

Recenzja rozprawy doktorskiej

mgr. Michała Pietruszki

z tytułowanej:

*Bridging the Generalization Gap: A Methodological Approach to Building and Benchmarking
NLP Models*

1. Rozprawa od strony formalnej

W myśl Ustawy z dnia 20 lipca 2018 r. „Prawo o szkolnictwie wyższym i nauce”, przedłożoną rozprawę stanowi „zbiór opublikowanych i powiązanych tematycznie artykułów naukowych”. W autoreferacie w tabeli I (str. 10) Doktorant przedstawił listę 9 prac ze swoim współautorstwem, na której jednak znalazły się również dwie prace nieopublikowane (nr 5 i 9). Należy odnotować, że sposób prezentacji zbioru prac w tabeli 1 nie odpowiada zwyczajowym standardom, ani nie ułatwia pracy recenzenta. Część prac ma niekompletne rekordy bibliograficzne, a we wszystkich podano tylko pierwszego autora, pomimo, iż wszystkie są wieloautorskie. Do autoreferatu nie dołączono kopii żadnej z prac (zamieszczono jedynie skrótowe opisy), a przynajmniej jedna z prac nie jest dostępna z podanego odnośnika (wymaga opłaty za dostęp), tj. praca nr 2, prace nr 7 i 8 nie mają odnośników (i recenzent nie miał pewności czy odnalazł w sieci właściwe wersje), a do pracy nr 6 podano odnośnik do jej kopii w repozytorium Semantic Scholar, zamiast do oryginalnej publikacji. Zamieszania dopełnia fakt, iż w tabeli I pierwsza kolumna zawiera symbole rozdziałów autoreferatu, które się powtarzają, a nie numery porządkowe, więc trzeba ręcznie odliczać kolejne prace.

Ze względu na sposób zaprezentowania prac ze zbioru i fakt braku ich jednoznacznego przedłożenia, można by rozważyć, czy rozprawę nie należy oceniać w kategoriach „praca pisemnej, w tym monografia naukowa”. Ponieważ jednak przedłożona praca pisemna dotyczy rozległej tematyki, którą omawia dosyć wybiórczo, skrótowo i z bardzo nielicznymi odwołaniami do literatury (tzn. bibliografia zamieszczona w pracy jest obszerna, ale wzmiankowanych jest bardzo wiele zagadnień i wiele z nich nie ma oparcia w literaturze, np. str. 11), recenzent przyjął interpretację korzystniejszą dla Doktoranta i potraktował tę pracę jako autoreferat wprowadzający do zbioru publikacji.

Wszystkie 7 prac (prace nieopublikowane wyłączamy z dalszych rozważań) zostały opublikowane na bardzo dobrych konferencjach naukowych, tj. CNLL, ICDAR (2), EACL, ACL, AAAI i ICCV. Wszystko to uznane światowe konferencje o wysokiej randze i wysokiej polskiej punktacji wg listy MNiSW.

Wszystkie prace ze zbioru mają charakter wieloautorski. Swoją wkład Doktorant opisał w ramach skrótowego przedstawienia poszczególnych prac w podrozdziałach autoreferatu. Utrudnia to analizę własnego wkładu Doktoranta w całość zbioru stanowiącego rozprawę. Ponieważ jednak większość prac ze zbioru została już przedstawiona wcześniej jako elementy zbioru stanowiącego rozprawę doktorską dr Łukasza Borchmanna, to należałoby zacząć od ustalenia faktycznego zakresu niniejszej rozprawy. Poniżej porównamy prace z obu zbiorów, w kolejności ich wymienienia w niniejszej rozprawie, pod

względem wkładu twórczego wykazanego w obu rozprawach dr Lukasa Borchmanna (obronionej w 2022 roku) i zadeklarowanego w ramach niniejszej rozprawy.

