SEMANTIC SIMILARITY FOR ASPECT-BASED SENTIMENT ANALYSIS

Blinov P. D. (blinoff.pavel@gmail.com)

Kotelnikov E. V. (kotelnikov.ev@gmail.com)

Vyatka State Humanities University, Kirov, Russian Federation

The paper investigates the problem of automatic aspect-based sentiment analysis. Such version is harder to do than general sentiment analysis, but it significantly pushes forward the limits of unstructured text analysis methods. In the beginning previous approaches and works are reviewed. That part also gives data description for train and test collections.

In the second part of the article the methods for main subtasks of aspect-based sentiment analysis are described. The method for explicit aspect term extraction relies on the vector space of distributed representations of words. The term polarity detection method is based on use of pointwise mutual information and semantic similarity measure. Results from SentiRuEval workshop for automobiles and restaurants domains are given. Proposed methods achieved good results in several key subtasks. In aspect term polarity detection task and sentiment analysis of whole review on aspect categories methods showed the best result for both domains. In the aspect term categorization task our method was placed at the second position. And for explicit aspect term extraction the first result obtained for the restaurant domain according to partial match evaluation criteria.

Key words: SentiRuEval, aspect-based sentiment analysis, machine learning, distributed representations of words, semantic similarity

1. Introduction

In the last few years sentiment analysis became an important task in the field of natural language processing. The task is interesting for researchers because of its intricate properties. Business community is attracted by the task because it opens potentially vast opportunity to analyze unstructured text and keep track of target audience attitude to a product or brand.

Formulation of sentiment analysis problem is evolving rapidly with respect to granularity: from whole text and sentences to phrase level (Feldman, 2013). The last level of analysis is the most detailed version that is capable to disentangle complex opinions in reviews. Opinions and sentiments are analyzed with respect to specific aspects of reviewed object, for example, aspects *food*, *service* and *price* of an object *restaurant*. Such detailed task is called aspect-based sentiment analysis (Liu, 2012). For simplification the task can often be split into following subtasks:

1) aspect term extraction;
2) aspect term polarity detection;
3) aspect category polarity detection.

In this article we present new methods for addressing these subtasks. The methods are mainly based on distributed representations of words and notion of semantic similarity.

The rest of the paper is structured as follows. Section 2 gives the overview of previous works. The characteristics of train and test text data are given in Section 3. Section 4 contains method descriptions and results for proposed subtasks. The final conclusions are given in Sections 5.

2. Related work

There are many research papers for sentiment analysis problem, fewer about aspect-based version of it. As for the language, plenty of works were carried out for English (Liu, 2012) and less fewer for Russian (Blinov, Kotelnikov, 2014). Recently there was a burst of research interest to the task because of SemEval-2014 Workshop (Pontiki et al., 2014), where one of the key topics was an aspect-based sentiment analysis. Here we give a brief analysis of applied approaches and methods regarding two main subtasks: aspect term extraction and aspect term polarity detection.

To address aspect term extraction problem participants resorted to two main approaches (Liu, 2012):

1) frequency-based approach;
2) machine learning approach.

Perhaps the first and most famous work from the first approach is (Hu, Liu, 2004). In a nutshell, the general idea of the approach is to find nouns and noun phrases and by some technique filter them out to left only relevant aspect terms. Statistical criteria are often used as such filters (Schouten et al., 2014). Rule-based and dependency parsing methods constitute another group of such filtering techniques (Pekar et al., 2014; Zhang et al., 2014).

The given task can be easily formulated in terms of information extraction tasks, so another popular approach is based on sequence labeling methods. SemEval-2014 Workshop's participants widely used well known Conditional Random Fields (CRF) method (Kiritchenko et al., 2014; Chernyshevich, 2014). In fact the best results in aspect term extraction task were attained by this method with common named entity recognition features and features based on various name lists and word clusters (Toh, Wang, 2014). Each word can be described in terms of features, so traditional machine learning methods for classifications are also used to address the task (Brun et al., 2014; Gupta, Ekbal, 2014).

For the aspect term polarity detection task the most of the solutions exploit external sentiment resources. (Bornebusch et al., 2014) used Stanford sentiment trees to detect terms' sentiments. The best results (Wagner et al., 2014) were obtained by SVM classifier and features based on combination of four rich sentiment lexicons.

3. Text data

This year sentiment analysis evaluation was organized in Russian and was called *SentiRuEval* (Loukachevitch et al., 2015). The evaluation included two types of tasks: aspect-oriented sentiment analysis of users' reviews and object-oriented sentiment analysis of Russian tweets. The article deals with the first of these tasks.

