PARAMETRICAL WORDS IN THE SENTIMENT LEXICON

Dr. Elena G. Brunova, Head of the Department of Foreign Languages and Cross-Cultural Communication, Tyumen State University, Russia, Tyumen
E-mail: egbrunova@mail.ru

UDK: 811.161.1'37

Abstract: In this paper, the main features of parametrical words within a sentiment lexicon are determined. The data for the research are client reviews in the Russian language taken from the bank client rating; the domain under study is bank service quality. The sentiment lexicon structure is presented; it includes two primary classes (positive and negative words) and three secondary classes (increments, polarity modifiers, and polarity anti-modifiers). This lexicon is used as the main tool for the sentiment analysis carried out by two methods: the Naïve Bayes classifier and the REGEX algorithm.

Parametrical words are referred to as the words denoting the value of some domain-specific parameter, e.g. the client's time consuming. To distinguish the main features of parametrical words, the parameters relevant for the bank service quality domain are determined. The revised lexicon structure is proposed, with a new class (decrements) added. The results of the research demonstrate that parametrical words express implicit opinions, since parameters are not usually named directly in reviews. Only a small number of parametrical words can be ranged into the primary classes (positive or negative), but this ranging is domain-specific. It is the parameter that determines the domain specificity of such words. Most parametrical words are ranged into the secondary classes, and this ranging can be considered universal. The parametrical words denoting the increase of a parameter should be ranged into the increment class, as they intensify positive or negative emotions. The parametrical words denoting the decrease of a parameter should be ranged into the decrement class, as they reduce positive or negative emotions. The evident progress on the way to the sentiment lexicon universalization can be achieved by classifying parametrical words within the sentiment lexicon.

Key words: cognitive linguistics, natural language processing, sentiment analysis, lexicon, domain, parametrical words, increment, decrement.

1. INTRODUCTION

Sentiment analysis is one of the rapidly developing methods of natural language processing. The first works were published in early 2000s (Nasukawa & Yi, 2003; Pang et al., 2002; Turney, 2002; Wiebe et al., 2001), and since then much has been done in this field. Sentiment lexicons have been built; algorithms have been developed (Gamon et al., 2005; Hu & Liu, 2004; Liu, 2010; Manning et al., 2008; Pang & Lee, 2008). All these successful studies were focused on the English language, and it seemed logical to apply their results to other natural languages, translating the lexicons and modifying the tools for syntactic analysis. However, the attempts to build a universal sentiment lexicon, the principal sentiment analysis tool, failed.

A *sentiment lexicon* is a set of words which are used to express opinions and emotions in sentiment documents (reviews, etc.), it is generally divided into two classes: the positive and negative ones (Pang et al., 2002). After numerous experiments, it is evident that such a lexicon should be both language-specific, and domain-specific.

The problem of language specificity concerns the differences in the morphological structure of natural languages, while the problem of domain-specificity is a semantic one. Some words from sentiment lexicons appear domain-specific (Ganapathibhotla & Liu, 2008: 242), e.g. the word *long* can be ranged into the positive lexicon when evaluating the battery operation (the smartphone domain), but it can be ranged into the negative lexicon in evaluating the client's time consuming (the bank service quality domain). In this paper, such ambiguous words as *long* are called *parametrical words*.

Parametrical words are referred to as the words denoting the amount of some domain-specific parameter (battery life, the client's time consuming, etc.).

The purpose of this paper is to determine the main features of parametrical words within a sentiment lexicon.

