Recenzja pracy doktorskiej mgr. inż. Łukasza Borchmanna „Span Identification and Key Information Extraction Beyond Sequence Labeling Paradigm"

Krzysztof Jassem

29 stycznia 2022

1 Wstęp

Celem niniejszej recenzji jest stwierdzenie, czy rozprawa doktorska mgr. inż. Łukasza Borchmanna zatytułowana „Span Identification and Key Information Extraction Beyond Sequence Labeling Paradigm" spełnia wymagania ustawowe (Art.187. Ustawy „Prawo o szkolnictwie wyższym i nauce"). Ustawa stwierdza w punkcie 3., że „Rozprawę doktorską może stanowić ... zbiór opublikowanych i powiązanych tematycznie artykułów naukowych". Właśnie ta forma prezentacji wyników została wybrana przez Doktoranta.

Doktorant przedstawia do recenzji zbiór ośmiu artykułów:

1. Łukasz Borchmann, Andrzej Gretkowski, Filip Graliński, „Approaching nested named entity recognition with parallel LSTM-CRFs"

   Artykuł dotyczy zagadnienia rozpoznawania zagnieżdżonych jednostek nazewniczych w tekstach języka polskiego. Przedstawia rozwiązanie autorskie, które zajęło pierwsze miejsce w konkursie organizowanym w ramach warsztatu PolEval 2018.

2. Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, Filip Graliński, „On RoBERTa-CRF, Span CLS and Whether Self-Training Helps Them"

   Artykuł omawia ekstrakcję fragmentów stanowiących teksty propagandowe. Opisuje rozwiązanie, które wzięło udział w konkursie organizowanym w ramach międzynarodowej konferencji Sem-Eval 2020, zajmując pierwsze miejsce w kategorii rozpoznawania zastosowanej metody propagandy i drugie miejsce w kategorii wykrywania lokalizacji fragmentów propagandowych w tekście. Artykuł został wyróżniony na tej konferencji nagrodą *Best Paper Award.*

3. Łukasz Borchmann, Dawid Wiśniewski, Andrzej Gretkowski, Izabela Kosmala, Dawid Jurkiewicz, Łukasz Szałkiewicz, Gabriela Pałka, Karol Kaczmarek, Agnieszka Kaliska, Filip Graliński, „Contract Discovery: Dataset and a Few-Shot Semantic Retrieval Challenge with Competitive Baselines"

   Artykuł dotyczy wyszukiwania w tekstach prawniczych klauzul określonego typu. Zadanie to nastawione jest na spełnienie konkretnej potrzeby biznesowej: redaktorzy tekstów prawniczych chcą mieć szybki dostęp do fragmentów zawartych umów, które dotyczą zagadnień pokrewnych. Autorzy artykułu przygotowali oznaczony zestaw danych, który umożliwił im wytrenowanie

rozwiązania bazowego. Następnie przygotowane przez siebie dane upublicznili, umożliwiając w ten sposób opracowanie rozwiązań wyższej jakości niż proponowane w artykule.

4. Łukasz Borchmann , Dawid Jurkiewicz , Filip Graliński, Tomasz Górecki, „Dynamic Boundary Time Warping for sub-sequence matching with few examples"

   Artykuł prezentuje zastosowanie algorytmu opracowanego do analizy podobieństwa szeregów czasowych w ekstrakcji informacji z tekstów. Rozwiązanie poszerza standardowy scenariusz wyszukiwania: wejściem do procesu może być nie jedno zapytanie, lecz zbiór zapytań.

5. Michał Pietruszka, Łukasz Borchmann, Łukasz Garncarek „Sparsifying Transformer Models with Trainable Representation Pooling"

   Artykuł wprowadza modyfikację architektury sieci neuronowej typu „transformer". Metoda polega na wczesnej selekcji informacji, która przetwarzana jest w kolejnych warstwach sieci. Usprawnienie to znacząco wpływa na zmniejszenie złożoności algorytmu i przekłada się na zwiększenie wydajności czasowej i pamięciowej. Autorzy wykazują zasadność zastosowania swojego rozwiązania w zadaniu sumaryzacji tekstów. Proponowane usprawnienie może mieć zastosowanie w innych zadaniach związanych z ekstrakcją informacji.

6. Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, Gabriela Pałka, „Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer"

   Artykuł wprowadza zastosowanie architektury „transformer", w którym przetwarzane są jednocześnie warstwa tekstowa i wizualna dokumentów. Autorzy wykazują, że wprowadzone przez nich usprawnienie poprawia jakość metod dla zadań takich jak ekstrakcja kluczowych informacji z tekstu czy odpowiadanie na pytania.

