarity lexicon for German, Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), Lisbon, pp. 7–13.
3. *Esuli A., Sebastiani F.* (2007), PageRanking wordnet synsets: An application to opinion mining, ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, pp. 424–431.
4. *Esuli A., Sebastiani F.* (2007), Random-walk models of term semantics: An application to opinion-related properties, Proceedings of LTC 2007, Poznań, pp. 221–225.
5. *Harabagiu S. M., Miller G. A., Moldovan D. I.* (1999), WordNet 2—a morphologically and semantically enhanced resource, Proceedings of SIGLEX99: Standardizing Lexical Resources, College Park, pp. 1–8.
6. *Hatzivassiloglou V., McKeown K.* (1997), Predicting the semantic orientation of adjectives, Proceeding EACL ‘97 Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Madrid, pp. 174–181.
7. *Loukachevitch N. V., Chetviorkin I. I.* (2014) Refinement of Russian sentiment lexicons using RuThes thesaurus), Selected Papers of XVI AllRussian Scientific Conference “Digital libraries: Advanced Methods and Technologies, Digital Collections”, Dubna, pp. 61–65.

8. *Pang B., Lee L.* (2004), A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, pp. 271–278.
9. *Segalovich I.* (2003), A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine, MLMTA, Las Vegas, pp. 273–280.
10. *Steinberger J., Ebrahim M., Ehrmann M., Hurriyetoglu A., Kabadjov M. A., Lenkova P., Steinberger R., Tanev H., Vazquez S., Zavarella V.* (2012), Creating sentiment dictionaries via triangulation, Decision Support Systems, Vol. 53(4), pp. 689–694.
11. *Wu Q., Tan S., Cheng X.* (2009), Graph ranking for sentiment transfer, ACL/IJCNLP (Short Papers), pp. 317–320.