RADA NAUKOWA DYSCYPLINY
INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ
zaprasza na
PUBLICZNĄ OBRONĘ ROZPRAWY DOKTORSKIEJ
mgr. inż. Tomasza Stanisławka

która odbędzie się w dniu **4 kwietnia 2022 roku o godzinie 11 :00** w trybie hybrydowym *

Temat rozprawy doktorskiej:

,,Ekstrakcja informacji z dokumentów o bogatej strukturze graficznej ”

Promotor: dr hab. inż. Przemysław Biecek - Politechnika Warszawska

Recenzenci: prof. dr hab. inż. Andrzej Czyżewski – Politechnika Gdańska

prof. dr hab. inż. Krzysztof Jassem – Uniwersytet Adama Mickiewicza w Poznaniu

dr hab. inż. Agnieszka Mykowiecka –Instytut Podstaw Informatyki PAN w Warszawie

*** Obrona odbędzie się w sali 40 w budynku Wydziału Matematyki i Nauk Informacyjnych Politechniki Warszawskiej. Możliwość udziału w trybie zdalnym dotyczy tylko recenzentów.** Kontakt do sekretarza komisji doktorskiej : dr hab. Maria Ganzha; m.ganzha@mini.pw.edu.pl

Z rozprawą doktorską i recenzjami można zapoznać się w Czytelni Biblioteki Głównej Politechniki Warszawskiej, Warszawa, Plac Politechniki 1.

Streszczenie rozprawy doktorskiej i recenzje są zamieszczone na stronie internetowej: https://bip.pw.edu.pl/Postepowania-w-sprawie-nadania-stopnia-naukowego/Doktoraty/Wszczete-po-30-kwietnia-2019-r/Dyscyplina-informatyka-techniczna-i-telekomunikacja-dziedzina-nauk-inzynieryjno-technicznych/mgr-inz.-Tomasz-Stanislawek.

Przewodniczący Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej

dr hab. inż. Jarosław Arabas

Streszczenie rozprawy

Tomasz Stanisławek

Bardzo szybki rozwój dziedziny przetwarzania języka naturalnego (ang. *Natural Language Processing*), a w szczególności pojawienie się nowych modeli języka (BERT, RoBERTa, T5, GPT-3) spowodował gwałtowny wzrost skuteczności w rozwiązywaniu standardowych problemów. Wpłynęło to również znacząco na jakość wyników w tematyce ekstrakcji informacji ze zwykłego tekstu. Przykładowo, dla zadania wykrywania jednostek nazewniczych (ang. *Named Entity Recognition, NER*) w samym tylko 2018 roku udało się osiągnąć przyrost o 1.88 pp miary F-1 dla zbioru CoNLL 2003 (wcześniej na taki przyrost trzeba było czekać 11 lat). Te sukcesy spopularyzowały użycie technik ekstrakcji informacji w celu automatyzacji procesów biznesowych, gdzie większość dokumentów posiada bogatą strukturę graficzną. Celem rozprawy doktorskiej było zbadanie możliwości istniejących metod wykorzystywanych do ekstrakcji informacji z dokumentów o bogatej strukturze graficznej, konceptualizacja problemów, jakie występują w tej dziedzinie, oraz zaproponowanie własnego mechanizmu, który poprawia jakość dotychczasowych rozwiązań. Efektem końcowym rozprawy doktorskiej było utworzenie nowego modelu LAMBERT, który dzięki wstrzyknięciu informacji o pozycji tokenów na stronie osiąga znacząco lepsze wyniki na trzech zbiorach domenowych: Kleister NDA, Charity oraz SROIE.

Prof. dr hab. inż. Andrzej Czyżewski
Politechnika Gdańska, Wydział ETI
Katedra Systemów Multimedialnych

19. 01. 2022 r.

