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

Poznań, 26 sierpnia 2024 r.

Recenzja rozprawy doktorskiej

mgr inż. Gabrieli Nowakowskiej

Named entity recognition and information extraction from various documents

1 Problem badawczy i jego znaczenie

W ostatnich latach jesteśmy świadkami ogromnego postępu w dziedzinie przetwarzania języka naturalnego (*NLP, Natural Language Processing*), będącego konsekwencją rozwoju nowoczesnych metod uczenia maszynowego, wzrostu mocy obliczeniowej i wolumenu dostępnych w internecie danych. Dzięki narzędziom NLP, inteligentne systemy są w stanie analizować bardzo złożone dokumenty tekstowe i ekstrahować z nich informację niezbędną do wykonania różnorakich zadań, takich jak tłumaczenie, podsumowywanie, odpowiadanie na pytania do tekstu, itp. Wydaje się, że dzięki dużym modelom językowym, takim jak GPT-4o czy Llama 2, posiadającym niezwykle szerokich możliwości rozumienia i generacji języka naturalnego, NLP stanowi najbardziej obiecujący kierunek w rozwoju sztucznej inteligencji.

Recenzowana rozprawa dotyczy dwóch podstawowych zagadnień związanych z tą dziedziną. Jedno z nich to rozpoznawanie jednostek nazwanych (NER, *Named Entity Recognition*), dotyczące identyfikacji i kategoryzacji nazw własnych pojawiających się w tekście, dotyczących osób, organizacji, miejsc, wydarzeń, itp. Rola NER wydaje się kluczowa w przetwarzaniu dokumentów i ekstrakcji informacji, ponieważ umożliwiając identyfikację nazw własnych pozwala na lepsze zrozumienie treści dokumentu, a także ułatwia klasyfikowanie i organizowanie informacji. Drugim zagadnieniem, któremu poświęcona jest rozprawa, jest ekstrakcja informacji (IE, *Information Extraction*), dotycząca wyodrębniania danych i informacji z nieustrukturyzowanych źródeł tekstowych, takich jak dokumenty cyfrowe czy posty w mediach społecznościowych. IE pozwala na automatyzację wielu procesów pozyskiwania danych i ma ogromne znaczenie w wielu dziedzinach przemysłu i biznesu. Oba badane zagadnienia stały się jeszcze ważniejsze w

kontekście zastosowań współczesnych systemów sztucznej inteligencji, a ich znaczenie w nadchodzących latach będzie tylko wzrastać. Stąd temat pracy Doktorantki wydaje się świetnie trafiać we współczesne trendy badawcze i może prowadzić do istotnych praktycznych zastosowań.

2 Struktura pracy i cele badawcze

W porównaniu do wcześniej recenzowanych przeze mnie rozpraw doktorskich, rozprawa mgr inż. G. Nowakowskiej ma specyficzny charakter, ponieważ jest kulminacją doktoratu wdrożeniowego. Prace badawcze prowadzone były w ramach współpracy Uniwersytetu im. Adama Mickiewicza i firmy Applica, pod kierownictwem prof. UAM dr. hab. Tomasza Góreckiego, a ich cel jest z założenia ukierunkowany na zastosowania. Nie zmienia to jednak faktu, że same wyniki rozprawy mają charakter w pełni naukowy i zostały opublikowane w recenzowanych materiałach renomowanych konferencji naukowych.

Rozprawa jest napisana w języku angielskim. Merytoryczny wkład Doktorantki w dyscyplinę Informatyka to cztery powiązane ze sobą tematycznie publikacje, stanowiące treść rozdziałów 2–5. Rozdział pierwszy to wprowadzenie do tematyki rozprawy, opis analizowanych zagadnień, zarys motywacji do podjęcia badań w tym zakresie, zdefiniowanie celów badawczych, a następnie bardziej szczegółowy opis każdej z czterech publikacji. Każdorazowo, Autorka opisała swój wkład w pisanie pracy, udokumentowany oświadczeniami współautorów zawartych w załącznikach. W załącznikach znajdziemy również oficjalne potwierdzenie wysokich wyników zespołu, w skład którego wchodziła Doktorantka, w konkursach związanych z tematyką pracy. Ponieważ struktura pracy jest w zasadzie podporządkowana „publikacyjnemu" charakterowi rozprawy, trudno mieć do niej zastrzeżenia.

