Automatic Detection and Attribution of Quotations

Richard FILIPČÍK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
richard@filipcik.sk

Although the text is one of the oldest ways of information preservation and information interchange, it still plays the key role in these tasks. In the present times thanks to the Internet however there are far easier ways for publishing and spreading textual works all over the globe. The only problem of this easy way of information spread is that we are often overloaded by information, hence it is almost impossible for us to get only the information we are currently looking for.

Natural language processing (NLP) is the field where we can look up the help. There are a lot of issues the NLP can help us with, one of which, little explored, but still interesting, is automatic detection and attribution of quotations. Those processes can be very useful in many domains since outputted data can be used for additional post-processing in a range of various ways.

The aim of our work is to propose a method for automatic detection and attribution of a direct or possibly even an indirect quotations in Slovak texts coming from the Internet sources such as newspaper articles. The output of our method should consist of a list of quotations extracted from the unstructured input text as well as names of their attributed originators.

We can divide our goal into three smaller parts. First and most important part of our goal is to detect quotations in an unstructured input text written in Slovak language with as high accuracy as possible. Although this can be seen as a relatively simple task to achieve, we have to take into account variability of punctuation semantics as well as possible improper use of punctuation characters by author of the text. There has already been published several studies about how to accomplish this task with relatively high accuracy [2]. Despite related studies are mostly dedicated to English languages, they still can be very helpful in several ways for our purpose.

Because quotation itself is almost meaningless without knowing its actual originator, our second "sub-goal" is to attribute detected quote to its proper originator. Generally, there are multiple approaches on how to achieve this. The character of the

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering

Spring 2015 PeWe Workshop, April 11, 2015, pp. 89–90.

input text plays very important goal here, as every method has different accuracy of attribution for different genres, e.g. a newspaper article or a novel [1].

The last and probably the hardest problem we would like to struggle with is detection of indirect quotations – paraphrases. Since there is currently almost no related research even for any non-Slovak language on this task, we take this as our secondary objective, however, we would like to make at least some progress on it.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG 1/0646/15.

References

[1] O'Keefe, T, Pareti, S and Curran, JR. A sequence labelling approach to quote attribution. In Proc. of the 2012 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012, pp. 790–799.

[2] Pouliquen, B, Steinberger, R and Best, C. Automatic detection of quotations in multilingual news. Proc. of Recent Advances in Natural Language Processing. 2007, pp. 487–492.