Warszawa, 1.02.2022

dr hab. Agnieszka Mykowiecka
Instytut Podstaw Informatyki PAN
Jana Kazimierza 5, Warszawa

Recenzja rozprawy doktorskiej mgr. inż. Tomasza Stanisławka

Ekstrakcja informacji z dokumentów o bogatej strukturze graficznej

Recenzja wykonana jest na zlecenie Rady Naukowej dyscypliny informatyka techniczna i telekomunikacja Politechniki Warszawskiej. Praca doktorska Tomasza Stanisławka zrealizowana została pod opieką dr hab. inż. Przemysława Biecka na Wydziale Matematyki i Nauk Informacyjnych i składa się ze zbioru czterech jednotematycznych artykułów w języku angielskim opublikowanych w recenzowanych materiałach z konferencji o zasięgu międzynarodowym. Wszystkie konferencje należą do najbardziej cenionych w zakresie analizy dokumentów lub ogólniej neuronowych metod rozwiązywania problemów: ICDAR (A), NeurIPS (kategoria A*), oraz CONLL (A). Tematem pracy jest wydobywanie informacji z dokumentów, które nie składają się jedynie z ciągłego tekstu, ale mają złożoną strukturę graficzną. Prace prowadzone były w trakcie studiów doktoranckich, ale miały także bezpośrednie powiązanie z możliwością wdrożenia modułu stanowiącego rozwiązania badanego problemu w firmie Applica i podniesieniem jakości proponowanych przez nią narzędzi.

Rozprawa ma formę dokumentu, który opisuje cel i wyniki prowadzonych badań, pełni zatem rolę autoreferatu. Załącznikami są arykuły stanowiące główną treść rozprawy. Rozdział pierwszy wprowadza czytelnika do tematyki rozprawy, przedstawia opis problemu, motywację podjęcia badań oraz cel pracy. Rozdział drugi zawiera prezentację głównych wyników rozprawy i jest podzielony na cztery podrozdziały, z których każdy odnosi się do jednego z artykułów stanowiących treść rozprawy. Rozdział trzeci obejmuje prezentację dorobku naukowego doktoranta, a rozdział czwarty zawiera podsumowanie wyników i wkładu doktoranta w rozwój dziedziny.

Artykuły stanowiące główną treść pracy, zawarte w załącznikach, to:

1. Stanislawek, T., A. Wróblewska, A. Wójcicka, D. Ziembicki i P. Biecek: **Named Entity Recognition - Is There a Glass Ceiling?** Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), 2019
2. Stanisławek, T., F. Graliński, A. Wróblewska, D. Lipiński, A. Kaliska, P. Rosalska, B. Topolski i P. Biecek, **Kleister: Key Information Extraction Datasets Involving Long Documents with Complex Layouts**, In International Conference on Document Analysis and Recognition (ICDAR) 2021, pp. 564-579
3. Borchmann, Ł., M. Pietruszka, T. Stanislawek, D. Jurkiewicz, M. Turski, K. Szyndler i F. Graliński: **DUE: End-to-End Document Understanding Benchmark**, Thirty-fifth Conference on Neural Information Processing Systems Datasets, 2021
4. Garncarek, Ł., R. Powalski, T. Stanisławek, B. Topolski, P. Halama, M. Turski, ... **LAMBERT: Layout-Aware Language Modeling for Information Extraction** International Conference on Document Analysis and Recognition, ICDAR 2021, pp. 532-547 (wyróżnienie w kategorii *Best Industry Related Paper Award*)

Sformułowanym w rozprawie celem pracy było zaproponowanie własnego mechanizmu poprawiającego skuteczność ekstrakcji informacji z dokumentów o bogatej strukturze graficznej. Realizacja tego zadania wymagała realizacji także celów pomocniczych takich jak

zbadanie możliwości istniejących metod oraz konceptualizacja problemów, jakie występują w tej dziedzinie.