Głównym celem pracy był rozwój metod i technologii we wspomnianych już dziedzinach rozpoznawania jednostek nazwanych oraz ekstrakcji informacji. W szczególności, mgr inż. Gabriela Nowakowska uzyskała w rozprawie następujące wyniki:

1. Propozycja nowych i twórcze wykorzystanie istniejących metod rozpoznawania jednostek nazwanych w celu rozwiązania problemu NER problemu zdefiniowanego w ramach konkursu na międzynarodowej konferencji.

2. Utworzenie nowych modeli lematyzacji jednostek nazwanych, również zastosowane do rozwiązania konkursu w ramach warsztatu naukowego.

3. Zaproponowanie modelu TILT, umożliwiającego ekstrakcję informacji z dokumentów o dwuwymiarowej strukturze (warstwa tekstowa oraz wizyjna).

4. Konstrukcja modelu STable, będącego rozwinięciem modelu TILT, umożliwiającego ekstrakcję danych tabelarycznych.

3 Ocena wkładu oryginalnego

Rozprawa jest oparta na czterech artykułach, których współautorką jest Doktoranta:

[1] A. Nowakowski, G. Pałka, K. Guttmann, M. Pokrywka: Adam Mickiewicz University at WMT 2022: NER-Assisted and Quality-Aware Neural Machine Translation. *WMT 2022 (EMNLP 2022)* [Punkty MNiSW: 140]

[2] G. Pałka, A. Nowakowski: Exploring the Use of Foundation Models for Named Entity Recognition and Lemmatization Tasks in Slavic Languages. *Slavic NLP 2023 (EACL 2023)*. [Punkty MNiSW: 140]

[3] 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* [Punkty MNiSW: 140]

[4] M. Pietruszka, M. Turski, Ł. Borchmann, T. Dwojak, G. Nowakowska, K. Szyndler, D. Jurkiewicz, Ł. Garncarek: STable: Table Generation Framework for Encoder-Decoder Models. *EACL 2024* [Punkty MNiSW: 140]

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

Wszystkie prace zostały opublikowane na konferencjach mających 140 punktów MNiSW, co wskazuje na ich wysoką pozycję naukową w dziedzinie przetwarzania języka naturalnego. Sumaryczne 560 punktów jest wynikiem dobrym jak na dorobek w trakcie doktoratu, tym bardziej biorąc pod uwagę wdrożeniowy jego charakter. Nie będę odnosił się tutaj do współczynnika *Impact Factor*, ponieważ materiały konferencyjne najprawdopodobniej go nie posiadają. Prace doczekały się już 177 cytowań (wg. *Google Scholar*, odczytane w sierpniu 2024), co jest osiągnięciem bardzo dobrym z uwagi na nieodległe daty publikacji. Warto zwrócić uwagę na bardzo dużą liczbę cytowań pracy [3] (155 cytowań), mimo, że od jej publikacji minęły zaledwie trzy lata. Z tej perspektywy dorobek uznaję więc za wartościowy.

Aspekt mogący budzić pewne wątpliwości, to znacząca liczba współautorów w pracach [3] i [4], w których Doktorantka nie odgrywała też wiodącej roli. Myślę jednak, że praca w dużych zespołach związana była (a zapewne wręcz nieodzowna) z wdrożeniowym charakterem badań i nie powinna być tutaj traktowana negatywnie, stąd moim zdaniem nie obniża całościowej oceny prac.

Rozprawa zawiera wiele oryginalnych i nowatorskich wyników, które zostały już pozytywnie zweryfikowane na etapie przyjmowania prac na konferencje, oraz poprzez osiągnięcie wysokich wyników w konkursach dziedzinowych.

- W pracy [1] zaprezentowano rozwiązanie zgłoszone na konkurs *WMT 2022 General MT Task*, dotyczące tłumaczenia między językami ukraińskim i czeskim. Zadanie to było trudne z uwagi na licznie występujące w korpusie jednostki nazwane, których tłumaczenie stanowi często duże wyzwanie i może zdecydować o sukcesie bądź nie całego rozwiązania. Autorzy zastosowali w tym celu model z grupy BERT (Slavic BERT) do języka czeskiego, oraz narzędzie Stanza Named Entity Recognition do języka ukraińskiego, nie tylko wykrywając jednostki nazwane, ale również przypisując im „czynniki źródłowej" (tzn. określając, dotyczą osoby, miejsca, organizacji, lub czegoś innego). Pełne rozwiązanie jest bardzo złożone

