Recenzja pracy doktorskiej mgr. inż. Tomasza Stanisławka

„Ekstrakcja informacji z dokumentów o bogatej strukturze graficznej"

Krzysztof Jassem

22 stycznia 2022

1 Wstęp

Celem niniejszej recenzji jest stwierdzenie, czy rozprawa doktorska mgr. inż. Tomasza Stanisławka zatytułowana „Ekstrakcja informacji z dokumentów o bogatej strukturze graficznej" spełnia wymagania ustawowe (Art.187. Ustawy „Prawo o szkolnictwie wyższym i nauce"). Ustawa stwierdza w punkcie 3., że „Rozprawę doktorską może stanowić ... zbiór opublikowanych i powiązanych tematycznie artykułów naukowych". Właśnie ta forma prezentacji wyników została wybrana przez Doktoranta. Ustawa zezwala (punkt 2.), aby przedmiotem rozprawy było zastosowanie wyników w sferze gospodarczej. Tomasz Stanisławek jest uczestnikiem doktoranckich studiów wdrożeniowych i przedmiotem jego prac jest wdrażanie wyników badawczych w działalności firmy Applica, w której jest zatrudniony na pełen etat.

Wszystkie artykuły wchodzące w skład recenzowanej Rozprawy są publikacjami zbiorowymi. Fakt ten utrudnia ocenę wymagań stawianych w punktach 1. i 2. art. 187: „Rozprawa doktorska prezentuje ...umiejętność **samodzielnego** prowadzenia pracy naukowej..." oraz „Przedmiotem pracy doktorskiej jest ...oryginalne rozwiązanie w zakresie zastosowania **własnych** badań naukowych."

W ramach poniższej recenzji, będę starał się zatem odpowiedzieć na następujące pytania:

1. Czy zestaw publikacji podany w Recenzji stanowi zbiór artykułów naukowych powiązanych tematycznie?
2. Czy merytoryczny poziom pracy jest adekwatny do wymagań stawianych rozprawom doktorskim?
3. Czy Rozprawa wykazuje umiejętność samodzielnego prowadzenia pracy naukowej?
4. Czy prezentowane zastosowania są efektem oryginalnych i własnych rozwiązań Doktoranta?

Ponadto, część recenzji poświęcona będzie stronie formalnej pracy, a w tym stosowanej terminologii i jej tłumaczeniu na język polski.

Recenzję kończy podsumowanie zawierające rekomendację.

2 Czy zestaw publikacji poddany recenzji stanowi zbiór artykułów naukowych powiązanych tematycznie?

W skład Rozprawy wchodzą cztery artykuły naukowe, które omówione są po kolei w sekcji 2. Rozprawy.

2.1 Omówienie artykułu „Named Entity Recognition – Is there a glass ceiling?"

Artykuł jest przeglądem stosowanych współcześnie metod i mechanizmów rozpoznawania jednostek nazewniczych w tekstach. Celem autorów była analiza typów błędów popełnianych przez stosowane rozwiązania, aby opracować narzędzia o wyższej skuteczności. W ramach pracy opracowano taksonomię błędów rozpoznawania jednostek nazewniczych i odtworzono eksperymenty raportowane we współczesnych pracach. Pozwoliło to na określenie przyczyn błędów i sformułowanie wniosku, aby w przyszłych eksperymentach brać pod uwagę kontekst szerszy niż jedno zdanie oraz strukturę graficzną dokumentu.

Jednostka nazewnicza to występująca w tekście fraza, która jednoznacznie identyfikuje obiekt lub byt. Rozpoznawanie jednostek nazewniczych jest niezbędnym składnikiem ekstrakcji informacji z tekstów, który pozwala na określenie, czego tekst dotyczy. Można więc z przekonaniem stwierdzić, że artykuł jest powiązany z tematem Rozprawy.

