Keyphrases extraction from multiple documents

Maria Lomaeva
University of Potsdam
M.Sc. Data Science

Abstract

Keyphrases extraction is a task which allows to retrieve the most important topic phrases from a massive datasets of tweets, emails, reviews or scientific articles. Reading and understanding them requires effort, time and resources which most of the businesses can not allow themselves nowadays. Automating this process and achieving accurate predictions could benefit countless areas of industry and academia, as well as make private lives of many people easier.

After reading several articles humans can easily summarise and distinguish the main idea they were trying to convey. The problem becomes much more difficult, sometimes even unsolvable, when the number of the articles grows into hundreds or thousands. Unlike the humans, machines are very capable of processing a very big number of documents in a short time. However, understanding the natural language and making inferences based on text is not one of computers' strength. Accurate keyphrases extraction seeks to combine natural language understanding with the ability to process large data.

The first part is to identify the main topic of each article (or bunch or articles if they are short). The second is to rank the keyphrases based on the chosen metrics according to the domain, e.g. frequency or similarity to a certain topic. The main difficulty is to identify and summarise the most important words in the linguistic structure of text. So far both the traditional (Mihalcea, Tarau (2004); Bougouin et al. (2013)) and the most recent state-of-the-art methods (Zhang et al. (2020); Sun et al. (2019)) for solving this problem involve graphs, their algorithms and metrics. The assumption is, the natural language can be more accurately structured as a network of words, rather than their linear combination.

The datasets used for the model training contain news articles (Gallina et al. (2019)) and abstracts (Meng et al. (2017)) of scientific papers in English along with the corresponding keyphrases.

Keywords— Natural Language Processing, Information Retrieval, Text Summarisation, Keyphrases Extraction

References

Bougouin Adrien, Boudin Florian, Daille Béatrice. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction // Proceedings of the Sixth International Joint Conference on Natural Language Processing. Nagoya, Japan: Asian Federation of Natural Language Processing, X 2013. 543–551.

Gallina Ygor, Boudin Florian, Daille Beatrice. KPTimes: A Large-Scale Dataset for Keyphrase Generation on News Documents // Proceedings of the 12th International Conference on Natural Language Generation. Tokyo, Japan: Association for Computational Linguistics, X–XI 2019. 130–135.

Meng Rui, Zhao Sanqiang, Han Shuguang, He Daqing, Brusilovsky Peter, Chi Yu. Deep Keyphrase Generation // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, VII 2017. 582–592.

Mihalcea Rada, Tarau Paul. TextRank: Bringing Order into Text // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain: Association for Computational Linguistics, VII 2004. 404–411.

Sun Zhiqing, Tang Jian, Du Pan, Deng Zhi-Hong, Nie Jian-Yun. DivGraphPointer: A Graph Pointer Network for Extracting Diverse Keyphrases // CoRR. 2019. abs/1905.07689.

Zhang Haoyu, Long Dingkun, Xu Guangwei, Xie Pengjun, Huang Fei, Wang Ji. Keyphrase Extraction with Dynamic Graph Convolutional Networks and Diversified Inference. 2020.