The organizers provide the train data for two domains: restaurant and automobile reviews. Each reviewed object was broken down into several aspects (also referred as aspect categories). For a restaurant there were four aspects: *Food*, *Interior*, *Service* and *Price*. And an automobile was analyzed by six aspects: *Comfort*, *Appearance*, *Reliability*, *Safety*, *Driveability* and *Costs*. In addition each aspect list was supplemented with aspect *Whole* to represent object itself.

The train reviews were manually annotated with mentioned aspect terms according to aspects listed above. There are different types of aspect terms (Loukachevitch et al., 2015), but in our study we focus only on explicit aspect terms. Assessors also were asked to specify sentiment toward terms using four-point scale: *positive*, *negative*, *neutral* and *both*. Thus each aspect term incorporates information about aspect category and polarity. All marked texts were stored in xml format documents. Detailed quantitative characteristics of explicit terms for the train and test data for both domains are given in Table 1. By analyzing the table one can see the usual peculiarity of sentiment analysis tasks: significant skewness toward positive class.

Table 1. Explicit aspect and sentiment distribution

		Number of terms			
		Restaurant		Automobile	
		Absolute	%	Absolute	%
Train	Positive	1,679	69.5	1,513	48.0
	Negative	380	13.5	858	27.2
	Neutral	714	25.3	690	21.9
	Both	49	1.7	91	2.9
	Total	**2,822**	**100**	**3,152**	**100**
Test	Positive	2,478	70.7	1,706	54.9
	Negative	509	14.5	844	27.1
	Neutral	440	12.5	454	14.6
	Both	79	2.3	105	3.4
	Total	**3,506**	**100**	**3,109**	**100**

Besides marked data the organizers provide unlabeled text data for each domain: 19,034 reviews for restaurant domain and 8,271 reviews for automobile domain. All text was preprocessed by morphology analyzer Mystem[1].

[1] Morphological analyzer for Russian mystem. URL: http://tech.yandex.ru/mystem.

4. Aspect-based sentiment analysis

Distributed representations of words show ability to cluster semantically similar words (Mikolov et al., 2013). This property can be useful for solving main subtasks of aspect-based sentiment analysis. In our methods for obtaining distributed representations we use skip-gram model (Mikolov et al., 2013) in the implementation of Gensim library[2]. That model gives us whole vector space in which word vectors are embedded. To produce 300-dimensional word vectors the context window of five words was used. The only texts provided by the organizers were used as the input data for the skip-gram model. But more unlabeled texts lead to better word representations which certainly facilitate performance of proposed method.

4.1. Explicit aspect term extraction method

In the workshop SentiRuEval there were two tasks related to aspect term extraction. Our method deals only with explicit aspect term extraction—task A.

Since the train collection is labeled with aspect terms the initial sets of seed words can be constructed for each aspect. All single-word terms (nouns and verbs) were selected.

For an unknown word-vector $\vec{a} = (a_1, ..., a_n)$ similarity to particular aspect *asp* specified by seed word-vectors $\vec{b}_i = (b_1, ..., b_n)$ can be calculated via cosine similarity in the vector space (Manning et al., 2008):

$$sim(\vec{a}, asp) = \sum_{i=1}^{k} \frac{\vec{a} \cdot \vec{b}_i}{\|\vec{a}\| \cdot \|\vec{b}_i\|}, \vec{b}_i \in B_{asp}, \quad (1)$$

where B_{asp} is the set of seed words for aspect *asp* and $|B_{asp}| = k$ is the number of seed words.

If that similarity exceeds a threshold then the word is marked as aspect term. Thresholds for each aspect category were defined by 10-fold cross validation.

However such procedure can find only single word aspect terms. But multi-word terms form a significant part of all aspect terms, especially for particular aspects, for example *Food*. By our estimate on the restaurant train collection about a fifth part of all terms are multi-word terms. And even greater proportion is preserved for automobile train collection. Probably the multi-word terms can be proceeded naturally by distributed representations but it requires additional preprocessing step to reveal such phrases (with high accuracy) before streaming them to skip-gram model. Very likely it also will require more amount of unlabeled texts. Such improvements lay beyond our current experiments and we resorted to more simple technique to tackle multi-word term issue.

A set of rules was applied to join single terms into a complex one. Sequentially marked words were merged and the ones conjoined by prepositions also merged in a single aspect term. For example, *котлетки из лосося* (*meatballs from salmon*) or *роллы на гриле* (*rolls on grill*). Another set of rules handles aspect terms of category

[2] Topic modeling library gensim. URL: http://radimrehurek.com/gensim.

