SENTIRUEVAL: ТЕСТИРОВАНИЕ СИСТЕМ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ ПО ОТНОШЕНИЮ К ЗАДАННОМУ ОБЪЕКТУ

Лукашевич Н. В. (louk_nat@mail.ru)[1],
Блинов П. Д. (blinoff.pavel@gmail.com)[2],
Котельников Е. В. (kotelnikov.ev@gmail.com)[2],
Рубцова Ю. В. (yu.rubtsova@gmail.com)[3],
Иванов В. В. (nomemm@gmail.com)[4],
Тутубалина Е. (tlenusik@gmail.com)[4]

[1]МГУ им. М. В. Ломоносова, Москва, Россия;
[2]Вятский государственный гуманитарный университет, Киров, Россия;
[3]Институт систем информатики им. А. П. Ершова СО РАН, Новосибирск, Россия;
[4]Казанский федеральный университет, Казань, Россия

Статья описывает данные, правила и результаты SentiRuEval — тестирования систем автоматического анализа тональности русскоязычных текстов по отношению к заданному объекту или его свойствам. Участникам были предложены два задания. Первое задание было аспектно-ориентированный анализ отзывов о ресторанах и автомобилях; основная цель этого задания была найти слова и выражения, обозначающие важные характеристики сущности (аспектные термины), и классифицировать их по тональности и обобщенным категориям. Второе задание заключалось в анализе влияния твитов на репутацию заданных компаний. Такие твиты могут либо выражать мнение пользователя о компании, ее продукции или услугах, или содержать негативные или позитивные факты, которые стали известны об этой компании.

Ключевые слова: анализ тональности текстов, оценка качества, разметка коллекций, оценочные слова

SENTIRUEVAL: TESTING OBJECT-ORIENTED SENTIMENT ANALYSIS SYSTEMS IN RUSSIAN

Loukachevitch N. V. (louk_nat@mail.ru)[1],
Blinov P. D. (blinoff.pavel@gmail.com)[2],
Kotelnikov E. V. (kotelnikov.ev@gmail.com)[2],
Rubtsova Y. V. (yu.rubtsova@gmail.com)[3],
Ivanov V. V. (nomemm@gmail.com)[4],
Tutubalina E. (tlenusik@gmail.com)[4]

[1]Lomonosov Moscow State University, Moscow, Russia;
[2]Vyatka State Humanities University, Kirov, Russia;
[3]A. P. Ershov Institute of Informatics Systems, Novosibirsk, Russia;
[4]Kazan Federal University, Kazan, Russia

The paper describes the data, rules and results of SentiRuEval, evaluation of Russian object-oriented sentiment analysis systems. Two tasks were proposed to participants. The first task was aspect-oriented analysis of reviews about restaurants and automobiles, that is the primary goal was to find word and expressions indicating important characteristics of an entity (aspect terms) and then classify them into polarity classes and aspect categories. The second task was the reputation-oriented analysis of tweets concerning banks and telecommunications companies. The goal of this analysis was to classify tweets in dependence of their influence on the reputation of the mentioned company. Such tweets could express the user's opinion or a positive or negative fact about the organization.

Keywords: sentiment analysis, users review, collection labeling, aspect words, evaluation

1. Introduction

During last years the task of automatic sentiment analysis of natural language texts, that automatic extraction of opinions expressed in texts, attracts a lot of attention of researchers and practitioners. This is due to the fact that this task has a lot of useful applications. So the analysis and representation of users' opinions about products and services are of interest to their producers and competitors as well as to new users. Social opinion processing is important for authorities for better government.

The initial approaches to automatic sentiment analysis tried to determine the overall sentiment of the whole texts or sentences (Pang et al., 2002). This level of analysis presupposes that each document expresses opinions on a single entity (for example, a single product). Later, the task of object-oriented sentiment analysis appeared, when the system should reveal sentiment towards a specific entity mentioned in the text (Amigo et al., 2012; Jiang et al., 2011).

Finally, an author of a text can have different opinions relative to specific properties (or aspects) of an entity. To reveal these opinions, so called aspect-based sentiment analysis should be fulfilled (Liu, 2012; Bagheri et al., 2013; Glavaš et al., 2013; Popescu, Etzioni, 2005; Zhang, Liu, 2014). Aspects are expressed in texts with aspect terms and usually can be classified into categories. For example, "Service" aspect category in restaurant reviews can be expressed such terms as *staff*, *waiter*, *waitress*, *server*.

Automatic sentiment analysis is a complex problem of natural language processing. Several evaluation initiatives were devoted to study the best methods in sentiment analysis and related applications. These initiatives include Blog Track within TREC conference (Macdonald et al., 2010), TAC Opinion QA Tasks (Dang, Owczarzak, 2008), opinion tracks at NTCIR conferences (Seki et al., 2008), reputation management tracks at CLEF conference (Amigo et al., 2012), Twitter and review sentiment analysis tasks within SemEval initiative (Nakov et al., 2013; Rosenthal et al., 2014), etc.

In this paper we present results of SentiRuEval evaluation focusing on entity-oriented sentiment analysis of Twitter and aspect-oriented analysis of users' reviews in Russian. This evaluation is the second Russian sentiment analysis evaluation event in Russian after ROMIP sentiment analysis tracks in 2011–2013. This year in SentiRuEval we had two types of tasks. The first task is aspect-oriented sentiment analysis of users' reviews. The data included reviews about restaurants and automobiles. The second task was object-oriented sentiment analysis of Russian tweets concerning two varieties of organizations: banks and telecommunications companies.

The structure of this paper is as follows. In Section 2 we consider related evaluation initiatives in sentiment analysis. Section 3 describes tasks, data and principles of labeling in aspect-based review analysis. Section 4 describes the data and the task in the entity-oriented sentiment analysis of Twitter. Section 5 discusses results obtained by participants.

2. Related work

Several evaluation initiatives were devoted to sentiment analysis tasks similar to current SentiRuEval evaluation.

Last years in the framework of SemEval conference two types of sentiment analysis evaluations have been organized: sentiment analysis in Twitter and aspect-based sentiment analysis of reviews. In the Twitter task one of the subtasks was a message-level task, that is participating systems should classify if the message has positive, negative, or neutral sentiment (Nakov et al., 2013; Rosenthal et al., 2014). The task is directed to reveal, namely, the author opinion in contrast to neutral or objective information.

In the framework of CLEF initiative (http://www.clef-initiative.eu/) in 2012–2014 RebLab evaluations devoted to monitoring of reputation-oriented tweets were organized. The tasks included the definition of the polarity for reputation classification. The goal was to decide if the tweet content has positive or negative implications for the company's reputation. The organizers stress that the polarity for reputation is substantially different from standard sentiment analysis that should differentiate subjective

from objective information. When analyzing polarity for reputation, both facts and opinions have to be considered to determine what implications a piece of information might have on the reputation of a given entity (Amigo et al., 2012; Amigo et al., 2013).

Evaluation of aspect-based review analysis at SemEval was organized in 2014 for the first time (Pontiki et al., 2014). The dataset included isolated, out of context sentences (not full reviews) in two domains: restaurants and laptops. 3K sentences were prepared for training in each domain. Set of aspect categories for restaurants included: *food, service, price, ambience, anecdotes/miscellaneous*.

In 2015 SemEval evaluations the aspect-based sentiment analysis of reviews (http://alt.qcri.org/semeval2015/task12/) is focused on entire reviews. Aspect categories of terms became more complicated and now consist of Entity-Attribute pairs (E#A). The E#A inventories for the restaurants domain contains 6 Entity types (RESTAURANT, FOOD, DRINKS, SERVICE, AMBIENCE, LOCATION) and 5 Attribute labels (GENERAL, PRICES, QUALITY, STYLE_OPTIONS, MISCELLANEOUS). The Laptops domain contains 22 Entity types and 9 Attribute labels.

In 2011–2013 two evaluation events of Russian sentiment analysis systems were organized. The first evaluation was devoted to extraction of overall sentiment of users' reviews in three domains: movies, books and digital cameras. For training, reviews from recommendation services were granted to participants. The evaluation was fulfilled on blog posts extracted with the help of the Yandex blog service (Chetviorkin et al., 2012). The second evaluation offered two new tasks for participants, namely: extraction of the overall sentiment of quotation (direct or indirect speech) from news articles and sentiment-oriented information retrieval in blogs when for a query (from the abovementioned domains) user opinions in blog posts should be found (Chetviorkin, Loukachevitch, 2013).

3. Ways to express opinions about aspects

Aspect terms also can be subdivided into several categories. They can be classified into three subtypes: **explicit aspects**, **implicit aspects** and **sentiment facts**.

Explicit aspects denote some part or characteristics of a described object such as *staff, pasta, music* in restaurant reviews. Explicit aspects are usually nouns or noun groups, but in some aspect categories we can meet explicit aspects expressed as verbs. For example, in restaurants the important characteristics of the service quality is time of order waiting, so this characteristic can be mentioned with verb *wait* (*ждать*): *ждали больше часа—waited for more than an hour.*

Implicit aspects are single words or single words with sentiment operators that contain within themselves as specific sentiments as the clear indication to the aspect category. In restaurant reviews the frequent implicit aspects are such words as *tasty* (*positive+food*), *comfortable* (*positive+interior), not comfortable* (*negative+interior*). The importance of these words for automatic systems consists in that fact that implicit aspects allow a sentiment system to reveal user's opinion about entity characteristics even if an explicit aspect term is unknown, written with an error or referred in a complicated way.

Sentiment facts do not mention the user sentiment directly, formally they inform us only about a real fact, however, this fact conveys us a user's sentiment as well as the aspect category it related to. For example, sentiment fact *отвечала на все вопросы* (*answered all questions*) means positive characterization of the restaurant service; this expression is enough frequent in restaurant reviews.

In the SentiRuEval labeling we annotated these three subtypes of aspect terms and our tasks for participants were not only to extract explicit aspect terms but also to extract all aspect terms (see Section 4).

An opinion about aspects can be expressed in several ways.

The **direct way of conveying the opinion** is through using opinion words such as *good*, *bad*, *excellent*, *awful*, *like*, *hate*, etc.

Opinions can be formulated as **comparisons** with other entities, previous cases or opinions of other people (Liu, 2012; Jindal, Liu, 2006). The problem of automatic analysis in these cases arise because used positive or negative words can be not relevant to the current review. In addition, comparison can be delivered in various ways not only using comparative constructions. For example, in the following extract from a restaurant review the comparison is marked with word *another,* and positive words *enjoyed* and *wonderful* characterize a restaurant distinct from the restaurant under review:

> *We decided not to have dessert and coffee there, but instead went to another restaurant where we **enjoyed** a **wonderful** end to our evening.*

We can formulate our opinion as **recommendation** (the constructive or suggestive opinion—see (Arora, Srinivasa, 2014)) or description of a **desirable situation** or characteristics of an entity, so called *irrealis factors* (Taboada et al., 2011; Kusnetsova et al., 2013). In these cases mentioned positive words can conceal the negative opinion.

At last, the opinion can be expressed with means of **irony or sarcasm** (Barbieri, Saggion, 2014; Riloff et al., 2013). In such cases the opinion can look like positive or at least medium one, but in fact it is strongly negative as in the following example: "***Excellent** translation, I don't understand anything*".

In the SentiRuEval labeling we marked these subtypes of opinions for further research (see Section 4).

4. Labeling and tasks of aspect-based analysis of reviews at SentiRuEval

For evaluation of aspect-oriented sentiment analysis systems we chose two domains: restaurant reviews and automobile reviews. In restaurant reviews aspect categories include: FOOD, SERVICE, INTERIOR (including atmosphere), PRICE, GENERAL. For automobiles aspect categories are: DRIVEABILITY, RELIABILITY, SAFETY, APPEARANCE, COMFORT, COSTS, GENERAL.

The length of reviews can vary drastically from one brief sentence to a long narrative. There can be also shifts to one or the other particular aspect. As an experiment, for labeling in the restaurant domain we tried to extract the most typical reviews

from our collection. To achieve it, the following procedure was performed. We represented each review as a bag-of-word vector and calculated the global collection's vector by averaging all the individual vectors. Then we imposed restrictions on min and max review length and chose most similar reviews according to the cosine similarity between global vector and single review vectors. As a result, most typical review representatives were selected for the labeling.

The labeling of training and test data was conducted with BRAT annotating tool (Stenetorp et al., 2012). Annotators had access to review collections through web interface. To unify and agree the annotation procedure, an assessor manual was prepared[1]. It is based on the SemEval-2014 (Pontiki et al., 2014) annotation guidelines.

The annotation task was to mark up two main types of tokens: aspect terms within a review and aspect categories attached to whole reviews. The aspect categories were labeled with the overall score of sentiment expressed in the text: positive, negative, both or absent.

According to the above-described categorization of opinions and aspect terms, the annotation of aspect terms within a text included several dimensions:

1. At first annotators should indicate explicit aspects, implicit aspects or sentiment facts in review texts and assign them their relevant type (explicit, implicit or fact).
2. All aspects terms should be assigned to aspect categories of the target entity.
3. Annotators marked the polarity of the aspect term: positive, negative, neutral, or both.
4. Annotators marked the relevance of the term to the review:
    a. *Rel—relevant* (to the current review),
    b. *Cmpr—comparison*, that is the term concerns another entity,
    c. *Prev—previous*, that is the term is related to previous opinions,
    d. *Irr—irrealis*, that is the term is the part of a recommendation or description of a desirable situation,
    e. *Irn—irony.*

So, for example, the annotation of word *девушка* (*girl*) in context *милая девушка* (*nice girl*) in a restaurant review includes sentiment orientation—*positive*, aspect category—*service*, aspect mark—*relevant,* aspect type—*explicit.*

Such detailed annotation process is very labor consuming. Therefore, each review was labeled only by a single assessor. However, to check the quality of aspect labeling two procedures were fulfilled after the labeling was finished. First, all labeled aspect terms were extracted from the markup according to their types and categories and were looked through; so some accidental mistakes were found and corrected. Second, we compared the aspect sentiment assigned to the review as a whole and sentiments of specific terms within this review. In cases of the differences between these two types of labeling the markup of the review was additionally verified.

During the annotation procedure, no balancing according to sentiment or aspect terms was performed; we tried to keep natural distributions specific for reviews in a given domain. Some statistics about relevant terms (Rel) are shown in Table 1.

---

[1] The manual is available at http://goo.gl/Wqsqit.

Table 1. Corpus statistics

		Restaurants		Automobiles	
		Train	Test	Train	Test
Number of reviews		201	203	217	201
Number of terms which are	explicit	2,822	3,506	3,152	3,109
	implicit	636	657	638	576
	fact	523	656	668	685
Number of terms which are	positive	2,530	3,424	2,330	2,499
	negative	684	865	1,337	1,300
	neutral	714	445	691	456
	both	53	85	100	115

The labeled data allowed us to offer the following tasks to the participants:

- **Task A**: automatic extraction of explicit aspects,
- **Task B**: automatic extraction of all aspects including sentiment facts,
- **Task C**: extraction of sentiments towards explicit aspects,
- **Task D**: automatic categorization of explicit aspects into aspect categories,
- **Task E**: sentiment analysis of the whole review on aspect categories.

To evaluate automatic systems the following quality measures were utilized.

For task A and B we applied macro F1-measure in two variants: exact matching and partial matching. Macro F1-measure means in this case calculating F1-measure for every review and averaging the obtained values.

To measure partial matching for every gold standard aspect term t we calculate precision and recall in the following way:

$$\text{Precision}_t = \frac{|t \cap t_s|}{|t_s|},$$

$$\text{Recall}_t = \frac{|t \cap t_s|}{|t|},$$

where t_s is an extracted aspect term that intersects with term t, $t \cap t_s$ is the intersection between terms t and t_s, $|t|$ is the length of the term in tokens. So F1-measure is calculated for every term and then we average the values for all gold standard terms.

For sentiment classification of aspect terms (task C) both variants of F1-measure (macro- and micro-) were utilized. Calculation of macro F1-measure is based on separate calculation of precision, recall, and F-measure for every category under consideration, then the obtained values are averaged. This allows us to evaluate the quality of categorization equally for every category. Micro F1-measure is calculated on the global confusion matrix, this measure greatly depends on the disbalance in the class distribution.

For aspect categorization of terms (task D) and the sentiment analysis of whole reviews (task E) macro F1-measure was used.

Table 2. Results in aspect-oriented review analysis (Restaurant domain)

Task	Measure	Baseline	Participants' results	Participant identifier
A	Exact matching, Macro F	0.608	**0.632**	**2**
			0.627	1
A	Partial matching, Macro F	0.665	**0.728**	**4**
			0.719	1
B	Exact matching, Macro F	0.587	**0.600**	**1**
			0.596	2
B	Partial matching, Macro F	0.619	**0.668**	**1**
			0.645	1
C	Macro F	0.267	**0.554**	**4**
			0.269	3
C	Micro F	0.710	**0.824**	**4**
			0.670	3
D	Macro F	0.800	**0.865**	**8**
			0.810	4
E	Macro F	0.272	**0.458**	**4**
			0.372	10

For all tasks we prepared baseline runs. The baseline system for tasks A and B extracts the list of labeled terms from the training collection, lemmatizes them and apply them to the lemmatized representation of the test collection. If more than one term matches the same word sequence, then a longer term is preferred.

The task C and D baseline systems attribute an aspect term to its most frequent category in the training collection. If a term is absent in the training collection then the most frequent aspect category is applied. The task E baseline is the most frequent sentiment category for the given aspect category (positive in all cases).

Altogether 12 participants with 21 runs were participated in the review sentiment analysis tasks. Due space limitations here we represent only two best results in each task and only primary F-measure, the full results are available at http://goo.gl/Wqsqit. Table 2 presents the participants' results for restaurant reviews, Table 3 contains the results for automobile reviews. Automobile reviews obtained much less attention from participants.

From the Tables 2, 3 it can be seen that the baselines for extracting aspect terms (tasks A and B) are quite high, which means the considerable agreement between annotation of training and testing collections. The best methods in these tasks were based on distributional approaches augmented with a set of rules (participant 4) and recurrent neural nets (participant 1). For the exact aspect matching, the best results were achieved by sequence labeling with SVM on the rich set of morphological, syntactic and semantic features (participant 2).

Table 3. Results in aspect-oriented review analysis (Automobile domain)

Task	Measure	Baseline	Participants' results	Participant identifier
A	Exact matching, Macro F	0.594	**0.676**	**2**
			0.651	1
A	Partial matching, Macro F	0.697	**0.748**	**1**
			0.730	2
B	Exact matching, Macro F	0.589	**0.636**	**2**
			0.630	1
B	Partial matching, Macro F	0.674	**0.714**	**1**
			0.704	1
C	Macro F	0.264	**0.568**	**4**
			0.342	1
C	Micro F	0.619	**0.742**	**4**
			0.647	1
D	Macro F	0.564	**0.652**	**8**
			0.607	4
E	Macro F	0.237	0.439	4

The best result in the analysis of sentiment towards aspect terms (task C) was obtained with Gradient Boosting Classifier (participant 4). The features were based on the skip-gram model exploiting word contexts for learning better vector representations and pointwise mutual information. In the task of categorization of explicit aspect terms (task D) the best results were obtained by SVM with features based on pointwise mutual information (participant 8). The second-place result is obtained by the method relying on the term similarity in the space of distributed representations of words (participant 4). For task E the best results were achieved by integration of the results obtained in tasks A, C and D (participant 4).

5. Object-oriented sentiment analysis of tweets

The goal of Twitter sentiment analysis at SentiRuEval was to find sentiment-oriented opinions or positive and negative facts about two types of organizations: banks and telecom companies. This task is quite similar to the reputation polarity task at RepLab evaluation (Amigo et al., 2013).

The training and test tweet collections were provided with fields corresponding all possible organizations for that tweets were extracted. A concrete organization mentioned in a given tweet was indicated with "0" label, denoting "neutral" as a default value. Annotators and participating systems should to leave this value unchanged if the tweet was considered as neutral or replace the value with "1" (positive) or "−1" (negative). The annotators also could label tweets with "−−", which means =meaningless=, or with "+−", which means positive and negative sentiments in the same tweet. Both latter cases were excluded from evaluation.

For training and testing collections assessors labeled 5,000 tweets in each domains (20000 tweets were labeled altogether). It is important to stress, that the training and testing collections were issued during different time intervals. The tweets of the training collection were written in 2014, the tweets of the testing collection were published in 2013.

Table 4. Results of the voting procedure in labeling of the tweet testing collection

Domain	The number of tweets with the same labels from at least 2 assessors	Full coincidence of labeling	The final number of tweets in the testing collection
Banks	4,915 (98.30%)	3,816 (76.36%)	4,549
Telecom companies	4,503 (90.06%)	2,233 (44.66%)	3,845

Analyzing the markup of the training collection we found that the estimation of some tweets can arise considerable discussion on their sentiment. To lessen the subjectivity of labeling and also accidental mistakes the testing collection was labeled by three assessors, and the voting scheme was applied to obtain the results of manual labeling. Finally, from the collection irrelevant tweets were removed. Results of the preparing the collection are presented in Table 4.

The participating systems were required to perform a three-way classification of tweets: positive, negative or neutral. As the main quality measure we used macro-average F-measure calculated as the average value between F-measure of the positive class and F-measure of the negative class. So we ignored Fneutral because this category is usually not interesting to anybody. But this does not reduce the task to the two-class prediction because erroneous labeling of neutral tweets negatively influences on Fpos and Fneg. Additionally micro-average F-measures were calculated for two sentiment classes.

Table 5. Results of participants in tweet classification tasks. The identifiers of participants in review and Twitter tasks are different

Domain	Measure	Baseline	Participant results	Participant identifier
Telecom	Macro F	0.182	**0.488**	**2**
			0.483	2
			0.480	3
Telecom	Micro F	0.337	**0.536**	**2**
			0.528	10
			0.510	3
Banks	Macro F	0.127	**0.360**	**4**
			0.352	10
			0.335	2
Banks	Micro F	0.238	**0.366**	**2**
			0.364	2
			0.343	8

In Table 5 we present the best results of tweet sentiment analysis for each domain and measure. Most best approaches in this task utilized SVM classification method. The features of the participant 2 comprised syntactic links presented as triples (head word, dependent word, type of relation). Participant 3 applied a rule-based method accounting syntactic relations between sentiment words and the target entities without any machine learning.

Additionally, one of participants fulfilled independent expert labeling of telecom tweets and obtained Macro-F—0.703, and Micro F—0.749, which can be considered as the maximum possible performance of automated systems.

The analysis of the obtained results showed that the most participants solved the general (not entity-oriented) task of tweet classification; entity-oriented approaches did not achieve better results in comparison with general approaches on tweets mentioned several entities.

6. Conclusion

In this paper we described the data, rules and results of SentiRuEval, evaluation of Russian object-oriented sentiment analysis systems. We offered two tasks to participants. The first task was aspect-oriented analysis of reviews about restaurants and automobiles, that is the primary goal was to find word and expressions indicating important characteristics of an entity (aspect terms) and then classify them into polarity classes and aspect categories.

The second task was the reputation-oriented analysis of tweets concerning banks and telecommunications companies. The goal of this analysis was to classify tweets in dependence of their influence on the reputation of the mentioned company. Such tweets could express the user's opinion or a positive or negative fact about the organization.

In each task about ten participants from universities and the industry took part. They have applied various machine-learning approaches including SVM, gradient boosting, CRF, recurrent neural networks and others. Given the participants' results, it can be concluded that the object-oriented sentiment analysis is poorly addressed by the applied methods. And most systems and methods need to be significantly improved to perform better on such tasks.

In the review collections interesting linguistic phenomena were also marked up. In particular, we have labeled comparisons with other entities or with previous opinions, desirable but not existing situations, irony. So the study of the markup can be useful also for linguists. All prepared materials are accessible for research purposes (reviews: http://goo.gl/Wqsqit and tweets: http://goo.gl/qHeAVo).

Acknowledgements

This work is partially supported by RFBR grants No. 14-07-00682, No. 15-07-09306 and by the Russian Ministry of Education and Science, research project No. 586.

