EFFECTIVE CORPORA CREATION FOR SENTIMENT ANALYSIS

Peter KONCZ, Ján PARALIČ

Technical University of Kosice
Faculty of Electrical Engineering and Informatics
Letná 9/B, 042 00 Košice, Slovakia
peter.koncz@tuke.sk,jan.paralic@tuke.sk

Abstract. Sentiment analysis is currently a popular research area which methods are usually divided into two main types. Both of them, methods based on machine learning as well as dictionary based methods, are dependent on manually annotated corpora. These corpora contain manually annotated documents which are necessary for the training and evaluation of machine learning based methods and for the evaluation of dictionary based methods. The aim of our research summarized in this work is to provide methods to make this annotation more effective. The main methods used to achieve this are automatic corpora creation, active learning and annotation suggestions.

1. Introduction

Sentiment analysis is a research area which is devoted to the methods for automatic quantification of subjective evaluations expressed in textual content [1]. These methods can be divided into two main groups, where the first one consist of methods utilizing machine learning, while the second one consist of methods based on use of dictionaries and rules for their application [2]. However the corpora, as sets of manually annotated documents, are usually mentioned in context of machine learning based methods, they are necessary also for the evaluation of dictionary based methods. Hence a part of our research related to the sentiment analysis is devoted to the methods facilitating the creation of these corpora. In the following section we will shortly describe our results in their automatic creation as well as facilitation of the annotation process.

2. Automatic creation of corpora

The first way how to overcome the necessity of the manual annotation of documents is to utilize the existing sources of annotated documents. Examples of such a source are review sites, where the textual evaluations as well as the corresponding numerical evaluation are present. In the work [3] we evaluated the possibility of automated creation of corpora from online reviews. The reviews were crawled, scrapped and used as a training corpus. Moreover the training corpus, created as a selection from these reviews, was adapted to the content of the testing corpus. The results of the experiments showed that the proposed solution offers relatively good estimates of the emotional orientation of the text, although the increased precision of the classification through consideration of thematic similarities of documents within the training and testing corpora wasn't confirmed. The impact of the so called domain dependency of classifiers we analyzed also in the work [4]. We first used clustering in order to identify thematically similar evaluations and then we compared the within and between cluster sentiment analysis precision. The achieved results confirmed the possibility to use clustering according to the topic in order to increase the precision of sentiment analysis. This can be subsequently used for the selection of automatically gathered online reviews to the corpora which better fits the topic of the target domain.

3. Document annotation facilitation

The creation of corpora can be supported also by smart selection of documents for annotation using active learning methods. However active learning is a common strategy used in text mining, currently it is a lack of systematical research devoted to its application in the context of sentiment analysis. Hence in our work [1] we evaluated six active learning strategies applicable in the context of sentiment analysis, from which three were based on classification uncertainty and three on sentiment dictionaries. The results of experiments confirmed the increase of the corpus quality in terms of higher classification accuracy achieved on the test set for most of the evaluated strategies. In case of active learning strategy based on support vector machines we achieved more than 20% higher accuracy in comparison to the random strategy for selection of documents for annotation. Another way how to facilitate the annotation process is trough annotation suggestions. In our work [5] we proposed an annotation tool providing annotation suggestions based on naïve Bayes classifier, which can be accepted or corrected by the user. The proposed tool was adapted especially for the needs of aspect-based sentiment analysis, which is a sentiment analysis at the level of particular aspects of the evaluations. In our later works [6][7] we combined the annotation suggestions with active learning strategy based on uncertainty of naïve Bayes classifier and confirmed the increase of annotation effectiveness.

4. Conclusions

In this work we summarized our research devoted to the facilitation of corpora creation, either by automation of its creation as well as by incorporation of active learning for effective selection of documents for annotation. The achieved results confirmed the possibility of utilisation of booth groups of strategies.

Acknowledgement: This work was partially supported by the Slovak Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/1147/12 and partially by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

to other papers publishing the results that are summarized here

[1] Koncz, P., Paralič, J.: Active Learning Enhanced Document Annotation for Sentiment Analysis. In: Cuzzocrea, A., Kittl, C., Simos, D., Weippl, E., and Xu, L. (eds.) *Availability, Reliability, and Security in Information Systems and HCI*. pp. 345–353. Springer Berlin Heidelberg (2013).

[2] Machová, K., Koncz, P.: Metódy dolovania v konverzačnom obsahu so zameraním na analýzu sentimentu. *Datakon a Znalosti 2013*. pp. 2–14. VŠB-TUO, Ostrava (2013). (in Slovak)

[3] Koncz, P., Paralič, J.: Automated creation of corpora for the needs of sentiment analysis. *Proceedings of the 3nd RapidMiner Community Meeting and Conference (RCOMM 2012)*. pp. 107–113. Aachen: Shaker Verlag, Budapest, Hungary (2012).

[4] Koncz, P., Paralič, J.: Využitie zhlukovania na základe témy hodnotení v úlohách analýzy sentimentu. *Znalosti 2012: Sborník příspěvků 11. ročníku konference*. pp. 149–152. Praha: MATFYZPRESS, Mikulov (2012). (in Slovak)

[5] Smatana, M., Koncz, P.: Semi-automatic annotation tool for aspect-based sentiment analysis. Presented at the *13th Scientific Conference of Young Researchers of Faculty of Electrical Engineering and Informatics Technical University of Košice*. pp. 196-198. Košice: Elfa, Herľany (2013).

[6] Koncz, P., Smatana, M.: Active learning enhanced semi-automatic annotation tool for aspect-based sentiment analysis. Presented at the *SISY 2013: IEEE 11th International Symposium on Intelligent Systems and Informatics (2013)*. pp. 191-194. Subotica (2013).

[7] Koncz, P., Paralič, J.: Anotačný nástroj pre podporu aspektovo-orientovanej analýzy sentimentu s využitím aktívneho učenia. *Datakon a Znalosti 2013*. pp. 113–116. VŠB-TUO, Ostrava (2013). (in Slovak)