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 marca 2024 r.

Recenzja rozprawy doktorskiej

Tomasza Ziętkiewicza

zatytułowanej

Zaprojektowanie oraz implementacja systemu automatycznej korekcji błędów i normalizacji wyjścia z systemu rozpoznawania mowy

1. Problem badawczy i jego znaczenie

Celem rozprawy, zrealizowanej jako doktorat wdrożeniowy we współpracy pomiędzy Uniwersytetem im. Adama Mickiewicza a firmą Samsung Electronics Poland, było opracowanie efektywnej i kontrolowalnej metody korekty błędów systemu rozpoznawania mowy oraz opracowanie metod normalizacji wyjścia z systemu rozpoznawania mowy (ograniczonych do przywracania znaków interpunkcyjnych).

Zaproponowany problem badawczy jest ważny zarówno w kontekście konkretnego wdrożenia w firmie współpracującej z uczelnią przy realizacji badań, jak i z punktu widzenia użytkownika systemów rozpoznawania mowy, coraz powszechniejszych w naszej codzienności. Z tego względu cel badań uznaję za wystarczająco szczegółowo sprecyzowany, a zadanie przedstawione przez Doktorantem odpowiednie dla doktoratu wdrożeniowego.

2. Zawartość pracy

Praca przedłożona do oceny napisana jest w języku polskim i ma układ klasyczny. Na 119 stronach Autor zawarł wprowadzenie, rozdział opisujący wdrożeniowe uwarunkowania badań, rozdział prezentujący stosowane metody uczenia maszynowego oraz ewaluacji, trzy rozdziały opisujące badania własne, rozdział podsumowujący, przykłady danych, bibliografię oraz spis rysunków i tabel.

Rozdział 1 zarysowuje motywację Autora, przedstawia cel i zakres pracy, opisuje metodę badawczą i przedstawia znaczenie rozprawy i jej strukturę. Rozdział 2 prezentuje specyfikę rozwoju systemów dialogowych oraz ograniczenia związane z pracą w środowisku produkcyjnym. Rozdział 3 opisuje użyte modele oraz metody i miary ewaluacyjne. Rozdziały 4–6, zawierające opis zasadniczych prac Autora, mają zbliżoną strukturę – prezentują przegląd literatury i istniejących rozwiązań, po którym następuje opis opracowanego rozwiązania, użytych danych, przeprowadzonych eksperymentów, zostają podane wyniki i wnioski. Rozdział podsumowujący, niezwykle krótki, przedstawia także kilka pomysłów na kontynuację prac.

3. Zasadnicza treść pracy

Rozdział 1, oprócz wymienionych już elementów wprowadzających i porządkujących opis, zawiera trzy istotne informacje dotyczące udziału Autora w powiązanych tematycznie z treścią pracy wyzwaniach PolEval 2020 i 2021 oraz organizacji zbliżonego wyzwania Center for Artificial Intelligence Challenge on Conversational AI Correctness (CAICCAIC). Jest to ciekawe uzupełnienie wdrożenia opracowanej metody w systemach dialogowych firmy Samsung o dodatkowy, stricte naukowy, wymiar ewaluacyjny.

Rozdział 2 także należy uznać za wprowadzający; Autor prezentuje w nim zwyczajową architekturę systemów dialogowych umożliwiających interakcję użytkownika z komputerem w języku naturalnym: dekoder wejścia, moduł analizy języka naturalnego, menedżera dialogu, generowania języka naturalnego oraz kodowania wyjścia. Sytuuje prowadzone badania w zakresie modułów korekty i normalizacji oraz wyjaśnia sposób ich wdrożenia w systemach produkcyjnych w procesie wytwórstwa oprogramowania. Zgodnie z tym kontekstem prace obejmują zbieranie danych, cykle trenowania, testowania oraz wdrożenia. Autor zwraca też uwagę na ograniczenia pozafunkcjonalne, dotyczące zasobów dyskowych i pamięciowych używanych przez modele oraz parametry środowiska uruchomieniowego.

Rozdział 3 prezentuje kilka modeli uczenia maszynowego oraz metod ewaluacyjnych używanych na dalszych etapach prac. Modele te, wykorzystywane do etykietowania sekwencji, to regułowy tager Brilla, modele probabilistyczne Conditional Random Field (CRF), neuronowe modele BERT i XLM-RoBERTa oraz biblioteka Flair. Opis metod ewaluacyjnych rozpoczyna zagadnienie normalizacji danych w celu uniknięcia wpływu nieistotnych różnic między porównywanymi ciągami. Autor przedstawia następnie popularne miary oceny modeli ASR (ang. Automatic Speech Recognition), takie jak Word Error Rate (WER), Word Recognition Rate (WRR), Sentence Error Rate (SER) i Sentence Recognition Rate (SRR) oraz miary oceny zadania wykorzystującego system rozpoznawania mowy, takie jak trafność klasyfikacji domeny, trafność klasyfikacji zamiaru, miara F1 czy ogólna efektywność modelu w danym zadaniu (EMA, ang. Exact Match Accuracy).

