How to Evaluate Opinionated Keyphrase Extraction?

Gábor Berend
University of Szeged
Department of Informatics
Árpád tér 2., Szeged, Hungary
berendg@inf.u-szeged.hu

Veronika Vincze
Hungarian Academy of Sciences
Research Group on Artificial Intelligence
Tisza Lajos krt. 103., Szeged, Hungary
vinczev@inf.u-szeged.hu

Abstract

Evaluation often denotes a key issue in semantics- or subjectivity-related tasks. Here we discuss the difficulties of evaluating opinionated keyphrase extraction. We present our method to reduce the subjectivity of the task and to alleviate the evaluation process and we also compare the results of human and machine-based evaluation.

1 Introduction

Evaluation is a key issue in natural language processing (NLP) tasks. Although for more basic tasks such as tokenization or morphological parsing, the level of ambiguity and subjectivity is essentially lower than for higher-level tasks such as question answering or machine translation, it is still an open question to find a satisfactory solution for the (automatic) evaluation of certain tasks. Here we present the difficulties of finding an appropriate way of evaluating a highly semantics- and subjectivity-related task, namely opinionated keyphrase extraction.

There has been a growing interest in the NLP treatment of subjectivity and sentiment analysis – see e.g. Balahur et al. (2011) – on the one hand and on keyphrase extraction (Kim et al., 2010) on the other hand. The tasks themselves are demanding for automatic systems due to the variety of the linguistic ways people can express the same linguistic content. Here we focus on the evaluation of subjective information mining through the example of assigning opinionated keyphrases to product reviews and compare the results of human- and machine-based evaluation on finding opinionated keyphrases.

2 Related Work

As the task we aim at involves extracting keyphrases that are responsible for the author's opinion toward the product, aspects of both keyphrase extraction and opinion mining determine our methodology and evaluation procedure. There are several sentiment analysis approaches that make use of manually annotated review datasets (Zhuang et al., 2006; Li et al., 2010; Jang and Shin, 2010) and Wei and Gulla (2010) constructed a sentiment ontology tree in which attributes of the product and sentiments were paired.

For evaluating scientific keyphrase extraction, several methods have traditionally been applied. In the case of exact match, the gold standard keywords must be in perfect overlap with the extracted keywords (Witten et al., 1999; Frank et al., 1999) – also followed in the SemEval-2010 task on keyphrase extraction (Kim et al., 2010), while in other cases, approximate matches or semantically similar keyphrases are also accepted (Zesch and Gurevych, 2009; Medelyan et al., 2009). In this work we applied the former approach for the evaluation of opinion phrases and made a thorough comparison with the human judgement.

Here, we use the framework introduced in Berend (2011) and conducted further experiments based on it to point out the characteristics of the evaluation of opinionated keyphrase extraction. Here we pinpoint the severe differences in performance measures when the output is evaluated by humans compared to strict exact match principles and also examine the benefit of hand-annotated corpus as opposed

Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pages 99–103,
Jeju, Republic of Korea, 12 July 2012.

to an automatically crawled one. In addition, the extent to which original author keyphrases resemble those of independent readers' is also investigated in this paper.

3 Methodology

In our experiments, we used the methodology described in Berend (2011) to extract opinionated keyphrase candidates from the reviews. The system treats it as a supervised classification task using Maximum Entropy classifier, in which certain n-grams of the product reviews are treated as classification instances and the task is to classify them as proper or improper ones. It incorporates a rich feature set, relying on the usage of SentiWordNet (Esuli et al., 2010) and further orthological, morphological and syntactic features. Next, we present the difficulties of opinionated keyphrase extraction and offer our solutions to the emerging problems.