Treść pracy

Artykuł [1], opisany w podrozdziale 2.1, przedstawia badania, w których poddano ocenie istniejące metody rozpoznawania nazw własnych (NER) w tekstach. Temat NER był bardzo często podejmowany, opracowane metody są często stosowane także do ekstrakcji informacji, a wyniki są łatwe do oceny ze względu na istniejące zbiory tekstów z oznaczeniami typów nazw i stosunkowo proste reguły anotacji. Wnioski wyciągnięte na podstawie analizy tych rozwiązań mogą być zatem bardzo przydatne także przy rozwiązywaniu zadania ekstrakcji informacji.

W pracy zanalizowano rozwiązania wykorzystujące różne, udostępnione publicznie, metody identyfikacji nazw własnych i ich typów, począwszy od najskuteczniejszej niegdyś metody CRF (rozwiązanie zaproponowane przez Stanford) poprzez sieci LSTM z warstwą CRF (CMU), sieci BiLSTM w modelu ELMO, do modelu języka typu BERT opartego na transformerach oraz modelu Flair. Wszystkie modele zostały przez autorów użyte do oznaczenia danych angielskich upowszechnionych na konferencji CoNLL w 2003 roku. Uzyskane wyniki zostały zanalizowane, ustalono klasyfikację popełnianych przez systemy błędów, a następnie przydzielono wszystkie wykryte błędy do odpowiednich kategorii. W wynikach analizy widać, że duża liczba błędów wynika z tego, że do prawidłowego przypisania typu nazwy potrzebna jest wiedza nie tylko z bezpośredniego kontekstu, ale z całego zdania lub całego dokumentu. Dlatego najwięcej błędów stwierdzono w wynikach systemu wykorzystującego tylko CRF, a wyniki ELMO okazały się lepsze niż z BERTa. Naturalnym wnioskiem jest zatem konieczność uzupełniania danych wejściowych do modelu NER o informacje dotyczące miejsca konkretnego fragmentu w całym dokumencie. Istotnym elementem prac było też zwrócenie uwagi na błędy anotacji w opublikowanych danych testowych. W przypadku niewielkich różnic w efektywności porównywanych systemów błędy te mogą zaburzyć ocenę. Przeprowadzone eksperymenty wykazały, że wyniki na poprawionych zbiorach testowych nie różniły się zbytnio od tych uzyskanych na danych z błędami (były nieco lepsze), ale kolejność wyników poszczególnych modeli czasami ulegała zmianie. Widać zatem, że jakość danych testowych jest istotna i na pewno tym większa im mniejszym zbiorem testowym dysponujemy.

Artykuł [2] dotyczy kolejnego kroku koniecznego przy opracowywaniu metod maszynowego uczenia – zebrania odpowiednich danych treningowych i testowych. Z uwagi na bardzo niewielką dostępność danych zawierających obrazy całych sformatowanych dokumentów, a nie tylko sam tekst, przygotowano dwa takie zbiory danych angielskich. Pierwszy zawiera umowy o zachowaniu poufności, które pochodzą z bazy EDGAR. Drugi, sprawozdania roczne fundacji charytatywnych (https://register-of-charities.charitycommission.gov.uk/). Ponieważ zbiory te miały przypisane informacje na poziomie dokumentów opracowano zestaw reguł – wyrażeń regularnych – dzięki którym przypisanie to zostało przeniesione na poziom fragmentów tekstu. Poza stworzeniem zbioru danych, w artykule przedstawiono też wyniki osiągnięte dla tych danych przez różne modele wytrenowane do zadania ekstrakcji informacji (FLAIR, BERT, RoBerta, LayoutLM i Lambert. Przetestowano także procent zgodności anotacji dwóch osób na próbce 100 dokumentów, który okazał się bardzo wysoki (ponad 97%). Zgodnie z oczekiwaniami modele, które uwzględniają jakiś sposób pozycję w tekście radziły sobie lepiej z wyznaczonym zadaniem. Wciąż jednak najlepsze osiągane wyniki były znacznie poniżej zgodności anotatorów (ok.85%).