References

1. *Amigo E., Corujo A., Gonzalo J., Meij E., de Rijke M.* (2012), Overview of RepLab 2012: Evaluating Online Reputation Management Systems, CLEF 2012 Evaluation Labs and Workshop Notebook Papers, Rome.
2. *Amigo E., Albornoz J. C., Chugur I., Corujo A., Gonzalo J., Martin T., Meij E., de Rijke M, Spina D.* (2013), Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems, CLEF 2013, Lecture Notes in Computer Science Volume 8138, pp. 333–352.
3. *Arora R., Srinivasa S.* (2014), A Faceted Characterization of the Opinion Mining Landscape, COMSNETS Workshop on Science and Engineering of Social Networks, Bangalore, pp. 1–6.
4. *Bagheri A., Saraee M., de Jong F.* (2013), An Unsupervised Aspect Detection Model for Sentiment Analysis of Reviews, in Natural Language Processing and Information Systems, Springer, Berlin, Heidelberg, pp. 140–151.
5. *Barbieri F., Saggion H.* (2014), Modelling Irony in Twitter: Feature Analysis and Evaluation, Proceedings of LREC, pp. 4258–4264.
6. *Chetviorkin I. I., Braslavski P. I., Loukachevitch N. V.* (2012), Sentiment Analysis Track at ROMIP 2011, Proceedings of International Conference Dialog, pp. 739–746.
7. *Chetvirokin I., Loukachevitch N.* (2013), Sentiment Analysis Track at ROMIP 2012, Proceedings of International Conference Dialog, volume 2, pp. 40–50.
8. *Dang H. T., Owczarzak K.* (2008), Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks, Proceedings of the First Text Analysis Conference.
9. *Glavaš G., Korencic D., Šnajder J.* (2013), Aspect-Oriented Opinion Mining from User Reviews in Croatian, Proceedings of the 4th Workshop on Balto-Slavonic Natural Language Processing, pp. 18–22.
10. *Jiang L., Yu M., Zhou M., Liu X., Zhao T.* (2011), Target-dependent Twitter Sentiment Classification, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 151–160.
11. *Jindal N., Liu B.* (2006), Mining Comparative Sentences and Relations, Proceedings of the 21st National Conference on Artificial Intelligence, Boston, pp. 1331–1336.
12. *Kusnetsova E., Loukachevitch N., Chetviorkin I.* (2013), Testing Rules for a Sentiment Analysis System, Proceedings of International Conference Dialog, pp. 71–80.
13. *Liu B.* (2012), Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies, Vol. 5(1).
14. *Macdonald C., Santos R., Ounis I., Soboroff I.* (2010), Blog Track Research at TREC, ACM SIGIR Forum, Vol. 44(1), pp. 58–75.
15. *Mohammad S. M., Kiritchenko S., Zhu X.* (2013), NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets, Proceedings of 7th International Workshop on Semantic Evaluation Exercises (SemEval-2013), Atlanta, pp. 321–327.
16. *Nakov P., Kozareva Z., Ritter A., Rosenthal S., Stoyanov V., Wilson T.* (2013), Semeval-2013 Task 2: Sentiment Analysis in Twitter, Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval-2013), Atlanta, pp. 312–320.

17. *Pang B., Lee L., Vaithyanathan S.* (2002), Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Vol. 10, pp. 79–86.
18. *Pontiki M., Galanis D., Pavlopoulos J., Papageorgiou H., Androutsopoulos I., Manandhar S.* (2014), SemEval-2014 Task 4: Aspect Based Sentiment Analysis, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, pp. 27–35.
19. *Popescu A. M., Etzioni O.* (2005), Extracting Product Features and Opinions from Reviews, Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Vancouver, pp. 339–346.
20. *Riloff E., Qadir A., Surve P., De Silva L., Gilbert N., Huang R.* (2013), Sarcasm as Contrast between a Positive Sentiment and Negative Situation, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 704–714.
21. *Rosenthal S., Ritter A., Nakov P., Stoyanov V.* (2014), SemEval-2014 Task 9: Sentiment Analysis in Twitter, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, pp. 73–80.
22. *Seki Y., Evans D. K., Ku L. W., Sun L., Chen H. H., Kando N.* (2008), Overview of Multilingual Opinion Analysis Task at NTCIR-7, Proceedings of NTCIR-7 Workshop Meeting, Tokyo, pp. 185–203.
23. *Stenetorp P., Pyysalo S., Topić G., Ohta T., Ananiadou S., Tsujii J.* (2012), BRAT: a Web-based Tool for NLP-assisted Text Annotation, Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, pp. 102–107.
24. *Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.* (2011), Lexicon-Based Methods for Sentiment Analysis, Computational Linguistics, Vol. 37(2), pp. 267–307.
25. *Zhang L., Liu, B.* (2014), Aspect and Entity Extraction for Opinion Mining, in Data Mining and Knowledge Discovery for Big Data, Springer, Berlin, Heidelberg, pp. 1–40.

ASPECT EXTRACTION AND TWITTER SENTIMENT CLASSIFICATION BY FRAGMENT RULES

Vasilyev V. G. (vvg_2000@mail.ru),
Denisenko A. A. (denisenko_alec@mail.ru),
Solovyev D. A. (dmitry_soloviev@bk.ru)

OOO «LAN-PROJECT», Moscow, Russia

The paper deals with approaches to explicit aspect extraction from user reviews of restaurants and sentiment classification of Twitter messages of telecommunication companies based on fragment rules. This paper presents fragment rule model to sentiment classification and explicit aspect extraction. Rules may be constructed manually by experts and automatically by using machine learning procedures. We propose machine learning algorithm for sentiment classification which uses terms that are made by fragment rules and some rule based techniques to explicit aspect extraction including a method based on filtration rule generation. The article presents the results of experiments on a test set for twitter sentiment classification of telecommunication companies and explicit aspect extraction from user review of restaurant. The paper compares the proposed algorithms with baseline and the best algorithm to track. Training sets, evaluation metrics and experiments are used according to SentiRuEval. As our future work, we can point out such directions as: applying semi-supervised methods for rule generation to reduce the labor cost, using active learning methods, constructing a visualization system for rule generation, which can provide the interaction process with experts.

Key words: fragment rules, sentiment classification, aspect extraction, opinion mining

1. Introduction

Opinion mining and sentiment extraction is an actively developing sub discipline of data mining and computational linguistics. A promising approach to automatic sentiment extraction is based on extraction of specific product features — aspects and on the determination of those polarities. Usually the problem is solved in three stages. At first aspects and those polarities are extracted. Then aspects gears to categories if they are predefined. Otherwise a set of aspects is clustered and representative aspects are selected. The final stage includes category polarity classification based on polarities of individual aspects.

In this paper we present a rule-based approach which exploits fragment rule model to explicit aspect extraction from user reviews and to sentiment classification of twitter messages. The main advantage of the approach is its good interpretability.

On the one hand, there is an opportunity to use expert knowledge in the model by means of constructing rules manually. On the other side, you can build the model automatically or get the interpretable model within a procedure, which includes interaction of an expert and a system.

In paper [7] approaches to sentiment classification of movie reviews are described. These approaches based on counting the number of the proposed positive and negative words and using Naive Bayesian classifier, maximum entropy classification, support vector machine. Using support vector machine raises accuracy to 82%. Another two methods of classification gives accuracy 75–80%. In paper [1] twitter sentiment classification based on support vector machine is described. The words, phrases and part of speech are used as features. The results shown in this paper are the same as results shown in the previous paper and stressed that using part of speech does not increased accuracy.

In paper [2] two approach to sentiment classification movie review. The first approach based on the number of positive and negative terms, intensification terms, and reverses the semantic polarity of a particular term. The second approach uses a machine learning algorithm, support vector machines. Using the first approach gives accuracy about 65–70%. Using the second approach raises accuracy to 85%. Combination the two approaches not increase accuracy.

In paper [3] authors propose approach to sentiment classification with polarity shifting detection. Polarity-shifted and polarity-unshifted sentences are used as features for classification based on support vector machine. This approach allows a few to improve the quality compared to the baseline.

In addition to the vocabulary and the vector approach for sentiment classification a number of papers propose special probabilistic models, for example, tree-based sentiment classification and using relationship between words [6]. Also, a number of papers the authors clearly define the rules of assessment texts. Particularly, in paper [7] different rule for determining the scope inverse word such as “no” are formulated. Thus, in the work on sentiment classification are used as standard methods for text classification, and modified methods, which take into account polarity shifted terms, the syntactic structure of sentences, the relationship between words.

In current paper approach to twitter sentiment classification based on features extracted by using fragment rules. Thus obtained features with proper setting of rules form the space of smaller dimension and have good descriptive power, as was shown in [10].

Aspect-based opinion mining has been widely researched. There are some known approaches to this task [4]: (1) frequency-based approach, (2) rule-based approach, (3) supervised learning techniques, (4) topic modelling techniques.

Frequency-based approach uses the fact that 60–70% of the aspects are explicit nouns [4]. It is argued that people writes reviews in aspect language because they also read other reviews and take the terminology. Rule-based approach uses the assumption that there is some kind of relation between aspects and polarities expressed in a text. A relation can be formalized by using rules. There is also a hybrid approach expressed in using rules for filtration of extracted noun phrases.

The problem may be considered as sequence labelling problem according to some suggested supervised machine learning methods. In particular, Hidden Markov Model

and Conditional Random Fields can be used. Topic modelling techniques use the natural assumption that topics of reviews are corresponding aspects.

In this paper, a rule-based approach to aspect extraction is proposed. There are two main rule models: grammar-based and fragment-based. Grammar models include the application of context-free grammars for example Tomita parser [8]. The other model is based on using special fragments from text and represents a number of operations under these fragments. A rule in this case is a declarative description of extracted information. Our model is an example of the last approach.

Due to the fact, that recall of aspect extraction can be achieved by using various dictionaries like thesaurus and domain-specific dictionaries, an important issue is improving precision. In this case, the improvements expressed in using special filtration mechanisms for extracted aspects. Here particularly fragment rules can be used. The purpose of participation in the track was testing fragment rule-based approaches to aspect extraction and tweet classification. In addition, we attempted to use methods for automatic fragment rule generation.

The remainder of the article is as follows. In section 2 a formal description of the fragment rule language and a description of proposed approaches is given. In section 3 obtained results are analyzed; a comparison with Baseline results and the best track results is given. Section 4 presents conclusion and future work.

2. Methods

2.1. Fragment rules model

In this work for describing text features and classification rules we used a mathematical model based on defining operations on sets of text fragments [9].

Let we have the text $D = (d_1, ..., d_n)$, where the $d_i \in T$ — single element of the text, $T = \{t_1, ..., t_m\}$ — the set of all elements, n — the length of the text, m — number of different elements of the text.

Definition 1

The set $\mathbb{F} = \{ (p, q) \mid 1 \le p \le q \le n \}$ will be called the set of all parts of the text length n. Fragments of the text will be called the single elements of the set $f = (f_l, f_r) \in \mathbb{F}$, that specify left f_l and right f_r border fragment (number of the first and last elements in fragment).

Definition 2

Let $f = (f_l, f_r) \in \mathbb{F}$ and $g = (g_l, g_r) \in \mathbb{F}$, then $|f| = f_r - f_l + 1$ — length of the fragment;
$g \sqsupset f$, if $g_l \le f_l \le f_r \le g_r$ and $f \ne g$ — inclusion relation;
$g \ll f$, if $g_l < f_l$ or $g_l = f_l \,\&\, f_r < g_r$ — order relation.

Definition 3

The set of fragments F will be called reduced, if there is no such $f \in F$, that $g \sqsupset f$. $R(F)$ denote reduced set of fragments based on the set F, R — reduce operation.

Definition 4

The distance between the fragments $f = (f_l, f_r) \in \mathbb{F}$ and $g = (g_l, g_r) \in \mathbb{F}$ is determined as follows:

$$d(f,g) = \begin{cases} g_l - f_r, f < g, \\ f_l - g_r, g < f, \\ g_l - f_r, g = f. \end{cases}$$

Definition 5

The result of the a rule Q for the text D is the set $F_Q \subset \mathbb{F}$, containing all of the fragment relevant this rule. If $F_Q \neq \emptyset$, then call the text D relevant rule Q.

Definition 6

Basic rules is a rule $Q = t, t \in T$ whose result is $F_Q = \{f_1, ..., f_l\}$ — reduced set of fragments, the elements that stand out in a single operation. Complex rule is a rule Q, which is obtained by performing operations on other rules $Q_1, ..., Q_k$.

Let us now determine the possible operations to build complex rules of Q from the basic rules $Q_1, ..., Q_k$.

Definition 7

$Q = Q_1 \nabla Q_2$ — binary operation OR, $F_Q \equiv R\left(F_{Q_1} \nabla F_{Q_2}\right)$,
$F_{Q_1} \nabla F_{Q_2} = \left\{f \in \mathbb{F} | \exists f_1 \in F_{Q_1}, f \sqsupset f_1 or\ \exists f_2 \in F_{Q_2}, f \sqsupset f_2\right\}$.

For example, the rule *good best quality* extract fragments relevant the appearance of these words in the text.

Definition 8

$Q = Q_1 \Delta_{n_1} Q_2$ — binary operation AND with limit on distance between fragments, $F_Q \equiv R\left(F_{Q_1} \Delta_{n_1} F_{Q_2}\right)$, $F_{Q_1} \Delta_{n_1} F_{Q_2} = \left\{f \in \mathbb{F} | \exists f_1 \in F_{Q_1} and\ \exists f_2 \in F_{Q_2}, that\ f \sqsupset f_1,\ f \sqsupset f_2 and\ d(f_1, f_2) \leq n_1\right\}$.

For example, the rule *beeline &4w LTE* extract fragments, in which distance between "*beeline*" and "*LTE*" less than 4 words. This operation can be used without any limits on the distance between the words.

Definition 9

$Q = Q_1 \square_{n_1,n_2} Q_2$ — binary operation of sequence with limit on distance between fragments, $F_Q \equiv R\left(F_{Q_1} \square_{n_1,n_2} F_{Q_2}\right)$, $F_{Q_1} \square_{n_1,n_2} F_{Q_2} = \{f \in \mathbb{F} | \exists f_1 \in F_{Q_1} and\ \exists f_2 \in F_{Q_2},\ that\ f_1 < f_2,\ d(f_1, f_2) > 0,\ f \sqsupset f_1, f \sqsupset f_2\ and\ n_1 \leq d(f_1, f_2) \leq n_2\}$.

For example, the rule *@Company: 3w (sale discount)* extract fragments, which after the name of the company at a distance of 3 words are words of "sale" or "discount". This operation can be used without any limits on the distance between the words.

Definition 10
$Q = \bowtie (Q_1, ..., Q_k)$ — multiple operation sequences of neighbouring elements (select of neighbouring fragments), $F_Q \equiv R\left(\bowtie\left(F_{Q_1}, \dots, F_{Q_k}\right)\right), \bowtie\left(F_{Q_1}, \dots, F_{Q_k}\right) =$
$= \{f \in \mathbb{F} | \exists f_i \in F_{Q_i}, i \in \overline{1,k}\colon\ f_i \prec f_{i+1},\ d(f_i, f_{i+1}) = 1, i \in \overline{1,k-1} and\ f \sqsupset f_i,$
$i \in \overline{1,k}\}$.

For example, the rule "*(boss head director chief) (mts beeline megafon)"* extract phrases corresponding to different telecom executives.

Definition 11
$Q = Q_1 \lozenge Q_2$ — binary operation finding the intersection of fragments,
$F_Q \equiv \{f \in \mathbb{F} | f \in F_{Q_1} \wedge f \in F_{Q_2}\}$.

For example, the rule *[Chapter $SentBegin] extract words* "*Chapter*", that are written in the beginning of the sentence.

Definition 12
$Q = Q_1 \lhd_{n_1,n_2}$ — unary operator imposes limitations on length of the fragment,
$F_Q \equiv \{f \in F_{Q_1} | n_1 \le |f| \le n_2\}$.

For example, the rule *(beeline & mts) #IN #INTERVAL(2w/3w)* extract fragments containing specific words in length from 2 to 3 words.

To be able to construct rules include negation and conditional statements (when the presence of the expression is checked, but it is not included in the final fragment) are special variants of binary rules $\nabla, \Delta, \square, \bowtie, \lozenge, \lhd,\ \Delta_{n_1}, \square_{n_1 n_2}$, in which one of the operands is considered negative or conditional. For example, $\square^{\rightarrow}_{n_1,n_2}$ is operator finding the sequence in which the second operand is taken from the negation; $\square^{\nleftarrow}_{n_1,n_2}$ is operator finding the sequence in which the first operand is taken from the negation; $\square^{\leftarrow}_{n_1,n_2}$ — is operator finding the sequence in which the first operand is conditional. The rule $\square^{\nrightarrow}_{n_1,n_2}$ defined as $Q = Q_1 \square^{\nrightarrow}_{n_1,n_2} Q_2\ F_Q \equiv \{f \in F_{Q_1} | \exists! f_2 \in F_{Q_2}\colon f \prec f_2, 0 < n_1 \le d(f, f_2) \le n_2\}$.

For example, the rule *no ^:3 (good best quality)* extract the word "*good*", "*best*" and "*quality*" before which there is no word "no" at distance of three words.

#define command sets the named expression. In the pre-treatment rules text expression is substituted into the rule text. These expressions are used to avoid repeating elements in complex rules. *#set* command s used to set the saved variables. Unlike #define command at the first reference to the variable is made save search results and on subsequent calls text processing is not performed. To use named expressions or saved variables in the rule is necessary to use operators @ and @@.

For example*, #define Good (good best quality)* sets the named expression *Good,* which should be handled *@Good.*

2.2. Sentiment classification

For sentiment classification we used a hybrid approach which is based on combining rule-based feature extraction and classifier training by machine learning methods. Classifier induction includes training set pre-processing, feature extraction by using predefined set of fragment rules, training classifier by using selected machine learning methods.

Texts in the training set are pre-processed by using the following procedures:

1. Graphematical analysis (tokenization, sentence boundary detection, phonetic coding, word descriptors extraction).
2. Linguistic analysis (lemmatization, part of speech tagging, word sense disambiguation, collocation extraction, syntactic features extraction).
3. Low level indexes construction (inverted index of source word forms, inverted index of lemma word forms, inverted index of word descriptors).

The general scheme of the learning algorithm has the following form.

1. Building vector representation of texts by using the set of fragment rules.
2. Dimension reduction and feature weights calculation.
3. Training and evaluation of the classifier on the training set.

At the first step the predefined set of 100 special fragment rules are used for features extraction.

Example of fragment rule:
@@COND^:5((@@NEG^:5\s(@@INTENS^:5\s($Adj $Verb $Noun $Adv)))
&5\s? @@OBJECT),

where @@COND — condition words ("if"), @@NEG — negative words, @@INTENS — intensive words ("very", "far", "purely"), @@OBJECT — object ("mts", "megafon", "beeline").

At the second step we used common methods for dimension reduction and feature weights calculation.

At the third step two classifiers are trained, one classifier for the positive class and one for the negative class. For classifier training we used our robust realization of the following standard machine learning methods:

1. Bayesian classifier based on multivariate Gaussian distribution (gmm),
2. K-nearest neighbours classifier (knn),
3. Von Mises-Fisher classifier (vmfs),
4. Roccio classifier (roccio),
5. Support vector machines classifier (svm).

Trained positive and negative classifiers are used for building the final decision rule of the following form:

$$d'(u) = \begin{cases} 1, d_{pos}(u) > d_{neg}(u) \mid d_{pos}(u) = d_{neg}(u) = 1, w_{pos}(u) > w_{neg}(u) \\ -1, d_{pos}(u) < d_{neg}(u) \mid d_{pos}(u) = d_{neg}(u) = 1, w_{pos}(u) < w_{neg}(u) \\ 0, d_{pos}(u) = d_{neg}(u) = 0 \end{cases}$$

where $d'(u) \in \{-1, 0, 1\}$ is the final decision rule, $d_{pos}(u) \in \{0, 1\}$ and $d_{neg}(u) \in \{0, 1\}$ is the decision rules for positive and negative class, $w_{pos}(u) \in [0, 1]$ and $w_{neg}(u) \in [0, 1]$

and degree of compliance positive or negative class (for probabilistic classifiers it is the probability assignment to the corresponding class, for svm it is the distance to corresponding hyperplane etc.), u — the set of features in the text.

2.3. Rule-based explicit aspect extraction

There are two types of aspects defined in aspect-based opinion mining: explicit and implicit. Explicit aspects are concepts that explicitly mentioned in a sentence. Implicit aspects are expressed indirectly. This section proposes a number of approaches to explicit aspect extraction based on fragment rules. Preliminary let $A = \{a_1, ..., a_n\}$ be a set off unique aspects extracted by experts and represented in the training set. Training set has been provide by SentiRuEval organizers [5].

Multiple operation OR

Basically for the purpose of explicit aspect extraction this kind of fragment rule can be used:

$$Q = Q_{\nabla}(a_1, a_2, ..., a_n),\ a_i \in A.$$

Here Q_{∇} — is a rule, where operation OR acts as a connector between unique aspects. In fact, an appropriate set of fragments is extracted for each aspect. The result of the operation is a reduced united set of fragments.

Multiple operation OR with maximizing reduction

In the concerned case, the following situation may arise. Instead of a whole aspect, structural parts can be extracted. For example, there are three extracted aspects HOT, DISH, HOT DISH. A standard reduction method will delete the biggest fragment HOT DISH, and we'll have two aspects instead of one. In this regard, it was decided to modify the reduction method and to exclude fragments which are included in other fragments. Also it should be noted that neighbouring fragments may be one aspect. Therefore overlapping fragments and neighbouring fragments should be combined. As a result, fragments of the maximum length are extracted.

Rule-based filtration

Also it seems appropriate to use rule-based filtration for aspect extraction. The extraction algorithm constructed as follows. At first using aspects selected by experts fragments from an aspect to the nearest adjective are extracted. Then, the most common rules based on the extracted fragments (templates) are formed. Here in the feature space is defined previously. The generated rules are applied to filter the set of extracted candidate-aspects by counting support and removal of candidates with support below a threshold. As already mentioned, recall may be achieved by using appropriate dictionaries. In this case, the filtration process is necessary to improve precision. Definition of the context of some aspects allows to separate situations where the term is not an aspect.

Let (a_i) be a rule, a result is a set of fragments from the aspect a_i to the nearest adjective. The aspect extraction algorithm for each aspect selected by experts generates a set of aspect contexts $Q(a_i)$ by applying rule $Q(a_i)$ to the training set L.

Then the rule generation algorithm builds templates of these contexts. In each review candidate-aspects are extracted and filtered by using these templates. Finally, we have a set of extracted explicit aspects.

Algorithm2. Explicit aspect extraction with filtration

Input. A_L — set of aspects selected by experts
I — hierarchy of features,
L — train set,
R — test set

Output. A_T — extracted explicit aspects.

Step 1. For all $a_i \in A_L$
$GenerateRules(I, Q_L(a_i))$

Step 2. For all $r \in R$
For all $a_i \in A$
$A_T \leftarrow A_T \cup FilterAspects(Q_r(a_i))$

There are a number of classical algorithms for searching frequent item sets which used for generating rules such as *Apriori*, *FP-growth*, *Eclat*. One important difference between these algorithms is a method of data representation. Basically there are two approaches—horizontal and vertical representation. In the vertical representation it's necessary to have lists of fragments that match elements of a rule. In the horizontal representation each fragment corresponds to a set of rule elements. Vertical representation is more practical in case of the fragment model. In this context, it is possible to apply one of the known algorithms — Eclat [11]. Especially because support of rules is determined by the intersection of sets of fragments.

Rules of the form $Q_1 \square_{1,1} Q_2 \square_{1,1} \ldots \square_{1,1} Q_n$ are used for filtration. Searching of rules is based on a feature hierarchy. As elements of the hierarchy you may have parts of speech descriptors, single words, etc. Sequentially from the descriptor $Any (any word) a rule is expanding and specifying. A selection criterion is a degree of specificity of rules and a minimal support threshold. The specificity of the rules increases depending on a number of elements and their place in the hierarchy. The more elements and the lower the place of elements in the hierarchy then specificity is higher. In this case, the rules are eliminated with support below a threshold. As a result, every aspect is associated with set of rules. In such a way, filtration is done when there are only those candidate-aspects which match at least one rule.

