Refinement of Russian Sentiment Lexicons Using RuThes Thesaurus

© N. V. Loukachevitch © I. I. Chetviorkin
Lomonosov Moscow State University
louk_nat@mail.ru ilia2010@yandex.ru

Abstract

The paper describes a combined approach to extraction of a domain-specific sentiment lexicon. At first, an initial version of a domain-specific lexicon is obtained by application of a supervised model. At the second stage, the ordered list of sentiment words is refined using the thesaurus information. This combined model is applied to several domains and at last the domain-specific sentiment lexicons are united to create an improved version of the Russian sentiment lexicon in the generalized domain of products.

1 Introduction

Automatic sentiment analysis of texts is a fast-developing technology in natural language processing. The task of automatic sentiment lexicon construction and improvement is a basic task for sentiment analysis of texts. There are no freely available sentiment lexicons for many languages or the quality of such lexicons is desired to be better. For example, in Russian only one automatically extracted sentiment lexicon has been published [1].

Besides, sentiment analysis of domain-specific texts requires adaptation of machine-learning models or sentiment lexicons to the target domain [6]. So, some sentiment words can loss their polarity in specific domains. For example, such word as *evil* in the movie domain usually refers to the movie plot, but not a user opinion.

Other words can obtain the sentiment polarity in a specific domain. For example, *word киношный* (adjective to Russian word *кино (movie))* can have the negative polarity with the meaning *"far from the real life"*. Another example - word *атмосферный (*adjective to word *атмосфера (atmosphere))* has the positive polarity in art-related domains denoting "creation of a special mood or feeling" (as *atmospheric* in English) – this is a relatively new sense of this word for Russian, not described in Russian dictionaries.

Proceedings of the 16th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" — RCDL-2014, Dubna, Russia, October 13–16, 2014.

Automatic extraction of sentiment words can be based on corpus-based or resource-based (dictionary, thesauri) approaches. In this paper we offer a combined approach to extracting sentiment lexicons. At first, an initial version of a domain-specific lexicon is obtained by application of a supervised model on the basis of statistical and linguistic features of sentiment words. This lexicon is presented as a list of words ordered by the decreased probability of their sentiment orientation. At this stage we obtain some sentiment words that are absent in dictionaries or having the domain-specific sentiment polarity. We extract sentiment-oriented words without any positive or negative labels because we consider this process as the first step to further polarity lexicon generation.

At the second stage, the ordered list of sentiment words is refined using the thesaurus information, in our case, newly published thesaurus of Russian language RuThes[1]. We trained a supervised model and tuned a combined model in the movie domain. Then this augmented model was utilized in four other domains. Finally, extracted sentiment lexicons from five domains are united to generate a high quality lexicon in the general product domain for Russian (ProductSentiRus+).

The reminder of this article is organized as follows. In Section 2 we review methods for generating sentiment lexicons. Section 3 briefly presents the structure of RuThes thesaurus, the Russian newly published thesaurus intended for natural language processing. Section 4 presents an approach for extracting sentiment words in various domains. Section 5 describes the refinement of the lexicon in the general product domain. To evaluate the quality of the obtained general resource extrinsically, we conduct the experiments on the tweet subjectivity classification task.

2 Related Work

There are two main approaches to sentiment lexicon extraction: corpus-based and dictionary-based methods.

Corpus-based methods utilize co-occurrence of words with each other [5, 9, 10], or appearance them in specific collocations or lexico-syntactic patterns [4]. Contemporary corpus-based approaches exploit a large

[1] labinform.ru/ruthes/index.htm

or hundreds of thousands of user reviews as in [6].

Dictionary-based methods utilize available electronic dictionaries and thesauri and usually begin their work from a set of seed words. In [3] SentiWordNet resource is described. It is the result of the automatic annotation of all the synsets of WordNet where each synset is associated to three numerical scores that indicate how positive, negative, and neutral the terms contained in the synset are. Different senses of the same word may thus have different opinion-related properties.

In [8] authors study semi-supervised approaches to label the polarity of words in a graph of lexical relations such as WordNet. They apply several methods: MinCut, Randomized MinCut, Label Propagation algorithm, described in [11], and show that Label Propagation algorithm achieves the best results. These and similar graph-based algorithms are also utilized in corpus-based approaches to sentiment lexicon extraction [4, 9].

In many studies domain-specific sentiment lexicons are created with corpus-based approaches using various types of propagation from a seed set of words, usually a general sentiment lexicon [6]. An important problem of such approaches is to determine an appropriate seed lexicon, which can depend on the domain.

In our study we create a domain-specific sentiment lexicon from medium-size datasets using multiple features of words and several collections without any co-occurrences between words. Then we improve an initial sentiment lexicon using sentiment labeling of the thesaurus concepts in a specific domain practically without pre-determined seed words. We use only two fixed seed opinionated words (*bad, good*), other potential sentiment words are obtained automatically from a ranked list of a sentiment lexicon (words ordered by the probability of their sentiment orientation) extracted from domain-specific collections.

