CROSS-DOMAIN OPINION WORD EXTRACTION MODEL

Ilia Chetviorkin
Faculty of Computational Mathematics and Cybernetics,
Lomonosov Moscow State University
e-mail: ilia2010@yandex.ru

Natalia Loukachevitch
Research Computing Center
Lomonosov Moscow State University
e-mail: louk_nat@mail.ru

Abstract

In this paper we consider a new approach for domain-specific opinion word extraction in Russian. We propose a set of statistical features and algorithm combination that can discriminate opinion words in a particular domain. The extraction model is trained in a movie domain and then applied to four other domains. We evaluate the quality of obtained sentiment lexicons intrinsically. Finally, our method is adapted to a movie domain in English and demonstrates comparable results.

Keywords: *Sentiment Analysis, Opinion Words, Domain Adaptation*

1. INTRODUCTION

In recent years increased attention is paid to domain adaptation in sentiment analysis research and in natural language processing in general. The reason for this is low generalization ability to new domains of supervised machine learning algorithms, which demonstrate a good quality of solving sentiment analysis problems within one domain. Such behavior of the algorithms trained in one domain and applied to some others can be explained by the differences in lexicons and their complexity [Ponomareva and Thelwall, 2012].

To overcome this issue various adaptation methods are proposed, like ensemble of classifiers [Aue and Gamon, 2005] or graph-based approach [Wu et al., 2009]. Nevertheless, such approaches do not work well for domains whose lexicons differ significantly and recent studies are focused on bridging the gap between domain-specific words [Pan et al., 2010].

For this reason we address the problem of automatic sentiment lexicon construction for various domains. Indeed, sentiment lexicons adapted to a particular domain or topic have been shown to improve task performance in a number of applications, including opinion retrieval [Jijkoun et al., 2010], and expression-level sentiment classification [Choi and Cardie, 2009].

Such lexicons are very useful [Taboada et al., 2011] because they contain sentiment-bearing words typical only for the specific domain. For

example, "must-see" is a strongly opinionated word in the movie domain, but neutral in the digital camera domain. This fact is also pointed out in [Blitzer et al., 2007]. In addition, opinion word extraction from a text collection enables to find slang, vulgarity and non-vocabulary words, which can be strong opinion predictors.

In the current study a new supervised method for domain-specific opinion word extraction is presented. We train our model in one domain and then apply it to several others.

Our approach is based on several text collections, which can be automatically formed for many domains, such as: a collection of product reviews with authors' evaluation scores, a text collection of product descriptions and a contrast corpus (for example, a general news collection). For each word in the review collection we calculate a set of linguistic and statistical features using the aforementioned collections and then use machine learning algorithms for term classification.

To evaluate performance of the proposed method we conduct experiments on data collections in the five different domains: movies, books, computer games, digital cameras and mobile phones. Our approach demonstrates the ability to identify opinion words specific for a particular domain (for example "fabricated" in the movie domain) and reasonable lexicon quality for further automatic utilization of this knowledge.

The reminder of this article is organized as follows. In Section 2 we observe the state-of-the-art methods for sentiment lexicon generation; Section 3 describes the data collections and features involved in the model. In Section 4 we apply our approach to four the other domains. Finally, in Section 5 we extract and evaluate opinion words in English.

2. RELATED WORK

Sentiment lexicons play an important role in most, if not all, sentiment analysis applications, including opinion retrieval, opinion question answering, summarization and opinion mining [Ding et al., 2008]. Even though supervised machine learning techniques have been shown to be effective for sentiment classification task [Pang and Lee, 2008], authors in [Choi and Cardie, 2009] demonstrate that including features from sentiment lexicons boosts classification performance significantly.

Generally there are two main approaches to the automatic identification of opinion words in texts: dictionary-based and corpus-based.

