КОНТЕКСТНО-ЗАВИСИМЫЙ ПЕРЕВОД СЛОВАРЯ ОЦЕНОЧНЫХ СЛОВ ПРИ ПОМОЩИ ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ

Уланов А. В. (alexander.ulanov@hp.com),
Сапожников Г. А. (gsapozhnikov@gmail.com)

Hewlett-Packard Labs Russia, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Ключевые слова: анализ мнений, оценочные слова, машинный перевод, классификация

CONTEXT-DEPENDENT OPINION LEXICON TRANSLATION WITH THE USE OF A PARALLEL CORPUS

Ulanov A. V. (alexander.ulanov@hp.com),
Sapozhnikov G. A. (gsapozhnikov@gmail.com)

Hewlett-Packard Labs Russia, St. Petersburg State University, St. Petersburg, Russia

The paper deals with multiligual sentiment analysis. We propose a method for projecting an opinion lexicon from a source language to a target language with the use of a parallel corpus. We can make sentiment classification in a target language using an opinion lexicon even if we have no labeled dataset. The advantage of our method is that it captures the context of a word and thus produces a correct translation of it. We apply our method to the language pair English-Russian and conduct sentiment classification experiments. They show that our method allows creating high-quality opinion lexicons.

Keywords: opinion mining, sentiment analysis, opinion words, machine translation

1. Introduction

Sentiment analysis is one of the most popular information extraction tasks both from business and research prospective. It has numerous business applications, such as evaluation of a product or company perception in social media. From the standpoint of research, sentiment analysis relies on the methods developed for natural language processing and information extraction. One of the key aspects of it is the opinion word lexicon. Opinion words are such words that carry opinion. Positive words refer to some desired state, while negative words — to some undesired one. For example, "good" and "beautiful" are positive opinion words, "bad" and "evil" are negative. Opinion phrases and idioms exist as well. Many opinion words depend on context, like the word "large". Some opinion phrases are comparative rather than opinionated, for example "better than". Auxiliary words like negation can change sentiment orientation of a word.

Opinion words are used in a number of sentiment analysis tasks. They include document and sentence sentiment classification, product features extraction, subjectivity detection etc. [12]. Opinion words are used as features in sentiment classification. Sentiment orientation of a product feature is usually computed based on the sentiment orientation of opinion words nearby. Product features can be extracted with the help of phrase or dependency patterns that include opinion words and placeholders for product features themselves. Subjectivity detection highly relies on opinion word lists as well, because many opinionated phrases are subjective [14]. Thus, opinion lexicon generation is an important sentiment analysis task. Detection of opinion word sentiment orientation is an accompanying task.

Opinion lexicon generation task can be solved in several ways. The authors of [12] point out three approaches: manual, dictionary-based and corpus-based. The manual approach is precise but time-consuming. The dictionary based approach relies on dictionaries such as WordNet. One starts from a small collection of opinion words and looks for their synonyms and antonyms in a dictionary [10]. The drawback of this approach is that the dictionary coverage is limited and it is hard to create a domain-specific opinion word list. Corpus-based approaches rely on mining a review corpus and use methods employed in information extraction. The approach proposed in [9] is based on a seed list of opinion words. These words are used together with some linguistic constraints like "AND" or "OR" to mine additional opinion words. Clustering is performed to label the mined words in the list as positive and negative. Part of speech patterns are used to populate the opinion word dictionary in [21] and Internet search statistics is used to detect semantic orientation of a word. Work [7] extends the mentioned approaches and introduces a method for extraction of context-based opinion words together with their orientation. Classification techniques are used in [2] to filter out opinion words from text. The approaches described were applied in English. There are some works that deal with Russian. For example, paper [4] proposes to use classification. Various features, such as word frequency, weirdness, and TF-IDF are used there.