2. MATERIALS AND METHODS

The data for the research are the client reviews in the Russian language on bank service quality from the bank client rating taken from (www.banki.ru). The domain under study is bank service quality. To build a sentiment lexicon, 20 reviews (10 positive and 10 negative ones) were randomly selected. From this content, the seed, i.e. basic, lexicon containing 100 words was constructed manually. Then the seed lexicon was extended up to about 500 words, using synonyms, antonyms, and the sentiment consistency technique (Liu, 2010). This technique first proposed in (Hatzivassiloglou & K. McKeown, 1997) uses a list of seed opinion adjective words and a set of linguistic constraints (*and, but, either-or, neither-nor*) to identify other opinion words and their polarity. For instance, in the sentence *This i-phone is beautiful and easy to use*, if *beautiful* is known to be positive, it can be inferred that *easy* is also positive. On the contrary, in the sentence *This i-phone is beautiful, but expensive*, if *beautiful* is known to be positive, it can be inferred that *expensive* is negative. The seed words with the linguistic constraints were entered to the Google search engine with the search limitation within (www.banki.ru).

All the words in the lexicon were stemmed for easier processing.

The structure of the lexicon is presented in Table 1.

Table 1. The structure of the sentiment lexicon (bank service quality)

Lexicon classes				
Primary classes		Secondary classes		
Positive	Negative	Increments	Polarity Modifiers	Polarity Anti-Modifiers
Безопасный (safe), бесплатный (free), вежливый (polite), компетентный (competent), четкий (clear), эффективный (efficient) …	Агрессивный (aggressive), безвыходный (hopeless), грубый (rude), досадный (annoying), обидный (offensive), трудный (difficult) …	Очень (very), совершенно (absolutely), никогда (never)* …	Не (no), нет (not), без (without) …	Так (so), такой (such) …

* In English lexicons, such words as *never*, *nobody*, etc. should be ranged into the polarity modifiers. In Russian lexicons, however, due to the occurrence of double negation in the Russian syntax, such words are not polarity modifiers, but increments.

As Table 1 demonstrates, the sentiment lexicon includes two primary classes: *positive* and *negative* words denoting positive and negative opinions, respectively. Besides, it includes three secondary classes: *increments*, *polarity modifiers*, and *polarity anti-modifiers*.

Increments are referred to as the words intensifying the polarity of the other words within a sentence without changing it into the opposite one, e.g. in the contexts *Это очень надежный банк*. (This is a very reliable bank) and *Это очень плохие условия кредита* (These are very poor credit terms), the word *очень* (very) is an increment which intensifies the positive and negative opinions, respectively.

Polarity modifiers are referred to as the words which change the polarity of the other words within a sentence into the opposite one, e.g. in the context *Сами работники банка не грубые и не злые* (The bank operators themselves are not rude and aggressive) the positive opinion is expressed, though the context includes negative words *грубые* (rude), *злые* (aggressive); the word *не* (not) is a polarity modifier which changes their polarity into the positive one.

Polarity anti-modifiers are referred to as the words which cancel the change in the polarity in spite of the occurrence of polarity modifiers within a sentence. Compare two contexts: 1) *Меня никогда не обманывали* (I have never been cheated) 2) *Меня никогда так не обманывали* (I have never been cheated in such a way). In spite of almost complete similarity of the words, these contexts express opposite opinions: the positive one and the negative one, respectively. The difference is that in the first context, the word *никогда* (never) implies *never in this bank*, and in the second one it implies *never except this bank*. The word *так* (such) is a polarity anti-modifier, which cancels the change in the polarity in the second example, and it remains negative, as the context contains the negative word *обманывали* (cheated).

To carry out the sentiment analysis, the REGEX algorithm was developed. The algorithm included 11 formal grammar rules and the corresponding syntactic models, being a sort of regular expressions which detect certain text elements, simplify each sentence, and present the text as a formal model. One of these rules is presented below.

Rule 1. If between the beginning of the sentence, or a punctuation mark, or a conjunction (*and/or*) and the next punctuation mark, or a conjunction (*and/or*), or the end of the sentence, there is a polarity modifier, then the polarity of all the words referred to the sentiment lexicon within this segment is changed into the opposite one. The sequence of the elements (a polarity modifier, a positive/negative word, any other word) does not matter.