Opinia o rozprawie doktorskiej mgr inż. **Tomasza Stanisławka**
pt.: „Ekstrakcja informacji z dokumentów o bogatej strukturze graficznej"
przygotowanej w Pol. Warszawskiej pod kier. dr hab. inż. Przemysława Biecka, prof. uczelni

Rozprawa składa się z części wprowadzającej o objętości ok. 25 stron tekstu wraz z pięcioma rysunkami, siedmioma tablicami oraz z wykazu literatury obejmującego 58 pozycji. Tę część uzupełnia wykaz dorobku naukowego, związanego z tematem rozprawy, który obejmuje cztery współautorskie publikacje konferencyjne. W przypadku dwóch z nich, mgr inż. Tomasz Stanisławek jest pierwszym autorem, w przypadku dwóch pozostałych, trzecim współautorem. Autor rozprawy wspomina na zakończenie powyżej wspomnianego podsumowania także o sześciu innych publikacjach. Nie są one bezpośrednio związane z tematem rozprawy, ale wchodzą w skład jego ogólnego dorobku publikacyjnego.

W odniesieniu do publikacji wchodzących w skład rozprawy, ich łączna punktacja według listy MEiN wynosi 620 (3 x 140 +200), mają one jednak we wszystskich przypadkach licznych współautorów (w liczbie od 5 do 8), stąd też niezbędne okazało się załączenie oświadczeń współautorów, w których zgodnie określają oni rolę mgr inż. Tomasza Stanisławka jako autora wiodącego dwóch publikacji, których jest on pierwszym autorem (rola pomysłodawcy, autora metody, implementacji, głównego eksperymentatora). W trzeciej wymienionej publikacji jego rola określana jest w oświadczeniach jako m. in. uczestnika dyskusji, osoby pracującej nad zbiorami danych i benchmarkami. Jako współautor czwartej, ostatniej, z wchodzących w skład rozprawy publikacji, mgr inż. T. Stanisławek ponownie wymieniany jest jako inicjator głównej idei, współautor koncepcji i metodyki, eskperymentator. Tym niemniej, dorobek wynikający z opublikowania wspomnianych czterech referatów, według załączonych oświadczeń, w znacznym stopniu dzieli się pomiędzy licznych współautorów. Pozostałą część rozprawy stanowią kopie wspomnianych czterech publikacji konferencyjnych z lat 2019-2021.
Dalsza część mojej opinii zredagowana jest w formie odpowiedzi na pytania stosowane zwyczajowo w toku oceny rozpraw naukowych.

1. Jaki problem naukowy (teza) został rozwiązany i przedstawiony w rozprawie?

Praca dotyczy dziedziny przetwarzania języka naturalnego. Jest to jeden z centralnych problemów współczesnej informatyki technicznej, którego rozwiązywanie w ogólności powinno umożliwić postęp w automatycznym przetwarzaniu dokumentów. Autor podkreśla, że typowe dokumenty składają się nie tylko z warstwy tekstowej, ale także z włączonych w nią treści innej natury, np. obrazów. W związku z powyższym, zagadnienie to komplikuje się oraz wypada zauważyć, że przynosi w pewien sposób

ograniczone korzyści praktyczne, zwłaszcza gdy podchodzi się do niego bez użycia wiedzy związanej z automatyzacją semantycznego opisu obrazów. W tym kontekście można zauważyć dążenie przez Doktoranta do wyabstrahowania stosunkowo wąskiego podejścia do ekstrakcji informacji z dokumentów o bogatej formie, gdyż w pracy nie wspomina się na przykład o dynamicznie rozwijających się technikach automatycznej semantycznej adnotacji obrazów .

W rozprawie nie sformułowanowo w sposób wyraźny tez, które byłyby formalnie udowadniane. Wprawdzie obecnie coraz częściej spotyka się rozprawy odbiegające od klasycznego stylu hipoteza-teza-dowód, ale taki sposób podejścia do konstruowania rozprawy doktorskiej ogranicza możliwości wnikliwej oceny składających się na nią osiągnięć. Z kolei zaletą wąskiego spojrzenia na zagadnienie naukowe jest spójny i dość przejrzysty charakter narracji, co ma zastosowanie do opracowanego przez Doktoranta przewodnika po publikacjach. Aktualna wartość inżynierska udokumentowanych prac jest znaczna, aczkolwiek z uwagi na szybką ewolucję algorytmów i zbiorów tekstów, wkład naukowy, mający głównie charakter tymczasowych przyczynków, może nie okazać się trwały. Nie zmienia to oceny, że przedstawione koncepcje, potwierdzone eksperymentalnie odzwierciedlają podejście do zagadnienia nie tylko inżynierskie, ale również naukowe.