Whole. Because reviewers often refer to a restaurant by name which is contained in review's metadata, the full match with that string in the text of review is marked as an aspect term.

The baseline method for that task memorizes aspect terms from the train reviews and look for the same terms in the test reviews. Table 2 shows baseline results, best results and results of our method with respect to exact and partial matching evaluation criteria (Loukachevitch et al., 2015). We apply following notion (here and for other tasks' results): **bold** for the best result and *italic* for our method's result. F_1-measure was a primary measure for the tasks.

Table 2. Results of explicit aspect term extraction task (task A)

		Exact matching (macro)			Partial matching (macro)		
	run_id	**Precision**	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**
Restaurant	baseline	55.70	69.03	60.84	65.80	69.60	66.51
	2_1	72.37	57.38	**63.19**	80.78	61.65	68.91
	4_1	*55.06*	*69.01*	*60.70*	*68.86*	*79.16*	***72.84***
Automobile	baseline	57.47	62.87	59.41	74.49	67.24	69.66
	2_1	76.00	62.18	**67.61**	85.61	65.51	73.04
	3_1	66.19	65.60	65.13	79.17	72.72	**74.82**
	4_1	*55.77*	*63.55*	*58.63*	*74.17*	*68.87*	*70.16*

Our method shows the best result in term extraction for the restaurant domain according to partial matching, but for exact matching the result is worse. For both variants of evaluation the method shows higher recall values then precision. This means that the method found many terms similar to aspect terms which in fact are not.

For the automobile domain our results are near baseline. This is probably due to small amount of unlabeled additional data. To obtain good vector space one need as much text data as possible. But for the automobile domain additional collection was four times smaller than for restaurant domain. Different aspect term compositionality is another possible explanation of such poor results. For example, in this domain there are mixed terms containing numbers and words such as *Двигатель 2.5 литра* (*The engine of 2.5 liters*), *ваз 2114* (*VAZ 2114*), etc. But our algorithm doesn't take this into account.

In general the baseline benchmarks for each domain are pretty high and even the best participants' results exceed them marginally (all gains are less than 10%). One of the possible reasons of relatively simple applied baseline algorithms' high results (Loukachevitch et al., 2015) is high-quality train collection, which covers a lot of aspect term lexicon which is rather limited.

4.2. Aspect term polarity detection method

The task C was to determine sentiments toward predefined aspect terms. The train examples were classified into four-point scale: *positive*, *negative*, *neutral* and

both. But the evaluation was performed only on three-point scale: *positive*, *negative* and *both*. So we prepared solution to that scale only.

In most cases sentiment of an aspect term is defined by its context words. To represent this context from sentiment perspective sentiment lexicon was created for each domain. All verbs and adjectives are the units of such resource. Only one type of negation (as most common) is handled: <*not*> + <*adjective* or *verb*>. To associate sentiment with each unit we use two types of weighting: based on semantic similarity and based on pointwise mutual information (PMI). The reason of using of two kinds of scores is that two different sources of sentiment information allow better estimate actual sentiment.

For semantic similarity weighting we apply the same procedure for sum similarity calculation (1) for each sentiment unit (represented by real-valued vector $\vec{a}$). The only difference in the task A is the set of words. Now these words are etalon for *positive* or *negative* sentiment. From two sum similarities (to positive and negative classes) the largest by absolute value with appropriate sign became sentiment score for a unit. Examples of such estimation are: *приятный* (*+7.1*) (*nice*); *прекрасный* (*+6.5*) (*lovely*); *стильный* (*+5.9*) (*stylish*); *неуместный* (*–4.8*) (*inappropriate*); *пошлый* (*–4.4*) (*vulgar*); *жуткий* (*–4.2*) (*spooky*); etc.

PMI scores for the same dictionary units were calculated based on collection of reviews with general scores. Collections for PMI calculation previously were filtered out to save most positive (restaurant domain: $score \geq 7 \rightarrow +1$ and automobile domain: $score \geq 4 \rightarrow +1$) and most negative (restaurant and automobile domain: $score \leq 3 \rightarrow -1$) reviews. The score for a unit *w* is defined as (Islam, Inkpen, 2006):

$$score\,(w) = PMI\,(w, pos) - PMI\,(w, neg). \tag{2}$$

Mutual information between unit *w* and, for example, *positive* sentiment class *PMI*(*w*, *pos*) (and for the *negative* class *PMI* was calculated in a similar way) is defined as (Islam, Inkpen, 2006):

$$PMI(w, pos) = \log_2 \frac{count(w, pos) \cdot N}{count(w) \cdot count(pos)}, \tag{3}$$

where *count*(*w*, *pos*)—count of unit *w* in positive reviews, *N* is total number of tokens in corpus, *count*(*w*)—count of unit *w* in all reviews, *count*(*pos*) is a total amount of terms in positive reviews.

