Aspect Extraction from Reviews Using Conditional Random Fields

Yuliya Rubtsova[1(✉)] and Sergey Koshelnikov[2]

[1] A.P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, 630090 Novosibirsk, Russia
yu.rubtsova@gmail.com
[2] Independent Developer, Novokuznetsk, Russia
koshelnikovsa@gmail.com

Abstract. This paper describes an information extraction and content analysis system. The proposed system is based on a conditional random field algorithm and intended to extract aspect terms mentioned in the text. We use a set of morphological features for machine learning. The system is used for automatic extraction of explicit aspects and also to automatic extraction of all aspects (explicit, implicit and sentiment facts), and tested on two domains: restaurants and automobiles. We show that our system can produce quite a high level of precision which means that the system is capable of recognizing aspect terms rather accurately. The system demonstrates that even a small set of features for conditional random field algorithm can perform competitively and shows good results.

Keywords: Aspect detection · Aspect extraction · CRF · Information retrieval · Information extraction · Content analysis

1 Introduction

With the popularity of blogs, social networks and users' reviews sites 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 the businessmens who are more interested in the practical application of reputation marketing in general and sentiment analysis in particular. Automatic sentiment analysis is mostly used at the following levels:

- Document level [1–3],
- Sentence or phrase level [4],
- Aspect level [5–7].

Generally people express their opinions not on the product or service as a whole but on some part, feature or characteristic thereof that is the aspect that shall be extracted from the text and subjected to sentiment analysis. The aspect

© Springer International Publishing Switzerland 2015
P. Klinov and D. Mouromtsev (Eds.): KESW 2015, CCIS 518, pp. 158–167, 2015.
DOI: 10.1007/978-3-319-24543-0_12

in our terms represents the opinion target. Simply aspect means a feature of a product. The aspect-level sentiment analysis can give us much more useful information on the authors opinion on various features of the product or service under analysis than sentiment analysis of the whole text.

System described in this paper took a part into Dialogue Evaluation section – SentiRuEval (Dialogue conference 2015): evaluation of sentiment analysis systems for the Russian language [8]. The participants of the evaluation were required 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.

Statistical information about train and test collections such as a number of reviews or amount of explicit terms in different domains can be find in Table 1. Almost equal size of train and test collections can be explained by the fact that SentiRuEval organizers first provided training collections for develope and train classifiers, later a test collections for participation in evaluation was granted. The collections consists of users reviews of restaurants or automobiles depending on domain.

Table 1. Collections statistics

	Restaurants		Automobiles	
	Train	Test	Train	Test
Number of reviews	201	203	217	201
Number of explicit terms	2822	3506	3152	3109
Number of implicit terms	636	657	638	576
Number of fact terms	523	656	668	685

This paper describes the system that was used to perform Tasks A and B during SentiRuEval evaluation. 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 the descriotion of system is given. Section 4 demonstrates the performance of system as compared to the results of systems of other SentiRuEval participants. We discuss errors made by presented system in section 5. Section 6 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 another 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 [9]. The precision level of that algorithm later has been improved by 22% with the recall decreasion 3% only [10]. 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 [11]. 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 [9]. 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 of a sequence of words. The most wide-spread methods of supervised machine learning are hidden Markov modeling (HMM) [12] and conditional random fields (CRF) [13–15]. 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 has its probability distribution [16,17]. Numerous works on aspect extraction with the use of topic modeling approach are based on the methods of probabilistic latent semantic analysis (pLSA) model [18] and latent Dirichlet allocation (LDA) model [19]. To perform complex tasks such as simultaneous aspect extraction and sentiment analysis or simultaneous aspect extraction and categorization, one can employ combination of different approaches such as maximum entropy and latent Dirichle allocation [20] or semi supervised model with the topic modeling approach when user provides some seed words for a few aspect categories [21].