2. W jaki sposób doktorant rozwiązał problem, jakich użył metod i jakich to wymagało umiejętności?

Z oświadczeń licznych współautorów referatów przedstawionych na czterech konferencjach wynika, że Doktotant inicjował, inspirował te referaty, bądź konsultował ich treści w toku dyskusji, a ponadto wykazał się m. in. kompetencjami z zakresu gromadzenia i obróbki danych testowych, implementacji oprogramowania do przetwarzania tego typu danych. Umiejętności tych nie sposób precyzyjnie ocenić, ani też zakwestionować na podstawie dokumentacji ograniczonej do stosunkowo krótkich treści wieloauorskich referatów, których objętości zawierają sie w przedziale 10-15 stron wydruku przy współautorstwie 5-8 osób.

3. Na czym polega oryginalny dorobek autora i jakie jest jego znaczenie poznawcze lub przydatność praktyczna dla nauki bądź techniki?

Referat konferencyjny, zatytułowany "Named Entity Recognition--Is there a glass ceiling?" prezentuje szczegółową analizę typów błędów występujących w nowoczesnych metodach uczenia maszynowego. Pokazano w nim słabe i mocne strony modeli: Stanford, CMU, FLAIR, ELMO i BERT, a także ich wspólne ograniczenia. Zaproponowano nowe techniki poprawy adnotacji, procesów treningowych oraz sprawdzania jakości i stabilności modelu. Prezentowane wyniki oparte są na zbiorze danych CoNLL 2003 dla języka angielskiego. Rezultatem tego jest nowa, wzbogacona semantyczna adnotacja błędów dla tego zbioru danych oraz nowe zestawy danych diagnostycznych.

Kolejny referat konferencyjny zatytułowany "Kleister: Key Information Extraction Datasets Involving Long Documents with Complex Layouts" podkreśla znaczenie zadania ekstrakcji kluczowych informacji w procesie przetwarzania języka naturalnego i prezentuje dwa nowe zbiory danych nazwane: Kleister NDA i Kleister Charity. Obejmują one kolekcję obszernych formalnych dokumentów w języku angielskim. W tych zbiorach danych system przetwarzania jezyka naturalnego ma za zadanie znaleźć lub wnioskować o różnych typach podmiotów, wykorzystując zarówno tekstowe, jak i strukturalne cechy dokumentów. Autor wraz ze współpracownikami przetestował kilka systemów bazowych z dziedziny ekstrakcji informacji kluczowych (Flair, BERT, RoBERTa, LayoutLM, LAMBERT), co pokazało, że opracowane zbiory danych są stosunkowo trudne do analizy przy użyciu znanych modeli. Opracowane zbiory dokumentów zostały udostępnione środowisku naukowemu.

Trzeci referat, zatytułowany "DUE: End-to-End Document Understanding Benchmark" dotyczy rozumienia dokumentów o na tyle bogatych układach, że pozostaje ono trudnym zadaniem w opinii środowisk badawczych, które zajmują sie przetwarzaniem języka naturalnego. Kwantyfikację postępu w tej dziedzinie utrudnia brak powszechnie akceptowanego benchmarku, stąd motywacja do wprowadzenia opisanego w pracy benchmarku pod nazwą Document Understanding Evaluation (DUE), składającego się z dostępnych i przeformułowanych zbiorów danych, pozwalających porównywać możliwości systemów w warunkach działania na rzeczywistych dokumentach. Zproponowany benchmark obejmuje zadania ekstrakcji kluczowych informacji oraz maszynowego czytania ze zrozumieniem dokumentów tekstowych, wzbogaconych o liczne elementy dodatkowe, inne niż tekst. Referat w sposób systematyczny prezentuje i porównuje aktualnie dostępne bazy danych i najnowsze osiągnięcia w modelowaniu języka z uwzględnieniem układu graficznego. Benchmarki oraz implementacje referencyjne zostały udostępnione do wykorzystania przez środowisko naukowe.

