Warszawa, 31.01.2024

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

Recenzja rozprawy doktorskiej mgr. Tomasza Ziętkiewicza

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

Recenzja rozprawy doktorskiej mgr. Tomasza Ziętkiewicza, zrealizowanej pod opieką prof. UAM dr hab. Jacka Marciniaka, oraz opiekuna pomocniczego dr. Marka Kubisa, wykonana została na zlecenie Rady Naukowej dyscypliny informatyka Uniwersytetu Adama Mickiewicza w Poznaniu. Przedstawiona rozprawa jest doktoratem wdrożeniowym zrealizowanym we współpracy z firmą Samsung.

Zawartość pracy

Celem rozprawy, będącej wynikiem współpracy między Uniwersytetem Adama Mickiewicza a firmą Samsung, było opracowanie i wdrożenie metod automatycznej korekty błędów w wynikach analizy mowy, oraz normalizacji uzyskiwanego wyniku do prawidłowej postaci tekstowej poprzez dodanie znaków interpunkcyjnych. Kryteria, zgodnie z którymi oceniane miały być rozwiązania to łatwość interpretacji i modyfikacji oraz efektywność obliczeniowa. Prowadzone prace przebiegały zgodnie ze standardowym schematem polegającym na przeglądzie istniejących metod i zasobów, pozyskaniu zbioru danych, które mogą być wykorzystane przy budowie i ewaluacji opracowanych metod, a następnie opracowaniu nowych rozwiązań i ich ewaluacja. Celem przedstawionych w rozprawie badań było między innymi ulepszenie systemu dialogowego wykorzystywanego przez użytkowników sprzętu Samsung. W systemie takim rola modułu rozpoznawania mowy jest bardzo istotna, gdyż od jego jakości zależy czy system będzie mógł reagować adekwatnie na wypowiedzi użytkownika. Z tego powodu wyniki otrzymywane bezpośrednio z systemu ASR poddawane są dalszej obróbce polegającej na korekcie błędów i normalizacji. Ten proces obecnie jest wykonywany przy wykorzystaniu modeli wytrenowanych na danych pochodzących przykładowo z ręcznie poprawionych zapisów przebiegu użycia systemu.

Przedstawiona do oceny praca składa się ze wstępu i siedmiu rozdziałów, w których omówiono kolejno kontekst przeprowadzonych badań, modele uczenia maszynowego, zaproponowane przez autora metody rozwiązania postawionych w pracy problemów korekty błędów i normalizacji tekstu oraz zastosowanie opracowanych metod w praktyce. Rozdział drugi poświęcony jest charakterystyce wdrożeniowego kontekstu badań. Zawiera on opis ogólnego schematu systemu dialogowego i wskazanie elementów istotnych z punktu widzenia komercyjnego użytkownika – szybkości, niezawodności i kontroli zachowania systemu. W rozdziale trzecim autor przedstawia bardzo skrótowo wybrane metody budowy aplikacji realizujących zadania NLP i modele językowe: transformacyjny tagger Brilla, CRF, model BERT, Flair i XML-RoBERTa oraz miary wykorzystywane do ewaluacji wyników. Rozdziały czwarty, piąty i szósty stanowią zasadniczą część pracy. Rozdział czwarty poświęcony jest automatycznej korekcji błędów popełnianych przez systemy rozpoznawania mowy: przeglądowi istniejących rozwiązań, opisowi rozwiązania zaproponowanego przez autora, opisowi

sposobu przygotowania danych i konkursów, w których autor uczestniczył. Kolejny, piąty rozdział, poświęcony jest problemowi normalizacji tekstu. Jako normalizacja rozumiane jest tu ujednolicenie stosowania wielkich i małych liter, interpunkcji, oraz sposobu zapisu wyrażeń, które, tak jak liczby czy daty, zapisywane są w sposób symboliczny, nie odzwierciedlający bezpośrednio sposobu w jakim są wymawiane. Końcowy rozdział opisuje nieco odmienny, ale ściśle skorelowany z tematem ogólnym rozprawy problem badania wpływu różnego rodzaju błędów ASR na efektywność modeli rozumienia języka naturalnego.