3 System Description

We participated into two evaluation tasks:

- 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 automobiles 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 authors 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

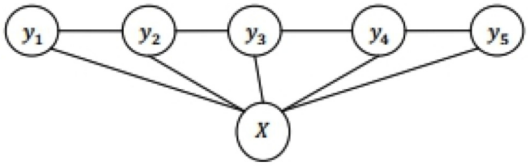

Fig. 1. An example of conditional random field

related to aspect-based Sentiment Analysis, two best results have been obtained by systems that were based on CRF [22].

Conditional Random field is a type of discriminative undirected probabilistic graphical model. It is used to encode known relationships between observations and construct consistent interpretations. Let $X = (x_1, ..., x_n)$ be the sequence of observed data (speaking in terms of our tasks these are tokens of a review). Let $Y = (y_1, ..., y_n)$ be the sequence of random variables associated with the vertices of the graph G (labels we want to learn to predict). Therefore in our case a graphics model looks as it shown in Fig. 1.

CRFs 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) [5]. Than the conditional distribution p(Y|X) can be formalized as Formula 1:

$$P(Y|X) = \frac{1}{Z(X)} exp(\sum_{c \in C} \lambda_c f_c(y_c, X)) \quad , \tag{1}$$

where, C is a set of all graphs cliques, f_c set of all features, λ_c is its corresponding weight. $Z(x)$ is a normalization function (Formula 2):

$$Z(X) = \sum_{y} exp(\sum_{c \in C} \lambda c f_c(y_c, X)). \tag{2}$$

There are two main advantages of CRFs:

1. their conditional nature, resulting in the relaxation of the independence assumptions of the observed variables,
2. CRFs avoid the label bias problem. As CRF has a single exponential model for the joint probability of the entire sequence of labels given the observation sequence (not only one given state). Hence, even if some data is missing, the observation sequence can still be labeled with less number of features. That is usefull for us as trainig collection is limited.

We utilized the Mallet tool as a software implementation of CRF [23].

3.1 Labeling

Jakob and Gurevych [13] 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. 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 morphological features (e.g. POS and lemma) described in the next section, we used TreeTagger for the Russian language [24]. We also noticed that automobile 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 4, 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. However we show a drawback of lowercase converting in the section error analysis.

3.2 Features

Word. We used the token itself and its neighboring words in a [-1, +1] window 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 morphological 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 utilized a TreeTagger.

3.3 Architecture

We built the system which was tested under two conditions:

- **Condition 1:** CRF with all the above-mentioned labels. We used s-e, c-e and O labels for explicit aspect extraction to perform the Task A and s-e, c-e, s-i, c-i, s-f, c-f, O labels to extract all the aspect terms for the Task B.
- **Condition 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 (but not explicit). Task A was performed using only condition 1 and Task B – using both conditions.

Further in the paper we would use shortness "system 1" for system under condition 1 and "system 2" for system under condition 2.

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 in this case means 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 for every term. Tables 2 to 6 demonstrate the System performance of Task A and Tables 7 to 9 refer to performance of Task B. The results of the System were compared to the baseline and the two best results of SentiRuEval participants.

Table 2. Task A results, Restaurant domain, exact matching

System	Precision	Recall	F-measure
baseline	0.5570	0.6903	0.6084
*No*1	0.7237	0.5738	0.6319
*No*2	0,6358	0.6327	0.6266
Word+POS	0.6610	0.5150	0.5704
+Lemma	0.6674	0.5417	0.5899

Table 3. Task A results, Restaurant domain, partial matching

System	Precision	Recall	F-measure
baseline	0.6580	0.696	0.6651
*No*1	0.8078	0.6165	0.7280
*No*2	0.7458	0.7114	0.7191
Word+POS	0.7380	0.5630	0.6277
+Lemma	0.7485	0.5937	0.6520

Table 4. Task A results, automobile domain, exact matching

System	Precision	Recall	F-measure
baseline	0.5747	0.6287	0.5941
*No*1	0.7600	0.6218	0.6761
*No*2	0.6619	0.6560	0.6513
Word+POS	0.7109	0.5454	0.6075
+Lemma	0.7040	0.5785	0.6256