Ostatni za zaprezentowanych referatów nosi tytuł "LAMBERT: Layout-Aware Language Modeling for Information Extraction". Przedstawiono w nim nowe, proste podejście do problemu rozumienia dokumentów, w których nietrywialny układ wpływa na lokalną semantykę. Model został oceniony w zadaniu ekstrakcji informacji z wykorzystaniem publicznie dostępnych zbiorów danych. Wykazano, że opracowany model osiąga wyższą wydajność na zbiorach danych składających się z dokumentów bogatych wizualnie, a także przewyższa bazowy model RoBERTa na dokumentach o płaskim układzie. Rozwiązanie to zajęło pierwsze miejsce w rankingu publicznym dla zadania Ekstrakcja kluczowych informacji z zestawu danych SROIE, a ponadto zostało wyróżnione jako Best Industry Related Paper Award na konferencji ICDAR 2021 w Lozannie.

4. Jaka jest szansa dalszego wykorzystania wyników rozprawy?

Potrzeba rozwiązywania problemów wynikających z ekstrakcji dokumentów o bogatej strukturze graficznej (ang. Visually Rich Documents, VRDs) oraz opracowywania rozwiązań uwzględniających aspekty struktury dokumentu mają znaczenie dla rozwoju technologicznego podmiotów, w których obieg informacji odbywa się poprzez różnego rodzaju dokumenty elektroniczne. Bezpośrednio wykorzystane przez społeczność naukową, zajmującą się przetwarzaniem języka naturalnego mogą być opracowane i udostępnione publicznie zbiory tekstów. Na platformie hostingowej oprogramowania Github Doktorant i współpracownicy umieścili implementacje referencyjne. Szansę wykorzystania ma ponadto opracowany benchmark pod nazwą Document Understanding Evaluation (DUE).

5. Jakiej wiedzy, umiejętności oraz kompetencji i na jakim poziomie nabył doktorant w wyniku realizacji rozprawy?

Analiza przewodnika po publikacjach i treści załączonych publikacji pozwala zauważyć, że Autor posiadł umiejętności z zakresu badań literaturowych, tworzenia i usprawniania algorytmów, programowania komputerowego, budowy i adnotowania zbiorów danych i tekstów wykorzystywanych do analizy algorytmicznej.

6. Czy rozprawa obejmuje najnowsze osiągnięcia nauki i świadczy o znajomości współczesnej literatury z dyscypliny naukowej, której dotyczy?

Rozprawa jest ściśle powiązana z nową i z najnowszą literaturą dotyczącą zagadnień przetwarzania języka naturalnego. Autor wykazał się orientacją na temat nowoczesnych algorytmów i zbiorów testowych. W publikacjach udokumentowane są eksperymenty porównawcze, oparte na zastosowaniu najnowszych algorytmów ekstrakcji informacji tekstowych z dokumentów, w tym dokumentów o bogatej strukturze.

7. Czy doktorant w trakcie pracy nad rozprawą wykazał się kompetencjami społecznymi niezbędnymi do prowadzenia pracy naukowo-badawczej?

Zespołowy charakter publikacji w pewnym stopniu utrudnia śledzenie indydwidualnego wkładu Doktoranta, jednak należy uznać, że współczesne zaawansowane projekty informatyczne w częstych przypadkach wymagają pracy zespołowej, do której prowadzenia konieczne jest posiadanie przez członków zespołu określonych kompetencji społecznych. Biorąc udział w zespołowej pracy eksperymentalnej i w opracowaniu wieloautorskich publikacji, Doktorant w trakcie pracy nad rozprawą wykazał się kompetencjami społecznymi niezbędnymi do prowadzenia pracy naukowo-badawczej.

Wniosek

Rozprawa pana mgr inż. Tomasza Stanisławka została zrealizowana w sposób odzwierciedlający wymagane kwalifikacje jej Autora, wystarczający nakład pracy badawczej, implementacyjnej i eksperymentalnej, jak również jego potwierdzony oświadczeniami współautorów referatów udział w opublikowaniu wyników na wysokopunktowanych konferencjach.

