Dr hab. Hung Son Nguyen
Wydział Matematyki, Informatyki i Mechaniki
Uniwersytet Warszawski
email: son@mimuw.edu.pl

Warszawa, 27/4/2019

Recenzja rozprawy doktorskiej

Tytuł:

Metoda identyfikacji schematów walencyjnych w tekstach polskich

Autor rozprawy: **Mgr Konrad Gołuchowski**

Promotor: **dr hab. Adam Przepiórkowski**

Rozprawa doktorska została wykonana w:

Instytucie Podstaw Informatyki, Polskiej Akademii Nauk

Rozprawa doktorska mgr. Konrada Gołuchowskiego przedstawia system identyfikacji schematów walencyjnych w tekstach polskich. Autor zaproponował nowoczesną technikę przetwarzania języka naturalnego wykorzystującą różne metody uczenia się. Główna teza rozprawy polega na tym, że identyfikacja schematów walencyjnych metodami maszynowego uczenia się bez ręcznie przygotowanego korpusu treningowego jest możliwa dzięki automatycznej identyfikacji potencjalnych podrzędników predykatów oraz dzięki wydobyciu preferencji selekcyjnych z automatycznie stworzonego korpusu przykładów użyć schematów walencyjnych.

1. ZAKRES ROZPRAWY

Poruszony problem w rozprawie jest typowym zagadnieniem w dziedzinie automatycznego przetwarzania tekstów w języku naturalnym. Dotyczy to problemu wykrywania części mowy w polskim tekście. W recenzowanej rozprawie, autor przyjmował zadanie automatycznej identyfikacji schematów walencyjnych predykatów czasownikowych. Jest to zadanie bardzo złożone ze względu na ilość czasowników i ich schematów walencyjnych. Nadzieje na

praktyczne rozwiązanie tego zadania są związane z techniką uczenia maszynowego, ale nadal pozostaje problem przygotowywania zbiorów treningowych dla tych algorytmów. Dlatego jednym z kluczowych celów jest opracowanie metody automatycznego tworzenia zbiorów treningowych dla algorytmów uczących się.

W mojej ocenie, podjęta problematyka rozprawy jest bardzo ważna i aktualna. Cel pracy został też bardzo jasno i poprawnie sformułowany

2. STRUKTURA ROZPRAWY

Rozprawa doktorska Pana mgr. Konrada Gołuchowskiego liczy 214 stron i składa się z trzech rozdziałów czterech dodatków i spisu literatury.

Pierwszy rozdział przedstawia podstawowe pojęcia i metody w przetwarzaniu języka naturalnego. Autor zdefiniował zadanie identyfikacji podrzędników predykatów oraz zadanie wyboru schematu walencyjnego. To są dwa podzadania głównego problemu rozprawy, czyli zadanie identyfikacji schematu walencyjnego w tekstach polskich. Oprócz tego, autor przedstawił przegląd dostępnych zasobów lingwistycznych użytych do realizacji zadań w rozprawie oraz listę informatycznych narzędzi służących do przetwarzania tekstów polskich.

W drugim rozdziale Pan mgr Konrad Gołuchowski przedstawił autorską metodę identyfikacji schematów walencyjnych czasowników w zdaniach języka polskiego. Proponowana metoda składa się z dwóch kroków, tj. (1) identyfikacji podrzędników predykatów oraz (2) wyboru schematów walencyjnych. Pierwszy krok wykrywa predykaty oraz syntaktyczne i semantyczne elementy głównych podrzędników i ich typów. Drugi krok składa z wielu algorytmów wybierających schemat walencyjny. Są to:

- Algorytm GreatestSchema (i jego wariant) wybierający schemat walencyjny wyłącznie na postawie informacji o znalezionych podrzędnikach za pomocą heurystyki Match.
- Algorytm LexSchema wyszukujący w zdaniu argument leksykalizowany, argument o stałym tekście lub argument będący złożonym przyimkiem. Takie argumenty są na ogół poprawnymi schematami.
- Algorytmy BayesSchema i NegBayesSchema wybierające schematy walencyjne w niejednoznacznych sytuacjach za pomocą klasyfikatora Naive Bayes.
- Algorytmy SRVoteSchema i SROneClassSchema wybierające schematy walencyjne w niejednoznacznych sytuacjach za pomocą głosowania. Pierwszy algorytm głosowania działa w oparciu o trzy algorytmy preferencji selekcyjnych. Można traktować algorytmy preferencji selekcyjnych jako mecze między dwoma schematami walencyjnymi, a algorytm głosowania SRVoteSchema jako cały turniej pomiędzy wszystkimi schematami. Schemat z największą liczbą wygranych wygra rywalizację. Drugi algorytm głosowania korzysta z algorytm klasyfikacji, który każdy schemat walencyjny jest klasyfikowany do jednej z trzech klas: pozytywny, neutralny i negatywny.

