prof. dr hab. Marek R. Ogiela
Akademia Górniczo-Hutnicza
al. Mickiewicza 30
30-059 Kraków

Kraków, dnia 08.01.2016 r.

Recenzja rozprawy doktorskiej

Pana mgr inż. Pawła Chrząszcza
pt. „*Automatyczna ekstrakcja i klasyfikacja semantyczna wielosegmentowych jednostek leksykalnych języka naturalnego*”

Przedmiotem niniejszej recenzji jest rozprawa doktorska Pana mgr inż. Pawła Chrząszcza zatytułowana „*Automatyczna ekstrakcja i klasyfikacja semantyczna wielosegmentowych jednostek leksykalnych języka naturalnego*”. Recenzja została napisana na podstawie zlecenia Prodziekana Wydziału Informatyki, Elektroniki i Telekomunikacji AGH, prof. dr hab. inż. Katarzyny Zakrzewskiej (pismo z dnia 10.07.2015 r.).

1. Zakres rozprawy i jej cel

Opiniowana rozprawa dotyczy zagadnień związanych z komputerową analizą i ekstrakcją wielosegmentowych wyrażeń leksykalnych w tekstach języka polskiego, w oparciu o klasyfikację semantyczną.

Podstawowym celem naukowym badań przedstawionych w niniejszej rozprawie doktorskiej, było opracowanie zestawu nowych metod wyodrębniania takich jednostek (nazywanych w pracy wyrazami wielosegmentowymi) za pomocą odpowiedniej konstrukcji etykiet semantycznych, zdefiniowania nowych formatów baz danych i tworzonych z pomocą prezentowanych w rozprawie metod, słowników wyrażeń wielosegmentowych. Zaprezentowane metody uwzględniają różnorodne aspekty lingwistyczne badanych jednostek tekstu, co w kolejnych krokach pozwoliło na opracowanie i zbadanie kilku podstawowych algorytmów ekstrakcji, a następnie stworzenie rozwiązania hybrydowego osiągającego najwyższe parametry skuteczności i dokładności ekstrakcji poszukiwanych jednostek wielosegmentowych, a także pozwoliło na stworzenie słownika wyrażeń wielosegmentowych.

Badania Autora skoncentrowały się na wykazaniu następujących tez badawczych:

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

2. Możliwe jest utworzenie w sposób automatyczny słownika wyrazów wielosegmentowych z haseł Wikipedii oraz wyrazów wielosegmentowych wyekstrahowanych przy pomocy algorytmu opisanego w Tezie 1.

Ponieważ w prowadzonych badaniach nad ekstrakcją wielosegmentowych wyrażeń dla języka polskiego problem ten nie został jeszcze całkowicie rozwiązany, podjęte przez Autora badania są w pełni uzasadnione, a niniejsza rozprawa prezentuje autorskie rozwiązania, które znacząco wzbogacają dostępne wcześniej algorytmy analizy i ekstrakcji

jednostek wielosegmentowych. Przeprowadzone badania pozwoliły stwierdzić, z jaką dokładnością można wydzielić poszukiwane jednostki w zasobach Wikipedii, ale także z nieuporządkowanych tekstów języka polskiego bez użycia wcześniej „otagowanych" zbiorów treningowych.

Ważną częścią realizowanej tematyki badawczej były praktyczne implementacje i testy algorytmów pozwalające dokonywać ekstrakcji jednostek wielosegmentowych, które w rezultacie pozwoliły na zdefiniowanie najbardziej skutecznej metody będącej połączeniem metod bazowych i uzyskującej wysokie procentowe wskaźniki precyzji i pełności ekstrakcji. Opracowane algorytmy, jako nowatorskie stały się także punktem odniesienia do dalszych badań w tym zakresie i pozwalają na utworzenie brakujących zasobów zawierających jednostki wielosegmentowe dla języka polskiego.