W mojej opinii treść rozprawy mgr inż. Tomasza Stanisławka spełnia zatem wymogi Prawa o Szkolnictwie Wyższym i Nauce, z dnia 20 lipca 2018 r. (Dz. U. 30. 08. 2018 r. Poz. 1668), stawiane kandydatom do stopnia naukowego doktora.

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.

Warszawa, 1.02.2022

dr hab. Agnieszka Mykowiecka
Instytut Podstaw Informatyki PAN
Jana Kazimierza 5, Warszawa

Recenzja rozprawy doktorskiej mgr. inż. Tomasza Stanisławka

Ekstrakcja informacji z dokumentów o bogatej strukturze graficznej

Recenzja wykonana jest na zlecenie Rady Naukowej dyscypliny informatyka techniczna i telekomunikacja Politechniki Warszawskiej. Praca doktorska Tomasza Stanisławka zrealizowana została pod opieką dr hab. inż. Przemysława Biecka na Wydziale Matematyki i Nauk Informacyjnych i składa się ze zbioru czterech jednotematycznych artykułów w języku angielskim opublikowanych w recenzowanych materiałach z konferencji o zasięgu międzynarodowym. Wszystkie konferencje należą do najbardziej cenionych w zakresie analizy dokumentów lub ogólniej neuronowych metod rozwiązywania problemów: ICDAR (A), NeurIPS (kategoria A*), oraz CONLL (A). Tematem pracy jest wydobywanie informacji z dokumentów, które nie składają się jedynie z ciągłego tekstu, ale mają złożoną strukturę graficzną. Prace prowadzone były w trakcie studiów doktoranckich, ale miały także bezpośrednie powiązanie z możliwością wdrożenia modułu stanowiącego rozwiązania badanego problemu w firmie Applica i podniesieniem jakości proponowanych przez nią narzędzi.

Rozprawa ma formę dokumentu, który opisuje cel i wyniki prowadzonych badań, pełni zatem rolę autoreferatu. Załącznikami są artykuły stanowiące główną treść rozprawy. Rozdział pierwszy wprowadza czytelnika do tematyki rozprawy, przedstawia opis problemu, motywację podjęcia badań oraz cel pracy. Rozdział drugi zawiera prezentację głównych wyników rozprawy i jest podzielony na cztery podrozdziały, z których każdy odnosi się do jednego z artykułów stanowiących treść rozprawy. Rozdział trzeci obejmuje prezentację dorobku naukowego doktoranta, a rozdział czwarty zawiera podsumowanie wyników i wkładu doktoranta w rozwój dziedziny.

Artykuły stanowiące główną treść pracy, zawarte w załącznikach, to:

1. Stanislawek, T., A. Wróblewska, A. Wójcicka, D. Ziembicki i P. Biecek: **Named Entity Recognition - Is There a Glass Ceiling?** Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), 2019
2. Stanisławek, T., F. Graliński, A. Wróblewska, D. Lipiński, A. Kaliska, P. Rosalska, B. Topolski i P. Biecek, **Kleister: Key Information Extraction Datasets Involving Long Documents with Complex Layouts**, In International Conference on Document Analysis and Recognition (ICDAR) 2021, pp. 564-579
3. Borchmann, Ł., M. Pietruszka, T. Stanislawek, D. Jurkiewicz, M. Turski, K. Szyndler i F. Graliński: **DUE: End-to-End Document Understanding Benchmark**, Thirty-fifth Conference on Neural Information Processing Systems Datasets, 2021
4. Garncarek, Ł., R. Powalski, T. Stanisławek, B. Topolski, P. Halama, M. Turski, ... **LAMBERT: Layout-Aware Language Modeling for Information Extraction** International Conference on Document Analysis and Recognition, ICDAR 2021, pp. 532-547 (wyróżnienie w kategorii *Best Industry Related Paper Award*)

Sformułowanym w rozprawie celem pracy było zaproponowanie własnego mechanizmu poprawiającego skuteczność ekstrakcji informacji z dokumentów o bogatej strukturze graficznej. Realizacja tego zadania wymagała realizacji także celów pomocniczych takich jak

zbadanie możliwości istniejących metod oraz konceptualizacja problemów, jakie występują w tej dziedzinie.

