Gabriela Nowakowska
Named entity recognition and information extraction from various documents

Abstract

The thesis presents a novel application of named entity recognition and information extraction methods during the processing of documents of various types. The thesis consists of four scientific articles that have been published and presented at conferences of international scope.

Chapter 1 describes the research problem, motivation and results obtained, as well as the structure and scope of the thesis. It also includes an overview and a brief summary of the included articles. Each description is preceded by information about the authors, the venue and type of presentation, and the contribution of the thesis author.

Chapters 2 and 3 present research work related to the application of named entity recognition methods, which served as part of the solution to problems defined in competitions held at international conferences. Chapter 2 includes a description of the translation system developed as part of the WMT 2022 conference. Chapter 3 presents new models for lemmatization of named entities that were used in the solution of the competition organized as part of the Slavic NLP 2023 workshop.

Chapters 4 and 5 focus on articles presenting neural network models developed as part of participation in the Industrial PhD program. Chapter 4 describes the TILT model created as part of the work on extracting information from documents with a two-dimensional structure (text and vision layer). Chapter 5 presents the STable model, which is an evolution of the TILT model and is used to extract tabular data.

At the end of the thesis, the appendices include three certificates received from the organizers of the ICDAR 2019, WMT 2022 and Slavic NLP 2023 conferences, as well as the first pages of two patents obtained related to the TILT and STable models. Lastly, declarations of the contributions of the co-authors of each article presented in the thesis are included.

Gabriela Nowakowska