11.10.2015

dr inż. hab. prof. (UAM) Grażyna Demenko
Instytut Językoznawstwa UAM
Zakład Fonetyki
al. Niepodległości 4
61-874 Poznań

Recenzja rozprawy doktorskiej mgr inż. Pawła Chrząszcza

Automatyczna ekstrakcja i klasyfikacja semantyczna wielosegmentowych jednostek leksykalnych języka naturalnego

1. Ocena wyboru tematu, tezy i zakresu pracy

Modelowanie struktur gramatycznych języka na potrzeby różnorodnych aplikacji technologii języka takich jak np. automatyczne tłumaczenie tekstu, streszczanie, indeksowanie dokumentów, wyszukiwanie informacji w tekstach ciągle nie jest wykorzystywane w zadawalający sposób. Jedną z podstawowych przyczyn niskiej efektywności modelowania lingwistycznego są trudności związane z formalizacją opisu gramatyki języka, zwłaszcza języka polskiego. W związku z tym, zauważa się w ostatnich latach rozwiązania oparte głównie na metodach korpusowych umożliwiających uczenie systemów na bazie statystyczno-matematycznych algorytmów bez konieczności bezpośredniego uwzględniania złożoności językowych struktur. Należy więc stwierdzić, że ukierunkowanie pracy na automatyczne wyodrębnianie leksykalnych jednostek wielosegmentowych na podstawie badań korpusowych, które przyjął Doktorant jest słuszne, zgodne z obowiązującymi tendencjami na świecie

i dające możliwość dalszego rozwoju. Temat rozprawy *Automatyczna ekstrakcja i klasyfikacja semantyczna i wielosegmentowych jednostek leksykalnych języka naturalnego* jest bardzo ważny zarówno z uwagi na teoretyczne aspekty przetwarzania informacji lingwistycznych, jak i praktyczne implementacje. Rozprawa dotyczy badań odnoszących się do specyfiki języka polskiego, w związku z tym można rozważyć uściślenie tematu rozprawy, który wydaje się w tym przypadku zdecydowanie zbyt ogólny.

W ocenie metodologii, muszę podkreślić, że mgr inż. Paweł Chrząszcz łączy w swojej pracy dobrze stosowane podejście teoretyczne w zakresie przetwarzania informacji lingwistycznych z rzetelnymi badaniami eksperymentalnymi i weryfikuje je systematycznymi testami w opracowanym przez siebie środowisku programistycznym.

Dysertacja mgra Chrząszcza dotyczy efektywnego wyodrębniania i rozpoznawania wielosegmentowych jednostek języka i jest pracą, która wpisuje się bardzo dobrze w nurt nowoczesnej inżynierii lingwistycznej. Zamierzeniem Autora, oprócz przeprowadzenia na przykładzie języka angielskiego analizy wybranych dotychczasowych metod ekscerpcyjnych leksykalnych jednostek wielosegmentowych (zwanych dalej WW), przedstawienia dyskusji dotyczących narzędzi do przetwarzania tekstu (ich możliwych typologii oraz zastosowań itp.), jest zaproponowanie na podstawie uzyskanych danych własnej metody tworzenia słownika WW, jego przetestowanie oraz implementacja. Dane te pozyskiwane są zawsze ze względu na pewne żądanie ekscerpcyjne. To żądanie w pracy mgra Chrząszcza określić można w skrócie: jest nią wielowyrazowa jednostka języka (inne nazwy, jakie pojawiają się w literaturze przedmiotu, to: wielowyrazowiec, kolokacja, idiom, jednostka nieciągła czy ostatnio, np. w pracach W. Chlebdy – reprodukt). Bardziej precyzyjnie można powiedzieć, że jest to wielowyrazowa jednostka języka w jego odmianie starannej, niespontanicznej, wystandaryzowanej. Do tego celu Autor dysertacji przyjmuje specyficzny korpus danych, tj. Wikipedię. To źródło danych, jak uzasadnia w rozdziale 4, zapewnia mu dostatecznie reprezentatywny zasób językowy, dla którego rozwija swoje metody ekscerpcyjne.

