Politechnika Warszawska
Wydział Elektroniki i Technik Informacyjnych

dr hab. inż. Artur Janicki, prof. uczelni
Instytut Telekomunikacji
Warszawa, 16.09.2024

RECENZJA ROZPRAWY DOKTORSKIEJ

mgr. **MICHAŁA TURSKIEGO**
pt. „*Utilizing Structured Resources in Neural Language Models*", przedłożonej Radzie Naukowej Informatyki Uniwersytetu im. Adama Mickiewicza w Poznaniu

PRZEDMIOT ROZPRAWY, GŁÓWNE CELE PRACY

Praca Pana **mgr. Michała Turskiego** pt. „*Utilizing Structured Resources in Neural Language Models*" dotyczy zagadnienia rozumienia dokumentów o bogatej strukturze, tj. zawierających np. nagłówki, tabele, grafikę, podpisy itp.

Przedstawiona rozprawa jest pracą badawczą z elementami praktycznymi. W swojej rozprawie Autor nie formułuje konkretnych tez, natomiast formułuje dwa następujące cele:

- Zmierzenie (sic!) aktualnego stanu wiedzy w dziedzinie zrozumienia dokumentów o bogatej strukturze poprzez ocenę wydajności istniejących modeli;
- Udoskonalanie istniejących rozwiązań poprzez rozwój i ocenę nowatorskich podejść, które zwiększają dokładność, wydajność i użyteczność narzędzi do rozumienia dokumentów;

Rozprawa została napisana w języku angielskim i ma formę ciągu 5 publikacji poprzedzonych krótkim wstępem. Na końcu rozprawy znajdują się oświadczenia współautorów o wkładzie do poszczególnych publikacji. Rozprawa składa się z 6 rozdziałów i zajmuje 134 strony.

Rozdział 1. wprowadza w tematykę rozprawy, przedstawia cele rozprawy i krótkie streszczenia poszczególnych publikacji.

Rozdział 2. stanowi publikacja „*DUE: End-to-End Document Understanding Benchmark*", opisująca proponowany sposób ewaluacji rozumienia dokumentów.

Rozdział 3. stanowi publikacja „*Arxiv Tables: Document Understanding Challenge Linking Texts and Tables*", prezentująca wyzwanie dot. rozumienia dokumentów o bogatej strukturze pochodzących z repozytorium arXiv.

Nowowiejska 15/19,
00-665 Warszawa
tel.: 22 234 77 22
e-mail:
Artur.Janicki
@pw.edu.pl

Rozdział 4. zawiera publikację „*CCpdf: Building a High Quality Corpus for Visually Rich Documents from Web Crawl Data*", opisująca metodologię tworzenia korpusu dokumentów PDF w celu m.in. uczenia dwuwymiarowych modeli językowych.

Rozdział 5. stanowi publikacja „*LAMBERT: Layout-aware language modeling for information extraction*", prezentująca nowatorską metodę dwuwymiarowego modelowania dokumentów.

Rozdział 6. stanowi publikacja „*STable: Table Generation Framework for Encoder-Decoder Models*", opisująca zaproponowaną metodę konwersji tekstu na tabelę.

OCENA OSIĄGNIĘĆ DOKTORANTA

Przedłożona rozprawa ma formę połączonych pięciu publikacji, poprzedzonych krótkim wprowadzeniem, zawierającym m.in. streszczenia poszczególnych artykułów. Wszystkie publikacje są bardzo dobre – były przyjęte i wygłoszone na topowych konferencjach dotyczących sieci neuronowych, analizy dokumentów czy lingwistyki obliczeniowej (NeurIPS 2021, ICDAR 2021 & 2023, EACL 2024). Potwierdzeniem ich wartości dla środowiska badaczy zajmujących się przetwarzaniem języka naturalnego jest liczba cytowań: na dzień pisania recenzji jest ich łącznie 116, co wg mnie jest wynikiem bardzo wysokim dla kandydatów do stopnia doktora.