Most of the research done in the field of sentiment analysis relies on the presence of annotated resources for a given language. However, there are methods

which automatically generate resources for a target language, given that there are tools and resources available in the source language. Different approaches to multilingual subjectivity analysis are studied in [14] and [1] and are summarized in [3]. In one of them, subjectivity lexicon in the source language is translated with the use of a dictionary and employed for subjectivity classification. This approach delivers mediocre precision due to the use of the first translation option and due to word lemmatization. Another approach suggests translating the corpus. This can be done in three different ways: translating an annotated corpus in the source language and projecting its labels; automatic annotation of the corpus, translating it and projecting the labels; translating the corpus in the target language, automatic annotation of it and projecting the labels. Language Weaver[1] machine translation was used on English-Roman and English-Spanish data [3]. Classification experiments with the produced corpora showed similar results. They are close to the case when test data is translated and annotated automatically. This shows that machine translation systems are good enough for translating opinionated datasets. It is also confirmed by the authors of [19] when they used Google Translate[2], Microsoft Bing Translator[3] and Moses[4].

Multilingual opinion lexicon generation is considered in the recent paper [19] that presents a semi-automatic approach with the use of triangulation. The authors use high-quality lexicons in two different languages and then translate them automatically into a third language with Google Translate. The words that are found in both translations are supposed to have good precision. It was proven for several languages including Russian with the manual check of the resulting lists. The same authors collect and examine entity-centered sentiment annotated parallel corpora [20].

In this paper we develop the idea of multilingual sentiment analysis. We propose a method for projecting an opinion lexicon from a source language to a target language with the use of a parallel corpus. We apply it to the language pair English-Russian having a collection of a parallel and a pseudo-parallel review corpora. The method is evaluated against the baseline, which is a translation of the opinion word lexicon with Goolge Translate. Sentiment classification experiments are conducted to evaluate the quality of the lexicons. The advantages of our method are the following. It captures the context of opinion words thus producing correct translations. It doesn't require a machine translation tool, as in [19] or a bilingual dictionary as in [14]. However, machine translation tool may be employed in the absence of parallel corpus or for better recall. The opinion lexicon is needed only in one language, unlike in work [19] where 2 lexicons are required.

---

1 http://www.sdl.com/products/automated-translation/

2 http://translate.google.com/

3 http://www.bing.com/translator

4 http://www.statmt.org/moses/

2. Approach

The idea of our approach is to use a parallel corpus to construct an opinion lexicon in a target language, given that there is an opinion lexicon in a source language. A parallel corpus is a text with its translation to the target language. We suppose that it contains opinionated sentences. An opinion lexicon is a set of words carrying opinion. It is not necessarily divided into positive/negative or other groups. The opinion lexicon for the target language is extracted from the parallel corpus by translating the words from the opinion lexicon in the source language. The algorithm of the method is as follows:

1. Collect a corpus of parallel reviews, align sentences
2. Compute word lexical translation probabilities
3. Collect opinion words translations and normalize them

Let us consider the mentioned steps in greater details. The task of parallel corpus acquisition and preparation is a well-studied area of research [8]. One collects or crawls data that is available in different languages. Parallel documents are determined by some identifier, e.g. name, time, or specific number. Documents are split into sentences by the sentence splitter, paragraphs are kept preserved. The resulting text is processed by the sentence aligner. A parallel corpus with opinionated texts can be obtained from the sites that post reviews in different languages (manually translated). Usually, such reviews are editorial. They contain opinionated text; however opinion words there tend to be more polite than in forums or user reviews. The size of the corpus is less important than the coverage of words from the source opinion lexicon. In the absence of a natural parallel corpus, a pseudo-parallel corpus can be used [20], which is a text along with its translation done by an automatic translation system.

Lexical translation probabilities of words are computed on the aligned corpus:

$$p_s(t) \text{ and } p_t(s),$$

where *t* is a word in the target language, *s* is a word in the source language. Lexical translation is a translation of a word in isolation. To compute it, one has to count how many times a certain word was translated into different options within the aligned sentences. The ratios of these counts and the count of that word represent the distribution of lexical translation probabilities. This operation is performed in both translation directions, i. e. $t \to s$ and $s \to t$.

Opinion word translations are collected for a given opinion word list in the source language. Correct translation of a source opinion word is determined as follows:

$$\exists t, s: p_t(s) = \max_i p_i(s) \;\; and \;\; p_s(t) = \max_j p_j(t)$$

In other words, to make translation of a source word, we choose a word with a maximum translation probability and check that it translates to the same word with a maximum probability as well. The translated words are normalized.