Po zapoznaniu się z rozprawą można nabrać pewności, że zastosowanie autorskich rozwiązań opisanych w pracy do analizy i ekstrakcji wyrażeń wielosegmentowych znacząco wzbogaca wcześniej znane metody lingwistyki komputerowej wykorzystywane do podobnych zadań. Opracowane algorytmy stanowią doskonałe narzędzia pozwalające na automatyczną ekstrakcję takich jednostek w obszernych repozytoriach internetowych lub bazach danych zawierających teksty języka polskiego.

2. Forma realizacji i układ rozprawy

Rozprawa doktorska jest obszernym opracowaniem składającym się ze 170 stron zawierającym 13 rozdziałów w tym podsumowanie i dodatek oraz na końcu bibliografię.

Rozdziały 1-3 to wstęp do pracy, który zawiera wprowadzenie do zagadnień związanych z określeniem wielosegmentowych jednostek leksykalnych oraz przedstawia tezy pracy, a także rozdział dotyczący stanu badań z zakresu podjętej tematyki, i rozdział opisujący narzędzia słownikowe opracowane na użytek prowadzonych badań.

Rozdziały 4-10 są poświęcone na prezentację metodologii badawczej oraz autorskich rozwiązań zaproponowanych na użytek ekstrakcji wyrażeń wielosegmentowych w różnych zasobach lingwistycznych, oraz przy użyciu różnych metod badawczych. Doktorant prezentuje w tych rozdziałach kolejne zadania ekstrakcji oraz uzyskane rezultaty dzięki zastosowaniu kolejno definiowanych algorytmów ekstrakcji.

Rozdział 11 prezentuje kierunki dalszych badań i możliwości rozwoju zaprezentowanych technik, natomiast rozdział 12 jest podsumowaniem dysertacji pokazującym, w jaki sposób zostały wykazane początkowe tezy pracy oraz prezentuje wszystkie osiągnięcia Autora w trakcie prowadzonych badań.

Ostatni rozdział 13 jest dodatkiem prezentującym funkcjonalność opracowanych procedur i kody źródłowe implementowanych procedur.

3. Oryginalne osiągnięcia badawcze

Za najważniejsze, oryginalne osiągnięcia Autora, przedstawione w recenzowanej rozprawie, uważam:

1. Porównanie możliwości i skuteczności ekstrakcji jednostek wielosegmentowych z zasobów Wikipedii i Wikisłownika i wykazanie większej użyteczności tekstów Wikipedii do prowadzenia skutecznej ekstrakcji takich jednostek.

2. Adaptacja wcześniej opracowanego algorytmu ekstrakcji etykiet semantycznych, dla haseł Wikipedii, który uzyskał ponad 94% dokładność ekstrakcji etykiet semantycznych.

3. Zaproponowanie obszernej metodologii ekstrakcji wyrażeń wielosegmentowych, pełniących rolę rzeczownika i mających ustaloną strukturę (np. wyrażenia niedekomponowalne) oraz stałe znaczenie. Do ekstrakcji takich wyrażeń Autor zaproponował i przebadał kolejno następujące metody:

    - Algorytm DM (Dictionary Matching) rozpoznający wyrazy wielosegmentowe z haseł Wikipedii w oparciu o przejścia automatu skończonego.
    - Metodę pDM uwzględniającą wzorce odmiany oraz wykorzystującą linki przychodzące do haseł.
    - Algorytm SM rozpoznawania wzorców syntaktycznych, który wykorzystuje wzorce odmiany oraz strukturę kontekstu wyrażenia wielosegmentowego. Metoda ta po wyznaczeniu wzorców syntaktycznych poszczególnych wyrazów i ich kontekstów określa także statystyki ich wystąpień. Algorytm ten pozwala także na dobór parametrów analizy mających wpływ na skuteczność ekstrakcji. Technika ta może także stanowić wzorcowy leksykon i posłużyć do rozszerzenia obecnie istniejących słowników fleksyjnych języka polskiego (SFJP). Autor wykorzystał tą metodę do budowy słownika wyrazów wielosegmentowych.
    - Metodę SDM dokonującą ekstrakcji wyrażeń wielosegmentowych w oparciu o rozszerzony słownik generowany przez algorytm SM.

4. Połączenie wybranych bazowych metod ekstrakcji i określenie metody hybrydowej dającej najlepsze rezultaty przy badaniu wyrażeń wielosegmentowych.