Należy jednak zauważyć, że wkład Doktoranta do niektórych publikacji wydaje się stosunkowo niewielki. Tylko w jednej publikacji (Rozdział 4, z konferencji ICDAR 2023, 140 pkt. MNiSW) Doktorant jest pierwszym autorem, a w kolejnej (Rozdział 6, z konferencji EACL 2024, 140 pkt. MNiSW) jest co prawda drugim, ale jego wkład jest określony jako równy wkładom pierwszego i trzeciego autora.

Tymczasem udział Doktoranta w publikacji stanowiącej Rozdział 2 (z bardzo dobrej konferencji NeurIPS 2021, 200 pkt. MNiSW) oraz stanowiącej Rozdział 5 (z konferencji ICDAR 2021, 140 pkt. MNiSW), gdzie Doktorant jest odpowiednio 5. i 6. autorem, jak wynika z deklaracji zamieszczonych w Załączniku A, wydaje się ograniczać do ważnych, ale jednak pobocznych zadań, takich jak przygotowanie zbiorów diagnostycznych, kontrolowanie procesu ręcznej annotacji czy implementacja procesu uczenia.

Kluczowe zadania w tych dwóch rozdziałach/ publikacjach (m.in. opracowanie koncepcji rozwiązania, wybór składowych zbiorów danych do zbioru DUE, pomysł zastosowania modelu BERT do ekstrakcji kluczowych informacji z dokumentu, wreszcie przeprowadzenie eksperymentów i napisanie artykułów) zostały wykonane przez innych, zwykle pierwszych autorów tych publikacji. Zresztą wspomniane wyżej dwie publikacje weszły wcześniej w skład rozprawy doktorskiej dr. inż. Tomasza Stanisławka.

Nowowiejska 15/19,
00-665 Warszawa
tel.: 22 234 77 22
e-mail:
Artur.Janicki
@pw.edu.pl

Politechnika Warszawska
Wydział Elektroniki i Technik Informacyjnych

Dlatego, żeby nie oceniać osiągnięć innych badaczy, w dalszej części recenzji skupię się na tych częściach rozprawy, w których Doktorant zadeklarował najbardziej istotny udział. Tutaj jednym z największych osiągnięć Doktoranta jest opracowanie metodyki tworzenia zbiorów dokumentów PDF. Metodyka ta obejmuje m.in. wyszukiwanie lokalizacji dokumentów PDF, rozpoznawanie języka, usuwanie spamu oraz rozpoznawanie, czy dokument wymaga optycznego rozpoznawania znaków (OCR). Wykorzystując zaproponowaną metodykę oraz zasoby Common Crawl, Doktorant stworzył zbiór dokumentów PDF, który nazwał CCpdf. Zbiór ten zawiera łącznie ponad 1,3 mln dokumentów PDF dla 11 języków. Zbiór ten znalazł zastosowanie m.in. w przemyśle, gdzie służy do uczenia modeli rozumienia dokumentów – fakt praktycznego wykorzystania wyników prac Doktoranta uważam za bardzo ważny. Skrypty do tworzenia i pobrania zbioru zostały udostępnione społeczności badawczej na serwerze GitHub, co również wg mnie zasługuje na uznanie.

Kolejne ważne osiągnięcie Doktoranta to współudział w opracowaniu koncepcji tzw. STables, której celem jest uformowanie wyjścia modelu językowego do postaci tabelarycznej. Zaproponowany model zamienia tekst na tabelę, umieszczając kolejną wartość komórki w wierszu lub kolumnie, kierując się najniższym prawdopodobieństwem błędu. Metoda ta została przetestowana na ośmiu różnych zbiorach i w większości przypadków okazała się lepsza od wcześniejszych rozwiązań. Jako znaczący sukces należy uznać też uzyskanie amerykańskiego patentu dla tego rozwiązania. Należy tutaj nadmienić, że Doktorant kierował projektem zajmującym się opracowaniem tej metody, brał udział w jej implementacji oraz zaproponował dodanie uczenia wstępnego modelu z wykorzystaniem m.in. algorytmu TILT. Natomiast z deklaracji współautorów publikacji stanowiącej Rozdział 6 zrozumiałem, że pomysł metody STable należy jednak do p. Michała Pietruszki. Doktoranta będę więc znów prosił o potwierdzenie stopnia udziału w tym niewątpliwie ważnym osiągnięciu.