3.1 Author keyphrases

In order to find relevant keyphrases in the texts, first the reviews have to be segmented into analyzable parts. We made use of the dataset described in Berend (2011), which contains 2000 product reviews each from two quite different domains, i.e. mobile phone and video film reviews from the review portal `epinions.com`. In the free-text parts of the reviews, the author describes his subjective feelings and views towards the product, and in the sections *Pros and cons* and *Bottomline* he summarizes the advantages and disadvantages of the product, usually by providing some keyphrases or short sentences. However, these pros and cons are noisy since some authors entered full sentences while others just wrote phrases or keywords. Furthermore, the segmentation also differs from review to review or even within the same review (comma, semicolon, ampersand etc.). There are also non-informative comments such as *none* among cons. For the above reasons, the identification of the appropriate gold standard phrases is not unequivocal.

We had to refine the pros and cons of the reviews so that we could have access to a less noisy database. Refinement included segmenting pros and cons into keyphrase-like units and also bringing complex phrases into their semantically equivalent, yet much simpler forms, e.g. instead of *'even I found the phones menus to be confusing'*, we would like to have *'confusing phones menus'*. Refinement was carried out both automatically by using handcrafted transformation rules (based on POS patterns and parse trees) and manual inspection. The annotation guidelines for the human refinement and various statistics on the dataset can be accessed at `http://rgai.inf.u-szeged.hu/proCon`.

	$Auth.$	Ann_1	Ann_2	Ann_3
$Auth.$	–	0.415	0.324	0.396
Ann_1	0.601	–	0.679	0.708
Ann_2	0.454	0.702	–	0.713
Ann_3	0.525	0.690	0.688	–

Table 1: Inter-annotator agreement among the author's and annotators' sets of opinion phrases. Elements above and under the main diagonal refer to the agreement rates in Dice coefficient for pro and con phrases, respectively.

3.2 Annotator keyphrases

The second problem with regard to opinionated keyphrase extraction is the subjectivity of the task. Different people may have different opinions on the very same product, which is often reflected in their reviews. On the other hand, people can gather different information from the very same review due to differences in interpretation, which again complicates the way of proper evaluation.

In order to evaluate the difficulty of identifying opinion-related keyphrases, we decided to apply the following methodology. We selected 25 reviews related to the mobile phone Nokia 6610, which were also collected from the website `epinions.com`. The task for three linguists was to write positive and negative aspects of the product in the form of keyphrases, similar to the original pros and cons. In order not to be influenced by the keyphrases given by the author of the review, the annotators were only given the free-text part of the review, i.e. the original *Pros and cons* and *Bottomline* sections were removed. In this way, three different pro and con annotations were produced for each review, besides, those of the original author were also at hand. The inter-annotator agreement rate is in Table 1.

Concerning the subjectivity of the task, pro and con phrases provided by the three annotators and

Eval	Ref	Top-5	Top-10	Top-15
$3Ann_{\cup}$	man	32.14	44.66	53.92
$3Ann_{\cup}$	auto	27.68	38.17	45.78
$Merged_{\cup}$	man	28.52	41.09	52.18
$Merged_{\cup}$	auto	27.39	37.67	46.34
$3Ann_{\cap}$	man	34.89	43.31	44.92
$3Ann_{\cap}$	auto	29.96	34.34	35.54
$Merged_{\cap}$	man	24.75	26.12	22.22
$Merged_{\cap}$	auto	21.39	20.94	21.89
$Author$	man	27.14	33.5	35.24
$Author$	auto	20.61	22.34	25.03

Table 2: F-scores of the human evaluation of the automatically extracted opinion phrases. Columns Eval and Ref show the way gold standard phrases were obtained and if they were refined manually or automatically.

the original author showed a great degree of variety although they had access to the very same review. Sometimes it happened that one annotator did not give any pro or con phrases for a review whereas the others listed a bunch of them, which reflects that the very same feature can be judged as still tolerable, neutral or absolutely negative for different people. Thus, as even human annotations may differ from each other to a great extent, it is not unequivocal to decide which human annotation should be regarded as the gold standard upon evaluation.

