dr hab. Maciej Ogrodniczuk, prof. IPI PAN

Instytut Podstaw Informatyki
Polskiej Akademii Nauk

Jana Kazimierza 5
01-248 Warszawa

e-mail: maciej.ogrodniczuk@ipipan.waw.pl
tel. 533 675 675

29 lipca 2024 r.

Recenzja rozprawy doktorskiej

Gabrieli Nowakowskiej

zatytułowanej:

Named entity recognition and information extraction from various documents

1. Problem badawczy i jego znaczenie

Opisane w rozprawie prace dotyczą zastosowania metod przetwarzania języka naturalnego w kilku zadaniach: tłumaczenia maszynowego z wykorzystaniem modelu rozpoznawania nazw własnych, samodzielnego rozpoznawania i lematyzacji nazw własnych w wielu językach oraz ekstrakcji informacji z warstwy tekstowej i wizualnej dokumentów, dodatkowo w wersji wzbogaconej o dane tabelaryczne.

Zadania te są niezwykle istotne pod względem praktycznym, czyli także w kontekście wdrożeniowym firmy specjalizującej się wówczas w tworzeniu systemów do analizy i zarządzania dokumentami cyfrowymi. Efektywne rozpoznawanie nazw własnych i ekstrakcja informacji pozwalają na automatyzację wielu złożonych zadań, takich jak identyfikacja trendów, odkrywanie powiązań międzytekstowych, tworzenie i aktualizowanie specjalistycznych baz wiedzy, streszczanie wielodokumentowe czy tworzenie zaawansowanych systemów wyszukiwania. W kontekście naukowym prace wpisują się w badania szeroko znanej szkoły tłumaczenia maszynowego oraz rozwiązywania problemów biznesowych metodami eksperckimi Centrum Sztucznej Inteligencji UAM.

2. Zawartość pracy

Praca przedłożona do oceny ma postać zbioru czterech jednotematycznych artykułów w języku angielskim opublikowanych w recenzowanych materiałach z warsztatów organizowanych przy największych międzynarodowych konferencjach z dziedziny NLP (dwa pierwsze artykuły – WMT na EMNLP 2022 i Slavic NLP na EACL 2023) oraz jako long papers na samych konferencjach ICDAR 2021 i EACL 2024 (artykuł trzeci i czwarty).

Pierwszy artykuł opisuje prace nad systemem tłumaczenia maszynowego między językiem ukraińskim a czeskim z uwzględnieniem tłumaczenia nazw własnych. Powstałe rozwiązanie wykorzystuje model

Slavic BERT i moduł Stanzy do oznaczenia nazw własnych odpowiednio w tekście czeskim i ukraińskim, obliczenia dla nich wektorów zanurzeń, a następnie włączeniu tak przetworzonej informacji o nazwach własnych w proces tłumaczenia poprzez dodanie wektorów nazw własnych do wektorów słów w tłumaczonych tekstach. Zastosowana technika doprowadziła do istotnej poprawy ogólnej jakości tłumaczenia (w miarach chrF i BLEU).

Drugi artykuł przedstawia nowe modele rozpoznawania nazw własnych i lematyzacji dla języka polskiego, czeskiego i rosyjskiego. Autorzy wykorzystali jedno- i wielojęzyczne modele BERT w zadaniu rozpoznawania nazw własnych oraz modele T5 do lematyzacji. Przeprowadzili również eksperyment z wykorzystaniem tłumaczenia maszynowego danych uczących z istniejących leksykonów polskich nazw własnych na języki czeski i rosyjski, choć ich użycie nie poprawiło wyników dla tych języków. Przedstawione rozwiązanie osiągnęło najlepszy wynik w zadaniu lematyzacji i było konkurencyjne w zadaniu rozpoznawania nazw własnych.