3. Experiments

3.1. Opinion lexicon projection

We conducted several experiments to validate the proposed approach. Two parallel datasets are used in our experiments. The first one consists of Russian and English reviews collected from the Mobile Review site[5]. We downloaded all pages from the English editorial of the site. Then we downloaded Russian versions of these pages using English links without the token "-en". We will refer to this dataset as to "MR".

The second one consists of the first 5,000 lines from the reviews of books, cameras and films taken from ROMIP 2011 sentiment analysis dataset [5] and 1,000 lines of iPhone4 reviews from Yandex Market[6] along with their Russian translation produced by Google Translate. We will refer to it as to "ROMIP-GT". The datasets are split into sentences with Freeling[7] and aligned with Microsoft Bilingual Sentence Aligner [18]. After the above mentioned, the aligned "MR" contains 579,559 Russian and 726,798 English words, the aligned "ROMIP-GT" contains 714,533 Russian and 820,241 English words. We use GIZA++ [15] for creating word lexical translation tables. English opinion word lists are downloaded from Bing Liu's homepage[8]. There are 4,818 negative and 2041 positive words. We will refer to this list as to "BL" dictionary. Mystem[9] is used to normalize the Russian words. They are transformed to singular, masculine, nominative, present time forms.

We produce 4 opinion lexicons in Russian in total. During lexicons construction we remove all words containing spaces and minuses, and which are shorter than 3 symbols. "BL-GT" lexicon contains translated and normalized opinion words from "BL". "BL-GT filtered" lexicon was constructed in the following way. Words from "BL" were translated to Russian and then back to English using Google Translate. We collected only those Russian translations that produced English translation equal to its English original.

"MR" lexicon is created by application of our method to "MR" parallel corpora. "ROMIP-GT" lexicon is created using our method with the "ROMIP-GT" dataset. "ROMIP-GT merged" lexicon is produced in the following way. We applied our method to 3 subsets of "ROMIP-GT", i.e. books, films and cameras. Then the resulting lists were merged. The number of opinion words in each lexicon is listed in Table 1. Table 2 shows intersections of the lexicons.

---

5 http://mobile-review.com/

6 http://market.yandex.ru/

7 http://nlp.lsi.upc.edu/freeling/

8 http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

9 http://company.yandex.ru/technologies/mystem/

Table 1. Opinion words number

Lexicon	Positive	Negative	Total
BL (English)	*2,041*	*4,818*	*6,859*
BL-GT	1,443	3,067	4,510
BL-GT filtered	907	2,037	2,944
MR	163	182	345
ROMIP-GT	706	1,311	2,017
ROMIP-GT merged	1,057	1,812	2,869
Union:	1,993	4,040	6,033

The lexicon "BL-GT" is the biggest with almost 4.5 thousand words. However, it is less than the original list by 34%. This is due to the fact that some words were translated to the same surface form (27%), due to phrases removal (they contain spaces) and due to normalization. There is a small portion of untranslated words as well. "BL-GT filtered" is almost a half of the original dictionary. It is interesting to see, however, that so many words are translated from English to Russian and back to English with the original form.

"MR" lexicon that was produced from the Mobile Review parallel corpus is rather small. This is because it contains a different English lexicon than the opinion word list "BL". The "MR" texts were written by a limited number of persons, while the opinion lexicon "BL" contains contributions from a lot of people.

Interestingly, "ROMIP-GT merged" is 30% bigger than "ROMIP-GT" and is almost as big as "BL-GT filtered". Table 2 suggests that "ROMIP-GT merged" has 1222 or 45% of words in common with "BL-GT filtered". This is because the words in the latter case were translated in isolation while in the first case they were translated within the context.

We can get as many as 6,033 opinion words if we merge all lists, which is 89% of the original English list.

Table 2. Opinion words intersection

Intersection		Words		
		pos	neg	total
MR	ROMIP-GT merged	118	88	206
MR	BL-GT	132	178	310
ROMIP-GT merged	BL-GT	626	1,006	1,632
ROMIP-GT merged	BL-GT filtered	436	786	1,222

We made a manual assessment of the lexicons. Table 3 shows their precision. "BL-GT filtered" is the most accurate. This can be explained by the fact that it contains just the right English words translated unambiguously without context. Also, we compared "MR" and "ROMIP-GT" lists. The first was derived from professional reviews, the second from user reviews. It is interesting to note that "MR" contains "specific" opinion words and "ROMIP" contains emotional words.

