Kraków, 01.08.2022 r.

Marek Skomorowski (prof. dr hab. inż.)

Instytut Informatyki i Matematyki Komputerowej
Uniwersytetu Jagiellońskiego

Recenzja rozprawy doktorskiej

Pana mgra inż. Zbigniewa Kalety

zatytułowanej

Regułowy algorytm automatycznego rozstrzygania wieloznaczności leksykalnej na granicy części mowy w tekstach języka polskiego

Przewód doktorski jest prowadzony przez Radę Dyscypliny Naukowej Informatyka Techniczna i Telekomunikacja Akademii Górniczo-Hutniczej w Krakowie (RDNITiT AGH). Recenzja została napisana na zlecenie Przewodniczącego RDNITiT AGH, Pana prof. dra hab. inż. Marka Kisiela-Dorohinickiego (pismo z dnia 13.06.2022 r.).

1. Problem badawczy, jego znaczenie i aktualność

Język polski, podobnie jak inne języki naturalne, zawiera słowa wieloznaczne. Czytając teksty lub słuchając wypowiedzi zawierające słowa wieloznaczne, najczęściej jesteśmy w stanie poprawnie je zinterpretować na podstawie kontekstu zdaniowego, pozazdaniowego (niewchodzącego w zakres zdania) lub pozajęzykowego (doświadczenie, wiedza). Jednak problem automatycznej interpretacji wieloznaczności w językach naturalnych nadal pozostaje wyzwaniem. Problem ten jest traktowany jako otwarty i jest określany jako *AI-complete* (*Artificial Intelligence-complete*), co uzasadnia prowadzenie badań w tym obszarze. Z punktu widzenia automatycznej analizy tekstu najistotniejsze są: wieloznaczność syntaktyczna i leksykalna. Wieloznaczność syntaktyczna to sytuacja, w której zdanie możemy interpretować na różne sposoby. Wieloznaczność leksykalna oznacza, że pojedyncze słowo ma więcej niż jedno znaczenie.

Recenzowana rozprawa doktorska dotyczy problemu automatycznego rozstrzygania wieloznaczności leksykalnej na granicy części mowy w tekstach pisanych w języku polskim. Zastosowania algorytmów rozstrzygania wieloznaczności leksykalnej są szerokie, na przykład: wyszukiwanie informacji w tekście, komunikacja człowiek-komputer w języku naturalnym, automatyczne tłumaczenie tekstu czy też rozpoznawanie i synteza mowy. Tematyka rozprawy lokuje się w dynamicznie rozwijającym się obszarze przetwarzania języka naturalnego (*Natural Language Processing,* NLP) w ramach sztucznej inteligencji.

Podsumowując, podjęta w rozprawie problematyka jest ważna i aktualna, zarówno z poznawczego jak również praktycznego punktu widzenia.

2. Tezy rozprawy

Celem rozprawy, sformułowanym w rozdziale pierwszym, było wykazanie, że:

1. Jeśli w tekście polskim wystąpi forma fleksyjna reprezentująca jednostki słownika należące do dwu lub więcej różnych części mowy, to wieloznaczność tego typu może być rozstrzygana za pomocą schematów syntaktycznych czasowników.

2. W miejsce schematu syntaktycznego dla czasownika występującego w tekście można z dobrym skutkiem użyć schematu jego odpowiednika aspektowego.

3. Pary aspektowe czasowników można wykryć w sposób automatyczny, dysponując zbiorem czasowników z przypisanymi aspektami i korzystając z wiedzy o słowotwórstwie.

3. Zawartość rozprawy

Rozprawa liczy 174 strony i zawiera streszczenie (w języku polskim i angielskim), spis treści, 5 rozdziałów, wykaz bibliografii (101 pozycji), wykaz rysunków, wykaz tabel i 5 dodatków. Dodatki zawierają informacje dotyczące wykazu form wieloznacznych, gramatyk formalnych, drzewa przedrostków i końcówek słowotwórczych czasownika jak również wybranych drzew derywacyjnych algorytmu wykrywającego pary aspektowe.

W rozdziale pierwszym omówiono genezę problematyki badawczej, przedstawiono motywację podjętych badań i tezy rozprawy. Rozdział pierwszy zawiera również opis zawartości rozprawy.