i robi spore wrażenie: użyto tu metod *transfer learning*, wykorzystując duży korpus tłumaczeń między językami czeskim i angielskim; wprowadzono tzw. *noisy back-translation*, umożliwiające rozszerzenie zbioru danych dodając syntetyczne przykłady powstające z podwójnego tłumaczenia z pośrednim dodaniem szumu do warstwy wyjściowej; użyto czynników źródłowych z zadania NER na dwa sposoby: konkatenując lub sumując ich zanurzenia z zanurzeniami samych tokenów; tłumaczenia na poziomie dokumentów, a nie tylko zdań, wykorzystując coraz dłuższy kontekst; techniki tworzenia zespołów (*ensembling*) modeli uczonych w innych konfiguracjach; pozyskiwania dużej liczby kandydatów tłumaczeń, a następnie wyboru najlepszego poprzez tzw *quality-aware decoding*, składający się z metod *T-RR* oraz *MBR decoding*; wreszcie zastosowanie licznych metod przetwarzania końcowego w celu podniesienia jakości wyników.

Widoczne jest więc duże zaangażowanie autorów, aby osiągnąć cel (wygraną w konkursie), a także ogrom pracy włożony w konstrukcję i analizę eksperymentalną rozwiązania. Przedstawione zostały wyniki nie tylko ostatecznego rozwiązania, ale również wpływ włączania kolejnych wyżej opisanych jego elementów.

Interesujące jest, że użycie metod *quality-aware decoding* przyczyniło się do wzrostu jakości względem miary COMET, natomiast równocześnie spodowało spadek na mierze BLEU. Czy Doktorantka ma pomysł, jaka była tego przyczyna?

- W artykule [2] przedstawiono model użyty w zadaniu lematyzacji jednostek nazwanych w językach słowiańskich (polskim, czeskim i rosyjskim) w ramach konkursu SlavNER. Zadanie to postawiło wiele poważnych wyzwań z uwagi na złożoność języków słowiańskich (bogata fleksja, swobodny szyk zdań, itp.), a także konieczność ekstrakcji jednostek nazwanych na poziomie dokumentów, a nie pojedynczych słów czy fraz. W rozwiązaniu problemu NER użyto szeregu modeli z rodziny BERT: HerBERT dla języka polskiego, Czert dla języka czeskiego, RuBERT dla języka rosyjskiego, dodatkowo porównując jest z modelami Multilingual BERT oraz XML-RoBERTa. Do modeli została na wyjściu dodana warstwa CRF (*Conditional Random Field*). Z kolei do problemu lematyzacji użyto modelu bazującego na architekturze T5, zarówno w wersjach jedno-językowych jak i wielo-językowych. Dodatkowo, w obu problemach użyto zewnętrznych zbiorów danych, pozwalających na znaczące rozszerzenie zbioru treningowego. Autorzy uzyskali bardzo dobre wyniki na danych konkursowych. Przedstawili też wartości metryk uzyskanych przez modele w różnych konfiguracjach.

  W zadaniu lematyzacji dla języka polskiego Autorzy zauważyli, że mniejszy model radził sobie lepiej, niż model bazowy (choć został później pokonany przez duży model). Jest to dość zaskakująca zależności, czy Autorka potrafi wskazać potencjalną przyczynę tego zjawiska?

- W pracy [3], w celu rozwiązania problemu rozumienia dokumentów, wprowadzony został model TILT, który jednocześnie uczy się informacji o układzie dokumentu, cech wizualnych oraz semantyki tekstu. TILT jest więc multimodalnym modelem wykorzystującym architekturę *encoder-decoder*, wprowadzający transformery ze uwzględnieniem informacji przestrzennej poprzez zastosowanie dodatkowego parametru w mechanizmie uwagi,

którzy bierze pod uwagę relatywną przestrzenną pozycję tokenów. W celu reprezentacji wizualnych cech dokumentu wykorzystano też sieć o architekturze *U-Net*. Zanurzenia otrzymywane z tej sieci były dodawane do zanurzeń wynikających z informacji semantycznej. Podczas uczenia użyto szeregu technik regularyzacji, opartych na augmentancji danych.

Całościowy model został nauczony bardzo dużym korpusie dokumentów, a następnie dostrojony na docelowych zadaniach. Autorom udało się na dwóch zadaniach uzyskać wyniki przewyższające pod względem trafności najlepsze istniejące rozwiązania.

Od publikacji pracy minęły już trzy lata, a w międzyczasie pojawiły się duże modele językowe, takie jak GPT-4o. Nasuwa się tu więc oczywiście pytanie, jakie wyniki mogłyby zostać osiągnięte przez taki model na zadaniach rozwiązywanych przez Autorów?