W artykule [3] zaprezentowany został benchmark opracowany z myślą o wszystkich zadaniach związanych z przetwarzaniem dokumentów o bogatej strukturze graficznej. Autorzy przejrzeli ponad 30 zbiorów danych, z których wybrano 7 spełniających kryteria najwyższej jakości (tylko anotacja manualna), trudności (duża różnica między najlepszym rozwiązaniem a poziomem trudności dla człowieka) oraz dostępności. Uwzględnione zbiory odpowiadają różnorodnym zadaniom od ekstrakcji informacji do zadawania pytań do treści znajdujących się w tabelach czy infografikach.

W ostatnim artykule [4] zaproponowano opracowaną przy dużym udziale doktoranta architekturę modelu do ekstrakcji informacji uwzględniającą strukturę dokumentu. Architektura ta oparta jest na strukturze bardzo popularnego modelu neuronowego BERT uzupełnionej o dodatkowe wejście opisujące położenie segmentów na stronie. Zmodyfikowano wagi atencji poprzez dodanie używanego w modelu T5 relatywnego kodowania pozycji oraz rozszerzono to kodowanie o dwa dodatkowe parametry, związane z relatywną pozycją dwóch segmentów względem odległości od siebie w poziomie oraz w pionie. Przeprowadzona seria eksperymentów wykazała, że model ten osiągnął wyniki lepsze (w granicach kilku procent) niż model bazowy RoBERTa oraz model LayoutLM . Zgodnie z często obserwowanymi zależnościami, im większe i lepsze dane (odfiltrowane z niskiej jakości dokumentów) oraz dłuższy trening, tym model LAMBERT jest skuteczniejszy.

Ocena

Przedstawiony cykl artykułów dobrze opisuje typową drogę, jaką należy przebyć próbując opracować nowe rozwiązanie jednego z zadań NLP. Krok pierwszy to analiza już znanych rozwiązań tego, lub pokrewnych zadań, następnie wybór bądź konstrukcja danych treningowych i testowych, a potem próba znalezienia rozwiązania, które pozwoli na zmniejszenie liczby błędów w sytuacjach, z którymi dotychczas istniejące systemy sobie nie radziły. Osiągniecie dobrego wyniku związane musi być na ogół z identyfikacją powodów dla których dotychczasowe rezultaty nie są wystarczające. Doktorant, wspólnie ze współautorami, zrealizowal wszystkie te etapy dochodząc do rozwiązania osiągającego bardzo wysoką skuteczność. W przedstawionej rozprawie cenne jest, że przeprowadzono w tym przypadku dokładną analizę i klasyfikację popełnianych błędów, a nie tylko porównywanie ogólnych wyników takich jak miara F1 dla poszczególnych kategorii. Doktorant, wraz ze współpracownikami, poświęcił też sporo uwagi zebraniu i ujednoliceniu wielu zbiorów treningowych dla zbliżonych do rozwiązywanego w pracy zadań. A przy szybko rozwijającej się dziedzinie NLP to właśnie zbiory danych stanowią często trwalszy wkład w jej rozwój, niż szybko zmieniające się modele neuronowe. Trochę szkoda, że autor nie pokusił się o zebranie choć jednego takiego zbioru z danymi dla języka polskiego.

Konstrukcja doktoratu ze zbioru artykułów zwykle stanowi pewne wyzwanie. Przedstawienie prac, które maja wielu autorów zawsze budzi pewne wątpliwości co do tego, na ile Doktorant uczestniczył w prezentowanych badaniach i jakie idee pochodzą od niego, a jakie od współautorów. Uważam jednak, że odnajdywanie się w grupie badawczej osiągającej dobre wyniki i potwierdzony oświadczeniami wkład Doktoranta w prace wystarczająco dowodzą jego osobistych umiejętności i osiągnięć. Stworzony przez Doktoranta opis towarzyszący artykułom stanowi dobry przewodnik po wykonanych pracach i we właściwy sposób podkreśla wyznaczony cel rozprawy. Tekst jest dość dobrze napisany, drobne literówki czy pomyłki są nieliczne. Jego niedostatkiem, wynikającym jednak z obranej formy rozprawy, jest to, że zawiera on mniej informacji niż załączone artykuły, a chciałoby się by zawierał ich więcej, tak by przeprowadzane eksperymenty, konwersje danych, alternatywne rozwiązania opisane były dokładniej niż można to zrobić w krótkim artykule konferencyjnym. Odnosząc