Rozdział 4 (s. 35–65), stanowiący zasadniczą część pracy, rozpoczyna uzasadnienie wykorzystania w prowadzonych pracach metody przetwarzania końcowego (ang. *post-processing*), po którym następuje przegląd literatury i przedstawienie działającego na tej zasadzie autorskiego rozwiązania „Otaguj i popraw”. Działa ono w dwóch krokach: poprzez oznaczenie tokenów etykietami umożliwiającymi ich ew. korektę oraz następnie poprawienie błędnych tokenów na podstawie przypisanych im etykiet operacji edycyjnych. Autor omawia również szczegółowo sposób generowania referencyjnych operacji edycyjnych na podstawie danych uczących oraz przetestowane modele tagowania. Przedstawia także sposób aplikowania operacji edycyjnych umożliwiający precyzyjną kontrolę nad procesem korekty błędów. Rozdział przedstawia także użyte zbiory danych oraz wyniki eksperymentów i wnioski z ewaluacji na danych firmy Sansung oraz ewaluacji zewnętrznej w ramach wyzwań „Open Challenge for Correcting Errors of Speech Recognition Systems” 2019 i „Post-editing and rescoring of automatic speech recognition results” w konkursie PolEval 2020. Najważniejsze w kontekście prac wdrożeniowych eksperymenty na danych firmy Samsung uwzględniały po 2 modele tagowania (sieć neuronowa typu BERT oraz tager z biblioteki Flair) dla języka niemieckiego, francuskiego i hiszpańskiego. Obie testowane metody (BERT i Flair) istotnie zmniejszyły liczbę błędów, osiągając względną redukcję Word Error Rate na poziomie 21%–25%.

Rozdział 5 opisuje wykorzystanie autorskiej metody z poprzedniego rozdziału w zadaniu automatycznej normalizacji tekstu, a właściwie jednego z podproblemów tak postawionego zadania –

odwrotnej normalizacji tekstu zwracanego przez system rozpoznawania mowy w zakresie przywracania w nim znaków interpunkcyjnych. Dodatkowo, zadanie to zostało ograniczone specyfiką wyzwania „Punctuation restoration from read text" z konkursu PolEval 2021, czyli poprzez ograniczenie zbioru operacji do siedmiu rodzajów znaków interpunkcyjnych. Dane trenujące dla modelu tagowania zostały przygotowane poprzez usunięcie znaków interpunkcyjnych z tekstów referencyjnych i przypisanie operacji edycyjnych dla sąsiednich wyrazów, a model tagowania został przygotowany na bazie modelu HerBERT, dotrenowanego na podzbiorze danych udostępnionych przez organizatorów wyzwania.

Rozdział 6 omawia zastosowania opracowanej metody do badania wpływu błędów ASR na efektywność modeli w zadaniu rozumienia języka (NLU, ang. *natural language understanding*), zmaterializowane w postaci rozwiązania bazowego w konkursie CAICCAIC. Autor był w tym zadaniu odpowiedzialny za przygotowanie danych polskich, angielskich i hiszpańskich na bazie istniejącego korpusu Leyzer oraz ewaluację wyników uczestników, w tym definicję i zaimplementowanie metryk ewaluacyjnych.

4. Uwagi

Przedstawiony materiał potwierdza wiedzę doktoranta zarówno w kwestii metody naukowej, znajomości dyscypliny, jak i kwestii wdrożeniowych. Zarówno struktura pracy, jak i poszczególnych rozdziałów, jest optymalna: od prezentacji motywacji, pytań badawczych i metodologii przez propozycję kilku rozwiązań, ich formalną ewaluację i dyskusję wyników. Ważnym, a rzadko spotykanym elementem składowym pracy jest udział opracowanych rozwiązań w zadaniach konkursowych, co potwierdza ich skuteczność w porównaniu z konkurencyjnymi rozwiązaniami tworzonymi przez innych uczestników, często tak ze środowiska naukowego, jak i komercyjnego.