Treść pracy

Artykuł [1], opisany w podrozdziale 2.1, przedstawia badania, w których poddano ocenie istniejące metody rozpoznawania nazw własnych (NER) w tekstach. Temat NER był bardzo często podejmowany, opracowane metody są często stosowane także do ekstrakcji informacji, a wyniki są łatwe do oceny ze względu na istniejące zbiory tekstów z oznaczeniami typów nazw i stosunkowo proste reguły anotacji. Wnioski wyciągnięte na podstawie analizy tych rozwiązań mogą być zatem bardzo przydatne także przy rozwiązywaniu zadania ekstrakcji informacji.

W pracy zanalizowano rozwiązania wykorzystujące różne, udostępnione publicznie, metody identyfikacji nazw własnych i ich typów, począwszy od najskuteczniejszej niegdyś metody CRF (rozwiązanie zaproponowane przez Stanford) poprzez sieci LSTM z warstwą CRF (CMU), sieci BiLSTM w modelu ELMO, do modelu języka typu BERT opartego na transformerach oraz modelu Flair. Wszystkie modele zostały przez autorów użyte do oznaczenia danych angielskich upowszechnionych na konferencji CoNLL w 2003 roku. Uzyskane wyniki zostały zanalizowane, ustalono klasyfikację popełnianych przez systemy błędów, a następnie przydzielono wszystkie wykryte błędy do odpowiednich kategorii. W wynikach analizy widać, że duża liczba błędów wynika z tego, że do prawidłowego przypisania typu nazwy potrzebna jest wiedza nie tylko z bezpośredniego kontekstu, ale z całego zdania lub całego dokumentu. Dlatego najwięcej błędów stwierdzono w wynikach systemu wykorzystującego tylko CRF, a wyniki ELMO okazały się lepsze niż z BERTa. Naturalnym wnioskiem jest zatem konieczność uzupełniania danych wejściowych do modelu NER o informacje dotyczące miejsca konkretnego fragmentu w całym dokumencie. Istotnym elementem prac było też zwrócenie uwagi na błędy anotacji w opublikowanych danych testowych. W przypadku niewielkich różnic w efektywności porównywanych systemów błędy te mogą zaburzyć ocenę. Przeprowadzone eksperymenty wykazały, że wyniki na poprawionych zbiorach testowych nie różniły się zbytnio od tych uzyskanych na danych z błędami (były nieco lepsze), ale kolejność wyników poszczególnych modeli czasami ulegała zmianie. Widać zatem, że jakość danych testowych jest istotna i na pewno tym większa im mniejszym zbiorem testowym dysponujemy.

Artykuł [2] dotyczy kolejnego kroku koniecznego przy opracowywaniu metod maszynowego uczenia – zebrania odpowiednich danych treningowych i testowych. Z uwagi na bardzo niewielką dostępność danych zawierających obrazy całych sformatowanych dokumentów, a nie tylko sam tekst, przygotowano dwa takie zbiory danych angielskich. Pierwszy zawiera umowy o zachowaniu poufności, które pochodzą z bazy EDGAR. Drugi, sprawozdania roczne fundacji charytatywnych (https://register-of-charities.charitycommission.gov.uk/). Ponieważ zbiory te miały przypisane informacje na poziomie dokumentów opracowano zestaw reguł – wyrażeń regularnych – dzięki którym przypisanie to zostało przeniesione na poziom fragmentów tekstu. Poza stworzeniem zbioru danych, w artykule przedstawiono też wyniki osiągnięte dla tych danych przez różne modele wytrenowane do zadania ekstrakcji informacji (FLAIR, BERT, RoBerta, LayoutLM i Lambert. Przetestowano także procent zgodności anotacji dwóch osób na próbce 100 dokumentów, który okazał się bardzo wysoki (ponad 97%). Zgodnie z oczekiwaniami modele, które uwzględniają jakiś sposób pozycję w tekście radziły sobie lepiej z wyznaczonym zadaniem. Wciąż jednak najlepsze osiągane wyniki były znacznie poniżej zgodności anotatorów (ok.85%).

