Poznań, 15 grudnia 2023r.

dr hab. Agata Filipowska, prof. UEP
Katedra Informatyki Ekonomicznej
Uniwersytet Ekonomiczny w Poznaniu

Recenzja rozprawy doktorskiej mgra Dawida Jurkiewicza pt. „Novel Methods and Datasets for Intelligent Document Processing"

Promotor rozprawy: prof. UAM dr hab. Filip Graliński

Recenzja sporządzona na podstawie Ustawy z dnia 20 lipca 2018r. Prawo o szkolnictwie wyższym i nauce (Dz.U. z 2021r. poz. 478).

1. Tematyka rozprawy doktorskiej

Przedstawiona rozprawa doktorska, na którą składa się cykl sześciu publikacji powiązanych wspólnym tematem, związana jest z domeną analizy języka naturalnego i dotyczy inteligentnego przetwarzania dokumentów, a dokładniej rozpoznawania fragmentów tekstu (ang. Span Identification) oraz rozumienia dokumentów (ang. Document Understanding). Trzy z przedstawionych publikacji dotyczą zagadnienia rozpoznawania fragmentów tekstu, a trzy pozostałe rozumienia dokumentów.

Doktorant w publikacjach dokonuje przeglądu literatury, proponuje zbiory danych, na których można tworzyć lub dokonywać porównania efektywności istniejących metod, przeprowadza analizę wskaźnikową działania opisanych w literaturze metod (ang. benchmark analysis), a także proponuje nowe metody rozwiązujące wskazane problemy naukowe.

Pierwsza z załączonych w pracy publikacji dotyczy tematyki wyszukiwania informacji prawnej, dostarczając zbioru danych, na jakim można przeprowadzać eksperymenty, a który został przygotowany przez autorów artykułu. Następnie, autorzy dokonują porównania metod z wykorzystaniem uczenia półnadzorowanego, proponując autorski potok przetwarzania danych na potrzeby wykrywania klauzul w tekstach prawnych.

Drugi z artykułów dotyczy zaproponowania metody wyszukiwania informacji w długich frazach temporalnych. Autorzy wykazują w niej, że wykorzystanie w trakcie przeszukiwania tekstów wielu zapytań równocześnie daje lepsze rezultaty niż tworzenie z nich podzbioru i wyznaczania dystansów między zdaniem konsensusu a przeszukiwanymi frazami. Zaproponowana w artykule metoda jest oryginalna i stanowi wkład do nauki.

Trzecia z publikacji jest raportem dotyczącym rozwiązania przygotowanego w ramach konkursu SemEval-2020. Jest to jedyna publikacja z załączonych w dysertacji, w której Doktorant jest pierwszym autorem. Artykuł prezentuje system dla klasyfikacji faz i artykułów propagandowych, wskazując na szczegóły implementacyjne rozwiązania oraz wyniki jego

działania. Opracowany przez autorów system zajął czołowe miejsca (pierwsze i drugie) w przytoczonym konkursie.

W kolejnym z artykułów, autorzy prezentują podejście do rozumienia/podsumowania dokumentów na podstawie ich struktury graficznej, cech wizualnych raz semantyki, wykorzystujące nowatorskie rozwiązania bazujące na architekturze Transformer. Dokonują bardzo szczegółowego testu opracowanej metody, z wykorzystaniem podejść z zakresu m.in. klasyfikacji dokumentów, ekstrakcji informacji faktograficznej z dokumentów, jak i analizy odpowiedzi na zadawane pytania (ang. question answering). Publikacja wskazuje na autorskie rozwiązanie przygotowane przez jej autorów.

Piąty z artykułów dotyczy zaproponowania sposobu oceny (benchmarku) dla metod dokonujących rozumienia dokumentów. W artykule opisano sposób pozyskania i przygotowania zbiorów danych, zadania związane z podsumowaniem tekstów (celem precyzyjnej późniejszej ewaluacji metod), a także zaprojektowano i zaimplementowano podejścia bazowe (ang. baseline), względem których będą mogli porównywać się kolejni badacze, i wskazano na wyzwania w dziedzinie. Autorzy w ten sposób podsumowali pracę wykonaną na rzecz społeczności zajmującej się tematyką rozumienia dokumentów (ang. Document Understanding).

