Dr hab. inż. Agnieszka Ławrynowicz, prof. PP
Instytut Informatyki
Politechnika Poznańska
ul. Piotrowo 2, 60-965 Poznań
alawrynowicz@cs.put.poznan.pl

30 września 2024 r.

Recenzja rozprawy doktorskiej

Michała Turskiego

zatytułowanej:

Utilizing Structured Resources in Neural Language Models

1. Problem badawczy i jego znaczenie

Głównym problemem badawczym podjętym w pracy jest opracowanie rozwiązań, które rozszerzą najnowsze modele językowe o wykorzystanie informacji strukturalnych, aby poprawić jakość przetwarzania dokumentów. Rozwiązania te mają znaczenie dla rozwoju modeli rozumienia dokumentów, szczególnie w zakresie analizy dokumentów o bogatej strukturze. Dokumenty takie są powszechnie spotykane jak np. artykuły naukowe, dokumenty prawne czy też techniczne instrukcje. Wykonane prace badawcze zostały motywowane poprzez problemy i potrzeby zidentyfikowane w przemyśle.

2. Wkład autora

Badania przedstawione w tej rozprawie są zgodne ze standardową metodologią eksperymentalną stosowaną w dziedzinie uczenia maszynowego

Rozprawa składa się z pięciu artykułów o charakterze naukowym podzielonych na dwie grupy tematyczne: mierzenie stanu zrozumienia dokumentu (ang. *measuring state of document understanding*) i modelowanie języka ustrukturyzowanego w 2D (ang. *2D-structured language modeling*). W ramach pierwszej grupy zagadnień zaproponowano nowe zbiory danych i wyzwania (ang. *benchmark*), z wykorzystaniem tabeli jako dodatkowego źródła informacji. W ramach drugiej grupy podejmowane zostały wyzwania związane z rozumieniem dokumentów mające na celu poprawę jakości istniejących modeli poprzez dokonanie poprawy w trzech zadaniach: przygotowanie danych, reprezentację struktur wejściowych oraz generowanie ustrukturyzowanego wyjścia z modelu.

Mierzenie stanu zrozumienia dokumentu

1. DUE: End-to-End Document Understanding Benchmark

Publikacja wprowadza test porównawczy (*benchmark*) do mierzenia aktualnego stanu rozumienia dokumentu. Benchmark obejmuje wizualne odpowiadanie na pytania, ekstrakcję kluczowych informacji i zadania polegające na czytaniu maszynowym w różnych domenach i układach dokumentów obejmujących tabele, wykresy, listy i infografiki. Na *benchmark* składa się siedem zbiorów danych. Zaproponowano także nowy format reprezentacji zbiorów danych, który uspójnia prezentację różnych zadań rozumienia dokumentów.

Praca została opublikowana na bardzo prestiżowej konferencji NeurIPS w 2021 roku.

Wkład autora dotyczył wzbogacenia zbiorów danych diagnostycznymi anotacjami aby dokonać wglądu w mocne i słabe strony każdego zgłoszonego modelu, organizację i kontrolę procesu anotacji przez ludzi oraz edycję publikacji.

2. Arxiv Tables: Document Understanding Challenge Linking Texts and Tables

Publikacja prezentuje zbiór danych dla zadania rozumienia tekstu w kontekście towarzyszącej jemu tabeli. Dane to prace naukowe, zawierające znaki specjalne i inną nomenklaturę domenową. Wykorzystano zarówno kody źródłowe Latex, jak i wizualizacje graficzne artykułów oraz połączono tabele z odniesieniami w tekście głównym, aby stworzyć quasi-zestaw danych do zadania odpowiadania na pytania poprzez maskowanie wybranych fragmentów dostępnych w tabeli. W czasie publikacji zbiór ten był największym dostępnym dla przedstawionego zadania.

Praca została opublikowana na warsztatach towarzyszących konferencji ICDAR w 2023 roku.

Wkład autora dotyczył propozycji modeli bazowych w celu umożliwienia porównywania różnych rozwiązań, przeprowadzenia eksperymentów, opisu powiązanych prac.

Modelowanie języka ustrukturyzowanego w 2D

3. CCpdf: Building a High Quality Corpus for Visually Rich Documents from Web Crawl Data

Publikacja prezentuje przygotowanie dużej skali, zróżnicowanego i wielojęzycznego korpusu dla celów trenowania wstępnego (pretrenowania) modeli rozumienia dokumentów. Źródłem dokumentów jest *Common Crawl*, otwarte repozytorium danych *web*. Opracowano proces, który zawiera detektory dokumentów w danych, detektory języka, filtry w celu zapewnienia różnorodności dokumentów oraz narzędzia OCR.

Praca została opublikowana na konferencji ICDAR w 2023 roku.

