Politechnika Wrocławska

Prof. dr hab. inż. Halina Kwaśnicka
Katedra Inteligencji Obliczeniowej
Wydział Informatyki i Zarządzania
Politechnika Wrocławska
Wyb. Wyspiańskiego 27, 50 370 Wrocław
Tel: (48)(71) 320 35 34, Fax: (48)(71) 321 10 18
E-mail: halina.kwasnicka@pwr.edu.pl

28 stycznia 2018.

Recenzja rozprawy doktorskiej

Tytuł rozprawy: ALGORITHMS FOR AUTOMATIC GRAMMATICAL ERROR CORRECTION

Autor rozprawy: mgr Roman Grundkiewicz

Recenzja wykonana jest na zlecenie Rady Wydziału Matematyki i Informatyki Uniwersytetu im. Adama Mickiewicza w Poznaniu, z dnia 27 października 2017r., pismo z 30 października 2017r.

Promotorem rozprawy jest dr hab. Krzysztof Jassem, prof. UAM
Promotorem pomocniczym jest dr Marcin Junczys-Dowmunt

1. Obszar problemowy pracy

Przez wiele wieków funkcję *lingua franca* pełniła greka i łacina, potem język francuski, a obecnie rolę tę przejął język angielski. Niezależnie od dużej liczby osób, dla których jest to podstawowy język, to coraz większa grupa uczy się języka angielskiego jako drugiego języka. Według raportu opublikowanego przez British Council w 2013 r. język angielski jest używany na 'poziomie użytecznym' przez 1,75 miliarda ludzi na całym świecie. Rick Noack i Lazaro Gamio, wykorzystując m.in. wyniki 15-letnich badań prof. Ulricha Ammona z University of Düsseldorf, w artykule "The world's languages, in 7 maps and charts"[1] podaje, że jeśli chodzi o naukę języków, to zdecydowanie najwięcej osób uczy się języka angielskiego, jest to 1,5 mld osób. Drugie miejsce w tym zakresie zajmuje język francuski (82 mln ludzi), trzecią lokatę zajmuje język chiński z 30 mln uczących się, kolejno język hiszpański i niemiecki (po 14,5 mln), włoski 8 mln, japoński 3 mln. Nauka nowego języka nigdy nie jest łatwa. Różnice między językiem podstawowym a angielskim przysparzają uczącym się trudności, co powoduje, że błędy popełniane przez uczących się różnią się od błędów pisanych przez 'native speakerów'.

Błędy popełniane przez uczących się języka angielskiego jako drugiego języka są przedmiotem zainteresowania świata naukowego co najmniej od lat 80. poprzedniego wieku. Wyniki prowadzonych badań pokazały m.in., że połowa z dziesięciu najczęściej popełnianych błędów przez 'native speakerów' była "nieistotna" w tekstach tworzonych przez ESL (English as a Second Language learners). Automatyczne weryfikowanie i korygowanie błędów popełnianych przez osoby piszące w języku angielskim, dla których jest on drugim językiem,

[1] https://www.washingtonpost.com/news/worldviews/wp/2015/04/23/the-worlds-languages-in-7-maps-and-charts/?utm_term=.f1df8126b572

znajduje się w kręgu zainteresowań z uwagi na ciekawy problem badawczy i potencjał komercyjny. Automatyczna korekta błędów gramatycznych jest popularnym przedmiotem badań w obszarze przetwarzania języka naturalnego (NLP, Natural Language Processing).

Recenzowana praca dotyczy wymienionego wyżej obszaru. Autor postanowił zaproponować **efektywne algorytmy i metody do weryfikacji i poprawy gramatycznej tekstów w sposób automatyczny** (system GEC).

Doktorant postawił hipotezę, że **połączenie generatywnych statystycznych modeli translacji maszynowej (na podstawie fraz) oraz komponentów klasyfikujących stosowanych do automatycznej korekcji błędów gramatycznych , da lepsze efekty od każdego z tych podejść oddzielnie.** W celu weryfikacji postawionej hipotezy badawczej sformułowano cztery zadania do realizacji. Aby zastosować podejście statystyczne niezbędne jest zgromadzenie przykładów naturalnie występujących błędów gramatycznych. Jest to zadanie pierwsze. Kolejne zadanie dotyczy wyboru automatycznej miary oceny, która najlepiej odpowiada ocenie ludzkiej. Mając zrealizowane poprzednie zadania, można przystąpić do trzeciego zadania – budowy wydajnego, zautomatyzowanego systemu korekcji błędów gramatycznych, opartego na statystycznym tłumaczeniu maszynowym. Ostatnie zadanie, pozwalające na weryfikację hipotezy badawczej, to integracja składników dyskryminacyjnych (klasyfikatorów) z modelem statystycznym.