3. Evaluation

3.1. Twitter sentiment classification

Used for teaching training set consisting of 3,846 tweets of telecommunications companies. Each company which was mentioned on Twitter rated on a scale {−1, 0, 1}.

Test set consists of 5,322 tweets about telecommunications companies. The objective of the testing was to include every mention of the company to one of three classes: positive, negative or neutral. Indicators macro *F*-measure and micro *F*-measure used to assess the quality. Test results are shown in Table 1. The table shows the best method, Baseline and 5 runs:

9_1 Bayesian classifier based on a mixture of multivariate normal distributions (*gmm*),
9_2 classifier k-nearest neighbours (*knn*),
9_3 Bayesian classifier based on the distribution of von Mises-Fisher (*vmfs*),
9_4 centroid classifier Roccio (*roccio*),
9_5 classifier based on support vector machines (svm).

Baseline refers all tweets to the most frequent class, in this case a negative. Used for teaching training set consisting of 3,846 tweets of telecommunications companies. Each company which was mentioned on Twitter rated on a scale {−1, 0, 1}.

Indicators macro *F*-measure and micro *F*-measure used to assess the quality [5].

Table 1. Evaluation of the quality of sentiment classification tweets

Algorithm	Macro *F*-measure	Micro *F*-measure
9_1 (gmm)	0,3158	0,3331
9_2 (knn)	0,2328	0,2626
9_3 (vmfs)	0,3305	0,3371
9_4 (roccio)	0,3310	0,3501
9_5 (svm)	0,3527	0,3765
Baseline	0,1823	0,3370
2_B	0,4829	0,5362

Evaluating the quality of classification are at Baseline micro *F*-measure and substantially higher macro *F*-measure. This can be explained feature Baseline and calculation rule micro and macro *F*-measure. Macro *F*-measure — is the average amount of standard *F*-measure that calculated separately for the three classes. Baseline algorithm has zero *F*-measure for two classes (positive and neutral), but *F*-measure negative class has a value of about 55%. By averaging the three classes *F*-measure is found to be 18%.Our algorithm solves these problems. The algorithm based on support vector machines shown best quality. The algorithm based on k-nearest neighbours showed the worst result. As we can see our result are comparable with result of other participants.

3.2. Explicit aspect extraction

Performance evaluation was made against the training set (gold standard), provided by organizers. The set consists of 202 annotated reviews in Russian. We used standard measures: precision, recall and F-measure. In official results the method based on multiple operation OR with maximizing reduction has identifier — 11.1.

Table 2. Evaluation results for explicit aspect extraction

Method	Strong demands			Weak demands		
	P	R	F1	P	R	F1
OR	49%	71%	58%	59%	72%	65%
Multiple operation OR with maximizing reduction [11.1]	51%	73%	60%	61%	74%	66%
Rule-basedfiltration	60%	64%	62%	66%	69%	67%
Baseline	55%	69%	61%	65%	70%	67%
[2.1] The best result/strong	72%	57%	63%	81%	62%	69%
[4.1] The best result/weak	55%	69%	61%	69%	79%	73%

In general, participants in the official track had comparable results. It turns out that the approach based on transferring aspects from the train set to the test set with normalization shows the same results as approaches used sophisticated models for training.

The results show that the modification of multiple OR operation generally contributes to the performance. It can be argued that maximizing reduction showed an advantage compared to minimizing reduction when there are only those fragments that contain no other. This reduction is applied in solving text classification tasks and offers advantages in terms of speed of execution of classification rules. In the future, different types of reduction can take the form of individual operations instead of using in default.

Application of rules in filtration also has a positive effect on the result, but there are a number of issues that require further study. Along with increasing precision recall decreases. To solve this problem it is advisable to consider other criteria of rule selection to find suitable experimental values of boundary parameters for rule specificity and support of candidate-aspects to achieve a minimum reduction of recall.

4. Conclusions and Future work

The paper deals with approaches to explicit aspect extraction and sentiment classification. The algorithm based on support vector machines shown best quality. The algorithm based on k-nearest neighbours showed the worst result. The results are at the level of the average results presented in sentiment analysis track. The algorithm based on SVM using as features normalized lemma and syntactic links shown the best results on the track. In the efforts to extract the aspects we can say that the simplest approach shows comparable with the rest of the results. The use of filtering rules

to improve the accuracy while reducing completeness. In this regard, it is necessary to separately evaluate the effect of boundary parameters on the result.

As our future work, we can point out such directions as: applying semi-supervised methods for rule generation to reduce the labor cost, using active learning methods, constructing a visualization system for rule generation, which can provide the interaction process with experts. Also expanding of the fragment rule model can give new expressive possibilities.

References

1. *Go A., Huang L., Bhayani R.* (2009), Twitter sentiment classification using distant supervision, CS224N Project Report, Stanford.
2. *Kennedy A., Inkpen D.* (2006), Sentiment Classification of Movie Reviews Using Contextual Valence Shifters, Computational Intelligence, vol. 22(2), pp. 110–125.
3. *Li S., Lee S. Y. M., Chen Y., Huang C.-R., Zhou G.* (2010), Sentiment Classification and Polarity Shifting, Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, China, pp. 635–643.
4. *Liu B.* (2012), Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers.
5. *Loukachevitch N., Blinov P., Kotel'nikov P., Rubtsova Ju., Ivanov V., Tutubalina E.* (2015), SentiRuEval: Testing Object-Oriented Sentiment Analysis Systems in Russian.
6. *Nakagawa T., Inui K., Kurohashi S.* (2010), Dependency tree-based sentiment classification using CRFs with hidden variables, The 2010 Annual Conference of the North American Chapter of the ACL, Los Angeles, USA, pp. 786–794.
7. *Pang B., Lee L., Vaithyanathan S.* (2002), Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP, Philadephia, Pennsylvania, USA, pp. 79–86.
8. *Tomita parser: https://tech.yandex.ru/tomita/*
9. *Vasilyev V. G.* (2011), Fragment extraction and text classification by logical rules [Klassifikatsija i vydelenie fragmentov v tekstah na osnove logicheskih pravil] Digital libraries: Advanced Methods and Technologies, Digital Collections RCDL'2011, Voronezh, pp. 133–139.
10. *Vasilyev V. G., Davidov S. U.* (2013), Sentiment classification by combined approach. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2013", available at: www.dialog-21.ru/digests/dialog2013/materials/pdf/VasilyevVG.pdf
11. *Zaki M. J., Parthasarathy S., Ogihara M., Li W.* (1997), New Algorithms for Fast Discovery of Association Rules, Proceedings of the Third International Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, USA, pp. 283–286.

АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ ТОНАЛЬНОСТИ ОБЪЕКТОВ С ИСПОЛЬЗОВАНИЕМ СЕМАНТИЧЕСКИХ ШАБЛОНОВ И СЛОВАРЕЙ ТОНАЛЬНОЙ ЛЕКСИКИ

Поляков П. Ю. (pavel@rco.ru),
Калинина М. В. (kalinina_m@rco.ru),
Плешко В. В. (vp@rco.ru)
ООО «ЭР СИ О», Москва, Россия

Ключевые слова: определение тональности, анализ мнений, тональность объектов, тональность атрибутов, синтактико-семантический анализ, семантические шаблоны

AUTOMATIC OBJECT-ORIENTED SENTIMENT ANALYSIS BY MEANS OF SEMANTIC TEMPLATES AND SENTIMENT LEXICON DICTIONARIES

Polyakov P. Yu. (pavel@rco.ru),
Kalinina M. V. (kalinina_m@rco.ru),
Pleshko V. V. (vp@rco.ru)
RCO LLC, Moscow, Russia

This paper studies use of a linguistics-based approach to automatic object-oriented sentiment analyses. The original task was to extract users' opinions (positive, negative, neutral) about telecom companies, expressed in tweets and news. We excluded news from the dataset because we believe that formal texts significantly differ from informal ones in structure and vocabulary and therefore demand a different approach. We confined ourselves to the linguistic approach based on syntactic and semantic analysis. In this approach a sentiment-bearing word or expression is linked to its target object at either of two stages, which perform successively. The first stage includes usage of semantic templates matching the dependence tree, and the second stage involves heuristics for linking sentiment expressions and their target objects when syntactic relations between them do not exist. No machine learning was used. The method showed a very high quality, which roughly coincides with the best results of machine learning methods and hybrid approaches (which combine machine learning with elements of syntactic analysis).

Key words: sentiment analysis, object-oriented sentiment analysis, aspect-based sentiment analysis, opinion mining, syntactic and semantic analysis, semantic templates

1. Introduction

The task of automatic sentiment analysis of natural language texts has become extremely in demand. Many commercial companies producing goods and services are interested in monitoring social networking websites and blogs for users' opinions about their products and services. However, until recently there were no tagged text corpora in Russian on which developers could test and compare quality of their methods. This gap was filled by ROMIP and later SentiRuEval sentiment analysis evaluation conferences with their sentiment analysis tracks. However, the task of the previous conferences was to detect general sentiment of a text (for example, see Chetviorkin I., Braslavski P. I., Loukachevitch N. [2]), while at the present conference the task was brand new—object-oriented sentiment analysis, which is more difficult and requires more sophisticated algorithms; for, in case of general sentiment detection, selection of positive and negative terms and defining of their weights are important, while, in case of object-oriented sentiment detection, syntactic relations between a target object and a word expressing sentiment are also of great importance.

Such object-oriented method is not new for us; we have already used similar approach in our previous research. For instance, we evaluated sentiment-oriented opinions in regard to car makes on the material of the LiveJournal blog AUTO_RU (see description of the method in Ermakov A. E. [4]). It should be mentioned, however, that in all the previous cases results had only been evaluated by ourselves. Participation in SentiRuEval gave us a chance to have an independent evaluation of our method and compare our results with other participants'.

In this paper we present results of applying a linguistics-based approach involving syntactic and semantic analysis to the task of automatic object-oriented sentiment analysis. We confined ourselves to a linguistic method only, having excluded machine learning, because it was interesting to see what results a pure linguistic approach without machine learning methods would provide.

The task was to find sentiment-oriented opinions (positive and negative) about telecom companies in tweets.

2. Related Work

Usually object-oriented or aspect-oriented approaches either rely only on statistics-based algorithms, word distance count, machine learning, etc. to find opinion targets (starting with the first work on opinion target extraction by Hu and Liu [5]); or they may use shallow parsing to segment a sentence, find significant conjunctions, negations, and modifiers (ex., Kan D. [7]). Other approaches are looking for syntactic dependency between a sentiment term and its target (ex., Popescu A., Etzioni O. [9]), ignoring sentiment-bearing words which are not syntactically related to any target object. The distinctive feature of our approach is that using a deep linguistic method we take into account not only syntactically related sentiment terms (which provides high precision) but also independent sentiment-bearing words and phrases (which provides high recall).

Some researchers try combine statistical and linguistic methods in order to achieve the best results; for example, in Jakob N., Gurevych I. [6] authors use, among other, the dependency parse tree to link opinion expressions and the corresponding targets; and the experiments show that adding the dependency path based feature yields significant improvement to their method. However, their algorithm is searching for short and direct dependency relations only; therefore, their approach has difficulties with more complex sentences. Furthermore, they do not distinguish between a target object (ex., *camera*), its attributes or parts (ex., *lens cap, strap*), and its qualities (ex., *usability*); and, hence, they label the closest noun phrase as a target of the opinion. In contrast, we use a very basic ontology to distinguish between a target object, attributes, and qualities; and having found a sentiment related to an attribute or quality our algorithm goes down the dependency parse tree searching for a target object. If not found syntactically, the target object is being searched for by a heuristic, based on the clause distance. When the target object is found, the sentiment labeled to its attribute is assigned to the object.

3. Methods

To perform the task we based on our previous researches and solutions. Detailed description of these methods can be found in Ermakov A. E., Pleshko V. V. [3] and Ermakov A. E. [4]. New to the approaches described in [3] and [4] was adding so-called 'Free Sentiment Detection', which will be described in Section 3.2.

The text analysis algorithm has the following stages in regard to the sentiment detection task:

1) Tokenization;
2) Morphological analysis;
3) Object extraction;
4) Syntactic analysis;
5) Fact extraction (use of semantic templates);
6) Free sentiment detection.

Stages 1, 2, and 4 were implemented by standard RCO tools for general text analysis. At stage 3 we paid more attention to the objects concerning the given subject (names of mobile companies, telecom terminology, etc.). Stages 5 and 6 were core to the sentiment detection task and, therefore, will be described in detail.

3.1. Semantic Templates

The main method of sentiment analysis involved usage of semantic templates.

Semantic template is a directed graph representing a fragment of a syntactic tree with certain restrictions applied to its nodes. The syntactic tree of a sentence contains semantic and syntactic relations between words, which are defined by the syntactic parser. The restrictions in the templates can be applied to a part of speech, name, semantic type, syntactic relations, morphological forms, etc. Fact extraction is performed by finding a subgraph in the syntactic tree of a sentence which is isomorphic to the template (with all restrictions applied).

RCO syntactic analyzer, based on the dependency tree approach, has been used. The semantic network built by the syntactic parser is invariant to the word order and voice; for example, sentences (1) *Оператор украл деньги со счета* and (2) *Деньги украдены оператором со счета* will have the same semantic net. Such semantic network constitutes an intermediate representation level between the semantic scheme of a situation and its verbal expression, that is, a deep-syntactic representation, abstracted from the surface syntax.

Settings of the semantic interpreter allow filtering negative and 'unreal' (imperative, conditional, etc.) statements, which don't correspond to real events and should not be analyzed. As a result, examples like (3) *если Билайн будет плохо работать; сеть якобы падает; связь бы обрывалась; не Билайн плохо работает* can be excluded from the sentiment detection.

To decrease the number of templates describing semantic frames, we have so-called auxiliary templates, which add new nodes and relations into the semantic network. In the process of semantic analysis and fact extraction auxiliary templates work before all other templates, so that semantic templates can base on the net built by both the syntactic analyzer and the auxiliary templates. For example, if we interpret phrases like (4) *X does Y, X begins to do Y*, and (5) *X decides to do Y* as equal for a particular semantic frame, instead of creating a semantic template for each example we can have one auxiliary template, which will mark the subject of the main verb as the subject of the subordinate verb, and one simple semantic template—(4) *X does Y*.

Semantic templates can have so-called 'forbidding nodes' which impose restrictions on the context, defining in which context the template should not match. For example, (6) *У Билайна надежная связь* is a positive statement, while adding the adverb *наименее* changes its sentiment to opposite: (7) *У Билайна наименее надежная связь*. By the means of forbidding nodes we can distinguish between these two sentences, stating that the adjective should not be modified by the adverb *наименее*. Usage of forbidding nodes significantly increases the precision of sentiment analysis.

Fig. 1 demonstrates a semantic template used to detect sentiment expressed by a verb or adverb in sentences like: (8) *Билайн ловит хорошо; Интернет летает*.

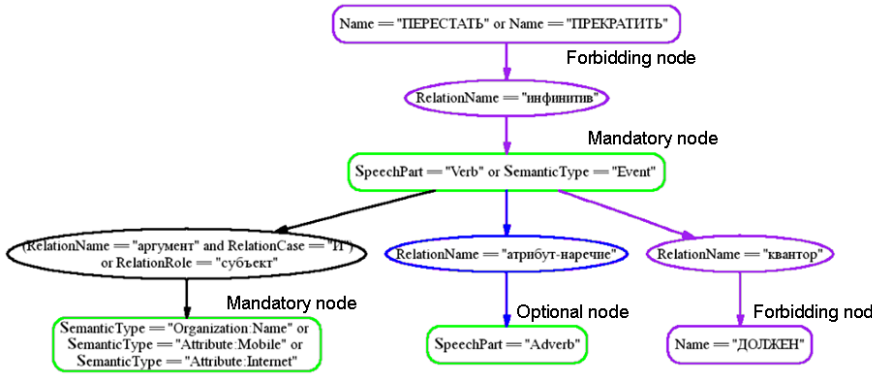

Fig. 1. Example of a semantic template

Nodes contain restrictions on parts of speech (*SpeechPart == "Verb"; SpeechPart == "Adverb"*), lexical items (*Name == "ПЕРЕСТАТЬ" or Name == "ПРЕКРАТИТЬ"*), semantic categories (*SemanticType == "Organization:Name" or SemanticType == "Attribute:Mobile"*). Restrictions on semantic and syntactic relations between words include: relation name (*RelationName == "аргумент»; RelationName == «квантор»*), semantic role (*RelationRole == "субъект»*), case (*RelationCase == "И"*). Forbidding nodes state that the verb expressing sentiment should not be controlled by the verbs *перестать* or *прекратить* or modified by the predicative *должен*. Thus, this template will match the sentence (8) *Билайн хорошо ловит* (which is positive), but not (9) *Билайн перестал хорошо ловить* (which is negative) or (10) *Билайн должен хорошо ловить* (which we consider neutral).

Restrictions of the semantic templates were enriched by the use of special dictionaries (so-called filters), containing vocabulary for positive and negative appraisals. This vocabulary includes nouns, adjectives, verbs, adverbs, and collocations. A word from a filter must be syntactically related to the target of evaluation. Selection of terms for the filters was manual, performed by a linguistic expert. Examples of positive terms: *супербыстрый, шустро, красота, крутяк, блистать, радовать, обеспечивать уверенный прием*. Examples of negative terms: *завышенный, препротивнейший, позорище, тормознутость, обдирать, терять соединение, фигово.*

For example, a set of particular words from the semantic filters are applied to the template in Fig.1 as restrictions: verbs or verbal nouns parameterize the node with the restriction *SpeechPart == "Verb" or SemanticType == "Event"*; adverbs parameterize the node with the restriction *SpeechPart == "Adverb"*, both these nodes have the semantic role 'Appraisal'.

Ultimate targets of evaluation were main Russian mobile phone providers (Beeline, Megafon, MTS, Rostelecom, Tele2), but also users' appraisals of providers' attributes were taken into account (communication quality, mobile Internet, customer service, etc.).

Analyzing users' comments and opinions on social networking sites and forums experts defined a set of attributes which were most frequently mentioned by mobile phone users. Thus, a list of most important things for users was made. Given attributes were divided into three classes: 1) Mobile Attributes—terms strictly connected to the mobile telephony: *SMS, MMS, 3G, LTE, SIM-card, roaming, etc.*; 2) Internet Attributes—terms strictly connected to the Internet: *Internet, ping, etc.*; 3) General Attributes—terms often used related to the mobile telephony but which can also refer to other domains: *call center, signal, network, customer support, balance, etc*. Each list was extended by synonyms and spelling variants (*интернет=инет=и-нет; lte=лте =lтешечка =лте-шечка; баланс счета=состояние счета=средства на счету=деньги на счету, etc.*). When a sentiment related to a certain attribute was detected, given sentiment was also ascribed to the corresponding mobile provider.

In Fig.1 the node with the restriction *SemanticType == "Organization:Name" or SemanticType == "Attribute:Mobile" or SemanticType == "Attribute:Internet"* is parameterized by names of mobile operators, mobile attributes or Internet attributes; the semantic role of the node is 'Target Of Evaluation'.

This method provides a very high precision, though not so high recall.

3.2. 'Free' Sentiment

Although usage of semantic templates provides very good accuracy, this method has its disadvantage—a word expressing sentiment must be in the same sentence as the target of evaluation and must be syntactically related to it. As it is not always so in natural texts, some cases of clearly expressed sentiment will be omitted by this method, and the recall will suffer. This problem becomes extremely significant when we analyze informal texts—forums, social networking websites, blogs, etc. Writing an informal text message, users often disregard punctuation and spelling rules, mistype, because of which the syntactic parser may fail to correctly analyze the structure of a sentence and build a semantic network. Users often express their sentiment through interjections, which are not a part of the syntactic tree; hence the semantic templates are of no use in this case. We call words that express sentiment but have no syntactic relation to the target of evaluation (or such relation has not been built by the parser) 'free sentiment'.

To solve this problem another method has been applied. We used an algorithm which is looking for free sentiment in the text using dictionaries (or profiles) of positive and negative lexicon, and if such sentiment has been found tries to relate it to the target object.

These two methods complement each other, with the semantic template method working first. In this regard, the classifier 'ignores' terms already found and related to the target object by templates, because we assume that the accuracy provided by the semantic templates is close to 100%.

As profiles for positive and negative classes we used corresponding filters, having removed context-dependent sentiment words and leaving only explicit emotional or evaluative vocabulary. For example, we removed verbs *УМЕРЕТЬ, ПРОИГРЫВАТЬ*, because although they are obviously negative in the context like: (11) *интернет умер;* (12) *оператор X проигрывает оператору Y*; but in another context, not related to the mobile telephony, they may be neutral and just state a fact. At the same time we enriched our profiles with interjections and other emotional expressions which cannot be syntactically related to the object of evaluation, for example: (13) *не надо так! что за нах; ни фига себе; ну как так можно, etc*.

Having found a sentiment, our algorithm was looking for an object of evaluation—a name of a mobile company—in the given text and ascribed this sentiment to the target. If several mobile operators were mentioned in the text, the appraisal was ascribed to the nearest operator. If both positive and negative sentiment was detected related to the same mobile provider mentioned, we gave preference to the negative sentiment, regarding positive expressions as sarcasm.

No machine learning had been used. The methods applied were based on linguistic analysis only.

4. Dataset

The training and test collection granted by organizers consisted of 5,000 labeled and 5,000 not labeled tweets containing sentiment-oriented opinions or positive and negative facts about telecom companies.

As the main goal of social networks sentiment analysis is to find sentiment-oriented opinions, we labeled texts containing reprints of news and additionally measured sentiment detection quality for the training collection with news reprints excluded. We excluded news texts from the final dataset because we believe that the difference in structure and vocabulary between formal (news) and informal (posts, blogs, tweets) texts is crucial. As a rule, in news texts authors don't express their attitude openly; news is more likely to contain coverage of events and facts, which can be interpreted as positive or negative for the newsmaker, rather than explicit sentiment; and therefore analyzing news demand a different approach. Furthermore, vocabulary of informal texts is quite different from vocabulary of formal texts.

That is why we additionally estimated the method performance on the collection with news reprints and companies' press releases excluded from the dataset. Since our method is based on linguistic analysis only, we did not use training collection.

5. Results

Initially, for the purpose of estimation of coincidence between assessors we asked our expert to evaluate the test collection manually and marked each reference to mobile phone companies as being positive, negative or neutral. Results of our expert's evaluation are presented in Table 1. F1-measure macro- and micro-averaged was used as a primary evaluation metric [1]. Additionally, for convenience, recall and precision are also present in the tables. As shown in Table 1, the estimation of tweets by our expert differed from one granted by the organizers. We consider the score given by our expert as the highest possible for an automatic sentiment detection system for the given collection. The agreement between our expert and organizers' labeling was higher when we excluded news from the dataset, which confirms our assumption that a different approach should be used for sentiment analysis of news.

Table 1. The estimation of coincidence between expert and assessors

	Macro-average			**Micro-average**		
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
With news	0.722	0.686	0.703	0.771	0.728	0.749
Without news	0.785	0.694	0.737	0.831	0.735	0.780

The results of all participants are shown in Fig. 1, our results are highlighted by bold lines and are labeled as "RCO". It is interesting that several methods probably based on different approaches demonstrate very similar high scores of F1 (about 0.5), nevertheless, these scores are sufficiently less than theoretical maximum that corresponds to coincidence between assessors (see bars "Expert" on Fig. 1). It could prove that automatic sentiment detection task is still a challenging problem.

