Dr hab. Małgorzata Marciniak, prof. IPI PAN
Instytut Podstaw Informatyki PAN
ul. Jana Kazimierza 5
01-248 Warszawa

Warszawa, 16.09.2024

RECENZJA
rozprawy doktorskiej mgr. Michała Turskiego pt.
Utilizing Structured Resources in Neural Language Models

Rozprawa doktorska mgr. Michała Turskiego pod tytułem *Utilizing Structured Resources in Neural Language Models* składa się z cyklu pięciu artykułów. Są to artykuły wieloautorskie, a sama rozprawa ma charakter wdrożeniowy. Zdefiniowanym przez Doktoranta tematem przewodnim artykułów jest wykorzystanie informacji o strukturze dokumentów do poprawy rozumienia tekstów. Z tekstów o złożonej strukturze graficznej ekstrahowane są informacje będące odpowiedziami na interesujące użytkownika pytania. Tak sformułowane zadanie jest niezwykle złożone. Wymaga opracowania oprogramowania pozwalającego na właściwą, automatyczną interpretację struktury tekstu (w tym tabel i rysunków), wydobycie z owej struktury treści pozwalającej na automatyczne analizowanie znaczenia, oraz ustalenia odpowiedzi na postawione pytania. Do realizacji zadania wykorzystano neuronowe modele języka, których poprawne działanie uzależnione jest od istnienia danych treningowych. Spora część pracy poświęcona jest więc opracowaniu danych pozwalających wytrenować model języka i ocenić jego działanie.

Na wstępie wymienię artykuły wchodzące w skład rozprawy doktorskiej i bardzo skrótowo je omówię. W dalszej części będę odnosić się do nich według podanej numeracji. W nawiasach umieściłam liczbę cytowań według Google Scholar.

1. *DUE: End-to-End Document Understanding Benchmark*; Ł. Borchmann, M. Pietruszka, T. Stanisławek, D. Jurkiewicz, **M. Turski**, K. Szyndler oraz F. Graliński (48)

   Artykuł prezentuje wzorcowy zbiór danych (ang. benchmark)[1], który służy do oceny efektywności działania systemów rozumienia dokumentów. Zadaniem doktoranta było przygotowanie podzbioru diagnostycznego, który ułatwia analizę wyników oraz pozwala na identyfikację słabych miejsc w opracowywanych rozwiązaniach. Doktorant uczestniczył też w nadzorowaniu i ocenie anotacji ręcznej. Artykuł nie zawiera przykładów anotacji, co utrudnia ocenę tego procesu. Próby obejrzenia przykładowych danych, które powinny się znajdować pod adresem podanym w abstrakcie https://duebenchmark.com, nie powiodły się, gdyż adres nie jest aktywny. Doktorant powinien w tekście wprowadzającym wskazać nowy adres lub poinformować o braku możliwości dostępu do opisywanego zasobu. Należy podkreślić, że artykuł ma zaledwie 3 lata i szkoda, że zasób prezentowany w przedstawionej do oceny pracy jest niedostępny, przestał więc spełniać rolę benchmarku.

2. *Arxiv Tables: Document Understanding Challenge Linking Texts and Tables*; K. Konopka, **M. Turski**, F. Graliński (0)

   Artykuł poświęcony jest przetwarzaniu danych tabelarycznych znajdujących się w zasobach umożliwiających skorelowanie źródeł tekstu z wynikiem graficznym. Teksty pochodzą z zasobów

[1] Autor dość niefortunnie nazywa ten zbiór „wyzwaniem".

Arxiv. Metoda wykorzystuje możliwość czerpania informacji o danych zawartych w tabelach na podstawie tekstów źródłowych w LaTex. W artykule podany jest adres zasobu: https://gonito.net/challenge/arxivtables, niestety próba użycia powoduje następujący błąd: *Podczas łączenia z serwerem „gonito.net" wystąpił błąd*. Należy dodać, że artykuł jest pracą magisterską pierwszej autorki.

3. *CCpdf: Building a High Quality Corpus for Visually Rich Documents from Web Crawl Data*; **M. Turski**, T. Stanisławek, K. Kaczmarek, P. Dyda, F. Graliński (5)

   Jest to artykuł poświęcony opracowaniu korpusu wielojęzycznych tekstów o bogatej strukturze. Wobec problemów z prawami autorskimi, autorzy udostępniają korpus w postaci internetowych linków oraz skryptów je obsługujących. Każdy zainteresowany sam może zebrać wówczas dane. Taka forma udostępniania korpusów wydaje się dość zawodna, gdyż zasoby internetowe są mało stabilne. Najlepszym przykładem są dwa niedziałające linki, podane w dwóch pierwszych artykułach. Opisana procedura zbierania tekstów do korpusu jest typowa, choć niewątpliwie pracochłonna.