Jakkolwiek w każdym wypadku mgr Chrząszcz ma na uwadze automatyczne pozyskiwanie danych, tj. przedstawienie takich metod, w których wynikiem będzie realny wynik w postaci konkretnych połączeń wyrazów, to każde takie techniczne rozwiązanie muszą poprzedzić pewne rozważania teoretyczne. Należy odpowiedzieć na dwa istotne pytania: w jaki sposób wyodrębniać wielowyrazowe jednostki języka? I w drugim etapie: w jaki sposób zaaplikować automatycznie wiedzę o tych jednostkach w słownik wyrazów wielosegmentowych.

Autor dysertacji formułuje przejrzyście tezy pracy:

Teza 1: *Możliwe jest opracowanie algorytmu ekstrahującego w sposób automatyczny wyrazy wielosegmentowe z tekstu w języku polskim, wykorzystującego jako źródła danych słownik fleksyjny i Wikipedię.*

Teza 2: *Możliwe jest utworzenie w sposób automatyczny słownika wyrazów wielosegmentowych z haseł Wikipedii oraz wyrazów wielosegmentowych przy pomocy algorytmu opisanego w tezie 1.*

Podstawowym celem pracy było zbadanie możliwości ekstrakcji wyrazów wielosegmentowych, w zależności od specyficznych właściwości użytego w tekście języka. Autor dysertacji wyodrębnia cztery główne składowe rozwiązania tego problemu: metodę DM (dictionary matching), metodę pDM (z wykorzystaniem wzorców odmiany), metodę SM (ekstrakcję WW w oparciu o wzorce syntaktyczne) oraz SDM (w oparciu o rozszerzony słownik).

Autor załącza starannie opracowane, realne wyniki testów, które pozwalają ocenić wartość zaproponowanych metod. Do istotnych zalet recenzowanej rozprawy należy to, że samodzielnie wprowadza w każdym z rozdziałów podsumowanie szczegółowego rozważenia, które z omawianych rozwiązań mają jakie zalety, a jakie mankamenty. Co więcej, sam Autor czyni tego typu zestawienia porównawcze i wskazuje, jakie metody zasługują na szczególną uwagę badaczy z dziedziny przedmiotu, a które metody należy poddać dalszym ulepszeniom.

Tezy pracy zostały w zupełności udowodnione. Metodologia zaproponowana w dysertacji potencjalnie otwiera nowe możliwości, które mogłyby być wykorzystywane nie tylko w różnorodnych aplikacjach technologii języka i mowy, ale również w dalszych badaniach związanych z frazeologią oraz teorią i praktyką automatycznego przetwarzania tekstu. Warto podkreślić, że opracowane przez mgra Chrząszcza metodologie automatycznego wyodrębniania elementów wielosegmentowych stanowią dla języka polskiego pionierskie rozwiązania, mogą stanowić również podstawę dalszego rozwoju uogólnionych technik i narzędzi przetwarzania języków fleksyjnych.

Zakres rozprawy w pełni został określony, można byłoby rozważyć jego poszerzenie o aspekty lingwistyczne związane z problemami frazeologii w języku polskim, co oczywiście podniosłoby jeszcze walory rozprawy, jednakże zdecydowanie wykraczałoby poza wymagany zakres badań na uczelni technicznej. Skala trudności rozwiązywanych problemów, zakres badań w pełni kwalifikują niniejszą rozprawę jako dysertację doktorską.

2. Ocena osiągnięć i uwagi dyskusyjne

Przed przystąpieniem do omówienia poszczególnych rozdziałów należy podkreślić, że wszystkie etapy badań zostały bardzo starannie zrealizowane i połączone w spójną całość. Praca zawiera znaczny materiał informacyjny, który jest wartościową obudową treści oryginalnej stanowiącej własne osiągnięcia Doktoranta.