Table 5. Task A results, automobile domain, partial matching

System	Precision	Recall	F-measure
baseline	0.7449	0.6724	0.6966
*No*1	0.7917	0.7272	0.7482
*No*2	0.8561	0.6551	0.7304
Word+POS	0.7970	0.6047	0.6747
+Lemma	0.7908	0.6485	0.6991

As it can be observed from Table 1-4, the System demonstrated high precision level in both domains (2nd position in Task A for both domains: automobiles 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 useit may be concerned to pre-processing rules to the automobiles collection. In Task B system showed rather high precision level (see Table 5-8). In the domain of restaurants system 1 (condition 1) with word+pos+lemma features ranked 3rd amount all the participants by the partial matching case of F-measure.

Table 6. Task B results, Restaurant domain, exact matching

System	Precision	Recall	F-measure
baseline	0.5466	0.6477	0.5872
*No*1	0.6094	0.6006	0.6001
*No*2	0.7336	0.5132	0.5962
System 1 Word+POS +Lemma	0.6393	0.4563	0.5258
System 2 Word+POS +Lemma	0.6521	0.4585	0.5316
	0.6715	0.4916	0.5615

Table 7. Task B results, Restaurant domain, partial matching

System	Precision	Recall	F-measure
baseline	0.6716	0.5931	0.6193
*No*1	0.7562	0.6108	0.6679
*No*2	0.6687	0.6371	0.6452
System 1 Word+POS +Lemma	0.7104	0.4934	0.5692
System 2 Word+POS +Lemma	0.7099	0.5294	0.5953
	0.7246	0.4579	0.5478
	0.7524	0.4936	0.5851

Table 8. Task B results, automobile domain, exact matching

System	Precision	Recall	F-measure
baseline	0.5979	0.5896	0.5886
*No*1	0.7701	0.5535	0.6366
*No*2	0.6563	0.6164	0.6301
System 1 Word+POS +Lemma	0.6908	0.4763	0.5561
System 2 Word+POS +Lemma	0.6706	0.5187	0.5781
	0.7190	0.4821	0.5683
	0.7012	0.5204	0.5893

Table 9. Task B results, automobile domain, partial matching

System	Precision	Recall	F-measure
baseline	0.7833	0.6060	0.6743
*No*1	0.8143	0.6510	0.7148
*No*2	0.7954	0.6470	0.7042
System 1 Word+POS +Lemma	0.7936	0.532	0.6255
System 2 Word+POS +Lemma	0.7773	0.5848	0.6561
	0.8086	0.5100	0.6130
	0.7824	0.5582	0.6389

5 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 wont be able to observe third type of mistake. From Table 10, we can find that a major bunch of errors is related to not recognized aspect terms. We isolated four types of our systems' errors.

5.1 Technical Errors

Special Symbols. Our system does not perform well when dealing with sequences containing markup characters like ""Caesar" salad" ("Салат "цезарь"").. In such cases the system returns only a part of an aspect: ""Caesar salad" ("Салат "цезарь") without the closing """.

Table 10. Error type distribution for the task A (exact matching).

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

Lower Case. As it was mentioned in Section 3.1 all the capital letters were converted into lowercase. But we did not leave out that some of specific terms were lost. For instance "TO" (technical maintenance in the automobile domain) and "то" (the particle).

5.2 No Recognition

Shortcuts. Our system cannot find shotcuts i.e. rubles $\rightsquigarrow$ rub $\rightsquigarrow$ R. The dictionary of frequently used shortcuts could help to remedy this.

Listings. The system deals with listings quite poorle. It found some listed items but not all of them, i.e "Vegetables, **salads "Caesar"**, salmon" (the item that system has found is shown in bold).

5.3 Partial Recognition

Before Head Word. The system can better deal with nouns and more precisely extract nouns as an aspect term: "Добавляла **вина**" ("pour **wine**"). However we found several non-noun terms which were partly recognized by the system: "Официант **хамил**" ("The waiter **was rude**").

