Dr hab. Agnieszka Mykowiecka, prof. nadzw.
Instytut Podstaw Informatyki
Polskiej Akademii Nauk
Jana Kazimierza 5
Warszawa

Warszawa, 11.01.2018

Recenzja rozprawy doktorskiej mgr. Romana Grundkiewicza
***Algorithms for automatic grammatical error correction* przygotowanej pod kierunkiem promotora dr. hab. Krzysztofa Jassema i promotora pomocniczego dr. Marcina Junczys-Dowmunta**

Przedstawiona do recenzji praca zawiera uporządkowany, ujednolicony i rozszerzony materiał zawarty w 8 tekstach z konferencji międzynarodowych i składa się z 7 rozdziałów i 2 dodatków.

Rozdział pierwszy stanowi wprowadzenie do tematyki i założonego celu rozprawy, którym była analiza problemu automatycznej poprawy błędów gramatycznych oraz stworzenie efektywnego systemu korekty (GEC, ang. *Grammar error correction*) tekstów angielskich wykorzystującego system statystycznego tłumaczenia maszynowego. Realizacja tego zadania wiązała się ściśle z budową korpusu błędów, który sam w sobie stanowi ważny efekt prezentowanych prac.

Rozdział drugi rozprawy poświęcony jest opisowi problemu poprawy gramatycznej tekstu. Autor przedstawia typologię błędów i założenia przyjęte przy budowie własnego systemu GEC. Zgodnie z nimi autor zajął się poprawianiem błędów popełnianych przez osoby, dla których angielski nie jest pierwszym językiem. Dokładniej, są to błędy oznaczone przez anotatorów w korpusie *NUS Corpus of Learner English* (NUCLE). W kolejnej części rozdziału przedstawiona jest typowa architektura systemów poprawy tekstu składających się z modułu detekcji błędów, generowania możliwych propozycji popraw i ich rangowania. Opisano też najważniejsze problemy, jakie trzeba rozwiązać przy budowie systemu i przy ewaluacji jego wyników. Są to między innymi: możliwe różne poprawne wersje tego samego błędnego tekstu, utrudniająca proces uczenia niska częstość występowania błędów, różnice w typie popełnianych błędów i ich częstości w danych pochodzących z różnych źródeł i od różnych autorów.

Rozdział trzeci rozprawy zawiera opis istniejących zbiorów danych, które mogą służyć jako dane treningowe i testowe dla systemów GEC. Jednym z najpopularniejszych zbiorów danych z anotacjami błędów gramatycznych jest NUCLE zawierający około 60-tysięcy zdań z esejów napisanych przez studentów uczelni w Singapurze, dla których angielski nie jest językiem ojczystym. Eseje poprawione zostały przez profesjonalnych nauczycieli angielskiego. Anotowanych w nich jest 27 typów błędów (np. brak czasownika, zły użyty czas, zły porządek słów). Druga część tego rozdziału zawiera opis tworzenia korpusu WikEd Error Corpus – zbioru zawierającego (prawie) wszystkie poprawki zawarte w historii edycji angielskiej Wikipedii. Zbiór zawiera ponad 55 milionów par zdań zawierających ponad 71 mln poprawek wprowadzonych do tekstów Wikipedii. Aby zbiór danych był jak najbardziej użyteczny autorzy usunęli zdania zbyt krótkie i zbyt długie, poprawki polegające na usuwaniu celowych aktów wandalizmu, zdania ze znacznikami html-owymi oraz zdania z poprawkami dotyczącymi tylko wartości liczbowych. Utworzony korpus dostępny jest publicznie beż żadnych ograniczeń.

Rozdział czwarty poświęcony jest bardzo istotnemu tematowi jakimi są miary pozwalające dokonać ewaluacji metod automatycznej poprawy tekstu. Autor opisuje problemy jakie wiążą się z oceną wyników tego konkretnego zadania. Jedną z cech charakterystycznych jest tu niska częstość występowania błędów w tekście, przez co, przy niestarannej ewaluacji, nierozpoznanie nawet sporej