Rozprawa została dobrze napisana pod względem kompozycji całości, stanowi nawet dla czytelnika nie zorientowanego w dziedzinie przedmiotu dość przejrzystą lekturę. Została również bardzo starannie opracowana technicznie, zwłaszcza graficznie. W wielu jednakże przypadkach poprawność i staranność stylistyczna zawodzi i pojawiają się nieoczekiwane potknięcia, jak niejasne opisy, omyłki (do których z recenzenckiego obowiązku ustosunkuję się w uwagach szczegółowych).

Dysertacja obejmuje łącznie 12 rozdziałów oraz staranny dodatek ilustrujący przegląd funkcjonalności zrealizowanego przez mgra Chrząszcza projektu. Warto byłoby wziąć pod uwagę zweryfikowanie układu rozdziałów oraz podrozdziałów, których tytuły nie zawsze wydają się przejrzyste (np. 7.4 *Rozpoznawanie wzorców słownikowych w tekście* i podobnie brzmiący podrozdział 7.4.2 *Rozpoznawanie wzorców w tekście,* oraz 9.2 *Statystyka wystąpień wzorca syntaktycznego* i podobny, zbędnie utworzony dla tego rozdziału, podrozdział 9.2.1 *Tworzenie statystyki wystąpień*).

Dysertacja zawiera dobrze dobrany zestaw wykorzystanej literatury w zakresie normalizacji przetwarzania języka naturalnego, narzędzi, słowników nie tylko dla języka angielskiego, ale również polskiego. Co więcej, podano również kilka fundamentalnych prac zajmujących się problemami języka naturalnego (np. publikacje Chomsk'ego), co należy w odniesieniu do opracowania technicznego podkreślić z uznaniem.

Odnotowałam jednakże dość istotny mankament w wykorzystanej w rozprawie bibliografii związany z brakiem nawiązania do badań polskich językoznawców. Przykładowo prace wybitnych polskich językoznawców, np. prof. Grochowskiego, prof. Bogusławskiego, prof. Chlebdy ujmują dość szczegółowo podstawowe problemy związane z teoretycznymi i praktycznymi trudnościami wyodrębniania jednostek wielowyrazowych. Jest to jednak raczej mój konieczny komentarz niż zarzut, ponieważ naturalna wydaje się koncentracja Autora rozprawy wyłącznie na opracowaniach technicznych. Mając na uwadze jednak w przyszłości dalsze badania związane z uogólnieniem metodologii przetwarzania jednostek wielowy-

razowych, warto byłoby przyjrzeć się również językoznawczym opracowaniom poświęconym frazeologii polskiej.

Rozdział pierwszy: ***Wstęp*** (liczący 5 stron) zawiera wprowadzenie do zagadnienia ekstrakcji wyrazu wielosegmentowego. Autor wyjaśnia wykorzystane w pracy pojecie wyrazu wielosegmentowego, jako jednostki języka posiadające własne znaczenie, często inne niż odrębne znaczenia poszczególnych wyrazów. Tabela.1.1 ilustruje narzędzia umożliwiające przetwarzanie tekstu w języku polskim ze szczególnym uwzględnieniem przedmiotu rozprawy. Cele badań oraz tezy rozprawy zostały sformułowane przejrzyście. Duży niedosyt jednakże wywołują bardzo skąpe wyjaśnienia odnośnie problemu definicji wyrazu wielosegmentowego (do tego zagadnienia odniosę się bardziej szczegółowo, przy omówieniu rozdziału 6).