To samo dotyczy metody LAMBERT, która polega na rozszerzeniu modelu Transformer (w tym wypadku modelu RoBERTa) o kodowanie informacji o położeniu tokenów na stronie. Umożliwiło to podniesienie wartości F1-score na publicznym zbiorze SROIE z 97,81% dla rozwiązań konkurencyjnych do 98,17% dla proponowanego rozwiązania. Zgodnie z deklaracją współautorów, rolą Doktoranta w tym projekcie było przygotowanie uczącego zbioru danych i przeprowadzenie uczenia modelu. Chętnie poznałbym szczegóły zaangażowania Doktoranta w opracowanie samego algorytmu LAMBERT. Warto podkreślić, że artykuł prezentujący to rozwiązanie na konferencji ICDAR 2021 uzyskał nagrodę Best Industry Related Paper Award, co zasługuje na uznanie i potwierdza praktyczne znaczenie badań Doktoranta i całego zespołu.

Nowowiejska 15/19,
00-665 Warszawa
tel.: 22 234 77 22
e-mail:
Artur.Janicki
@pw.edu.pl

Doktorant był aktywnym uczestnikiem dwóch projektów naukowych: „Uniwersalna platforma robotyzacji procesów wymagających rozumienia tekstu o unikalnym poziomie automatyzacji wdrożenia i obsługi” oraz „Hiper-OCR – innowacyjne rozwiązanie do ekstrakcji informacji z dokumentów skanowanych” (oba finansowane z

Politechnika Warszawska
Wydział Elektroniki i Technik Informacyjnych

POIR). Przyczynił się także do rozwoju badań w dziedzinie przetwarzania języka naturalnego w międzynarodowym środowisku naukowym. Rozdział 3 opisuje opracowanie wyzwania dotyczącego rozumienia tekstów z tabelami z repozytorium arXiv. Jak rozumiem, w tym projekcie Doktorant zajmował się przygotowaniem modeli stanowiących odniesienie dla uczestników wyzwania. Nie znalazłem jednak informacji, czy to wyzwanie zostało upublicznione lub czy może jest to planowane.

Uważam, że rozprawa doktorska bez wątpienia prezentuje oryginalne rozwiązania w zakresie przetwarzania dokumentów w języku naturalnym o bogatej strukturze. Ponieważ jednak opisywane prace były przeprowadzane zespołowo, do wyjaśnienia pozostaje udział Doktoranta w opracowaniu tych rozwiązań. Niewątpliwie natomiast Doktorant wykazał się sprawnym posługiwaniem warsztatem badawczym oraz biegłą znajomość teoretyczną z dziedziny informatyki.

UWAGI SZCZEGÓŁOWE

- 9-stronicowy wstęp do ciągu artykułów jest bardzo przejrzysty i „esencjonalny", wydaje mi się jednak zbyt skrótowy. Być może właśnie tutaj Doktorant mógłby bardziej szczegółowo opisać swoje pomysły. Na pewno mógłby też bardziej szczegółowo omówić cele i znaczenie swojej pracy.
- Tytuł rozprawy nie do końca wydaje mi się odpowiadać tematyce rozprawy. „Utilizing" sugeruje wyłącznie wykorzystywanie struktur dokumentów, a Doktorant też takie struktury (np. tabele) generuje. Z kolei „Resources" wydaje mi się trochę za szerokie: w sumie chodzi „tylko" dokumenty, a nie np. o zbiory audio czy wideo.
- „*Measuring state of document understanding*" – czy nie chodzi po prostu o „*Evaluation of document understanding*"?
- Niektóre sformułowania są zbyt potoczne. Zwrot „*to measure the state-of-the-art (in the domain of structure-rich document understanding)*" jest raczej skrótem myślowym, a zapewne chodzi o ewaluację poziomu rozumienia dokumentów na aktualnym poziomie wiedzy. Zachęcam Doktoranta do dbałości o staranne, precyzyjne sformułowania w oficjalnych dokumentach.
- Rozdział 4 (str. 361): Fig. 10 – "*text occurs more frequently on the right side of a page*" – czy mapa ciepła nie pokazuje czegoś dokładnie przeciwnego?
- Rozdział 5 (str. 540): Tabela 1 – Niejasna jest dla mnie różnica między modelem LAMBERT (16M) oraz LAMBERT (75M), skoro oba mają wg tej tabeli 125M parametrów.