The first approach is based on information from a dictionary or a thesaurus. In this approach the sentiment lexicon is obtained through a bootstrapping process using WordNet or others semantic resources. The

basic principle of these approaches is that synonyms and antonyms of a sentiment-bearing word are also carry sentiment. Therefore, from the initial set of words, a new, more comprehensive set of opinion words can be constructed [Hu and Liu, 2004; Neviarouskaya et al., 2009]. In [Esuli and Sebastiani, 2005], glosses are used for opinion word extraction. The main idea of their method is that words with the same orientation have "similar" glosses.

The second approach is based on finding rules and patterns in the text collections [Kanayama and Nasukawa, 2006; Lu et al., 2011]. In [Velikovich et al., 2010] the authors examine viability of the web-derived polarity lexicon. They utilize a graph propagation framework to the phrase co-occurrence graph, which is built using 4 billion web pages. The obtained lexicon has superior performance to several previously published lexicons and contains wide range of spelling variations, slang and vulgarity.

There are also some works that combine corpus-based and dictionary-based approaches [Ding et al., 2008].

3. THE PROPOSED APPROACH

In this section we will describe our method in respect to the movie domain. We will construct and train the model on the movie data and then try to adapt it to the other domains.

3.1. Data Preparation

We collect 28773 film reviews of various genres from online recommendation service *www.imhonet.ru*. For each review, user's score on a ten-point scale is extracted. We name this collection the **review collection**.

Example of the movie review:

Nice and light comedy. There is something to laugh — exactly over the humor, rather than over the stupidity... Allows you to relax and gives rest to your head.

Besides the review collection we form two more datasets. In these collections the concentration of opinions should be as little as possible. For this purpose, we have collected 17680 movie descriptions (**description collection**) and second contrast corpus is a collection of two million news documents. We have calculated document frequency of each word in the news dataset and use only this frequency list further. This list is named the **news corpus**.

3.2. Collections with Higher Concentration of Opinions

We suggest that it is possible to extract some fragments (and form a collection from them) of reviews from the review collection, which have higher concentration of opinion words. The following fragments are extracted:

- Sentences ending with a "!";
- Sentences ending with a "…";
- Short sentences, with length less than 7 words;
- Sentences containing the word "movie" without any other nouns.

We name this collection — **small collection**.

3.3. Statistical Features

Our aim is to create a high quality list of opinion words based on the calculation of various discriminative features and their combination using machine learning algorithms. We propose the following set of features for each word:

- **Frequency-based.**
  - Collection frequency $f(w)$ (i.e. number of occurrences in all documents in the collection).
  - Document frequency.
  - Frequency of capitalized words.
  - Weirdness (relative frequency).
  - TFIDF.
- **Rating-based.**
  - Deviation from the average score.
  - Word score variance.
  - Sentiment category likelihood for each (*word*, *category*) pair.

We will consider some of them in more details.

Frequency of capitalized words. The frequency (in the review dataset) of each word starting with the capital letter and not located at the beginning of the sentence is calculated. With this feature we are trying to identify potential proper names, which are always neutral.

Weirdness (Relative frequency). To calculate this feature two collections are required: one with high concentration of opinion words and the other — contrast one. The main idea of this feature is that opinion words will be «strange» in the contexts of the contrast collection. This feature is calculated as follows [Ahmad et al, 1999]:

$$Weirdness = \frac{P_s(w)}{P_g(w)}$$

where $P_s(w)$ — probability of the word in a special corpus, $P_g(w)$ — probability of the word in a general corpus. Instead of the collection frequency one can use the document frequency for probability calculation.

Weirdness is calculated using the following collection pairs: *opinion-news*, *opinion-description*, *description-news* with document frequency and *small-description*, *opinion-description* with collection frequency;

TFIDF. There are many varieties of this feature. We used *TFIDF* variant described in [Callan et al., 1992] (based on BM25 function):

$$TFIDF = \beta + (1 - \beta) \cdot tf \cdot idf$$

$$tf(w) = \frac{f(w)}{f(w) + 2}$$

$$idf(w) = \frac{\log\left(\frac{|c| + 0.5}{df(w)}\right)}{\log(|c| + 1)}$$

where $f(w)$ — number of occurrences of term w in a collection, $df(l)$ — number of documents in a collection (e.g. description or news collection) where term w appears, β=0.4 and $|c|$ — total number of documents in a collection.