2.2 Omówienie artykułu „Key information extraction datasets involving long documents with complex layouts"

Celem badań omawianych w artykule jest opracowanie publicznie dostępnych zbiorów danych do trenowania systemów ekstrakcji informacji z dokumentów o bogatej strukturze graficznej. Opracowano dwa zbory danych: pierwszy z nich, o nazwie Kleister NDA, zawiera umowy o zachowaniu poufności (które oryginalnie były zapisane w postaci elektronicznej), a drugi, o nazwie Kleister Charity – sprawozdania organizacji charytatywnych (pozyskane z dokumentów papierowych za pomocą optycznego rozpoznawania graficznego). Każdy element obu zbiorów danych zawiera dokument w formacie PDF. W przypadku pierwszego zbioru każdy dokument oznaczony jest manualnie opracowaną listą jednostek nazewniczych, które powinny zostać wyekstrahowane z dokumentu. W drugim zbiorze danych informacje dotyczące poszczególnych organizacji charytatywnych, które mają być ekstrahowane z dokumentów, zostały pozyskane ze źródeł zewnętrznych (bez konieczności pełnej weryfikacji manualnej). Efektem prac są dwa wysokiej jakości oznaczone zbiory danych, udostępnione publicznie, które mogą służyć do trenowania mechanizmów rozpoznawania jednostek nazewniczych.

Oczekiwanym elementem publikowanego zbioru danych do trenowania systemów uczenia maszynowego jest tzw. rozwiązanie bazowe (ang. *baseline solution*), czyli przykładowy mechanizm wykonu-

jący zadanie, dla którego zbiór został stworzony. Dzięki temu badacze poszukujący nowych metod przetwarzania podanego zbioru danych mogą porównać jakość swoich rozwiązań z rozwiązaniem bazowym. Autorzy artykułu opracowali „mocne" rozwiązanie bazowe o nazwie LAMBERT którego jakość (wg miary F1) jest wyższa od wyników innych współcześnie stosowanych metod.

Omawiany artykuł jest zatem kolejnym krokiem w drodze do osiągnięcia celu Rozprawy. Poszerza zadanie omówione w pierwszym artykule o rozpoznawania jednostek w dokumentach graficznych. Ponadto, poprzez opublikowanie zbioru danych, umożliwia dalszy rozwój proponowanych metod.

2.3 Omówienie artykułu „DUE – benchmark do mierzenia postępów w dziedzinie rozumienia tekstów"

Kluczowym elementem ekstrakcji informacji jest rozumienie treści zapisanych w dokumentach. Termin „rozumienie dokumentów" obejmuje kilka zadań, które mogą być stosowane albo rozłącznie (jako cel sam w sobie) lub łącznie z innymi (w celu zintegrowania informacji uzyskanych przez kilka mechanizmów rozumienia dokumentów). Przykładowe zadania związane z rozumieniem dokumentów to:

- ekstrakcja informacji kluczowych z dokumentu,
- klasyfikacja tematyczna dokumentu,
- analiza układu dokumentu,
- odpowiadanie na pytania na podstawie informacji zawartych w dokumencie,
- wnioskowanie na podstawie informacji zawartych w dokumencie.

Autorzy twierdzą, że istniejące zbiory danych opracowywane są z myślą o zastosowaniu tylko w jednym z powyższych zadań. Z tego powodu przygotowali zbiór danych, o nazwie DUE, przeznaczony dla wielu zadań rozumienia dokumentów łącznie. Podobnie, jak w przypadku zbiorów Kleister, również dla zbioru DUE autorzy przetestowali skuteczność dostępnych metod przetwarzania dokumentów, wskazując rozwiązanie, które w momencie publikacji osiągało najwyższą skuteczność.

Omawiany artykuł jest naturalną kontynuacją eksperymentów prowadzonych przez Doktoranta. Rozszerza zakres prowadzonych eksperymentów w drodze do osiągnięcia celu nakreślonego w Rozprawie.

2.4 Omówienie artykułu „Lambert - Layout-aware language modelling for information extraction"

W artykule omawia się autorski model języka o nazwie LAMBERT, który jest uzupełnieniem modelu wprowadzonego przez badaczy firmy Facebook, o nazwie RoBERTa. W stosunku do pierwowzoru, do modelu wprowadzono dodatkowo informacje o strukturze dokumentu. Aby ocenić

jakość autorskiego modelu języka porównano jego skuteczność z innymi modelami w standardowych zadaniach ekstrakcji informacji. Eksperymenty przeprowadzono na kilku dostępnych zbiorach danych, m.in. na zbiorach omówionych w poprzednich pracach. Doświadczenia wykazały poprawność przyjętej metodologii: wprowadzenie do modelu informacji o strukturze dokumentów poprawiło skuteczność testowanych algorytmów.