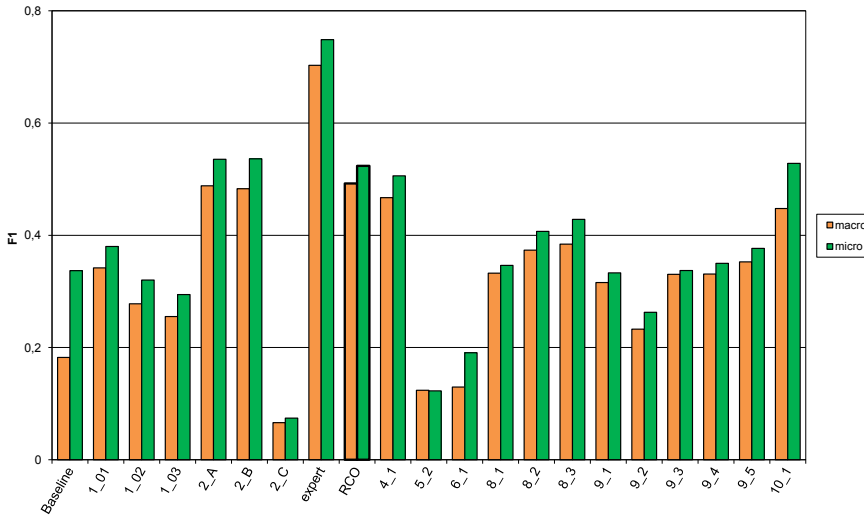

Fig. 2. Macro- and micro-averaged F1 measure calculated on test collection for all participants. The scores for our method are labeled as "RCO". The scores of expert's evaluation are labeled as "expert"

The detailed results of our method are presented in Table 2. We calculated recall, precision and F1 for original collection (labeled as "With news") and for collection with exclusion of messages contained news and press releases (labeled as "Without news"). For comparison, the best scores among the methods of all participants are presented.

Table 2. The performance of our method and best F1 measure among the methods of all participants

	Macro-average			Micro-average		
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
With news	0.436	0.566	0.480	0.451	0.585	0.509
Without news	0.465	0.562	0.492	0.475	0.583	0.524
Best result			0.492			0.536

6. Conclusion

Our combined linguistic method showed a very high quality, which roughly coincides with the best results of machine learning methods and hybrid approaches (combining machine learning with elements of syntactic analysis). In the future we are planning to add machine learning to our linguistic approach.

References

1. *Blinov P. D., Kotelnikov E. V.* (2014), Using distributed representations for aspect-based sentiment analysis, Dialog '14, Bekasovo.
2. *Chetviorkin I., Braslavski P. I., Loukachevitch N.* (2012), Sentiment analysis track at ROMIP 2011, Bekasovo.
3. *Ermakov A. E., Pleshko V. V.* (2009), Abstract Semantic Interpretation in Computer Text Analysis Systems [Semanticheskaya interpretatsiya v sistemakh kompyuternogo analiza teksta], Information Technologies [Informacionnye tehnologii], Vol. 6, pp. 2–7.
4. *Ermakov A. E.* (2009), Knowledge Extraction from Text and its Processing: Current State and Prospects [Izvlecheniye znaniy iz teksta i ikh obrabotka: sostoyaniye i perspektivy], Information Technologies [Informacionnye tehnologii], Vol. 7, pp. 50–55.
5. *Hu M., Liu B.* (2004), Mining and summarizing customer reviews, International Conference on Knowledge Discovery and Data Mining (ICDM).
6. *Jakob N., Gurevych I.* (2010), Extracting Opinion Targets in a Single-and Cross-Domain Setting with Conditional Random Fields, Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010).
7. *Kan D.* (2012), Rule-based approach to sentiment analysis at ROMIP '11 , Bekasovo.
8. *Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Yu., Ivanov V., Tutubalina E.* (2015), SentiRuEval Testing Object-Oriented Sentiment Analysis Systems in Russian.
9. *Popescu A., Etzioni O.* (2005), Extracting product features and opinions from reviews, Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP) .
10. *Polyakov P. Yu., Kalinina M. V., Pleshko V. V.* (2012), Research of applicability of thematic classification to the problem of book review classification. Dialog '12. Naro-Fominsk.
11. *Polyakov P. Yu., Frolov A. V., Pleshko V. V.* (2013), Using semantic categories in application to book reviews sentiment analysis, Dialog'13, Bekasovo.

АНАЛИЗ ТОНАЛЬНОСТИ ТВИТОВ О ТЕЛЕКОММУНИКАЦИЯХ И БАНКАХ НА ОСНОВЕ МЕТОДА МАШИННОГО ОБУЧЕНИЯ В РАМКАХ SENTIRUEVAL

Тутубалина Е. В. (tutubalinaev@gmail.com)[1],
Загулова М. А. (mazagulova@stud.kpfu.ru)[1],
Иванов В. В. (nomemm@gmail.com)[1, 2],
Малых В. А. (valentin.malykh@phystech.edu)[3]

[1]Казанский (Приволжский) Федеральный Университет (КФУ), Казань, Россия
[2]Институт информатики, Академия наук Татарстана, Казань, Россия
[3]ИСА РАН, Москва, Россия

Ключевые слова: анализ тональности текстов, SentiRuEval, твиттер, классификация твитов

A SUPERVISED APPROACH FOR SENTIRUEVAL TASK ON SENTIMENT ANALYSIS OF TWEETS ABOUT TELECOM AND FINANCIAL COMPANIES

Tutubalina E. V. (tutubalinaev@gmail.com)[1],
Zagulova M. A. (mazagulova@stud.kpfu.ru)[1],
Ivanov V. V. (nomemm@gmail.com)[1, 2],
Malykh V. A. (valentin.malykh@phystech.edu)[3]

[1]Kazan Federal University (KFU), Kazan, Russia
[2]Institute of Informatics, Tatarstan Academy of Sciences, Kazan, Russia
[3]Institute for Systems Analysis RAS, Moscow, Russia

This paper describes a supervised approach for solving a task on sentiment analysis of tweets about banks and telecom operators. The task was articulated as a separate track in the Sentiment Evaluation for Russian (SentiRuEval-2015) initiative. The approach we proposed and evaluated is based on a Support

Vector Machine model that classifies sentiment polarities of tweets. The set of features includes term frequency features, twitter-specific features and lexicon-based features. Given a domain, two types of sentiment lexicons were generated for feature extraction: (i) manually created lexicons, constructed from *Pros* and *Cons* reviews; (ii) automatically generated lexicons, based on pointwise mutual information between unigrams in a training set.

In the paper we provide results of our method and compare them to results of other teams participated in the track. We achieved 35.2% of macro-averaged F-measure for banks and 44.77% for tweets about telecom operators. The method described in the paper is ranked second and fourth among 7 and 9 teams, respectively. The best SVM setting after tuning parameters of the classifier and error analysis with common types of errors are also presented in this paper.

Key words: sentiment analysis, sentirueval, twitter, social media, tweet sentiment classification

1. Introduction

Sentiment analysis has received much attention in recent years due to its capability to identify people's opinions about products, named entities, facts (or events), and companies. This field of study has become important, especially due to the rapid growth of microblogging services such as Twitter, in which people talk about their personal experiences.

The goal of this task is to determine whether a given tweet is positive, negative or neutral according to its influence on the reputation of telecom or financial company. It is generally difficult to implement traditional sentiment analysis of user reviews since tweets collection could be noisy and each message is limited in length and could contain misspelling, slang and short forms of words. There have been a large number of research studies in the area of sentiment classification of short informal texts that are well described in (Martínez-Cámara, 2014). State-of-the-art papers have applied various feature sets from traditional text classification features (e.g., ngrams, part of speech tags, stems) to twitter-specific features (e.g., emoticons, hashtags, abbreviations) to handle the task in supervised manner (Kiritchenko et al., 2014). Since sentiment analysis in English has been explored in depth, there are not much research on sentiment classification of users' reviews in Russian. The recent works have focused on solving a task on sentiment analysis during ROMIP sentiment analysis tracks in 2011–2013 (Chetviorkin and Loukachevitch, 2013; Kotelnikov and Klekovkina, 2012; Blinov et al., 2013; Frolov et al., 2013).

In this study we report our submission to the SentiRuEval task. The approach is based on a Support Vector Machine model. The set of features includes term frequency features i.e. word ngrams, character ngrams; twitter-specific features and lexicon-based features. Since lexicon-based features are the most useful features for sentiment classification of tweets in English, we generated two types of sentiment lexicons. These two types are: manually created lexicons, constructed from *Pros* and *Cons* reviews in a particular domain; automatically generated lexicons, based on pointwise

mutual information between unigrams in training set. We achieve 44.77% of macro-average F-measure of for tweets about telecommunications companies and 35.2% for banks domain, that give improvements of 26.54% and 22.53% in macro F1-measure over official baseline results, respectively.

The rest of the paper is organized as follows. In Section 2 we introduce related work on sentiment classification of short informal texts. In Section 3 we describe proposed classifiers with a set of text classification features and twitter-specific features. Section 4 presents results of experiments. Section 5 provides error analysis. Finally, in Section 6 we discuss the results and future extensions of our work.

2. Related Work

Extracting information from short informal texts, such as tweets or sms messages, has received much attention in sentiment analysis (Go, 2009; Kiritchenko et al., 2014; Sidorov et al., 2013), event detection (Sakaki et al., 2010), problem extraction (Gupta, 2013), sarcasm detection (Davidov et al., 2010) and public sentiment tracking (O'Connor et al., 2010). Traditional approaches of sentiment classification were based on the presence of words or emoticons that indicated positive or negative polarity (Turney, 2002; Taboada, 2010; O'Connor et al., 2010). State-of-the-art papers have implemented hybrid approaches based on the use of machine learning techniques and lexical resources such as sentiment lexicons (Mohammad et al., 2013; Zhu et al., 2014; Kiritchenko et al., 2014; Evert, 2014). Recent studies showed that important machine learning features are bag-of-words unigrams and bigrams, and the use of tweet syntax features (e.g., hashtags, retweets and links) can improve the classification results (Barbosa and Feng, 2010). In (Kiritchenko et al., 2014) authors showed the importance of determining the sentiment of words in the presence of negation. They used separate lexicons for terms in affirmative and negated contexts.

Much work in sentiment analysis involves the use of existing sentiment lexicons and generation of lexical resources capturing the sentiment of words (Martínez-Cámara, 2014). The generation of lexicons range from manual approaches of annotating lexicons to fully automated approaches. In (Evert, 2014) authors used manual extension of existing sentiment lexicons and dictionaries of emoticons and internet slang. In (Mohammad et al., 2013) authors created automatically generated hashtag lexicon estimating sentiment scores for terms based on pointwise mutual information between terms and tweets with polarities. Inspired by these works, that describe supervised methods top-ranked in the SemEval-2014 task about sentiment analysis of tweets in English, we decided to create sentiment lexicons in similar way.

Sentiment analysis of texts in Russian is less studied. In (Chetviorkin and Loukachevitch, 2013) authors describe the first open sentiment task about sentiment classification of users reviews in Russian. Supervised methods, based on SVM classifier in a combination of manual or automatic dictionaries or rule-based systems, are top-ranked for reviews about movies, books, and digital cameras in the task. In (Frolov et al., 2013) authors proposed an approach based on special dictionaries and fact semantic filters in sentiment analysis of user reviews about books. In (Blinov et al.,

2013) authors used manual emotional dictionaries for each of three domains and showed benefits of machine learning method over lexical approach for user reviews in Russian. They reported that it was difficult to select particular machine learning method with the best results in all review domains.

3. Twitter-based Sentiment Classification

The task determines whether each tweet about a telecommunication companies (ttk) or banks contains a positive, negative, or neutral sentiment. We applied a machine-learning approach, based on bag-of-words model and a set of twitter-specific, lexicon-based features that are described in section 3.3.

The following examples illustrate situations in which different types of classification features appear in a tweet. Tweets such as "Лучи дикой ненависти вашей организации, ГОРИТЕ В АДУ *бешусь*" ("Sending rays of wild hatred to your organization, BURN IN HELL *rage*") contain strong negative polarities with regards to words with all characters in upper case. Tweets such as "Почему у дебетовой карты списали деньги просто так?!" ("Why was money from my debit card taken out with no reason?!") and "Сеть прыгает из E в 3G и обратно каждые 5 минут ((" ("Network shifts from E to 3G every 5 minutes ((") do not contain any positive and negative words. Therefore, a human annotator detects negative sentiment in each tweet with regards to the context of the tweet and whether the last symbols are emoticons, exclamation or question marks. Emoticons indicate positive or negative sentiment in short tweets, e.g. "@ sberbank всё спасибо, готово :)" ("@sberbank thank you, it is done :)") and "сбербанк продлил рассмотрение дела до 160 дней :(" ("Sberbank has prolonged consideration of the case till 160 days :("). Complex sentiment analysis in tweets such as "Проехать полгорода и узнать, что карта в другом из банков. Всегда мечтал .__." ("Crossed half the city to hear that my card is in another bank. I have always dreamed .__.") shows that some emoticons present sarcasm, which means that the opposite polarity of the positive word *мечтал* (*dreamed*) is denoted in the tweet. Presence of twitter-specific features such as URL or a retweet indicate to neutral context of tweets about news or informal messages, e.g. "mts коннект драйвер для android http://t.co/J3I5SNZuKM" ("mts connect driver for android *URL*") and "RT @Anna_Anna29: в билайне как узнать свой номер http://t.co/FpDZtLbdMZ" ("RT @Anna_Anna2: how to know your number in Beeline *URL*").

In the following examples we consider the use of sentiment lexicons, created manually and automatically. Manually created sentiment lexicons have been successfully applied in sentiment analysis in traditional approaches that detect whether a message contains positive or negative sentiment (Turney, 2002). The tweets such as "хреновый интернет, отвратительная работа с клиентами. Никогда не связывайтесь с этой шайкой" ("the lousy Internet, disgusting operation with clients. Never communicate with this gang") and "МТС пожелали хорошего дня, даже не попытались ничего продать. Уверовал в добро" ("MTS wished good day to me, didn›t even try to sell anything. I have believed in good") contain mention of domain-independent sentiment words like *отвратительный (disgusting)* and *хороший*

(good). Many tweets require deeper sentiment analysis due to difficult context of messages, e.g. the negative tweets "к вашему интернету хочется приложить подорожник" ("there is a wish to put a plantain to your internet") or "Билайн, отдай мне мой интернет" ("Beeline, give me my internet"). For these reasons, other sentiment lexicon is automatically created to cover such cases.

We tested three different learning algorithms: Naive Bayes, logistic regression (MaxEnt) and Support Vector Machine model (SVM). The squared euclidean norm L2 is selected as the standard regularizer for linear models. Based on the results obtained on the training sets we select SVM with default parameters[1] for tweet classification in banks domain.

3.1. Two Types of Sentiment Lexicons

We explore two main methods to construct sentiment lexicons: manual and automatic.

In the manual method we collected user rated reviews from otzovik.com: 3357 reviews about banks and 1928 reviews about telecom companies. To make corpus more accurate, we included only *Pros* reviews into positive corpus and *Cons* reviews into negative corpus. *Pros* (*Преимущества*) and *Cons* (*Недостатки*) are parts of a review that describe strong reasons why an author of the review likes or dislikes the product aspect, respectively. For each domain we selected the top K adverbs, adjectives, verbs, and nouns which have the highest frequencies in each corpus. Then we reduced noun words, expressing explicit aspects in a user review of particular domain due to neutral polarity of these aspects (e.g., *связь* (*connection*), *услуга* (*service*), *платеж* (*payment*), *скорость* (*speed*), *сотрудник* (*employee*)). In addition, we reduced the most common adjectives (e.g., *российский* (*russian*), *большой* (*big*), *абонентский* (*subscriber*)) and verbs expressing an action (e.g., *использовать* (*use*), *написать* (*write*), *подключать* (*connect*)). For each word we added other word forms. The dictionary consists of about 139 positive and 131 negative words in banks domain. The dictionary consists of about 68 positive and 168 negative words in telecom companies domain.

Following Mohammad et al. (2013) and other state-of-art approaches, automatically generated lexicons are based on sentiment score for each term w in the training test:

$$score\,(w) = PMI\,(w, pt) - PMI\,(w, nt)$$

$$PMI\,(w, pt) = log_2 \frac{p\,(w, pt)}{p\,(w) \times p\,(pt)}$$

where *PMI* is pointwise mutual information, *pt* denotes positive tweets, *nt* denotes negative tweets, *p (w)*, *p (pt)*, and *p (w, pt)* are probabilities of *w* occurs in positive corpus. The words with strong sentiment polarities have statistically significant difference between *PMI (w, pt)* and *PMI (w, nt)* in contrast to neutral words. For example, the pair of values (*PMI (w, pt)*, *PMI (w, nt)*) computed over the tweets in banks domain

[1] We have used the scikit-learn library in Python.

equals (−0.8016, 0.1450) for the neural word *еда* (*food*); (−15.2438, 1.5649) for the negative word *ущерб* (*loss*) and (2.1839, −19.2026) for the positive word *выгодный* (*profitable*). Since tweets contain low-frequency noisy words, we ignored terms that occurred less than three times in the training set.

3.2. Preprocessing for Short Informal Texts

Since raw tweets are usually informal and very noisy, the following preprocessing steps are performed. User mentions are normalized to @username. The morpho-syntactic analyzer[2] is applied to replace the words in the tweet with the base forms. We define negated context as a part of tweet between a negation (e.g., a particle *не* (*no*), a predicative expression *нет* (*not*)) word and a punctuation mark. Words with related negations (the words after negations) are modified in conjunction with the negation tag "neg_". We identify emoticons and replace them with corresponding sentiment expressions[3] (e.g., we replace ':-)' with *happy*, 'o_0' with *surprise* and ' ;-]' with *wink*).

3.3. Classification Features for Sentiment Classification of Tweets

Each tweet is represented as a feature vector; brief descriptions of the features that we use are presented below:

- **word n-grams**: unigrams (single words) and bigrams (multiword expressions) extracted from a tweet are used as the features. Features with document frequency greater than two are selected.
- **character n-grams**: lowercased characters n-grams for $n = 2, \ldots, 4$ with document frequency greater than two were considered for feature selection.
- **all-caps words**: the feature counts the number of words which contain all capitalized characters. Abbreviations of companies (e.g., *МТС* (*MTS*), *ВТБ* (*VTB*)) are excluded.
- **punctuation**: the features count the number of marks in sequences of exclamation marks, question marks, or a combination of these marks and the number of marks in contiguous sequences of dots. Sequences that consisted of more than one mark are considered for feature selection.
- **last symbol**: a binary feature indicates whether the last symbol of a tweet is an exclamation mark or a bracket.
- **emoticons**: four features are extracted: the number of positive emoticons; the number of negative emoticons; two binary features that indicate whether a last symbol of a tweet is a positive or negative emoticon, respectively.
- **twitter-specific features**: three binary features that indicate whether a tweet contains mentions of a twitter user, a retweet, and a presence of URL.

---

[2] We have used Mystem tool, url: https://tech.yandex.ru/mystem/

[3] We have used some sentiment expressions from http://en.wikipedia.org/wiki/List_of_emoticons

- **lexicon-based features:** for each of the two generated lexicons, the features are calculated as follows:
  - for the manual created lexicon we count the number of positive sentiment words, negative sentiment words. Sentiment words with negations change the sentiment polarity, e.g. a positive word with a negation suffix consider as a negative word.
  - for the automatically created lexicon four features are added: the count of words with non-zero scores; the sum of the words' sentiment scores normalized by words' count; the maximal sentiment score and minimum sentiment score in a tweet. Sentiment words with negations shift the sentiment score towards the opposite polarity.

4. Experimental Results

We used the training set of 5,000 annotated tweets for each domain provided for the SentiRuEval task. The final number of tweets in the testing collection is 4,549 tweets about banks and 3,845 tweets about telecom companies.

The official results obtained by our classifiers on the testing set are presented in Table 1. The table shows the official baseline results and the results of the method, ranked first according to macro-average F-measure as the main quality measure in the task (Loukachevitch et al., 2015). Macro-average F-measure is calculated as the average value between F-measure of the positive class and F-measure of the negative class. The classifier was trained to predict all three classes (positive, negative, and neutral), but this macro-averaged measure does not consider any correctly classifying neutral tweets. Our method is second among 7 teams with 14 runs in banks domain. The method is ranked fourth among 9 teams and fifth among 19 runs in telecom companies domain. The best approach has a 0.007% improvement in macro F1-measure over our approach in banks domain.

Table 1. Performance metrics in tweet classification task in two domains: telecom companies and banks

	telecom companies		**banks**	
	micro F	**macro F**	**micro F**	**macro F**
Best	0.536	0.488	0.343	0.359
Our approach	0.528	0.448	0.337	0.352
Official baseline	0.337	0.182	0.238	0.127

We also present feature ablation experiments on the testing set, removing one each individual feature category from the full set. Table 2 shows the results of the ablation experiments, each row shows macro-average precision, macro-average recall, and macro-average F-measure, calculated as the average value between corresponding measures of the positive and the negative classes. The most effective features are word n-grams for tweets about telecom companies. The most effective features are

based on character n-grams and emoticons in banks domain. The method also archives an improvement of 0.021% in F-measure after reducing word n-grams in banks domain and an improvement of 0.041% in F-measure after reducing word automatic lexicons in ttk domain. These improvements could be caused by a dynamic context of tweet messages about companies. The tweets of the training set were published in 2014, the tweets of the testing set were written in 2013.

Table 2. Experimental Results for the ablation experiments in two domains

	telecom companies (ttk)			**banks**		
	macro P	**macro R**	**macro F**	**macro P**	**macro R**	**macro F**
All features	0.443	0.471	0.447	0.538	0.279	0.352
w/o character n-grams	0.447	0.413	0.405	0.444	0.233	0.301
w/o emoticons	0.413	0.450	0.406	0.489	0.274	0.335
w/o both lexicons	0.419	0.553	0.475	0.496	0.276	0.337
w/o last symbol	0.458	0.379	0.390	0.509	0.274	0.340
w/o lexicon (manual ver.)	0.379	0.505	0.432	0.516	0.270	0.340
w/o lexicon (automatic v.)	0.427	0.569	0.488	0.426	0.292	0.343
w/o all-caps words	0.446	0.447	0.436	0.498	0.293	0.349
w/o punctuation	0.429	0.429	0.412	0.522	0.286	0.350
w/o twitter syntax features	0.447	0.441	0.443	0.491	0.289	0.351
w/o word n-grams	0.390	0.412	0.373	0.507	0.316	0.373

We also analyzed the significance of SVM tuning to our method. After shifting SVM's regularized regression method to elastic net that linearly combines the L1 and L2 penalties and the regularization term's alpha to 0.0001, the classifier had the improvements of 4–5% in macro F1-measures over our results with SVM's default parameters in both domains. The tuned classifier achieves a macro-average F-measure of 39.46% for banks domain and of 50.6% for tweets about telecommunications companies. The results show that careful tuning of the machine learning algorithm could obtain much better results.

5. Error Analysis

After error analysis we identify the following types of most frequent errors in tweet classification:

- misspelling and difficulty with transliteration of English text into Russian
- multiple hashtags
- emotional discussion of neutral topics
- insufficient size of sentiment lexicons (presence of out-of-lexicon words in the testing set)

From Table 3 shows that most of the errors are caused by insufficient information about context in positive or negative tweets about companies.

Table 3. Error types distribution

	Misspelling and transliteration	Multiple hashtags	Emotional discussion	Insufficient size of sentiment lexicons
telecom companies	20.40%	8%	14.90%	43%
banks	9%	1%	11%	64%

Tweets such as "Билайну труба короче" ("Beeline›s game›s over") contain hidden negative meaning like "game›s over" with the word "труба" ("a pipe"). Negative tweets such as "Самый безалаберный банк!" ("The most disorganized bank!") are missclassified due to low-frequency words like "безалаберный" that are not contained in the training set nor created lexicons.

We haven›t applied error correlation for cases of orthographic errors like *ацтой* (*rubbish*) and *чорд* (*damn*), while the correct spellings of these words are included in manually created lexicons. Tweets such as "Билайн. Дисконнектинг пипл." ("Beeline. Disconnecting people.") with transliterated words with strong negative polarity in English were misclassified as neutral. The analysis shows that misspelling caused less errors to tweets than elongated, transliterated words, and presence of asterisk (star symbol) in foul language words.

Hashtags such as *#отстойсвязь* (*#yourconnectionsucks*), *#мтсумри* (*#mtsdie*), *#люблюего* (*#loveit*) contain strong sentiment orientation. 8% of errors in telecommunications would be eliminated by splitting hashtags into words and then calculated the sentiment scores of hashtags.

Fourth type the errors is related to neutral tweets about telecom companies or banks, that contain positive or negative polarity about other topics (e.g., tweets about a *company's dress code,* friendly conversation or flirting with a company's worker*). Other type of such tweets is* a tweet describing some daily company's event: "Матч штабквартиры Вымпелком — Сибирь. Пока ведем!!! :)" ("Match of Vympelcom›s headquarters Vs Siberia. We›re winning!!! :)"). In all these cases the tweet about the company is neutral. Our classifiers haven›t considered such cases that affect up to 11% of errors about bank tweets, and 14.9% of errors in telecommunication tweets.