7. Tomasz Dwojak, Michał Pietruszka, Łukasz Borchmann, Jakub Chłedowski, Filip Graliński, „From Dataset Recycling to Multi-Property Extraction and Beyond"

   W artykule poruszone jest zagadnienie jednoczesnej ekstrakcji wielu informacji z dokumentu. Autorzy proponują metodę, która polega między innymi na bardziej starannym przygotowaniu danych treningowych. Na bazie istniejącego zbioru danych tekstowych o nazwie *WikiReading* opracowują zestaw *WikiReading Recycled* i wykazują, że wytrenowany na nim mechanizm ekstrakcji działa skuteczniej niż wytrenowany na zbiorze oryginalnym.

8. Łukasz Borchmann, Michał Pietruszka, Tomasz Stanisławek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, Filip Graliński, „DUE: End-to-End Document Understanding Benchmark"

   Artykuł wprowadza nowy zestaw danych dla zadań związanych z rozumieniem dokumentów, takich jak:

   - ekstrakcja informacji kluczowych z dokumentu,
   - klasyfikacja tematyczna dokumentu,

- analiza układu dokumentu,
- odpowiadanie na pytania na podstawie informacji zawartych w dokumencie,
- wnioskowanie na podstawie informacji zawartych w dokumencie.

Autorzy twierdzą, że istniejące zbiory danych opracowywane są z myślą o zastosowaniu tylko w jednym z powyższych zadań. Z tego powodu przygotowali zbiór danych, o nazwie DUE, przeznaczony dla wielu zadań rozumienia dokumentów łącznie. Autorzy przetestowali skuteczność dostępnych metod przetwarzania dokumentów, wskazując rozwiązanie, które w momencie publikacji osiągało najwyższą skuteczność.

2 Problem badawczy

Problemem badawczym, który przewija się we wszystkich artykułach zawartych w rozprawie, jest ekstrakcja informacji z dokumentów tekstowych. W poszczególnych artykułach rozważane są różne zadania związane z tym zagadnieniem. Autor podzielił te zadania na trzy grupy:

1. oznaczanie sekwencji tokenów,
2. identyfikacja zakresu poszukiwanej informacji,
3. wyszukiwanie informacji kluczowych.

Do pierwszej grupy Autor zaliczył artykuły 1. i 2., do drugiej – artykuły 3. i 4. i 5., a do trzeciej – artykuły 6, 7, 8. (artykuł 2. porusza zadania zarówno z grupy pierwszej, jak i z drugiej). Nie ulega wątpliwości, żeńń recenzowany zestaw jest zbiorem powiązanych tematycznie artykułów naukowych, spełniając w ten sposób podstawowy wymóg Ustawy.

Wyniki prezentowane w rozprawie są istotnym wkładem w rozwój dziedziny. Omawiane algorytmy i zbiory danych zostały udostępnione publiczne, umożliwiając odtworzenie eksperymentów i zweryfikowanie raportowanych danych. Metody badawcze prezentowane w rozprawie w sposób znaczący przyczyniają się do światowego postępu w dziedzinie ekstrakcji informacji z dokumentów.

Niezwykle istotne jest znaczenie praktyczne przeprowadzonych badań. Łukasz Borchmann jest zatrudniony w firmie Applica i jego wyniki badawcze wdrażane są w działalności tej firmy. Mam przekonanie, że badania raportowane w rozprawie motywowane były konkretnymi potrzebami biznesowymi.

3 Wkład autora

Wszystkie artykuły wchodzące w skład recenzowanej rozprawy są publikacjami zbiorowymi. Fakt ten może utrudniać ocenę wymagań stawianych w punktach 1. i 2. art. 187: „Rozprawa doktorska

prezentuje ...umiejętność **samodzielnego** prowadzenia pracy naukowej..." oraz „Przedmiotem pracy doktorskiej jest ...oryginalne nie w zakresie zastosowania **własnych** badań naukowych."

Ocenę recenzji prac zbiorowych powinny ułatwiać oświadczenia współautorów. Rozprawa zawiera podpisane oświadczenia wszystkich współautorów wszystkich artykułów. Niestety, w oświadczeniach tych zadania wykonywane przez poszczególnych autorów nakładają się na siebie. Na przykład w artykule 8. zadanie „conceptualization and methodology" wykonane zostało przez trzech autorów (na marginesie uważam, że sformułowanie „conceptualization" w pracy, w której nie wprowadza się nowych pojęć, jest niefortunne).