Ostatnia z publikacji zawartych w dorobku autora jest raportem z konkursu przeprowadzonego celem porównania metod pozwalających na rozumienie dokumentów. W artykule wskazano na sposób przygotowania konkursu (opis wyzwań, przygotowanie zbiorów danych, itp.), a także protokół oceny zgłoszonych do konkursu rozwiązań. W konkursie wzięło udział 3 uczestników, którzy zgłosili do oceny 6 metod, poddanych ewaluacji przez autorów.

Temat zaproponowanej rozprawy doktorskiej jest spójny z przedstawionym dorobkiem, a także stanowi wkład do dyscypliny informatyka.

2. Cele rozprawy doktorskiej

Celem głównym pracy było zaproponowanie nowych, opartych na przetwarzaniu języka naturalnego, metod oraz zbiorów danych dla zadań rozpoznawania fragmentów tekstu (ang. Span Identification) oraz rozumienia dokumentów (ang. Document Understanding), co oznaczało wkład do domeny inteligentnego przetwarzania tekstów.

Celami szczegółowymi rozprawy zaproponowanymi przez Doktoranta były:

1. Opracowanie zbioru danych oraz metod identyfikacji fragmentów w dokumentach.
2. Opracowanie metody/systemu pozwalającego na identyfikację fragmentów tekstu zawierających zdania propagandowe oraz opracowanie systemu klasyfikacji tekstów zawierających takie zdania.
3. Zbudowanie kompleksowego modelu dla rozumienia tekstów (bazującego zarówno na cechach semantycznych, jak i wizualnych tekstów).
4. Przygotowanie zbiorów danych dla zadań rozumienia dokumentów.

Cele pracy zostały sformułowane poprawnie i pozwalają na uzyskanie wkładu naukowego do dyscypliny informatyka.

3. Wkład badawczy dysertacji

Wkład badawczy Doktoranta wynika z publikacji opisanych w sekcji pierwszej recenzji. Podsumowując go w wymiarze ilościowym i jakościowym, dokonano:

1. Zaproponowania i opublikowania zbiorów danych:
    a. dla domeny informacji prawniczej i klauzul zawartych w dokumentach (zaanotowano 21 typów klauzul w 586 dokumentach),
    b. DUE: End-to-End Document Understanding Benchmark, na który składa się szereg powiązanych zbiorów danych uspójnionych przez autorów pracy,
    c. wykorzystywanych w konkursie przeprowadzonym w ramach konferencji IDCAR2023.
2. Zaproponowania metod:
    a. wykrywania klauzul umownych w tekstach prawnych (zadanie rozpoznawania fragmentów tekstów),
    b. wyszukiwania informacji w długich frazach temporalnych (zadanie rozpoznawania fragmentów tekstów),
    c. klasyfikacji faz i artykułów propagandowych (zadanie rozpoznawania fragmentów tekstów),
    d. rozumienia dokumentów na podstawie ich struktury graficznej, cech wizualnych raz semantyki, wykorzystujące nowatorskie rozwiązania bazujące na architekturze Transformer (zadanie rozumienia dokumentów).
3. Przygotowania inicjatyw dla społeczności:
    a. przygotowanie inicjatywy DUE: End-to-End Document Understanding Benchmark zawierającej oprócz zbioru danych także benchmark dla ewaluacji metod w zadaniach związanych z rozumieniem dokumentów,
    b. przeprowadzenia konkursu w ramach konferencji IDCAR2023 polegającego na ocenie metod rozumienia dokumentów bazujących na ich strukturze i cechach semantycznych.

