Extracting Domain-Specific Opinion Words for Sentiment Analysis

Ivan Shamshurin

National Research University – Higher School of Economics,
20 Myasnitskaya Ulitsa, Moscow, 101000, Russia
ivanshamshurin@gmail.com

Abstract. In this paper, we consider opinion word extraction, one of the key problems in sentiment analysis. Sentiment analysis (or opinion mining) is an important research area within computational linguistics. Opinion words, which form an opinion lexicon, describe the attitude of the author towards certain opinion targets, i.e., entities and their attributes on which opinions have been expressed. Hence, the availability of a representative opinion lexicon can facilitate the extraction of opinions from texts. For this reason, opinion word mining is one of the key issues in sentiment analysis. We designed and implemented several methods for extracting opinion words. We evaluated these approaches by testing how well the resulting opinion lexicons help improve the accuracy of methods for determining the polarity of the reviews if the extracted opinion words are used as features. We used several machine learning methods: SVM, Logistic Regression, Naïve Bayes, and KNN. By using the extracted opinion words as features we were able to improve over the baselines in some cases. Our experiments showed that, although opinion words are useful for polarity detection, they are not sufficient on their own and should be used only in combination with other features.

Keywords: Sentiment Analysis, Natural Language Processing, Machine Learning.

1 Introduction

In this paper, we consider opinion word extraction, one of the key problems in sentiment analysis. Sentiment analysis (or opinion mining) is an important research area within computational linguistics. It is mainly concerned with methods for determining the attitude of the author towards the subject of her text (so-called "polarity") by classifying documents as positive, negative, or neutral.

The increasing popularity of sentiment analysis is due to widespread opinion-rich user-generated resources such as online social networks, personal blogs, wikis, and review websites. [8] provides a general survey of topics in sentiment analysis, including the problem of determining the polarity of texts.

People often express their opinions of products, events, etc. by using subjective opinion words, such as "beautiful", "boring", "interesting", "banal", etc. Opinion

I. Batyrshin and M. González Mendoza (Eds.): MICAI 2012, Part II, LNAI 7630, pp. 58–68, 2013.

words, which form an opinion lexicon, describe the attitude of the author towards certain opinion targets, i.e., entities and their attributes on which opinions have been expressed. Hence, the availability of a representative opinion lexicon can facilitate the extraction of opinions from texts. For this reason, opinion word mining is one of the key issues in sentiment analysis [4, 7, 10–12].

In our paper we applied several methods of machine learning and used opinion words as features in order to determine text polarity (see section 5). The intuition suggests that using opinion words as features makes it possible to improve accuracy of documents classification.

The main issue in creating a list of opinion words is its dependency on the subject domain, e.g., a word can be used to express an opinion in one domain (*"original* ***Japanese*** *quality"*) and be neutral in another (*"****Japanese*** *literature"*). Also there is another case, when a word is an indicator of an opinion in both domains, but in the first one it is positive ("***old*** wine") and in the second one it is negative (*"an* ***old*** *car"*).

Hence when forming an opinion word list a better approach is to build domain-specific lists rather than one general-purpose list in the subject domain under consideration. Another reason for this is the fact that some opinion words are created by users and are not contained in dictionaries.

It is worth mentioning that opinion words can be divided into two types: "pure" opinion words (*"beautiful", "boring"*) and "conditional" opinion words, which indicate objective information (size, age, etc.: *"Japanese", "old"*), but which are used as estimative words. Furthermore, the polarity of "pure" opinion words does not depend on the context, whereas in the case of "conditional" opinion words the polarity may change to the opposite one (see example above: *"old wine", "an old car"*).

When extracting "conditionally" evaluative opinion words one should take into account the following fact. Among these adjectives there are both words with constant polarity in the subject domain under consideration, and words with variable polarity. Let us consider the case, when in a certain domain the term $word_1$ is generally used in a positive context, so we label this word as opinion word (i.e. positive). At the same time the term $word_2$ is used both in positive contexts and in negative contexts. The term $word_2$ is opinion word (contains an affective evaluation), but there is approximately equal number of supporters and opponents of $word_2$. In this case it is harder to extract $word_2$ than $word_1$, because the simple statistical approach does not work. In this paper we do not extract this type of opinion words.

