prof. dr hab. Dominik Ślęzak
Wydział MIM, Uniwersytet Warszawski
ul. Banacha 2, 02-097 Warszawa
e-mail: slezak@mimuw.edu.pl

2024-12-03

Recenzja rozprawy doktorskiej

mgr Gabrieli Nowakowskiej

zatytułowanej:

Named entity recognition and information extraction from various documents

(*Rozpoznawanie jednostek nazwanych i ekstrakcja informacji z dokumentów różnego typu*)

1. Problem badawczy i jego znaczenie

Recenzowana rozprawa odnosi się do nowatorskich badań Doktorantki w dwóch zakresach – metod rozpoznawania jednostek nazwanych (ang.: named entity recognition – NER) oraz metod ekstrakcji informacji (ang.: information retrieval – IR) z dokumentów różnego typu. Zakresy te wiążą się ze sobą, są one bowiem składowymi większej dziedziny przetwarzania języka naturalnego (ang.: natural language processing – NLP). Sprawia to, że omawiana rozprawa jest z jednej strony tematycznie spójna, z drugiej zaś strony cechuje się różnorodnością omawianych podejść i zastosowań.

Dziedzina NLP jest ważna z praktycznego punktu widzenia, w związku z ogólnym trendem tworzenia, tudzież przenoszenia dokumentów do formatów cyfrowych, dzięki czemu są one łatwo dostępne, gotowe do wyszukiwania, podsumowywania, dalszego przetwarzania. Nie jest to jednak oczywiście proste. Zastosowania NLP wymagają poprawnej interpretacji języka naturalnego, a także zrozumienia struktury i umiejętności identyfikowania najważniejszych aspektów analizowanych dokumentów. Nie jest zatem zaskoczeniem częste wykorzystywanie w NLP metod sztucznej inteligencji. Nie dziwi też fakt, że choć w dziedzinie tej osiągnięto już znaczące sukcesy, jest wciąż dużo do zrobienia.

W zebranym materiale możemy znaleźć opis opracowanych przez Doktorantkę metod lematyzacji i maszynowego tłumaczenia tekstów, ekstrakcji informacji z dokumentów z uwzględnieniem ich struktury, a wreszcie zapisu wydobytych informacji w przejrzystej i wygodnej do dalszych zastosowań formie tabelarycznej. Są to istotne przykłady pracy z tekstami i dokumentami na różnych etapach procesu ich przetwarzania. Autorka wykorzystuje przy tym wspomniane już metody sztucznej inteligencji, łącząc je w umiejętny sposób z innymi technikami. A zatem recenzowana rozprawa jest cenna zarówno ze względu na swoją tematykę, jak i na podejście do tej tematyki.

2. Wkład Autorki

Rozprawa bazuje na czterech poniższych publikacjach naukowych, prezentowanych na renomowanych – a co więcej wysoko punktowanych – konferencjach o zasięgu międzynarodowym:

1. Artur Nowakowski, **Gabriela Pałka**, Kamil Guttmann, Mikołaj Pokrywka: *Adam Mickiewicz University at WMT 2022: NER-Assisted and Quality-Aware Neural Machine Translation*. Seventh Conference on Machine Translation (Abu Dhabi, United Arab Emirates), **WMT 2022**. https://aclanthology.org/2022.wmt-1.26/
2. **Gabriela Pałka**, Artur Nowakowski: *Exploring the Use of Foundation Models for Named Entity Recognition and Lemmatization Tasks in Slavic Languages*. 9th Workshop on Slavic Natural Language Processing (Dubrovnik, Croatia), **SlavicNLP 2023**. https://aclanthology.org/2023.bsnlp-1.19/
3. Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, **Gabriela Pałka**: *Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer*. 16th International Conference on Document Analysis and Recognition (Lausanne, Switzerland), **ICDAR 2021**. https://link.springer.com/chapter/10.1007/978-3-030-86331-9_47
4. Michał Pietruszka, Michał Turski, Łukasz Borchmann, Tomasz Dwojak, **Gabriela Nowakowska**, Karolina Szyndler, Dawid Jurkiewicz, Łukasz Garncarek: *STable: Table Generation Framework for Encoder-Decoder Models*. The 18th Conference of the European Chapter of the Association for Computational Linguistics (St. Julian's, Malta), **EACL 2024**. https://aclanthology.org/2024.eacl-long.151