When formalized, the rule can be presented as below:

<S>|<Z>|& {ALT, *, Any POS} <Z>|&|</S>|<!/S>|<?/S>|<?!/S> →

→ <S>|<Z>* Any NEG *<Z>|&|</S>|<!/S>|<?/S>|<?!/S> →

→ nNEG → -n

where <S> is the beginning of the sentence;

| is the divisor of equally allowable elements,

<Z> is a punctuation mark;

& is the and/or conjunction;

ALT is a polarity modifier;

POS is a word from the positive lexicon,

NEG is a word from the negative lexicon;

* is any other word;

{A, B, C} is the group of the elements which can follow in any sequence;

Any is any number of elements;

</S>, <!/S>, <?/S>, <?!/S> are the ends of the sentence with a full point, an exclamation mark, an interrogation mark, or both, respectively;

The REGEX algorithm included successive application of the substitution rules according to the priorities obtained from the experiments with the documents from the training set. For example, the application of Rule 1 resulted in the following layout conversion:

Платежи проходят очень быстро, деньги не зависают. (The payments are processed quickly, money doesn't become hung)

<S>*POS <Z> * ALT NEG </S>→ <S>*POS <Z> * POS </S> →

→ 2POS → +2

At a certain step of the algorithm, the number of the POS and NEG wildcards was calculated in each sentence, then the draft sentence polarity was calculated (+2 in the example above). The group of the rules to correct the draft polarity was also applied. The output of the REGEX algorithm was the calculation of the document polarity normalized to the number of the words in the document.

To carry out the automatic sentiment analysis of the reviews, the SENTIMENTO system was implemented as an Internet application with an interface for the model testing and its adjustment (Brunova, E.G., Bidulya Yu.V. (2014) Algorithm with Formal Grammar Elements for Sentiment Analysis. *Tyumen State University Herald*. No. 1, in press). Fig. 1 demonstrates the window of the sentiment analysis module with the conclusion of the system.

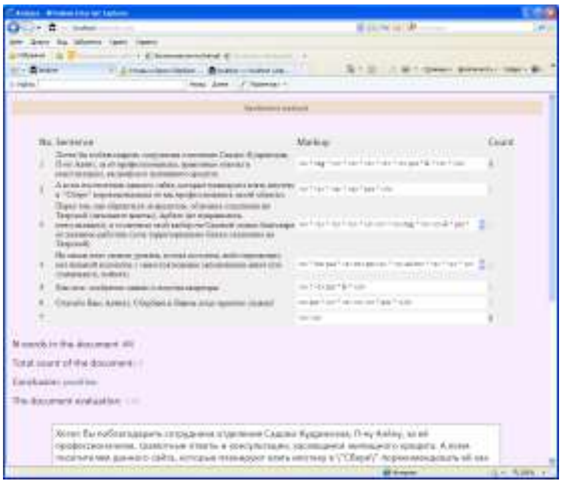

Fig.1. The *SENTIMENTO* software. The sentiment analysis module.

The system provides the opportunity for its users to confirm or reject the system conclusion, for this purpose, *Your conclusion* request is displayed with two buttons (*Positive* and *Negative*). The interface for entering human conclusions is presented in Fig.2. After the user presses a button, the system checks if the human conclusion matches the system one. In case it does, the document is included into the database. Besides, these results are used to calculate the system efficiency.