Table 3. Precision by manual assessment

Lexicon	Precision
BL-GT	0,79
BL-GT filtered	0,87
MR	0,76
ROMIP-GT	0,83
ROMIP-GT merged	0,82

3.2. Document Sentiment Classification

The number of words in the list doesn't mean its quality. We conducted several experiments to benchmark the produced opinion word lists. We decided not to check the words manually, but to use them in the real-world task, that is sentiment classification. The experiments are performed on the annotated part of ROMIP 2011 dataset [5]. It contains reviews of books, films and cameras. There are 750 positive and 124 negative review instances.

Counting the number of positive and negative words is the most straightforward way to text sentiment classification [13]. The one with the greater number of opinion words wins. The work [17] suggests that it is better to consider the presence of an opinion word in text rather than the number of appearances. We implement both approaches. We will refer to the first as to "Frequency voc" and to the second as to "Binary voc".

Supervised approaches to text sentiment classification were studied by Pang et al. [17]. We use a linear perceptron classifier with two types of feature computation: term frequencies and delta TF-IDF. The latter was proposed by Martineau et al. [11] and proven to be efficient for sentiment classification in Russian [16]. The experiment results of these methods were obtained after performing 10-fold cross validation. These results act as a base line of supervised classification that requires an annotated dataset. We compare them with dictionary-based classification that does not require class labels to train, because it has negative and positive words. Therefore, results of supervised classification are considered as a higher bound for a dictionary based.

Table 4. Experiment results

Lexicon	Method	MicroP	MicroR (Acc)	MacroR	MacroF1
	Perceptron	**0.84**	**0.84**	0.59	0.60
	Perceptron + TfIdf	**0.84**	**0.84**	**0.62**	**0.63**
Romip-GT	Binary Voc	0.76	0.68	0.59	0.58
	Frequency Voc	0.79	0.72	0.59	0.59
Romip-GT merged	Binary Voc	**0.84**	0.80	0.59	0.61
	Frequency Voc	**0.86**	**0.82**	0.59	**0.61**

Lexicon	Method	MicroP	MicroR (Acc)	MacroR	MacroF1
BL-GT	Binary Voc	0.65	0.60	**0.62**	0.54
	Frequency Voc	0.73	0.69	0.59	0.56
BL-GT filtered	Binary Voc	0.78	0.78	0.59	0.58
	Frequency Voc	0.77	0.72	0.58	0.58
MR	Binary Voc	0.67	0.52	0.50	0.49
	Frequency Voc	0.66	0.53	0.51	0.50

The experiment results are represented in Table 4. The binary approach provides the same weight to all of the words. Low performance of the binary approach as compared with the frequency approach means that the lexicon is of low quality. It may contain common words that can be found in the text (that rarely speak about subjectivity). So we can say that "BL-GT" is rather dirty. "ROMIP-GT merged" gives the best performance among the opinion lexicons. It has the same number of words as "BL-GT filtered", but the performance of the "ROMIP-GT merged" is higher, so we can say that its quality for sentiment classification is better. It is because the words in "ROMIP-GT merged" were translated with the use of context unlike the words in "BL-GT filtered". "BL-GT filtered" shows better results in manual assessment, but worse results in classification. We can explain this by the fact that "ROMIP-GT merged" contains such words that out of context may seem not opinion words or words that are more often used in user reviews as compared with words from "BL-GT filtered".

We supposed that the increase in the classification performance could be due to the fact that we used a part of the big dataset ROMIP 2011 to retrieve "ROMIP-GT merged", and the labeled dataset that was used for classification was also a part of ROMIP 2011. However, it turned out that the intersection between these parts did not exceed 1%, and it couldn't lead to the significant increase of the classification performance.

We use our lexicons as a list for feature selection as in [6], and train a linear perceptron classifier. It produces nearly the same results both for "ROMIP-GT merged" and "BL-GT filtered". This experiment shows that "BL-GT filtered" contains enough words that can be used as classification features. However, it also contains common words that have low weight in the supervised classifier, which does not happen when this lexicon is used in vocabulary classification.