Trzeci artykuł opisuje prace nad modelem Text-Image-Layout Transformer (TILT) powstałym w ramach prac nad ekstrakcją informacji z warstwy tekstowej i wizualnej przetwarzanych dokumentów. TILT, oparty na architekturze Transformer i T5, wspomaga wyszukiwanie informacji łącząc dane tekstowe, graficzne i określające strukturę dokumentu. Uczenie modelu obejmuje trzy etapy: pre-trening nienadzorowany, trening nadzorowany na różnych zadaniach oraz dostrajanie do konkretnego zadania. Skuteczność systemu TILT potwierdzono w konkursie DocVQA, gdzie model zajął pierwsze miejsce w zadaniu odpowiadania na pytania dotyczące infografik, osiągając wynik znacząco wyższy niż system wicelidera.

Czwarty artykuł prezentuje model STable będący rozwinięciem modelu TILT i służący do ekstrakcji danych tabelarycznych z dokumentów. Proponowane rozwiązanie wykorzystuje kompleksowe podejście do modeli neuronowych, z dekoderem opartym na permutacjach. Model tworzy strukturę tabeli w odpowiednim formacie na podstawie zapytania i wybiera najbardziej prawdopodobne ścieżki generowania komórek. Proces ekstrakcji informacji obejmuje przewidywanie liczby wierszy, generowanie potencjalnych wartości dla każdej komórki i iteracyjne wypełnianie tabeli. Model osiągnął wyniki przewyższające lub porównywalne z najlepszymi dostępnymi systemami tego rodzaju na standardowych zestawach danych testowych.

3. Poprawność rozwiązania

Zaproponowane rozwiązania i analizy oraz prezentacja wyników są zgodne z regułami sztuki, co dodatkowo potwierdza fakt ich przyjęcia na cenione konferencje i warsztaty oraz udział (z powodzeniem) w zadaniach ewaluacyjnych. Kontekst wdrożeniowy rozwiązań oraz uzyskane patenty gwarantują ich gotowość do wykorzystania w praktyce.

Nie sposób jednak mimo wszystko nie zauważyć, że mimo poprawności przedstawionych rozwiązań praca rozpada się na dwie luźno powiązane części, chociaż zawiera tak ściśle powiązanie zadania, jak rozpoznawanie nazw własnych i ekstrakcja informacji. Zabrakło mi spajającego je elementu – najlepiej w postaci dodatkowego artykułu, a może choćby wyjaśnienia tej zależności we wprowadzeniu. Skoro rozpoznawanie nazw własnych w tradycyjnym podejściu stanowiło pierwszy krok w procesie ekstrakcji informacji, zakładam, że taką rolę odgrywa właśnie pośrednio w pierwszym artykule dot. tłumaczenia maszynowego. Wydaje się też, że nazwy własne pomagają w strukturyzacji danych, która jest kluczowym aspektem ekstrakcji informacji. Znów, echo tego stwierdzenia pobrzmiewa w artykule czwartym, chociaż oczywiście metody tej strukturyzacji są dziś inne i powiązania relacji czy jednostek są zawarte bezpośrednio w sieci neuronowej.

4. Wiedza i wkład kandydatki

W związku z formą pracy, stanowiącej serię artykułów, istniejący stan wiedzy omawiany jest po części w każdym z nich. Odwołując się do ówczesnego stanu wiedzy Kandydatka potwierdza bardzo dobrą orientację i stan wiedzy w zakresie informatyki.

Pierwszy artykuł wchodzący w skład pracy został zaprezentowany na najlepszym warsztacie dot. tłumaczenia maszynowego (WMT 22), a opisywany w nim system praktycznie zwyciężył w konkursie ewaluacyjnym (w tłumaczeniu z czeskiego na ukraiński przegrał wyłącznie z jednym komercyjnym systemem dostępnym online). Drugi artykuł opisuje system, który w zadaniu rozpoznawania nazw własnych na warsztacie Slavic NLP (na konferencji EACL 2023) zajął drugie, a w zadaniu ich normalizacji – pierwsze miejsce (na trzy zgłoszenia). Wiedza i umiejętności Kandydatki zostały zatem wnikliwie zweryfikowane zarówno przez warunki konkursowe, jak i przez recenzentów opisów tych systemów. Podobnie należy traktować pozostałe prace, recenzowane na najlepszych konferencjach z dziedziny przetwarzania języka naturalnego (ICDAR 2021 i EACL 2024).