We calculate the *TFIDF* using the collection pairs: *small-news*, *small-description*, *opinion-news*, *opinion-description* and *description-news*;

3.4. Review Rating-Based Features

As we mentioned above we has collected user's numerical score (on a ten point scale) for each review. Let C={1…10} to be the set of rating categories in the review collection. First, we want to give some definitions, which will be used further.

Definition 1.

i. The probability of rating category c for a given word w:

$$P(c \mid w) = \frac{f(w, c)}{\sum_{c_i \in C} f(w, c_i)}$$

ii. The probability of word w for a given rating category c:

$$P(w \mid c) = \frac{f(w, c)}{\sum_{w_i \in C} f(w_i, c)}$$

Definition 2.

i. Expected category for a given word:

$$E(c \mid w) = \sum_{c_i \in C} c_i \cdot P(c_i \mid w)$$

ii. Expected category in the review collection:

$$E(c) = \sum_{c_i \in C} c_i \cdot P(c_i)$$

Using these definitions we propose the following features:

Deviation from the average score.

$$Dev(w) = \left| E(c \mid w) - E(c) \right|$$

This feature can discriminate words appearing in wide range of rating categories.

Word score variance. One more useful predictor is a word score variance. If a word has a small variance then it might be used in reviews with similar scores and has high probability to be an opinion word.

$$Var(w) = E(c^2 \mid w) - E(c \mid w)^2$$

Scaled likelihood. To get some intuition about how likely a word w is to appear in each sentiment class, we define a scaled log-likelihood:

$$Lhc(w) = \log \frac{P(w \mid c)}{P(w)}$$

Scalability of this feature is required to be comparable between words. We have also added some features aggregating *Lhc* values, like maximum and average.

3.5. Morphological Features

Some linguistic features are also added to our system because they can play crucial role in improving sentiment lexicon extraction.

- Four binary features indicating the part of speech (noun, verb, adjective and adverb).
- Two binary features reflecting the POS ambiguity (i.e. it can have various parts of speech depending on a context) and if this word is contained in a dictionary of a POS tagger.
- Predefined list of prefixes. This feature is a strong predictor for words starting with negation (e.g. *unfunny* — "*nesmeshnoi*").

For Russian and English we used the same morphological parser *Cir_morph*[1].

3.6. Algorithms and Evaluation

To train supervised machine learning algorithms we require a set of labeled opinion words. For our experiments we manually label all words from the review collection with frequency greater than three (18362 words). We consider a word as an opinion one when we can imagine it in any opinion context in the movie domain. All words are tagged by two authors. As a result of our labeling procedure we obtain the list of 4079 opinion words in the movie domain.

Using the tagged data we solve a two class classification problem: separating all words into opinion and neutral categories. For this purpose

[1] http://ru-eval.ru/participants.html#Cirmorph

Weka[2] data mining tool is used. We consider the following algorithms: *Logistic Regression*, *LogitBoost* and *Random Forest*. For all experiments 10 fold cross-validation is used.

Using these algorithms we obtain word lists, ordered by the predicted probability of their opinion orientation. To measure the quality of these lists we use *Precision@n* metric. This metric is very convenient for measuring the quality of list combinations and it can be used with different thresholds. To compare quality of the algorithms in different domains we choose n=1000. This level is not too large for manual labeling and demonstrates the quality in an appropriate way.

The results of classification in the movie domain are in Table 1.

Table 1. *Precision*@1000 in the movie domain

Logistic Regression	***LogitBoost***	***Random Forest***	***Average***
75.7%	75.3%	72.4%	81.5%

We notice that the list of opinion words extracted using each algorithm differs significantly from the others. So we decide to take word weight average in these three lists. The result of such summation can be found in the last column of the Table 1. We will apply and evaluate only this meta-algorithm in the other domains.