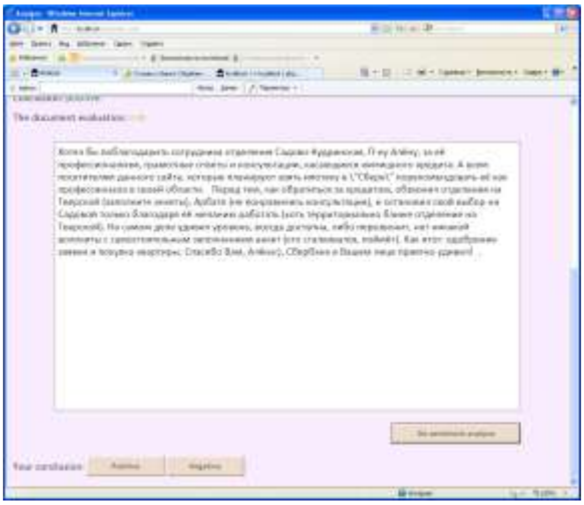

Fig.2. The interface for entering human conclusions

The efficiency of the proposed algorithm was evaluated in comparison with the efficiency of the Naïve Bayes Classifier (Webb et al., 2005).

The sentiment analysis experiments with the SENTIMENTO software revealed a number of problems, in particular, concerning parametrical words. For instance, a user evaluated the context *Предлагают маленький процент по вкладу* (A small deposit interest was offered) as negative, while the system evaluated it as neutral, since it did not detect any negative words in it. As for the context *Очередь была совсем маленькая* (The queue was quite small), the human conclusion was positive, while the system conclusion was negative, as it detected a negative word *очередь* (queue).

Thus, the behavior of parametrical words in reviews differs from that of negative or positive words, and ignoring this fact leads to incorrect analysis results.

3. RESULTS AND DISCUSSION

Researchers notice that some words, e.g. *очень* (very), *совершенно* (absolutely), *долго* (*long*), *медленно* (slowly), demonstrate their ambiguous nature in the process of sentiment analysis. N. Lukashevich and I. Chetverkin distinguish operators affecting the semantic polarity, however, their operators include rather negation words (*не* (not), *нет* (no)) or adjective increments (*очень* (very), *самый* (most, least)), than adjectives themselves (Lukashevich & Chetverkin, 2011: 77). Nevertheless, the adjectives, adverbs, and even nouns (e.g. *максимум* maximum) expressing the amount of a parameter could be included into the sentiment lexicon. Such words express the intensity of the domain attribute, or parameter, e.g. the client's time saving.

Depending on the parameter, a positive or negative opinion can be expressed, while it increases or decreases. For instance, the word *high* spoken or written about the speed of service (the parameter is the client's time saving) is evaluated as positive, but the word *high* spoken or written about the price or credit interest (the parameter is the client's money costs) is evaluated as negative. It is the parameter that determines the domain specificity of such lexicon units.

To determine the main features of parametrical words, the contexts containing the words meaning *large*, *small*, *long*, *short*, *maximum*, *minimum*, etc. were extracted from the corpus of the 70 client reviews randomly selected from (www.banki.ru). The study of these contexts enabled the domain-specific parameters to be determined.

Consider the parameters relevant for the bank service quality domain, below a context per each parameter is cited, the parametrical words are underlined, the translation into English is given in brackets:

Positive opinions

1) Increase in the parameter

a) The client's positive emotions: *хочется отметить оперативность в работе и готовность оказать максимум помощи даже потенциа-льным клиентам* (I'd like to emphasize the speed of operation and the readiness to offer the maximum of help even to potential clients)

b) The client's cost saving: *Карта с немалым лимитом* (The card with a considerable limit)

c) The client's time saving: *Наш кредит одобрили очень быстро* (Our credit was approved very fast)

d) The sufficiency of service information: *Много информации, листовки, плакаты с рекламой* (There is a lot of information, there are advertising leaflets and posters)

2. Decrease in the parameter

a) The client's negative emotions: *небольшой список замечаний* (short list of remarks)

b) The client's money costs: *маленький процент по кредиту* (low credit interest)

c) The client's time consuming: *Очередь была совсем маленькая* (The queue was quite small)

Negative opinions

1) Increase in the parameter

a) The client's negative emotions: *хитрости для большого обмана* (tricks for a great fraud)

b) The client's money costs: *Я и так плачу немалый процент за пользование кредитом* (Anyway, I pay a considerable credit interest)

c) The client's time consuming: *Банк для тех, у кого много лишнего времени* (The bank is for those who have much spare time)

2. Decrease in the parameter

a) The client's positive emotions: *толку мало* (little use.)

b) The client's cost saving: *Лимит маленький* (The credit limit is small.)

c) The client's time saving: *платежи проходят медленно* (The payments are processed slowly)

d) The sufficiency of service information: *информации мало* (there is little information)

The extracted parameters are summarized in the diagram (Fig. 3).