- Publikacja [4] proponuje nowe podejście, nazwane STable, do generowania tabel w ramach modeli typu *encoder-decoder*. Oparte jest na interesującym paradygmacie dekodowania opartego na permutacjach, które pozwala na generowanie danych tabelarycznych w dowolnej kolejności. Na etapie trenowanie sekwencja danych tabelarycznych dla każdego przykładu jest losowo permutowana, co uniezależnia algorytm od jakiejś specyficznej kolejności wypełniania tabeli. Z kolei na etapie predykcji algorytm w sposób zachłanny wypełnia kolejne elementy tabeli na podstawie zwracanych prawdopodobieństw.

  Podejście to eliminuje wiele problemów związanych z metodami sekwencyjnymi (autoregresywnymi), takich jak nawarstwianie się błędów z pierwszych komórek tabeli czy ograniczenia wynikające z sztywnej kolejności generowania, co zostało potwierdzone w eksperymentach obliczeniowych. Ciekawym pomysłem jest również mechanizm uwagi, wykorzystujące dodatkowe nauczalne funkcje „odległosci" mierzone po wierszach i kolumnach tabeli.

  Wyniki eksperymentów wskazują, że STable osiąga znaczącą poprawę wyników w porównaniu z metodą „liniową" (w określonej z góry kolejności) na części użytych zbiorów danych. W szczególności, na zestawie danych PWC autorzy uzyskali znaczący wzrost jakości, uzyskując wynik *state-of-the-art*. STable okazał się również bardzo poprawiać wyniki (nad „oryginalną" wersją TILT, na której był oparty) w zadaniach związanych z ekstrakcją informacji z rachunków, wyciągów bankowych i odcinków wypłat. Autorzy wykonali również tzw. *ablation studies*, w których starają się określić, które elementy nowego algorytmu są odpowiedzialne za wzrost jakości.

  Czy Doktorantka rozważała zmianę heurystyki wypełniania komórek tabeli na wersję, która jest mniej zachłanna, np. używając metody *beam search*? Oczywiście jeszcze bardziej skomplikuje to proces predykcji, ale interesujące wydawałoby się sprawdzenie, czy tego typu metody mogą jeszcze poprawić wyniki.

  Ciekawe wydaje mi się również studium z dodatku C, w który rozważane są inne kryteria wypełnienia, oceny i wyboru komórek. Autorzy eksperymentują tu z dość zaskakującymi miarami, np. aby wypełniać komórki używając *najmniejszej* wartości z modelu stowarzyszonej z tokenami („min"), a później wybierać komórkę o najmniejszej wartości spośród

kandydatów. Podejście to wydaje się zupełnie nieintuicyjne, a mimo to obniża tylko nieznacznie jakość modelu. Jakie mogą być tego przyczyny?

Uważam, że każda z wymienionych publikacji daje istotny wkład do dziedziny przetwarzania język naturalnego. Uznaję stąd, że wymieniony na wstępie cel pracy cele udało się Doktorantce osiągnąć.

4 Konkluzja końcowa

Recenzowaną rozprawę oceniam bardzo dobrze. Załączone prace potwierdzają wiedzę Doktorantki w zakresie przetwarzania języka naturalnego, a w szczególności rozpoznawania jednostek nazwanych i ekstrakcji informacji, oraz umiejętność prowadzenia prac badawczych. Problemy, z którymi zmierzyła się mgr inż. Gabriela Nowakowska, są ambitne i istotne dla postępu w dziedzinie. Wyniki konkursów, w których brał udział zespół Autorów prac, potwierdzają wysoką jakość zaproponowanych rozwiązań. Stwierdzam, że *przedstawiona do oceny Rozprawa Doktorska spełnia warunki określone w Art. 187 Ustawy z dnia 20 lipca 2018 r./Prawo o szkolnictwie wyższym i nauce /(Dz.U. z 2022 r. poz. 574 z późn. zm.), oraz uzasadnia nadanie stopnia naukowego doktora w dyscyplinie informatyka.*

W związku z tym **rozprawę oceniam jako spełniającą wymogi stawiane pracom doktorskim i wnoszę o dopuszczenie mgr inż. Gabrieli Nowakowskiej do dalszych etapów postępowania w sprawie nadania stopnia doktora**.

Dr hab. inż. Wojciech Kotłowski

PODPIS ZAUFANY
WOJCIECH
KOTŁOWSKI
17.09.2024 17:30:38 [GMT+2]
Dokument podpisany elektronicznie
podpisem zaufanym