6. Conclusion

In this paper we described a supervised method for sentiment classification of financial or telecom twitter data with an emphasis on consumer experience. The proposed method exploits Support Vector Machines with term frequency features, twitter-specific features and lexicon-based features. Given a tweet the lexicon-based features were generated by checking whether a word is in sentiment lexicons, that were created both automatically and manually from user reviews. In order to produce an automatically

created lexicon, we used pointwise mutual information to calculate sentiment score and associate each word from a training set with a proper sentiment class.

We demonstrated that by using these features, classification performance increases from a baseline macro-averaged F-measures of 0.265 to 0.447 for telecoms and of 0.225 to 0.352 for banks. We plan to create large corpora of positive and negative tweets for the sake of improvement of the classifiers with automatically created lexicons.

Acknowledgments

This work was funded by the subsidy of the Russian Government to support the Program of competitive growth of Kazan Federal University and supported by Russian Foundation for Basic Research (RFBR Project 13-07-00773).

References

1. *Barbosa L., Feng J.* (2010), Robust sentiment detection on Twitter from biased and noisy data, Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, pp. 38–42
2. *Blinov P., Klekovkina M., Kotelnikov E., Pestov O.* (2013), Research of lexical approach and machine learning methods for sentiment analysis, Computational Linguistics and Intellectual Technologies, Vol. 2(12), pp. 48–58.
3. *Chetviorkin I., Loukachevitch N.* (2013), Evaluating Sentiment Analysis Systems in Russian, ACL 2013, p. 14.
4. *Davidov D., Tsur O., Rappoport, A.* (2010), Semi-supervised recognition of sarcastic sentences in twitter and amazon, Proceedings of the Fourteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, pp. 107–116.
5. *Evert S., Proisl T., Greiner P., Kabashi B.* (2014), SentiKLUE: Updating a Polarity Classifier in 48 Hours, SemEval 2014, Dublin, p. 551.
6. *Frolov A. V., Polyakov P. Yu., Pleshko V. V.* (2013), Using semantic filters in application to book reviews sentiment analysis, available at: www.dialog-21.ru/ digests/dialog2013/materials/pdf/FrolovAV.pdf
7. *Go A., Bhayani R., Huang L.* (2009), Twitter sentiment classification using distant supervision, CS224N Project Report, Stanford, pp. 1–12.
8. *Gupta N. K.* (2013), Extracting phrases describing problems with products and services from twitter messages, Computación y Sistemas, Vol. 17(2), pp. 197–206.
9. *Kiritchenko S., Zhu X., Mohammad S. M.* (2014), Sentiment analysis of short informal texts, Journal of Artificial Intelligence Research, Vol. 50, pp. 723–762.
10. *Kotelnikov, E. V., Klekovkina, M. V.* (2013), Sentiment analysis of texts based on machine learning methods [avtomaticheskij analiz tonal'nosti tekstov na osnove metodov machinnogo obuchenija]. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog, pp. 753–762.

11. *Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E.* (2015), SentiRuEval: testing object-oriented sentiment analysis systems in Russian, Proceedings of International Conference Dialog-2015, Moscow, pp. 3–9.
12. *Martínez-Cámara E., Martín-Valdivia M. T., Urena-López L. A., Montejo-Ráez A. R.* (2014), Sentiment analysis in twitter, Natural Language Engineering, Vol. 20(01), pp. 1–28.
13. *Mohammad S. M., Kiritchenko S., Zhu X.* (2013), NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets, Proceedings of the Second Joint Conference on Lexical and Computational Semantics (SEMSTAR'13), Atlanta, p. 2
14. *O'Connor B., Balasubramanyan R., Routledge B. R., Smith N. A.* (2010), From tweets to polls: Linking text sentiment to public opinion time series, ICWSM-11, Barcelona, pp. 122–129.
15. *Sakaki T., Okazaki M., Matsuo Y.* (2010), Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of the 19th international conference on World wide web, ACM, pp. 851–860.
16. *Sidorov G., Miranda-Jiménez S., Viveros-Jiménez F., Gelbukh A., Castro-Sánchez N., Velásquez F., Gordon J.* (2013), Empirical study of machine learning based approach for opinion mining in tweets, Advances in Artificial Intelligence, Vol. 7629, pp. 1–14.
17. *Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.* (2011), Lexicon-based methods for sentiment analysis, Computational linguistics, Vol.37(2), pp. 267–307.
18. *Turney P. D.* (2002), Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, Proceedings of the 40th annual meeting on association for computational linguistics, Philadelphia, pp. 417–424.
19. *Wilson T., Wiebe J., Hoffmann P.* (2005), Recognizing contextual polarity in phrase-level sentiment analysis, Proceedings of the conference on human language technology and empirical methods in natural language processing, pp. 347–354.

ВЫДЕЛЕНИЕ АСПЕКТНЫХ ТЕРМИНОВ В ОТЗЫВАХ С ИСПОЛЬЗОВАНИЕМ МОДЕЛИ УСЛОВНЫХ СЛУЧАЙНЫХ ПОЛЕЙ

Рубцова Ю. В. (yu.rubtsova@gmail.com),
Кошельников С. А. (koshelnikovsa@gmail.com)

Ключевые слова: извлечение аспектов, CRF, извлечение мнений, отзывы пользователей

ASPECT EXTRACTION USING CONDITIONAL RANDOM FIELDS

Rubtsova Y. V. (yu.rubtsova@gmail.com),
Koshelnikov S. A. (koshelnikovsa@gmail.com)

This paper describes the aspect extraction system that was presented at SentiRuEval-2015: aspect-based sentiment analysis of users' reviews in Russian. The proposed system uses a conditional random field algorithm for extracting aspects mentioned in the text. We used a set of morphological and syntactic features for machine learning and demonstrated that using lemmas as a feature can improve aspect extraction results. The system was used to perform two subtasks, Task A—automatic extraction of explicit aspects and Task B—automatic extraction of all aspects (explicit, implicit and sentiment facts), and tested on two domains—restaurants and cars. Both subtasks, A and B, in both domains have been completed with quite a high level of precision which meant that the system was capable of rather accurate recognition of aspect terms. But lower recall results implied that the system found enough aspect terms that could not be treated as aspects according to the gold standard. Our systems performed competitively and showed the results comparable to those of the other 10 participants.

Key words: aspect detection, aspect extraction, CRF, opinion mining, reviews

1. Introduction

With the popularity of blogs, social networks, and sites with user reviews of products and services growing every year web users post more and more reviews. As a result an enormous pool of reviews, evaluations, and recommendations in various domains has been accumulated that data attracts attention of both the researchers dealing with opinion mining, sentiment analysis and trend recognition and businessmen who are more interested in the practical application of reputation marketing.

Automatic sentiment analysis is mostly used at the following levels:

- Document level (Turney, 2002; Pang et.al, 2002; Rubtsova, 2014),
- Sentence or phrase level (Wilson et.al, 2009),
- Aspect level (Liu 2012; Zhang, Liu, 2014; Marrese-Taylor et.al, 2014).

As a rule people express their opinions not on the product or service as a whole but on some part, feature or characteristic thereof and that is the aspect that shall be extracted from the text and subjected to sentiment analysis. The aspect-level sentiment analysis can give us much more useful information on the author's opinion on various features of the product or service under analysis than sentiment analysis of the whole text.

Dialogue conference included Dialogue Evaluation section: evaluation of sentiment analysis systems for the Russian SentiRuEval (Loukachevitch et.al, 2015). The participants of the evaluation were to perform the following 5 subtasks:

A. Extract explicit aspects from the offered review,
B. Extract all the aspects from the offered review,
C. Perform sentiment analysis of the explicit aspects,
D. Categorize the aspects terms by predefined categories,
E. Evaluate the aspects categories as related to the offered review in general.

This paper describes the system that was used to perform Tasks A and B during SentiRuEval competition.

The rest of the paper is structured as follows. In Section 2 we discuss the current state of the art and different mechanisms of aspects extraction from product reviews. In Section 3 we describe our system. Section 4 demonstrates the performance of our system as compared to the results of systems of other SentiRuEval participants. Section 5 presents details conclusions and prospects of the future development.

2. Related work

There are four major approaches to extract aspects from texts. The first one is based on the frequency of nouns and/or noun phrases. Commonly people use similar terms to describe the features and their attitude to the products and differing terms used to describe other details (situation, required accompanying information) in their comments. Thus, counting frequency of the most common nouns and/or phrases in the texts of the same domain helps to extract explicit aspect terms from a large number of reviews (Hu and Liu, 2004). The precision level of that algorithm later has been improved by 22% (Popescu and Etzioni, 2005). As common words appear frequently in texts and are often defined as aspects, a filtering mechanism was invented to exclude most common non-aspect nouns and/or phrases from the analysis results (Moghaddam and Ester, 2011).

The second approach is based on simultaneous extraction of both sentiment words (user opinions) and aspects. As any opinion is expressed in relation to an object, by looking for sentiment words we can find aspects they relate to. Hu and Liu used this approach to find low-frequency aspects (Hu and Liu, 2004). Another approach is supervised machine learning. Generally, for the purposes of aspect extraction

supervised machine learning is focused on sequence labeling tasks because aspects and opinions on the products are often interrelated and constitute a sequence of words. The most wide-spread methods of supervised machine learning are hidden Markov modeling (HMM) (Jin et al., 2009) and conditional random fields (CRF) (Lafferty et al., 2001; Sutton and McCallum, 2006; Jakob and Gurevych, 2010). The fourth approach is unsupervised machine learning or topic modeling. Topic modeling assumes that each document consists of a mixture of topics and each topic is a probability distribution (Titov and McDonald, 2008; Brody and Elhadad, 2010). The most works on aspect extraction with the use of topic modeling approach are based on the methods of extended probabilistic latent semantic analysis (pLSA) model (Hofmann, 2001) and latent Dirichlet allocation (LDA) model (Blei et al., 2003).

To perform complex tasks such as simultaneous aspect extraction and sentiment analysis or simultaneous aspect extraction and categorization, one can use combination of different approaches such as max entropy и latent Dirichle allocation (Zhao W. X. et al, 2010) or semi supervised model with the topic modeling approach when user provides some seed words for a few aspect categories (Mukherjee and Liu, 2012).

3. System description

We participated into two evals:

- Extract the explicit aspects, i.e. extract a part of the object under analysis or one of its characteristics such as *engine* for the domain of cars or *service* for the domain of restaurants,
- Extract all the aspects of the object under analysis that includes extraction of explicit aspects, implicit aspects (an aspect + the author's unambiguous opinion on the aspect) and sentiment facts (when the author uses no opinion expressions but specifies a fact that unambiguously reveals his or her attitude to the object).

To extract opinion targets or aspects from sentences containing opinion expressions, we utilized CRF. CRF shows comparatively good results for the task of aspect extraction from reviews. For instance, for SemEval-2014 shared task related to aspect-based Sentiment Analysis, two best results have been obtained by systems that were based on CRF (Pontiki rt al., 2014).

Conditional Random fields is proposed as an undirected sequence model, which models a conditional probability p(Y|X) over hidden sequence Y given observation sequence X. That is, the conditional model is trained to label an unknown observation sequence X by selecting the hidden sequence Y which maximizes p(Y|X). As a software implementation of CRF, we utilized the Mallet tool (McCallum, 2002).

3.1. Pre-processing

Jakob and Gurevych (Jakob and Gurevych, 2010) represented the possible labels following the Inside-Outside-Begin (IOB) labelling schema: B-Target, identifying the

beginning of an opinion target; I-Target, identifying the continuation of a target, and O for other (non-target) tokens. Therefore as we used sequential labeling, we assigned a label to each word in the sentence where *s-e* indicated the start of an explicit aspect term, *c-e* indicated the continuation of an explicit aspect term, *s-i* indicated the start of an implicit aspect term, *c-i* indicated the continuation of an implicit aspect term (just as for facts-terms: *s-f* for start fact, *c-f* for continuation fact) and *O* indicated a non-aspect term.

To extract syntactic features (e.g. POS and lemma) described in the next section, we used TreeTagger for the Russian language (Sharoff et al., 2008).

We also noticed that car brands are often written in the Latin alphabet and/or contain numbers such as Nissan Micra or VAZ 2109. So for the collection of cars we added the rules that made it possible to recognize a full car name (or brand) as a single explicit term. As you can see in Table 3, this had some positive results—the System was ranked 3rd by the exact matching variant of F-measure.

We also converted all the capital letters into lowercase as the software tools may take *Engine* and *engine* as two different aspects, which is not true.

3.2. Features

Word

Strings of the current token were used as features. We extracted one previous and one subsequent word and used them as additional word features to get more information on the context the word is used in.

POS

The part-of-speech (POS) tag of the current token was used as a feature. Aspect terms are often expressed by nouns. POS tagging adds useful information on the part of speech the word belong to. To determine the part of speech we used TreeTagger—a tool that performs complete syntactic analysis. We reduce complete morphologic analysis up to the parts of speech such as *N* for *engine* and *V* for *driving*.

Lemma

The lemma of the current token was used as a feature. Due to the enormous number of word-forms in Russian language we added the normal form of word as a feature. To extract lemmas we also use a TreeTagger.

3.3. Architecture

We built two systems:

- System 1: CRF with all the above-mentioned labels. We used *s-e, c-e and O* labels for explicit aspect extraction to perform Task A and *s-e, c-e, s-i, c-i, s-f, c-f, O* to extract all the aspects for Task B.

- System 2: Combination of the results of two CRFs—CRF for extraction of explicit aspect terms and CRF for extraction of implicit aspect terms + sentiment facts terms (not explicit).
  Task A was performed using System 1 and Task B—using both systems.

4. Results

The results of Tasks A and B were evaluated by F-measure. Two cases of F-measure were calculated: exact matching and partial matching. Macro F1-measure means in this case calculating F1-measure for every review and averaging the obtained values. To measure partial matching, the intersection between gold standard and extracted term was calculated. Tables 1 to 4 demonstrate how the System performance of Task A and Tables 5 to 8 refer to performance of Task B. The results of the System were compared to the baseline and the two best results of SentiRuEval participants.

As you can see from Table 1 to 4, the System demonstrated high precision level in both domains (2nd position in Task A for both cars and restaurants by Precision metrics). It shall be noted that in the domain of cars the results were better when lemma feature was not in use—it may be concerned to pre-processing rules to the car collection. In Task B both systems built also showed a rather high precision level (see Table 5–8). In the domain of restaurants system 1 with word+pos+lemma features ranked 3rd amount all the participants by the partial matching case of F-measure.

Table 1 Task A results, Restaurant domain, exact matching

System	Precision	Recall	F-measure
baseline	0.557	0.6903	0.6084
№1	0.7237	0.5738	0.6319
№2	0.6358	0.6327	0.6266
Word+POS	0.661	0.515	0.5704
+lemma	0.6674	0.5417	0.5899

Table 2 Task A results, Restaurant domain, partial matching

System	Precision	Recall	F-measure
baseline	0.658	0.696	0.6651
№1	0.8078	0.6165	0.728
№2	0.7458	0.7114	0.7191
Word+POS	0.738	0.563	0.6277
+lemma	0.7485	0.5937	0.652

Table 3 Task A results, Car domain, exact matching

System	Precision	Recall	F-measure
baseline	0.5747	0.6287	0.5941
№1	0.76	0.6218	0.6761
№2	0.6619	0.656	0.6513
Word+POS	0.7109	0.5454	0.6075
+lemma	0.704	0.5785	0.6256

Table 4 Task A results, Car domain, partial matching

System	Precision	Recall	F-measure
baseline	0.7449	0.6724	0.6966
№1	0.7917	0.7272	0.7482
№2	0.8561	0.6551	0.7304
Word+POS	0.797	0.6047	0.6747
+lemma	0.7908	0.6485	0.6991

Table 5 Task B results, Restaurant domain, exact matching

System	Precision	Recall	F-measure
baseline	0.546577	0.647729	0.587201
№1	0.609432	0.600621	0.600128
№2	0.733599	0.513197	0.596179
System 1 Word+POS	0.639256	0.456334	0.52577
+lemma	0.639798	0.487202	0.546905
System 2 Word+POS	0.652145	0.458471	0.531644
+lemma	0.67152	0.491622	0.56153

Table 6 Task B results, Restaurant domain, partial matching

System	Precision	Recall	F-measure
baseline	0.671626	0.593093	0.619285
№1	0.756213	0.610754	0.667928
№2	0.668677	0.637097	0.645234
System 1 Word+POS	0.710428	0.493393	0.5692
+lemma	0.709915	0.529354	0.595303
System 2 Word+POS	0.724649	0.457863	0.547813
+lemma	0.752364	0.493553	0.585126

Table 7 Task B results, Car domain, exact matching

System	Precision	Recall	F-measure
baseline	0.597886	0.589612	0.588623
№1	0.7701	0.553546	0.636623
№2	0.656321	0.616423	0.630149
System 1 Word+POS	0.690826	0.476309	0.556107
+lemma	0.670594	0.518742	0.578086
System 2 Word+POS	0.718995	0.482064	0.568331
+lemma	0.701193	0.520375	0.589311

Table 8 Task B results, Car domain, partial matching

System	Precision	Recall	F-measure
baseline	0.783254	0.605976	0.674288
№1	0.814283	0.650998	0.714762
№2	0.795431	0.646999	0.704189
System 1 Word+POS	0.793637	0.53216	0.625502
+lemma	0.777257	0.584768	0.656113
System 2 Word+POS	0.808562	0.509979	0.61308
+lemma	0.782394	0.558153	0.638947

4.1. Error analysis

An analysis of the errors indicated some common mistakes: not recognized and excessively recognized. In general there is one more type of error for the task of aspect extraction—partially recognized aspect terms. Due to provided evaluation scripts we won't be able to observe third type of mistake. From Table 9, we can find that a major bunch of errors is related to not recognized aspect terms.

Table 9. Error type distribution for the task A (exact matching)

	Restaurants	Cars
Word+POS		
not recognized	65%	68%
excessively recognized	35%	32%
Word+POS+lemma		
not recognized	63%	65%
excessively recognized	37%	35%

We can also observe that adding Lemmas as a CRF feature leads to increasing excessively recognized terms. We compared two our systems and find out that second one can better deal with collocation. For instance it extracted “duck soup” («суп из утки») instead of just “soup” («суп») extracted by system 1. However collocations extraction is also a drawback of system 2 because occasionally it extracts to much irrelevant terms. For example “sea food pasta to husband” («пасту с морепродуктами, мужу»).

In the future, we would like to experiment with additional statistical and lexical features of CRF. Using additional text collections and topic modeling preprocessing can also make further improvements.

5. Conclusions

We presented two aspect extraction systems built on the basis of conditional random field algorithm. Realization of these systems demonstrated that preprocessing and use of lemmas for the Russian language as a CRF feature shows comparatively good the overall F-measure. The performance of our systems was comparable to the best results of SentiRuEval participants. Subsequently we are going to add statistical methods as a CRF feature. We are also planning to make a research and find a way to improve the recall results without reduce a precision.

References

1. *Blei D. M., Ng A. Y., Jordan M. I.* (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993–1022.
2. *Brody S., Elhadad N.* (2010). An unsupervised aspect-sentiment model for online reviews. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 804–812.
3. *Hofmann T.* (2001). Unsupervised learning by probabilistic latent semantic analysis. Machine learning, 42(1–2), pp. 177–196.
4. *Hu M., Liu B.* (2004). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168–177.
5. *Jakob N., Gurevych I.* (2010). Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, ACL, pp. 1035–1045.
6. *Jin W., Ho H. H., Srihari R. K.* (2009, June). OpinionMiner: a novel machine learning system for web opinion mining and extraction. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1195–1204.
7. *Lafferty J., McCallum A., Pereira F.* (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. In Proceedings of International Conference on Machine Learning (ICML-2001).

8. *Liu B.* (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), pp. 1–167.
9. *Loukachevitch N. V., Blinov P. D., Kotelnikov E. V., Rubtsova Yu. V., Ivanov V. V., Tutubalina E.* (2015), SentiRuEval: Testing Object-oriented Sentiment Analysis Systems in Russian, Proceedings of International Conference Dialog.
10. *Marrese-Taylor E., Velásquez J. D., Bravo-Marquez F.* (2014). A novel deterministic approach for aspect-based opinion mining in tourism products reviews. Expert Systems with Applications, 41(17), pp. 7764–7775.
11. *McCallum A. K.* (2002). MALLET: A Machine Learning for Language Toolkit.
12. *Moghaddam S., Ester M.* (2011). ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 665–674
13. *Mukherjee A., Liu B.* (2012). Aspect extraction through semi-supervised modeling. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pp. 339–348
14. *Pang B., Lee L., Vaithyanathan S.* (2002). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79–86.
15. *Pontiki M., Papageorgiou H., Galanis D., Androutsopoulos I., Pavlopoulos J., Manandhar S.* (2014). Semeval-2014 task 4: Aspect based sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval 2014, pp. 27–35.
16. *Popescu A. M., Etzioni O.* (2007). Extracting product features and opinions from reviews. In Natural language processing and text mining, pp. 9–28.
17. *Rubtsova Yu. V.* (2014). Development and research domain independent sentiment classifier, In SPIIRAS Proceedings, 5(36), pp. 59–77.
18. *Sharoff S, Kopotev M, Erjavec T, Feldman A, Divjak D.* (2008) Designing and evaluating Russian tagsets. In LREC.
19. *Sutton C., McCallum A.* (2006). An introduction to conditional random fields for relational learning. Introduction to Statistical Relational Learning. MIT Press.
20. *Titov I., McDonald R.* (2008). Modeling online reviews with multi-grain topic models. In Proceedings of the 17th international conference on World Wide Web, ACM, pp. 111–120.
21. *Turney P. D.* (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics, ACL, pp. 417–424.
22. *Wilson T., Wiebe J., Hoffmann P.* (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational linguistics, 35(3), 399–433.
23. *Zhang L., Liu B.* (2014). Aspect and entity extraction for opinion mining. In Data mining and knowledge discovery for big data, pp. 1–40.
24. *Zhao W. X., Jiang J., Yan H., Li X.* (2010). Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, ACL, pp. 56–65.

ВЫСОКОТОЧНЫЙ МЕТОД ИЗВЛЕЧЕНИЯ АСПЕКТНЫХ ТЕРМИНОВ ДЛЯ РУССКОГО ЯЗЫКА

Майоров В. (vmayorov@ispras.ru),
Аванесов В. (avanesov@ispras.ru),
Андрианов И. (ivan.andrianov@ispras.ru),
Астраханцев Н. (astrakhantsev@ispras.ru),
Козлов И. (kozlov-ilya@ispras.ru),
Турдаков Д. (turdakov@ispras.ru)

Институт Системного Программирования РАН, Москва, Россия

Ключевые слова: извлечение аспектных терминов, анализ эмоциональной окраски, извлечение именованных сущностей, автоматическое извлечение терминов

A HIGH PRECISION METHOD FOR ASPECT EXTRACTION IN RUSSIAN

Mayorov V. (vmayorov@ispras.ru),
Andrianov I. (ivan.andrianov@ispras.ru),
Astrakhantsev N. (astrakhantsev@ispras.ru),
Avanesov V. (avanesov@ispras.ru),
Kozlov I. (kozlov-ilya@ispras.ru),
Turdakov D. (turdakov@ispras.ru)

Institute for System Programming of RAS, Moscow, Russia

This paper presents a work carried out by ISPRAS on aspect extraction task at SentiRuEval 2015. Our team submitted one run for Task A and Task B and got best precision for both tasks for all domains among all participants. Our method also showed the best F1-measure for exact aspect term matching for task A for automobile domain and both for Task A and Task B for restaurant domain.

The method is based on sequential classification of tokens with SVM. It uses local, global, syntactic-based, GloVe, topic modeling and automatic term recognition features. In this paper we also present evaluation of significance of different feature groups for the task.

Key words: Aspect Extraction, Sentiment Analysis, NERC, Syntax Trees, Topic modeling, GloVe, Automatic Term Recognition

Introduction

This paper describes participation in aspect extraction tasks of SentiRuEval 2015, which focuses on detecting aspect terms in reviews for restaurant and cars.