4. Conclusion

We proposed a novel method for opinion lexicon projection from a source language to a target language with the use of a parallel corpus. The method was applied to different datasets and evaluated against the baseline. The quality of created lexicons was evaluated in sentiment classification benchmark. The experiments showed that the lexicons are of high quality. They can be used for sentiment annotation of a corpus in a target language as well.

Out future work is related to enhancement of the method and conducting more experiments. We plan to work with opinion phrases, investigate other translation

options instead of the most probable ones. We will apply our method to other language pairs, apart from English-Russian. Additionally, it will be interesting to explore how the method can be applied to other tasks, such as subjectivity lexicon projection and, more general, multilingual projection of document features.

References

1. *Banea C., Mihalcea R., Wiebe J., Hassan S.* Multilingual subjectivity analysis using machine translation. EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2008.
2. *Breck, E., Y Choi, and C. Cardie.* Identifying expressions of opinion in context. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-2007), 2007.
3. *Carmen Banea, Rada Mihalcea, and Janyce Wiebe.* Multilingual Sentiment and Subjectivity, in Multilingual Natural Language Processing, editors Imed Zitouni and Dan Bikel, Prentice Hall, 2011.
4. *Chatviorkin Ilya, Lukashevich Natalia.* Automatic Extraction of Domain-Specific Opinion Words. Proceedings of the International Conference Dialog, 2010.
5. *Chatviorkin Ilya, Braslavski Pavel, Lukashevich Natalia.* Sentiment analysis track at ROMIP 2011. Proceedings of the International Conference Dialog, 2012.
6. *Dang Y., Zhang Y., Chen H.* A lexicon-enhanced method for sentiment classification: An experiment on online product reviews, IEEE 2010.
7. *Ding, X., B. Liu, and P. Yu.* A holistic lexicon-based approach to opinion mining. In Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008), 2008.
8. *Eisele A., Chen Y.* MultiUN: A Multilingual Corpus from United Nation Documents. In Language Resources and Evaluation, 2010.
9. *Hatzivassiloglou, V. and K. McKeown.* Predicting the semantic orientation of adjectives. In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-1997), 1997.
10. *Hu, M. and B. Liu.* Mining and summarizing customer reviews. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), 2004.
11. *J. Martineau and T. Finin.* Delta TFIDF: An Improved Feature Space for Sentiment Analysis. In Proceedings of the Third AAAI Internatonal Conference on Weblogs and Social Media, San Jose, CA, 2009.
12. *Liu, Bing.* Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data. 2nd ed. 2011, XX, 622 p.
13. *Melville P., Gryc W., and Lawrence R.* Sentiment analysis of blogs by combining lexical knowledge with text classification. KDD 2009.
14. *Mihalcea R., Banea C. and Wiebe J.* Learning Multilingual Subjective Language via Cross-Lingual Projections, in Proceedings of the Association for Computational Linguistics (ACL 2007), Prague, June 2007.

15. *Och F. J., Ney H.* A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, volume 29, number 1, pp. 19–51 March 2003.
16. *Pak A., Paroubek P.* Language independent approach to sentiment analysis (Limsi participation in romip'11) Proceedings of the International Conference Dialog, 2012.
17. *Pang B., Lee L.* Thumbs up? Sentiment Classification using Machine Learning Techniques, In Proc. of the conference on the Empirical Methods 2002
18. *Robert C. Moore*. Fast and Accurate Sentence Alignment of Bilingual Corpora. In Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users (2002), pp. 135–144
19. *Steinberger J., Ebrahim M., Ehrmann M., Hurriyetoglu A., Kabadjov M., Lenkova P., Steinberger R., Tanev H., Vázquez S., Zavarella V.* Creating Sentiment Dictionaries via Triangulation. Decision Support Systems, May 2012.
20. *Steinberger J., Lenkova P., Kabadjov M., Steinberger R., van der Goot E.* Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora. In Proceedings of the International Conference Recent Advances in Natural Language Processing, 2011.
21. *Turney, P.* Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2002), 2002.