W artykule [3] zaprezentowany został benchmark opracowany z myślą o wszystkich zadaniach związanych z przetwarzaniem dokumentów o bogatej strukturze graficznej. Autorzy przejrzeli ponad 30 zbiorów danych, z których wybrano 7 spełniających kryteria najwyższej jakości (tylko anotacja manualna), trudności (duża różnica między najlepszym rozwiązaniem a poziomem trudności dla człowieka) oraz dostępności. Uwzględnione zbiory odpowiadają różnorodnym zadaniom od ekstrakcji informacji do zadawania pytań do treści znajdujących się w tabelach czy infografikach.

W ostatnim artykule [4] zaproponowano opracowaną przy dużym udziale doktoranta architekturę modelu do ekstrakcji informacji uwzględniającą strukturę dokumentu. Architektura ta oparta jest na strukturze bardzo popularnego modelu neuronowego BERT uzupełnionej o dodatkowe wejście opisujące położenie segmentów na stronie. Zmodyfikowano wagi atencji poprzez dodanie używanego w modelu T5 relatywnego kodowania pozycji oraz rozszerzono to kodowanie o dwa dodatkowe parametry, związane z relatywną pozycją dwóch segmentów względem odległości od siebie w poziomie oraz w pionie. Przeprowadzona seria eksperymentów wykazała, że model ten osiągnął wyniki lepsze (w granicach kilku procent) niż model bazowy RoBERTa oraz model LayoutLM . Zgodnie z często obserwowanymi zależnościami, im większe i lepsze dane (odfiltrowane z niskiej jakości dokumentów) oraz dłuższy trening, tym model LAMBERT jest skuteczniejszy.

Ocena

Przedstawiony cykl artykułów dobrze opisuje typową drogę, jaką należy przebyć próbując opracować nowe rozwiązanie jednego z zadań NLP. Krok pierwszy to analiza już znanych rozwiązań tego, lub pokrewnych zadań, następnie wybór bądź konstrukcja danych treningowych i testowych, a potem próba znalezienia rozwiązania, które pozwoli na zmniejszenie liczby błędów w sytuacjach, z którymi dotychczas istniejące systemy sobie nie radziły. Osiągniecie dobrego wyniku związane musi być na ogół z identyfikacją powodów dla których dotychczasowe rezultaty nie są wystarczające. Doktorant, wspólnie ze współautorami, zrealizował wszystkie te etapy dochodząc do rozwiązania osiągającego bardzo wysoką skuteczność. W przedstawionej rozprawie cenne jest, że przeprowadzono w tym przypadku dokładną analizę i klasyfikację popełnianych błędów, a nie tylko porównywanie ogólnych wyników takich jak miara F1 dla poszczególnych kategorii. Doktorant, wraz ze współpracownikami, poświęcił też sporo uwagi zebraniu i ujednoliceniu wielu zbiorów treningowych dla zbliżonych do rozwiązywanego w pracy zadań. A przy szybko rozwijającej się dziedzinie NLP to właśnie zbiory danych stanowią często trwalszy wkład w jej rozwój, niż szybko zmieniające się modele neuronowe. Trochę szkoda, że autor nie pokusił się o zebranie choć jednego takiego zbioru z danymi dla języka polskiego.

Konstrukcja doktoratu ze zbioru artykułów zwykle stanowi pewne wyzwanie. Przedstawienie prac, które maja wielu autorów zawsze budzi pewne wątpliwości co do tego, na ile Doktorant uczestniczył w prezentowanych badaniach i jakie idee pochodzą od niego, a jakie od współautorów. Uważam jednak, że odnajdywanie się w grupie badawczej osiągającej dobre wyniki i potwierdzony oświadczeniami wkład Doktoranta w prace wystarczająco dowodzą jego osobistych umiejętności i osiągnięć. Stworzony przez Doktoranta opis towarzyszący artykułom stanowi dobry przewodnik po wykonanych pracach i we właściwy sposób podkreśla wyznaczony cel rozprawy. Tekst jest dość dobrze napisany, drobne literówki czy pomyłki są nieliczne. Jego niedostatkiem, wynikającym jednak z obranej formy rozprawy, jest to, że zawiera on mniej informacji niż załączone artykuły, a chciałoby się by zawierał ich więcej, tak by przeprowadzane eksperymenty, konwersje danych, alternatywne rozwiązania opisane były dokładniej niż można to zrobić w krótkim artykule konferencyjnym. Odnosząc