Aspect extraction is a part of object-oriented sentiment analysis. An author of a text can have different opinions relative to specific properties of an object called aspects. Aspect terms represent these aspects in particular text.

Organizers of the competition divided all aspect terms into three types: Explicit aspects, Implicit aspects, Sentiment facts (Lukashevich N. V. et. al. 2015). According to the task definition, «Explicit aspects denote some part or characteristics of a described object such as staff, pasta, music in restaurant reviews. [...] Implicit aspects are single words or single words with sentiment operators that contain within themselves as specific sentiments as the clear indication to the aspect category. In restaurant reviews the frequent implicit aspects are such words as tasty (positive+food) [...] Sentiment facts do not mention the user sentiment directly, formally they inform us only about a real fact, however, this fact conveys us a user's sentiment as well as the aspect category it related to. For example, sentiment fact *отвечала на все вопросы* (*answered all questions*) means positive characterization of the restaurant service".

SentiRuEval dataset was annotated with these three subtypes of aspect terms and participants were asked to extract separately only explicit aspect terms and all aspect terms. In the rest of the paper we will refer to explicit aspect extraction task as "Task A" and all aspect extraction task as "Task B".

Our aspect extraction system uses supervised machine learning with support vector machines (SVM) to classify each token of a review into classes which denote beginning or middle of an aspects or term outside aspect. We train our classifier only on explicit aspect terms in order to perform Task A, and use union of results of three different classifiers trained for extraction of each type of aspects separately.

Main challenge was search of good feature space. We define three groups of features: local features computed in the bounds of one sentence; global features calculated for one document; and features that use external resources.

The paper is organized as follows: Section 1 gives brief overview of the related work; in Section 2 we present full description of our method and feature space it uses; Section 3 provides evaluation for different combination of features for each task; in the final section we make conclusion for this work.

1. Related work

Aspect extraction task has been widely studied in recent years. There are four main approaches (Liu, 2012) for this task. The first approach is to extract frequent nouns and noun phrases (Hu & Liu, 2004) (Popescu & Etzioni, 2007) (Scaffidi et al., 2007). The second one utilizes opinion word and target relations (Hu & Liu, 2004) (Qiu et al., 2011) (Poria et al. 2014). These methods are based on the idea that opinion words (i.e. words or phrases that specify sentiment) are related to aspect expressions in reviews. The third approach uses topic modeling (Mei et al., 2007) (Branavan et al., 2008) (Li, Huang & Zhu,

2010). The last approach is based on supervised machine learning. The most effective methods were shown to be sequential learning, namely Hidden Markov Models (Jin & Ho, 2009) and Conditional Random Fields (Jakob & Gurevych, 2010) (Choi & Cardie, 2010).

2. Method description

2.1. Overview

User's opinion could be expressed in several ways. Each aspect in datasets provided by organizers was marked with one of five types of expression: *relevant* (aspect term mention is relevant for current review object), *comparison* (aspect term is mentioned in comparison with another object), *previous* (aspect term is mentioned in comparison with previous experience), *irrealis* (aspect term is mentioned to describe hypothetical not materialized state of things) and *irony* (aspect term is mentioned with irony). We merged all marks except *relevant* to one class *"other"* due to relatively small number of aspects with marks *comparison*, *irony* etc.

At first we tokenize all reviews and transform task into sequence labelling task: given list of tokens assign sequence of tags to each element of sequence. Our method assigns one of five following classes to each token:

1. Out of aspect term
2. Beginning of *relevant* aspect term
3. Middle of *relevant* aspect term
4. Beginning of *other* aspect term
5. Middle of *other* aspect term

Each token is classified using SVM with L2 regularization. Used features are briefly described below.

We use Texterra system (Turdakov et. al., 2014) as general NLP tasks solution for text tokenization, PoS tagging and morphological analysis. Also we use MaltParser (Nivre et al., 2007) trained on SynTagRus[1] corpora for syntactic parsing.

2.2. Local features

Local features are features that are computed using only sentence. The main local feature used in our method is classification labels of tokens in left window of size 2.

We note that aspect extraction task is very similar to named entity recognition task (NERC). So, we use some features that are successfully used in supervised machine learning NERC method (Zhang & Johnson, 2003). Used NERC features are described in section 2.2.1.

Because Russian language has free word order, we decided to use sentence syntactic structure based features (see section 2.2.2).

[1] http://www.ruscorpora.ru/instruction-syntax.html

2.2.1. NERC features

We note that aspect extraction task is very similar to named entity recognition task. So, as basic features we choose following features that are described in (Zhang & Johnson, 2003).

Token prefixes and suffixes of length 1–4; token word forms, POS tags, morphological properties, lemmas in sentence window of size 2; whether a token placed at start of a sentence; token mask (all digits in token are replaced to a special character) and some token spelling features in window of size 2 (are all characters in uppercase / digits or punctuation marks / non letters / digits or letters; is any character a digit; is first character in uppercase).

2.2.2. Syntactic features

We use following features based on sentence syntactic structure. Distance in sentence syntactic tree between current token and other tokens in window of size 3. Lemma, POS tag and token morphological properties for parent token (in terms of syntactic tree) and for each child token. Classification labels assigned to parent and children tokens in left window.

2.3. Global features

Global features are features that are computed using the whole document. We use some of features used for supervised machine learning based NERC method (Ratinov & Roth, 2009): relative frequency of classification labels for all tokens having an equal word form with current one in left window of size 1000; relative frequency of having upper case first character for all tokens having an equal word form with current one in left window of size 200; relative frequency of POS tags, morphological properties and lemmas for all tokens having an equal word form with current one in left window of size 200.

2.4. Features based on external resources

2.4.1. Glove

We also use word to vector space embedding as features. In order to obtain the embedding to 50-dimensional vector space we train GloVe (Pennington, 2014) on Russian Wikipedia. Unfortunately, the vectors assigned to words are non-interpretable but they are known to be similar (in terms of Euclidean distance) for similar words. In order to obtain interpretable features we discover clusters of words using a fuzzy clustering approach—Gaussian Mixture Model (GMM) with 200 clusters—the number of clusters is optimized via Bayesian Information Criterion which is known to be a sufficient estimate for GMM (Roeder and Wasserman, 1995). And finally, the posterior distribution of clusters given for the vector embedding of a word is used as features.

2.4.2. Topic Modeling

Topic modeling is a fuzzy clustering approach usually used to clusterize documents by topics. The very basic topic model—Probabilistic Latent Semantic Analysis (Hofmann, 1999) was employed. This model assumes that every document was drawn

from a mixture of multinomial distributions over words. The components of the mixture are referred as topics. So, as a result of topic modeling, we obtain a distribution of words given the topic. Using Bayes' theorem we can easily compute the distribution of topics given the words. Finally, this distribution is used as a feature. The model was trained using a large unlabelled dataset of user's reviews. The tm[2] implementation was used.

2.4.3. Automatic Term Recognition

Since aspects are usually expressed by domain-specific terms, we check if the particular word-candidate is a part of domain-specific term. To do so, we apply methods for Automatic Term Recognition. Most of them, including those used by us, work as follows: take domain-specific text collection as an input; extract term candidates (n-grams filtered by the pre-specified part of speech patterns); compute features (e.g. frequency of term occurrences or tf-idf); and finally, classify or rank term candidates based on their feature vectors. In this work we skip the last step, i.e. we obtain the feature vector for each term candidate and then use it as follows: during a review text processing, we greedily search term candidates among word token sequences so that the longest appropriate term candidate is chosen, then we attach the corresponding feature vector to each word token from the matched sequence.

In particular, as an input text collection we use a combination of train and test data sets and also a set of documents crawled from the Web—namely, 44567 docs (82.6 Mb) from restoclub.ru for Restaurant domain and 7590 reviews (28.5 Mb) from otzovik.com for Automobile domain.

The following features are taken: 3 well-known features: Frequency; TF-IDF; C-Value (Frantzi et al., 2000) in modification that supports single-word terms (Lossio-Ventura et al., 2013); and 4 our features (Astrakhantsev, 2014): ExistsInKB—a boolean feature indicating if a term candidate is presented in Wikipedia; Link Probability—a probability of term candidate to be a hyperlink in Wikipedia; Key concept relatedness—a semantic relatedness value computed over Wikipedia to automatically found key concepts; PUATR—result of probabilistic Positive-Unlabeled classifier trained on top 100 term candidates (found by special method based on frequencies of nested occurrences) as positives and other candidates as unlabeled with all previously described features.

3. Evaluation

3.1. SVM parameter estimation

For SVM parameter estimation we perform 10-fold cross-validation on available training data with C parameter from 0.001 to 0.2 with step 0.001 in two settings (see Fig. 1). First settings is testing on training data (red line), the second settings is normal cross-validation (green line). As one can see, when to $C < 0.045$ F1 score grow for both train and test data.

For $C > 0.45$ F1 measure for train is grow and for test data it is stay almost same, thus we decided that this is frontier between over and underfitting. Thus we set C equals to 0.45

[2] https://github.com/ispras/tm

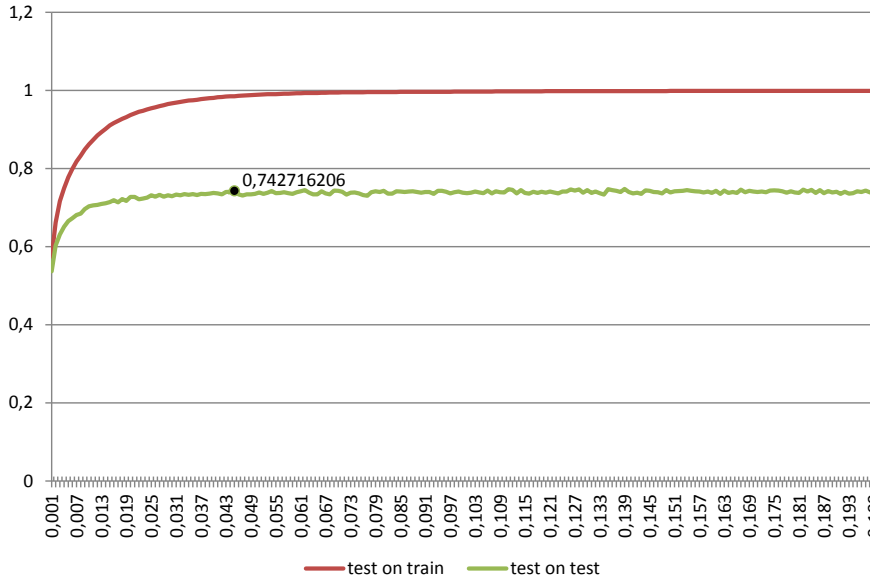

Fig 1. Method performance with different SVM parameter

3.2. Evaluation of feature groups impact

In order to understand impact of each feature group we sequentially remove each group from our feature set and measure method quality for task A. For quality measurement we perform repeated 10 times 10-fold cross-validation and compute 95% confidence interval for each quality metric. Results for automobile domain is presented in Table 1. Table 2 presents results for restaurant domain.

Table 1. Quality results (95% confidence intervals) for different features sets for Automobile domain (Task A)

	exact matching			partial matching		
features set	**precision**	**recall**	**f1**	**precision**	**recall**	**f1**
all	(0.7061; 0.7197)	(0.6500; 0.6618)	(0.6773; 0.6885)	(0.8080; 0.8200)	(0.6975; 0.7114)	(0.7493; 0.7604)
all—GloVe	(0.7107; 0.7249)	(0.6467; 0.6584)	(0.6775; 0.6891)	(0.8139; 0.8257)	(0.6888; 0.7015)	(0.7467; 0.7573)
all—TM	(0,7031; 0,7166)	(0,6427; 0,6548)	(0,6720; 0,6832)	(0,8061; 0,8181)	(0,6882; 0,7016)	(0,7431; 0,7540)
all—ATR	(0,7032; 0,7165)	(0,6414; 0,6537)	(0,6713; 0,6826)	(0,8066; 0,8185)	(0,6915; 0,7059)	(0,7452; 0,7565)
all—global	(0,7046; 0,7185)	(0,6509; 0,6633)	(0,6771; 0,6888)	(0,8068; 0,8190)	(0,6990; 0,7129)	(0,7496; 0,7609)

features set	exact matching			partial matching		
	precision	recall	f1	precision	recall	f1
all—syntactic	(0,7132; 0,7276)	(0,6582; 0,6706)	(0,6850; 0,6968)	(0,8155; 0,8268)	(0,7069; 0,7203)	(0,7579; 0,7685)
all—NERC	(0,6373; 0,6535)	(0,5120; 0,5253)	(0,5682; 0,5810)	(0,7655; 0,7798)	(0,5812; 0,5968)	(0,6611; 0,6747)

Table 2. Quality results (95% confidence intervals) for different features sets for Restaurant domain (Task A)

features set	exact matching			partial matching		
	precision	recall	f1	precision	recall	f1
all	(0,7122; 0,7260)	(0,6546; 0,6692)	(0,6830; 0,6942)	(0,7894; 0,8024)	(0,7012; 0,7143)	(0,7439; 0,7530)
all—GloVe	(0,7146; 0,7284)	(0,6529; 0,6672)	(0,6831; 0,6943)	(0,7956; 0,8080)	(0,6963; 0,7093)	(0,7438; 0,7528)
all—TM	(0,7140; 0,7281)	(0,6450; 0,6591)	(0,6786; 0,6896)	(0,7912; 0,8045)	(0,6884; 0,7017)	(0,7375; 0,7467)
all—ATR	(0,7106; 0,7247)	(0,6514; 0,6662)	(0,6805; 0,6920)	(0,7887; 0,8020)	(0,6972; 0,7106)	(0,7414; 0,7507)
all—global	(0,7118; 0,7256)	(0,6551; 0,6696)	(0,6831; 0,6941)	(0,7893; 0,8017)	(0,7045; 0,7177)	(0,7458; 0,7545)
all—syntactic	(0,7101; 0,7249)	(0,6570; 0,6713)	(0,6833; 0,6949)	(0,7947; 0,8076)	(0,7009; 0,7144)	(0,7461; 0,7554)
all—nerc	(0,6325; 0,6488)	(0,5109; 0,5265)	(0,5656; 0,5795)	(0,7426; 0,7571)	(0,5775; 0,5929)	(0,6504; 0,6627)

As one can see, only NERC features make a meaningful contribution to the method. Other feature groups are not so significant.

3.3. Method performance on SentiRuEval testing dataset

The quality of proposed method trained on all available training data with all described feature groups are presented in table 3 for task A and in table 4 for Task B. These results are obtained by SentiRuEval organizers.

Table 3. SentiRuEval Task A experiment results

Domain	exact matching			partial matching		
	precision	recall	f1	precision	recall	f1
Automobile	0.760041	0.621793	0.676118	0.856055	0.655098	0.730366
Restaurant	0.723656	0.573800	0.631871	0.807759	0.616549	0.689096

Table 4. SentiRuEval Task B experiment results

Domain	exact matching			partial matching		
	precision	**recall**	**f1**	**precision**	**recall**	**f1**
Automobile	0.770100	0.553546	0.636623	0.866178	0.549210	0.659989
Restaurant	0.733599	0.513197	0.596179	0.814496	0.479988	0.590601

Conclusion

We have described aspect term extraction system, which employs SVM with a broad set of features. This system perform with high precision and good F1-measure on all settings and showed one of the best results among 21 runs received for aspect extraction tasks of SentiRuEval.

In addition, we made evaluation of impact of different feature groups and found that features used for named entity recognition are most useful for aspect extraction too. We also found that removing some features could slightly improve results of cross-validation. One of the reasons for such phenomena is sparsity of feature set. Therefore we can guess that feature selection and dimensionality reduction could improve quality of the proposed method. In addition, we should note that due to lack of time, we estimated SVM parameter only on full feature set and use it for all experiments. However SVM parameter estimation for each feature combination can improve overall performance of the system. This make a slot for future improvement of the proposed method.

References

1. *Astrakhantsev N.,* (2014), Automatic term acquisition from domain- specific text collection by using Wikipedia, The Proceedings of ISP RAS [Trudy ISP RAN], vol. 26, issue 4, P. 7–20.
2. *Fangtao L., Huang M., Zhu X.,* (2010), Sentiment analysis with global topics and local dependency, in Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2010).
3. *Frantzi K., Ananiadou S., Mima H.,* (2000), Automatic recognition of multi-word terms:. the c-value/nc-value method, International Journal on Digital Libraries, 3(2), 115–130.
4. *Jin Wei, Hung Hay Ho,* (2009), A novel lexicalized HMM-based learning framework for web opinion mining, in Proceedings of International Conference on Machine Learning (ICML-2009).
5. *Hofmann T.,* (1999), Probabilistic latent semantic indexing, in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, (pp. 50–57). ACM.

6. *Liu B.,* (2012), Sentiment analysis and opinion mining, Synthesis Lectures on Human Language Technologies, *5*(1), 1–167.
7. *Lossio-Ventura J. A., Jonquet C., Roche M., Teisseire M.,* (2013), Combining c-value and keyword extraction methods for biomedical terms extraction, In LBM'2013: 5th International Symposium on Languages in Biology and Medicine (pp. 45–49).
8. *Mei Qiaozhu, Xu Ling, Matthew Wondra, Hang Su, ChengXiang Zhai,* (2007), Topic sentiment mixture: modeling facets and opinions in weblogs, in Proceedings of International Conference on World Wide Web (WWW-2007).
9. *Niklas J., Gurevych I.,* (2010), Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields, in Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010).
10. *Nivre J., Hall J., Nilsson J., Chanev A., Eryigit G., Kübler S., Marsi E.,* (2007), MaltParser: A language-independent system for data-driven dependency parsing, Natural Language Engineering, *13*(02), 95–135.
11. *Pennington J., Socher R., Manning C. D.,* (2014), Glove: Global vectors for word representation, Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014), *12*.
12. *Poria S., Cambria E., Ku L. W., Gui C., Gelbukh A.,* (2014), A rule-based approach to aspect extraction from product reviews, SocialNLP 2014, 28.
13. *Popescu A. M., Etzioni O.,* (2007), Extracting product features and opinions from reviews, In Natural language processing and text mining (pp. 9–28), Springer London.
14. *Qiu G., Liu B., Bu J., Chen C.,* (2011), Opinion word expansion and target extraction through double propagation, Computational linguistics, 37(1), 9–27.
15. *Ratinov L., Roth D.,* (2009) Design challenges and misconceptions in named entity recognition, Proceedings of the Thirteenth Conference on Computational Natural Language Learning / Association for Computational Linguistics, pp. 147–155.
16. *Roeder K., Wasserman L.,* (1997), Practical Bayesian density estimation using mixtures of normals, Journal of the American Statistical Association, 92(439), 894–902.
17. *Scaffidi C., Bierhoff K., Chang E., Felker M., Ng H., Jin C.,* (2007), Red Opal: product-feature scoring from reviews, In Proceedings of the 8th ACM conference on Electronic commerce, pp. 182–191
18. *Turdakov D., Astrakhantsev N., Nedumov Y., Sysoev A., Andrianov I., Mayorov V., Fedorenko D., Korshunov A., Kuznetsov S.* (2014), Texterra: A Framework for Text Analysis, Proceedings of the Institute for System Programming of RAS [Trudy ISP RAN], volume 26, Issue 1, pp. 421–438.
19. *Yejin C., Cardie C.,* (2010), Hierarchical sequential learning for extracting opinions and their attributes, in Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2010).
20. *Zhang T., Johnson D.* (2003), A robust risk minimization based named entity recognition system, Proceedings of the seventh conference on Natural language learning at HLTNAACL 2003-Volume 4 / Association for Computational Linguistics, pp. 204–207.

ГЛУБОКИЕ РЕКУРРЕНТНЫЕ НЕЙРОННЫЕ СЕТИ ДЛЯ АСПЕКТНО-ОРИЕНТИРОВАННОГО АНАЛИЗА ТОНАЛЬНОСТИ ОТЗЫВОВ ПОЛЬЗОВАТЕЛЕЙ НА РАЗЛИЧНЫХ ЯЗЫКАХ

Тарасов Д. С. (dtarasov3@gmail.com)

Интернет-портал reviewdot.ru, Казань, Россия

Ключевые слова: рекуррентные нейронные сети, анализ тональности, извлечение аспектных терминов, унифицированный подход

DEEP RECURRENT NEURAL NETWORKS FOR MULTIPLE LANGUAGE ASPECT-BASED SENTIMENT ANALYSIS OF USER REVIEWS

Tarasov D. S. (dtarasov3@gmail.com)

Reviewdot research,Kazan,Russian Federation

Deep Recurrent Neural Networks (RNNs) are powerful sequence models applicable to modeling natural language. In this work we study applicability of different RNN architectures including uni- and bi-directional Elman and Long Short-Term Memory (LSTM) models to aspect-based sentiment analysis that includes aspect terms extraction and aspect term sentiment polarity prediction tasks. We show that single RNN architecture without manual feature-engineering can be trained to do all these subtasks on English and Russian datasets. For aspect-term extraction subtask our system outperforms strong Conditional Random Fields (CRF) baselines and obtains state-of-the-art performance on Russian dataset. For aspect terms polarity prediction our results are below top-performing systems but still good for many practical applications.

Keywords: recurrent neural networks, sentiment polarity, aspect term extraction, unified approach

1. Introduction

In many practical natural language processing (NLP) systems, it is desirable to have one architecture that can be quickly adapted to different tasks and languages without the need to design new feature sets. Recent success of deep neural networks in general and deep RNNs in particular offers hope that this goal is now within reach. RNNs were applied to a number of English NLP problems, demonstrating their superior capabilities in slot-filling task [Mesnil et al, 2013] and opinion mining [Irsoy and Cardie, 2014].

While these results are promising it is still unclear if RNNs can now be used to replace other models in practical multi-purpose NLP system and if single RNN architecture can efficiently perform many different tasks.

Our work evaluates a number of RNN architectures on three different datasets: ABSA Restaurants (English) dataset from SemEval-2014 [Pontiki et al, 2014] and two Russian datasets (Restaurants and Cars) from SentiRuEval-2015.

We show that RNN performance on aspect terms extraction is close to state-of-the art and results on sentiment prediction, while being significantly behind top performing systems, outperform strong baselines and offer sufficient performance for use in practical applications. We discuss factors that contribute to RNNs results and suggest possible directions to further improve their performance on these tasks.

2. Related work

Sentiment analysis or opinion mining is the computational study of people's attitudes toward entities. In user reviews analysis two principal tasks are aspect terms extraction and aspect sentiment polarity prediction.

Aspect term extraction methods could roughly be divided into supervised and unsupervised approaches. In supervised approach aspect extraction is usually seen as sequence labeling problem, and often solved using variants of conditional random field (CRF) [Ganu et al, 2009;Breck and Cardie, 2007] methods, including semi-CRF systems, that operate at the phrase level and thus allow incorporation of phrase-level features [Choi and Cardie, 2010]. Such systems currently hold state-of-the arts results in term extraction from user reviews [Pontiki et al, 2014]. However, success of CRF and semi-CRF approaches depends on the access to rich feature sets such as dependency parse trees, named-entity taggers and other preprocessing components, that are often not readily available in underresourced languages such as Russian. Unsupervised approaches to term extraction attempts to cut cost and effort associated with manual feature selection and annotation of training data. These approaches typically utilize topic models such as Latent Dirichlet Allocation to learn aspect terms [Brody and Elhadad, 2010]. Their performance however, is below that of supervised systems trained on in-domain data.

Quite recently recurrent neural network models were proposed to solve sequence tagging problems, including similar opinion mining task [Irsoy and Cardie, 2014], demonstrating results superior to all previous systems. Importantly, these results were obtained using only word vectors as features, eliminating the need for complex feature-engineering schemes.

Similarly, sentiment polarity prediction subtask is solved within supervised and unsupervised learning frameworks. State-of-the-art performance on term polarity detection is currently obtained by using support vector machines (SVM) with rich feature sets that include parse trees and large opinion lexicons, together with preprocessing to resolve negation [Pontiki et al, 2014]. Unsupervised methods in sentiment analysis usually focus on construction of polarity lexicons for which number of approaches currently exists [Brody and Elhadad, 2010], and then applying heuristics to determine term polarity.

