Warszawa, 5.09.2022

dr hab. Agnieszka Mykowiecka
Instytut Podstaw Informatyki PAN
Jana Kazimierza 5, Warszawa

Recenzja rozprawy doktorskiej mgr. Arkadiusz Janza

Metoda rozpoznawania semantycznej charakterystyki słów z wykorzystaniem lingwistycznie wzbogaconej reprezentacji tekstu

Recenzja rozprawy doktorskiej mgr. Arkadiusza Janza, zrealizowanej pod opieką dr hab. inż. Macieja Piaseckiego, wykonana została na zlecenie Rady Naukowej dyscypliny informatyka techniczna i telekomunikacja Politechniki Wrocławskiej.

Sformułowanym w pracy celem rozprawy było opracowanie metody rzutowania leksykalnych elementów tekstu na semantyczne bazy wiedzy poprzez zapewnienie reprezentacji kontekstów pozwalającej na efektywne określanie znaczenia wyrażeń niejednoznacznych w tekście. Realizacja tego celu podzielona została na następujące problemy:

- Opracowanie algorytmu eksploracji sieci leksykalno-semantycznych dostosowanego do zadania ujednoznaczniania znaczeń słów selektywnego w zakresie wykorzystania sieci i kontekstu ujednoznaczniania.
- Opracowanie rozwiązania heterogenicznego pozwalającego na elastyczne połączenie pogłębionej eksploracji sieci leksykalno-semantycznych z pretrenowanymi modelami językowymi.
- Wykazanie, że opracowana metoda uzyskuje skuteczność porównywalną lub wyższą od stosowanych w literaturze rozwiązań.
- Opracowanie zunifikowanego wzorcowego zbioru treningowo-ewaluacyjnego zintegrowanego ze współczesnym repozytorium znaczeniowym na potrzeby projektowania i ewaluacji modeli ujednoznaczniania znaczeń słów dla języka polskiego.

Rozprawa mgr. Janza składa się z 8 rozdziałów. Pierwszy – wstęp – przedstawia cel i zakres pracy oraz formalną definicję problemu. Autor wprowadza tu krótko problemy związane z niejednoznacznością słów języka naturalnego. Zadanie ujednoznaczniania przedstawione jest jako zadanie rzutowania sekwencji tekstowych na bazę wiedzy – repozytorium zawierające pojęcia jednoznaczne powiązane relacjami semantycznymi. Rozdział 2 – semantyka języka naturalnego – zawiera wprowadzenie podstawowych pojęć z zakresu semantyki leksykalnej oraz semantyki dystrybucyjnej. Wprowadzone są pojęcia semantycznych baz wiedzy i dystrybucyjnych osadzeń/wektorów słów. W rozdziale 3 przedstawiono kluczowe angielskie i polskie zasoby semantyczne (słowniki i anotowane dane) wykorzystywane w projektowaniu i ewaluacji systemów automatycznego ujednoznaczniania. Rozdział 4 – metody ujednoznaczniania – to przegląd metod rozpoznawania znaczeń z uwzględnieniem różnorodnych technik reprezentacji tekstu i informacji o znaczeniach.

Kolejne rozdziały rozprawy zawierają opis prac doktoranta. W rozdziale 5 autor omawia opracowaną przez siebie metodę eksploracji sieci leksykalno-semantycznych. Rozdział 6 zawiera opis wyników prac nad zbudowaniem jednolitych danych etykietowanych znaczeniami dla języka polskiego. Rozdział 7 zawiera opis wykorzystanych danych, zarówno angielskich, jak i polskich oraz wyniki eksperymentów porównujących opracowane przez autora metody z tymi znanymi z literatury. Rozdział 8 to podsumowanie pracy. W ramach bardzo bogatej

literatury przedmiotu doktorant cytuje sześć swoich publikacji (pięć konferencyjnych i jedna w czasopiśmie) związanych z tematem pracy.