części błędów może i tak dać wyniki w okolicach 90-95% poprawności. Poza tym jeden błąd można naprawić na wiele sposobów, a tekst, z którym porównujemy wyniki zawiera tylko jedną z możliwych wersji. Sposób opisu błędu prowadzący do uzyskania tego samego efektu także może być różny, różna może być też interpretacja liczby miejsc poprawnych (takich, w których mógł wystąpić błąd, ale go nie stwierdzono). W dalszej części tego rozdziału autor przedstawia standardowe metryki: dokładność, precyzję, pełność, miarę F, a następnie miarę MaxMatch i I-WAcc oraz dwie miary stosowane na ogół w systemach maszynowego tłumaczenia: BLEU i METEOR. Miara MaxMatch (skracana do M^2) mierzy poprawność edycji na poziomie zdania poprzez znalezienie ciągu edycji maksymalizującego przecięcie testowanej hipotezy i zdania referencyjnego. Miara M^2 uznawana jest za standardową miarę dla systemów GEC, mimo iż posiada też negatywne własności. W szczególności nie karze ona w żaden sposób za wprowadzanie nowych błędów. Ponieważ jednak żadna z istniejących miar nie jest wystarczająca by oddać rzeczywiste wrażenia osób czytających poprawiony tekst, ostateczna ocena opiera się na ogół na weryfikacji dokonywanej przez ludzi. W rozdziale 4.3 autor rozprawy opisuje opracowaną procedurę oceny wielu systemów GEC przez anotatorów. Ponieważ ocena wielu odpowiedzi dla wielu testowanych zdań jest bardzo żmudna i łatwo w niej popełnić błędy, procedura oceny wspomagana jest komputerowo przez adaptowana wersję systemu Appraise. Porównanie dotyczyło 13 systemów z konkursu CoNLL-2014. Ich odpowiedzi często są bardzo do siebie zbliżone lub nawet pokrywają się (tylko w jednym przypadku otrzymano 13 różnych propozycji, a średnia liczba różnych propozycji poprawy to 5.7). Autor rozprawy opracował dość skomplikowany system wyboru maksymalnie 5 zdań, które prezentowane są anotatorom. Wybór zdań dokonywany jest w taki sposób, by prezentowane propozycje reprezentowały możliwie dużo porównywanych systemów. W pracy nie jest dokładnie opisane dlaczego zdecydowano się na takie ograniczenie – autor tylko cytuje źródło dotyczące warsztatu MT, które wspiera tę decyzję. Procedura wyboru zdań, a następnie wnioskowania o wszystkich możliwych uporządkowaniach jest starannie opracowana, w szczególności uwzględnia to, że wyniki różnych systemów są często mało rozróżnialnie i drobne różnice liczbowe tworzą raczej arbitralny niż faktyczny porządek. Końcowy wynik poza listą rankingowa obejmuje też grupowanie, które pozwala na lepsze odróżnienie systemów o różnym istotnie poziomie efektywności. Wyniki przedstawionych analiz różnią się od tych zaprezentowanych oficjalnie na CoNLL-2014, ale wydaje się, że ten porządek lepiej skorelowany jest z oceną anotatorów. W szczególności lepsze niż w tych oryginalnych rankingach miejsce (na ogół jest to zmiana z trzeciego miejsca na pierwsze) przypada systemowi, w którego tworzeniu udział miał autor pracy.

W rozdziale piątym pracy przedstawiony jest opracowany przez autora (wraz z promotorem pomocniczym) w 2016 roku system poprawy gramatycznej tekstu wykorzystujący model statystycznego tłumaczenia maszynowego. Autorzy skupili się tu na właściwym doborze parametrów modelu, tak by był on bardziej dopasowany do zadania GEC. Jako dodatkowe cechy modelujące relacje między „językiem źródłowym" i „docelowym" wybrano odległość Levenshteina między frazami (liczoną na poziomie słów) oraz licznik operacji edycji. Elementem rozwiązania jest też kilka modeli języka. Pierwszym z nich jest model n-gramowy dla słów (5-elementów z wygładzaniem typu Knesser-Ney). Nie jest niestety podane na jakich danych ten model był budowany. Drugi model, to model dla kategorii gramatycznych słów wyliczony na danych Wikipedii oznaczonych przez Stanford Log-Linear tagger. Kolejny model, także wyliczony na tekście z Wikipedii, to model klas słów wyliczony poprzez grupowanie wektorów word2vec. Ponadto zbudowano także tłumaczeniowy model operacji, które trzeba wykonać na tekście by otrzymać tekst wynikowy oraz neuronowy model dwujęzyczny.

System korekcji błędów został wytrenowany jako system SMT przy w użyciu aplikacji Moses (z wyłączoną funkcją zmiany porządku słów). Dobór parametrów był optymalizowany odrębnie ze względu na maksymalizację miary BLEU i M^2. W tym drugim przypadku wykorzystano dodatkowy zbiór zawierający anotowane błędy językowe, a zatem pewną wiedzę lingwistyczną, brak jednak dokładniejszego opisu tego zbioru. System dostosowany do miary BLEU miał gorsze wyniki dla

miary M^2, system dostosowany do miary M^2 wykazywał wyższą stabilność. Autor przetestował wykorzystanie różnych zbiorów treningowych, dodanie opisanych wyżej cech dodatkowych i różne rozmiary wytrenowanych modeli językowych.