Podsumowując tę część rozprawy stwierdzam, że Autor **podjął ważny i aktualny problem naukowy. Zaproponował autorskie rozwiązania, przeprowadził analizę istniejących podejść i możliwych rozwiązań, oraz ważne badania eksperymentalne**, pozwalające ocenić zaproponowane rozwiązania.

2. Kompozycja i zawartość pracy

Praca składa się z wstępu (numerowanego jako rozdział 1.), pięciu rozdziałów oraz podsumowania. Zawiera też dwa załączniki oraz spisy skrótów, symboli, tabel i rysunków. Na końcu zamieszona jest bibliografia. Spisy przydałyby się na początku pracy, zwłaszcza spis skrótów i symboli. Załączniki, jak sama nazwa wskazuje, umieszcza się po całej pracy, a bibliografia jest jej niezbędną częścią, powinna być po podsumowaniu.

We wstępie Autor zamieścił **motywacje** zajęcia się tym tematem, **przedstawił cel i zadania** badawcze, podkreślił wkład Autora w rozwój dziedziny, przedstawił bardzo krótko zwartość pracy. Zamieścił też spis ośmiu publikowanych prac konferencyjnych, których jest autorem bądź współautorem.

Rozdział drugi prezentuje podstawową wiedzę naukową, potrzebną do zrozumienia pozostałych części pracy. Omawia proces automatycznej korekty błędów gramatycznych, oraz taksonomię podejść do korekty tekstów opartych na typie błędów. Modele języka są ważnym elementem w każdym zadaniu przetwarzania języka naturalnego, Autor przybliżył również tę tematykę, podobnie jak podejścia dyskryminujące i statystyczne do maszynowego tłumaczenia. Autor dobrze wprowadził w obszar automatycznej korekcji błędów gramatycznych, wskazał, że **celem jest wykrywanie i korygowanie całego zakresu błędów gramatycznych, które mogą być tworzone przez użytkowników języka**. Dokonując przeglądu podejść wskazał na ich mocne i słabe strony. Rozdział jest kompletny i dobrze prezentujący obszar badań będący przedmiotem recenzowanej pracy. Na stronie 9. Pojawia się pierwszy raz skrót 'POS' (POS

tagging), dobrze by było tutaj podać jego pełne znaczenie (part-of-speech tagging). Termin „part-of-speech tagging" jest użyty dalej na tej stronie, ale jest to dalej i niekoniecznie kojarzy się natychmiast z wcześniej użytym skrótem POS.

Rozdział trzeci jest poświęcony zbiorom danych. Każde podejście wykorzystujące dane musi korzystać z odpowiednio przygotowanych zestawów danych. Niektóre błędy mogą być wykrywane w oparciu o reguły heurystyczne, formułowane przez ekspertów, jednak wykrywanie wielu rodzajów błędów wymaga korzystania z dużych zasobów danych – korpusów. Autor prezentuje w tym rozdziale rodzaje i źródła danych wykorzystywanych przy tworzeniu systemów automatycznej korekty błędów gramatycznych. Ważnym osiągnięciem jest **opracowanie i udostępnienie publicznie korpusu *WikEd Error Corpus***. Jest to duży zasób, który może być wykorzystywany także przez innych badaczy, również do innych zadań w obszarze przetwarzania języka naturalnego. Korpus powstał przy pomocy **opracowanej metody ekstrakcji błędów z historii poprawek** Wikipedii, metoda jest niezależna od języka. Nie mam uwag do tego rozdziału.