As the baseline for our experiments we use word lists ordered by frequency in the review collection and deviation from the average score. *Precision*@1000 in these lists was 26.9% and 35.5% accordingly. Thus, our algorithm gives significant improvements over the baseline. All the other features and their quality estimates can be found in Table 2.

Let us look at some examples of opinion words with the high probability value in the resulting list: *Trogatel'nyi* (*affective*), *otstoi* (*trash*), *fignia* (*crap*), *otvratitel'no* (*disgustingly*), *posredstvenniy* (*satisfactory*), *predskazuemyi* (*predictable*), *ljubimyj* (*love*) etc.

Table 2. *Precision*@1000 for different features

Feature	***Collection***	***Precision@1000***
TFIDF	small-news	38.5%
TFIDF	small-news	36.4%
TFIDF	review-news	30.5%
TFIDF	review-descr	39.8%
Weirdness	review-news (doc.freq)	31.7%

[2] http://www.cs.waikato.ac.nz/ml/weka

Feature	*Collection*	*Precision@1000*
Weirdness	review-descr (doc.freq)	48.1%
Weirdness	small-descr (freq)	49.1%
Weirdness	review-descr (freq)	46.6%
Dev	review	35.5%
Var	review	21.5%
Lhc	review	33.0%
Frequency	review	26.9%
Frequency	small	31.9%
DF	review	27.8%

4. MODEL ADAPTATION

In the previous section we constructed the opinion word extraction model for the movie domain. The next step of the current research is using this model in the four different domains and evaluating the quality of obtained results.

4.1. Additional Datasets

We collected data in the four selected domains. The structure of the datasets is the same as for movie domain. Data collection characteristics for each collection can be found in Table 3.

Table 3. Characteristics of the data collections

	Review Collection	*Description Collection*	*Source*
Books	23,883	22,321	Imhonet
Games	7,928	1,853	Imhonet
Digital Camera	10,208	920	Yandex Market
Mobile Phone	30,620	890	Yandex Market

In further experiments we use the same news corpus as for movie domain.

4.2. Model Utilization and Evaluation

For all words in a particular field (excluding low frequent ones) we compute feature vectors (see Sections 3.3–3.5) and construct a domain word-feature matrix using them. We applied our classification model to these word-feature matrixes and manually evaluated first thousand of the most probable opinion words in each domain. The results of the evaluation are in Table 4.

Table 4. The results of domain adaptation

	Average
Books	86.0%
Games	72.2%
Digital Camera	62.0%
Mobile Phone	73.2%

The latter three domains have a significant drop in the classification quality. We connect this issue with poorness of the description collection in each domain and excess of a neutral domain-specific terminology (especial in the digital camera domain, e.g. "aperture"), which our algorithm tends to give high weights.

To overcome this problem more comprehensive neutral description collections are required, which should decrease the influence of such neutral, very specific words.

5. MULTI-LINGUAL APPLICATION

We assume that the proposed algorithm is rather language-independent. To confirm this thesis we conduct the experiment on the movie reviews in English. We use the review dataset from [Blitzer et al., 2007], but take only reviews from the movie domain. As contrast collections we used plot dataset freely available on the IMDb[3] web site and Reuters-21578[4] news collection. Summing up we had 34,180 movie reviews, 40,000 movie descriptions and a document frequency list calculated using Reuters-21578 collection.

Using these datasets we compute the word-feature matrix following the previously described procedure and apply our model trained on the Russian movie reviews. The evaluated quality of obtained lexicon is **70.5%** according to P@1000 measure.

The most probable opinion words are: remarkable, recommended, overdo, understated, respected, overlook, lame, etc. Some of these words (for example overlook) are opinion words only in the movie domain. Such words can not be found in general sentiment lexicons available for English .

The drop in quality is caused by the simplicity of English POS-tagger used in current research. More accurate English tagger could solve this issue.