2 Related Work

Existing approaches to opinion word extraction can be divided into two categories: corpora-based [3, 5, 15] and dictionary-based [4, 6, 14] methods. We follow the corpora-based paradigm, which makes it possible to extract domain-specific opinion words.

In [3] the method for determining polarities of opinion adjectives using corpus was proposed for the first time. Different conjunction patterns were studied with conjunctions such as *“and”, “or”*. The idea is that we can define the polarity of the words in conjoined pairs if we know the polarity of the second word in the pair. As this method relies on conjunctions, the algorithm does not allow the extraction of isolated, not conjoined opinion words.

In [13] a double-propagation method is proposed which outperforms some of state-of-the-art methods. The main idea of the double propagation approach is the following: we start by fixing a set of known opinion words, which will be used as a seed in the subsequent process. Then on each iteration we extract new opinion words (and opinion targets) using words from this seed, as well as words extracted in previous iterations, through some predefined syntactic relations. One of the advantages of this method is that it does not require any additional text corpora and dictionaries. The polarity assignment of the newly extracted words is also implemented.

[16] is the first work in which the task of extracting opinion words-nouns is considered. There it is proposed to determine an opinion word-noun as a noun which can be found more often either in a positive or in a negative context. By context we mean neighborhood of the term (neighbour words), which may contain positive or negative opinion words-adjectives (we assume that evaluative adjectives are known).

In [1] a machine learning based approach for automatic extraction of opinion words (both adjectives and non-adjectives) is proposed. The method requires additional two corpora: a corpus of the neutral descriptions of opinion targets and neutral contrast corpus (news). There are 17 attributes for machine learning, which depend on the corpora. One of them, the deviation index, is used in this paper.

In this paper, we considered several methods for automatic extraction of domain-specific opinion words. For one of these methods, double propagation [13], we proposed a new technique for building a seed to be used as a starting point for opinion word extraction. Syntactic rules of this method were adapted for Russian.

In order to evaluate the methods of opinion word extraction we used several machine learning methods for determining text polarity. For this, we developed several text representations using various features and compared them empirically.

3 Methods of Opinion Word Extraction

We designed and implemented several methods for extracting opinion words.

3.1 Double Propagation Approach

The first approach is based on the double propagation technique proposed in [13], where it is claimed to outperform some of the state-of-the-art methods.

We devised several syntactic rules and tested the approach on reviews in Russian using the Semantic analyzer from AOT project[1] as a dependency parser. Rules for opinion words (OW) and opinion targets (OT) extraction are listed below. Most of them are based on the ideas of the rules from [13].

1. $OW \rightarrow OT$
   (a) $OW - OT$, table 1

Table 1. WT_1 rule

RuleID	Semantic Pattern
WT_1	OW – $PROPERTY$ – OT
Example (transliteration): *"V etom filme igraet izvestnaya aktrisa."* Example (translation): *"A famous actress played in this movie."* Output: *famous* → *actress.* Having *"famous"* we obtain *"actress"*.	

   (b) $OW - H - OT$, i.e. OW depends on OT through H, table 2

Table 2. WT_2 rule

RuleID	Semantic Pattern
WT_2	OT – F-ACT/S-ACT – IS/ARE – S-ACT/F-ACT – OW
Example (transliteration): *"Film klasny." "Film – klasny."* These two sentences mean *"The movie is great."* Output: *great* → *movie.*	

   Here F-ACT, S-ACT are semantic roles (first and second actants respectively[2]). Here, as we often can see in Russian, the predicate is omitted, but the semantic analyzer understands it and adds *"est"*, which means *"is"/"are"*. So we can identify this kind of relations.
2. $OT \rightarrow OW$
   (a) $OT - OW$, table 3.
   (b) $OT - H - OW$, i.e. OT depends on OW through H, table 4.
3. $OW \rightarrow OW$
   (a) $OW - OW$, table 5.
4. $OT \rightarrow OT$
   (a) $OT - OT$, table 6.
   $BELNG$ is a semantic variable, which means one thing is the part of another one. In this example this relation is identified not as simple as

