Dr hab. inż. Wojciech Kotłowski, prof. PP
Instytut Informatyki
Politechnika Poznańska
ul. Piotrowo 2, 60-965 Poznań
tel: (+48) 61 665 2936
wkotlowski@cs.put.poznan.pl

Poznań, 12 listopada 2023 r.

Recenzja rozprawy doktorskiej

mgr. inż. Dawida Jurkiewicza

Novel Methods and Datasets for Intelligent Document Processing
(Nowe metody i zbiory danych do inteligentnego przetwarzania dokumentów)

1 Problem badawczy jego znaczenie

W dzisiejszym, zdominowanym przez informację, świecie kluczowe staje się automatyczne przetwarzanie dużej ilości tekstu. Dziedzina inteligentnego przetwarzania dokumentów (ang. *Intelligent Document Processing*), będąca tematyką rozprawy doktorskiej mgr. inż. Dawida Jurkiewicza, stanowi fundamentalny element w radzeniu sobie z tymi wyzwaniami. W ostatnim czasie duże modele językowe, takie jak GPT-4 czy Llama 2 pokazały, że budując probabilistyczne modele języka na dużych korpusach dokumentów można uzyskać inteligentne systemy o niezwykle szerokich możliwościach rozumienia i generacji języka naturalnego. Obecnie wydaje się, że przetwarzanie języka, a więc dokumentów, stanowi najbardziej obiecujący kierunek w rozwoju sztucznej inteligencji.

W szczególności, Doktorant podejmuje w rozprawie badania nad dwoma zagadnieniami dziedziny – identyfikacji relewantnych fragmentów tekstu (ang. *Span Identification*) oraz rozumienia dokumentów (ang. *Document Understanding*). Identyfikacja relewantnych fragmentów tekstu dotyczy umiejętności systemu do precyzyjnego identyfikowania i wyodrębniania kluczowych fragmentów dokumentu, zawierających poszukiwanie przez użytkownika informacje. Szersza problematyka rozumienia dokumentów obejmuje zdolność systemu do zrozumienia kontekstu, związków i znaczenia całego dokumentu, włączając w to informacje wizualne, co pozwala na wydobywanie istotnych informacji (np. odpowiedź na

pytania). Obie problematyki stały się niezwykle ważne w kontekście zastosowań współczesnych systemów sztucznej inteligencji, a ich znaczenie w nadchodzących latach będzie tylko wzrastać. Stąd temat pracy Doktoranta wydaje się świetnie trafiać we współczesne trendy badawcze i może prowadzić do istotnych praktycznych zastosowań oraz przyczynić się do ogólnego postępu nad tworzenie systemów inteligentnych.

Tematyka inteligentnego przetwarzania dokumentów (tak jak i dwa konkretne zagadnienia rozważane w pracy) nie jest oczywiście nowa, a w ostatnich latach rozwijana jest bardzo intensywnie, z ogromną liczbą prac badawczych publikowanych na konferencjach i w czasopismach dziedziny. Pomimo tego Doktorantowi udało się znaleźć szereg problemów, które nie zostały dotąd przebadane w wystarczającym stopniu. W szczególności, skupił się on nad wyzwaniami związanymi z małą ilością dostępnych danych, powszechnych w różnych zastosowaniach biznesowych i przemysłowych, kiedy to system musi nauczyć się rozwiązać zadanie, przed którym nie był do tej pory postawiony, mając jedynie niewielką liczbę (nie przekraczającą pięciu) przykładów rozwiązań tego zadania (ang. *few-shot learning*). W pracy rozważane są również problemy rozumienia dokumentów bogatych wizualnie, w których wzajemne umiejscowienie fragmentów tekstu, oraz ich otoczenie elementami graficznymi, stanowi kluczową informację w ich zrozumieniu. Mamy tu więc do czynienia z danymi multimodalnymi, zawierającymi zarówno informację obrazową, jak i tekstową, a model musi zrozumieć semantykę tekstu, ale również cechy wizualne i całościową strukturę dokumentu. Dodatkowo, Doktorant przyczynił się do rozwoju dziedziny, przedstawiając nowe zbiory danych, a nawet zestawy zbiorów danych dla identifikacji relewantnych fragmentów i rozumienia dokumentów z informacją wizualną, jak również proponując i nadzorując konkurs dla problematyki znajdywania odpowiedzi na pytania dla wielodomenowych, wielobranżowych i wielostronicowych dokumentów. Wymienione powyżej zadania są z pewnością wystarczająco ambitne, aby stały się tematyką rozprawy doktorskiej.