3 RuThes Linguistic Ontology

In our study we use RuThes Thesaurus of Russian language. RuThes is a linguistic ontology for natural language processing, i.e. an ontology, where the majority of concepts are introduced on the basis of actual language expressions. For a long time RuThes has been manually developed within various NLP and information-retrieval projects, and now it is available for public use. The publicly available version of RuThes contains around 100 thousand Russian words and expressions [7].

If compared to WordNet-style resources RuThes is organized as a united semantic net where different parts of speech (nouns, verbs, adjectives) can be text entries of the same concepts. Each concept has a unique unambiguous name. Concepts can be connected with several types of conceptual relations. In addition, RuThes includes a lot of multiword expressions useful for applications and terms of so-called Sociopolitical domain – a broad domain of contemporary social relations, which includes terms from political, economic, military, sports and other fields [7].

Ambiguous words in RuThes are described similar to WordNet-style resources through attachment to several concepts. For example, in the current version of RuThes word *пресный* is attached to three concepts:

- *ПРЕСНАЯ ВОДА (fresh water)*;
- *ПРЕСНЫЙ, БЕЗВКУСНЫЙ (tasteless, bland in taste);*
- *ПРЕСНЫЙ (НЕИНТЕРЕСНЫЙ) (uninteresting).*

The first concept is neutral and not relevant to the movie domain. The second concept is negative but also irrelevant to the domain. Last concept is negative and relevant to the domain.

4 Extraction of Sentiment Lexicons

In this section an algorithm for extraction of sentiment words in a specific domain is described. The results of this algorithm are refined using the iterative procedure on the basis of RuThes thesaurus to obtain a high quality domain-specific sentiment lexicon.

Such a method is applied to four other domains without additional manual labeling and the results are combined in a sentiment lexicon in a generalized product domain ProductSentiRus+.

Table 1. Domain-specific collection statistics

Domain	Reviews	Descriptions
Movies	28, 773	17, 680
Books	23, 883	22, 321
Games	7, 928	1, 853
Digital Cameras	10, 208	920
Mobile Phones	30, 620	890

4.1 Extraction of domain-specific sentiment lexicon based on multiple features

At the first stage sentiment words are extracted with a corpus-based method utilizing a trained machine-learning model applied to several domain-specific text collections.

The first domain-specific collection (with high concentration of sentiment words) is a collection of user reviews in the domain (review collection) with numeric scores specified by their authors. In these experiments collections were gathered from the online services *imhonet.ru* and *market.yandex.ru* in five domains: movies, books, computer games, mobile phones and digital cameras. The second domain-specific collection (with low concentration of sentiment words) is a text collection of object descriptions (e.g. plots for movies). The overall collection statistics can be found in Table 1.

Another contrast corpus was a collection of two million news documents. Such a collection is useful for correct classification of general neutral words frequent in news.

Using such collections the feature representation is calculated for each word. The set of features includes the following feature types [1]:

- Frequency-based: collection frequency, document frequency, frequency of capitalized words, frequency of co-occurrence with polarity shifters (*no, not*), TFIDF;
- Score-based: deviation from the average score, word score variance, sentiment category likelihood for each (word, category) pair;
- Linguistic: Four binary features indicating the word part of speech, two binary features reflecting POS ambiguity, predefined list of prefixes of a word.

To train supervised machine learning algorithms, all words with the frequency greater than three in the movie review collection were labeled manually by two assessors. If there was a disagreement about the sentiment of a specific word, the collective judgment after discussion was used as the final ground truth. As a result of the assessment procedure the list of 4079 sentiment words was obtained. The best quality of classification using labeled data was shown by the ensemble of three classifiers: Logistic Regression, LogitBoost and Random Forest from WEKA programming package.

The result of this corpus-based method is a ranked list of domain-specific words ordered by the probability of their sentiment orientation – further *sentiment weights.* The algorithm boosts sentiment words to have high weights (to be closer to the beginning of the list) and neutral words to have low weights.

So in the movie domain in the list of more than 18 thousand words the following words are located in the first positions:

трогательны (affective), отстой (trash), фигня (crap), отвратительно (disgustingly), посредственный (satisfactory), предсказуемый (predictable), любимый (loved).

Word *атмосферный (atmospheric)* takes 830th, high-opinionated position in the list.

Evident sentiment adjectives of the movie domain *пресный* and *безвкусный (*both are translated into English as *tasteless*) take even higher opinionated positions: 139th and 193th . But their noun derivations *пресность, безвкусие, безвкусность,* and *безвкусица* are less successful. *Пресность, безвкусие, безвкусность,* are absent from the list because of low frequency; *безвкусица* takes 1515th place in the list. So thesaurus-based improvements may be quite possible.