[1] http://aot.ru

[2] http://aot.ru/docs/SemRels.htm

Table 3. TW_1 rule

RuleID	Semantic Pattern
TW_1	the same as in WT_1
Example: the same as in WT_1 rule Output: *actress* $\rightarrow$ *famous.*	

Table 4. TW_2, TW_3 and TW_4 rules

RuleID	Semantic Pattern
TW_2	the same as WT_2
Example: the same as in WT_2. Output: *movie* $\rightarrow$ *great.*	
TW_3	$OT - CONJ(OW_1, OW_2)$
Example (transliteration): *"Istoriya neobychnaya i intriguyshaya."* Example (translation): *"The story is unusual and intriguing."* Output: *story* $\rightarrow$ *unusual, intriguing.*	
TW_4	$OT - S\text{-}ACT/F\text{-}ACT - IS - F\text{-}ACT/S\text{-}ACT - AND(OW_1, OW_2)$
Example (transliteration): *"Film horoshy i interesny."* Example (translation): *"The movie is good and interesting."* Output: *movie* $\rightarrow$ *good, interesting.*	

Table 5. WW_1 and WW_2 rules

RuleID	Semantic Pattern
WW_1	$CONJ(OW_1, OW_2, ...OW_n)$
Example (transliteration): *"Smeshnaya i neobychnaya komediya."* Example (translation): *"A funny and unusual comedy."* Output: *funny* $\rightarrow$ *unusual.*	
WW_2	$COMMA(OW_1, OW_2, ...OW_n)$
Example (transliteration): *"Krasivaya, zhivaya, rozhdestvenskaya komediya."* Example (translation): *"Nice, lively, Christmas story."* Output: *nice* $\rightarrow$ *lively, Christmas.*	

.

in English, because in English we use "of the" (which can help identify the relation), but, in Russian, possessive case is used for this purpose, which does not require prepositions.

(b) $OT - H - OT$, table 7.
In the TT_4 examples in Russian sentences the predicate is omitted again, but the semantic analyzer understands it and adds "IS".

Table 6. TT_1 and TT_2 rules

RuleID	Semantic Pattern
TT_1	$CONJ(OT_1, OT_2, ...OT_n)$
Example (transliteration): *"Horoshie rezhiser i aktyori."* Example (translation): *"A good director and actors."* Output: *artist* $\rightarrow$ *director.*	
TT_2	$OT - BELNG - OT$
Example (transliteration): *"Interesny suzhet filma..."* Example (translation): *"An interesting plot of the movie..."* Output: *movie* $\rightarrow$ *plot.*	

Table 7. TT_3 and TT_4 rules

RuleID	Semantic Pattern
TT_3	$OT - HAS - OT$
Example (transliteration): *"Rezhiser imeet horoshuyu filmografiu."* Example (translation): *"The director has a good filmography."* Output: *director* $\rightarrow$ *filmography.*	
TT_4	$OT{-}F\text{-}ACT/S\text{-}ACT{-}IS{-}S\text{-}ACT/F\text{-}ACT{-}OT$
Example (transliteration): *"Titanic eto luchyiy film." "Titanic – luchyiy film."* The translation of these sentences is *"The "Titanic" is the best movie."* Output: *movie* $\rightarrow$ *Titanic.*	

3.2 A Method Based on Conditional Probability

This approach is based on conditional probability: if the ratio between the conditional probability of the word occurrence in a positive (negative) review and the conditional probability of the word occurrence in a negative (positive) review is higher than a certain threshold, we label it as an opinion word. In our experiments, we observed that the optimal threshold was 1.25.

3.3 A Method Based on a Pointwise Mutual Information

We also tried two other scoring methods to select opinion words: one is based on the Pointwise Mutual Information [2].