2 Cele badawcze

Mgr inż. Dawid Jurkiewicz postawił w rozprawie następujące cele badawcze:

1. Rozwiązanie nowego problemu uczenia się identyfikacji relewantnych fragmentów tekstu z małej liczby przykładów w dziedzinie dokumentów prawniczych.

2. Wprowadzenie nowego algorytmu dopasowania fragmentów długiej sekwencji tekstu do zbioru innych sekwencji, inspirowanego technikami *time warping* z dziedziny dopasowywania szeregów czasowych, na potrzeby identyfikacji relewantnych fragmentów tekstu z małej liczby przykładów.

3. Utworzenie systemu o wysokiej trafności klasyfikacji metod propagandy zawartych w dokumentach tekstowych, oraz identyfikacji fragmentów tekstu zawierających treści propagandowe, przewyższającego lub konkurującego z najlepszymi rozwiązaniami dziedziny.

4. Propozycja nowego, multimodalnego modelu rozumienia dokumentów opartego na architekturach *encoder-decoder* i *transformer*, umożliwiającego korzystanie z informacji o strukturze dokumentu, wzajemnym położeniu fragmentów tekstu, cechach wizualnych oraz semantyce tekstu.

5. Wprowadzenie dużej kolekcji zbiorów dokumentów, wielodziedzinowych, wielomodalnych, o wysokiej jakości, stanowiących benchmark pozwalający na mierzenie i śledzenie postępów w dziedzinie.

6. Zaproponowanie konkursu w dziedzinie rozumienia dokumentów z obszernym zbiorem danych dokumentów multimodalnych, wielostronicowych, pochodzących z wielu dziedzin, w celu zintesyfikowania rozwoju metod umożliwiających odpowiadanie na pytania dotyczące treści dokumentów.

W mojej opinii, powyższe cele są bardzo dobrze uzasadnione, interesujące, o dużym wymiarze praktycznym, a przede wszystkim nowatorskie i ambitne. Każdy z tych celów doprowadził do oryginalnych i istotnych wyników badawczych opublikowanych na prestiżowych konferencjach i w czołowych czasopismach dziedziny.

3 Struktura i zawartość pracy

Recenzowana rozprawa jest napisana czytelnie bardzo dobrym językiem angielskim. Liczy 139 stron i składa się z siedmiu rozdziałów wraz z dodatkami, poprzedzonych streszczeniem w języku angielskim i polskim. Ma charakter syntezy treści powiązanych ze sobą tematycznie artykułów naukowych, zbiorczo załączonych jako treść rozdziałów 2-7, natomiast pierwszy rozdział to wprowadzenie do tematyki pracy, określenie celów badawczych, streszczenie wyników, oraz podsumowanie. Doktorant starannie opisał i dobrze umotywował dziedzinę badawczą inteligentnego przetwarzania dokumentów, a także wprowadził czytelnika w rozważane problematyki identyfikacji relewantnych fragmentów tekstu oraz rozumienia dokumentów. Streścił również każdą z załączonych prac, równocześnie opisując wzajemne relacje między ich wynikami. Zawarł również wszystkie potrzebne dane bibliograficzne dotyczące publikacji. W dodatkowych rozdziałach na końcu pracy znajdziemy również informacje o uczestnictwie i wynikach w konkursach, o udziale w projektach badawczych, a także deklaracje potwierdzające wkład do prac każdego z współautorów. Ponieważ struktura pracy jest w zasadzie podporządkowana „publikacyjnemu" charakterowi rozprawy, trudno mi ją ocenić inaczej niż pozytywnie.