3.3 Evaluation methodology

Since the comparison of annotations highlighted the subjectivity of the task, we voted for smoothing the divergences of annotations. We wanted to take into account all the available annotations which were manually prepared and regarded as acceptable. Thus, an annotator formed the union and the intersection of the pro and con features given by each annotator either including or excluding those defined by the original author. With this, we aimed at eliminating subjectivity since in the case of union, every keyphrase mentioned by at least one annotator was taken into consideration while in the case of intersection, it is possible to detect keyphrases that seem to be the most salient for the annotators as regards the given document. Thus, four sets of pros and cons were finally yielded for each review depending on whether the unions or intersections were determined purely on the phrases of the annotators excluding the original phrases of the author or including them. The following example illustrates the way new sets were created based on the input sets (in italics):

> $\boldsymbol{Pro_1}$: *radio, organizer, phone book*
> $\boldsymbol{Pro_2}$: *radio, organizer, loudspeaker*
> $\boldsymbol{Pro_3}$: *radio, organizer, calendar*
> **Union:** radio, organizer, calendar, loudspeaker, phone book
> **Intersection:** radio, organizer
> $\boldsymbol{Pro_{author}}$: *clear, fun*
> **Merged_Union**: radio, organizer, calendar, loudspeaker, phone book, clear, fun
> **Merged_Intersection**: $\emptyset$

The reason behind this methodology was that it made it possible to evaluate our automatic methods in two different ways. Comparing the automatic keyphrases to the union of human annotations means that a bigger number of keyphrases is to be identified, however, with a bigger number of gold standard keywords it is more probable that the automatic keywords occur among them. At the same time having a larger set of gold standard tags might affect the recall negatively since there are more keyphrases to return. On the other hand, in the case of intersection it can be measured whether the most important features (i.e. those that every annotator felt relevant) can be extracted from the text. Note that our strategy is similar to the one applied in the case of BLEU/ROUGE score (Papineni et al., 2002; Lin, 2004) with respect to the fact that multiple good solutions are taken into account whereas the application of union and intersection is determined by the nature of the task: different annotators may attach several outputs (in other words, different numbers of keyphrases) to the same document in the case of keyphrase extraction, which is not realistic in the case of machine translation or summarization (only one output is offered for each sentence / text).

3.4 Results

In our experiments, we used the opinion phrase extraction system based on the paper of Berend (2011). Results vary whether the manually or the automatically refined set of the original sets of pros and cons were regarded as positive training examples and also whether the evaluation was carried out

	Mobiles			Movies		
A/A	9.95	9.55	8.61	7.58	7.1	6.24
A/M	13.51	12.73	11.2	9.95	9.05	7.72
M/A	10.15	9.7	8.69	7.52	6.92	5.97
M/M	15.27	14.11	12.17	12.22	10.63	8.67

Table 3: F-scores achieved with different keyphrase refinement strategies. A and M as the first (second) character indicate the fact that the training (testing) was based on the automatically and manually defined sets of gold standard expressions, respectively.

against purely the original set of author-assigned keyphrases or the intersection/union of the manual annotations including and excluding the author-assigned keyphrases on the 25 mobile phone reviews. Results of the various combinations in the experiments for the top 5, 10 and 15 keyphrases are reported in Table 2 containing both cases when human and automatic refinement of the gold standard opinion phrases were carried out. Automatic keyphrases were manually compared to the above mentioned sets of keyphrases, i.e. human annotators judged them as acceptable or not. Human evaluation had the advantage over automated ones, that they could accept the extracted term '*MP3*' when there was only its mistyped version '*MP+*' in the set of gold standard phrases (as found in the dataset).

Table 3 presents the results of our experiments on keyphrase refinement on the mobiles and movies domains. In these settings strict matches were required instead of human evaluation. Results differ with respect to the fact whether the automatically or manually refined sets of the original author phrases were utilized for training and during the strict evaluation. Having conducted these experiments, we could examine the possibility of a fully automatic system that needs no manually inspected training data, but it can create it automatically as well.