- [P4.1] “From Dataset Recycling to Multi-Property Extraction and Beyond”:
 - Dr Borchmann określił swój wkład jako, m.in., „Conceptualization and methodology ... curation of human-annotation process”,
 - wkład zadeklarowany przez Doktoranta to „My main contribution to the article is the redefinition of the dataset,”
 - wydaje się być pomocniczy i realizować podzadanie na rzecz głównych celów badawczych.
- [P5.1] “Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer”:
 - Dr Borchmann: „Conceptualization and methodology, implementation and experiments with model prototypes, running experiments with the final model, writing the paper”,
 - Doktorant: „My primary contributions to TILT were twofold. First, I facilitated our collaboration with the LAMBERT team to develop a generalizable architecture for Key Information Extraction, with a focus on layout understanding. Second, I was responsible for sourcing, transforming, and evaluating datasets for supervised multimodal QA, which significantly improved our training process”.
 - Praca opisująca metodę LAMBERT jest wspomniana w artykule [P5.1] w odniesieniach do literatury, ale nie jest wymieniana jako punkt wyjścia do metody opracowanej w [P5.1].
 - W zakresie danych w [P5.1] wykorzystano gotowe zbiory danych, kwestii danych poświęcono krótką sekcję i nie są raportowane żadne działania na danych poza wykorzystaniem tych zbiorów w ewaluacji i filtracją części elementów w jednym ze zbiorów.
 - Wkład Doktoranta należy uznać za pomocniczy w ramach realizacji głównych celów [P5.1].
- [P5.2] „STable: Table Generation Framework for Encoder-Decoder Models”:
 - Trudno znaleźć informacje na temat przedstawienia [P5.2] w ramach innej rozprawy.
 - Doktorant: „My contributions to this work include the main idea behind the algorithm—both for training and inference—as well as the architectural changes, implementations, and related ablation studies.”
 - Wkład doktoranta należy uznać za znaczący, ale warto zauważyć, że w tekście [P5.2] trzech współautorów jest oznaczonych jako wiodących z równym udziałem, a łącznie współautorów jest ośmiu.
- [P6.1] “Sparsifying Transformer Models with Trainable Representation Pooling”:
 - Dr Borchmann: “Conceptualization and methodology, idea of using subset selection, design of Transpooler, performing experiments, analysis of the results, implementation of several pooling strategies, writing the paper, implementation of LSH and Efficient transformer baselines”
 - Doktorant: “My contributions to this work were substantial, including the design of the successive halving algorithm, the integration of this algorithm into the DeepPyramidion model, conducting all experiments and ablation studies, implementing half of the baseline models, preparing the datasets, and managing the overall scope of the research project.”
 - Z porównania opisów widać, że Doktorant pełnił pomocniczą rolę w zakresie opracowania metody jak również w zakresie przeprowadzenia eksperymentów.

- [P7.1] “DUE: End-to-End Document Understanding Benchmark”:
 - Dr Borchmann: “Conceptualization and methodology (including participation in regular discussions with the core team), writing the paper, methodology of datasets selection process, implementation of baselines, results analysis, organization and controlling of the human annotation process, creation of the accompanying website and evaluation scripts, reformulation of TabFact and WTQ datasets”,
 - Doktorant: “I was one of three primary authors who developed the concept of the benchmark and the methodology for evaluating and selecting the datasets. Additionally, I was responsible for identifying suitable datasets, preparing the majority of them, participating in the implementation of baseline models, and organizing the code, models, and datasets for the final release.”.
 - Z porównania wynika, że Doktorant uczestniczył w opracowaniu idei wzorcowego zbioru testowego i wniósł pomocniczy twórczy wkład w budowę tego zbioru.
- [P7.2] „Document Understanding Dataset and Evaluation (DUDE)”:
 - Dr Borchmann: jest współautorem, cytuje tę pracę, ale nie wykazał jej jako element rozprawy,
 - Doktorant: “my primary contributions encompassed the manual collection of documents, enhancement of our implementation of the T5 model family, experimentation to improve baseline models, and the achievement of state-of-the-art. results.”
 - Wkład Doktoranta można uznać za istotny w rozwój zbioru danych i jego ewaluacji, ale warto zauważyć, że zespół współautorów jest duży (13 osób), międzynarodowy i naukowców z kilku uczelni i firm.
 - Dodatkowo opublikowana została praca [P7.2: ICDAR 2023] podsumowująca organizację konkursu naukowego i opublikowana na prestiżowej konferencji. Wkład Doktoranta nie został określony, ale ze względu na jej charakter, należy scharakteryzować jako istotny w zakresie rozwoju dyscypliny.
- [P8.1] ani nie została opublikowana, ani też nie została zaakceptowana do publikacji, przynajmniej do w momentu składania rozprawy. Upubliczniony został jedynie jej *preprint*. Dodatkowo na str. 34 przypisano [P8.1] kategorię „CORE: A* | MNiSW: 200”, co jest całkowicie bezpodstawne i mylące, bo nie była to opublikowana praca naukowa (sic!). Ponieważ mamy do czynienia z rozprawą doktorską, a nie nieformalnym CV Doktoranta, to należy potraktować włączenie [P8.1] do zbioru jako błąd formalny, który ponadto utrudnia pracę recenzenta. Praca [P8.1] nie będzie traktowana jako element zbioru stanowiącego rozprawę.
- Na liście publikacji stanowiących zbiór na str. 10, oprócz właściwej publikacji [P6.1] został wymieniony wcześniejszy rozszerzony abstrakt tej samej publikacji. Co prawda jego punktacja MNiSW została określona jako 0, ale ma inną nazwę i został włączony do zbioru. Należy to potraktować jako błąd formalny i kolejne utrudnienie dla pracy recenzenta, chyba że był to test na uważność recenzenta.