Przedstawione osiągnięcie

Głównym celem autora było znalezienie skutecznych i praktycznie użytecznych metod poprawiania wyników analizy mowy. Po krytycznym przeanalizowaniu możliwości metod regułowych i rozwiązań typu end-to-end, swoją uwagę Doktorant poświęcił metodom sekwencyjnego etykietowania tekstu. Przy tym podejściu każdy segment może zostać zaklasyfikowany jako poprawny lub niepoprawny. W tym drugim przypadku etykieta musi zawierać propozycje rozwiązania - zmiany tekstu. Przedstawione w pracy rozwiązanie problemu poprawy tekstu polegało na na wytrenowaniu modelu proponującego jedną z ustalonych wcześniej reguł poprawy. Autor przetestował kilka metod: tagger Brilla, model języka typu BERT oraz model Flair, w którym sieć typu BERT uzupełniona jest warstwą LSTM, a końcowy wybór etykiety dokonywany jest za pomocą metody CRF. Zaproponowana przez autora metoda korekty działa na poziomie znaków ograniczając odwołania do słów do przypadków większych pomyłek. Pozwoliło to znacznie ograniczyć liczbę reguł, które trzeba zastosować by otrzymać poprawny tekst. Wyniki uzyskane przez tagger były jeszcze filtrowane przez zestaw manualnie zdefiniowanych reguł. Na potrzeby organizacji konkursu Ópen Challenge for Correcting Errors of Speech Recognition Systems" (OCCESRS) przygotowano specjalny korpus z nagraniami zdań z polskiej części Wikinews. Eksperymenty na danych firmy Samsung wykazały paroprocentowy spadek liczby błędów (w naturalny sposób był on większy dla modelu o mniejszej skuteczności początkowej). Wykorzystanie biblioteki Flair dało nieznacznie lepsze rezultaty niż sam BERT, ale opłacone znacznie dłuższym czasem trenowania i dłuższym czasem reakcji sytemu. W konkursie Poleval zaproponowana metoda zajęła drugie miejsce, a wyniki na danych wyzwania OCCESRS dały podstawę do wniosków podobnych do tych, które można było wyciągnąć na podstawie wyników dla danych wewnętrznych w innych językach. Osiągnięta niższa poprawa procentowa mogła wynikać z małej liczby danych treningowych. Ogólnie na różnych danych uzyskano usunięcie od 8 do 23 procent błędów.

Podobne podejście Doktorant zaproponował rozwiązując problem normalizacji tekstu, a dokładniej problem przywracania znaków interpunkcyjnych w tekście stanowiacym wynik analizatora mowy. Został on sformułowany jako zadanie etykietowania słów oznaczeniami reprezentującymi operacje wstawienia poszczególnych znaków interpunkcyjnych lub pozostawieniu słowa bez zmian. Proces etykietowania wykonywany jest przez model HerBERT wytrenowany dodatkowo na odpowiednio przygotowanych zdaniach z WikiNews i WikiTalks. Przywracanie znaku kropki okazało się zadaniem stosunkowo łatwym dla takiego modelu (F1=0.9), wyniki dla pozostałych znaków interpunkcyjnych są znacznie gorsze. Opracowane podczas badań metody zostały wdrożone w produkcyjnych systemach dialogowych oraz wykorzystane do przygotowania rozwiązania wyzwań Poleval 2020 i Poleval 2021.

Opracowana metoda przywracania znaków interpunkcyjnych znalazła też zastosowanie w przeprowadzonych przez autora razem ze współpracownikami pracach nad wpływem różnego rodzaju błędów ASR na efektywność modeli rozumienia języka naturalnego (NLU). To zagadnienie, będące przedmiotem szóstego rozdziału pracy, wydaje się pokazywać ciekawy

kierunek prac, w których zamiast poprawiać wszystkie błędy, można się by skupić nad dokładniejszą analizą takich miejsc, które są istotne z punktu widzenia rozwiązania końcowego. Ciekawym wynikiem jest tu też zbiór danych, który może służyć do porównywania różnych systemów pod kątem rodzajów błędów, które są przez nie popełniane. Zaszumiony korpus został utworzony za pomocą metody wstecznej transkrypcji, czyli uzyskano wyniki rozpoznawania mowy z mowy zsyntetyzowanej dla pierwotnego poprawnego tekstu.