4 Discussion and conclusions

Both human and automatic evaluation reveal that the results yielded when the system was trained on manually refined keyphrases are better. The usage of manually refined keyphrases as the training set leads to better results (the difference being 5.9 F-score on average), which argues for human annotation as opposed to automatic normalization of the gold standard opinion phrases. Note, however, that even though results obtained with the automatic refinement of training instances tend to stay below the results that are obtained with the manual refinement of gold standard phrases, they are still comparable, which implies that with more sophisticated rules, training data could be automatically generated.

If the inter-annotator agreement rates are compared, it can be seen that the agreement rates between the annotators are considerably higher than those between a linguist and the author of the product review. This may be due to the fact that the linguists were to conform to the annotation guidelines whereas the keyphrases given by the authors of the reviews were not limited in any way. Still, it can be observed that among the author-annotator agreement rates, the con phrases could reach higher agreement than the pro phrases. This can be due to psychological reasons: people usually expect things to be good hence they do not list all the features that are good (since they should be good by nature), in contrast, they list negative features because this is what deviates from the normal expectations.

In this paper, we discussed the difficulties of evaluating opinionated keyphrase extraction and also conducted experiments to investigate the extent of overlap between the keyphrases determined by the original author of a review and those assigned by independent readers. To reduce the subjectivity of the task and to alleviate the evaluation process, we presented our method that employs several independent annotators and we also compared the results of human and machine-based evaluation. Our results reveal that for now, human evaluation leads to better results, however, we believe that the proper treatment of polar expressions and ambiguous adjectives might improve automatic evaluation among others.

Besides describing the difficulties of the automatic evaluation of opinionated keyphrase extraction, the impact of training on automatically crawled gold standard opinionated phrases was investigated. Although not surprisingly they lag behind the ones obtained based on manually refined training data, the automatic creation of gold standard keyphrases can be a much cheaper, yet feasible option to manually refined opinion phrases. In the future, we plan to reduce the gap between manual and automatic evaluation of opinionated keyphrase extraction.

Acknowledgments

This work was supported in part by the NIH grant (project codename MASZEKER) of the Hungarian government.

References

Alexandra Balahur, Ester Boldrini, Andres Montoyo, and Patricio Martinez-Barco, editors. 2011. *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*. ACL, Portland, Oregon, June.

Gábor Berend. 2011. Opinion expression mining by exploiting keyphrase extraction. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1162–1170, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Andrea Esuli, Stefano Baccianella, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Domain-specific keyphrase extraction. In *Proceeding of 16th International Joint Conference on Artificial Intelligence*, pages 668–673. Morgan Kaufmann Publishers.

Hayeon Jang and Hyopil Shin. 2010. Language-specific sentiment analysis in morphologically rich languages. In *Coling 2010: Posters*, pages 498–506, Beijing, China, August. Coling 2010 Organizing Committee.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 21–26, Morristown, NJ, USA. ACL.

Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, and Hao Yu. 2010. Structure-aware review mining and summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 653–661, Beijing, China, August. Coling 2010 Organizing Committee.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. ACL.

Olena Medelyan, Eibe Frank, and Ian H. Witten. 2009. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1318–1327, Singapore, August. ACL.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the ACL*, pages 311–318, Philadelphia, Pennsylvania, USA, July. ACL.

Wei Wei and Jon Atle Gulla. 2010. Sentiment learning on product reviews via sentiment ontology tree. In *Proceedings of the 48th Annual Meeting of the ACL*, pages 404–413, Uppsala, Sweden, July. ACL.

Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Kea: Practical automatic keyphrase extraction. In *ACM DL*, pages 254–255.

Torsten Zesch and Iryna Gurevych. 2009. Approximate Matching for Evaluating Keyphrase Extraction. In *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing*, pages 484–489, September.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 43–50, New York, NY, USA. ACM.