4 Ocena wkładu oryginalnego

Rozprawa jest oparta na sześciu artykułach, których współautorem jest Doktorant:

[1] Ł. Borchmann, D. Wiśniewski, A. Gretkowski, I. Kosmala, D. Jurkiewicz, Ł. Szałkiewicz, G. Pałka, K. Kaczmarek, A. Kaliska, F. Graliński. Contract Discovery: Dataset and

a Few-Shot Semantic Retrieval Challenge with Competitive Baselines. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4254–4268, 2020. Association for Computational Linguistics [Punkty MEiN: 140]

[2] Ł. Borchmann, D. Jurkiewicz, F. Graliński, T. Górecki. Dynamic Boundary Time Warping for subsequence matching with few examples. *Expert Systems with Applications*, 169:114344, 2021 [Punkty MEiN: 140]

[3] D. Jurkiewicz, Ł. Borchmann, I. Kosmala, F. Graliński. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, Span CLS and Whether Self-Training Helps Them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, COLING: Barcelona, 2020. International Committee for Computational Linguistics [Punkty MEiN: 140]

[4] R. Powalski, Ł. Borchmann, D. Jurkiewicz, T. Dwojak , M. Pietruszka, G. Pałka. Going Full- TILT Boogie on Document Understanding with Text-Image-Layout Transformer. In *Document Analysis and Recognition – ICDAR 2021*, pages 732–747, Cham, 2021. Springer International Publishing [Punkty MEiN: 140]

[5] Ł. Borchmann, M. Pietruszka, T. Stanisławek, D. Jurkiewicz, M. Turski, K. Szyndler, F. Graliński. DUE: End-to-End Document Understanding Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Currant, 2021 [Punkty MEiN: 200]

[6] J. Landeghem, R. Tito, Ł. Borchmann, M. Pietruszka, D. Jurkiewicz, R. Powalski, P Józiak, S. Biswas, M. Coustaty, T. Stanisławek. ICDAR 2023 Competition on Document UnderstanDing of Everything (DUDE). Accepted for *Document Analysis and Recognition – ICDAR 2023* [Punkty MEiN: 140]

W dalszej części recenzji będę się do prac odnosił używając powyższej numeracji.

Wszystkie prace zostały opublikowane na konferencjach lub czasopismach mających co najmniej 140 punktów MEiN. Na uwagę zasługuje praca [5], opublikowana na zapewne najbardziej prestiżowej konferencji uczenia maszynowego *Neural Information Processing Systems (NeurIPS)*, wycenianej na 200 punktów. Sumaryczna liczba punktów MEIN w powyższym dorobku jest więc równa 900, co jest wynikiem świetnym jak na dorobek w trakcie doktoratu. Jedyne na liście czasopismo *Expert Systems with Applications* posiada z kolei współczynnik *Impact Factor* na poziomie 8.665. Prace doczekały się już 126 cytowań (wg. *Google Scholar*, bez autocytowań, zebrane pod koniec czerwca), co jest osiągnięciem bardzo dobrym z uwagi na nieodległe daty publikacji. Dorobek uznaję za więc za bardzo wartościowy.

Jedyny element, mogący budzić wątpliwości, to duża liczba współautorów w załączonych pracach, przy czym z deklaracji udziału załączonych do rozprawy wynika, że nie we wszystkich pracach Doktorant odgrywał główną rolę. Ponieważ jednak liczba i jakość prac z nawiązką spełnia wymogi do finalizacji postępowania w sprawie nadania stopnia doktora, nie traktuje powyższej uwagi jako istotnej z punktu widzenia całościowej oceny prac.

Rozprawa zawiera wiele oryginalnych i nowatorskich wyników, które zostały już pozytywnie zweryfikowane i docenione przez recenzentów na etapie przyjmowania prac na konferencje do czasopism. Poniżej przedstawiam ich podsumowanie:

- Rozwiązanie problemu uczenia się identyfikacji relewantnych fragmentów tekstu z małej liczby przykładów w dziedzinie dokumentów prawniczych [1]: wprowadzono nowy zbiór danych 600 dokumentów z 2 500 adnotacji, wspólny sposób ewaluacji metod wraz z oceną jakości w postaci miękkiej miary F, rozważono wiele wariantów metod rozwiązania problemu opisanych za pomocą wspólnego potoku przetwarzania, przedstawiono ich wyniki wraz z obszerną dyskusją i zależnością jakości względem liczby przykładów zadania przedstawionych algorytmom.

- Wprowadzenie nowatorskiej metody dopasowania fragmentów długiej sekwencji tekstu do zbioru innych sekwencji, inspirowanej technikami *time warping*, ale stanowiącej istotną i twórczą modyfikację związaną z potrzebą dynamicznego wyboru granic dopasowanej sekwencji oraz wzięcia pod uwagę dopasowania do wielu przykładowych sekwencji na raz (metoda *Dynamic boundary time warping*) [2]. Wprowadzony algorytm cechuje się niską złożonością obliczeniową i wypada pozytywnie pod względem trafności identyfikacji fragmentów tekstu w porównaniu z rozwiązaniami bazowymi. Został przetestowany na zbiorze danych pochodzącym z poprzedniej publikacji, jak również na problemie *Named Entity Recognition*, jako moduł pełnej (*end-to-end*) metody bazującej na zanurzeniach pochodzących z modelu GPT-1 z wyborem określonych metryk odległości w przestrzeni zanurzeń, oraz metodą ważenia zanurzeń SIF (*Smoothed Inverse Frequency*).