Rozdział drugi: ***Stan badań*** (14 stron) stanowi przegląd stanu wiedzy o problemach przetwarzania języka naturalnego, aspektach klasyfikacji i ekstrakcji wyrażeń wielosegmentowych, w tym także dla języka polskiego. Zdecydowanie jest zbyt syntetyczny, warto byłoby ten fragment poszerzyć o krótkie omówienie gramatycznych uwarunkowań polskiego. Modelowanie struktury lingwistycznej języków fleksyjnych stanowi poważny problem, szczególnie języka polskiego, w którym liczne niejednoznaczności (np. syntaktyczne, semantyczne) i relatywnie swobodny szyk frazy utrudniają analizy. Wyrazy zawierają mniej lub bardziej wyraźnie oddzielony temat oraz końcówki wyrażające odmienne kategorie gramatyczne. Dodatkowo, w odmianie, temat wyrazu często uzależniony jest od licznych i nieregularnych przekształceń, ta sama końcówka w języku polskim może pełnić odmienne funkcje.

Autor rozważa charakterystyki analizowanych języków naturalnych, odnosi się do wyników badań dla angielskiego, jednakże zbyt syntetycznie zwraca uwagę na specyficzne uwarunkowania języka polskiego (np. problem wieloznaczności) i w rezultacie brak jest komentarza odnośnie lingwistycznych analiz nieprawidłowych rozpoznań jednostek WW w ogólnej ewaluacji zaprezentowanej w dysertacji metodologii.

Cel ogólny badań (str.17): *naszym celem jest ekstrahowanie dowolnego rodzaju wyrazów wielosegmentowych, bez ograniczeń i kategoryzacji* sformułowany jest zbyt ogólnie i niejasno.

Autor rozprawy szczegółowo dyskutuje również możliwość zastosowania narzędzia SPEJD do rozpoznawania wielosegmentowych jednostek w tekście. Uzasadnia potrzebę

opracowania leksykonu takich jednostek dla języka polskiego. Kolejno, w rozdziale drugim przedstawia przekonywującą dyskusję na temat wykorzystania Wikipedii, jako źródła danych leksykalnych i semantycznych w odniesieniu do jednostek WW. Autor świadomie jednak podkreśla mankamenty tego rozwiązania i konieczność posiłkowania się dodatkowymi źródłami danych.

Rozdział 3: N*arzędzia słownikowe* (8 stron) poświęcony jest zasobom słownikowym stanowiącym podstawę algorytmów zaimplementowanych w automatach rozpoznających jednostki WW. Autor przekonywująco uzasadnia wybór *Słownika Fleksyjnego Języka polskiego (SFJP)* oraz biblioteki CLP stanowiącej interfejs programistyczny do tego słownika. Dyskutuje uzupełnienie zasobów poprzez włączenie słownika Morfologika dla optymalizacji opracowywanych przez siebie technik i umiejscawia właściwie swoje narzędzia badawcze.

Mimo, iż Autor dysertacji zajmuje się głównie problemami technicznymi ekstrakcji jednostek WW, automatycznym tworzeniem słownika, należy z uznaniem podkreślić, iż poprawnie odnosi się również do wybranych analiz językowych i uświadamia sobie trudności związane z analizą specyficznych informacji lingwistycznych.

W łącznej ocenie rozdziałów 1-3 wstępnych stwierdzam, że są one przykładem samodzielnego syntetycznego wprowadzenie do zagadnień inżynierii lingwistycznej i świadczą o dobrym zorientowaniu Autora rozprawy w dziedzinie przedmiotu.

W **rozdziale 4: *Ekstrakcja danych, z Wikipedii i Wisłownika*** (8 stron), oraz w **rozdziale 5: *Ekstrakcja etykiet semantycznych dla haseł Wikipedii*** (10 stron) mgr Chrząszcz szczegółowo zapoznaje czytelnika z problemem pozyskania właściwego materiału lingwistycznego i uzasadnia wybór Wikipedii, jako źródła pozyskiwania jednostek WW. Można by odnieść wrażenie, że opis ten jest tylko przeglądem problematyki wynikającym z pewnego obowiązku dysertacyjnego. Tak nie jest, ponieważ P. Chrząszcz dzięki niemu może umiejscowić swój własny korpus badawczy, wskazać zarówno zalety jak i mankamenty wybranego rozwiązania.