Zgodnie z deklaracją Kandydatki jej wkład w dwie pierwsze prace należy ocenić jako kluczowy, obejmujący zarówno pomysł, jak i implementację rozwiązania, prowadzenie eksperymentów oraz autorstwo artykułu. W przypadku dwóch kolejnych tekstów jej odpowiedzialność jest już mniejsza i ogranicza się do przeglądu i przygotowania danych, prowadzenia eksperymentów i redakcji tekstu („Review and preparation of the datasets, running experiments. editing the manuscript"), co potwierdzają oświadczenia współautorów. Zastanawiający jest fakt, że w artykule trzecim przy pięciu z sześciu jego autorów pojawia się dopisek „Contributed equally", z którego wyłączona jest właśnie Kandydatka. W związku z tym udział Kandydatki w ich powstaniu uważam za znacząco mniejszy niż głównych autorów.

Wątpliwość tę musiałem rozwiać kontaktując się z jednym z głównych autorów prac dotyczących ekstrakcji informacji i uzyskałem jego potwierdzenie, że w przypadku modelu TILT Kandydatka była zaangażowana w przegląd i przygotowanie zbiorów danych, które można potencjalnie wykorzystać do etapu treningu nadzorowanego modelu oraz eksperymenty na niektórych zbiorach ewaluacyjnych (RVL-CDIP). Nie była zaangażowana w samą architekturę, zaś dodatkowy wkład związany z TILT-em miała w późniejszą aplikację modelu na zadaniu InfographicsVQA w ramach konkursu przy konferencji ICDAR. W przypadku STable wykonywała zaś część eksperymentów oraz była zaangażowana w prace wokół zbiorów danych, ich przygotowywania, wyboru oraz część implementacyjną ich obsługi (wczytywanie i reprezentacja danych dla modelu). W obu przypadkach, mimo że jej wkład był mniejszy niż „głównych" autorów, nie był wyłącznie techniczny, ale także naukowy.

Na osobną uwagę zasługuje natomiast udział Kandydatki w uzyskaniu patentów związanych z modelami TILT i STable.

5. Inne uwagi

Jeśli chodzi o drobne błędy w samej treści, ograniczają się one wyłącznie do części wprowadzających. Moją uwagę zwróciły trzy kwestie, oczywiście bez znaczenia dla oceny całej pracy:

1. wyrażenie „Our priority was to develop the first open model for lemmatization of the Polish language based on the T5 architecture." (p. 8) – chodzi zapewne o lematyzację nazw własnych, bo ogólne modele lematyzacji dla polszczyzny istnieją już od jakiegoś czasu,

2. link do repozytorium https://huggingface.co/amu-cai – powinien wskazywać konkretny model, a nie ogólne repozytorium Centrum Sztucznej Inteligencji,
3. w streszczeniu przywołana jest konferencja ICDAR 2019, podczas gdy certyfikat dotyczy konferencji ICDAR 2021.

Jeśli chodzi o punktację poszczególnych artykułów, nadmienię dla porządku, że w mojej opinii materiały z warsztatów (dwa pierwsze artykuły) nie są niestety punktowane, ale oczywiście to także kwestia drugorzędna, zwłaszcza w kontekście planowanych zmian w ewaluacji jednostek.

6. Podsumowanie

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) moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

A. Czy rozprawa zawiera oryginalne rozwiązanie problemu naukowego? (wybierz jedną opcję stawiając znak **X**)

☐	☒	☐	☐	☐
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

B. Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka (w dziedzinie nauk ścisłych i przyrodniczych)?

☒	☐	☐	☐	☐
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

C. Czy kandydat posiada umiejętność samodzielnego prowadzenia pracy naukowej?

☐	☒	☐	☐	☐
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

Maciej Ogrodnicki

Podpis