W drugim etapie rozwoju system został wzbogacony o elementy pozwalające na sterowanie wyborem rozwiązania takie jak klasyfikator i zestaw cech rzadkich. Cechy te to wiele binarnych funkcji reprezentujących konkretne zjawiska, takie jak na przykład zamiana 'a' na 'the', dotyczące samego słowa lub jego bezpośredniego kontekstu. Opis modyfikacji systemu i jego ewaluacja zawarte są w rozdziale szóstym pracy. Wykazała ona ograniczoną wartość zastosowanej klasteryzacji i pozytywny, choć niezbyt wielki, wpływ cech rzadkich. Najlepszy z opracowanych systemów przewyższa zwycięzcę konkursu CoNLL-2014 oraz trenowany na tych samych danych konkurencyjny system (Rozovskaya i Toth, 2016) zarówno co do wartości precyzji, pełności, jak i miary M^2 (52,21 w stosunku do 47,4).

Tematyka rozprawy jest bardzo aktualna i ma potencjalnie duże znaczenie praktyczne, gdyż korektory składni są naturalnym i potrzebnym uzupełnieniem bardzo obecnie popularnych korektorów ortograficznych. Wybrana metoda realizacji tego zadania nie jest sama w sobie nowatorska, ale w ramach prac dokonano wielu prac adaptacyjnych i uzupełniających pozwalających na wykorzystanie jej potencjału w sposób znacznie szerszy niż w prezentowanych do tej pory innych rozwiązaniach. Praca napisana jest bardzo dobrym językiem angielskim, jej układ formalny jest prawidłowy, a podział tematyczny logiczny. Chociaż, zapewne poprzez łączenie kilku prac, autor nie ustrzegł się jednak pewnych powtórzeń, czy rozproszenia informacji na ten sam temat. Przykładowo, zbiory danych opisane są zarówno w rozdziale 3 jak i 5. Rozprawa jest dość skondensowana i opis opracowanych rozwiązań jest czasem zbyt skrótowy lub zawiera tylko odesłanie do literatury, np. model językowy opary na klasach słów zbudowany został poprzez grupowanie wektorów, ale w pracy nie ma informacji dotyczących użytego algorytmu grupującego. Praca zawiera bardzo bogatą, adekwatną do tematyki bibliografię. Jedynym niedosytem, jaki pozostaje po przeczytaniu pracy, jest brak analizy błędów popełnianych przez opracowany system GEC. Prezentowane wyniki precyzji na poziomie 0,4-0,7 i pełności na poziomie 0,2-0,3 wskazują, że problem nie jest jeszcze rozwiązany w sposób zadowalający i dobrze byłoby sprawdzić czy dotyczy to w równym stopniu wszystkich rodzajów błędów, czy błędów określonego rodzaju.

Podsumowując, rozprawa doktorska Romana Grudkiewicza dotyczy wybranej aplikacji i wybranej metodologii jej rozwiązania, zawiera szczegółową analizę związanych z tym problemów, różne propozycje ich rozwiązania i bardzo staranną ewaluację osiągniętych wyników. Według mojej opinii jest to praca w pełni spełniającą wymogi stawiane przed rozprawami doktorskimi w zakresie informatyki. Ze względu na moje osobiste doświadczenie zawodowe mam mniejsze kwalifikacje do oceny prac z dziedziny nauk matematycznych niż technicznych, ponieważ jednak w prezentowanej rozprawie podstawowymi wykorzystywanymi i rozwijanymi metodami są metody statystyczne, a autor wykazuje głębokie zrozumienie ich matematycznych podstaw rozszerzając je o dodatkowe cechy, a także przedstawiając swoje propozycje rozwiązania procedury ręcznej ewaluacji w sposób formalny, uważam, że praca spełnia kryteria pracy doktorskiej z dziedziny nauk matematycznych.

Konkludując, stwierdzam, że praca *Algorithms for automatic grammatical error correction* spełnia wymagania, jakie ustawa o stopniach i o tytule naukowym przewiduje dla rozpraw doktorskich i na tej podstawie wnoszę o dopuszczenia jej Autora - mgr Romana Grundkiewicza - do publicznej obrony.