Począwszy od **rozdziału 6: *Problem ekstrakcji wyrazów wielosegmentowych*** (6 stron) Autor zapoznaje Czytelnika już z własnymi rozważaniami dotyczącymi kategoryzacji i ekstrakcji wyrażeń wielosegmentowych. Wyróżnia 4 kategorie jednostek WW

przyjętych dla języka angielskiego i na tej podstawie precyzuje kategorie dla języka polskiego stanowiące przedmiot rozprawy. Ogranicza się do: *wyrażeń o ustalonej strukturze, funkcjonujących, jako jednostka słownikowa posiadająca własne znaczenie i odmieniająca się poprzez odpowiednie formy gramatyczne* i zawęża badane klasy do wyrażeń pełniących role rzeczownika. W tabeli 6.1 podaje przykłady wyrażeń wielosegmentowych których ekstrakcja jest przedmiotem rozprawy.

Przydatne byłoby umieszczenie w tej części dysertacji szerszego omówienia w odniesieniu do problemu semantycznej interpretacji wyrażenia wielowyrazowego dla języka polskiego. W szczególności warto zwrócić uwagę na związki frazeologiczne z jednostkami jedno i wielosegmentowymi języka.

Jak pisze prof. Maciej Grochowski (str.*34 Sbornik praci filozoficke fakulty brnenske untverzity studia minora facultatis philosophicae universitatis brunensis, a 29, 1981):*

„Jedyna różnica, jaka istnieje między jednosegmentowymi wyrazami (będącymi jednostkami języka) a związkami frazeologicznymi, dotyczy ciągłości tych obiektów: związek frazeologiczny jest nieciągłą (tzn. zawierającą pauzy) sekwencją elementów diakrytycznych. Jest to, więc różnica wyłącznie zewnętrzna, irrelewantna z punktu widzenia zarówno semantycznego, jak i syntaktycznego. Nieuzasadnione jest, zatem rejestrowanie związków frazeologicznych w artykułach hasłowych poświęconych wyrazom przypadkowo tylko zbieżnym graficznie z różnymi segmentami związków. Np. jednostka dziadek do orzechów nie ma nic wspólnego z jednostką dziadek, tak samo, jak jednostka nie ma za grosz nie pozostaje w żadnej relacji do jednostki grosz, a fakt, że kształt danej jednostki jednosegmentowej jest identyczny z kształtem części pewnej innej jednostki wielosegmentowej, nie wynika z żadnych, ani gramatycznych, ani semantycznych, reguł języka.
Związki frazeologiczne, tak samo, jak jednosegmentowe wyrazy uznane za jednostki języka, są niepodzielne na mniejsze odcinki znaczące, stanowią budulec produktów językowych, służą do generowania tekstu, mogą charakteryzować' się określonymi własnościami walencyjnymi."

Autor dysertacji podaje definicję wyrazu wielosegmentowego przyjętą na potrzeby przeprowadzanych badań: *wyrazy wielosegmentowe WW można zdefiniować jako wyrażenia odmienne, o dokładnie zdefiniowanej ustalonej strukturze, odmieniające się w całości jak rzeczowniki, pełniące w tekście rolę rzeczowników i posiadające określone, stałe znaczenie.*
Wyróżnia 2 zasadnicze kategorie w odniesieniu do rzeczownika wielosegmentowego: segmenty odmienne tworzące główną cześć definicji oraz nieodmienne do których zalicza pozostałe segmenty, których forma nie zmienia się niezależnie od formy gramatycznej całego wyrazu.

Powyżej przyjęte ograniczenia gramatyczne pozwalają uniknąć niejasności wynikających z problemów segmentacji jednostek wyrazowych ujmowanych w polskiej frazeologii.

Rozdziały 7 i 8 dotyczą metod ekstrakcji wyrazów wielosegmentowych na bazie haseł z Wikipedii: metody DM (dictionary matching) oraz metody pDM (z wykorzystaniem wzorców odmiany).