się zaś do samej rozprawy w wybranej formie – nie do końca przekonuje mnie uporządkowanie artykułów – umieszczenie artykułu dotyczącego modelu LAMBERT po zawierających wyniki testów tego modelu. Zapewne jest to wynik prowadzenia równolegle badań nad samym modelem i próbami jego weryfikacji, a zatem z trudnością w uszeregowaniu prac (trzy z nich opublikowane zostały w roku 2021).

Temat ekstrakcji informacji z danych sformatowanych jest ważny, gdyż bardzo często mamy do czynienia z dokumentami, w których istnieje co najmniej narzucona struktura poszczególnych części, a nawet są one wprost formularzami. Operowanie tylko na czystym tekście powoduje zatem brak możliwości odwoływania się do informacji, która może znacząco wpłynąć na jakość wyniku. Doktorant przedstawił w rozprawie zarówno analizy wskazujące na potrzebę dodawania informacji o pozycji w tekście jak i model osiągający dzięki jej uwzględnieniu lepsze wyniki. Zaakceptowanie proponowanych rozwiązań na wiodących konferencjach z tej dziedziny i nagroda dla jednej z publikacji świadczą o tym, że zostały one uznane za wartościowe przez społeczność międzynarodową.

Zbiór paru odrębnych artykułów oczywiście ma drobne wady związane ze spójnością. Jeśli patrzymy na niego jako na całość, to zauważamy, że w artykule [3] używanymi modelami są T5 i T5+2D i żadnym za artykułów nie ma komentarza o ewentualnych porównaniach (być może nie wprost) zaproponowanego modelu LAMBERT i T5+2D. Jeśli zaś popatrzymy na dane, to trochę brak dyskusji na temat znaczenia pozycji etykietowanych segmentów w stosunku nie tyle do strony co do innych tekstów, takich jak treść stałych elementów formularzy, zwłaszcza w przypadku gdy wielkość poszczególnych pól jest zmienna, a zatem informacje te nie zawsze znajdują się w tym samym miejscu dokumentu. W szczególności dotyczy to sytuacji, w których zbiór typów obiektów, które chcemy wykrywać w dokumentach jest liczniejszy i bardziej różnorodny. Być może gorsze wyniki uzyskiwane dla analizy tekstów długich są z tym jakoś powiązane. W pracy brak sugestii jakie mogą być przyczyny tego faktu, a jego zbadanie byłoby na pewno ciekawe. Oczywiście te ogólne uwagi nie umniejszają mojej pozytywnej oceny opisanych w pracy osiągnięć przy budowie modelu uwzględniającego pozycję analizowanej informacji w tekście, który w pewnym momencie był jednym z najlepszych proponowanych dla tego zadania na świecie.

Wniosek końcowy

Stwierdzam, iż przedłożona mi do recenzji rozprawa, której autorem jest mgr Tomasz Stanisławek, zawiera ważne osiągnięcia w dziedzinie konstruowania modeli neuronowych do rozwiązywania zadań NLP, w szczególności ekstrakcji informacji ze sformatowanych tekstów. Doktorant wykazał się sporą wiedzą w tematyce rozprawy oraz znajomością metod badawczych. Przedstawiony zestaw artykułów cechuje spójność tematyczna i wyraźnie nakreślony kierunek w dochodzeniu do realizacji wyznaczonego celu badawczego. Recenzowana praca spełnia wymagania ustawowo stawiane rozprawom doktorskim, zatem wnoszę o dopuszczenie magistra Tomasza Stanisławka do publicznej obrony.

Agnieszka Wykowska