- Utworzenie metod identyfikacji fragmentów tekstu oraz ich klasyfikacji w kontekście dokumentów zawierających treści propagandowe [3]. Metody te wygrały w prestiżowym konkursie SemEval-2020 (Task 11) w kategorii klasyfikacji, oraz zajęły drugiej miejsce w kategorii identyfikacji. W twórczy sposób zaproponowano w pracy wykorzystanie danych nienadzorowanych poprzez użycie metodyki *self-training* (algorytm sam etykietuje sobie nienadzorowane przykłady), co miało istotny wpływ na finalną trafność metod. Sam problem identyfikacji treści propagandowych bazował na architekturze RoBERTa-CRF i problemie etykietowania sekwencji. Z kolei problem klasyfikacji propagandy został rozwiązany poprzez dodanie mniejszego modelu o architekturze transformer korzystającego z reprezentacji poprzednio nauczonej sieci, wykorzystaniu techniki *self-training*, a także ważenia obserwacji, aby poradzić sobie z niezrównoważonymi klasami. W pracy znalazły sie obszerne analizy wyników, w tym tzw. *ablation studies*, określający wpływ poszczególnych modułów rozwiązania na finalną jego jakość, analiza błędów ze względu na obecność pewnych cech sekwencjach tekstu, oraz analiza macierzy pomyłek.

- Wprowadzenie architektury sieci neuronowej TILT w celu rozwiązania problemu rozumienia dokumentów w kontekście zadań ekstrakcji informacji i odpowiadania na pytania [4]. TILT to multimodalny model wykorzystujący architekturę *encoder-decoder*,

wprowadzający transformery ze świadomością informacji przestrzennej poprzez zastosowanie członu obciążenia (*bias*) w mechanizmie uwagi, biorącego pod uwagę relatywną przestrzenną pozycję tokenów. Dodatkowo, wykorzystano również sieć o architekturze *U-Net*, umożliwiającą wzięcie pod uwagę cech wizualnych dokumentu, której wynikowe zanurzenia dodawane są do zanurzeń wynikających z informacji semantycznej. Model był wstępnie trenowany na dużym korpusie dokumentów, a następnie dotrenowywany na docelowych zadaniach. Autorom udało się na dwóch zadaniach uzyskać wyniki przewyższające pod względem trafności najlepsze istniejące rozwiązania.

- Wprowadzenie obszernej kolekcji zbiorów dokumentów tekstowych pod nazwą *Document Understanding Evaluation (DUE)* [5] – wybrano siedem zbiorów danych, niektóre z nich zostały przeformułowane i poprawione celem zwiększenia ich jakości. Autorzy dokonali też ręcznej adnotacji wielu dokumentów. Zbiory dotyczą szeregu zadań rozumienia dokumentów, takich jak odpowiadanie na pytania z informacją wizualną, ekstrakcji kluczowej informacji, czy weryfikacji informacji. Przedstawiono również szereg rozwiązań bazowych, wraz z wynikami ich zastosowania do zebranych zbiorów danych.

- Zaproponowanie konkursu *Document UnderstanDing of Everything (DUDE)* w ramach konferencji ICDAR 2023, w dziedzinie rozumienia dokumentów [6]. Zbiór danych, stanowiący bazę konkursu, składa się z 5 tysięcy dokumentów z treściami wizualnymi, wraz z przyporządkowanymi do nich 40 tysiącami pytań i odpowiedzi. Dokumenty są multimodalne, wielostronicowe, i pochodzą z wielu różnych dziedzin. Opisano również zasady konkursy, sposób ewaluacji i wyniki poszczególnych uczestników.

Każda z wymienionych publikacji daje istotny wkład do badanej dziedziny. Część prac wprowadza nowatorskie metody z sukcesem rozwiązujące obrany problem badawczy, zweryfikowane eksperymentalnie względem najlepszych istniejących rozwiązań. Ostatnie prace proponują również kolekcję nowych zbiorów danych i konkurs rozumienia dokumentów, które są moim zdaniem nie mniej istotne od nowych rozwiązań, gdyż pozwalają śledzić postęp w dziedzinie.