There was no notion of a threshold for PMI scores and each unit of the lexicon assigned to some score. Examples are: *классный* (*+3.1*) (*cool*); *добротный* (*+2.6*) (*mighty*); *выдающийся* (*+1.6*) (*outstanding*); *тошнить* (*–2.7*) (*to puke*); *не дружелюбный* (*–3.8*) (*not friendly*); *хамский* (*–4.5*) (*boorish*); etc.

With the help of weighted dictionary units each aspect term is presented in near (three nearest words) and far (six words) contexts as feature vector. In such form train data is used as an input to gradient boosting classifier (Friedman, 2001).

The sentiment class *both* is presented by very small set of samples (see Table 1). And it is a problem for the classifier to learn such minor-represented class. By observing

"both" aspect terms simple regularity was revealed: for the great number of "both" terms there are "but" conjunction in the sentence. And rule "to assign *both* sentiment to a term if there is a 'but' conjunction in the sentence" was applied to resolve the issue.

The baseline method for this task was a very simple one: to assign a major sentiment for a term based on stats from the train collection (mostly *positive*). Results of baseline, our method and second place participants are given in Table 3.

Table 3. Results of aspect term polarity detection (task C)

		Micro-averaging			Macro-averaging		
	run_id	**Precision**	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**
Restaurant	baseline	71.04	71.04	71.04	32.09	25.06	26.71
	4_1	*82.49*	*82.49*	***82.49***	*58.72*	*55.69*	***55.45***
	3_1	66.96	66.96	66.96	32.23	24.30	26.96
Automobile	baseline	61.92	61.92	61.92	29.49	26.85	26.48
	4_1	*74.28*	*74.28*	***74.28***	*57.25*	*56.67*	***56.84***
	1_2	65.31	65.31	65.31	35.63	32.97	34.22

4.3. Aspect term classification method

Goal of task D was to categorize predefined set of terms into aspect categories. Some methods can extract terms and at the same time define its aspect category. In this paper, term categorization task taken out into separate stage.

To solve task D we again resorted to similarity between words. In such meaning this task is opposite to task A. The solution is to compute similarity (1) to seed sets of words and choose aspect category that maximize the similarity. For multi-word term single vector representation can be found by averaging out words of the term (since each word is represented by its vector).

The baseline for that task is identical to baseline in task C: assign most frequent category for a term. With described method our team occupied the second place in this task (Table 4).

Table 4. Results of aspect term categorization (task D)

	run_id	**P**	**R**	**F_1**
Restaurant	baseline	87.42	77.37	79.96
	8_1	89.60	84.14	**86.53**
	4_1	*86.27*	*79.63*	*81.10*
Automobile	baseline	66.72	51.89	56.36
	8_1	68.54	63.55	**65.21**
	4_1	*71.46*	*57.50*	*60.77*

It is interesting that for automobile domain the metrics are much lower than for restaurant domain. Probably it is because the lexicon of automobile review is more intertwined and context dependent. For some terms it is hard to decide to which category it belongs to. For example, *руль* (*steering wheel*) belongs to aspect *Drivability* and *Comfort*; *обзор* (*visibility*) occurs in aspect *Comfort* and *Safety*; etc. And in general number of aspect categories are greater for automobile domain: seven whereas there are only five for restaurants.

4.4. Sentiment analysis of whole review on aspect categories

The task E was to define sentiments about aspect categories. Such sentiments related to the whole review rather than individual aspect terms.

As the solution of polarity detection task is performed in three-point scale the task E is automatically addressed in this scale also. By this point each review has a list of aspect terms with defined sentiment and categories. Following mapping was used to cast sentiments to numbers: +1—*positive*, –1—*negative*, 0—*both*. For each category summation over terms sentiment gives total sentiment of aspect category. If there are no terms for some aspect category it is left with "*absence*" value. If at least one category's term has *both* sentiment the entire category is assign to it.

There were not many participants in this task. Again the baseline is just an assignment of the most frequent sentiment for a particular aspect category. Results are shown in Table 5.

Table 5. Results of sentiment analysis of the whole review on aspect categories (task E)

	run_id	F_1
Restaurant	baseline	27.20
	4_1	***45.82***
	10_1	37.28
Automobile	baseline	23.68
	4_1	***43.90***

The obtained results are the lowest for this task (comparing with other tasks) because of its complexity. The method can be misled by incorrectly extracted aspect term or wrongly detected term's sentiment.