After Head Word. We have also found that not only terms before the head word cause difficulties for the system, but also terms after it i.e. "**местечко** углу" ("**a place** in the corner"). It should be noted that there were relatively fewer mistakes than the mistakes of the previous type.

5.4 Excessive Recognition

Our system does not always precisely deal with named entities, i.e. sometimes it extracts names like "Александр" ("Alexander") which are not an aspect term.

It can also be observed from the Table 10 that adding Lemmas as a CRF feature leads to increasing excessively recognized terms. We compared system under two conditions and found out that the second one can better deal with

collocations. For instance, it extracted "duck soup" ("суп из утки") instead of just "soup" ("суп") extracted by the system under Condition 1. However collocations can be problematic even under Condition 2 because occasionally the system can extract too many 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 can also make further improvements.

6 Conclusion

We presented aspect extraction system built on the basis of conditional random field algorithm. Realization of these system demonstrated that preprocessing and using even a small set of features for CRF shows comparatively good results by the overall F-measure. The performance of our system was comparable to the best results of SentiRuEval participants. Subsequently we are going to add statistical informaion as a CRFs' feature. We are also planning to make a research and find a way to improve the recall results without reduce a precision.

References

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86 (2002)
2. Rubtsova, Y.V.: Development and research domain independent sentiment classifier. SPIIRAS Proceedings **5**(36), 59–77 (2014)
3. Turney, P.D.: 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, pp. 417–424. ACL (2002)
4. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational linguistics **35**(3), 399–433 (2009)
5. Zhang, L., Liu, B.: Aspect and entity extraction for opinion mining. In: Data Mining and Knowledge Discovery for Big Data, pp. 1–40 (2014)
6. Liu, B.: Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies **5**(1), 1–167 (2012)
7. Marrese-Taylor, E., Velásquez, J.D., Bravo-Marquez, F.: A novel deterministic approach for aspect-based opinion mining in tourism products reviews. Expert Systems with Applications **41**(17), 7764–7775 (2014)
8. Loukachevitch, N.V., Blinov, P.D., Kotelnikov, E.V., Rubtsova, Y.V., Ivanov, V.V., Tutubalina, E.: SentiRuEval: testing object-oriented sentiment analysis systems in russian. In: Proceedings of International Conference Dialog (2015)
9. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177 (2004)
10. Popescu, A.M., Etzioni, O.: Extracting product features and opinions from reviews. In: Natural Language Processing and Text Mining, pp. 9–28 (2007)

11. Moghaddam, S., Ester, M.: 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 (2011)
12. Jin, W., Ho, H.H., Srihari, R.K.: 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 (2009)
13. Jakob, N., Gurevych, I.: 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, pp. 1035–1045. ACL (2010)
14. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of International Conference on Machine Learning (ICML-2001) (2001)
15. Sutton, C., McCallum, A.: An introduction to conditional random fields for relational learning. In: Introduction to Statistical Relational Learning. MIT Press (2006)
16. Brody, S., Elhadad, N.: 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 (2010)
17. Titov, I., McDonald, R.: Modeling online reviews with multi-grain topic models. In: Proceedings of the 17th International Conference on World Wide Web, pp. 111–120. ACM (2008)
18. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine learning **42**(1–2), 177–196 (2001)
19. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research **3**, 993–1022 (2003)
20. Zhao, W.X., Jiang, J., Yan, H., Li, X.: Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 56–65. ACL (2010)
21. Mukherjee, A., Liu, B.: Aspect extraction through semi-supervised modeling. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, vol. 1, pp. 339–348 (2012)
22. Pontiki, M., Papageorgiou, H., Galanis, D., Androutsopoulos, I., Pavlopoulos, J., Manandhar, S.: Semeval-2014 task 4: aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval 2014, pp. 27–35 (2014)
23. McCallum, A.K.: MALLET: A Machine Learning for Language Toolkit (2002)
24. Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A., Divjak, D.: Designing and evaluating russian tagsets. In: LREC (2008)