Wkład autora to prace konceptualizacyjne i metodologiczne, zaprojektowanie i konceptualizacja narzędzia do zarządzania korpusem, analizy i eksploracji korpusu, uruchamianie procesu generowania korpusu, pisanie i edycja publikacji oraz zarządzanie projektem.

4. LAMBERT: Layout-Aware Language Modeling for information extraction

Publikacja wprowadza nową metodę modelowania dokumentów o skomplikowanej strukturze, gdzie skomplikowane elementy układu mają wpływ na semantykę. W ramach pracy zmodyfikowano architekturę kodera Transformer, aby wykorzystać informacje o układzie uzyskane z systemu OCR, bez potrzeby ponownego uczenia modelu semantyki języka od podstaw. Dane wejściowe modelu zostały wzbogacone jedynie o współrzędne ramek ograniczających tokeny, dzięki czemu nie są używane surowe obrazy. W efekcie powstaje model językowy uwzględniający układ tekstu, który można dostroić do kolejnych zadań.

Praca została opublikowana na konferencji ICDAR w 2021 roku i nagrodzona jako najlepsza publikacja powiązana z przemysłem.

Wkład autora to utworzenie zbioru danych do trenowania modelu wraz z metodologią odfiltrowywania dokumentów, które nie dotyczą biznesu lub prawa, implementacja procesu trenowania modeli LAMBERT oraz edycja publikacji.

5. STable Table Generation Framework for Encoder-Decoder Models

W publikacji zaproponowano framework dla modeli neuronowych typu „text-to-table" mających zastosowanie np. do ekstrakcji elementów zamówienia, wyodrębniania wspólnych podmiotów i relacji lub populacji baz wiedzy. Publikacja prezentuje nową metodę generowania wyjścia w strukturze tabeli z modelu języka. Polega ona na generowaniu komórek tabeli na podstawie miary pewności (ang. *confidence*) i estymacji błędu, aby najpierw wypełniać komórki gdzie prawdopodobieństwo błędu jest najmniejsze.

Publikacja została zaprezentowana na konferencji EACL 2024.

Wkład autora to konceptualizacja i metodyka pracy badawczej, pomysły dotyczące pretrenowania, przygotowania zbiorów danych implementacja rozwiązań bazowych, przeprowadzenie eksperymentów, przygotowanie publikacji oraz zarządzanie projektem.

3. Poprawność

W przedstawionych artykułach nie zidentyfikowano istotnych błędów.

4. Wiedza kandydata

Kandydat posiada znajomość teorii i know-how w obszarze przetwarzania i rozumienia języka naturalnego, który wchodzi w zakres dyscypliny informatyka. Kandydat korzysta z różnych technik opisanych w literaturze, wskazując na odniesienia literaturowe i wykazując się biegłością w korzystaniu z różnych narzędzi i zbiorów narzędzi. Zademonstrował tę wiedzę i umiejętności w zadaniach takich jak konstrukcja wyzwań (ang. *benchmark*) w zakresie rozumienia dokumentów oraz modelowanie języka z uwzględnieniem informacji strukturalnych w celu rozumienia dokumentów o skomplikowanej strukturze.
Kandydat zdobył doświadczenie w pracy badawczej w przemyśle (w firmie Applica.ai) oraz w ramach dwóch projektów badawczych finansowanych z funduszy Unii Europejskiej w tematyce: 1) robotyzacji procesów wymagających rozumienia tekstu oraz 2) rozwiązania do ekstrakcji informacji z dokumentów skanowanych.
Kandydat jest także współautorem patentu.

5. Inne uwagi

Wraz z rozwojem dużych modeli języka, obszary pracy doktorskiej stają się kluczowymi obszarami naukowymi i technologicznymi w zakresie sztucznej inteligencji. Doktorant zaprojektował i opracował różnorodne rozwiązania, w tym nowe zbiory danych i wyzwania oraz rozwiązania mające na celu poprawę jakości modeli rozumienia dokumentów.
Kandydat uczestniczył w badaniach zespołowych. Raz jest pierwszym, głównym autorem przedstawionej publikacji, raz jest jednym z takich autorów oraz raz występuje w roli autora korespondującego.
Wyniki prac znalazły praktyczne zastosowanie w systemach produkcyjnych w przemyśle.

6. Podsumowanie

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez art. 187 Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (z późniejszymi zmianami)[1] moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

A. Czy rozprawa zawiera oryginalne rozwiązanie problemu naukowego? (wybierz jedną opcję stawiając znak **X**)

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

B. Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

C. Czy kandydat posiada umiejętność samodzielnego prowadzenia pracy naukowej?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

A. Łangrzeszyn

Podpis

[1] http://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20190000276