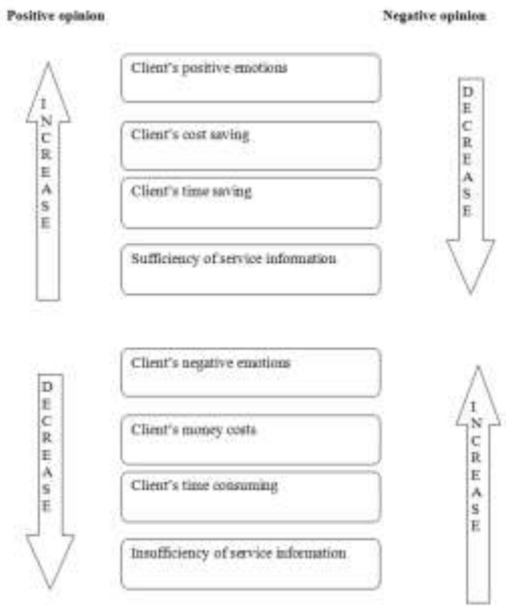

Fig. 3. The parameters of the sentiment analysis for bank service quality extracted from the review contexts

As it can be seen from the parameters determined (Fig. 3), the increase of a certain parameter results in a negative or a positive opinion, and the decrease of the same parameter results in the opposite opinion. For instance, the increase in the client's time saving evokes positive emotions and results in positive opinions, its decrease results in negative emotions and negative opinions. On the other hand, the increase in the client's money costs results in negative emotions and negative opinions, its decrease evokes positive emotions and results in positive opinions. Thus, parametrical words are not only domain-specific, but they demonstrate their ambiguous nature even within a single domain. This is confirmed by their occurrence within the same, mainly negative, context, cf. *Много слов, но мало дела* (There are many words, but little work) *Дают быстро, отдают долго* (They give quickly, but return slowly) *Большой минус и маленький плюс* (A large minus and a small plus).

The results of the analysis demonstrate that ignoring parametrical words in sentiment analysis results in incorrect conclusions, so they should be included into the sentiment lexicon and ranged into one of its classes.

Only a small number of parametrical words can be ranged into the primary classes, e.g. the word *быстро* (fast) is ranged into the positive class, and the words *долго* (long) and *медленно* (slowly) are ranged into the negative one; this ranging is definitely domain-specific.

The parametrical words denoting the increase of a parameter (meaning *large*, *many*, *much*, *maximum*, etc.) should be ranged in the increment class along with the words meaning *very*, *absolutely*, etc., as they intensify positive or negative emotions. As it was mentioned above, increments are the words intensifying the polarity of the other words within a sentence without changing it into the opposite one. The parametrical words denoting the decrease of a parameter (meaning *small*, *little*, *few*, *minimum*, etc.) should be ranged in a new class which may be referred to as the decrement class. *Decrements* are the words decreasing the polarity of the other words within a sentence without changing it into the opposite one. Thus, most parametrical words are ranged into the secondary classes; this means that they do not express the direct opinion, but affect the intensity of the opinion expressed by other words.

The revised structure of the sentiment lexicon is presented in Table 2, the parametrical words are underlined.