Krótko podsumowując, uznaję, że wymienione na wstępie pracy cele udało się Doktorantowi w pełni osiągnąć.

5 Dalsze uwagi

Nie kwestionując wartości całościowych wyników zawartych w rozprawie, chciałbym zgłosić poniżej kilka uwag, głównie w formie pytań bądź dyskusji.

- W pracy [1] nie było dla mnie jasne, jak został skonstruowany *segmenter* – pozostałe metody zostały tam opisane szerzej, natomiast segmenterowi poświęcono jedynie dwa krótkie zdania. Ciekawi mnie, jak wygląda podział na sekwencje w dokumencie i

– w konsekwencji – jakie sekwencje mogą został później dopasowane i zwrócone jako zidentyfikowany fragment.

- W pracy [2] miałem trudności ze zrozumieniem równań (1-3): występuje w nich po lewej stronie równania indeks i, ale z prawej strony równania ten sam indeks występuje tylko jako licznik w sumie, co oznacza, że prawa strona jest niezależna od i. Wydaje mi się, że nie o to Doktorantowi w tych równaniach chodziło i będę wdzięczny za wyjaśnienie.

  Podobnie, miałem trudności z interpretacją pętli w algorytmie 4: proszę zwrócić uwagę, że indeks j nie jest w obrębie pętli w ogóle zwiększany, przez co sekwencja u wydaje się „stać w miejscu"

- Praca [3]: zaproponowano sposób radzenia sobie z niezrównoważonymi klasami poprzez ważenie odwrotnościami ich częstości, co jest jak najbardziej sensowne, ale nie zawsze działa najlepiej (może podnosić wariancję modelu szczególnie w przypadku mało-licznych klas). Czy Autor próbował również innych podejść do tego problemu, np. dopróbkowywania klasa mniejszościowej (*resampling*), tworzenia sztucznych przykładów (np. przez parafrazowanie), czy użycia dedykowanych funkcji błędu (np. *focal loss*)?

- Uwaga do pracy [4]: Wydaje się, że pewne elementy wizualne mogłyby zostać wzięte pod uwagę w transformerach semantycznych, niezależnie od użytej sieci U-Net. Przykładowo, odległość w pionie lub poziomie można by próbować liczyć liniami tekstu (bądź uzależnić ją od rozmiaru użytego fontu). Potencjalnie mogłaby również pomóc normalizacja jasności czy kontrastu w dokumencie.

- Na koniec pytanie ogólne. Od niedawna duży model językowy GPT-4 pozwala swoim użytkownikom na przetwarzanie, obok tekstu, również informacji wizualnej. Z pewnością możemy się w wkrótce spodziewać innych dużych systemów rozumiejących i przetwarzających taką informację. Jak zdaniem Doktoranta kształtuje się jakość takich systemów w rozwiązywaniu poszczególnych problemów inteligentnego przetwarzania dokumentów względem rozwiązań dedykowanych?

6 Konkluzja końcowa

Rozprawę oceniam bardzo dobrze, co w powyższej recenzji podkreśliłem wielokrotnie. Problemy badawcze, z którymi zmierzył się mgr inż. Dawid Jurkiewicz, są ambitne i istotne dla postępu w dziedzinie inteligentnego przetwarzania dokumentów, a w szczególności identyfikacji relewantnych fragmentów tekstu oraz rozumienia dokumentów. Sama rozprawa charakteryzuje się wysokim poziomem merytorycznym i zawiera wiele interesujące rezultatów, jednoznacznie potwierdzających użyteczność i praktyczność zaproponowanych algorytmów. W mojej opinii Doktorant wykazał się bardzo dobrymi umiejętnościami prowadzenia badań naukowych. Wszystkie postawione w rozprawie cele zostały osiągnięte.

W związku z tym **rozprawę oceniam jako spełniającą wymogi stawiane pracom doktorskim i wnoszę o dopuszczenie mgr. inż. Dawida Jurkiewicza do dalszych etapów postępowania w sprawie nadania stopnia doktora, równocześnie proponując wyróżnienie pracy.**

Wojciech Kotłowski

Dr hab. inż. Wojciech Kotłowski