The most probable opinion words are: *remarkable*, *recommended*, *overdo*, *understated*, *respected*, *overlook*, *lame*, etc. Some of these words

[3] Information courtesy of The Internet Movie Database (http://www.imdb.com). Used with permission.

[4] http://www.daviddlewis.com/resources/testcollections/reuters21578

(for example overlook) are opinion words only in the movie domain. Such words can not be found in general sentiment lexicons available for English[5].

The drop in quality is caused by the simplicity of English POS-tagger used in current research. More accurate English tagger could solve this issue.

6. CONCLUSIONS

In this paper, we presented a new method for opinion word extraction in a particular domain on the basis of several specific text collections. We applied our algorithm in various domains and demonstrated that it had good generalization abilities. Finally, it was shown that the proposed algorithm could be easily adapted to the other languages and obtained reasonable results.

7. ACKNOWLEDGMENTS

This work is partially supported by RFBR grant N11-07-00588-a.

REFERENCES

1. **Ahmad K., Gillam L., Tostevin L.** 1999. University of Surrey participation in Trec8: Weirdness indexing for logical documents extrapolation and retrieval In the Proceedings of Eigth Text Retrieval Conference (Trec-8).
2. **Aue, A., Gamon, M.** 2005. Customizing sentiment classifiers to new domains: A case study. In: Proceedings of RANLP
3. **Blitzer, J., Dredze, M., Pereira, F.** 2007 Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of ACL 2007, pp. 440–447
4. **Callan J.P., Croft W.B., Harding S.M.** 1992. The INQUERY Retrieval System Proc. of Database and Expert System Applications DEXA-92, 3rd International Conference on Database and Expert Systems Applications / A.M. Tjoa and I. Ramos (eds.). – Springer Verlag, New York, pages 78–93.
5. **Choi Y. and Cardie C.** 2009 Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In EMNLP '09, pages 590–598.
6. **Ding X., Liu B., and Yu P. S.** 2008. A holistic lexicon-based approach to opinion mining. In WSDM '08, pages 231–240.
7. **Esuli A., Sebastiani F.** 2005 Determining the Semantic Orientation of Terms through Gloss Classification. In: Conference of Information and Knowledge Management
8. **Hu M., Liu B.** 2004. Mining and Summarizing Customer Reviews. KDD
9. **Jijkoun V., de Rijke M., and Weerkamp W.** 2010. Generating focused topic-specific sentiment lexicons. In ACL '10, pages 585–594
10. **Kanayama H. and Nasukawa T.** 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In EMNLP '06, Morristown, NJ, USA. pages 355–363
11. **Lu Y., Castellanos M., Dayal U. and Zhai C.** 2011. Automatic Construction of a Context-Aware Sentiment Lexicon: An Optimization Approach In Proceedings of the World Wide Web Conference (WWW)

5 http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

12. **Neviarouskaya A., Prendinger H., and Ishizuka M.** 2009. Sentiful: Generating a reliable lexicon for sentiment analysis. In ACII, pages 1–6
13. **Pan, S. J., Ni, X., Sun, J-T, Yang, Q. and Chen, Z.** 2010. Cross-Domain Sentiment Classification via Spectral Feature Alignment. In Proceedings of the World Wide Web Conference (WWW)
14. **Pang B., Lee L.** 2008. Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval. Now Publishers
15. **Ponomareva N. and Thelwall M.** 2012. Biographies or Blenders: Which Resource Is Best for Cross-Domain Sentiment Analysis? Computational Linguistics and Intelligent Text Processing. Springer Berlin / Heidelberg. pages 488–499
16. **Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.** 2011. Lexicon-based methods for Sentiment Analysis. Computational linguistics, v 37 (2), pages 267-307.
17. **Velikovich, L., Blair-Goldensohn, S., Hannan, K. and McDonald, R.** 2010. The viability of web-dervied polarity lexicons. In Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics.
18. **Wu, Q., Tan, S., Cheng, X.** 2009 Graph ranking for sentiment transfer. In: Proceedings of ACL-IJCNLP 2009, pages. 317–320