się zaś do samej rozprawy w wybranej formie – nie do końca przekonuje mnie uporządkowanie artykułów – umieszczenie artykułu dotyczącego modelu LAMBERT po zawierających wyniki testów tego modelu. Zapewne jest to wynik prowadzenia równolegle badań nad samym modelem i próbami jego weryfikacji, a zatem z trudnością w uszeregowaniu prac (trzy z nich opublikowane zostały w roku 2021).

Temat ekstrakcji informacji z danych sformatowanych jest ważny, gdyż bardzo często mamy do czynienia z dokumentami, w których istnieje co najmniej narzucona struktura poszczególnych części, a nawet są one wprost formularzami. Operowanie tylko na czystym tekście powoduje zatem brak możliwości odwoływania się do informacji, która może znacząco wpłynąć na jakość wyniku. Doktorant przedstawił w rozprawie zarówno analizy wskazujące na potrzebę dodawania informacji o pozycji w tekście jak i model osiągający dzięki jej uwzględnieniu lepsze wyniki. Zaakceptowanie proponowanych rozwiązań na wiodących konferencjach z tej dziedziny i nagroda dla jednej z publikacji świadczą o tym, że zostały one uznane za wartościowe przez społeczność międzynarodową.

Zbiór paru odrębnych artykułów oczywiście ma drobne wady związane ze spójnością. Jeśli patrzymy na niego jako na całość, to zauważamy, że w artykule [3] używanymi modelami są T5 i T5+2D i żadnym za artykułów nie ma komentarza o ewentualnych porównaniach (być może nie wprost) zaproponowanego modelu LAMBERT i T5+2D. Jeśli zaś popatrzymy na dane, to trochę brak dyskusji na temat znaczenia pozycji etykietowanych segmentów w stosunku nie tyle do strony co do innych tekstów, takich jak treść stałych elementów formularzy, zwłaszcza w przypadku gdy wielkość poszczególnych pól jest zmienna, a zatem informacje te nie zawsze znajdują się w tym samym miejscu dokumentu. W szczególności dotyczy to sytuacji, w których zbiór typów obiektów, które chcemy wykrywać w dokumentach jest liczniejszy i bardziej różnorodny. Być może gorsze wyniki uzyskiwane dla analizy tekstów długich są z tym jakoś powiązane. W pracy brak sugestii jakie mogą być przyczyny tego faktu, a jego zbadanie byłoby na pewno ciekawe. Oczywiście te ogólne uwagi nie umniejszają mojej pozytywnej oceny opisanych w pracy osiągnięć przy budowie modelu uwzględniającego pozycję analizowanej informacji w tekście, który w pewnym momencie był jednym z najlepszych proponowanych dla tego zadania na świecie.

Wniosek końcowy

Stwierdzam, iż przedłożona mi do recenzji rozprawa, której autorem jest mgr Tomasz Stanisławek, zawiera ważne osiągnięcia w dziedzinie konstruowania modeli neuronowych do rozwiązywania zadań NLP, w szczególności ekstrakcji informacji ze sformatowanych tekstów. Doktorant wykazał się sporą wiedzą w tematyce rozprawy oraz znajomością metod badawczych. Przedstawiony zestaw artykułów cechuje spójność tematyczna i wyraźnie nakreślony kierunek w dochodzeniu do realizacji wyznaczonego celu badawczego. Recenzowana praca spełnia wymagania ustawowo stawiane rozprawom doktorskim, zatem wnoszę o dopuszczenie magistra Tomasza Stanisławka do publicznej obrony.

Agnieszka Wykowska