W każdym z wariantów przedstawionych rozwiązań Autor opracował kilka rozwiązań wykorzystujących różne modele tagowania i porównał ich wyniki stosując opisane wcześniej standardowe miary ewaluacyjne. Co ważne w kontekście wdrożeniowym, analiza wyników uwzględniała także wpływające na stosowalność rozwiązania parametry niefunkcjonalne, takie jak czas trenowania modeli oraz ich latencję (czas potrzebny na poprawienie jednego zdania) wraz z oceną ich akceptowalności w warunkach przemysłowych.

To powiedziawszy, nie można mimo wszystko nie zauważyć, że niektóre opisane w pracy metody (jak np. tager Brilla z 1992 roku, stanowiący m.in. podstawę narzędzia PANTERA do tagowania części mowy w języku polskim z 2009 r.) mocno się zestarzały. To oczywiście wynik niesłychanie szybko zachodzących zmian technologicznych, które sprawiają, że datę rozpoczęcia prac nad doktoratem i moment jego zakończenia dzieli technologiczna przepaść. Nawet więcej: na s. 39 Autor odwołuje się do przeglądu literatury przeprowadzonego około roku 2014, co zapewne nawiązuje do jego wcześniejszych aktywności, także związanych z tematyką pracy, a prowadzonych podczas zatrudnienia w firmie Samsung.

Sam Autor zdaje sobie sprawę z tego uwarunkowania, pisząc we wprowadzeniu, że „badania były prowadzone w dynamicznie zmieniającym się kontekście naukowym i technologicznym, uwzględniając zawsze aktualny stan badań. Zaproponowane autorskie metody (...) wpisywały się w ówczesny rozwój dziedziny, będąc w trakcie powstawania nowatorskimi rozwiązaniami" (s. 10). Nie mam zamiaru czynić Autorowi zarzutu z wykorzystania ówczesnych metod, zwłaszcza że użyte modele typu BERT ze względu na jest ich dwukierunkowość są wciąż uznawane za lepiej nadające się do zadań NLU niż nawet modele z rodziny GPT-3.

Na uwagę zasługuje na pewno kontekst wdrożeniowy badań, także często podkreślany przez Autora, a więc „łatwość interpretacji i modyfikacji działania metody oraz efektywność obliczeniowa umożliwiająca uruchomienie modeli w środowiskach o ograniczonych zasobach obliczeniowych" (s. 9), co również może nie pozostać bez wpływu na wybór użytych metod. Z tego zapewne powodu nieczęsto zdarza mi się czytać prace, które np. oprócz typowo wykorzystywanych w zadaniach NLP miar oceny narzędzi biorą pod uwagę lepiej przydatne użytkownikowi miary takie jak np. ogólny „wskaźnik powodzenia zadania".

Jeśli chodzi o zagadnienia normalizacji, szkoda, że Autor nie idzie o krok dalej i nie porusza go w szerszym kontekście. Początek rozdziału 5 obiecuje wiele, podając różne przykłady normalizacji, np. rozwijanie skrótów, co faktycznie stanowiłoby nietrywialny problem generowania odpowiednich form gramatycznych (p. np. „pociąg PKP odjedzie z toru 1 przy peronie 2"). Niestety, zagadnienie normalizacji tekstu zostało bardzo mocno ograniczone – do przywracania znaków interpunkcyjnych.

Bardzo dziwnie brzmi też tłumaczenie ze s. 74: „Niestety, ze względu na udostępnienie przez organizatorów danych trenujących w dwóch lokalizacjach, autor nie użył danych z podzbioru »rest«, co znacznie ograniczyło liczbę danych użytych do trenowania modeli." Problem jest dla mnie trudny do zrozumienia – w repozytorium są wszystkie pliki, także wspominany plik „rest", który daje się pobrać. Czy to znaczy, że Autor przyznaje się do błędu nieuwzględnienia dodatkowych danych, które jednak były dostępne?

Pierwsza część rozdziału 6 opisuje ograniczone zaangażowanie Autora w prace badawcze związane z przedmiotem zadania ewaluacyjnego. Samo przygotowanie danych oraz ewaluacja wyników nie są pracami naukowymi, nie wpisują się też w kontekst wdrożeniowy. Autor rozprawy nie jest też autorem rozwiązania bazowego (dość silnego, bo znamienny jest przecież fakt, że żadnemu ze zgłoszonych systemów nie udało się pobić jego wyników dla języka angielskiego oraz zadania klasyfikacji zamiarów i domen). Na szczęście rozdział zawiera jeszcze zadanie oceny stopnia wpływu błędów ASR na zadanie NLU, wykonane zgodnie z regułami sztuki, z wykorzystaniem dwóch modeli przetwarzania mowy (aczkolwiek niestety wytrenowanych na tym samym zbiorze danych).

