CROSS-LANGUAGE PLAGIARISM DETECTION: A CASE STUDY OF EUROPEAN UNIVERSITIES ACADEMIC WORKS

Oleg Bakhteev[1], Yury Chekhovich[1], Georgy Gorbachev[1], Tatyana Gorlenko[1], Andrey Grabovoy[1], Kirill Grashchenkov[1], Alexander Kildyakov[1], Andrey Khazov[1], Vladislav Komarnitsky[1], Artemiy Nikitov[1], Aleksandr Ogaltsov[1], Alexandra Sakharova[1]

[1] *The Antiplagiat Company, Moscow, Russia*

The present report examines the problem of detecting cases of plagiarism in academic works with the use of automated plagiarism detection systems.

Over the past two decades, the research of methods of cross-language plagiarism detection has been rapidly evolving (Potthast et al., 2011; Franco-Salvador et al., 2016). The key prerequisites for such development are, on the one hand, a significant improvement in the methods of machine translation (Vaswani et al., 2017) that facilitate the generation of translated texts, and, on the other hand, in natural language processing methods (Belinkov et al., 2019), especially those using the deep learning (Li et al., 2018).

However, the scope of their application in the plagiarism detection systems oriented towards the verification of works on the commercial scale was quite limited until recently. The leading producers were either not announcing such opportunities or this feature was implemented nominally. The ambiguity of translation, high requirements to equipment, and significant time inputs for building indexes, configuring the algorithm, and processing a single document during the research were the most significant obstacles towards the broad-scale use. A number of studies were aimed at developing the methods based on the analysis of bibliometric data, such as title, author(s), abstract, bibliography (for example, see Mazov et al., 2016; Mazov and Gureev, 2017). These methods are characterized by significantly lower requirements to equipment and time inputs, but the scope of their application is also rather limited. In general, the opportunities provided by the cross-language plagiarism cases have been considered by the leading experts as accidents rather than as the result of a targeted research.

Since 2017 the developers of the Antiplagiat system, which is widely used in universities in Russia and the former Soviet countries (Nikitov et al., 2012), have been working on algorithms and services for the translated plagiarism detection (Bakhteev et al., 2019), which are used to process large amounts of verifiable documents that are compared with commercial scale source databases (with hundreds of millions of source documents). First, an algorithm was developed that allowed to detect text reuse from English-language sources in Russian texts; then other language pairs were added, with a unique algorithm for each pair configured separately. In 2020, the cross-language plagiarism detection algorithm was developed to trace text reuse by 100 languages.

The technology for detecting translated plagiarism cases, implemented in the Antiplagiat system, is implemented in two stages: finding the so-called candidate texts and comparing text pieces in the verified document with the candidate documents. The shingles method for document search in a large collection of documents is used at the stage of candidate selection. For each document in the collection, the text is normalized, split into n-grams, and the hashes of these n-grams are then saved in the index. During the search for cross-language plagiarism cases, an automatic machine translation system translated the document into a language from the collection. At this stage, the requirements to the quality of machine translation are not high, which is why the chain of translation tools is used to cover all possible language pairs made by 100 supported languages. Multilingual methods of sentence vectorization are used for document comparison: all the sentences from the verified document and the documents in the collection selected at the

first stage are placed in the vector space using the deep learning models. As such deep learning model, the distilled version of Language-agnostic BERT Sentence Embedding model used (Feng et al., 2020). This model showed a high quality in many natural language processing tasks related to multilingual document analysis. The model assumes that if the vectors of some sentences are located next to each other in the vector space, they are similar in meaning, and therefore can be considered as an instance of text reuse.

The present study is aimed at searching for previously undetected cases of cross-language plagiarism in the papers published by European universities in their open access repository. We test the hypothesis stipulating that some authors, who wanted to benefit from the imperfection of plagiarism detection tools, used translated parts of texts by including them in their works and not providing the reference to actual authors.

In this research, we used the scientific papers from the repositories of the 25 leading universities in the countries with a high level of education, where English is not the official language: France, Germany, Portugal, Spain and Sweden. More than 10 thousand works were analyzed during the research. The analyzed collection of papers is balanced across the considered countries and mainly contains papers written in the most common language of each country. The experiment is conducted by comparing the collection of 10 thousand multilingual documents against the large web collection of documents. The size of the web collection is 50 million and it contains mainly documents written in English, Russian, and other European languages. We analyze the obtained results and classify detected cases into several groups such as improper text reuse, self-citation, bibliographic source citation, and legal documents citation. The analysis of detected cases is provided in the report.

REFERENCES

Bakhteev O., Ogaltsov A., Khazov A., Safin K., and Kuznetsova R. (2019) CrossLang: the system of cross-lingual plagiarism detection. In *Proc. of the KDD Workshop on Deep Learning for Education*. https://truth-discovery-kdd2019.github.io/papers/crosslang.pdf

Belinkov, Y., and Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7, (pp. 49–72).

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic BERT sentence embedding. *arXiv preprint*, arXiv:2007.01852.

Franco-Salvador, M., Gupta, P., Rosso, P., and Banchs, R. E. (2016). Cross-language plagiarism detection over continuous-space-and knowledge graph-based representations of language. *Knowledge-based systems*, 111, (pp. 87–99).

Potthast M., Barrón-Cedeño A., Stein B., and Rosso P. (2011). *Cross-language plagiarism detection*. 45(1), (pp. 45–62). https://doi.org/10.1007/s10579-009-9114-z

Mazov N., Gureev V., and Kosyakov D. (2016) On the development of a plagiarism detection model based on citation analysis using a bibliographic database. *Scientific and Technical Information Processing*, 43(4), (pp. 236-240). https://doi.org/10.3103/S0147688216040092

Mazov N., and Gureev V. (2017) Study results for the detection of translated plagiarism using bibliometric databases. *Nauchnye i tekhnicheskie biblioteki-scientific and technical libraries* 12, (pp. 87–96). https://doi.org/10.33186/1027-3689-2017-12-87-96

Li Z., Jiang X., Shang L., and Li H. (2018) Paraphrase Generation with Deep Reinforcement Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. (pp. 3865–3878). http://dx.doi.org/10.18653/v1/D18-1421

Nikitov A., Orchakov O., and Chehovich Y. (2012) Plagiarism in works of undergraduate and graduate students: problem and methods of counteraction. *University Management: Practice and Analysis*, 5 (81). (pp. 61–68).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010). https://dl.acm.org/doi/10.5555/3295222.3295349