Ze względu na kolejność współautorów, na szczególną uwagę zasługuje w mojej ocenie pozycja nr 2, w której przypadku Doktorantka była odpowiedzialna za opracowanie pełnej koncepcji podejścia, metodologię badań, zaprojektowanie całościowego rozwiązania informatycznego, implementację modułu NER, przeprowadzenie eksperymentów i wreszcie prace edycyjne nad artykułem. Punktem odniesienie dla tej pracy był konkurs przy konferencji SlavicNLP 2023, dotyczący przetwarzania wybranych języków słowiańskich. W zaproponowanym rozwiązaniu Autorka zastosowała nowe modele lematyzacji jednostek nazwanych. Wkładem Autorki było tu nowatorskie wykorzystanie tzw. modeli fundamentalnych, które są obecnie bardzo ważnym trendem w dziedzinie sztucznej inteligencji. Zastosowano między innymi różne warianty modelu BERT oraz T5. Wykazano, że poprzez dodanie do modeli BERT warstwy warunkowego pola losowego (ang.: conditional random field – CRF), rozumianej jako metoda przetwarzania danych sekwencyjnych, można znacząco poprawić jakość systemów NER. Podejście to ma duży potencjał praktyczny, co zostało pozytywnie zweryfikowane w konkursie.

Wkład Autorki w pozostałe prace również jest istotny i dobrze udokumentowany. Artykuł nr 1 dotyczy – wraz z pracą nr 2 – dziedziny NER. Także w tym przypadku Autorka była osobą odpowiedzialną za koncepcję, metodologię, implementację i eksperymenty. Tu również mamy do czynienia z konkursem, co podkreśla nastawienie Doktorantki na rozwiązywanie problemów praktycznych, często poprzez stosowanie zaawansowanych podejść i technik. Zadanie opisane w pracy nr 1 dotyczyło tłumaczenia tekstów. Na uwagę zasługuje tutaj wykorzystanie dobrze przemyślanego, autorskiego łańcucha przetwarzania dokumentów, gdzie współdziałają ze sobą dwa poziomy procesu tłumaczenia.

Artykuły nr 3 i 4 dotyczą dziedziny IR. Doktorantka deklaruje w nich wkład w przygotowanie zbiorów danych, przeprowadzanie eksperymentów, prace edycyjne i uczestnictwo w dyskusjach naukowych. Mogłoby się wydawać, że udział Doktorantki w tych pracach jest – w porównaniu z publikacjami nr 1 i 2 – mniej znaczący. A jednak odnoszę wrażenie – w szczególności na podstawie stopnia dojrzałości i dokładności, z jaką zagadnienia te zostały opisane w rozprawie – że Pani Nowakowska jest w dziedzinie IR wysokiej klasy ekspertem i że jej wkład w prace nr 3 i 4 był porównywalny z pracami nr 1 i 2.

Artykuły nr 3 i 4 opisują nowatorskie modele sieci neuronowych, opracowane i zaimplementowane w ramach prac wdrożeniowych. Podobnie jak w pracach nr 1 i 2, mamy tu nawiązanie do konkursów konferencyjnych (w których Autorka odnosiła niebagatelne sukcesy!), ale fakt podjęcia wdrożeniowej współpracy z przemysłem zasługuje na dodatkowe uznanie. Artykuł nr 3 dotyczy badań nad ekstrakcją informacji z dokumentów o dwuwymiarowej strukturze (warstwa tekstowa oraz wizyjna / graficzna). Natomiast publikacja nr 4 rozwija te badania w kierunku ekstrakcji danych tabelarycznych, tzn. takiego domknięcia całości procesu przetwarzania dokumentów (także z potencjalnym wykorzystaniem metod wprowadzonych w pracach nr 1 i 2), które skutkuje w wysokim stopniu ustrukturyzowaną informacją końcową, mogącą stanowić wejście do systemów raportowania bądź eksploracji danych.

Warto też dodać, że metody opisane w publikacjach nr 3 i 4 stały się podstawą dla uzyskania dwóch patentów w USA, w których Doktorantka występuje jako jeden z wynalazców. Podkreślmy, że nie są to tylko same zgłoszenia patentowe (ang.: patent applications), z którymi mamy do czynienia dość często. W istocie, są to już przyznane patenty (ang.: granted patents), co jest znaczącym sukcesem, szczególnie biorąc uwagę fakt, że mówimy o dziedzinie informatyki, o niezwykle popularnym zakresie zastosowań praktycznych, jakim jest IR, a wreszcie że patenty te dotyczą rozwoju metod sztucznej inteligencji, gdzie coraz trudniej jest dokonać czegoś istotnie wyróżniającego się na tle nieustająco przybywających na całym świecie publikacji. Jednym słowem, oceniam to jako wielkie osiągnięcie Doktorantki.