Podsumowując, nie ma podstaw, aby nie uznać dużego wkładu badawczego Doktoranta w pracę [P5.2] oraz istotnego wkładu w pracę [P7.2] jak i twórczej współpracy przy pracy [P7.1]. Ponadto Doktorant wniósł wkład w pracę [P7.2: ICDAR 2023], która jednak dotyczy organizacji konkursu i w zakresie zbioru danych i metod ewaluacji powieła rozwiązania z [P7.2]. W pozostałych pracach Doktorant wykonywał pomocnicze zadania i prace te zostały już wcześniej przedstawione jako elementy rozprawy dr Łukasza Borchmanna. Oznacza to, że z listy 9 prac w tabeli na str. 10, trzem pracom należy się przyjrzeć uważnie, jako pracom definiującym oryginalny merytoryczny rezultat rozprawy. Pozostałe

prace będą scharakteryzowane krótko, jako prace uzupełniające, chociaż zadeklarowane jako należące do zbioru. Prace nieopublikowane nie będą wcale wzięte pod uwagę w niniejszej recenzji, jako że na mocy ustawy nie stanowią one części rozprawy.

Poza zbiorem publikacji, rozprawa zawiera autoreferat, który powinien pokazać jak prace ze zbioru realizują temat i cele badawcze rozprawy. Ogólnie struktura autoreferatu spełnia to założenie. Sekcja "P3 Background and Notation" ma w zamierzeniu wprowadzić podstawowe pojęcia, założenia i koncepcje będące punktem wyjścia do procesu badawczego, a po niej następuje pięć sekcji prezentujących kolejne etapy badawcze. Jednak zaprezentowana w sekcji P3 całościowa koncepcja procesu badawczego i działań odbiega od rzeczywistej problematyki badawczej poszczególnych prac, do czego wrócimy w dalszej części recenzji.

W kolejnych sekcjach autoreferatu zamieszczono również syntetyczne skrócone opisy prac ze zbioru dobrze prezentujące poszczególne problemy badawcze. Szkoda, że zabrakło kopii poszczególnych prac ze zbioru w powiązaniu z przedstawionymi ich opisami.

Autoreferat zawiera również podsumowanie, ale niestety skupia się ono na koncepcjach wprowadzonych w sekcji P3, a nie na problemach badawczych rozwiązywanych w poszczególnych pracach ze zbioru.

2. Problem badawczy i jego znaczenie

Temat rozprawy mówi o zmniejszaniu luki generalizacyjnej (ang. „generalization gap”) poprzez metodologiczne podejście do budowania i oceny modeli z obszaru przetwarzania języka naturalnego na podstawie wzorcowych zbiorów danych. Pojęcie luki generalizacyjnej zostało szeroko zdefiniowane (str. 4) jako problem techniczny („technical problem”) charakteryzowany przez 5 aspektów („dimensions”): ewaluacyjny, architektoniczny (*de facto* algorytmiczny, bo odnoszący się do sposobu działania programu, którego częścią jest sieć neuronowa, określana mianem „model”), praktyczny (*de facto* wydajnościowy lub złożoności algorytmu), złożoności danych (czyli w sumie tożsamy z ewaluacyjnym, bo odnosi się do rodzajów danych testowych) oraz ‘integracji umiejętności’ (ang. „skill integration”, czyli tożsamy z algorytmicznym, bo dotyczy rodzaju zadań przetwarzanych i ich kompleksowości). Poszczególne aspekty nie są rozłączne (co zostało zasygnalizowane w komentarzach powyżej) i są raczej wyjaśnione opisowo, niż zdefiniowane. Opisy są dosyć intuicyjne, ale nie zastępują poprawnych definicji.

Pojęcie metodologicznego podejście („methodological approach”) pojawiające się w temacie, powtórzone na str. 4 i występujące wśród celów badawczych (sekcja P1.1) jest niejasne i nie zostało w rozprawie zdefiniowane. Metodologia zwykle oznacza naukę o metodach lub stosowanych procedurach. W rozprawie „metodologia” wydaje się być interpretowana jako sekwencja konkretnych etapów procesu opracowywania innowacyjnego rozwiązania, gdzie w ramach poszczególnych etapów określone są oczekiwane działania. Wydaje się, że w pracy proponowany jest określony sposób postępowania, więc określona metoda. Tym bardziej niejasny jest termin „Methodological Roadmap”.