Nie jest jasne, jaki udział miał w tym zadaniu Autor – w pracy wielokrotnie pojawia się sformułowanie o pracy wykonanej „przez autora wraz ze współpracownikami". To akurat ważne pytanie, bo dotyczy treści artykułu przyjętego na prestiżową konferencję EMNLP 2023 (Conference on Empirical Methods in Natural Language Processing), jedną z najlepszych na świecie w dziedzinie przetwarzania języka naturalnego (aktualnie 140 pkt na liście ministerialnej, a pod koniec 2023 roku jej kategoria na liście CORE została nawet podniesiona z A na A*). W artykule tym, opublikowanym w ACL Anthology, znajdziemy niektóre z tabel przytoczonych w pracy, np. *Tabela 5: NLU models performance* to kopia *Tabeli 6.5: Wyniki ewaluacji modeli NLU przed i po procedurze back-transcription (BT)* z pracy, a *Table 7: Top 20 most frequent errors* i *Table 8: Top 20 errors that deteriorate* R_{123} z artykułu to odpowiednio kopie *Tabeli 6.6: 20 najczęstszych błędów ASR w zbiorze* i *Tabeli 6.7: 20 błędów najbardziej pogarszających wynik NLU* z pracy. Mimo to nie traktuję tego przypadku jako autoplagiatu, lecz jako ilustrację wkładu Autora w obie prace, treściowo jednak różne.

W częściach końcowych rozdziałów Autor wymienia wiele ciekawych pomysłów, jak np. „zastosowanie do trenowania modelu NLU danych augmentowanych za pomocą metody *back-transcription* i zbadanie odporności takiego modelu na błędy w porównaniu z modelem trenowanym na czystych danych" (s. 87), ale niestety nie przeprowadza takiego eksperymentu. Tym razem trudno czynić mu z tego zarzut, zwłaszcza w kontekście wdrożeniowym, który rządzi się swoimi prawami i ograniczeniami, także czasowymi, skutkującymi koniecznością zakończenia prac w określonym terminie.

Kontynuując moją prywatną krucjatę korektorską. muszę niestety kolejnemu autorowi recenzowanej przeze mnie pracy zwrócić uwagę na konieczność (nawet wielokrotnego) sczytania tekstu po jego napisaniu – lub powierzeniu tego zadania specjaliście. Trudno mi uwierzyć, że nikt z osób czytających pracę (czyli przede wszystkim Autor po jej napisaniu) nie zauważył, że brakuje przypisów ze znacznikami umieszczonymi w tabelach (4 i 8–10), a sam tekst zawiera wiele literówek, które wykryje każdy program do korekty tekstu. Nie będę już wspominać o takich drobiazgach jak niekonsekwentnie stosowane cudzysłowy, użycie dywizów zamiast myślników czy (akurat tu konsekwentne, ale niepoprawne) użycie kropki zamiast przecinka w roli separatora dziesiętnego. Szkoda, że pracy zabrakło tego ostatniego szlifu, bo sam tekst czyta się bardzo dobrze, co nie jest aż tak znowu częste w przypadku prac z dyscypliny informatycznej.

5. Ocena wiedzy Doktoranta i osiągnięcie celu rozprawy

Doktorant potwierdza bardzo dobrą orientację i stan wiedzy w zakresie informatyki – z dokładnością do stanu wiedzy sprzed 2–3 lat, który to okres wyznaczał zapewne najważniejsze momenty prac nad rozprawą.

Lista pozycji bibliograficznych zawiera 8 artykułów autorstwa Doktoranta, w tym 3 samodzielne i 4 we współautorstwie z promotorem pomocniczym. Jeden z artykułów ukazał się na wspomnianej konferencji EMNLP 2023. co także należy uwzględnić przy ocenie dorobku Doktoranta.

Podczas lektury pracy miałem oczywiście momenty zawahania nad oceną jej zakresu. Część badawcza składa się w zasadzie z dwóch rozdziałów merytorycznych opisujących dwa warianty tego samego rozwiązania. Z jednej strony to bardzo mało jak na pracę doktorską, z drugiej – rozwiązanie to jest centrum całego ekosystemu automatycznej korekty błędów. Oprócz samego rozwiązania praca dokumentuje także historię długoletniego zaangażowania Autora w budowę rozwiązań z zakresu korekty, opisuje wyniki testów zaproponowanego rozwiązania w środowisku zewnętrznym, prezentuje jego adaptację, a w końcu wdrożenie w przemysłowym systemie dialogowym firmy komercyjnej. W kontekście doktoratu wdrożeniowego takie rozwiązanie uznaję za wystarczające.

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?

[ ]	[**X**]	[ ]	[ ]	[ ]
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 Ogrodniczuk

Podpis