Neural network based methods were developed recently to detect document level and phrase-level sentiment, including tree-based autoencoders [Socher et al, 2011;2013] and convolutional neural networks [dos Santos and Gatti, 2014;Blunsom et al, 2014] and Elman-type RNNs were applied to sentence-level sentiment analysis with promising results [Wenge et al, 2014].

3. Methodology

3.1. Datasets

SemEval-2014 ABSA Restaurants dataset [Pontiki et al, 2014] was downloaded through MetaShare (http://metashare.ilsp.gr:8080/). This dataset is a subset of (Ganu et al, 2009) dataset. It contains English statements from restaurants reviews (3,041 in training and 800 sentences in test set) annotated for aspect terms occurring in the sentences, aspect term polarities, and aspect category polarities.

Russian Restaurants dataset and corresponding Cars dataset released by SentiRuEval-2015 organizers to participants consist of similarly annotated reviews in Russian with a number of important differences. These datasets contain whole reviews, rather than individual sentences and are annotated with three categories of aspect terms “explicit” (roughly equivalent to SemEval-2014 notion of aspect term), “implicit” and so called “polarity facts”—statements that don't contain explicit judgments but nevertheless tell something good or bad about aspect in question.

Auxiliary dataset for training Russian unsupervised word vectors was constructed from concatenation of unannotated cars and restaurants reviews, provided by SentiRuEval-2015 organizers and 300,000 user reviews of various consumer products from reviewdot.ru database (obtained by crawling more than 200 online shops and catalogs).

3.2. Evaluation of human disagreement

As a part of this work we decided to evaluate human disagreement on SentiRuEval-2015 Restaurants dataset because we found many examples that seemed ambiguous. To do this we split dataset in two parts (70/30) and appointed two human judges. Human judges were given “annotation guidlines” sent by SentiRuEval organizers and 70% of annotated dataset. They then were asked to annotate remaining 30% with aspect terms (explicit, implicit and polar facts) and results were compared to original annotation using evaluation metrics described in “metrics” section.

3.3. Recurrent neural networks

A recurrent neural network [Elman, 1990] is a type of neural network that has recurrent connections. This makes them applicable for sequential prediction tasks, including NLP tasks. In this work, we consider simple Elman-type networks and Long-Short Term Memory architectures.

3.3.1. Simple recurrent neural network

In an Elman-type network (Fig. 1a), the hidden layer activations $h(t)$ at time step t are computed by transformation of the current input layer $x(t)$ and the previous hidden layer $h(t-1)$. Output $y(t)$ is computed from the hidden layer $h(t)$.

More formally, given a sequence of vectors {x (t)} where t = 1..T, an Elman-type RNN computes memory and output sequences:

$$h(t) = f(Wx(t) + Vh(t-1) + b) \tag{1}$$

$$y(t) = g(Uh(t) + c) \tag{2}$$

where f is a nonlinear function, such as the sigmoid or hyperbolic tangent function and g is the output function. W and V are weight matrices between the input and hidden layer, and between the hidden units. U is the output weight matrix, b and c are bias vectors connected to hidden and output units. $h(0)$ in equation (1) can be set to constant value that is chosen arbitrary or trained by backpropagation.

Deep RNN can be defined in many possible ways [Pascanu et al, 2013], but for the purposes of this work deep RNNs were obtained by stacking multiple recurrent layers on top of each other.

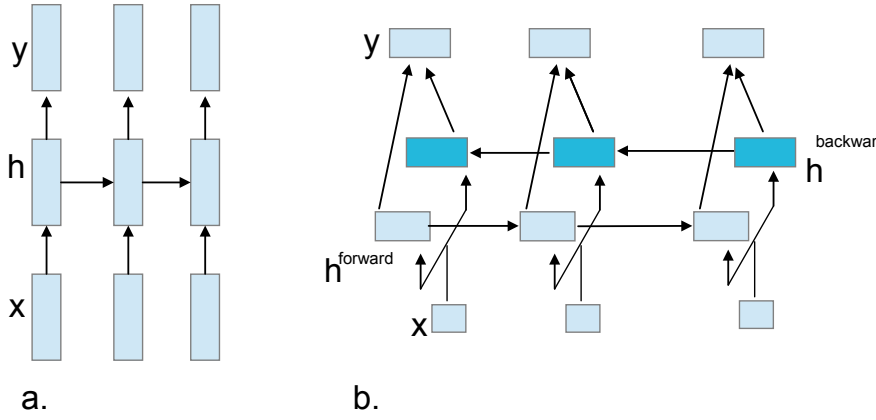

Figure 1. Recurrent neural networks, unfolded in time in three steps
a. Simple recurrent neural network
b. Bidirectional recurrent neural network

3.3.2. Long Short Term Memory

The structure of the LSTM [Hochreiter and Schmidhuber, 1997] allows it to train on problems with long term dependencies. In LSTM simple activation function f from above is replaced with composite LSTM activation function. Each LSTM hidden unit is augmented with a state variable $s(t)$ The hidden layer activations correspond to the 'memory cells' scaled by the activations of the 'output gates' o and computed in following way:

$$h(t) = o(t) * f(c(t)) \tag{3}$$

$$c(t) = d(t) * (c(t-1) + i(t)) * f(Wx(t) + Vh(t-1) + b) \tag{4}$$

where * denotes element-wise multiplication, $d(t)$ is dynamic activation function that scales state by "forget gate" and $i(t)$ is activation of input gate.

3.3.3. Bidirectional RNNs

In contrast with regular RNN that can only consider information from past states, bidirectional recurrent neural network (BRNN) [Schuster and Kuldip, 1997] can be trained using all available input data in the past and future. In BRNN (Fig. 1b) neuron states are split in a part responsible for positive time direction (forward states) and and a part for the negative time direction (backward states):

$$h(t)^{forward} = f(W^{forward} x(t) + V^{forward} h^{forward}(t-1) + b^{forward}) \tag{5}$$

$$h(t)^{backward} = f(W^{backward} x(t) + V^{backward} h^{backward}(t+1) + b^{backward}) \tag{6}$$

$$y(t) = g(U^{forward} h^{forward} + U^{backward} h^{backward} + c) \tag{7}$$

3.3.4. Training

All networks were trained using backpropagation through time (BPTT) [Werbos, 1990] algorithm with mini-batch gradient descent with one sentence per mini-batch as suggested in [Mesnil et al, 2013]. For sequence labeling tasks loss function was evaluated at every timestep, while for classification tasks such as term polarity prediction, loss function was only evaluated at the position corresponding to terms whose polarity was being predicted.

3.3.5. Regularization

To prevent overfitting small Gaussian noise was added to network inputs. Large networks were also regularized with dropout [Hinton et al, 2012] a recently proposed technique that omits certain proportion of the hidden units for each training sample.

3.4. Word embeddings

Real-valued embedding vectors for words were obtained by unsupervised training of Recurrent Neural Network Language Model (RNNLM) [Mikolov et al, 2010]. English embeddings of size 80 trained on 400M Google News dataset were downloaded

from RNNToolkit (http://rnnlm.org/) website. Russian embeddings of same size were trained using auxiliary dataset described above, using same method. Russian text was preprocessed by replacing all numbers with #number token and all occurrences of rare words were replaced by corresponding word shapes.

3.5. Evaluation metrics

For term extraction tasks where term boundaries are hard to identify even for humans, it is generally recommended to use soft measures like Binary Overlap that counts every overlapping match between a predicted and true expression as correct [Breck et al, 2007], and Proportional Overlap that computes partial correctness proportional to the overlapping amount of each match [Johansson and Moschitti, 2010].

From the description of SemEval-2014 task it appears that exact version of F- measure was used (only exact matches count), even though organizers note that "In several cases, the annotators disagreed on the exact boundaries of multi-word aspect terms".

For Russian SentiRuEval-2015 datasets, due to somewhat different annotation approach, multi-word (4 and 5 word terms) are quite common and human disagreement is quite large (as will be shown below). SentiRuEval-2015 organizers adopt two metrics for aspect-term extraction—main (based on exact count) and secondary (based on proportional overlap).

In SentiRuEval-2015 datasets all terms are tagged as "relevant" (related to target entity), or irrelevant (related to something else) and official metrics only count identification of relevant terms as correct. We feel that identification of aspect term and classification it as "relevant" or not are two fundamentally different tasks and should be measured separately. Due to extremely low presence (less than 5%) of irrelevant terms, their exclusion is quite hard for machine learning algorithm to achieve, and finding algorithms that do that well is a problem of significant theoretical interest. Such systems cannot be identified using official metrics, since contribution of "relevance" detection to overall F1 value is rather small.

For the purposes of this paper unless otherwise stated, we apply F-measure based on proportional overlap to facilitate comparison of results obtained on different datasets. For English Restaurants ABSA dataset F-measure is computed on Test dataset of 800 sentences (that was not used in development of models). For Russian datasets, as test data were not available at the time of this work, we separate development set of 5000 words and use 7-fold cross-validation on remaining data, similar to [Isroy and Cardie, 2014] approach. Since we participated in a number of SentiRuEval-2015 tracks, official results according to SentiRuEval-2015 metrics are also shown for comparison and discussion purposes.

For classification tasks such as sentiment polarity and aspect category detection tasks, macro average of F-measure cannot be used due to the fact that some categories (such as "conflict" polarity, named "both" in Russian dataset) are extremely rare (Russian Restaurant dataset contains less than 80 instances of "both" polarity per 3000 instances of aspect terms). F-measure for such categories is subject to huge sampling error, and can also be undefined (with zero precession and recall), making macro

average value undefined also. To prevent this problem from occurring SemEval-2014 uses Accuracy instead of F-measure. SentiRuEval-2015 organizers use F1 micro average in addition to macro average. In this paper, for classification tasks we show overall accuracy, computing macro-average as additional measure where possible.

3.6. Baselines

For term extraction task we consider several baseline systems: simple feed-forward multi-layer perceptron (MLP), frame-level MLP (a feed-forward MLP with inputs of only word embedding features within a word context window), logistic regression using word embedding features, and CRF using stemmed words and POS-tags as features.

4. Results and Discussion

4.1. Aspect term extraction task

Tables 1–3 summarize our results on aspect term extraction. Initially, for Russian Restaurant dataset, we found it very difficult to improve upon simple CRF baseline. Manual examination of annotation revealed a number of inconsistent decisions in provided training data, for example in one place term “официантка Любовь” (“servant Lubov”) was tagged as a whole, while in other similar case servant name was not tagged as part of the term. That led us to evaluation of human disagreement that appeared to be very close to baseline results, making term extraction very formidable challenge.

Nevertheless, we found that augmented forward RNN outperforms CRF baseline on explicit aspect extraction and deep LSTM model outperforms both CRF and Frame-NN baselines on all subtasks, while simple BRNN while providing reasonable good results, failed to improve on these baselines in contrast with English dataset. We think that inconsistent annotation in training set leads to over-fitting in simple BRNNs, because complex local models are learned before long time dependencies in the data can be discovered.

Overall, as shown in Table 2, our system obtains best result in extraction of all aspects terms according to proportional measure and best result in extraction of all aspect terms on cars dataset according to exact measure, while holding second-best result on restaurants dataset. These good results, should, however, be interpreted with caution due to relatively small number of participants, general lack of strong competitors and poor quality of the data (at least in Restaurant domain).

Therefore, to better understand system capabilities we evaluated our system on English dataset of SemEval-2014. The advantage of this dataset is that it is carefully cleaned from errors and also results of state-of-the-art systems are readily available for comparison. Table 3 demonstrates that in this dataset our system did not obtain top results. Still, LSTM performance is quite good (equivalent to 6th best result of 28 total participants).

Table 1. F-measure (proportional overlap) on SentiRuEval dataset, evaluated using 7-fold cross-validation

Mehod	SentiRuEval Restaurants dataset				SentiRuEval Cars dataset			
	Explicit	Implict	Fact	Macro average	Explicit	Implict	Fact	Macro average
Human Judge 1	69.1	58.7	33.0	53.6	—	—	—	—
Human Judge 2	65.0	62.3	27.0	51.4	—	—	—	—
CRF baseline	68.2	57.7	24.0	49.96	—	—	—	—
Logistic regression	54.0	43.0	3.0	33.3	70.1	75.4	15.2	53.6
MLP	64.5	53.6	18.2	45.3	75.8	82.2	34.8	64.2
Frame-NN	67.9	61.4	**26.1**	51.8	76.0	**83.0**	33.0	64.0
Simple RNN	68.4	58.5	20.0	48.9	75.2	81.3	30.1	62.2
Simple RNN augmented with one future word	68.9	60.0	25.3	51.4	75.8	82.0	31.4	63.1
Simple RNN augmented with one future word + dropout	71.1	56.0	20.1	49.06	76.0	82.1	24.3	60.8
Bidirectional RNN	69.8	61.2	19.1	50.3	76.1	81.5	32.1	63.2
Bidirectional LSTM	**73.5**	**64.3**	23.5	**53.76**	**77.0**	82.5	**36.3**	**65.3**

Table 2. F-measure on SentiRuEval Test dataset (according to SentiRuEval results)

Method	SentiRuEval Restaurants dataset				SentiRuEval Cars dataset			
	Proportional		Exact		Proportional		Exact	
	Explicit	All	Explicit	All	Explicit	All	Explicit	All
BRNN	67.2	52.2	57.5	64.5	71.7	70.4	61.7	59.9
LSTM	71.9	**60.0**	62.6	**66.8**	—	—	—	—
LSTM, Depth 2	—	—	—	—	**74.8**	**71.4**	65.1	63.0
Other systems best result	**72.8**	59.6	**63.1**	59.5	73.0	65.9	**67.6**	**63.6**

Table 3. Results on English SemEval ABSA Restaurant dataset (computed by us, using SemEval official metrics), reference results are taken from [Pontiki et al, 2014]

Method	F1 value
baseline	47.15
CRF with words and POS tags features	75.20
6th-best result	79.60
Top result	84.01
BRNN	76.20
LSTM	79.80

4.2. Sentiment polarity prediction task

Tables 4–6 summarize sentiment polarity results. Here more complex systems generally obtain superior results to simpler methodologies.

Using SentiRuEval-2015 official metrics we obtain second-best result in explicit aspect term polarity prediction on cars-dataset and third-result in restaurants dataset (unfortunately, results from our top systems were not included in official results due to errors that we made in data format. This error only became apparent after release of test sets and thus impossible to correct). Also, relatively poor results are partially explained by the fact that our system was optimized to all-term polarity prediction task, leading to suboptimal performance on explicit-term only task (information about official metrics were released by organizers with delay and we were not able to adapt all systems due to time and resource constraints). On English ABSA Restaurant dataset we obtain accuracy of 69.7, significantly below best results, but still reasonable.

Even through our results here are below top systems, they are reasonable good and have some theoretical value in demonstrating that exactly same architecture can be used both for sequence tagging and polarity prediction tasks. It also worth noting, that we used neither sentiment lexicon, nor special preprocessing steps for negation (we found that RNNs under certain conditions are capable to learn negation just from training data). Another important finding here that using hidden layer activations of RNNLM model as features instead of word vectors considerably improves overall system performance. Our hypothesis is that next-word prediction task of RNNLM includes the need to understand word dependencies—a knowledge that shown to be crucial in aspect-term polarity prediction task. This knowledge from unsupervised model can thus be leveraged by supervised RNN to enhance performance.

Table 4. Results on all-terms polarity prediction task on SentiRuEval dataset (F1 macro average on positive and negative classes and overall accuracy over all terms)

	Restaurants		Cars	
Method	**Macro F1**	**Accuracy**	**Macro F1**	**Accuracy**
TDNN N=3	61.0	57.4	55.2	56.2
RNN	63.1	59.2	57.1	57.1
BRNN	67.4	60.3	60.3	56.9
LSTM	70.2	61.1	62.4	58.0
LSTM + RNNLM features *	74.1	62.5	65.0	59.1

* Obtaining by using hidden layer activations of RNNLM

Table 5. Results on explicit-only terms polarity classification (according to SentiRuEval-2015 official results)

Method	**Restaurants**	**Cars**
BRNN	61.9	64.7
LSTM + RNNLM features	—	65.3
Top result	82.4	74.2

Table 6. Results for English terms polarity classification on ABSA Restaurants SemEval-2014 dataset (according to our evaluation metrics)

Method	Accuracy
Baseline	64.00
Sentiment lexica over dependency graphs *	69.50
BRNN	65.10
LSTM	69.70
Top result	82.92

* Value taken from [Wettendorf et al, 2015]

5. Conclusions

In aspect term extraction task recurrent neural networks models demonstrate excellent perfomance. On Russian SentiRuEval-2015 dataset our system obtained best result in extraction of all aspects terms according to proportional measure and best result in extraction of all aspect terms on cars dataset according to exact measure, while holding second-best result on restaurants dataset. On English SentEval-2014 dataset, we obtained reasonable good results, equivalent to 6th best known result on this dataset. From all RNN models, best results were obtained with deep bidirectional LSTM with 2 hidden layers.

For aspect term polarity predictions, we obtained second best result on SentiRuEval-2015 car dataset and third best result on SentiRuEval-2015 car restaurants dataset. We also obtained good results on all terms polarity prediction. To our knowledge, this is first time when LSTM models were applied to aspect term polarity prediction with reasonable good results.

Overall, our work demonstrates that RNN models are useful in aspect-based sentiment analysis and can be utilized for rapid prototyping and deployment of opinion mining systems in different languages.

Acknowledgments

Author want to thank Ekatertina Izotova for help with data format conversion, anonymous reviewers for helpful comments and SentiRuEval organizers for preparing and running evaluation and thus making this work possible.

References

1. *Blunsom, P., Grefenstette, E., & Kalchbrenner, N.* (2014). A convolutional neural network for modelling sentences. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.
2. *Breck E., Choi Y., Cardie C.* (2007). Identifying expressions of opinion in context. In IJCAI, pp. 2683–2688.
3. *Brody S., Elhadad N.* (2010). An unsupervised aspect-sentiment model for online reviews. In Proceedings of NAACL, pp. 804–812, Los Angeles, California
4. *Choi Y., Cardie C.* (2010). Hierarchical sequential learning for extracting opinions and their attributes. In Proceedings of the ACL 2010 Conference Short Papers, pp. 269–274.
5. *dos Santos, C. N., & Gatti, M.* (2014). Deep convolutional neural networks for sentiment analysis of short texts. In Proceedings of the 25th International Conference on Computational Linguistics (COLING), Dublin, Ireland.
6. *Elman J.* (1990). Finding structure in time. Cognitive science, 14(2):179–211.
7. *Ganu, G., Elhadad, N., & Marian, A.* (2009, June). Beyond the Stars: Improving Rating Predictions using Review Text Content. In WebDB (Vol. 9, pp. 1–6).
8. *Hinton G. E., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinov R.* (2012). Improving neural networks by preventing coadaptation of feature detectors. arXiv preprint arXiv:1207.0580
9. *Hochreiter, S., & Schmidhuber, J.* (1997). Long short-term memory. Neural computation, 9(8), 1735–1780.
10. *Irsoy O., Cardie C.* Opinion Mining with Deep Recurrent Neural Networks (2014). EMNLP, Doha, Qatar. pp. 720–728
11. *Johansson R., Moschitti A.* (2010). Syntactic and semantic structure for opinion expression detection. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning, pp. 67–76. Association for Computational Linguistics.
12. *Mesnil, G., He, X., Deng, L. & Bengio, Y.* (2013). Investigation of recurrent neural network architectures and learning methods for spoken language understanding. In INTERSPEECH pp. 3771–3775 : ISCA.
13. *Mikolov T., Karafi´at M., Burget L., Cernock´y J., Khudanpur S.* (2010). Recurrent neural network based language model. In INTERSPEECH, pp. 1045–1048.
14. *Pascanu, R., Gulcehre, C., Cho, K., & Bengio, Y.* (2013). How to construct deep recurrent neural networks. arXiv preprint arXiv:1312.6026.
15. *Pontiki M., Papageorgiou, H., Galanis, D., Androutsopoulos, I., Pavlopoulos, J., & Manandhar, S.* (2014). Semeval-2014 task 4: Aspect based sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (pp. 27–35).
16. *Schuster M., Kuldip K. P.* (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11):2673–2681.
17. *Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., & Manning, C. D.* (2011, July). Semi-supervised recursive autoencoders for predicting sentiment distributions. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 151–161). Association for Computational Linguistics.

18. *Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C.* (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the conference on empirical methods in natural language processing (EMNLP) (Vol. 1631, p. 1642).
19. *Wenge R., Baolin P., Yuanxin O., Chao Li, Zhang X.* (2004) Structural information aware deep semi-supervised recurrent neural network for sentiment analysis. Frontiers of Computer Science, pp. 1–14, http://dx.doi.org/10.1007/s11704-014-4085-7
20. *Werbos, P. J.* (1990). Backpropagation through time: what it does and how to do it. Proceedings of the IEEE, 78(10), 1550–1560.
21. *Wettendorf C., Jegan R., Korner A., Zerche J.* (2014) SNAP: A Multi-Stage XML-Pipeline for Aspect Based Sentiment Analysis In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 578–584

ИЗВЛЕЧЕНИЕ АСПЕКТОВ, ТОНАЛЬНОСТИ И КАТЕГОРИЙ АСПЕКТОВ НА ОСНОВАНИИ ОТЗЫВОВ ПОЛЬЗОВАТЕЛЕЙ О РЕСТОРАНАХ И АВТОМОБИЛЯХ

Иванов В. В. (nomemm@gmail.com),
Тутубалина Е. В. (tutubalinaev@gmail.com),
Мингазов Н. Р. (nicrotek547@gmail.com),
Алимова И. С. (alimovaIlseyar@gmail.com)

Казанский Федеральный Университет, Казань, Россия

Ключевые слова: анализ тональности текстов, SentiRuEval, отзывы пользователей, извлечение аспектов, категории аспектов

EXTRACTING ASPECTS, SENTIMENT AND CATEGORIES OF ASPECTS IN USER REVIEWS ABOUT RESTAURANTS AND CARS

Ivanov V. V. (nomemm@gmail.com),
Tutubalina E. V. (tutubalinaev@gmail.com),
Mingazov N. R. (nicrotek547@gmail.com),
Alimova I. S. (alimovaIlseyar@gmail.com)

Kazan Federal University, Kazan, Russia

This paper describes a method for solving aspect-based sentiment analysis tasks in restaurant and car reviews subject domains. These tasks were articulated in the Sentiment Evaluation for Russian (SentiRuEval-2015) initiative. During the SentiRuEval-2015 we focused on three subtasks: extracting explicit aspect terms from user reviews (tasks A), aspect-based sentiment classification (task C) as well as automatic categorization of aspects (task D).

In aspect-based sentiment classification (tasks C and D) we propose two supervised methods based on a Maximum Entropy model and Support Vector Machines (SVM), respectively, that use a set of term frequency features in a context of the aspect term and lexicon-based features. We achieved 40% of macro-averaged F-measure for cars and 40,05% for reviews about restaurants in task C. We achieved 65.2% of macro-averaged F-measure for cars and 86.5% for reviews about restaurants in task D. This method ranked first among 4 teams in both subject domains. The SVM classifier is based on unigram features and pointwise mutual information to calculate category-specific score and associate each aspect with a proper category in a subject domain.

In task A we carefully evaluated performance of a method based on syntactic and statistical features incorporated in a Conditional Random Fields model. Unfortunately, the method did not show any significant improvement over a baseline. However, its results are also presented in the paper.

Key words: aspect-based sentiment analysis, sentirueval, user reviews, aspect extraction, aspect categories

1. Introduction

Over the past decade, opinion mining (also called sentiment analysis) has been an important concern for Natural Language Processing (NLP). Since online reviews significantly influence people's decisions about purchases, sentiment identification has a number of applications, including tracking people's opinions about movies, books, and products, etc.

In this study we describe our approaches for solving a task on sentiment analysis, which was formulated as a separate track in the Sentiment Evaluation for Russian (SentiRuEval-2015) initiative. The SentiRuEval task concerns aspect-based sentiment analysis of user reviews about restaurants and cars. The task consists of several subtasks: aspect extraction (tasks A and B), sentiment classification of explicit aspects (task C), and detection of aspects categories and sentiment summarization of a review (tasks D and E). The primary goal of the SentiRuEval task is to find words and expressions indicating important aspects of a restaurant or a car based on user opinions and to classify them into polarity classes and aspect categories (Loukachevitch et al., 2015).