W sekcji P1.1 sformułowane zostały trzy główne cele badawcze („research objectives”) rozprawy:

1. Scharakteryzowanie kluczowych barier dla generalizacji („Characterize the Key Barriers to Generalization”), w odniesieniu do zdefiniowanych aspektów, które stanowią o luce pomiędzy akademicką („academic”), a przemysłową („industrial”) jakością rozwiązań w dziedzinie przetwarzania języka naturalnego (PJN, ang. NLP).
2. Opracowanie i ewaluacja docelowych rozwiązań technicznych („Develop and Validate Targeted Technical Solutions”) – opracowanie rozwiązań przekraczających zidentyfikowane bariery i demonstrujące poprawioną zdolność modelu językowego do generalizacji.
3. Wykazanie spójnego podejście metodologicznego („Demonstrate a Cohesive Methodological Approach”) – pokazanie, że uwzględnianie poszczególnych aspektów generalizacji w

rozwiązaniach w dziedzinie PJN powoduje, że są bardziej niezawodne, wydajne i stosowalne w praktycznych zastosowaniach przemysłowych.

Pierwszy z celów ma charakter naukowy, ale jest trudny do osiągnięcia bo wymagałby bardzo szeroko zakrojonych badań oraz zdefiniowanie miar oceny generalizacji w jej rozumieniu w pracy. Wyzwanie to nie zostało podjęte, ani w autoreferacie, ani w żadnej publikacji ze zbioru, np. nie sformułowano miary generalizacji rozpisanej na wszystkie pięć aspektów.

Pozostałe dwa cele nie mają charakteru naukowego z wielu powodów. Wystarczy wymienić jeden: oba mogą być osiągnięte w trywialny sposób.

Szukając dalej co było celem rozprawy możemy znaleźć stwierdzenia, że odnoszące się do metodologii:

- "This dissertation proposes and follows a structured methodology for tackling the generalization gap."
- "This dissertation argues that bridging the generalization gap in NLP requires a systematic, multi-faceted technical approach."
- "we propose a methodological roadmap tailored to the unique characteristics of NLP development. The gap lies in the absence of a framework that specifically addresses the unique challenges faced by NLP technologies at each stage of their development."

Z jednej strony jest to oczywiste, że dobór odpowiedniego algorytmu (metody rozwiązania) do problemu w sposób biorący pod uwagę zmienność warunków jego stosowania (co jest częścią dobrego zdefiniowania problemu) oraz dobór odpowiednich metod ewaluacji, które pokażą wszystkie własności i ograniczenia tego algorytmu jest warunkiem koniecznym. Z drugiej strony, dalszy ciąg autoreferatu pokazuje, że chodzi jednak o określoną metodę, a nie o metodologię. Pomijając na razie kwestię czy metodę taką autoreferat przedstawia, to warto zwrócić uwagę, że żadna z prac ze zbioru nie dotyczy metody rozwiązywania problemów z dziedziny PJN (NLP). Tymczasem, w przypadku rozprawy w postaci zbioru publikacji, to te publikacje powinny realizować cele badawcze rozprawy.

Dodatkowo, na stronie 38, w podsumowaniu Doktorant zadeklarował, że:

„Ultimately, this dissertation’s primary contribution is a methodological one.”

Jest to sformułowanie całkowicie błędne od strony formalnej, ponieważ aspekt metodologiczny, jeżeli się już pojawia w którejś z prac ze zbioru, to jedynie śladowo. Żadna z prac ze zbioru nie dotyczy metodologii prowadzenia badań, a jedynie konkretnych problemów naukowych jak budowa zbiorów wzorcowych (ang. benchmark dataset) czy też opracowanie metody do wydobywania informacji kluczowej (key information extraction).

Również sam autoreferat nie realizuje postawionych celów badawczych.

Sekcja „P3 Background and Notation” przedstawia bardzo hasłowe wprowadzanie do zagadnień, których dotyczą prace omawiane w dalszej części autoreferatu, praktycznie bez odwołań do literatury, np. podsekcja dotyczą miar ewaluacji, bardzo istotnego zagadnienia w pracach ze zbioru, jest tak enigmatyczna, że w tej postaci całkowicie zbędna.

Wprowadzenie do dziedziny i zadania nie jest problemem w przypadku rzeczywistej tematyki samych prac ze zbioru, które mają wysoką jakość naukową i posiadają bardzo dobre bibliografie. Jednak, jeżeli autoreferat miał przedstawiać nową metodę rozwiązywania problemów z dziedziny PJN (NLP), to konieczne byłoby bardzo dobre osadzenie tej propozycji w literaturze przedmiotu i skrupulatne, naukowe podejście do tego wyzwania.

Dobór zagadnień w sekcji P3 wydaje się być bardzo wybiórczy, ale to wynika też z bardzo szerokiego i słabo określonego celu pracy. Wiele z tytułów krótkich sekcji tego rozdziału to całe działy sztucznej inteligencji, więc trudno zrozumieć czemu one mają służyć w strukturze pracy. *De facto* ich treść służy podkreśleniu najciekawszych elementów prac składających się na zbiór. Trochę szkoda, że Doktorant nie przebudował swojego autoreferatu w oparciu o te elementy, zamiast podążać za celami bardzo niejasno określonymi w sekcjach 1 i 2.