Zaprezentowane **w rozdziale 7: *Metoda DM*** (17 stron) wzorce słownikowe dla jednostek WW utworzono na podstawie automatu Moore'a według własnego algorytmu.

Samo stosowanie automatów skończonych, dla analizy połączeń wyrazowych jest w językoznawstwie korpusowym relatywnie często wykorzystywane. Jest to drugi, obok gramatyki, model obliczeń będący przedmiotem badań teorii języków. Analizowany jest model na podstawie którego możliwe jest sformułowanie algorytmu rozstrzygającego, czy dowolne słowo należy, czy też nie należy do języka rozpoznawanego przez ten automat.

Na podstawie zaproponowanego przez siebie automatu mgr inż. Chrząszcz przedstawia propozycję usystematyzowania metodologii, pozwalającej na relatywnie łatwe kształtowanie mechanizmu tworzenia wzorca WW, dopasowując go do konkretnego problemu.

Założono, że automat, powinien z każdym analizowanym segmentem dokonywać przejścia stanowego i pozwalać na rozpoznanie kilku wzorców w jednym stanie (z dodatkowym założeniem, że jeden wzorzec może być częścią innego). Z uwagi na te uwarunkowania przyjęto do analiz automat Moore'a, gdzie sygnał wyjściowy jest wyłącznie funkcją stanu.

Określony algorytm możne zrealizować różnymi strukturami automatów; czy Doktorant rozważał inne automaty, modele grafowe dla uzyskania optymalnej architektury. Biorąc pod uwagę, iż w przyszłości planowane jest wykorzystanie kontekstu czy efektywne byłoby zastosowanie innego automatu np. gdzie sygnał wyjściowy zależy od wejściowego i stanu automatu np. automatu Mealy'go.

W **rozdziale 8: *Metoda pDM*** (14 stron) uzupełniono metodę *DM* z uwzględnieniem sposób odmiany wyrazów w sposób automatyczny, na podstawie linków w Wikipedii. Szczegółowo omówiono tworzenie wzorców odmiany, optymalizacje wyznaczania wzorców oraz przedstawiono szczegółowe wyniki testów. Podkreślono wysoką precyzję metody pDM (ponad 90%), jednakże przy jakości rozpoznawania wynoszącej 4 na 10 wyrazów wielosegmentowych w tekście.

W **rozdziale 9:** ***Metoda SM*** (20 stron) przedstawiono ekstrakcję WW w oparciu o wzorce syntaktyczne oraz w **rozdziale 10:** ***Metoda SDM*** (12 stron) zoptymalizowano ekstrakcję jednostek WW w oparciu o poszerzony słownik.
Podana statystyka dotyczy wystąpień jednostek WW oraz tworzenia wzorców w syntaktycznym kontekście. Zwracają uwagę bardzo starannie opracowane podpunkty: 9.5 w którym szczegółowo omówiono procedurę rozpoznawania wzorców syntaktycznych oraz 9.7 przedstawiający systematycznie, rzetelnie przeprowadzone testy.
Ostateczne testy (punkt 10.2 oraz 10.3) zostały wykonane na bazie reprezentatywnych zbiorach danych. Analizowano wiele różnych wariantów łączenia metod, w rezultacie test metody łączonej został uznany za najbardziej efektywny. W poszczególnych podpunktach rozdziału 10 Autor prezentuje wyniki dla badanych zbiorów dokumentując szczegółowymi tabelami i wnioskami, który z wariantów daje najlepsze rezultaty.

W **rozdziale 11:** ***Możliwości dalszego rozwoju*** (8 stron) Autor po analizie osiągniętych rezultatów, proponuje preferowane metody dla każdego z rozpatrywanych wariantów ekstrakcji jednostek WW. Jako najbardziej obiecującą metodę uznaje SM w oparciu o strukturę wzorców syntaktycznych. W przypadku metod DM i pDM wykorzystujących słownik utworzony z haseł Wikipedii możliwość poprawy może zaistnieć poprzez wykorzystanie manualnie utworzonych reguł (potwierdza to Tabela 7.4 ilustrująca testy rozpoznawania).