Bardzo ważnym rozdziałem jest kolejny, poświęcony miarom oceny systemów automatycznej korekty błędów gramatycznych. Są dwa powody mojej wysokiej oceny ważności tego rozdziału. Po pierwsze, to rola miar w tworzeniu i ocenie tworzonych systemów przetwarzania języka naturalnego. Po drugie, pokazuje dojrzałość badawczą Autora. Rozdział zaczyna się od analizy trudności w ocenie systemów GEC, gdzie problemy są wypunktowane wraz z krótkimi komentarzami. Następnie zaprezentowane są poszczególne miary oceny. Praca jest zwarta, bez przegadania, mimo to czasami nie jest łatwa w czytaniu. Umieszczona na górze strony tabela 4.1 zawiera w ostatnim wierszu gwiazdkę, odwołanie do tabeli jest na dole strony, dopiero w następnym zdaniu jest wyjaśnienie co oznacza gwiazdka w tej tabeli. Umieszczenie tabeli na górze następnej strony ułatwiłoby (przynajmniej mnie) czytanie tej części pracy. Dla porządku, oznaczenie symbolu N we wzorze na A_{cc} powinno być podane pod wzorem (jest podane w spisie symboli). Omawiając miarę *MaxMatch* (M^2) Autor oznacza zbiór kandydatów zmian symbolami $c_i, \ldots, c_n$, zmiany referencyjne jako $r_i, \ldots, r_m$ (strona 34.). Następnie, we wzorach poniżej, występują sumy od i=1 do n wyrażeń zawierających r_i. To jak się ma n do m? Jak policzyć sumę $|r_i|$ dla i zmieniającego się od 1 do n, przy podanych kandydatach c od i do n a referencyjnych od i do m? Szkoda że wzory nie są numerowane, łatwiej byłoby się po nich „poruszać". Czym jest (poza ogólną informacją) i jak się liczy karę zwięzłości (ang. brevity penalty) musiałam sobie poszukać w innych źródłach. Podobnie Pen_{METEOR}. We wzorze na p_i na stronie 39. też miałam problem ze zrozumieniem czym jest E; aby nie pozostawać w sferze domysłów zajrzałam do spisu symboli. Warto podkreślić, że Autor, po dość dokładnej analizie, dostosował metody oceny ludzkiej WMT do GEC. Pokazał, że często używane wartości parametrów dla standardowych metryk mogą nie być najlepsze dla rozpatrywanego w pracy zadania. Metryka *MaxMatch* (M^2) okazała się najlepiej skorelowana z ludzkimi osądami spośród wszystkich testowanych miar. Zawarte przykłady w rozdz. 4.3.1.1 ułatwiają zrozumienie tej części pracy. Pokusiłam się o przeliczenie przykładu 4.2 – wynik końcowy wyszedł mi ten sam, choć licznik C_i wyszedł mi 353 a nie 354 (nie jest to istotne z punktu widzenia oceny rozprawy).

Dwa kolejne rozdziały dotyczą dwóch podejść łączonych w niniejszej pracy, w jednym systemie GEC: statystycznego tłumaczenia maszynowego (SMT, Statistical Machine Translation) oraz modele dyskryminacyjne (DM, Discriminative Models). Te pierwsze są modelami generatywnymi, na podstawie danych estymują rozkłady parametrów modelu i wykorzystują je do przewidywania nowych (nieznanych) danych. Modele dyskryminacyjne modelują granice pomiędzy możliwymi wyjściami systemu, na podstawie obserwowanych danych. W rozdziale piątym przedstawiony jest autorski wkład w GEC przy wykorzystaniu tłumaczenia maszynowego (korzystano z zestawu narzędzi Moses). Wprowadzono tu dwie ważne idee: **dostrojenie parametrów do specyficznej dla zadania metryki** oceny, oraz **eksploracja funkcji cech o różnych gęstościach**. Stwierdzono, że konfiguracja SMT opartego na frazach, z odpowiednią optymalizacją przewyższa wszystkie wcześniej opublikowane wyniki dla zestawu testowego CoNLL-2014. Najlepsze wyniki dał tu dobór parametrów w oparciu o miarę M^2. Lepsze zbiory danych i procedura optymalizacji z wykorzystaniem walidacji krzyżowej dały bardziej stabilne wyniki. Mechanizm strojenia umożliwił zbadanie cech specyficznych dla zadań, które poprawiają wyniki, np. liczba operacji edycji par fraz i różne modele językowe.