W szczególności, brakuje dobrego porównania do wielu innych podejść do opracowywania innowacyjnych rozwiązań i metod wytwarzania oprogramowania.

Patrząc na samą metodą, nie jest jasne jaka jest różnica pomiędzy podstawowymi zbiorami wzorcowymi (benchmark dataset) w fazie pierwszej oraz zbiorami wzorcowymi w fazie czwartej? W metodzie pominięta jest całkowicie kwestia zebrania zbioru wymagań czy też sformułowania opisu problemu. Wydaje się, że domyślnie zbiory wzorcowe są przyjmowane w autoreferacie jako definicja problemu, ale to nie jest prawdą.

Opis metody nie wychodzi poza intuicyjnie określone jej poszczególne fazy. Do każdej z faz podpięte są jedna lub dwie prace ze zbioru, w których zaproponowane są rozwiązania poszczególnych problemów naukowych. Można to traktować jednak jako przykłady działań w ramach poszczególnych faz, a nie jako szczegółową ich definicję.

W podsumowaniu Doktorant stwierdza, że w ramach rozprawy wykazano, że zaproponowana metoda zmniejsza lukę generalizacją, jednak realizację metody pokazano przy pomocy wybranych kilku przykładów działań. Co więcej, ilustracja ta jest częściowa, bo wyniki poszczególnych prac nie łączą się ze sobą w jeden proces. Dlatego metoda jest zarówno daleka od poprawnej definicji, jak i też od naukowej weryfikacji.

Sama treść autoreferatu nie ma wartości naukowej, co szczególnie staje się widoczne, jeżeli zauważymy, że ani nie ma dobrego porównania do literatury w zakresie metod, ani też nie ma żadnej próby oceny zaproponowanej metody w jakim stopniu ona redukuje lukę generalizacyjną.

Rozprawę jednak stanowi zbiór prac opublikowanych na bardzo prestiżowych konferencjach naukowych – prac o niewątpliwie bardzo dużej wartości naukowej. Dlatego należy przyrzeć się poszczególnym pracom pod kątem ich celów badawczych i osiągniętych wyników. Analizie poddane zostaną jedynie prace opublikowane oraz główny nacisk zostanie położony na prace ze znaczącym wkładem naukowym Doktoranta.

W pracy [P5.2] („STable: Table Generation Framework for Encoder-Decoder Models”) głównym celem było opracowanie neuronowego modelu językowego umożliwiającego generowanie tabel na podstawie danych tekstowych. Podjęto problem trenowania modelu w oparciu o dane prezentowane w dowolnej kolejności i zdolności modelu do odkrywania semantycznej struktury danych.

Celem pracy [7.1] “DUE: End-to-End Document Understanding Benchmark” było połączenie kilku dostępnych testowych zbiorów danych (ang. benchmark datasets) dla różnych zadań wydobywania informacji i rozumienia dokumentów w jeden nowy ujednolicony testowy wzorcowy zbiór danych.

Praca [7.2] stanowi logiczną kontynuację [7.1] i ukierunkowana była na opracowanie nowego testowego zbioru danych tego typu w zasadniczy sposób rozszerzający założenia poprzedniego pod wieloma względami. Celem pracy [7.2] było opracowanie koncepcji i zbudowanie nowego anotowanego testowego zbioru wzorcowego opartego na dokumentach tekstowych, jedno i wielostronicowych, obejmującego zarówno warstwę wizualną jak i tekstową oraz ukierunkowanego na możliwie szeroki zbiór zadań wydobywania informacji z dokumentów. Zbiór danych powinien reprezentować możliwie dużą liczbę dziedzin.

Dodatkowo celem pracy [P7.2: ICDAR 2023] było przygotowanie w oparciu o zbiór z [P7.2] konkursu naukowego dla rozwiązań w zakresie rozumienia multimodalnych dokumentów zorganizowanego przy konferencji naukowej najwyższej rangi.

Ponadto, w opublikowanych pracach, w których Doktorant odgrywał pomocniczą rolę, podejmowane były zagadnienia:

- [P4.1] przekształcenia wcześniejszego wzorcowego zbioru danych opartego na danych z Wikipedii w zbiór definiujący wydobywanie kluczowych informacji dla wielu własności jednocześnie,

- [P5.1] opracowania metody wydobywania kluczowych własności z dokumentów z wykorzystaniem jednoczesnym warstwy tekstowej i informacji o układzie dokumentu,
- opracowania wydajnej i efektywnej architektury głębokich sieci neuronowych do przetwarzania długich sekwencji tekstowych.

