

Tytuł pracy w jęz. pracy: Novel Methods and Datasets for Intelligent Document Processing

Tytuł pracy w jęz. polskim: Nowe metody i zbiory danych do inteligentnego przetwarzania dokumentów

Autor: Dawid Jurkiewicz

## Streszczenie

Dziedzina inteligentnego przetwarzania dokumentów (ang. Intelligent Document Processing) zyskuje na znaczeniu, ponieważ organizacje mają trudności z efektywnym wykorzystaniem swoich stale przybywających danych.

Niniejsza rozprawa ma na celu wnieść wkład w innowacyjne rozwiązania i zbiory danych dla dziedziny inteligentnego przetwarzania dokumentów.

Nacisk kładziony jest na dwa kluczowe obszary w ramach dziedziny inteligentnego przetwarzania dokumentów: identyfikację relewantnych fragmentów tekstu (ang. Span Identification) i problematykę rozumienia dokumentów (ang. Document Understanding). Identyfikacja relewantnych fragmentów tekstu zajmuje się lokalizacją fragmentów tekstu zawierających określone informacje, podczas gdy problematyka rozumienia dokumentów obejmuje różne zadania związane z pojmowaniem i wydobywaniem istotnych informacji z dokumentów bogatych wizualnie.

Duży nacisk położony jest na zmierzenie się z wyzwaniami związanymi z małą ilością dostępnych danych, które są powszechne w różnych zastosowaniach biznesowych. Aby rozwiązać ten problem, zaproponowano zbiór danych dla identyfikacji relewantnych fragmentów tekstu na podstawie kilku przykładów oraz unikatową metodę do wyszukiwania podsekwencji na podstawie kilku przykładów. Oprócz rozwiązań bazujących na kilku przykładach, przedstawiono metody do identyfikacji i klasyfikacji fragmentów tekstu zawierających propagandę.

Ponadto wprowadzono multimodalny model oparty na architekturze Transformer dla problematyki rozumienia dokumentów.

Model rozumie semantykę tekstu, cechy wizualne i strukturę dokumentu oraz potrafi odpowiadać na różne sformułowania w języku naturalnym dotyczące dokumentu. Dodatkowo zaproponowano pierwszy zestaw zbiorów danych pozwalający społeczności na dokładną obserwację postępów w dziedzinie rozumienia dokumentów. Na koniec zaprezentowano wymagający konkurs dla problematyki rozumienia dokumentów. Zadanie zawiera nowatorskie pary typów pytań i odpowiedzi dla wielodomenowych, wielobranżowych i wielostronicowych dokumentów, zachęcając do opracowywania rozwiązań, które mają znaczące możliwości uogólniania w przypadku małej ilości dostępnych danych.

*Dawid Jurkiewicz*