W ocenie indywidualnego wkładu Doktoranta przyjąłem założenie, że jest on głównym pomysłodawcą koncepcji przedstawionych w tych pracach, w których jest on pierwszym autorem. Spośród ośmiu artykułów Łukasz Borchmann wymieniony jest jako pierwszy autor w czterech (1., 3., 4. i 8.).

Moim zdaniem artykułem przełomowym w karierze naukowej Łukasza Borchmanna był artykuł nr 3. Wykazał on bowiem, że można z powodzeniem zastosować współczesne metody uczenia maszynowego do bardzo konkretnego zadania motywowanego potrzebami biznesowymi. Kolejne artykuły potwierdziły tę tezę.

4 Poprawność

Dla wszystkich eksperymentów opisywanych w pracy przeprowadzono automatyczną ewaluację. Miała ona na celu wykazanie poprawności zastosowanych metod poprzez porównanie ich skuteczności z innymi rozwiązaniami na poziomie światowego state-of-the-art. W przypadku prac 1. i 2. ewaluacja została przeprowadzona w środowisku zewnętrznym opracowanym przez organizatorów konkursów PolEval i SemEval. Wysoka jakość wyników została więc potwierdzona obiektywnie i bezspornie.

W pozostałych eksperymentach autorzy sami wykazywali skuteczność swoich rozwiązań. Przeprowadzali eksperymenty albo na zbiorach danych, które zostały wcześniej przygotowane przez innych badaczy, albo na przygotowanych przez siebie. W pierwszym przypadku oceniali skuteczność swoich rozwiązań za pomocą tych samych metryk, które były stosowane w rozwiązaniach konkurencyjnych. W przypadku eksperymentów przeprowadzanych na własnych zbiorach trenujących autorzy udostępniali te dane innym badaczom, umożliwiając w ten sposób odtworzenie swoich wyników.

W moim przekonaniu wszystkie eksperymenty zostały przeprowadzone poprawnie, zgodnie z przyjętymi na świecie standardami obowiązującymi dla zadań uczenia maszynowego. Dzięki temu można z przekonaniem stwierdzić, że rezultaty przedstawione w rozprawie są godne zaufania.

Rozprawa napisana jest w języku angielskim. Jest to wybór oczywisty, jeśli chodzi o artykuły, które z założenia przeznaczone były na konferencje międzynarodowe. Wydaje mi się jednak, że wprowadzenie autora lepiej czytałoby się w języku polskim. Być może pisząc w języku ojczystym, autor uniknąłby tendencji do tworzenia długich zdań, które utrudniają lekturę (już pierwsze zdanie

wstępu składa się z 32 wyrazów).

W rozprawie bardzo trudno doszukać się ewidentnych błędów językowych lub merytorycznych. Podpis do tabeli 4.2. na stronie 43 zawiera referencję do dodatku C, zamiast do dodatku A.

5 Wiedza kandydata

W każdym z recenzowanych artykułów bezspornie wykazano znajomość najnowszych badań. Omówione są one w sekcjach zatytułowanych „Overview of Existing Database", „Related Works", itp. O znajomości współczesnych badań świadczy również sposób przeprowadzania eksperymentów – wyniki autorów porównywane są ze współczesnymi eksperymentami z dziedziny ekstrakcji informacji. Bibliografia każdego artykułu jest bogata i zawiera w większości publikacje z kilku ostatnich lat. Z całym przekonaniem stwierdzam, że kandydat ma ogólną wiedzę w dziedzinie informatyka techniczna i telekomunikacja.

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:

6.1 Czy rozprawa zawiera oryginalne rozwiązanie problemu naukowego?

Zdecydowanie TAK

6.2 Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka techniczna i telekomunikacja?

Zdecydowanie TAK

6.3 Czy kandydat posiada umiejętność samodzielnego prowadzenia pracy naukowej?

Zdecydowanie TAK

6.4 Rekomendacja wyróżnienia

Ponadto, rekomenduję wyróżnienie rozprawy doktorskiej, motywując to następująco:

- Łączna liczba artykułów stanowiących rozprawę (8) jest wysoka.

- Większość artykułów została przyjęta na renomowane konferencje międzynarodowe z dziedziny ekstrakcji informacji.
- Wyniki przeprowadzanych eksperymentów wykazywały wyższość nad innymi współczesnymi rozwiązaniami w skali światowej.
- Skuteczność opracowanych metod badawczych została pozytywnie zweryfikowana w praktyce gospodarczej – poprzez wdrożenia w firmie Applica.

Podpis