- Algorytm VecSchema korzystający z wektorowej reprezentacji słów uzyskanych za pomocą np. algorytmu Word2vec oraz regresji logistycznej. Algorytm ten charakteryzuje się tym, że nie wymaga ręcznie oznaczonego korpusu tekstu.
- Algorytm RFSchema wybierający schemat walencyjny za pomocą techniki lasów losowych (ang. Random Forest).
- Algorytm wyszukiwania podobnych czasowników VR wyszukuje schemat walencyjny przez zastąpienie czasownika w zdaniu czasownikiem o podobnym znaczeniu, ale posiadającym więcej przykładów w korpusie treningowym.
- Meta-klasyfikator VoteMeta wybierający schemat walencyjny przez łączenie poprzednich algorytmów. Istnieje wiele schematów łączenie klasyfikatorów, a optymalna kombinacja składowych algorytmów została znaleziona za pomocą algorytmu symulowanego wyżarzania.

Trzeci rozdział zawiera szczegółowe wyniki ewaluacji opracowanych metod, omawianych w rozdziale drugim. Oba zagadnienia identyfikacji podrzędników predykatów, jak i wyboru schematów walencyjnych zostały ocenione pod względem miar dokładności (precision), pełności (recall) i miary F (F-measure).

Do porównywania skuteczności algorytmu identyfikacji schematu walencyjnego z literaturą, Autor wybrał analogiczne prace badawcze nad językiem czeskim. Mimo, że proponowana przez Autora metoda nie korzysta z oznaczonych danych treningowych, osiągnięte wyniki są porównywalne z algorytmem identyfikacji schematów walencyjnych dla języka czeskiego z wykorzystaniem oznaczonego zbioru treningowego i są istotnie lepsze niż metoda bez korzystania z oznaczonego zbioru.

Na zakończeniu, Autor podsumował wyniki rozprawy i przedstawił kierunki dalszych badań. Jednym z kierunków dalszych badań dotyczy zastosowania proponowanych metod dla innych języków słowiańskich. Autor charakteryzował również warunki potrzebne do tego, aby Jego metoda mogła być zastosowana dla nowego języka.

Rozprawa posiada również cztery dodatki. Dodatek A przedstawia metody wyznaczania preferencji selekcyjnych, stosowane w algorytmie SRVoteSchema. Dodatek B opisuje metodę Word2vec, która jest częścią algorytmu VecSchema. Dodatek C przedstawia szczegółowe opisy algorytmów uczenia się, stosowane w rozprawie. Są to metody: Naive Bayes, regresja logistyczna, liniowe warunki pola losowe (linear random field) i lasy losowe (random forest). Dodatek D zawiera instrukcję użytkową oprogramowania, które zostało stworzone na potrzebę rozprawy.

3. OCENA ZAWARTOŚCI ROZPRAWY

Rozprawa doktorska mgr. Konrada Gołuchowskiego przedstawia autorską metodę identyfikacji schematów walencyjnych w tekstach polskich. Zaproponował kompletny system komputerowy, w którym Autor zaimplementował zarówno wszystkie szczegółowe techniki wstępnego przetwarzania języka naturalnego jak i techniki uczenia się do automatycznej identyfikacji schematów walencyjnych z tekstów polskich. Głównymi wynikami rozprawy są

metody automatycznej identyfikacji predykatów i ich podrzędników oraz metody automatycznej budowy zbioru przykładów użyć schematów walencyjnych jako zbiór treningowy dla algorytmu wybierania właściwego schematu walencyjnego dla danego predykatu w zdaniu. Zaletą proponowanej metody jest skalowalność systemu, ponieważ nie wymaga ona ręcznie przygotowywanego korpusu treningowego z oznaczonymi schematami.

Przedstawione rozwiązanie jest pierwszą metodą identyfikacji schematów walencyjnych bez potrzeby korzystania z ręcznie zaznaczonego zbioru treningowego. Pod tym względem rozprawa doktorska jest unikalna w skali światowej.