Autor nie zadowolił się uzyskaniem dobrych na tle literatury wyników przy podejściu SMT. Historycznie, najlepsze wyniki dawały modele dyskryminacyjne, definiowane dla konkretnych typów błędów. Przedmiotem analizy zawartej w rozdziale 6. jest próba poprawy działania opracowanego systemu GEC, opartego na SMT, poprzez zintegrowanie go z modelem dyskryminacyjnym. Pokazano w jaki sposób można zintegrować modele dyskryminacyjne z modelem log-liniowym. Przedstawiono szablony funkcji używane do generowania zestawów cech dla każdej metody. Klasyfikator jest uczony na wyekstrahowanych danych i specjalnie zaprojektowanych cechach. Cechy wykorzystywane do treningu klasyfikatorów są konstruowane automatycznie w oparciu o *n*-gramy lub o operacje edycyjne. Miara klasyfikacji jest wykorzystywana jako funkcja cech w modelu log-liniowym. Uzyskane wyniki porównano z bazowymi z literatury (tabela 6.6). Generatywny system SMT wykorzystujący specyficzne dla zadania cechy dyskryminacyjne, przewyższa większość raportowanych w literaturze wyników dla dowolnego podejścia w GEC (zbiory CoNLL-2014 oraz GEC-10).

Ostatni rozdział podsumowuje całą pracę. Zawiera prezentację autorskiego wkładu w rozwój dziedziny oraz możliwych kierunków przyszłych prac.

Przed bibliografią umieszczone są dwa załączniki oraz spisy – skrótów, symboli (oba bardzo potrzebne), tabel i rysunków. Zdecydowanie wolałabym, by spisy te były przed właściwą pracą, a załączniki po bibliografii.

Pracę kończy Bibliografia, licząca blisko 200 pozycji literaturowych. Cytowane prace są różnorodne – książki, publikacje w czasopismach naukowych i konferencjach oraz strony www. Nie mam zastrzeżeń do spisu literatury, jest bardzo obszerny i na temat.

Podsumowując tę część recenzji stwierdzam, że **praca jest skonstruowana poprawnie**, jej język jest odpowiedni dla prac naukowych, jakim jest dysertacja doktorska. Autor biegle posługuje się formalizmem w precyzyjnym prezentowaniu omawianych problemów. Praca zawiera też dobrze dobrane, proste przykłady ilustrujące omawiane/definiowane pojęcia. Eksperymenty są przeprowadzone i omówione w sposób właściwy dla prac naukowych.

3. Oryginalność i waga osiągnięć

Postawiony w dysertacji cel jest **ważny i aktualny z naukowego punktu widzenia. Aspekt praktyczny** jest również obecny. Z uwagi na dużą liczbę osób posługujących się językiem angielskim jako językiem obcym, automatyczna weryfikacja i korekta ich tekstów jest użyteczna. Tematyka ta dobrze wkomponowuje się w nurt współczesnych badań nad przetwarzaniem języka naturalnego (NLP). Wśród osiągnięć Autora należy wymienić:

- opracowanie metody służącej do opracowania oraz opracowanie dużego (ok. 50 mln zdań) korpusu *WikEd Error Corpus*, który został udostępniony publicznie,
- dostosowanie metod 'ludzkich' z Workshop'u Machine Translation do potrzeb automatycznej korekcji błędów gramatycznych,
- przeprowadzenie analizy korelacji między standardowymi metrykami stosowanymi w systemach GEC i osądem 'ludzkim', zauważenie oraz pokazanie, że stosowane parametry mogą nie być dobre i można je dobrać lepiej, eksperymentalne pokazanie, że system oparty na SMT może być lepszy od innych podejść GEC, przy odpowiednim dostrojeniu parametrów,
- wykorzystanie mechanizmu dostrajania do analizy cech specyficznych dla zadania, co pozwala stwierdzić które cechy poprawiają wyniki systemu opartego na frazach.

Za najważniejsze osiągnięcie naukowe uważam ciekawe połączenie dwóch podejść – opartego na SMT i dyskryminacyjnego. Wykorzystanie klasyfikatora jako źródła dodatkowych cech do modelu log-liniowego jest ciekawym pomysłem. Sama praca jest skonstruowana w sposób logiczny, Autor przeanalizował poszczególne komponenty proponowanego systemu GEC dobierając możliwie najlepsze z nich. Ciekawa i ważna jest analiza przydatności poszczególnych miar do porównania wyników z osądem ludzkim.