Table 2. The revised structure of the sentiment lexicon (bank service quality)

Lexicon classes					
Primary classes		Secondary classes			
Positive	Negative	Increments	Decrements	Polarity Modifiers	Polarity Anti-Modifiers
Безопасный (safe), бесплатный (free), вежливый (polite), компетентный (competent), четкий (clear), эффективный (efficient), быстро (fast) …	Агрессивный (aggressive), безвыходный (hopeless), грубый (rude), досадный (annoying), обидный (offensive), трудный (difficult), долго (long), медленно (slowly)…	Очень (very), совершенно (absolutely), никогда (never), нигде (nowhere), много (much, many), максимум (maximum), большой (large), высокий (high) …	Мало (little, a few), минимум (minimum), маленький (small), низкий (low)	Не (no), нет (not), без (without) …	Так (so), такой (such) …

4. CONCLUSION

The general features of parametrical words within the sentiment lexicon are determined. The structure of the sentiment lexicon is revised; a new class (decrements) is added.

The results of this research demonstrate that the behavior of most parametrical words in reviews differs from that of negative or positive words, and ignoring this fact results in incorrect sentiment analysis results. Parametrical words generally express the implicit opinion: they do not express the opinion directly, but affect the intensity of the opinion expressed by other words. Besides, the parameters themselves are not usually named directly in reviews.

Parametrical words should be included into the sentiment lexicon as follows:

1) A small number of parametrical words can be ranged into the primary classes (positive or negative), but this ranging is domain-specific. It is the parameter that determines the domain specificity of such words.

2) Most parametrical words are ranged into the secondary classes (increments or decrements), and this ranging can be considered universal.

Thus, the evident progress on the way to the sentiment lexicon universalization can be achieved by classifying parametrical words within the sentiment lexicon.

REFERENCES

Brunova, E.G. (2012). Metodika Sostavleniya Otsenochnogo Leksikona dlya Kontent-Analiza Mneniy (Technique of Constructing a Sentiment Lexicon) *Language and Science*. No. 1. (Online). Available: http://www.utmn.ru/docs/9317.pdf (in Russian)

Gamon M., et al. (2005). Pulse: Mining Customer Opinions from Free Text. *Proc. of the 6th International Symposium on Intelligent Data Analysis (IDA).* P. 121-132.

Ganapathibhotla, M., Liu B. (2008). Mining Opinions in Comparative Sentences. *Proc. of the 22nd*

International Conference on Computational Linguistics. Manchester. P. 241–248.

Hatzivassiloglou V., McKeown K. (1997). Predicting the Semantic Orientation of Adjectives. *Proc. of the 35th Annual Meeting of ACL*, Madrid. P. 174-181.

Hu M., Liu B. (2004). Mining and summarizing customer reviews. *Proc. of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. P. 168-177.

Liu, B.(2010). Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing, Second Edition*. (Online). Available: http://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf

Lukashevich, N.B., Chetverkin I. I. (2011). Izvlecheniye i Ispolsovaniye otsenochnykh Slov v Zadache Klassifikatsii Otzyvov na Tri Klassa (Extracting and Appliction of Sentiment Words in the Task of Three-Class Review Classification). *Vychislitelnye Metody i Programmirovaniye*. Vol. 12. P. 73-81. (in Russian).

Manning C., Raghavan P, Schütze H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press. 544 p.

Nasukawa T., Yi J. (2003). Sentiment Analysis: Capturing Favorability Using Natural Language Processing. *Proc. of the 2nd International Conference on Knowledge Capture*. Florida. P. 70-77.

Pang B., Lee L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*. Vol. 2, No 1-2. P. 1–135.

Pang B., Lee L., Vaithyanathan S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proc. of EMNLP*. (Online). Available: http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf

Turney P. (2002). Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proc. of the 40th Annual Meeting on Association for Computational Linguistics*. P. 417-424.

Webb, G. et al. (2005). Not So Naive Bayes: Aggregating One-Dependence Estimators. *Machine Learning*. 58. P. 5-24.

Wiebe J., Wilson T., Bell M. (2001). Identifying Collocations for Recognizing Opinions. *Proc. of ACL/EACL 01 Workshop on Collocation*.

www.banki.ru