Semantic orientation is based on the concept Pointwise Mutual Information (PMI) with the words "excellent" and "poor".

$$PMI(word_1, word_2) = log_2 \frac{p(word_1 + word_2)}{p(word_1)p(word_2)} \quad (1)$$

$$SO(phrase) = PMI(phrase, \text{``}excellent'') - PMI(phrase, \text{``}poor'') \quad (2)$$

PMI-IR:

$$SO(phrase) = log_2 \frac{hits(phrase\ NEAR\ \text{“}excellent''})hits(\text{“}poor'')}{hits(phrase\ NEAR\ \text{“}poor'')hits(\text{“}excellent'')} \quad (3)$$

3.4 An Approach Based on a Deviation Index.

Deviation score based approach reflects the deviation of the average scores of the reviews where the word occurs from the average score of the entire dataset.

$$dev(l) = |\frac{\sum_{i=1}^{n} m_i k_i}{k} - \frac{\sum_{i=1}^{n} m_i}{n}|, \sum_{i=1}^{n} k_i = k \quad (4)$$

Here l is the considering term, n is the total number of reviews, m_i is the mark of the i-th review, k_i is the number of occurrences of the term in i-th review (0 if not occured).

4 Experimental Evaluation

4.1 Data Description

We used the four approaches to extract opinion lexicons from a real-life dataset that was obtained from `http://imhonet.ru/`. It is a customer review collection dated from January 2009 to December 2011 (87563 reviews). We evaluated these approaches by testing how well the resulting opinion lexicons help improve the accuracy of methods for determining the polarity of the reviews if the extracted opinion words are used as features. For this purpose, we collected reviews dated from January 2012 to March 2012 (500 positive and 500 negative reviews) each consisting of at least 15 words.

4.2 Determining Text Polarity

We used several machine learning methods: SVM, Logistic Regression, Naïve Bayes, and KNN, implemented as part of an open-source data mining tool Rapid-Miner[3].

We represent each document *doc* by a feature-count vector $(n_1(doc),...n_m(doc))$. Our experiments show that machine learning with the presence-based features provide us better results than the machine learning with frequency based features. Presence based features mean setting $n_i(doc)$ to 1 if feature f appears in *doc*, 0 - otherwise.

Baselines. Random-choice baseline provides us with the accuracy 0.5. Accuracies for other baselines are presented below (the average ten-fold cross-validation results):

Table 8. Baselines

	features	# features	Freq./Pres.	KNN	NB	SVM	LR
1	random-choice	0.5					
2	unigrams (2, 2)	741	Freq.	93.4	74.2	85.2	94.3
3	unigrams (2, 2)	741	Pres.	**95.2**	72.7	86.5	95.0
4	*OW* 15 pos, 15 neg	30	Pres.	66.2	59.8	72.5	72.7

Table 9. Manually created list of opinion words (translation presented)

Positive	Negative
beautiful wonderful delightful nice excellent perfect positive magnificent ideal cool high-quality worthful good fine superb	*dim scandalous vile dreadful ugly odious primitive disgusting cheerless idiotic bad vapid slight miserable poor*

In this table features of baselines 2 and 3 are unigrams which occured at least 2 times in positive and 2 times in negative reviews; features of the baseline 4 are manually created 15 positive and 15 negative opinion words (see Table 9).

The baseline feature representations—unigrams and human-selected opinion words—allow us to get 95.2% and 72.7% accuracy, respectively.

Using the Extracted Opinion Words for Improving Baseline. By using the extracted opinion words as features we were able to improve over the baselines in some cases. The average ten-fold cross-validation results are reported in the Table 11.

The description of the feature representation is presented in the Table 10.

Extracting Opinion Targets. Nouns that often co-occur with opinion words (both positive and negative) are likely to be opinion targets [12].

A list of nouns from the frequent bigrams with opinion words is presented below (32 words, frequency threshold is 6, translation is presented):

place, actor, action, impression, time, life, play, idea, story, picture, cinema, comedy, end, place, moment, music, image, feeling, viewing, serial, tale, meaning, trick, scene, plot, theme, thriller, movie, part, man, joke, effect, humour

With the frequency threshold 5 the following words were obtained additionally to the nouns above (12 words, translation is presented):

version, character, book, love, cartoon, evaluation, character, job, episode, situation, side, sense

A list of 21 newly extracted nouns for the frequency threshold 4 is presented below:

[3] `http://rapid-i.com/`

Table 10. Feature representation models for improving the baselines: description

	Description
1.1, 1.2	bigrams $Noun + Adj$, which occured at least k times in positive and k times in negative reviews
2	1.1, 1.2 + unigrams with the frequency threshold 2, 2
3	opinion words, obtained with the conditional probability based method with the frequency threshold 3, 3, reviews are represented as vektors
4	3 + unigrams with the frequency threshold 2, 2
5.1	Feature representation 2, described in 4.2; 9 opinion targets are manually created: *"movie"*, *"acting"*, *"actor"*, *"director"*, *"comedy"*, *"melodrama"*, *"cartoon"*, *"horror movie"*, *"plot"*; opinion words, obtained with the conditional probability based method with the frequency threshold 3, 3
5.2–5.4	similar to 5.1, but OT list of the length k is automatically created (see "Extracting Opinion Targets")
6.1–6.2	Feature representation 2, described in 4.2 and unigrams
7	300 terms with the highest and 300 terms with the lowest dev, described in 3.4
8	7 + unigrams
9	8 + Feature representation 2, described in 4.2
10	9 + 2 features: "!" and difference between the number of ")" and "("

Table 11. Feature representation models for improving the baselines: results

	Features	# feat.	Fr./Pr.	KNN	NB	SVM	LR
1.1	$N + Adj(1,1)$	23	Pres.	60.2	59.0	62.2	60.3
1.2	$N + Adj(0,0)$	42	Pres.	65.7	67.2	68.5	65.3
2	$N + Adj(0,0)$, $un._{(2,2)}$	783	Pres.	95.2	72.7	86.5	95.0
3	$OW_{(3,3)}$	1544	Pres.	92.4	73.2	86.3	90.7
4	$OW_{(3,3)}$, $un._{(2,2)}$	2285	Pres.	**95.7**	73.2	87.1	95.0
5.1	$Repr_2\ OT_9^m\ OW_{(3,3)}$	20	Pres.	50.6	63.2	68.5	68.1
5.2	$Repr_2\ OT_{45}^a\ OW_{(3,3)}$	92	Pres.	57.4	59.1	71.3	71.3
5.3	$Repr_2\ OT_{90}^a\ OW_{(3,3)}$	182	Pres.	60.8	61.0	68.4	71.9
5.4	$Repr_2\ OT_{385}^a\ OW_{(3,3)}$	772	Pres.	63.8	63.0	69.1	73.1
6.1	$Repr_2\ OT_9^m\ OW_{(3,3)}$, $un._{(2,2)}$	761	Pres.	94.6	73.0	87.0	94.9
6.2	$Repr_2\ OT_{385}^a\ OW_{(3,3)}$, $un._{(2,2)}$	1126	Pres.	94.3	62.8	85.7	**95.2**
7	dev (Top_{min}^{300}, Top_{max}^{300})	600	Pres.	74.4	56.7	73.1	73.3
8	dev (Top_{min}^{300}, Top_{max}^{300}), $un._{(2,2)}$	1341	Pres.	94.8	76.4	86.4	94.7
9	dev, $Repr_2\ OT_{385}^a$, $un._{(2,2)}$, $OW_{(3,3)}$	3657	Pres	**95.6**	80.6	88.6	**95.9**
10	9, Signs	3659	Pres	**95.8**	80.6	88.9	**96.1**

auditorium, variant, age, question, eye, graph, drama, viewer, quantity, thought, plan, advantage, continuation, role, word, shooting, level, outcome, worth, emotion, event

Feature Representation. In the current section two types of the object-attribute matrix for machine learning are considered.

Representation 1. Attributes are opinion words

Representation 2. An object-attribute matrix is created for machine learning according to the given in Table 12.

Table 12. An object-attribute matrix

	OT^+	OT^-	OT_1^+	OT_1^-	OT_2^+	OT_2^-	...	OT_n^+	OT_n^-
Doc_1	a_{10}^+	a_{10}^-	a_{11}^+	a_{11}^-	a_{12}^+	a_{12}^-	...	a_{1n}^+	a_{1n}^-
Doc_2	a_{20}^+	a_{20}^-	a_{21}^+	a_{21}^-	a_{22}^+	a_{22}^-	...	a_{2n}^+	a_{2n}^-
...	...	...	...	...	...	...	...	...	...
Doc_m	a_{m0}^+	a_{m0}^-	a_{m1}^+	a_{m1}^-	a_{m2}^+	a_{m2}^-	...	a_{mn}^+	a_{mn}^-

Here OT^+ and OT^- are the number of positive and negative opinion words in the document Doc_i, OT_i^+ and OT_i^- are the number of occurrences of the opinion target OT_i in the positive and negative contexts respectively, where OT_i is the word from the list opinion targets. By the term "context of OT_i" we mean the surrounding radius of the word by 5 words. The context can be called positive (negative) if it contains a positive(negative) opinion word.