Podsumowując, główne cele badawcze prac realizowanych przez Doktoranta to budowana anotowanych wzorcowych zbiorów danych dla rozwoju metod wydobywania kluczowych informacji z dokumentów oraz częściowego rozumienia dokumentów, jak również rozwiązywanie szczegółowych problemów w obszarze metod wydobywania informacji i wiedzy z treści dokumentów. Cele te układają się w spójną całość, a nawet wskazują na to, że prace ze zbioru stanowią elementy pewnego cyklu. Cele tych prac mają charakter badawczy i w momencie opublikowania tych prac wpisywały się w aktualne nurty rozwoju dziedziny. Tym bardziej dziwi niejasny tytuł całości rozprawy i praktyczny brak sformułowanych celów badawczych dla rozprawy.

3. Oryginalny wkład w rozwój dyscypliny

Wkład Doktoranta w rozwój dyscypliny można rozpatrywać na dwóch poziomach: wkładu wynikającego z realizacji celów całej rozprawy doktorskiej oraz poszczególnych publikacji ze zbioru. W tym pierwszym przypadku wkład ten jest trudny do uchwycenia, ponieważ cele główne zostały słabo określone i spośród celów głównych tylko pierwszy ma charakter badawczy, ale w bardzo ograniczony i pośredni sposób odnosi się do celów zrealizowanych w ramach publikacji ze zbioru.

Na poziomie publikacji z istotnym udziałem badawczym Doktoranta, czyli [P5.2], [P7.1] i [P7.2], w przypadku tych dwóch ostatnich opracowane zostały wzorcowe zbiory testowe (ang. benchmarks). Obydwa zostały udostępnione na otwartych licencjach. Na obydwu zbiorach przeprowadzono szereg testów różnych metod wydobywania informacji z dokumentów i pokazano, że zbiory dobrze reprezentują szereg różnorodnych trudnych problemów przetwarzania. Oryginalność zbioru z pracy [P7.1] była ograniczona, ponieważ został zbudowany poprzez agregację i unifikację szeregu istniejących już zbiorów danych. Jednak zbiór zaproponowany w pracy [P7.2] to już oryginalny zasób zbudowany od podstaw z otwartych danych źródłowych, na które zostały naniesione ręczne anotacje. Wzorcowe zbiory danych są obecnie kluczowe dla rozwoju przetwarzania języka naturalnego. Szczególnie wartościowym jest zbiór z [P7.2] ze względu na jego wielkość, wielość reprezentowanych dziedzin, multimodalnych charakter oraz bogatą, ręczną anotację. Jego cenną cechą jest włączenie pytań bez poprawnej odpowiedzi w dokumencie. Takie negatywne dane są bardzo cenne, a rzadko występują w zbiorach wzorcowych, szczególnie dla zadania odpowiadania na pytania. W [P7.2] zaproponowano też ciekawy, kaskadowy system ręcznej anotacji i jej weryfikacji oraz podzbiory weryfikacyjne opisane metadanymi diagnostycznymi. Te ostatnie są cenne dla analizy wyników metod testowanych na zbiorze wzorcowym.

W pracy [P5.2] zaproponowano metodę generowania tabel na podstawie informacji zawartej w tekście, gdzie połączono ciekawy sposób uczenia głębokiej sieci neuronowej typu transformer wraz zaproponowanymi rozszerzeniami do architektury sieci neuronowej typu transformer. W procesie uczenia wykorzystano wszystkie możliwe porządki predykcji. Transformer rozszerzono o dodatkowe elementy w mechanizmie uwagi i dodatkową warstwę do predykcji struktury tabeli. W oparciu o rozszerzony transformer zaproponowano heurystyczny algorytm rozpoznawania struktury tabeli i jej generowania. W specjalistycznym zadaniu wydobywania tabel z tekstu były to ciekawe rozwiązania, chociaż należy podkreślić, że w pracy [P5.2] zabrakło porównania z zastosowaniem dużych generatywnych modeli językowych w tym zadaniu.

Praca [P7.2: ICDAR 2023] przyczyniła się do rozwoju dyscypliny poprzez organizację konkursu naukowego o zasięgu międzynarodowym.

4. Poprawność i jakość badań

Jak już było to wspomniane, forma autoreferatu bardzo odbiega od standardów pracy naukowej i zawarte w nim propozycje odnośnie „metodologii” prowadzenia badań nie mają charakteru naukowego. Natomiast wszystkie opublikowane prace ze zbioru stanowiącego rozprawę reprezentują wysoki poziom naukowy i ich wyniki zostały poddane odpowiedniej ocenie. Niemniej, można znaleźć kilka niejasności oraz przyczynków do polemiki.