The obtained model was applied to four other domains (books, games, digital cameras, mobile phones) without any additional manual efforts. The quality of extracted sentiment lexicons was measured using precision measures and presented in the Baseline columns of Table 2.

4.2 Refinement of domain-sentiment sentiment lexicons using RuThes thesaurus

To increase the quality of extracted sentiment lexicons we refine them with general thesaurus for Russian language RuThes [7]. The input of the refinement algorithm is a ranked sentiment list obtained with the model described in the previous subsection; however, a similar input can be also generated with other methods.

Table 2. Precision of the domain-specific lexicons at levels 100 and 1000 first words in the sentiment lists

.Domain	Baseline P@100, %	+RuThes P@100, %	Baseline P@1000, %	+RuThes P@1000, %
Movie	99	**100**	81.5	**85.5**
Books	99	**100**	86.0	**86.2**
Games	97	**100**	72.2	**73.1**
Digital Cameras	85	**92**	65.8	**66.3**
Mobile Phones	85	**97**	73.2	**78.6**
General Product Domain	**100**	**100**	90.5	**95.2**

Words from the ranked sentiment list are quite different relative to RuThes descriptions. Some words are not described in RuThes, e.g. three of the most probable sentiment words in the movie domain are absent in RuThes, others are mentioned in text collections exactly in the same senses as described in RuThes, the thirds (e.g. *atmospheric*) are described in RuThes but have an additional (or the other) sentiment polarity. So we should try to correct the word order in the sentiment list carefully applying RuThes descriptions.

The main idea of the lexicon refinement is to label conceptual subgraphs of the thesaurus network as sentiment or neutral and use this labeling to reorder the initial sentiment list. This process in contrast to such a method as Label Propagation [8, 11] should be also regulated with previously obtained sentiment weights of words.

Let us denote a domain-specific lexicon with W_D where all words are ordered by their sentiment weights (*sw*). Initially the algorithm forms two sets of thesaurus concepts using words from the both sides of the list W_D: L_s – concepts supposed to be opinionated, L_n – neutral concepts. With this aim the initial average sentiment weights *csw* for all concepts containing words from W_D are calculated. Then the algorithm adds to L_s concepts with the high average weight ($csw_s > 0.85$) and also two pre-defined concepts, corresponding to senses of words *bad* and *good*.

Concepts with the low average weight ($csw_n < 0.05$) are added to the set of neutral concepts L_n, which formed without any pre-defined concepts. The thresholds for csw_s and csw_n are obtained from experiments.

Further, every set (L_s and L_n) is iteratively augmented with concepts using two conditions: the average sentiment weight threshold and the number of direct thesaurus relations to the existing sets. Formally, L_s and L_n are calculated as shown in Algorithm 1 listing. The algorithm uses also the following additional notation:

– *Adj (L)* is a set of direct-link neighbor concepts to set of concepts *L*;

– *nlink (C, L)* is a function returning the number of direct thesaurus relations between concept *C* and set *L*.

In the last step *sw* weights of all words corresponding to L_s concepts are modified by multiplying them by factor k_1 ($k_1 > 1$) and all words corresponding to L_n are multiplied by factor k_2 ($0 < k_2 < 1$). The resulting list is reordered by weight.

Low-frequent words (with the frequency less than 3) of the source domain collection are absent in the initial ranked sentiment list and therefore do not have any sentiment weights. The initial sentiment weights of such words are calculated as the average sentiment weights of concepts they related to. The weights of these concepts, in turn, are calculated from other, more frequent synonyms or from average weights of neighbor concepts in the labeling process.

Algorithm 1. Weights+Relations

```
Input: concept list with sentiment
weights csw
Output: Ls,  Ln
  Ls = {Cbad, Cgood}∪Chigh, Chigh={Ci:
csw(Ci)>0.85},
  Ln = Ln∪Clow, Clow = {Ci: csw(Ci)<0.05}
  θ = 0.1, Nlink = 3, Ls_iter = Ls, Ln_iter
= Ln
while θ<0.6
    for C ∈ Adj(Ls)
      if nlink(C, Ls) > Nlink &&
csw(C)>0.7-θ
        then Include(C,Ls);
  end

    for C ∈ Adj(Ln)
      if nlink(C, Ln) > Nlink && csw(C)<
θ
        then Include(C,Ln);
    end

    if Ln == Ln_iter && Ls == Ls_iter
        then Nlink = Nlink-1;
  if Nlink == 0
        then θ = θ + 0.05, Nlink = 3;
  Ln = Ln_iter, Ls = Ls_iter
end
```

All parameters of the algorithm are tuned in the movie domain and then applied to four other domains. The quality of domain specific sentiment word lists can be found in Table 2 in RuThes column.