2.5 Wniosek

Omawiane prace pozytywnie weryfikują przyjęty plan badawczy Doktoranta.

W kolejnych etapach pracy Doktorant:

- przeanalizował aktualny stan wiedzy i sformułował konstruktywne wnioski na temat obszarów problemowych zastanych metod;
- opracował zbiory danych, na których mógł testować skuteczność algorytmów ekstrakcji informacji z dokumentów;
- opracował model języka, w oparciu o który możliwym stało się opracowanie metod skuteczniejszych niż zastane.

Z pełnym przekonaniem stwierdzam, że zestaw publikacji poddany recenzji stanowi zbiór artykułów naukowych powiązanych tematycznie.

3 Czy merytoryczny poziom pracy jest adekwatny do wymagań stawianych rozprawom doktorskim?

Wszystkie artykuły wchodzące w skład Rozprawy zostały przyjęte na wysoko punktowane konferencje międzynarodowe, a mianowicie:

- Conference on Computational Natural Language Processing 2019 (CoNLL) – 140 punktów MEiN
- Document Analysis and Recognition 2021 – 140 punktów MEiN (dwie prace)
- Conference on Neural Information Processing Systems 2021 (NeurIPS) – 200 punktów MEiN

Ponadto, jedna z omawianych prac została wyróżniona na konferencji nagrodą w kategorii *Best Industry Related Paper*. Proces recenzyjny na wszystkich podanych konferencjach jest niezwykle wymagający. Na przykład na konferencji NeuRIPS przyjęcie artykułu wymaga czterech pozytywnych opinii niezależnych recenzentów, którzy publikują swoje wypowiedzi w powszechnie dostępnej platformie OpenReview.

Każdy artykuł z osobna został pozytywnie zweryfikowany przez międzynarodowe grono wybitnych ekspertów. Ponadto, jak wykazałem w poprzedniej części recenzji, zestaw omawianych pub-

likacji stanowi wynik zwartego i logicznie umotywowanego planu badawczego. Merytoryczny poziom pracy jest więc z całą pewnością adekwatny do wymagań stawianych rozprawom doktorskim.

4 Czy Rozprawa wykazuje umiejętność samodzielnego prowadzenia pracy naukowej?

Wszystkie omawiane artykuły zostały napisane w licznym gronie autorów – odpowiednio: pięciu, siedmiu, siedmiu i siedmiu. Ten stan rzeczy może budzić zaniepokojenie recenzenta. Prace badawcze z dziedziny sztucznej inteligencji wymagają współpracy osób z różnych środowisk, co jednak nie w pełni uzasadnia aż tak liczny udział autorów. Szczególnie taka sytuacja jest trudna do oceny, gdy prace zbiorowe mają stanowić podstawę przyznania stopnia naukowego. W skrajnym przypadku można by sobie wyobrazić, że ten sam zestaw kilku publikacji mógłby zostać uznany jako dorobek dla wielu rozpraw doktorskich.

W mojej opinii, jeśli rozprawa składa się wyłącznie z artykułów zbiorowych, to Doktorant powinien wykazać umiejętność pracy samodzielnej w inny sposób, na przykład omawiając realizację indywidualnego planu badawczego.

Doktorant wybrał inny sposób wykazania samodzielności – omówił „swoimi słowami" zawartość artykułów. W mojej opinii nie jest to trafiona koncepcja. Omówienie nie wnosi nowych treści w stosunku do artykułów, które stanowią integralną część Rozprawy.

W celu znalezienia odpowiedzi na pytanie zawarte w tej części recenzji, zasięgnąłem opinii współautorów artykułów. Ich jednoznacznie pozytywna opinia na temat umiejętności samodzielnego prowadzenia pracy naukowej przez Doktoranta oraz fakt pierwszeństwa na liście autorów dwóch publikacji przekonuje mnie do udzielenia opinii pozytywnej.