5. Poprawność

Nie mam żadnych znaczących uwag co do poprawności merytorycznej rezultatów zgromadzonych w rozprawie. Co więcej, jak już wspomniałem, bardzo pozytywne wrażenie wywarła na mnie jakość rozprawy. Kluczowy jest rozdział nr 1 stanowiący wprowadzenie w tematykę i syntetyczne omówienie najważniejszych wyników. (Dalsze cztery rozdziały odpowiadają czterem publikacjom wchodzącym w skład rozprawy, podzielonym na dwie części – NER i IR.) Rozdział ten jest niezwykle przejrzyście napisany, jego forma, dobór figur, przykładów i referencji – wszystko jest bez zarzutu. Jest to o tyle ważne, iż często zdarza się, że osoby przygotowujące swoje rozprawy w formie kolekcji publikacji nie dbają o rozdział wprowadzający, pewnego rodzaju wspólny mianownik wszystkich swoich wyników. W przypadku rozprawy Pani Nowakowskiej ten problem nie istnieje.

Z formalnych względów należy też wspomnieć o załącznikach umieszczonych na końcu rozprawy, a więc o certyfikatach związanych z konferencjami – w tym bardzo wysokimi miejscami w konkursach – o patentach oraz dokumentach świadczących o wkładzie autorów we wspólne publikacje. Ważna jest tu w szczególności deklaracja współautora publikacji nr 2, do której wrócę jeszcze na końcu.

6. Wiedza Kandydatki

Nie mam żadnych wątpliwości co do rozległej wiedzy Doktorantki w zakresie metod i zastosowań NLP, w tym dziedzin NER oraz IR omawianych w rozprawie. Nie posiadając takiej wiedzy, Autorka nie byłaby z pewnością w stanie osiągnąć tak znaczących sukcesów (konkursy, patenty), a także opisać swoich myśli w tak spójny i przejrzysty sposób (szczególnie w rozdziale nr 1 rozprawy). Korzystając z okazji, chciałbym zatem zadać Kandydatce kilka pytań, licząc na wartościową dyskusję:

A. Czy można podać przykłady konkretnych zastosowań praktycznych, w których warto byłoby wykorzystać naraz wszystkie metody wprowadzone w omawianych pracach nr 1, 2, 3, 4?
B. Czy istotnie (to znaczy czy nie jestem w błędzie uważając, że) rezultaty algorytmów z pracy nr 4, a więc ustrukturalizowane reprezentacje informacji wyekstrahowanych z oryginalnych dokumentów, mogą stanowić wejście do systemów raportowania i eksploracji danych?
C. Czy jest możliwe uwzględnianie w opracowanych metodach informacji zwrotnej o błędach? Czy np. błędy (albo podejrzenia błędów) wartości wykryte w rezultatach otrzymanych w formie tabelarycznej mogą być wykorzystywane do poprawy metod te rezultaty tworzących?
D. Jakie jest zdanie Autorki na temat powstających teraz niezwykle dynamicznie dużych modeli językowych (ang.: large language models – LLM)? Czy mogą one wpłynąć na dalszy rozwój metod opracowanych przez Autorkę, w tym na procesy przetwarzania dokumentów?

7. Podsumowanie

W świetle moich uwag zawartych w poprzednich sekcjach, uwzględniając ustawowe wymagania stawiane doktoratom w obszarze informatyki, oceniam rozprawę jednoznacznie pozytywnie. Na moją ocenę mają wpływ następujące czynniki:

- Merytoryczna i edycyjna jakość prac wchodzących w skład rozprawy, w tym wysoka renoma (i punktacja) konferencji, na których prace te zostały opublikowane.
- Przejrzystość rozprawy, które w mojej opinii dowodzi bardzo dobrze ugruntowanej wiedzy Doktorantki, w tym umiejętności uwypuklenia wspólnego mianownika wszystkich omawianych publikacji. (Jak już pisałem na wstępie, jest to tematyka spójna i zarazem różnorodna.)
- Wysoka skuteczność opracowanych metod w konkursach konferencyjnych o tematyce NLP.
- Dwa patenty przyznane w USA na podstawie badań opisanych w rozprawie. Patenty te są dla mnie kluczowe i w związku z nimi **chciałbym zawnioskować o wyróżnienie rozprawy**.

Na koniec chciałbym jeszcze raz skomentować to, że wszystkie prace wchodzące w skład rozprawy są wieloautorskie i że w niektórych przypadkach wkład Kandydatki może nie wydawać się w porównaniu z innymi autorami najważniejszy. Mimo wszystko uważam, że jest to wkład znaczący. W szczególności warto tu zwrócić uwagę na deklaracje autorów publikacji nr 2 – w tym przypadku wkład Doktorantki jest niepodważalnie kluczowy. Podsumowując, **wnoszę o dopuszczenie rozprawy doktorskiej mgr Gabrieli Nowakowskiej do kolejnych etapów postępowania**.

Dominik Ślęzak