The best accuracy, 96.1%, was obtained using simultaneously all opinion words extracted by all the four methods together with unigrams and features involving opinion targets.

5 Conclusion and Future Work

In this paper, we considered several methods for automatic extraction of domain-specific opinion words. For one of these methods, double propagation [13], we proposed a new technique for building a seed to be used as a starting point for opinion word extraction. Syntactic rules of the double propagation method were adapted for Russian.

We used several machine learning methods for determining text polarity. For this, we developed several text representations using various features and compared them empirically. Our experiments showed that opinion words are important for determining the polarity.

Using all unigrams as the only features during classification, we were able to obtain the accuracy of 95.2%. Adding opinion words to the list of features made it possible to improve the accuracy up to 96.1%. The gain in accuracy may not seem large at first glance, but note that the classification error was reduced by almost 20%. However, our experiments showed that, although opinion words are useful for polarity detection, they are not sufficient on their own and should be used only in combination with other features. Our results are better than those described in [10] and [9], which report the accuracy of 82.9% and 89% respectively. It remains to see whether this difference is due to our choice of the dataset or if, perhaps, the polarity of texts in Russian is easier to determine than that of texts in English, at least for some genres.

Acknowledgments. The results of the project “Mathematical Models, Algorithms, and Software Tools for Intelligent Analysis of Structural and Textual Data”, carried out within the framework of the Basic Research Program at the National Research University Higher School of Economics in 2012, are presented in this work.

References

1. Chetviorkin, I., Loukachevitch, N.: Automatic extraction of domain-specific opinion words. In: Proceedings of “Dialogue” International Conference (2010) (in Russian)
2. Church, K.W., Hanks, P.: Word association norms, mutual information and lexicography. In: Proceedings of the 27th Annual Conference of the ACL (1989)
3. Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. In: Proceedings of ACL 1997, Stroudsburg, PA, pp. 174–181 (1997)
4. Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. In: Proceedings of SIGKDD 2004, Seattle, Washington, USA, pp. 168–177 (2004)
5. Kanayama, H., Nasukawa, T.: Fully automatic lexicon expansion for domain-oriented sentiment analysis. In: Proceedings of EMNLP 2006, pp. 355–363 (2006)
6. Kim, S.-M., Hovy, E.: Determining the sentiment of opinions. In: Proceedings of COLING 2004, pp. 1367–1373 (2004)
7. Kobayashi, N., Inui, K., Matsumoto, Y.: Extracting aspect-evaluation and aspect-of relations in opinion mining. In: Proceedings of EMLP 2007 (2007)
8. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Springer, Berlin (2011)
9. Mullen, T., Collier, N.: Sentiment analysis using support vector machines with diverse information sources. In: Proceedings of EMNLP 2004 (2004)
10. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of EMNLP 2002 (2002)
11. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Now–Publishers Inc., Hanover (2008)
12. Popescu, A.-M., Etzioni, O.: Extracting Product Features and Opinions from Reviews. In: Proceedings of EMNLP 2005 (2005)
13. Qiu, G., Liu, B., Bu, J., Chen, C.: Opinion Word Expansion and Target Extraction through Double Propagation. Computational Linguistics 37(1), 9.27 (2011)
14. Takamura, H., Takashi, I., Manabu, O.: Extracting semantic orientations of words using spin model. In: Proceedings of ACL 2005, pp. 133–140 (2005)
15. Wiebe, J., Wilson, T., Bruce, R., Bell, M., Martin, M.: Learning subjective language. Computational Linguistics 30(3), 277–308 (2004)
16. Zhang, L., Liu, B.: Identifying Noun Product Features that Imply Opinions. In: Proceedings of ACL 2011 (2011)