5. Bardzo dokładne i systematyczne przestawienie części eksperymentalnej związanej z ewaluacją skuteczności klasyfikacji jednostek leksykalnych na wybranych zbiorach danych, a także opracowanie słownika wyrazów wielosegmentowych i jego udostępnienie środowisku naukowemu do dalszych badań.

4. Uwagi edycyjne

Recenzowana rozprawa jest napisana w języku polskim i charakteryzuje się dużą starannością oraz zwięzłym przedstawieniem tematyki prowadzonych badań. Tematyka ta została przedstawiona w sposób klarowny i systematyczny.

Autor nie ustrzegł się jednak przed popełnieniem kilku drobnych błędów edycyjnych związanych z redakcją pracy. Wśród nich da się zauważyć np.:

- Błędne określenie na str. 12 „wielosegmentowe języki leksykalne” powinno być „wielosegmentowe jednostki leksykalne”.
- Kilka drobnych błędów literowych dostrzeżonych w tekście.

Nieliczne błędy mają jednak znaczenie marginalne i nie wpływają na pozytywną ocenę merytoryczną osiągnięć Autora pracy.

5. Uwagi o charakterze dyskusyjnym

Po uważnym zapoznawaniu się z rozprawą da się łatwo zauważyć, że praca została napisana w sposób bardzo systematyczny.

Czytając rozprawę nasuwają się jedynie kolejne pytania dotyczące omawianych zagadnień, tj.:

1. W opisanych przykładach zastosowania zaproponowanych metod, Autor często podaje przykłady analizy wydzielanych jednostek leksykalnych. Czy procesy takie można całkowicie zautomatyzować dla wszystkich opisanych technik i realizować za pomocą opracowanych programów komputerowych, również dla całkowicie nowych zbiorów danych tekstowych?

2. W systemach ekstrakcji lub klasyfikacji wzorców (w tym przypadku wyrażeń wielosegmentowych) pojawia się kwestia możliwości dalszego uczenia lub rozszerzenia wykorzystywanych algorytmów działających lub wytrenowanych dla dostępnych zbiorów danych. Czy proponowane metody mogą dla nowych wzorców rozszerzać swoją funkcjonalność tzn. czy istnieją możliwości samo-uczenia się tych algorytmów?

3. Czy można określić złożoność czasową procesu ekstrakcji opisanymi metodami, jako funkcję zależną od wielkości danych wejściowych (np. liczebności próbek tekstów w badanej bazie danych lub korpusie)?

4. Czy można oszacować wpływ długości wydzielanych jednostek (bigramy, trigramy etc.) na skuteczność opisanych rozwiązań?

6. Podsumowanie

Postawione pytania nie umniejszają wartości merytorycznej opiniowanej rozprawy i nie wpływają na moją ostateczną, pozytywną opinię o wartościowych osiągnięciach Autora, uzyskanych w badaniach nad wykazaniem tez pracy. Przedstawione w rozprawie wyniki mają duże znaczenie zarówno teoretyczne jak i praktyczne. Wzbogacają one dostępne techniki komputerowej analizy i ekstrakcji wielosegmentowych jednostek leksykalnych, pozwalając na ich szybszą oraz skuteczną ekstrakcję z zasobów zawierających dane tekstowe. Osiągnięcia Autora należą do jednych z pierwszych w zakresie ekstrakcji wyrażeń wielosegmentowych pełniących rolę rzeczownikową, przez co mają charakter nowatorski, a wyniki prowadzonych prac zostały także przedstawione w publikacjach naukowych.

Podsumowując moją recenzję uważam, że opiniowana rozprawa doktorska spełnia wymagania przewidziane dla rozpraw doktorskich w aktualnie obowiązującej Ustawie o Stopniach Naukowych i o Tytule Naukowym, i wnoszę o dopuszczenie Pana mgr inż. Pawła Chrząszcza do dalszych faz przewodu doktorskiego, w celu nadania mu stopnia naukowego doktora nauk technicznych w dyscyplinie informatyka.