5 Czy prezentowane zastosowania są efektem oryginalnych i własnych rozwiązań Doktoranta?

W ramach Rozprawy przedstawione są oświadczenia wszystkich współautorów recenzowanych prac o ich wkładzie badawczym. W oświadczeniach tych niektóre zadania zostały przypisane do kilku autorów, przez co nie wskazują jednoznacznie roli Doktoranta. Na przykład w pierwszym artykule wkładem Doktoranta było m.in.:

- „conceptualization and methodology" – to samo zadanie było wykonywane przez dwóch innych autorów;
- „annotation of datasets" – to samo zadanie wykonywane było przez trzech innych autorów;
- „results analysis" – dwóch innych współautorów;
- „writing the paper" – dwóch innych współautorów.

Taki nakładający się podział zadań przyjęto dla wszystkich artykułów wchodzących w skład Rozprawy. Fakt ten utrudnia udzielenie odpowiedzi na postawione pytanie. W mojej opinii autor Rozprawy powinien zadbać, by jego osobisty wkład w badania był jednoznacznie i wyłącznie określony – albo w oświadczeniach współautorów, albo w indywidualnej części Rozprawy.

W zaistniałej sytuacji zwróciłem się drogą mailową do autora Rozprawy o jednoznaczne określenie oryginalnego wkładu w publikacji. Uzyskane wyjaśnienia pozwalają mi pozytywnie odpowiedzieć na postawione pytanie.

6 Formalna strona Rozprawy

Praca przedstawiona jest w bardzo czytelnym układzie graficznym. Omówienie każdego artykułu poprzedzone jest wprowadzeniem, wskazującym motywację badań. Całość Rozprawy zainicjowana jest krótkim przedstawieniem problemu badawczego oraz wdrożeniowego celu Rozprawy.

W indywidualnej części Rozprawy jest kilka drobnych uchybień językowych, np.

- „...zaproponowanie własnego mechanizmy...", str. 12;
- „Poziom skompilowania zadania...", str. 20; Autorowi zapewne chodziło o „skomplikowanie".

Nie jestem też zwolennikiem stosowania czasownika "wstrzykiwać informację", co Doktorant czyni kilkakrotnie (jest to zapewne kalka z języka angielskiego). Ja bym proponował czasowniki: „wprowadzać", „dodawać", lub „uzupełniać".

Na stronie 34. autor przedstawił listę angielskich terminów specjalistycznych i ich przyjętych przez siebie tłumaczeń na język polski. W słowniku znalazł się jeden błąd niespójności gramatycznej („trenować w trybie nienadzorowanym – ang. supervised training"). Ponadto sugerowałbym przyjęcie nieco innych tłumaczeń dwóch terminów:

- Termin „word embedding" lub „words embedding"proponowalbym tłumaczyć jako „wektor słowa", a nie „wektor słów"; tłumaczenie proponowane przez autora sugeruje, że każdym elementem wektora jest reprezentacja jednego słowa.
- Termin „attention head" sugerowałbym tłumaczyć jako „centrum atencji", a nie dosłownie: „głowa atencji".

Powyżej poczynione uwagi nie zmieniają mojej pozytywnej oceny strony formalnej pracy.

7 Podsumowanie recenzji

Zestaw publikacji poddany recenzji stanowi niewątpliwie zbiór artykułów naukowych powiązanych tematycznie. Merytoryczny poziom pracy jest z całą pewnością adekwatny do wymagań stawianych rozprawom doktorskim. Rozprawa wykazuje umiejętność samodzielnego prowadzenia pracy

naukowej w stopniu wystarczającym. Prezentowane zastosowania są w wystarczającym stopniu efektem oryginalnych i własnych rozwiązań Doktoranta.

Uważam, że Rozprawa **spełnia** wymagania ustawowe stawiane pracom doktorskim.

Podpisano przez /Signed by:
KRZYSZTOF JASSEM
Data /Date: 2022.02.02 12:54
mSzafir