Zawartość i ocena ogólna rozprawy

We wprowadzającym rozdziale 2, poza przedstawieniem samego zadania, autor wskazuje na praktyczne implikacje poprawnego rozwiązania problemu ujednoznaczniania słów wyliczając takie potencjalne zastosowania jak wyszukiwanie semantyczne, formułowanie rekomendacji, analiza wydźwięku emocjonalnego czy pełna analiza semantyczna tekstu. Następnie wprowadza podstawowe pojęcia charakteryzujące elementy tekstu (wyraz graficzny, wyraz tekstowy, jednostka leksykalna) i ich opis semantyczny (referencje, kompozycyjność), a także wybrane relacje semantyczne i typy zasobów semantycznych (słowniki, tezaurusy i sieci semantyczne). W kolejnym podrozdziale przedstawiona jest w skrócie metoda word2vec wyliczania statycznych wektorów słów oraz nowe, kontekstowe modele języka uzyskiwane poprzez uczenie sieci neuronowych na podstawie bardzo dużych danych tekstowych. W tym samym rozdziale opisane są też semantyczne miary istotności węzła w grafie wyliczane na podstawie struktury sieci semantycznej. Miary te wykorzystywane są przy znajdowaniu najbardziej pasującego w kontekście znaczenia słowa.

Rozdział wprowadzający ma dość trudną do rozszyfrowania strukturę. Wydaje sie,że jego postać mogłaby być istotnie poprawiona pod kątem uporządkowania i stopnia szczegółowości poszczególnych informacji. Opis modeli neuronowych przeplata sie tu z opisem cech sieci leksykalnych, niektóre pojęcia są wyjaśnione zbyt dokładnie (formy fleksyjne leksemu), inne niezbyt wydają się potrzebne (ukryty model Markowa), inne (modele kontekstowe, atencja) wydają się nie być wyjaśnione wystarczająco dokładnie. Niezależnie jednak od tych uwag, w treści pojawiają się wszystkie istotne dla konkretnego zadania elementy, co dowodzi orientacji doktoranta w szerszym kontekście rozprawy.

Kolejne dwa rozdziały rozprawy poświęcone są opisowi aktualnego stanu badań w dziedzinie będącej bezpośrednio tematem rozprawy. Autor omawia tu pokrótce wszystkie znaczące repozytoria znaczeniowe dla języków polskiego i angielskiego oraz istniejące anotowane znaczeniami korpusy tekstowe. Następnie opisane są różne podejścia do rozwiązywania omawianego problemu. Wymienione są metody słownikowe, wykorzystujące przede wszystkim tekstowe definicje znaczeń i przykłady ich użycia, metody frekwencyjne wykorzystujące informację o częstości użycia słów w poszczególnych znaczeniach, metody relacyjne opierające się na strukturze sieci semantycznych, a także metody wykorzystujące różne metody uczenia się maszyn, w których słowa reprezentowane są przez zbiory cech lub wektory rzeczywiste. W przypadku tych ostatnich metod ujednoznacznianie tekstu traktowane jest albo jako typowe zadanie klasyfikacji (przy znanym repertuarze znaczeń i zbiorze danych treningowych) lub jako zadanie wyszukiwania największego podobieństwa. Zadanie ujednoznaczniania słów jest zadaniem trudnym i ciągle nie ma zadowalającego rozwiązania. Wiele metod wykorzystujących tylko jedno źródło danych nie osiąga wyników istotnie wyższych niż 0.6 miary F1. Dlatego też dużym zainteresowaniem cieszą się metody hybrydowe, które łączą różne sposoby wykorzystania różnych źródeł informacji.

Rozdziały 3-4 dają dość dobry obraz dziedziny ujednoznaczniania słów w tekście. Przedstawione tu opisy metod nie są jednak zbyt szczegółowe.

Głównym celem doktoranta było opracowanie efektywnej metody przeszukiwania sieci semantycznych na potrzeby hybrydowej metody rozpoznawania znaczeń słów łączącej selektywne metody przeszukiwania sieci semantycznych z wyznaczaniem podobieństwa znaczeń i kontekstów przy wykorzystaniu reprezentacji wektorowej. Dwie metody, które służyły autorowi jako punkty odniesienia to propozycje z (Bevilacqua i Navilgi,2020) oraz (Wang i Wang, 2020) polegające na połączeniu reprezentacji kontekstowej ujednoznacznianego słowa