4. *LAMBERT: Layout-Aware Language Modeling for Information Extraction*; Ł. Garncarek, R. Powalski, T. Stanisławek , B. Topolski, P. Halama , **M. Turski**, F. Graliński (117)

   Artykuł przedstawia nowe podejście do problemu rozumienia dokumentów o złożonej strukturze. Metoda opiera się na zmodyfikowanej strukturze modeli transformacyjnych, która wykorzystuje dane uzupełnione o współrzędne informujące o miejscu tokenu na stronie. Jest to metoda pozwalająca zredukować nakłady związane z uczeniem. Zaproponowane rozwiązanie okazało się konkurencyjne do innych podejść. Dodatkowo jest proste w implementacji. Artykuł ten został doceniony przez organizatorów konferencji ICDAR 2021 i otrzymał nagrodę za najlepszy artykuł związany z przemysłem.

5. *STable Table Generation Framework for Encoder-Decoder Models* M. Pietruszka, **M. Turski**, Ł. Borchmann, T. Dwojak, G. Pałka, K. Szyndler, D. Jurkiewicz, Ł. Garncarek (12)

   W artykule zaproponowana jest metoda generowania tabel na podstawie tekstu. Autorzy zaproponowali metodę pozwalającą na skuteczniejsze generowanie poprawnych tabel, która opiera się na wyliczaniu prawdopodobieństwa popełnienia błędu i wypełnianiu tych pozycji, które wydają się najbardziej pewne. Autorzy twierdzą, że metoda pozwoliła na poprawę wyników aż o 15%. Zaproponowane rozwiązanie wydaje się bardzo przydatne w praktycznych zastosowaniach automatycznego prezentowania wyników.

Doktorant pełnił rolę koordynatora projektu w przypadku artykułów nr 3 i 5, co jest warte odnotowania w recenzji i wskazuje na umiejętność prowadzenia badań naukowych.

Ocena merytoryczna artykułów

Przedstawione do oceny artykuły zostały opublikowane na uznanych i dobrze punktowanych konferencjach. Przeszły wnikliwy proces recenzyjny, co sprawia, że zostały zweryfikowane przez społeczność naukową, nie zamierzam więc poddawać ich kolejnej recenzji. Dwa spośród artykułów zyskały w krótkim czasie sporą liczbę cytowań. Zawartość merytoryczną artykułów oceniam wysoko.

Struktura pracy

We wprowadzeniu Doktorant niezbyt trafnie dzieli artykuły na dwie grupy. Pierwsze dwa artykuły zyskały niezrozumiały dla mnie nadtytuł *Measuring state of document understanding*. Sugeruje on

prowadzenie pomiarów, choć np. w artykule nr 2 słowa pokrewne z *measure* pojawiają się jedynie w kontekście rozpoznawania jednostek miary w tekście. Następne trzy artykuły zyskały nadtytuł *2D-Structured Language Modeling*, choć nie do końca rozumiem kryteria, które sprawiły, że znalazły się w jednej grupie tematycznej.

Właściwe przedstawienie tematyki łączącej artykuły w cykl wymaga większego wysiłku włożonego w opracowanie tekstu towarzyszącego artykułom. Aktualny tekst sprawia, że tematyka przewodnia cyklu nie została wystarczająco precyzyjnie przedstawiona czytelnikowi.

Wkład

Wkład pracy Doktoranta nie jest łatwy do zidentyfikowania na podstawie dołączonych deklaracji.

W artykule nr 1 Doktorant deklaruje:

- „methodology and preparation of the diagnostic subset" co jest częściowo zbieżne z deklaracją p. Tomasza Stanisławka „methodology for creation of the diagnostic subset". Na podstawie artykułu nie umiem ustalić na czym polega przygotowanie tych podzbiorów i na ile się pokrywa z deklaracją Michała Pietruszki „preparation of datasets ... (tu wymienione)", przygotowanie innego zbioru deklaruje też p. Tomasz Stanisławek.
- „organizing and controlling the process of human annotation" deklarują panowie: Łukasz Borchmann, Michał Pietruszka, Tomasz Stanisławek oraz Doktorant.

W artykule nr 3 Doktorant i p. Tomasz Stanisławek zajmowali się „design and implementation of the tool for managing the corpus".