W pracy [P5.2] cała metoda wraz z wyspecjalizowaną architekturą sieci neuronowej i heurystycznym algorytmem sterującym są dostosowane pod jedno, konkretne zadanie transformacji opisu tekstowego do struktury tabeli. W bardzo wyraźny sposób brakuje porównania z podejściami opartymi na zastosowaniu generatywnych modeli językowych (LLM), wytrenowanych do wielu zadań.

Struktura sieci neuronowej typu transformer została rozszerzona, np. o dodatkowe czynniki w głowicach uwagi (tzn. „bias” w „attention heads”) oraz o dodatkową warstwę do predykcji liczby części składowych tabeli („groups”). Do realizacji samego zadania transformer został obudowany heurystycznym algorytmem generowania struktury tabeli, m.in., z wykorzystaniem strategii przeszukiwania zachłannego. Całość jest ciekawym rozwiązaniem, wyniki są bardzo dobre, przynajmniej według testów, ale brakuje bardzo porównania do działania dużych generatywnych modeli językowych (ang. LLMs) na tych samych zbiorach testowych, w tym z użyciem tych samych danych treningowych do dostrojenia (ang. fine tuning). Sama praca [P5.2] jest dosyć stara, jeśli chodzi o bibliografię (choć sama praca opublikowana została w roku 2024), ale można było takie badania wykonać na potrzeby rozprawy, tymczasem większość miejsca w autoreferacie jest wykorzystana na rozważania niezwiązane z celami badawczymi w publikacji ze zbioru.

Algorytm z [P5.2] został wytrenowany na specyficznym zbiorze pozyskanym ze stron WWW sformatowanych przy pomocy tabel. Natomiast ewaluacja została oparta w znacznym stopniu na zamkniętych zbiorach danych, a przez to ocena metody jest trudna do niezależnego powtórzenia.

Sekcja „2.5 Grammar-Constrained Decoding” wydaje się być niedokończona lub skrócona nadmiernie. Zawiera kilka dosyć oczywistych stwierdzeń i brakuje w niej odniesienia do rozwiązań, które zostały zaproponowane przez autorów. Nie ma nawet jasności czy zapowiadane w niej regularyzacje zostały wprowadzone. Prawdopodobnie chodzi o heurystyki zawarte w algorytmie sterującym generacją tabeli, który został przedstawiony w dodatkach, a które stanowią jedną trzecią długości artykułu (sic!). Tutaj również można było wykorzystać autoreferat, aby uporządkować narrację i lepiej przedstawić te osiągnięcia.

Ważne założenie upraszczające, że nagłówki tabeli do wydobycia z tekstu są podane do algorytmu pojawia się dopiero w załączniku.

W pracy [7.1] zunifikowano i zagregowano kilka istniejących wzorcowych zbiorów danych, co zaowocowało wartościową kompozycją wspierającą testy dla wielu zadań przetwarzania dokumentów, ale wszystkie dane i większość anotacji zawartych w zasobie to dane zagregowane ze zbiorów już istniejących wcześniej. Może to rodzić ryzyko, że dane te już są bardzo dobrze znane z punktu widzenia rozwoju różnych metod.

Bardzo dobrze, że zapewniono dodatkowe anotacje ręczne dla oszacowania dokładności ludzi w kilku zadaniach, ale dziwi brak próby zmierzenia zgodności pomiędzy anotatorami (IAA) i jej analizy ([P7.1], str. 8: „Each dataset was handled by two annotators; the average of their scores, ...”).

Brakuje podjęcia próby ewaluacji jakości zbudowanego zasobu, szczególnie pod kątem działań w kierunku oczyszczenia danych („data cleaning”) i poprawy podziałów danych.

Wzorcowy zbiór danych z [P7.2] został zbudowany od podstaw i robi bardzo dobre wrażenie pod względem wielkości, różnorodności dziedzin i reprezentowanych zadań. Opis selekcji materiału i filtrowania dokumentów pod kątem ich włączenia do zbioru jest dosyć enigmatyczny i w zasadzie

ograniczony do informacji, że dokonano tego ręcznie. Również nie jest jasne jak sprawdzano licencje gromadzonych materiałów. Nie zostało wyjaśnione jak została zaanotowana informacja w podzbiorach diagnostycznych („Diagnostic Subsets”). Dla poziomu odniesienia (*baseline*) w postaci wyników ludzi mimo wszystko brakuje analizy zgodności, choćby na próbkach, aby można było lepiej przeanalizować osiągnięte rezultaty. Bardzo niskie wyniki osiągnięte przez anotorów ekspertów w odpowiadaniu na pytania są zaskakujące. Jeżeli to problem metody porównania odpowiedzi ze wzorcem, jak zostało zasugerowane w pracy, to oznacza, że metoda porównania na potrzeby oceny nie jest dopasowana do zadania. Jeżeli przyczyną są rozbieżności w odpowiedziach, to powstają wątpliwości odnośnie spójności i tym jakości całości zbioru.