4. Uwagi i problemy do dyskusji

Praca jest napisana językiem zwięzłym, poprawnym naukowo. Zawarte przykłady ułatwiają zrozumienie omawianej problematyki. Postawiony w dysertacji cel jest ważny i aktualny z naukowego punktu widzenia. Jest to praca, do której nie mam istotnych uwag, wpływających w sposób istotny na jej ocenę. Nasuwające się uwagi redakcyjne zamieściłam w części omawiającej kompozycję i zawartość pracy. Mogę tu dodać, że ja wolę nagłówki tabel a nie podpisy (co oznacza, że 'podpis' tabeli jest nad tabelą), podpisy pod rysunkami, numerowane wzory, oraz legendę do symboli występujących w tabeli zawartą w jej nagłówku. Tu zamieszczam uwagi ogólne, bardziej jako problemy do dyskusji a nie zastrzeżenia do pracy.

1. Autor porównał opracowaną metodę z wynikami znanymi z literatury na kilku zbiorach testowych. Nie przeprowadzono statystycznej analizy (porównującej metody). Co można powiedzieć o istotności tych wyników? Na ile możemy je uogólniać?
2. Brakuje mi w pracy 'spojrzenia z góry' na zaproponowany system GEC. Przydałby się jakiś schemat (diagram), pokazujący przepływ danych w tym systemie, w całości. Zaczynając od danych wejściowych do systemu, przechodząc przez poszczególne bloki przetwarzania, do wyjścia systemu.
3. Praca pokazuje, m.in., problem z doborem odpowiedniego wektora cech, na których modele są uczone (z możliwie dużą precyzją, czy inną miarą, np. miarą F). Problem ten występuje również w analizie obrazów i w wielu innych zadaniach rozpoznawania wzorców. Uczenie

głębokie (deep learning) zwalnia twórcę systemu z uciążliwego i trudnego definiowania zestawu cech[2]. Czy Doktorant może skomentować ten temat?

4. Autor łączy dwa różne podejścia (SMT i klasyfikatory). Czy zdaniem Autora warto rozważać inne połączenia różnych podejść, np. homogeniczne lub heterogeniczne klasyfikatory złożone (ensembles)?

Wymienione usterki i problemy dyskusyjne nie zmieniają wrażenia, że praca jest merytorycznie wartościowa i zredagowana bardzo starannie.

5. Konkluzja

Podsumowując stwierdzam, iż przedłożona mi do recenzji rozprawa, której autorem jest mgr Roman Grundkiewicz, **zawiera oryginalne i ważne osiągnięcia w obszarze sztucznej inteligencji w zastosowaniach do weryfikacji i korekcji tekstów pisanych w języku angielskim przez użytkowników, dla których nie jest to język ojczysty**. Autor wykazał się **dużą wiedzą** w tematyce rozprawy, **dobrą umiejętnością pracy naukowej oraz dobrą znajomością metod badawczych**. Osiągnięte wyniki świadczą o dobrym **przygotowaniu Autora do pracy naukowej**. Wymienione w recenzji uwagi i zapytania nie wpływają w sposób znaczący na moją **wysoką ocenę osiągnięć naukowych** Doktoranta.

Doktorant wymienił osiem publikacji Jego autorstwa bądź współautorstwa, wszystkie to prace konferencyjne. Nie umniejszając rangi prac konferencyjnych, szkoda, że nie ma choć jednej pracy w liczącym się czasopiśmie naukowym, co umożliwiłoby mi wystąpienie o wyróżnienie recenzowanej pracy.

Recenzowana praca spełnia wymagania ustawowo stawiane rozprawom doktorskim, zatem wnoszę o to, by mgr Roman Grundkiewicz **został dopuszczony do publicznej obrony.**

Halina Kwaśnicka

[2] Cytat z jednej z prac: „Instead of using surface and shallow features (POS, parse information, etc.), we use deep features directly. In particular, we use bidirectional Gated Recurrent Units (GRUs) to represent context. Compared with traditional classifier approach for GEC, our new method does not require elaborated feature engineering for each error type. Deep context representations are learnt from large plain text corpora in an end-to-end fashion."