5. Conclusions

We described full stack of methods for main subtasks of aspect-based sentiment analysis. To achieve the best possible results the proposed methods actively use notion of semantic similarity between words, statistical measures and hand-crafted rules.

By partial matching evaluation criteria method for aspect term extraction showed the best results for the restaurant domain among fourteen methods. By exact matching the result is worse but still in the top among participants at the fourth position. The method of polarity term detection showed the best results in both domains among seven runs. For the task of aspect terms' categorization our method was placed at the second position. Also the first place for both domains earned the method for sentiment analysis by aspect categories. From the good results we can conclude that the proposed methods can be used for practical applications to perform detailed sentiment analysis of users' reviews.

Another conclusion that can be drawn is about complexity of sentiment analysis for Russian and English. Actually for one task—exact aspect term extraction—we can compare the results with analogous task from SemEval-2014 (Pontiki et al., 2014). There the best result by F_1 measure for the restaurant domain was 84% while in our competition the best result was only 63%. This leads us to the conclusion that aspect term extraction for Russian is more difficult than for English. The possible sources of the problem are free word order and more complex morphology. To overcome that machine learning methods with more extensive usage of linguistically specific knowledge can probably show the better results for object-oriented sentiment analysis.

Acknowledgements

We want to thank the organizers and assessors for their efforts in running such evaluation workshop. This work is supported by the Russian Ministry of Education and Science, research project No. 586.

References

1. *Blinov P. D., Kotelnikov E. V.* (2014), Using Distributed Representations for Aspect-Based Sentiment Analysis, Proceedings of International Conference Dialog, pp. 739–746.
2. *Bornebusch F., Cancino G., Diepenbeck M., Drechsler R., Djomkam S., Fanseu A., Jalali M., Michael M., Mohsen J., Nitze M., Plump C., Soeken M., Tchambo F., Toni, Ziegler H.* (2014), iTac: Aspect Based Sentiment Analysis using Sentiment Trees and Dictionaries, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, pp. 351–355.
3. *Brun C., Popa D., Roux C.* (2014), XRCE: Hybrid Classification for Aspect-based Sentiment Analysis, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, pp. 838–842.
4. *Chernyshevich M.* (2014), IHS R&D Belarus: Cross-domain Extraction of Product Features using Conditional Random Fields, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, pp. 309–313.
5. *Feldman R.* (2013), Techniques and Applications for Sentiment Analysis, Communications of the ACM, Vol. 56, pp. 82–89.

6. *Friedman J.* (2001), Greedy Function Approximation: A Gradient Boosting Machine, The Annals of Statistics, Vol. 29, pp. 1189–1232.
7. *Gupta D., Ekbal A.* (2014), IITP: Supervised Machine Learning for Aspect based Sentiment Analysis, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, pp. 319–323.
8. *Hu M., Liu B.* (2004), Mining and Summarizing Customer Reviews, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177.
9. *Islam A., Inkpen D.* (2006), Second Order Co-occurrence PMI for Determining the Semantic Similarity of Words, Proceedings of the International Conference on Language Resources and Evaluation, pp. 1033–1038.
10. *Kiritchenko S., Zhu X., Cherry C., Mohammad S.* (2014), NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, pp. 437–442.
11. *Liu B.* (2012), Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies, Vol. 5(1).
12. *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.
13. *Manning C., Raghavan P., Schütze H.* (2008), Introduction to Information Retrieval, Cambridge University Press., New York.
14. *Mikolov T., Sutskever I., Chen K., Corrado G., Dean J.* (2013), Distributed Representations of Words and Phrases and their Compositionality, Proceedings of NIPS, pp. 3111–3119.
15. *Pekar V., Afzal N., Bohnet B.* (2014), UBham: Lexical Resources and Dependency Parsing for Aspect-Based Sentiment Analysis, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, pp. 683–687.
16. *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.
17. *Schouten K., Frasincar F., Jong F.* (2014), COMMIT-P1WP3: A Co-occurrence Based Approach to Aspect-Level Sentiment Analysis, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, pp. 203–207.
18. *Toh Z., Wang W.* (2014), DLIREC: Aspect Term Extraction and Term Polarity Classification System, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, pp. 235–240.
19. *Wagner J., Arora P., Cortes S., Barman U., Bogdanova D., Foster J., Tounsi L.* (2014), DCU: Aspect-based Polarity Classification for SemEval Task 4, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, pp. 223–229.
20. *Zhang F., Zhang Z., Lan M.* (2014), ECNU: A Combination Method and Multiple Features for Aspect Extraction and Sentiment Polarity Classification, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, pp. 252–258.