There have been a large number of research studies in the area of aspect-based sentiment analysis, which are well described in Liu (2012) and Pand and Lee (2008). Traditional approaches in opinion mining are based on extracting high-frequency phrases containing adjectives from manually created lexicons (Turney, 2002; Popescu and Etzioni, 2007). State-of-the-art papers have implemented probabilistic topic models, such as Latent Dirichlet Allocation (LDA), and Conditional Random Field (CRF) for multi-aspect analysis tasks (Moghaddam and Ester, 2012; Choi and Cardie, 2010). Sentiment analysis in English has been explored in depth and there are many well-established methods and general-purpose sentiment lexicons that contain a few thousand terms. However, research studies of sentiment analysis in Russian have been less successful. In 2011–2013 studies have focused on solving a task on sentiment analysis during ROMIP sentiment analysis tracks (Chetviorkin and Loukachevitch, 2013; Kotelnikov and Klekovkina, 2012; Blinov et al., 2013; Frolov et al., 2013).

We use the Conditional Random Fields model applied to the aspect extraction task. In task C for aspect-based sentiment classification we propose a method based on a Maximum Entropy model that uses a set of term frequency features in a context of the aspect term and lexicon-based features. The classifier for aspect category detection is based on a SVM model with a set of category-specific features. We achieved 40% of macro-averaged F-measure for cars and 40,05% for reviews about restaurants in task C. We achieved 65.2% of F-measure for cars and 86.5% for reviews about restaurants in task D.

The rest of the paper is organized as follows. In Section 2 we introduce related work on sentiment analysis. In Section 3 we describe proposed approaches. Section 4 presents results of experiments. Finally, in Section 5 we discuss the results.

2. Related Work

In this paper, we focus on the detection of the three major cores in a review: aspect terms, sentiment about these aspects, and aspects' categories. During the last decade, a large number of methods were proposed to identify these elements.

Aspect term extraction. There are several widely used methods that treat the task as a classification problem (Popescu et al., 2005), as a sequence labeling problem (Jakob and Gurevych, 2010; Kiritchenko et al., 2014; Chernyshevich, 2014), as a topic modeling or a traditional clustering task (Moghaddam and Ester, 2012; Zhao et al., 2014). The classification problem is to determine whether nouns and noun phrases are target of an opinion or not. Popescu et al. (2005) used syntactic patterns in relation with sentiment from general-purpose lexicons to identify high-frequency noun phrases. Poria et al. (2014) proposed a rule-based approach, based on knowledge and sentence dependency trees. These approaches are limited due to lower results on extracting low-frequency aspects or hand-crafted dependency rules for complex extraction. In (Kiritchenko et al., 2014; Chernyshevich, 2014) the authors proposed two modifications of a standard scheme for sequence labeling models.

Aspect term polarity. Most of the early approaches for classifying aspects rely on seed words or a manually generated lexicon that contains strongly positive or strongly negative words. Turney (2002) proposed an unsupervised method, based on a sentiment score of each phrase that is calculated as the mutual information between the phrase and two seed words. Recent papers have widely applied machine learning methods to solve the tasks of sentiment classification (Pang et al., 2002; Pang and Lee, 2008; Blinov et al., 2013; Kiritchenko et al., 2014). Moghaddam and Ester (2012) proposed extensions of the LDA model to extract aspects and their sentiment ratings by considering the dependency between aspects and their sentiment polarities. However, topic models achieve lower performance on multi-aspect sentence classification than the SVM classifier in three different domains (Lu et al., 2011).

Aspect category detection. Automatic categorization of explicit aspects into aspect categories has been studied as the task of sentiment summarization. Moghaddam and Ester (2012) investigated it as a part of a latent aspect mining problem. There have been some works on grouping aspect terms from review texts for the sentiment analysis in the task 4 of the international workshop on Semantic Evaluation (SemEval-2014). The task was evaluated with the F-measure and the best results were achieved by SVM classifies with bag-of-words features and information from unlabeled reviews (Pontiki et al., 2014; Kiritchenko et al., 2014).

Several studies about sentiment analysis have been done in Russian, related to evaluation events of Russian sentiment analysis systems (Chetviorkin and Loukachevitch, 2013). Frolov et al. (2013) proposed a dictionary-based approach with fact semantic filters for sentiment analysis of user reviews about books. Blinov et al.

(2013) showed benefits of machine learning method over lexical approach for user reviews in Russian and used manual emotional dictionaries.

3. System description

In this section we describe our approaches for three tasks of aspect-based sentiment analysis of user reviews about restaurants and cars. The CRF model was used for automatic extraction of explicit aspects (task A). We applied machine-learning approaches for the tasks C and D, based on bag-of-words model and a set of lexicon-based features that are described in Section 3.2 and 3.3, respectively. The morpho-syntactic analyzer Mystem was used for text normalization at the preprocessing step.

3.1. Aspect Extraction

The goal of aspect extraction is to detect extract major explicit aspects of a product (task A). Since the task can be seen as a particular instance of the sequence-labeling problem, we employ Conditional Random Fields (Lafferty et al., 2001).

Explicit aspects denote some part or characteristics of a described object such as *передний привод* (*front-wheel drive*), *руль* (*steering wheel*), *динамика* (*dynamics*) in cars reviews; *столик* (*table*), *официант* (*waiter*), *блюдо* (*dish*) in reviews about restaurants. In the following examples we consider user phrases about explicit aspects.

We use Inside-Outside-Begin scheme and Passive Aggressive algorithm for training CRF; brief description of the features used to represent the current token w_i are presented below: the current token w_i, the current token w_i within a window ($w_{i-2}, \ldots, w_{i+2}$); the part of speech tag of the current token; the part of speech tag of the token within a tag window ($tag_{i-2}, \ldots, tag_{i+2}$); the number of occurrences of the tokens in the training set; the presence of the token in manually created domain-dependent dictionaries.

3.2. Aspect-based sentiment classification

The task of sentiment classification aims to predict polarity (positive, negative, neutral, or both) of each aspect from the product reviews. We applied the Maximum Entropy classifier with default parameters, based on a bag-of-words model and a set of lexicon-based features that are described in Section 3.2.2.

The following examples illustrate the aspects (marked in italic) with different polarities from the reviews. Some phrases like "*персонал* улыбчивый, приветливый." ("smiling, friendly staff"), "общее впечатление: отличная *машина*" ("overall impression: great car") or "просторный *салон*, *удобно* сидеть пассажиру сзади" ("*spacious interior*, a passenger could sit *comfortably* behind the driver's seat") contain strong positive or negative context near the aspect term. Therefore, such cases could

be correctly classified extracting bigrams in the phrases. Complex analysis of sentiment phrases such as "заказывал *бифштекс*, нет слов как *вкусно*" ("I ordered a beefsteak, there are no words to describe just how *tasty* this was") and "в городском цикле *компьютер* будет показывать очень неприятные цифры" ("in the city the *computer* will show very unpleasant figures") shows that there is a distance between the polarity words *вкусно (tasty), неприятные* (*unpleasant*) and the aspect terms. We use combinations of the aspect term and a context term to classify these cases. Difficult phrases with both sentiments such as "отмечу некоторую *жесткость сидений*, но привыкаешь, главное сидеть удобно" ("I note some *rigidity of the seats*, but you get used to it, the main thing is sit conveniently") or "*горячее* неплохое, но на гриль было непохоже" ("*hot dishes* are quite good, but not similar to a grill") could be recognized by presence of the conjunction word *но* (*but*).

Given a context of the aspect term, two types of word bigrams are generated for feature extraction: (i) context bigrams, using a text within a context window of the aspect term; (ii) aspect-based bigrams as a combination of the aspect term itself and a context word within the context window. The context window of the aspect term w_i denotes a sequence $(w_{i-4}, \ldots, w_{i+4})$.

3.2.1. Manually created sentiment lexicon

We collected user rated reviews from otzovik.com: 7,526 reviews about restaurants and 4,952 reviews about cars. To make corpus more accurate, we included only *Pros* reviews with an overall rating 5 into positive corpus and *Cons* reviews with an overall rating 1 or 2 into negative corpus. *Pros (Преимущества)* and *Cons (Недостатки)* are parts of a review that describe strong reasons why an author of the review likes or dislikes the product, respectively. For each domain we selected the top K adverbs, adjectives, verbs, reducing noun words that express aspects, action verbs and most common adjectives. The manually created dictionary consists of about 741 positive and 362 negative words in restaurants domain and includes 1,576 positive and 741 negative words in cars domain. We combine two dictionaries to achieve better evaluation results.

For lexicon-based features we use the following scores: each word in the sentence is weighted by its distance from the given aspect:

$$score(w) = \frac{sc(w)}{e^{|i-j|}}$$

where i, j is the positions of the aspect term and the word, $sc(w)$ is the sentiment word's score, that equals 1 for positive words and -1 for negative words, extracted from the sentiment dictionary.

3.2.2. Classification Features for Aspect Term Polarity

Each review is represented as a feature vector, for each aspect features are extracted from the aspect and its context in a sentence. A brief description of the features that we use is presented below:

- **character n-grams**: lowercased characters n-grams for $n = 2, \ldots, 4$ with document frequency greater than two were considered for feature selection.

- **lexicon-based unigrams**: unigrams from the sentiment lexicon are extracted for feature selection.
- **context n-grams:** unigrams (single words) and bigrams are extracted from the context window. We extract these n-grams for several combinations: (i) replacement of the aspect term with the word *aspect*; (ii) replacement of sentiment words with the polarity word *pos* or *neg*; (iii) replacement of sentiment words with a part of speech tag.
- **aspect-based bigrams:** bigrams generated as a combination of the aspect term itself and a word within the context window. We extract these bigrams for several combinations that described above.
- **lexicon-based features:** the features are calculated as follows: the maximal sentiment score; the minimum sentiment score; the sum of the words' sentiment scores; the sum of positive words' scores; the sum of negative words' scores. Sentiment words with negations shift the sentiment score towards the opposite polarity.

Due to limited size of the context window and difficulty in classifying the aspect with both negative and positive sentiment towards its term, we create hand-crafted rule for such cases: if the sentence *(s)* contains the aspect term, a conjunction word *но, а* (*but*) and the classifier predicts the neutral label for the aspect, we mark the aspect by the both label.

3.3. Automatic categorization of explicit aspects into aspect categories

The goal of task D is to classify each aspect to one of predefined categories. In restaurant reviews there are the following aspect categories: *food, service, interior, price, general*. For automobiles aspect categories are: *drivability, reliability, safety, appearance, comfort, costs, general.*

We describe the task of automatic categorization of explicit aspects in the following examples. Some aspects such as food products (e.g., *бифштекс* (*beefsteak*), *утка по-пекински* (*Peking duck*)) or car components (e.g., *гидроусилитель* (*power steering*), *двигатель* (*engine*)) are classified by a human annotator's explicit knowledge. The categories of food products and car components are *food* and *drivability,* respectively. The category label of some explicit aspects depends on a context of a user review. In the examples "*машина* свои деньги отработала полностью" ("the car is worth its price"), "пробовал отпускать руль *машина* едет ровно" ("have experimented with the driving wheel and the car running smoothly"), "*машина* предназначена для фанатов" ("the *car* is intended for fans") and "довольно красивая *машина*" ("quite beautiful *car*") the categories of the aspect term *машина* (*car*) are *costs*, *drivability*, *whole*, *appearance,* respectively.

We addressed the task as a text classification problem and trained the SVM classifier with the sequential minimal optimization (SMO). For each aspect term w_i we extracted the aspect term itself and the features from the context window $(w_{i-2}, \ldots, w_{i+2})$. Category-specific lexicons are based on a score for each term w in the training test:

$$score\,(w) = PMI\,(w, cat) - PMI\,(w, oth)$$

where *PMI* is pointwise mutual information, *cat* denotes all aspects' contexts in the particular category, *oth* denotes aspects' contexts in other categories.

The SVM classifier is based on bag-of-words model and other features described below:

- word n-grams: the aspect term and unigrams from the context of the aspect term are extracted for feature selection.
- category-specific features: the following features are calculated separately for each category: the maximal score in the context; the minimum score in the context; the sum of the words' scores in the context; the average of the words' scores in the context;

4. Experimental Results

For experimental purposes we used the training set of 200 annotated reviews and the testing set of 200 reviews for each domain provided by the organizers of the SentiRuEval task.

4.1. Performance results

The official results obtained by our approaches on the testing set are presented in Tables 1, 2a, 2b and 3. The tables show the official baseline results and the results of other participants according to macro-average F-measure as the main quality measure in the task (Loukachevitch et al., 2015).

For task A exact matching and partial matching were used to calculate F1-measure. Table 1a and 1b show that our method based on the CRF model did not have any significant improvement over a baseline.

For task C macro-averaged F-measure is calculated as the average value between F-measure of the positive class, negative class and F-measure of the both class. Tables 2a show that according to macro-averaged F1-measure, our classifier does not pay off when compared with the approach with run_id 4_1, that is based on a Gradient Boosting Classifier model. Our approach has 0.13% and 0.06% improvements in macro-averaged F1-measure over the approach with run_id 3_1, ranked second in restaurants and banks domain, respectively. Our runs could not be evaluated due to technical problems with the submission.

Table 3 shows the official baseline results and the results of the method, ranked second according to macro-averaged F-measure in task D. This method ranked first among 4 teams in both subject domains. The best approach has 0.06% and 0.09% improvements in macro F1-measure over the baseline in restaurants and cars domains, respectively.

Table 1a. Performance metrics in extraction of explicit aspects in restaurants domain (task A)

	Exact matching			Partial matching		
	Macro P	**Macro R**	**Macro F**	**Macro P**	**Macro R**	**Macro F**
Our method	0.3515	0.5331	0.5331	0.6507	0.4399	0.5109
An approach, ranked first	0.5506	0.6901	0.6070	0.6886	0.7916	0.7284
Official baseline	0.5570	0.6903	0.6084	0.6580	0.6960	0.6651

Table 1b. Performance metrics in extraction of explicit aspects in cars domain (task A)

	Exact matching			Partial matching		
	Macro P	**Macro R**	**Macro F**	**Macro P**	**Macro R**	**Macro F**
Our method	0.6411	0.5363	0.5749	0.7264	0.6117	0.6498
An approach, ranked first	0.6619	0.6560	0.6513	0.7917	0.7272	0.7482
Official baseline	0.5747	0.6287	0.5941	0.7449	0.6720	0.6966

Table 2a. Performance metrics in the classification task in restaurants domain (task C)

Run_id	**Micro P**	**Micro R**	**Micro F**	**Macro P**	**Macro R**	**Macro F**
Official baseline	0.7104	0.7104	0.7104	0.3209	0.2506	0.2671
1_1	0.6194	0.6194	0.6194	0.2517	0.2454	0.2379
1_2	0.6194	0.6194	0.6194	0.2517	0.2454	0.2379
3_1	0.6696	0.6696	0.6696	0.3223	0.2430	0.2696
4_1	0.8249	0.8249	0.8249	0.5872	0.5569	0.5545
Our approach	0.7671	0.7671	0.7671	0.4582	0.3729	0.4081

Table 2b. Performance metrics in the classification task in cars domain (task C)

Run_id	**Micro P**	**Micro R**	**Micro F**	**Macro P**	**Macro R**	**Macro F**
Official baseline	0.6192	0.6192	0.6192	0.2949	0.2685	0.2648
1_1	0.6471	0.6471	0.6471	0.3399	0.3194	0.3293
1_2	0.6531	0.6531	0.6531	0.3563	0.3297	0.3422
3_1	0.5589	0.5589	0.5589	0.3016	0.2621	0.2794
4_1	0.7428	0.7428	0.7428	0.5725	0.5667	0.5684
1_3	0.6252	0.6252	0.6252	0.3507	0.3262	0.3345
Our approach	0.7110	0.7111	0.7111	0.4481	0.3761	0.4001

Table 3. Performance metrics in categorization of aspects in both subject domains (task D)

	Restaurants			Cars		
	Macro P	**Macro R**	**Macro F**	**Macro P**	**Macro R**	**Macro F**
Our approach	0.8960	0.8414	0.8653	0.6854	0.6355	0.6521
Second result	0.8627	0.7963	0.8110	0.7146	0.5750	0.6077
Official baseline	0.8742	0.7737	0.7996	0.6672	0.5190	0.5636

4.2. Ablation Experiments

We performed ablation experiments to study the benefits of features, which are used for the CRF model and machine learning methods. Tables 4a, 4b and 5 show ablation experiments for tasks A and C on the testing set, removing one each individual feature category from the full set. Error analysis and Tables 4a and 4b show that the features on the set of two previous and two next tokens decrease our results in task A in restaurants domain. The most effective features for task C are based on aspect-based bigrams that include combinations of the aspect term and other words from the context window.

Table 4a. Results for the ablation experiments in aspect extraction about restaurants (task A)

	Exact matching			Partial matching		
	P	**R**	**F1**	**P**	**R**	**F1**
all features	0.3515	0.5331	0.5331	0.6507	0.4399	0.5109
w/o dictionaries	0.3382	0.4971	0.3961	0.3850	0.6921	0.4821
w/o frequencies	0.6503	0.4322	0.5068	0.7313	0.4755	0.5612
w/o all tokens within $(w_{i-2}, \ldots, w_i)$	0.6105	0.4065	0.4751	0.7118	0.4667	0.5471
w/o all tokens within $(w_i, \ldots, w_{i+2})$	0.6471	0.4375	0.5104	0.7272	0.4865	0.5681
w/o tokens that contained all features within $(w_{i-1}, \ldots, w_{i+1})$	0.7311	0.4801	0.5644	0.6476	0.4416	0.5120

Table 4b. Results for the ablation experiments in aspect extraction about cars (task A)

	Exact matching			Partial matching		
	P	**R**	**F1**	**P**	**R**	**F1**
all features	0.6411	0.5363	0.5749	0.7264	0.6117	0.6498
w/o dictionaries	0.6451	0.5421	0.5798	0.7303	0.6191	0.6556
w/o frequencies	0.6380	0.5364	0.5742	0.7148	0.6121	0.6455
w/o all tokens within $(w_{i-2}, \ldots, w_i)$	0.6281	0.5217	0.5609	0.7341	0.6077	0.6498
w/o all tokens within $(w_i, \ldots, w_{i+2})$	0.6144	0.5328	0.5624	0.7022	0.6197	0.6453
w/o tokens that contained all features within $(w_{i-1}, \ldots, w_{i+1})$	0.6414	0.5356	0.5742	0.7264	0.6091	0.6472

Table 5. Results for the ablation experiments in sentiment classification towards aspects (task C)

	Restaurants			Cars		
	macro P	**macro R**	**macro F**	**macro P**	**macro R**	**macro F**
All features	0.4582	0.3729	0.4081	0.4481	0.3761	0.4001
w/o character n-grams	0.4479	0.3659	0.4000	0.4480	0.3750	0.3994
w/o lexicon-based unigrams	0.4259	0.3651	0.3921	0.4213	0.3669	0.3869
w/o aspect-based bigrams	0.4261	0.3396	0.3728	0.4380	0.3746	0.3951
w/o context n-grams	0.4355	0.3586	0.3906	0.4370	0.3717	0.3941
w/o lexicon-based scores	0.4629	0.3681	0.4050	0.4374	0.3747	0.3959

Table 6. Results for feature ablation experiments in categorization of aspects (task D)

Combinations of features	Restaurants			Cars		
	P	**R**	**F**	**P**	**R**	**F**
word n-grams	0.7650	0.7193	0.7388	0.6554	0.6060	0.6219
word n-grams + single cumulative score	0.8185	0.7705	0.7914	0.6800	0.6296	0.6461
word n-grams + domain-specific scores	0.8960	0.8414	0.8653	0.6854	0.6355	0.6521

The experiments for task D are presented in Table 6. Through these feature ablation experiments we show that most important features are the domain-specific features, that are based on pointwise mutual information for the category and include four different calculations of scores in the context of the aspect term.

5. Conclusion

In this paper we described supervised methods for sentiment analysis of user reviews about restaurants and cars. In extraction of explicit aspects (task A) we proposed the method based on syntactic and statistical features incorporated in the Conditional Random Fields model. The method did not show any significant improvement over the official baseline. In extraction of sentiments towards explicit aspects (task C) our method was based on the Maximum Entropy model on a set of lexicon-based features and two types of term frequency features: context n-grams and aspect-based bigrams. We demonstrated that by using these features, classification performance increases from baseline macro-averaged F-measures of 0.267 to 0.408 for restaurants and of 0.265 to 0.4 for cars. In categorization of explicit aspects into aspect categories (task D) we proposed the SVM classifier, based on unigram features and pointwise mutual information to calculate category-specific score. We achieved 65.2% of macro-averaged F-measure for cars and 86.5% for reviews about restaurants in task D. This method ranked first among 4 teams in both subject domains. For future work we plan to provide error analysis of the described methods.

Acknowledgments

This work was funded by the subsidy of the Russian Government to support the Program of competitive growth of Kazan Federal University and supported by Russian Foundation for Basic Research (RFBR Project 13-07-00773).

References

1. *Blinov P., Klekovkina M., Kotelnikov E., Pestov O.* (2013), Research of lexical approach and learning methods for sentiment analysis, Computational Linguistics and Intellectual Technologies, Vol. 2(12), pp. 48–58.
2. *Chernyshevich M.* (2014), IHS R&D Belarus: Cross-domain Extraction of Product Features using Conditional Random Fields, SemEval 2014, pp. 309–313.
3. *Chetviorkin I., Loukachevitch N.* (2013), Evaluating Sentiment Analysis Systems in Russian, ACL 2013, p. 14.
4. *Choi Y., Cardie C.* (2010), Hierarchical sequential learning for extracting opinions and their attributes, Proceedings of the ACL 2010 conference short papers, pp. 269–274.
5. *Jakob N., Gurevych I.* (2010), Extracting opinion targets in a single-and cross-domain setting with conditional random fields, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 1035–1045.

6. *Frolov A. V., Polyakov P. Yu., Pleshko V. V.* (2013), Using semantic filters in application to book reviews sentiment analysis, available at: www.dialog-21.ru/digests/dialog2013/materials/pdf/FrolovAV.pdf
7. *Kiritchenko S., Zhu X., Cherry C., Mohammad S. M.* (2014), NRC-Canada-2014: Detecting aspects and sentiment in customer reviews, SemEval 2014, pp. 437–442.
8. *Lafferty J., McCallum A., Pereira F. C.* (2001), Conditional random fields: Probabilistic models for segmenting and labeling sequence data, Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001), pp. 282–289.
9. *Liu B.* (2012), Sentiment analysis and opinion mining, Synthesis Lectures on Human Language Technologies, vol. 5(1), pp. 1–167.
10. *Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E.* (2015), SentiRuEval: testing object-oriented sentiment analysis systems in Russian, Proceedings of International Conference Dialog-2015, pp. 3–9.
11. *Lu B., Ott M., Cardie C., Tsou B. K.* (2011), Multi-aspect sentiment analysis with topic models, Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference, pp. 81–88.
12. *Moghaddam S., Ester M.* (2012), On the design of LDA models for aspect-based opinion mining, Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 803–812.
13. *Pang B., Lee L., Vaithyanathan S.* (2002), Thumbs up?: sentiment classification using machine learning techniques, Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol. 10, pp. 79–86.
14. *Pang B., Lee L.* (2008), Opinion mining and sentiment analysis, Foundations and trends in information retrieval, vol. 2(1–2), pp. 1–135.
15. *Pontiki M., Papageorgiou H., Galanis D., Androutsopoulos I., Pavlopoulos J., Manandhar S.* (2014), Semeval-2014 task 4: Aspect based sentiment analysis, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 27–35.
16. *Popescu A. M., Etzioni O.* (2007), Extracting product features and opinions from reviews, Natural language processing and text mining, pp. 9–28.
17. *Poria S., Cambria E., Ku L. W., Gui C., Gelbukh A.* (2014), A rule-based approach to aspect extraction from product reviews, SocialNLP 2014, pp. 28–37.
18. *Turney P. D.* (2002), Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424.
19. *Zhao Y., Qin B., Liu T.* (2014), Clustering Product Aspects Using Two Effective Aspect Relations for Opinion Mining, Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, pp. 120–130.