Rozdział drugi jest wprowadzeniem do problemu wieloznaczności leksykalnej. Przedstawiono w nim model symbolu językowego. Opisano zagadnienia wieloznaczności w zakresie jednej części mowy i na granicy części mowy. Omówiono terminologię stosowną w rozprawie (słowo, napis, token). Przedstawiono skalę i znaczenie występowania form fleksyjnych wieloznacznych w języku polskim korzystając z następujących słowników: Słownik Fleksyjny Języka Polskiego (CLP) i PoliMorf. Opisano statystykę (w postaci tabel) wieloznaczności w następujących korpusach tekstów: korpusie książek beletrystycznych, korpusie Słownika Syntaktyczno-Generatywnego Czasowników Polskich (SSGC), korpusie Słownika Frekwencyjnego Polszczyzny Współczesnej (KF), Narodowym Korpusie Języka Polskiego (NKJP) i korpusie konkursu PolEval. Scharakteryzowano wykorzystywane korpusy językowe, jak również uzasadniono wybór NKJP i KF jako najbardziej reprezentatywnych z punktu widzenia problematyki rozprawy. Z przedstawionych danych (Tabela 2.1) wynika, że problem wieloznaczności na granicy części mowy w słownikach dotyczy od 0,63% (CLP) do 1,57% (PoliMorf) wszystkich form fleksyjnych, co wydaje się być nieistotne. Natomiast problem wieloznaczności na granicy części mowy w korpusach książek dotyczy 10,82% (Tabela 2.3) wszystkich form fleksyjnych, co jest już istotne i uzasadnia prowadzenie badań w tym obszarze. Omówiono zastosowania algorytmów rozstrzygania wieloznaczności.

W rozdziale drugim przedstawiono również stań badań w zakresie rozstrzygania wieloznaczności leksykalnej. Opisując znane podejścia podano za ich autorami (jeśli było to możliwe) następujące miary algorytmów: precyzja, czułość, miara F1 i dokładność. Omówiono takie pojęcia jak rozstrzyganie gruboziarniste i drobnoziarniste. Przedstawiono problem rozstrzygania wieloznaczności jako zadanie samodzielne i jako fragment innego zadania. Opisano sieć semantyczną WordNet wykorzystywaną przez metody WSD (*Word Sense Disambiguation*). Omówiono dwa następujące podstawowe podejścia algorytmów WSD: pierwsze z nich bazujące na wiedzy i drugie wykorzystujące uczenie maszynowe. Przedstawiono metody WSD stosowane do języka polskiego. Opisano także algorytmy tagowania morfosyntaktycznego, które rozstrzygają wieloznaczność w ograniczonym zakresie. Rozdział drugi kończy następujący komentarz uzasadniający podjęcie tematyki będącej przedmiotem rozprawy: zastosowanie algorytmu regułowego, opartego na wiedzy lingwistycznej zwiększa szansę na niezależność wyników od korpusu. Rozdział drugi stanowi niezbędne wprowadzenie do problematyki rozprawy.

W rozdziale trzecim zaproponowano oryginalny algorytm automatycznego rozstrzygania wieloznaczności na granicy części mowy w oparciu o schematy syntaktyczne czasowników. Założono, że wejściem dla zaproponowanego algorytmu jest zdanie w postaci pisanego tekstu w standardowym szyku (najpierw podmiot, później orzeczenie), poprawne pod względem ortografii i interpunkcji. Założono również, że wyjściem zaproponowanego algorytmu jest przyporządkowanie poszczególnym tokenom zdania wejściowego ich części mowy. Przedstawiono pojęcie schematu syntaktycznego (wzorca składniowego) związanego z danym czasownikiem. Omówiono słownik SSGC zawierający wzorce składniowe dla czasowników w języku polskim, determinujące sposoby budowania poprawnych zdań. Przedstawiono konwencję oznaczeń wykorzystywanych do opisu działania zaproponowanego algorytmu. Opisano schemat działania zaproponowanego algorytmu sprowadzającego się do następujących pięciu podstawowych etapów: tokenizacja; budowa hipotez; wykrywanie i interpretacja fraz rzeczownikowych; dopasowywanie hipotez do schematów; wybór hipotezy i rozstrzygnięcie. Każdy z etapów został szczegółowo omówiony w kolejnych podrozdziałach rozprawy. Przedstawiono w nich następujące zagadnienia: poprawki tokeniazacji; obsługi czasowników zanegowanych (słowo *nie* bezpośrednio przed czasownikiem); traktowania zaimków i interpretacji fraz rzeczownikowych; łączenia fraz rzeczownikowych; obsługi takich samych fraz rzeczownikowych (powtórzenie); obsługi zdania złożonego (więcej niż jeden czasownik); obsługi zdania w niestandardowym szyku i form nierozpoznanych.