i opisu struktury sieci w otoczeniu analizowanego znaczenia. W obu, reprezentacja znaczenia odbywa się w stały, a priori ustalony sposób. Zaproponowana przez doktoranta metoda eksploracji sieci leksykalno-semantycznych opisana jest w rozdziale 5. Przedstawione rozwiązanie to połączenie metody przeszukiwania wszerz i macierzowej metody propagacji aktywacji, która polega na wielokrotnym mnożeniu macierzy incydencji reprezentującej graf semantyczny przez macierz kodującą węzły startowe. Autor zaproponował różne metody ustalania jakie informacje z sieci semantycznej będą przy tym brane pod uwagę wprowadzając dodatkowe macierze pozwalające na wygaszanie bądź wzmacnianie sygnałów i na badanie zgodności kontekstów semantycznych. Pozwala to na znacznie bardziej ogólne i trenowalne sformułowanie zależności znaczenia od innych węzłów w grafie semantycznym. W standardowych metodach ograniczano się na ogół do bezpośredniego sąsiedztwa i explicite wskazanych relacji, co uprzywilejowywało miejsca w sieci o dużej liczbie węzłów sąsiednich, a taka charakterystyka wynika czasem wyłącznie ze sposobu tworzenia sieci, a nie odzwierciedla zależności semantycznych.

Kolejny rozdział rozprawy poświęcony jest wykorzystywanym w pracy danym. Etykietowanie tekstu znaczeniami słów jest bardzo pracochłonne, zatem odpowiednich danych nie jest wiele. Dodatkowym problemem w przypadku języka polskiego jest (paradoksalnie) ciągły rozwój sieć plWordnet, w której kolejnych, coraz bogatszych wersjach, nie umożliwiono powiazania z pojęciami z wersji poprzednich. Etykietowane dane stają się zatem mało wartościowe, gdyż ich konwersja wymaga pracy ręcznej. Kolejnym problemem jest to, że praktycznie wszystkie nowe prace dotyczące zadania przypisywania znaczeń dotyczą języka angielskiego. Aby uzyskać porównanie z innymi konkurencyjnymi rozwiązaniami, doktorant przeprowadzał zatem swoje eksperymenty zarówno na tekstach polskich jak i angielskich. Aby zgromadzić odpowiednią ilość oznaczeń autor dokonał zebrania dostępnych danych. Dla danych polskich dokonana została konwersja na oznaczenia zgodne z plWornet 4.2 poprzez automatyczne wyszukanie wątpliwych oznaczeń i ich ręczną korektę. Oznaczenia polskie zostały następnie zrzutowane na oznaczenia z Princetown WordNet 3.0 i BabelNet 4.0. Rzutowanie to objęło około 70% oznaczeń.

Ta część dokonanej pracy ma niewątpliwie dużą wagę dla dalszych prac nad opracowaniem metod przypisywania znaczeń, szczególnie dla prac dotyczących języka polskiego. Odrębną wartość mają też opisane w kolejnym rozdziale eksperymenty z powtórzeniem niektórych wyników obcych na tych właśnie danych, które będą mogły stanowić punkt odniesienia.

Rozdział 7 zawiera wyniki przeprowadzonych eksperymentów. Wiele z nich miało na celu dobór najlepszych parametrów. Po ich ustaleniu dokonane zostało porównanie wyników z tymi znanymi z literatury. W tym celu doktorant zebrał wszystkie wyniki uzyskane za pomocą zbliżonych metod, a niektóre z prac powtórzył. Wyniki uzyskiwane dzięki zaproponowanej metodzie były na ogół nieznacznie lepsze od dotychczasowych. Wobec niewielkiego (w stosunku do liczby pojęć) zasobu anotowanych danych i typowej sytuacji, w której najwięcej danych jest dla języka angielskiego, doktorant wykorzystał transfer międzyjęzykowy, w którym dzięki wielojęzycznym modelom językowym można dokonywać uczenia na danych w jednym języku (lub dwóch), a wykorzystywać model dla danych w drugim. Odrębne eksperymenty dotyczyły wykrywania znaczeń spoza zbiorów treningowych i znaczeń rzadkich. Sekcja eksperymentalna jest bardzfo obszerna. Chyba aż trochę za bardzo, wobec faktu, że uzyskiwane wyniki mieszczą się w stosunkowo wąskim przedziale.