Utworzone środowisko badawcze, integrujące wszystkie wymagane komponenty oraz dostarczające samodzielnie szereg dodatkowych narzędzi, pozwoliło to na sprawne weryfikowanie stawianych hipotez roboczych oraz badanie wpływu różnych technik ekstrakcji jednostek WW. Cenna dyskusja nad dalszym postępem ekstrakcji jednostek WW uwzględniająca różne metody (iteracyjne, nie iteracyjne), uczenie z nadzorem i bez nadzoru, stosowanie dodatkowych korpusów i słowników, zastosowanie etykietyzacji semantycznej świadczy o odpowiednich kompetencjach mgra Chrząszcza w dziedzinie przedmiotu.

W rozdziale 12 ***Podsumowanie*** (4 strony) Autor przedstawia efekty pracy. Jako najważniejsze osiągnięcia prowadzonych badań podaje: algorytm wyznaczający wzorce odmiany haseł Wikipedii, tworzenia wzorców słownikowych i syntaktycznych, algorytm tworzący i rozpoznający wzorce, automat ekstrahujący wyrazy wielosegmentowe.

Wskazuje, że użyty sposób ekstrakcji określonych jednostek językowych może być także zastosowany w innych niż klasyfikacja, problemach lingwistyki komputerowej. Uzy-

skane efekty otwierają szereg kierunków związanych z dalszym wykorzystaniem proponowanego podejścia w modelowaniu lingwistycznym np. w systemach syntezy oraz rozpoznawania mowy.

W obszernym **Dodatku** ilustrującym ***Przegląd funkcjonalności projektu*** (20 stron) autor odnosi się: w punkcie A.1 do *Tworzenia słownika CLPM,* w punkcie do 1..3.2. do *Ekstrakcji danych,* w punkcie *1.3.3.* do *Tworzenia etykiet semantycznych,* w punktach *13.4-13.9 do ekstrakcji WW* , w punkcie 13.9 do *Testowania metod,* w punkcie *13.10* do *podsumowania.*
Zwracają uwagę starannie dobrane przykłady ilustrujące funkcje opracowanego przez Doktoranta oprogramowania.

Uwagi szczegółowe:

1. Przydatny byłby spis tabel i rysunków. Przykładowo, nie odnalazłam w tekście tabeli 7.3.
2. Niejasne sformułowania. Praca pisana jest w liczbie mnogiej, nie zawsze więc wiadomo, kiedy Autor powołuje się na swoje badania, a kiedy na wyniki prac zespołowych.
3. Liczne błędy stylistyczne. Przykładowo:
   - *W przypadku języka polskiego póki co* .. str.21
   - *Jednak ciężko ten wynik odnieść*, str.21
   - *Podsumowując dotychczasowe wyniki możemy stwierdzić, ze nie znaleziono publikacji opisujących rozwiązanie problemu ekstrakcji MWE, gdy nie są dostępne bogate korpusy treningowe, zwłaszcza dla języka polskiego* str.28.
   - *Rozważano również użycie tagerów- co prawda wówczas, gdy rozpoczynano badania (rok 2009), podejmowane były jedynie próby tagowania języka polskiego tagerami utworzonymi głównie dla języka angielskiego, dające mieszane wyniki [39 , jednak od tego czasu został wykonany duży postęp* (str. 29)

Podsumowując, recenzowana praca doktorska jest oryginalnym, ambitnym rozwiązaniem zaprezentowanego w niej problemu badawczego mgr Chrząszcz podjął się opraco-

wania oryginalnych metod automatycznej analizy jednostek WW. Po przestudiowaniu dysertacji nasuwają się następujące pytania.

1) W jakim zakresie zaproponowane metody mogą się stosować ogólnie do języków fleksyjnych (prócz przykładów omówionych w pracy dla języka polskiego)?
2) Czy po przeprowadzonej analizie wyników można już dziś określić kierunek rozwoju metodologii zmierzającej do rozwiązania automatycznej analizy związku frazeologicznego (w odniesieniu do połączeń wielowyrazowych)?
3) Czy zaproponowana architektura automatu Moore'a będzie optymalna dla analiz innych wyrażeń (np. nie pełniących funkcji rzeczownika).
4) Jakie mogą być przyczyny prawdopodobnego niewykorzystywania metod automatycznych pozyskiwania danych w implementacji modeli lingwistycznych w systemach technicznych.?

Oryginalnym osiągnięciem mgra Chrząszcza są następujące elementy:

1. Opracowanie metod ekscerpcyjnych jednostek WW i algorytmu ekstrahującego w sposób automatyczny wyrazy wielosegmentowe z tekstu w języku polskim.
2. Prezentacja realnego wyniku ekscerpcyjnego w postaci *słownika* nienotowanych leksykograficznie wyrazów wielosegmentowych dla języka polskiego, który może okazać się niezwykle przydatny dla dalszych badań związanych z przetwarzaniem tekstu lub dla implementacji modeli lingwistycznych w systemach technologii języka i mowy.
3. Opracowanie środowiska programistycznego dla laboratoryjnego testowania metody *wyrazów wielosegmentowych*, analizy zastosowanych metryk i porównywania statystycznej istotności uzyskanych rezultatów
4. Opracowanie oryginalnych struktur automatów Moore'a pozwalających efektywnie wykorzystać zaproponowane metody ekscerpcyjne,
5. W pracy przedstawiono pierwsze dla polskiego narzędzie ekstrahujące z tekstu wyrazy wielosegmentowe pełniące role rzeczownika.
6. Przedyskutowanie osiągniętych wyników (autor nie absolutyzuje swoich metod, rzetelnie ujawnia ich mankamenty i zalety; ujawnia wydajność każdej z nich),

Mgr. Paweł Chrząszcz podjął się opracowania oryginalnych metod ekscerpcyjnych i przeprowadził analizy, które świadczą o dobrej znajomości przynajmniej dwu dyscyplin: inżynierii języka i informatyki stosowanej. Dodatkowo, dla realizacji zadanego celu, wymagana była wiedza z zakresu formalizacji opisu gramatyki języka polskiego. Publikacja dysertacji po uzupełnieniach i poprawkach niewątpliwie byłaby niezwykle przydatna nie tylko dla inżynierów zajmujących się implementacjami systemów technologii mowy, ale również dla lingwistów badających gramatykę języka polskiego.

Podsumowanie

W posumowaniu mojej oceny dysertacji doktorskiej zatytułowanej *Automatyczna ekstrakcja i klasyfikacja semantyczna wielosegmentowych jednostek leksykalnych języka naturalnego* mgr inż. Pawła Chrząszcza stwierdzam, co następuje.
Recenzowana praca doktorska jest oryginalnym, ambitnym rozwiązaniem zaprezentowanego w niej problemu badawczego. Autor podjął się nie tylko opracowania metod klasyfikacji dokumentów ale co jest szczególnie cenne, ich implementacji. Uważam opiniowaną rozprawę jako bardzo wartościową i stanowiąca ważne osiągnięcie dla teorii i praktyki w zakresie informatyki oraz inżynierii lingwistycznej.

Ze względu na systematyczność opracowanej przez Autora metody naukowej, ważnej dla dalszego ukierunkowania badań, jak i przyszłej użyteczności praktycznej oraz niezwykle systematycznie przeprowadzone analizy naukowe, uważam opiniowaną prace z naddatkiem spełniającą wymagania stawiane rozprawom doktorskim i w związku z tym wnoszę o dopuszczenie jej Autora do dalszych etapów przewodu doktorskiego.

Demenko

dr hab. inż. prof. UAM Grażyna Demenko
Uniwersytet Adama Mickiewicza
Poznań, dnia 11.10.2015