Doctoral School of Information and Biomedical Technologies
Polish Academy of Sciences (TIB PAN)

Topic:

Algorithms for Extracting Bibliographies from Footnotes of Scientific Texts in Polish Language

Supervisor:

Przemyslaw Korytkowski, professor at National Information Processing Institute – National Research Institute (e-mail: pkorytkowski@opi.org.pl)

Description:

The scientific goal of the project is to develop algorithms using artificial intelligence methods to extract citations from footnotes from scientific texts in Polish automatically.

International institutions such as UNESCO, the World Bank and the European Commission analyze the state of science in individual countries based on scientific publications. These analyses do not reflect the scientific standing of many Polish scientists, disciplines or journals. One important reason is that references to cited works are placed in footnotes in the social sciences and humanities, which citation extraction tools cannot analyze.

The project's research hypothesis: Algorithms using computer vision, natural language processing and machine learning methods will allow building a system to automatically extract and link references from footnotes from scientific articles and monographs written in Polish.

Extracting references from footnotes is a non-trivial task for several reasons. It requires a combination of computer vision (CV) techniques, which identify and extract text from footnotes, and natural language processing (NLP) techniques, which analyze and understand the meaning of these texts. The lack of unambiguous structure in footnotes makes the process of reference extraction difficult, as information about the author, title, publisher and other data are expressed in different forms and places in the footnote. Footnotes often contain more than one reference and may also refer to previous references and be dependent on each other, which requires tracking and understanding the relationship between them for correct extraction.

Recommended reading:

1. Kunnath, S., Pride, D., Herrmannova, D., and Knoth, P. (2021) A Meta-analysis of Semantic Classification of Citations. Quantitative Science Studies.
2. Xu, Y., Li, et al. (2020). LayoutLM: Pre-training of Text and Layout for Document Image Understanding. Proceedings of the 26th ACM SIGKDD.
3. Garncarek, Ł., et al. (2021). LAMBERT: Layout-Aware Language Modeling for Information Extraction. Lecture Notes in Computer Science, 532–547.
4. Li, J., Tan, G., Ke, X., Si, H., & Peng, Y. (2022). Object detection based on knowledge graph network. Applied Intelligence.
5. Fang, Y., et al. (2017). Object Detection Meets Knowledge Graphs. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence.