Utilizing Structured Resources in Neural Language Models

Michał Turski

Abstract

The majority of research in the field of Natural Language Processing is focused on processing plain text. While this paradigm is highly effective for numerous use cases, such as machine translation, summarization, and chatbots, it fails to fully harness the richness of many texts created by and for humans. Documents, on the other hand, convey meaning not only through their textual content but also through their structure and visual features. A key challenge tackled by this thesis is to develop solutions that combine recent advancements in language modeling with structural information to improve the processing and comprehension of documents.

This thesis comprises five scientific papers in the domain of document understanding, divided into two main sections. The first section focuses on evaluating document understanding models, introducing the first benchmark in this area and proposing a novel dataset in the scientific domain. The proposed benchmark includes a diverse range of document types and tasks, enabling a comprehensive evaluation of document understanding models. The novel dataset, designed specifically for scientific documents, assesses models' ability to reason over both tables and text simultaneously.

The second section of this thesis tackles various challenges in the document understanding domain, proposing innovative solutions to enhance model performance. These include a diverse, multilingual corpus for pretraining document-oriented language models, enabling improved understanding of documents across languages and domains; a novel architecture extending capabilities of Transformer model by using structural information, enhancing its ability to process and comprehend structured documents; and a framework for generating tables using a language model, enabling the creation of structured data from natural language input.

Overall, this thesis contributes to the development of more accurate and useful document understanding models, enabling improved processing and comprehension of rich, structured documents.

Michał Turski