W rozdziale trzecim podano również definicje następujących miar: precyzja, czułość, miara F1 i dokładność, za pomocą których dokonano oceny zaproponowanego algorytmu. Podczas testów przyjęto, że przypadkami testowymi są tylko tokeny wieloznaczne. Wyniki testów uzyskanych za pomocą zaproponowanego algorytmu porównano z wynikami testów uzyskanych za pomocą algorytmu KRNNT (autor [98] nie podaje rozwinięcia skrótu KRNNT) dla tych samych danych. W przeprowadzonych testach wykorzystano korpusy: SSGC i zmodyfikowane PolEval, NKJP i KF. W testach nie uwzględniano zdań nadmiernie złożonych, za które uznano takie, które zawierają co najmniej 40 tokenów lub 7 tokenów wieloznacznych. Podano przykładowe zdania odrzucone z powodu nadmiernej złożoności.

Podano także przykładowe zdania odrzucone z powodu braku czasownika. Wyniki przeprowadzonych testów przedstawiono zarówno w tabelach jak również na wykresach i szczegółowo skomentowano. Oszacowano także złożoność obliczeniową zaproponowanego algorytmu, mianowicie: pesymistyczna złożoność obliczeniowa wynosi $O(3^n)$, gdzie n oznacza liczbę tokenów w zdaniu. Przeprowadzono dyskusję dotyczącą modyfikacji zaproponowanego algorytmu w celu obniżenia jego złożoności obliczeniowej.

W rozdziale trzecim przedstawiono także wyniki testów rozstrzygania wieloznaczności z wykorzystaniem schematu syntaktycznego odpowiednika aspektowego czasownika (na przykład: odpowiednikiem aspektowym czasownika *czytać* jest *przeczytać*, a czasowniki *czytać* i *przeczytać* stanowią parę aspektową). Wyniki przeprowadzonych testów przedstawiono w tabelach i szczegółowo skomentowano.

Rozdział trzeci stanowi oryginalną pracę naukową Doktoranta, w którym wykazano dwie pierwsze tezy rozprawy.

W rozdziale czwartym przedstawiono informacje o aspekcie polskiego czasownika w zakresie niezbędnym z punktu widzenia problematyki rozprawy (badania nad aspektem czasowników tworzą obecnie dziedzinę językoznawstwa zwaną aspektologią). Opisano korzyści wynikające z rozwiązania problemu automatycznego wykrywania odpowiedników aspektowych czasownika. Omówiono stan badań w zakresie automatycznego generowania derywacji słów. Zaproponowano algorytm automatycznego wykrywania par aspektowych czasownika. Założono, że wejściem dla zaproponowanego algorytmu są czasowniki w postaci bezokoliczników. Założono również, że wyjściem zaproponowanego algorytmu są pary aspektowe czasowników wejściowych lub informacja że dany czasownik nie posiada odpowiednika aspektowego. Przedstawiono schemat działania zaproponowanego algorytmu sprowadzającego się do następujących trzech podstawowych etapów: rozpoznawania aspektów wszystkich czasowników (na podstawie słownika CLP); wykrywania derywacji czasowników; wyznaczania par aspektowych dla poszczególnych czasowników. Każdy z etapów został szczegółowo opisany w kolejnych podrozdziałach rozprawy. Omówiono w nich zagadnienia budowy drzew derywacyjnych i wybór par aspektowych. Wyniki przeprowadzonych testów przedstawiono zarówno w tabelach jak również na wykresie i szczegółowo skomentowano. Oszacowano także złożoność obliczeniową zaproponowanego algorytmu, mianowicie: $O(n^2)$, gdzie n jest liczbą wszystkich czasowników w słowniku. Opisano również inne, następujące metody rozwiązania problemu automatycznego wykrywania par aspektowych: algorytmy oparte na osadzeniach i algorytmy oparte na morfologii. Wyniki przeprowadzonych eksperymentów za pomocą tych algorytmów zostały porównane z wynikami uzyskanymi za pomocą algorytmu zaproponowanego przez Doktoranta. Wynika z nich, że algorytm zaproponowany przez Doktoranta daje lepsze wyniki.

