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

11 marca 2022 r.

Recenzja rozprawy doktorskiej

Łukasza Borchmanna

zatytułowanej:

Span Identification and Key Information Extraction
Beyond Sequence Labeling Paradigm

1. Problem badawczy i jego znaczenie

Opisane w rozprawie prace dotyczą metod efektywnej ekstrakcji informacji z tekstu, czy to w postaci wyłącznie oznaczenia fragmentu tekstu, czy przypisania danemu fragmentowi jakiejś etykiety, czy wreszcie powiązania danego klucza z wyekstrahowaną z tekstu wartością. Rozważany problem dotyczy z jednej strony możliwości zastąpienia tradycyjnych metod znakowania ciągu (ang. *sequence labelling*) metodami działającymi na fragmentach tekstu, co okazuje się ważne np. w przypadku niedostępności anotacji na poziomie tokenów czy błędów w analizowanym tekście, a z drugiej – wykorzystania do tego celu modeli koder-dekoder, które potencjalnie rozwiązują wszystkie wymienione rodzaje problemów, dodatkowo eliminując potrzebę specjalnego traktowania zagnieżdżonych wartości.

Oprócz dość oczywistego znaczenia praktycznego problem efektywnej ekstrakcji informacji z tekstu ma także charakter naukowy. Zarówno zmieniające się w ciągu ostatnich lat metody realizacji tego zadania, rozwijające się modele języka i powstające na potrzeby tego zadania zbiory danych wymagają tworzenia coraz to nowych metod z dziedziny przetwarzania języka naturalnego – ale także wykraczających poza tę dziedzinę.

2. Wkład autora

Praca przedłożona do oceny ma postać zbioru ośmiu jednotematycznych artykułów w języku angielskim opublikowanych w recenzowanych materiałach z konferencji lub warsztatów o zasięgu międzynarodowym oraz w czasopiśmie.

Dwa pierwsze artykuły (z rozdziałów 2 i 3) stosują metodę znakowania ciągu w zadaniach ekstrakcyjnych – wykrywania nazw własnych oraz technik propagandowych. Artykuł z rozdziału 4 podejmuje problem ekstrakcji klauzul prawnych o podobnym znaczeniu, proponując dostrojenie modelu neuronowego. Artykuł z rozdziału 5 proponuje ogólny algorytm znajdowania sekwencji na bazie przykładów, również stosując go do zadania wyszukiwania podobnych klauzul prawnych oraz wykrywania nazw własnych. W artykule z rozdziału 6 Autor kontynuuje temat wykrywania istotnych

fragmentów tekstu, tym razem na potrzeby algorytmu automatycznego streszczania. Proponuje metodę integrującą model abstraktywny z ekstrakcyjnym, w którym strata z modelu generatywnego propagowana jest do modelu bazującego na identyfikacji fragmentów tekstu (ang. *span identitication*).

Rozdziały 7 i 8 zajmują się problemem ekstrakcji kluczowych informacji (ang. *key information extraction*) w zadaniach wykrywania wartości dla danego klucza lub pytania. Artykuł z rozdziału 7 dotyczy kwestii ekstrakcji danych z dokumentów o bogatej strukturze graficznej, zaś z rozdziału 8 – ekstrakcji wielu par klucz–wartość z Wikipedii.

Rozdział 9 stanowi swoiste podsumowanie pracy, reformułując zadanie ekstrakcji informacji jako zadanie całościowego rozumienia dokumentu (ang. *document understanding*), z uwzględnieniem jego warstwy tekstowej, graficznej i strukturalnej oraz biorąc pod uwagę różny charakter zadań ekstrakcyjnych.

Wkład autora we wszystkie prace należy ocenić jako wielowymiarowy. Nie ogranicza się on wyłącznie do proponowania rozwiązań wybranego problemu czy zestawu problemów metodami informatycznymi, lecz proponuje całościowy system przetwarzania dokumentów, zajmując się w poszczególnych częściach pracy jego różnymi aspektami – od tworzenia zbiorów ewaluacyjnych, poprzez opracowywanie kryteriów oceny rozwiązań, aż po implementację nowatorskich algorytmów. Przykładami tego rodzaju wkładu są m.in. doceniona przez recenzentów SemEvalu metoda połączenia modeli transformerowych z metodami CRF czy opisana w rozdziale 6 jednoetapowa metoda łącząca krok ekstrakcji i generowania tekstu, znacząco różna od proponowanych wcześniej metod dwuetapowych, z osobno trenowanymi modułami ekstraktywnym i generacyjnym.

Zgodnie z deklaracjami Autora oraz wkład w poszczególne publikacje składowe dotyczy tworzenia koncepcji i metodologii rozwiązania (we wszystkich pracach), ponadto prowadzenia eksperymentów i analizy ich wyników (w większości prac), implementacji modeli i rozwiązań bazowych. Kandydat jest pierwszym autorem czterech z ośmiu publikacji, zaś w pozostałych jego udział w tworzeniu prac został oznaczony jako równy z pierwszym autorem lub należy go ocenić jako znaczący na podstawie zadeklarowanego wkładu.