Ocena pracy

Przetwarzanie tekstów w języku naturalnym (NLP) jest od około dziesięciu lat dziedziną ogromnie szybko się rozwijającą. W tej sytuacji prowadzenie badań eksperymentalnych jest trudne, gdyż nowe propozycje, rozwiązania pojawiają się bardzo często i są na ogół opracowywane przez firmy dysponujący dużym budżetem i dużymi zasobami danych. Z tego punktu widzenia doktoraty wdrożeniowe mają głęboki sens, gdyż zasoby firmy pozwalają na powiększenie zaplecza eksperymentalnego. Niestety często oznacza to jednocześnie, że nie ma wiele czasu czy motywacji do pracy nad formalną, teoretyczną częścią badań. Rodzi to pewne wątpliwości przy pracach w dyscyplinie informatyka, a nie informatyka techniczna, ale ze względu na dość mało precyzyjny podział między tymi dyscyplinami i wiele przykładów już tak zaklasyfikowanych prac, uważam, że w tym przypadku nie stanowi to problemu. Inny zwykle zauważalny problem to fakt, że praca nad rozprawą doktorską trwa zwykle lata i przy szybkim rozwoju danego obszaru badań wiele założeń początkowych się dezaktualizuje. Z pracy wynika jednak, że doktorant starał się podążać za najnowszymi osiągnięciami w dziedzinie i stosować aktualne narzędzia i metody, jednocześnie porównując ich wyniki do prostszych rozwiązań tradycyjnych.

Opis opracowanych metod i osiągniętych wyników jest precyzyjny, a przeprowadzone badania wydają się być pogłębione i wykonane przy użyciu aktualnych dostępnych w danym momencie narzędzi. Opracowane przez Doktoranta metody pozwalają na zmniejszenie liczby błędów w wyniku uzyskiwanym przez ASR w sposób, który jest możliwy do praktycznego zastosowania w czasie rzeczywistym. Jakkolwiek korekta ta nie jest bardzo skuteczna, to jednak pomaga uczynić tekst czytelniejszym i lepiej odbieranym przez człowieka. Z naukowego punktu widzenia najciekawsza jest końcowa część pracy, w której autor analizuje różne rodzaje cech danych i typów popełnianych przez system błędów starając się odkryć słabe strony poszczególnych rozwiązań.

Dorobek publikacyjny Doktoranta nie jest bardzo duży, ale wystarczający, a najnowszy z opublikowanych artykułów znajduje się w materiałach jednej z najważniejszych konferencji w dziedzinie przetwarzania danych językowych – EMNLP 2023. Opublikowane prace opisujące uzyskiwane wyniki są wieloautorskie, ale to jest obecnie standard przy tego rodzaju pracach.

Krytyczne uwagi dotyczą głównie strony prezentacyjnej. Oceniając początkową ogólną cześć pracy trudno nie zauważyć jej dość dużej skrótowości i bardzo ogólnego poziomu opisu poszczególnych pojęć i problemów. Być może ten ogólny poziom miał wpływ na to, że nie wszystko opisane jest z należytą starannością i dokładnością. W części prezentującej analizowany w pracy problem, opisy konkurencyjnych rozwiązań omawianych problemów są właściwie wyłącznie hasłowe. Można też zauważyć pewne błędy. W szczególności opis taggera Brilla odbiega od jego standardowej definicji. Jeśli zmiany zostały wprowadzone, to należałoby to dokładniej opisać, jeśli nie, to uściślić opis. Wartości funkcji f w definicji CRF standardowo nie są liczbami rzeczywistymi, ale pochodzą ze zbioru 0,1. Na stronie 32 opis obiektów typu FP został przyporządkowany do zduplikowanej etykiety FN. W definicji precyzji zamiast FP jest FS, a definicji czułości (*recall*) FP zamiast FN (opis jest poprawny). Autor nie ustrzegł się te też literówek, ale ich liczba jest raczej niewielka, np.

str. 45: obliczną zamiast obliczoną, pary indeksów (i_{S1}, i'_{S2}) zamiast (i_{S1}, i'_{S1}), czy str. 46 UNSOPPORTED zamiast UNSUPPORTED.

Powyższe uwagi te nie zmieniają mojej pozytywnej oceny rzetelności przeprowadzonych eksperymentów i uzyskanych wyników. Dodatkową wartością, poza zaproponowanymi metodami i poczynionymi wnioskami są, jak obecnie przy wielu tego typu pracach, opracowane dość unikalne zbiory danych, które mogą służyć przy prowadzeniu kolejnych eksperymentów.

Wniosek końcowy

Stwierdzam, iż przedłożona mi do recenzji rozprawa zawiera opis istotnych osiągnięć w dziedzinie poprawy wyników automatycznego rozpoznawania mowy i spełnia wymagania ustawowo stawiane rozprawom doktorskim. Wnoszę o dopuszczenie magistra Tomasza Ziętkiewicza do publicznej obrony.

Agnieszka Mykowiecka

Podpisany certyfikatem wystawionym dla Agnieszka Maria Mykowiecka (Certyfikat kwalifikowany). Utworzony w dniu: 2024-02-02 10:57:02 +0100