Prezentowane publikacje stanowią wkład naukowy do dyscypliny informatyka. Wyzwaniem wynikającym z przedstawionych publikacji jest fakt, że Doktorant jest głównym autorem tylko jednej z 6 publikacji wskazanych w rozprawie. Publikacja ta jest raportem z opracowania systemu do wykrywania fraz propagandowych oraz klasyfikacji dokumentów zawierających takie frazy. Dla pozostałych publikacji wykazano wkład merytoryczny autora, który jednak pokrywał się on w dużej części (w sensie opisu w tabelach zawartych w deklaracjach o współautorstwie w aneksie B) z wkładem innych autorów w prezentowane publikacje. Dodatkowo, procentowy wkład badawczy Doktoranta nie jest wykazany. Oprócz publikacji zawartych w rozprawie, zgodnie z Google Scholar, Doktorant jest autorem dodatkowych 9 artykułów naukowych.

Na podstawie przytoczonych publikacji i dorobku można stwierdzić, że Doktorant potrafi pracować w zespołach badawczych i dostarczać wartościowych wyników prac, natomiast konieczne jest precyzyjniejsze opisanie wykonanych zadań i dorobku uzyskanego przez

Doktoranta. Nawet w przypadku konkursów czy projektów badawczych wskazanych w aneksie A rozprawy nie opisano roli i wyników uzyskanych przez Doktoranta, co jest krytyczne dla oceny wkładu do nauki i umiejętności samodzielnego formułowania i rozwiązywania problemów badawczych.

4. Struktura dysertacji

Przedstawiona dysertacja składa się z czterech części. W pierwszej z nich zawarto wprowadzenie do tematyki pracy oraz scharakteryzowano cele badawcze podjęte przez Doktoranta. Sekcja ta, przygotowana na potrzeby zgłoszenia dysertacji do recenzji, stanowi ciekawe podsumowanie problematyki badawczej podjętej przez Doktoranta.

Drugą część pracy stanowią trzy artykuły stanowiące wkład do tematyki identyfikacji fragmentów tekstu. Dwa z nich są publikacjami w czasopiśmie, ostatni został zawarty w materiałach konferencyjnych, jednak wszystkie uzyskały 140 pkt. zgodnie z punktacją MEiN. Ostatni z artykułów uzyskał także nagrodę Best Paper Award. Struktura przedstawionych artykułów jest zgodna ze sztuką, choć w przypadku artykułu pt. „ApplicaAI at SemEval-2020 Task 11: On RoBERTa-CRF, Span CLS and Whether Self-Training Helps Them" brakuje poprawnego odwołania się do stanu wiedzy zawartego w literaturze przedmiotu.

Trzecia z części pracy dotyczy obszaru rozumienia dokumentów, z wykorzystaniem informacji o ich strukturze, nie tylko zawartości semantycznej. Trzy artykuły zawarte w tej sekcji pracy zostały opublikowane w materiałach konferencyjnych lub około-konferencyjnych, jednak każdy z nich uzyskał przynajmniej 140 pkt. zgodnie z punktacją MEiN (jest wśród nich jeden artykuł za 200 pkt.). Artykuły są wysokiej jakości i prezentują wiele odwołań bibliograficznych do aktualnej literatury.

W ostatniej z sekcji pracy zawarto dwa aneksy, jeden prezentujący konkursy i projekty badawcze, w jakich autor brał udział, zaś drugi zawiera deklaracje o współautorstwie związane z prezentowanymi w pracy artykułami. Krytyczne komentarze dot. aneksów zawarto w sekcji czwartej recenzji dot. wkładu badawczego dysertacji.

W pracy nie zawarto podsumowania odnoszącego się do celów pracy i wskazującego na stopień ich realizacji. Podsumowanie takie mogłoby podkreślić wkład Doktoranta i przyczynić się do rozwiązania wątpliwości związanych z badaniami wykonywanymi we współautorstwie, a niewystarczająco precyzyjnie opisanymi w deklaracjach zawartych w aneksie B rozprawy (tabela 1.1 na stronie 20 jedynie przytacza informacje z deklaracji, nie dokonując ich rozwinięcia).