Rozdział czwarty stanowi oryginalną pracę naukową Doktoranta, w którym wykazano trzecią (ostatnią) tezę rozprawy.

Rozdział piąty zawiera podsumowanie rozprawy i propozycje kierunków dalszych badań.

4. Oryginalny wkład naukowy Doktoranta

Przedstawiona w rozprawie analiza wyników przeprowadzonych eksperymentów potwierdza, że wszystkie tezy rozprawy zostały wykazane. Na oryginalny dorobek naukowy Doktoranta przedstawiony w rozprawie w zakresie problematyki automatycznego rozstrzygania wieloznaczności leksykalnej w tekstach pisanych w języku polskim składają się:

1. Zaproponowanie, implementacja i przetestowanie metody rozstrzygania wieloznaczności leksykalnej w tekstach pisanych w języku polskim. Podejście to wykorzystuje relacje semantyczne pomiędzy symbolami stanowiącymi zdanie. Relacje te uwidaczniają się w składni i są opisywane za pomocą schematów syntaktycznych. Zaproponowany algorytm rozpoznaje równolegle części mowy i frazy rzeczownikowe co jest nowatorskim podejściem do problemu. Stawiane następnie hipotezy odnośnie do fraz rzeczownikowych uwzględniają ograniczenia wynikające z możliwych opisów morfosyntaktycznych tokenów. Hipotezy te są uwierzytelniane na podstawie schematów zdaniowych. Najlepsza hipoteza wyznacza zakres fraz rzeczownikowych jak również części mowy tokenów wchodzących w ich skład. Zaproponowana metoda rozstrzyga wieloznaczność tylko na granicy części mowy, ale ze względu na złożoność problematyki ograniczenie to nie umniejsza osiągnięcia naukowego Doktoranta.

2. Wykazanie, że w przypadku braku schematów syntaktycznych dla danego czasownika występującego w tekście, można wykorzystać schemat odpowiednika aspektowego danego czasownika, co zwiększa przydatność zaproponowanej metody rozstrzygania wieloznaczności leksykalnej w tekstach pisanych w języku polskim.

3. Zaproponowanie, implementacja i przetestowanie algorytmu automatycznego wykrywania par aspektowych czasownika.

5. Podsumowanie

Podsumowując, stwierdzam że:

1. Tezy rozprawy zostały wykazane.

2. Rozprawa stanowi oryginalne rozwiązanie problemu naukowego w zakresie automatycznego rozstrzygania wieloznaczności leksykalnej w tekstach pisanych w języku polskim na co składają się osiągnięcia Doktoranta wymienione w rozdziale czwartym niniejszej recenzji, w punktach od 1 do 3.

3. Analiza aktualnego stanu wiedzy w zakresie problematyki rozprawy została przeprowadzona na podstawie literatury w sposób wyczerpujący, świadczący o dobrej znajomości tematyki.

4. Rozprawa dowodzi, że Doktorant posiada ogólną wiedzę teoretyczną w dyscyplinie informatyka techniczna i telekomunikacja.

5. Rozprawa dowodzi również, że Doktorant wykazał się umiejętnością samodzielnego prowadzenia pracy naukowej.

Na podstawie punktów od 1 do 5 podsumowania niniejszej recenzji stwierdzam, że recenzowana praca doktorska spełnia ustawowe wymagania stawiane rozprawom doktorskim. W związku z tym wnioskuję o przyjęcie rozprawy doktorskiej i dopuszczenie Pana mgra inż. Zbigniewa Kalety do publicznej jej obrony w dziedzinie nauk inżynieryjno-technicznych w dyscyplinie informatyka techniczna i telekomunikacja.

Marek Skomorowski