Wykorzystanie zasobów ustrukturyzowanych w neuronowych modelach języka

Michał Turski

Streszczenie

Większość badań w dziedzinie przetwarzania języka naturalnego koncentruje się na przetwarzaniu tekstu. Choć ten paradygmat jest bardzo skuteczny w wielu zastosowaniach, takich jak tłumaczenie maszynowe, automatyczne podsumowywanie i systemy dialogowe, nie potrafi w pełni wykorzystać bogactwa wielu dokumentów tworzonych przez i dla ludzi. Dokumenty przekazują znaczenie nie tylko przez warstwę tekstową, ale także poprzez Kluczowym wyzwaniem podejmowanym w tej pracy jest proponowanie rozwiązań rozszerzających najnowsze modele języka o wykorzystanie informacji strukturalnych celem poprawy jakości przetwarzania dokumentów.

Niniejsza rozprawa składa się z pięciu prac naukowych w domenie rozumienia dokumentów i jest podzielona na dwie główne sekcje. Pierwsza sekcja dotyka problemu oceny modeli rozumienia dokumentów, wprowadzając pierwsze wyzwanie (ang. benchmark) w tej domenie oraz proponuje nowy zbiór danych oparty na piśmiennictwie naukowym. Zaproponowane wyzwanie obejmuje różnorodny zakres dokumentów i zadań, umożliwiając kompleksową ocenę modeli rozumienia dokumentów. Nowy zbiór danych, zaprojektowany specjalnie dla dokumentów naukowych, ocenia zdolność modeli do rozumienia tekstu z wykorzystaniem tabeli jako dodatkowego źródła informacji.

Druga sekcja tej pracy podejmuje różne wyzwania w domenie rozumienia dokumentów, proponując innowacyjne rozwiązania mające na celu poprawę jakości modeli. Są to: różnorodny, wielojęzyczny korpus do uczenia modeli języka przeznaczonych dla dokumentów, umożliwiający lepsze rozumienie dokumentów w różnych językach i dziedzinach; nowa architektura rozszerzająca model Transformer o kodowanie informacji strukturalnych, co pozwala na przetwarzanie dokumentów o~bogatej strukturze; oraz metoda generowania tabel przy użyciu modelu języka, umożliwiająca tworzenie strukturalnych danych z wejścia w postaci tekstu.

Podsumowując, ta praca przyczynia się do rozwoju modeli rozumienia dokumentów, umożliwiając lepsze przetwarzanie i analizę dokumentów o bogatej strukturze.

Michał Turski