Przedstawiona do recenzji praca zawiera ciekawe rozwiązanie ważnego w dziedzinie przetwarzania języka naturalnego problemu ustalania znaczenia konkretnego użycia w tekście wyrażenia wieloznacznego. Doktorant wykazał się sporą wiedzą na ten temat oraz znajomością najnowszych metod badawczych. Sformułował własne rozwiązanie problemu, przygo-

towal odpowiednie środowisko eksperymentalne i przetestował różne warianty rozwiązania na danych w dwóch językach (angielskim i polskim) osiągając wyniki porównywalne lub nieco lepsze od najlepszych. Jakkolwiek wyniki te nie okazały się dużo lepsze od dotychczas opublikowanych, trudno było tego oczekiwać wobec stopnia skomplikowania zadania i wielu dotychczas podejmowanych prób. Wskazane przez doktoranta rozszerzenia metody wraz z rozwojem leksykalnej sieci semantycznej i zwiększeniem ilości dostępnych danych mogą potencjalnie doprowadzić do osiągnięcia lepszego rezultatu, ale przy obecnych zasobach semantycznych nawet przeszkolone osoby nie zawsze zgadzają sie co do wyboru jednego z zawartych w nich znaczeń, zatem sama ewaluacja ma ograniczoną wiarygodność. Ponadto, sieci leksykalno-semantyczne nie są kompletne, co także wpływa na to, że osiągnięcie stuprocentowo poprawnych wyników nie jest możliwe. W tej sytuacji automatyczne ustalenie choćby tylko w nieco ponad 70% przypadków właściwego znaczenia jest dużym osiągnięciem.

Merytoryczną treść rozprawy oceniam wysoko, pewne uwagi mam jednak do jej formy. Wybór tego co w pracy zostało umieszczone i na jakim poziomie szczegółowości budzi trochę wątpliwości. W szczególności metody, z którymi się autor porównuje, mogłyby zostać opisane bardziej szczegółowo. Własna metoda doktoranta mogłaby być zilustrowana jakimiś choćby wyrywkowymi sztucznymi przykładami. Prezentacja tej metody jest formalna, ale wobec niepełnego opisu wzorów zrozumienie jej wymaga nieco wysiłku. Mało tu opisu inuicji i jakichkolwiek przykładów pozwalających lepiej zrozumieć motywacje doktoranta. Rysunki i wzory umieszczone w pracy często nie są opisane, lub opisane wyrywkowo, do wzorów często nie ma żadnego odwołania w tekście, odwołanie zamiast do (6) kieruje do (identycznego) (42); (45) i (46) wyglądają identycznie. Bardziej uciążliwym brakiem jest jednak bardzo skąpy opis zamieszczonych w pracy ciekawych rysunków ilustrujących faktyczną strukturę wewnętrzną proponowanych sieci. Ewaluacja opera się wyłacznie na prezentacji wyników miary F1 dla całosci danych. Brak dokładniejszej analizy wyników, która mogłaby zawierać choćby analizę porównawczą wyników dla różnych metod na poziomie kategorii słów, czy wybranych słów o wielu znaczeniach. Mogłoby to pomóc w pszukiwaniu drogi do ulepszenia proponowanego rozwiązania.

Edytorsko praca jest poprawna, napisana jest dobrym językiem, nie ma w niej praktycznie błędów stylistycznych, a błędy pisowni są bardzo nieliczne. W ramach niedostatków wyliczyć można brak spisu treści. Ponadto, z drobnych uwag, tytuł paragrafu „Semantyczna analiza tekstu" nie wydaje się adekwatny do treści dotyczącej zasobów słownikowych i nakłada się znaczeniowo na kolejny, tym razem już adekwatny do treści tytuł „rozumienie języka naturalnego". Tekst na stronie 99 kończy się niespodziewanie w pół zdania.

Wniosek końcowy

Stwierdzam, iż przedłożona mi do recenzji rozprawa, której autorem jest mgr Arkadiusz Janz, zawierająca opis nowatorskiego rozwiązania zadania ustalania znaczenia słów (i części prostych fraz) w tekście, w mojej opinii spełnia wymagania ustawowo stawiane rozprawom doktorskim. W wnoszę o dopuszczenie magistra Arkadiusza Janza do publicznej obrony.