3. Poprawność

Stwierdzenia zawarte w rozprawie, ich uzasadnienia, zaproponowane rozwiązania i prezentacja wyników są zgodne z regułami sztuki i uwzględniają niezwykle cenne (i dziś coraz częściej wymagane na najlepszych konferencjach) elementy, takie jak analiza wpływu najważniejszych komponentów użytych modeli na jego działanie (ang. *ablation study*). Analiza błędów jest wnikliwa i zawiera jasne wnioski dla przyszłych użytkowników zaproponowanych rozwiązań.

Każda z prac składowych zawiera pożyteczną dyskusję zaproponowanego rozwiązania. Np. w rozdziale 2 Autor analizuje m.in. przetestowane, a nie włączone do tekstu rozwiązania (z wektorami FastText, czy algorytmem Layered-LSTM-CRF). W rozdziale 3 analizuje błędy predykcji i proponuje usprawnienia, także ze wskazaniem odwołań do istniejących prac, które mogłyby pomóc w poprawie uzyskanego wyniku.

Zaproponowane rozwiązania są niezwykle praktyczne i są gotowe do wykorzystania (a, biorąc pod uwagę karierę zawodową Autora, zapewne również wykorzystane) w rzeczywistych zastosowaniach z zakresu automatyzacji takich procesów biznesowych jak wykrywanie technik propagandowych i wyszukiwanie istotnych informacji w treści skanów czy dużych zbiorach tekstów internetowych. Na uwagę zasługuje udostępnienie tworzonych modeli i zbiorów danych w publicznych repozytoriach w celu umożliwienia ich dalszego rozwoju przez kolejnych badaczy.

4. Wiedza kandydata

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

Większość artykułów wchodzących w skład pracy zostało opublikowanych na jednych z najlepszych konferencji z dziedziny przetwarzania języka naturalnego, a przedstawione rozwiązania wygrywały konkursy ewaluacyjne w Polsce i za granicą. Wiedza i umiejętności Kandydata zostały zatem wnikliwie zweryfikowane już przez recenzentów tych prac. Na uwagę w tym kontekście zasługuje szczególnie nagroda Best Paper Award uzyskana przez pracę z rozdziału 3 na warsztacie SemEval 2020 na konferencji COLING za udane połączenie technik neuronowych z tradycyjnymi modelami i metodami uczenia maszynowego. Warto tu zauważyć, że wyróżniony w ten sposób artykuł był jedną z ok. 300 prac (!), gdyż nagroda dotyczyła wszystkich 12 zadań ewaluacyjnych z 2020 roku.

Bibliografia jest kompletna, z jedynie drobnymi mankamentami typograficzno-redakcyjnymi. Części z nich sam nie wykryłem nawet podczas pracy nad redagowanym przeze mnie tomem z materiałami konkursu PolEval 2018, w którym została zamieszczona praca z rozdziału 2.

Stwierdzam zatem, że kandydat posiada stosowną wiedzę w dyscyplinie Informatyka techniczna i telekomunikacja.

5. Inne uwagi

Streszczenie i rozdział 1, czyli przedmowa podsumowująca pracę, zawierają nieliczne literówki i wpadki, często zabawne w kontekście tematu i układu pracy, jak „neutral networks" czy różna liczba artykułów włączonych do rozprawy podana w różnych częściach wprowadzenia. Są to jednak kwestie zupełnie bez znaczenia.

Jedyne, na co może warto zwrócić uwagę, raczej w kontekście potrzeby ustalenia wspólnego stanowiska polskiego środowiska NLP niż zarzutu do Autora, to kwestia do podanej we wprowadzeniu punktacji poszczególnych artykułów (p. s. 7, pozycje w tabeli odpowiadające rozdziałom 3 i 4). Materiały z warsztatów, takich jak *Workshop on Semantic Evaluation* czy zamieszczone w *Findings of the ACL* w mojej opinii i zgodnie z obecną wykładnią nie są niestety punktowane.

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**)

[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 techniczna i telekomunikacja?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

C. Czy kandydat posiada umiejętność samodzielnego prowadzenia pracy naukowej?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

Ponadto, biorąc pod uwagę wysoki poziom artykułów wchodzących w skład pracy, z których jeden został opublikowany w czasopiśmie 140-punktowym, dwa na konferencjach 140-punktowych, jeden na konferencji 200-punktowej, a kolejny otrzymał nagrodę Best Paper Award na niepunktowanym, ale bardzo prestiżowym warsztacie po zadaniu SemEval, rekomenduję wyróżnienie rozprawy doktorskiej.

Maciej Ogrodniczuk

Podpis