Title: Novel Methods and Datasets for Intelligent Document Processing

Author: Dawid Jurkiewicz

Abstract

The field of Intelligent Document Processing (IDP) is gaining prominence as organizations struggle to utilize their ever-growing data effectively.
This thesis aims to contribute innovative solutions and datasets to the IDP domain. The focus is set on two key areas within IDP: Span Identification (SI) and Document Understanding (DU). Span Identification involves localizing relevant spans of text containing specific information, while Document Understanding encompasses various tasks related to comprehending and extracting meaningful information from visually rich documents.

Significant emphasis is placed on addressing the challenges posed by low-data scenarios, which are prevalent in various business use cases.
A few-shot SI dataset and a unique approach for sub-sequence matching with few examples are proposed to address this.
Besides the few-shot setting, methods for identifying and classifying propaganda spans are presented.

Furthermore, a multi-modal end-to-end Transformer-based model for Document Understanding is introduced. The model efficiently comprehends layout information, textual semantics, and visual cues present in the document and can answer various document-related questions posed in the natural language.
Additionally, the first DU benchmark is proposed, allowing the community to measure the DU field's state accurately.
Lastly, a challenging DU competition is showcased. The task features novel question and answer type pairs over multi-domain, multi-industry, and multi-page documents, encouraging the development of solutions with strong generalization capabilities in low-data regimes.

Dawid Jurkiewicz