STRONA EDYCYJNA PRACY

Rozprawa doktorska **mgr. Michała Turskiego** w większości składa się z publikacji, które zostały już wielokrotnie sprawdzone przez współautorów, recenzentów i edytorów materiałów konferencyjnych, tak więc ich forma jest bardzo staranna. Tym niemniej zauważam pewne niedociągnięcia natury językowej i typograficznej, m.in.:

Nowowiejska 15/19,
00-665 Warszawa
tel.: 22 234 77 22
e-mail:
Artur.Janicki
@pw.edu.pl

Politechnika Warszawska
Wydział Elektroniki i Technik Informacyjnych

- Mimo że poszczególne artykuły mają stanowić kolejne rozdziały, to ich numery nigdzie się nie pojawiają – brak chociażby stron tytułowych rozdziałów.
- Kłopotliwa jest nieciągła numeracja stron.
- Niespójna jest pisownia nazwy repozytorium arXiv.
- Niespójna jest pisownia metryki F1-score.
- Tabela 1 w Rozdziale 6 (str. 2459) jest mało przejrzysta. Stosowane metryki są wymieszane, a wyjaśnień trzeba szukać dopiero w tekście. O ile wyjaśnione jest znaczenie podkreślonych wartości, to znaczenia wartości wytłuszczonych należy się już tylko domyślać.
- (Rozdział 3, str. 99): Table [2] – chyba powinno być Table 2.
- (Rozdział 4, str. 360): Fig. 8 – chmura tagów słabo spełnia swoją rolę, gdyż słowa pisane mniejszą czcionką są nieczytelne, z kolei „Microsoft Word" pojawia się w wielu wersjach.
- (Rozdział 5, str. 532): pusty nawias przy nazwisku p. Garncarka.
- Brakujące przedimki, np.: "*for Kleister Charity dataset*" => "*for the Kleister Charity dataset*" (Rozdział 5, str. 541), "*presents dataset*" => "*presents a dataset*" (Rozdział 1, str. 4).
- "*3 different datasets*" => "*three different datasets*" (np. Rozdział 5, str. 532)

Wyżej wymienione uchybienia edycyjne nie umniejszają jednak dorobku Doktoranta.

WNIOSKI KOŃCOWE

W podsumowaniu stwierdzam, że cele badawcze postawione w rozprawie zostały osiągnięte. Doktorant zbadał aktualny stan wiedzy w dziedzinie zrozumienia dokumentów o bogatej strukturze poprzez ocenę wydajności istniejących modeli, a także udoskonalił istniejące narzędzia do rozumienia tych dokumentów. Rozprawa doktorska **mgr. Michała Turskiego** zawiera też oryginalne rozwiązania w zakresie zastosowania wyników własnych badań naukowych w sferze gospodarczej.

Stwierdzam, że przedstawiona rozprawa doktorska spełnia warunki określone w art. 187 ustawy Prawo o szkolnictwie wyższym i nauce (Dz.U. z 2022 r. poz. 574 z późn. zm.), dlatego niniejszym wnioskuję o dopuszczenie Doktoranta, Pana **mgr. Michała Turskiego**, do publicznej obrony jego rozprawy doktorskiej.

Nowowiejska 15/19,
00-665 Warszawa
tel.: 22 234 77 22
e-mail:
Artur.Janicki
@pw.edu.pl

dr hab. inż. Artur Janicki, prof. uczelni
Politechnika Warszawska

PODPIS ZAUFANY
ARTUR
JANICKI
30.09.2024 23:27:45 [GMT+2]
Dokument podpisany elektronicznie
podpisem zaufanym