W odniesieniu do testów z użyciem dużych generatywnych modeli językowych (LLM) brakuje informacji o wykorzystanych pobudzeniach (ang. prompts), ich ewentualnym dostrojeniu, schemacie stosowania oraz parametrach modeli.

5. Wiedza kandydata

Autoreferat można podzielić na dwie części: wprowadzenie do problemów poruszanych w rozprawie i dziedziny oraz krótkie przedstawienie każdej pracy ze zbioru. Pierwsza część jest bardzo skrótowa, ogólna i selektywna, co widać szczególnie w sekcji „P3 Background and Notation”. Pomijając wątpliwości odnośnie do samych celów rozprawy, to powiązane z nimi rozważania metodologiczne są oparte na ograniczonym i selektywnym wyborze odniesień do literatury. Ponieważ prace naukowe ze zbioru nie rozwijają tych wątków, to w tym zakresie wiedza Kandydata z dyscypliny informatyka i telekomunikacja jest słabo pokazana.

Inny obraz wyłania się z prac naukowych ze zbioru, które stoją na wysokim poziomie i wszystkie posiadają bardzo dobrze dobraną i bogatą bibliografię. Omówienie stanu badań jest w każdym przypadku syntetyczne, ale są to artykuły konferencyjne. Prace ze zbioru są jednak wszystkie wieloautorskie, więc zawarte w nich odniesienia do stanu badań należy traktować jako efekt wspólnych działań autorów.

6. Komentarze szczegółowe

- s 7: „a methodological roadmap.” – pojęcie to jest kluczowe dla wyводу w autoreferacie, pojawia się tutaj wytłuszczone, ale nie zostało nigdzie zdefiniowane (sic!), np. w jakim stopniu „metodologiczna mapa drogowa” różni się od metodologii; ponadto w całym autoreferacie jest proponowany jeden określony sposób postępowania, a więc *metoda* realizacji procesu, a nie metodologia.
- s 37: „complex, real-world industrial settings” – pojęcie to nie zostało zdefiniowane w całej rozprawie, a jest podstawą kluczowej opozycji
- s 37: „By developing the WikiReading Recycled benchmark, we broke the Benchmark-Model Development Loop,” – zbiór ten to jeden konkretny testowy zbiór danych i jego zastosowanie z pewnością nie uprawnia do tak ogólnego stwierdzenia; zdanie to to jeden z wielu przykładów braku naukowej precyzji i logiki wyводу jakich można znaleźć dużo w autoreferacie.
- p 38: „It argues that the path from academic theory to industrial application is not a random walk of chasing state-of-the-art. metrics, but a structured process that must be navigated with intent.” – poza widocznym subiektywnym wartościowaniem, zdanie to sugeruje, iż proces tworzenia innowacyjnych rozwiązań przemysłowych na podstawie wyników prac naukowych można

opisać w sposób deterministyczny i algorytmiczny. Na szczęście dla innowacji przemysłowych tak nie jest i zdanie to jest nienaukowe i nieprawdziwe.

- p 38: „This roadmap is not merely a theoretical construct; it has been validated through the concret technical contributions of this thesis.” – Rozprawa zawiera garść przykładów rozwiązań dla określonych ograniczonych problemów. Twierdzenie, że w ten sposób można coś wykazać (validate) jest czystym przykładem nieuprawnionej naiwnej indukcji,
- [P5.2], str. 2458: „the relative position of table cells” – nie jest jasne, czy “cell” odnosi się do element transformera, czy też do elementu tekstu, traktowanego jako element struktury tabeli; analogicznie w „Predicting Number of Groups” – pojęcie „group” nie zostało zdefiniowane w artykule.
- [P7.2], str. 19532: nie jest jasne co oznaczają podwójne indeksy dla dokumentów w definicji Div.

7. Konkluzja

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez art. 187 Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (z późniejszymi zmianami)¹ moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

- rozprawa jako całość i w zakresie wkładu Kandydata nie prezentuje oryginalnego rozwiązania problemu naukowego, a jedynie rozwiązania problemów częściowych,
- Kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie informatyka i telekomunikacja,
- oraz Kandydat nie wykazał się umiejętnością samodzielnego prowadzenia pracy naukowej, w szczególności w zakresie określenia celów naukowych i sformułowania planu badawczego.

Podsumowując, rozprawa nie spełnia wszystkich formalnych i zwyczajowych wymagań stawianych rozprawom doktorskim.

Rozprawa w postaci w jakiej została złożona musiałaby być poprawiona w zakresie wskazanym w recenzji, aby można było wnioskować o dopuszczenie Doktoranta do dalszych faz przewodu doktorskiego. Obecna postać rozprawy nie pozwala na postawienie takiego wniosku.



Maciej Ruralski
Podpis

¹ <http://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20190000276>