5. Ocena metodyki badawczej

W rozprawie brakuje dokładnego omówienia metodyki badawczej stosowanej przez autora, jednakże w poszczególnych publikacjach wskazano na sposób prowadzenia pracy badawczej. Opisane w publikacjach oryginalne metody są poprawnie wprowadzone

i przetestowane zgodnie ze sztuką, co więcej scharakteryzowano także ich pozycjonowanie względem obecnego stanu wiedzy.

O umiejętnościach Doktoranta z zakresu projektowania i oceny metod mogą stanowić dwie publikacje dot. podsumowania dokumentów: „DUE: End-to-end document understanding benchmark" oraz „ICDAR 2023 Competition on Document UnderstanDing of Everything (DUDE)", w których opisano proces przygotowania zbiorów danych dla potrzeb oceny metod (w tym dla konkursu), a także opracowania benchmarku dla zadań związanych z rozumieniem tekstu. Obydwie publikacje pokazują bardzo dobre zrozumienie metodyki prowadzenia i oceny wyników badawczych w dziedzinie inteligentnego przetwarzania dokumentów.

6. Kwestie do dyskusji

W trakcie przygotowania do obrony pracy warto byłoby, poza innymi punktami wskazanymi w recenzji, dodatkowo zwrócić uwagę na:

- Przenośność opisanych metod oryginalnych na inne zastosowania np. inne klauzule prawnicze lub zbliżony problem w innej domenie. Jaki nakład pracy (oraz zakres zadań) byłby potrzebny dla implementacji metody dla nowego problemu?
- Efektywność działania metod w przypadku zastosowań wielodomenowych.
- Wpływ przygotowanego zestawu danych oraz benchmarku w zakresie podsumowania dokumentów na kolejne badania / kolejnych autorów w dziedzinie.
- Wyzwania wynikające z wielodomenowości tekstów dla rozumienia dokumentów.

7. Podsumowanie

Przedmiotem rozprawy doktorskiej mgr Dawida Jurkiewicza jest inteligentne przetwarzanie dokumentów, a dokładniej rozpoznawanie fragmentów tekstu (ang. Span Identification) oraz rozumienie dokumentów (ang. Document Understanding). Na rozprawę składa się cykl sześciu wysoko punktowanych (140 pkt. i więcej zgodnie z punktacją MEiN) publikacji, w których Doktorant jest jednym z autorów. Rozprawę doktorską przygotowano w języku angielskim wraz z krótkim streszczeniem w języku polskim.

Rozprawa doktorska prezentuje wiedzę teoretyczną Doktoranta oraz umiejętność prowadzenia pracy naukowej w dziedzinie nauk ścisłych i przyrodniczych, w dyscyplinie informatyka. Przedstawione publikacje prezentują umiejętność dokonania syntezy oraz oceny wyników badawczych uzyskanych przez innych badaczy (przegląd literatury) oraz opracowania własnych rozwiązań postawionych problemów. Jakość przestawionych publikacji oceniam wysoko, a każda z nich stanowi oryginalne podejście do rozwiązywanego problemu. Głównym wyzwaniem wynikającym z oceny publikacji jest niewystarczająco precyzyjne określenie wkładu Doktoranta, wskazane w recenzji.

W rozprawie Doktorant, mimo zawartych w recenzji komentarzy krytycznych nie umniejszających wartości uzyskanych wyników, wykazał się rzetelną wiedzą i znajomością

technik prowadzenia badań w dyscyplinie naukowej oraz umiejętnością rozwiązywania podstawionych problemów badawczych. Cel rozprawy został zrealizowany, podobnie jak cele szczegółowe postawione w pracy.

Przedstawiona do recenzji rozprawa spełnia warunki określone w art. 187 Ustawy z dnia 20 lipca 2018r. Prawo o szkolnictwie wyższym i nauce (Dz.U. z 2021r. poz. 478) dla dziedziny nauk ścisłych i przyrodniczych, w dyscyplinie informatyka. **W związku z powyższym występuję z wnioskiem o uznanie rozprawy doktorskiej jako spełniającej wymagania obowiązującej ustawy oraz dopuszczenie Doktoranta do jej publicznej obrony.**