After application of this algorithm in the movie domain our example words *пресный, пресность, безвкусный, пресность, безвкусица* have the following places in the generated sentiment list: *пресный* – 81, *пресность* – 86, *безвкусный* – 115, *безвкусность* – 172, *безвкусие* – 173, *безвкусица* – 943.

The words related to the neutral sense of word *пресный – ПРЕСНАЯ ВОДА (fresh water)* preserved their very low positions in the sentiment list: *вода (water)* – 23059, *айсберг (iceberg)* – 26124.

5 Improvement of General Sentiment Lexicon Using RuThes Thesaurus

Integrating sentiment lexicons from various product-oriented domains it is possible to create a general sentiment lexicon in the broad domain of products and services. Such a lexicon for Russian was described in [1], it was called ProductSentiRus[2].

In that paper the lexicons of five domains were summed up using a formula intended to boost words that occur in many different domains and have high weights in each of them.

Thus, for combining multiple weighted word lists the following formula was used:

$$R(w) = \max_{d \in D}(prob_d(w)) \sum_{d \in D} \frac{1}{|D|}\left(1 - \frac{pos_d(w)}{|d|}\right),$$

where D – is the domain set with five domains, d is the sentiment word list for a particular domain and $|d|$ is the total number of words in this list. Functions $prob_d(w)$ and $pos_d(w)$ are the sentiment probability and position of the word in the list d. Precision@1000 of ProductSentiRus was reported as 90.5%. Similar combination of improved sentiment lexicons in the new resource (ProductSentiRus+) yields 95.2% in terms of Precision@1000 (Table 2).

We took 5000 of the most probable sentiment words of ProductSentiRus+ lexicon for further work (the same amount as in a previous version) and evaluated it in the tweet subjectivity classification task.

The evaluation is based on TEST data set described in [10], which include two thousand tweets in Russian. We assumed that ProductSentiRus+ comprises sentiment units of Internet language. A tweet was classified as subjective if it contained at least one word from the lexicon. Table 3 demonstrates that such a generalized lexicon can be useful also in tweet subjectivity analysis.

Table 3. Quality of tweet subjective classification

Lexicon	*P*	*R*	F_{subj}
Twitter-based lexicon from (Volkova, 2013)	–	–	61.0
ProductSentiRus (data from Volkova, 2013)	–	–	61.0
ProductSentiRus+	58.5	84.7	**69.2**

6 Conclusion

In this paper we described a combined approach to extraction of domain-specific sentiment lexicons. At first, an initial version of a domain-specific lexicon is obtained by application of a supervised model. At the second stage, the ordered list of sentiment words is refined using information described in RuThes

[2] http://www.cir.ru/SentiLexicon/ProductSentiRus.txt

thesaurus of Russian language, which was lately published.

This combined model is applied to several domains and at last domain-specific sentiment lists are united to create a sentiment word list in the generalized domain of products – ProductSentiRus+, which is an improved version of the only published Russian sentiment lexicon and will be also publicly available. The proposed approach can be applied to other languages and can utilize other thesauri.

Acknowledgments

This work is partially supported by RFBR grant 14-07-00682.

References

[1] Ilia Chetviorkin and Natalia V Loukachevitch. Extraction of Russian sentiment lexicon for product meta-domain. In COLING, 2012. P. 593–610.

[2] Yejin Choi and Claire Cardie. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Vol. 2, 2009. P. 590–598.

[3] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of LREC, vol. 6, 2006. P. 417–422.

[4] Song Feng, Jun Seok Kang, Polina Kuznetsova, Yejin Choi. Connotation Lexicon: A Dash of Sentiment Beneath the Surface Meaning. In Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics ACL-2013. 2013.

[5] Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the semantic orientation of adjectives. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, 1997. P. 174–181.

[6] Raymond Lau, Chun-Lam Lai, Peter Bruza, Kam-Fai Wong. Leveraging web 2.0 data for scalable semi-supervised learning of domain-specific sentiment lexicons. Proceedings of the 20th ACM international conference on Information and knowledge management. ACM. 2011.

[7] Natalia Loukachevitch and Boris Dobrov. RuThes Linguistic Ontology vs. Russian Wordnets. In Proceedings of Global Wordnet Conference. 2013.

[8] Delip Rao and Deepak Ravichandran. Semi-supervised polarity lexicon induction. In Proceedings of the 12th Conference of the European Chapter of the ACL, EACL-2009, 2009. P. 675–682.

[9] Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. The viability of web-derived polarity lexicons. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010. P. 777–785.

[10] Svitlana Volkova, Theresa Wilson, and David Yarowsky. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. In Proceedings of the 51st Annual Meeting of the Association of Computational Linguistics (ACL13), 2013. P. 505–510.

[11] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University. 2002.