Pan mgr Konrad Gołuchowski wykazał się dobrą znajomość wiedzy w przetwarzaniu języka naturalnego (NLP) i uczeniu maszynowym (ML). Można stwierdzić, że opracowana metoda ma istotny wkład w rozwoju informatyki, zwłaszcza w zastosowaniu technik uczenia maszynowego w NLP.

Autor też implementował wszystkie proponowane metody i stworzył system do identyfikacji schematów walencyjnych. Ten system może być używany we wszystkich aplikacjach, gdzie zachodzi potrzeba automatycznego wykrywania zdarzeń w tekstach. Za pomocą tego systemu, autor wykonał ogromną ilość eksperymentów w celu oceniania skuteczności proponowanych metod i porównywania tych metod ze sobą. Autor umieścił w rozprawie bardzo szczegółowe wyniki testowania dla każdego klasyfikatora pod względem trzech popularnych miar: dokładności (precision), pełności (recall) i miary F (F-measure).

Tak jak można było spodziewać, zespół klasyfikatorów VoteMeta okazał się najskuteczniejszą metodą dla wyboru schematów walencyjnych w niejednoznacznych sytuacjach.

Pan mgr Konrad Gołuchowski chciałby też porównywać swoją metodę z odpowiednimi metodami dla innych języków. Zdecydował się na porównywanie z metodami opracowanymi dla języka czeskiego ze względu na podobieństwo jego podobieństwo do języka polskiego. Proponowana metoda okazała się skuteczniejsza niż odpowiednie metody, które zostały opracowane dla języka czeskiego.

Praca została napisana poprawnym językiem naukowym. Autor bardzo dobrze opanował technikę pisania prac naukowych, między innymi, podział na rozdziały, tworzenie rysunków, wykresów i tabel oraz spis treści. Bibliografia zawiera około 100 pozycji literaturowych. Są to w większości najbardziej aktualne publikacje w dziedzinie badań.

4. KRYTYCZNE UWAGI

Pan mgr Konrad Gołuchowski jest autorem trzech publikacji, w tym tylko jednej dotyczącej tematyki rozprawy. Według mnie, to jest bardzo słaby dorobek naukowy Doktoranta.

W trzecim rozdziale, gdzie są pokazane wyniki eksperymentów, czasem Autor przedstawił również przedziały ufności tych miar. Niestety nie znalazłem wyjaśnienia jak te przedziały zostały obliczone, i na jakim poziomie ufności?

Autor podał (Tab. 3.27) listę najlepszych i najgorszych predykatów pod względem skuteczności na korpusie ewaluacyjnym. Rozbieżność między nimi jest ogromna, tj. 100% precyzji dla czasowników takich jak „powiedzieć" lub „powstać", a 8,9% dla „stanowić" . Nie znalazłem jednak wniosku lub komentarza na temat przyczyny tej rozbieżności i możliwe metody poprawiania skuteczności.

W całej rozprawie nie było dyskusji na temat złożoności obliczeniowej proponowanych algorytmów i całego systemu. Brakuje również analizy czasu obliczeniowego dla proponowanych metod.

4. UWAGI KOŃCOWE

Reasumując można stwierdzić, że recenzowana praca doktorska zawiera interesujące wyniki. Rozprawa stanowi samodzielne rozwiązanie przez Doktoranta problemu naukowego, gdyż Autor wykazał się umiejętnością identyfikacji problemów badawczych, formułowania celu badań, pracy nad badaniami literaturowymi w zakresie analizowanych problemów, konstruowania i doboru metod badawczych, przeprowadzenia badań, wnioskowania i prezentacji wyników,

Po lekturze rozprawy można również stwierdzić, że Autor wykazał się bardzo dobrą ogólną wiedzę w swojej dyscyplinie. Mimo, że w recenzji zostały wskazane pewne niedociągnięcia oraz zgłoszone pewne zastrzeżenia, praca stanowi jednak ciekawy, oryginalny przykład zastosowania teorii analizy danych w praktyce oraz stanowić cenny materiał dla praktycznych zastosowań.

Uwzględniając wszelkie uwagi - zarówno aprobujące, jak i krytyczne oraz mając świadomość istnienia w przedstawionej do recenzji pracy pewnych kwestii dyskusyjnych, stwierdzam, że praca mgr. Konrada Gołuchowskiego spełnia wymogi stawiane pracom doktorskim. Wnoszę zatem o dopuszczenie przedłożonej mi do recenzji rozprawy do publicznej obrony.

Hung Son Nguyen