W artykule nr 5:

- Nie wiem czym się różni zadanie „preparation of domain-specific pre-training datasets" wykonane przez Doktoranta od szerszego wydawałoby się zadania „review and preparation of the datasets" zadeklarowanego przez p. Nowakowską.
- "baselines implementation" jest zadeklarowana przez Doktoranta oraz czterech innych współautorów.
- Zadanie „ablation studies" zadeklarowane przez Doktoranta opisane jest również w zadaniach wykonanych przez Ł. Borchmanna jako „designing and conducting ablation studies".
- Niemal wszyscy autorzy zadeklarowali zadanie „running experiments".

Deklaracje wkładu poszczególnych autorów zawierają dwie wyraźnie różne czynności: „writing" oraz „editing". W przypadku dwóch artykułów (pierwszego i czwartego) Doktorant deklaruje jedynie ich edycję. Proszę podać przyczynę takiego stanu rzeczy.

Złożoność zadania oraz charakter wdrożeniowy pracy, w wyniku której powstało efektywnie funkcjonujące oprogramowanie, wymaga dużego zespołu współpracujących osób. Wyniki prac przedstawione są więc w artykułach napisanych przez wielu autorów. Wiem z doświadczenia, że w takim przypadku, nawet samym autorom jest czasem trudno precyzyjnie ustalić wkład poszczególnych osób. Nie zwalnia to jednak Doktoranta z precyzyjnego zdefiniowania swojego osiągnięcia przedstawionego do oceny w pracy doktorskiej. Oczekuję, że w trakcie obrony Doktorant dokładnie wyjaśni te kwestie.

Język pracy

Streszczenie pracy jest jedynym fragmentem tekstu rozprawy napisanym po polsku. Wydaje się ono automatycznie przetłumaczonym tekstem angielskiego streszczenia. Doktorant w tych kilku zdaniach

wykazał się daleko posuniętym niedbalstwem użycia języka polskiego, a zakładam, że jest to jego język ojczysty. Poniżej przykład: „Niniejsza rozprawa składa się z pięciu prac naukowych w domenie rozumienia dokumentów i jest podzielona na dwie główne sekcje. Pierwsza sekcja dotyka problemu oceny modeli rozumienia dokumentów, wprowadzając pierwsze wyzwanie (ang. benchmark) w tej domenie ..."

Kilka uwag szczegółowych:

- Przyjęło się, że informację o stanowisku profesora w instytucji naukowej umieszcza się po nazwisku, a nie przed jak to ma miejsce w przypadku strony tytułowej pracy.
- Razi mnie sformułowanie zawarte w opisie osiągnięć autora „Firstly, my paper titled LAMBERT..." w kontekście artykułu nr 4 podpisanego przez siedmiu autorów, dla którego Doktorant określił swój wkład w tworzenie tekstu słowem „editing".
- Abstrakt artykułu nr 5 zaczyna się od niezrozumiałego zdania: „Since the output structure of database-like tables can cover a wide range of NLP tasks, we propose a framework for text-to-table neural models applicable to, e.g...".

Konkluzja

Przedstawiona recenzja zawiera wiele uwag krytycznych odnoszących się do konstrukcji pracy i jej formy. W recenzji wskazałam wątpliwości związane z wyodrębnieniem wkładu Doktoranta w artykuły wchodzące w skład cyklu. Bardzo skromne wprowadzenie do cyklu nie pomaga w ustaleniu tego wkładu. Sformułowanie konkluzji było więc dla mnie niezmiernie trudne. Finalnie uznałam jednak, że wdrożeniowa praca doktorska rządzi się nieco odmiennymi regułami i postanowiłam dać szansę Doktorantowi na odniesienie się do wątpliwości sformułowanych w recenzji. Przedstawione artykuły stanowią bowiem istotny wkład merytoryczny w rozwój dziedziny. Doktorant wykazał się umiejętnością kierowania pracami skutkującymi powstaniem dwóch z pięciu artykułów. Jednym z nich jest artykuł nr 5, dla którego uważam, że przedstawiony pomysł i osiągnięta poprawa w rozwiązywaniu omawianego zadania jest imponująca. Ostatecznie postanowiłam więc ocenić pozytywnie pracę Doktoranta, licząc równocześnie na wyjaśnienie podniesionych w recenzji kwestii w trakcie obrony. Postuluję by dopuścić Pana mgr. Michała Turskiego do dalszych etapów postępowania, w tym publicznej obrony.

Małgorzata Marciniak