Warszawa, 21.2.2022

Prof. dr hab. inż. Mieczysław A. Kłopotek
Kierownik Zespołu Podstaw Sztucznej Inteligencji
Instytut Podstaw Informatyki
Polskiej Akademii Nauk, Warszawa

Recenzja rozprawy doktorskiej pana mgr. inż. Sławomira Dadasa

Niniejszą recenzję przygotowałem na zlecenie, dr. hab. inż. Jana W. Owsińskiego, Zastępcy Dyrektora ds. Naukowych Instytutu Badań Systemowych Polskiej Akademii Nauk.

Przedmiotowa rozprawa, przygotowana pod kierunkiem prof. dr. hab. inż. Witolda Pedrycza (IBS PAN, promotor) i dr. inż. Jarosława Protasiewicza (OPI PIB, promotor pomocniczy), zatytułowana „*Iteracyjne metody wykrywania hierarchicznych struktur jednostek nazewniczych*", o objętości 115 stron, składa się z 8 rozdziałów i kończy bibliografią obejmującą 150 pozycji, w tym 1 pozycję autorstwa i 1 współautorstwa doktoranta.

Rozprawa dotyczy istotnego zagadnienia z dziedziny przetwarzania języka naturalnego, mianowicie wykrywania w tekście jednostek nazewniczych, o dużym znaczeniu nie tylko dla samego przetwarzania języka naturalnego (w tym semantycznej anotacji), ale także ekstrakcji informacji, wyszukiwania informacji, uczenia maszyn, w systemach odpowiadania na zapytania, w systemach Semantic Web, Social Web itd. Zagadnienie to jest przedmiotem zainteresowań badaczy co

najmniej od lat 90tych minionego stulecia i nadal dalekie od definitywnych rozwiązań. Doktorant skupił się na jednym z trudniejszych problemów w tym obszarze, tj. identyfikacji jednostek nazewniczych hierarchicznie w sobie zagnieżdżonych.

W przedłożonej pracy własne wyniki badawcze Autora są zawarte w rozdziałach 5 (Metody iteracyjne), 6 (Dwukierunkowy algorytm iteracyjny) oraz 7 (Eksperymenty). Rozdziały 1-4 służą prezentacji materiału potrzebnego do wyjaśnienia idei zaproponowanej przez Doktoranta algorytmiki (jak np. zagadnienia modelowania sekwencji czy wektorowej reprezentacji tekstu), zaś rozdział 8 stanowi podsumowanie pracy. Za autorski wkład można uznać częściowo także rozdział 4, w którym prezentowana jest zaproponowana przez Doktoranta systematyzacja metod wykrywania hierarchicznych struktur jednostek nazewniczych.

Rozprawa ma charakter konceptualno-eksperymentalny. Wkład Autora w rozwój obszaru badawczego wykrywania hierarchicznych struktur jednostek nazewniczych to zaproponowanie iteracyjnego ujęcia tego problemu, zaproponowanie realizującego je dwukierunkowego algorytmu z funkcją selekcji i implementującej go architektury sieci neuronowej i metody uczenia tejże, a także badania eksperymentalne nad zaproponowaną metodą, algorytmem i architekturą sieci.

W szczególności rozdział 5 prezentuje propozycję rozszerzenia klasycznej metody tagowania sekwencji do procesu iteracyjnego, w którym między iteracjami są przekazywane stany wynikające z poprzedniej iteracji, które stanowią potem komponent „reguł" tagowania. Choć w literaturze można spotkać podejście do

wykrywania hierarchicznych jednostek nazewniczych polegające na wielokrotnym przetwarzaniu danej sekwencji, to oryginalny pomysł Autora polega na tym, by we wszystkich iteracjach stosować ten sam model tagowania, co upraszcza proces uczenia, uodparnia na nadmierne dopasowanie, redukuje potrzebny korpus uczący itd. Dalej Autor proponuje, by proces wykrywania iteracyjnego trwał tak długo, jak długo ulega zmianie wektor stanu, przez co nie jest narzucona jakaś określona hierarchia modeli. Doktorant proponuje ponadto ciekawy model trenowania iteracyjnej sieci neuronowej, w którym każda iteracja jest traktowana niezależnie, bez przenoszenia stanów z poprzedniego kroku iteracji. Dane wejściowe i wyjściowe (w tym stany) są przygotowywane w sposób sztuczny, a więc są takie, jakby oczekiwano tego w modelu idealnym. Takie podejście upraszcza proces uczenia i skraca go, redukując do znanych metod uczenia tagowania sekwencji. Dodatkowym zagadnieniem, które musiał Autor rozwiązać i zaproponować algorytm, to kwestia ujednoznaczniania przynależności tagów do poziomów hierarchii w szczególności w sytuacjach, gdy długość nadrzędnej jednostki nazewniczej i jednostki podrzędnej są takie same (w oparciu o statystyki częstości występowania). Autor proponuje wykorzystywanie dwóch trybów iteracyjnego tagowania – od szczegółu do ogółu oraz od ogółu do szczegółu.

Rozdział 6 doprecyzowuje algorytmiczną propozycję z rozdziału 5, m.in. eksplorując fakt, iż żaden z dwóch trybów iteracyjnego tagowania – od szczegółu do ogółu oraz od ogółu do szczegółu – nie jest dominujący we wszystkich rodzajach

hierarchicznych jednostek nazewniczych. Autor proponuje, by stworzyć algorytm wykorzystujący oba tryby tagowani i poprzez metodę selekcji poprawiać końcowy wynik tagowania. Postanowił zbadać 6 metod selekcji, dwie trywialne (kopia jednego z wyników iteracyjnego tagowania), część wspólną oraz sumę wyników iteracyjnego tagowania, metodę opartą o tzw. współczynnik prawności oraz bazującą na dodatkowo uczonym klasyfikatorze na bazie stanów sieci. Autor rozważa także propagację stanów tagowania między modelami – tylko na etapie końcowym albo tzw. podejście krzyżowe (współdzielenie stanów podczas całego procesu iteracyjnego). Rozdział kończy zestawieniem szacowanych złożoności procesów uczenia oraz predykcji dla literaturowych metod wykrywania hierarchicznych jednostek nazewniczych jak i własnych. Autorskie rozwiązania na pewno nie odbiegają w tym aspekcie od rozwiązań literaturowych.

Rozdział 7 prezentuje ewaluację zaproponowanych w pracy rozwiązań algorytmicznych. Rozważa się 4 korpusy danych w trzech językach (polskim, angielskim i niemieckim). Podział danych na uczące, alidacyjne i testujące jak i projekt procesu uczenia należy uznać za zgodny z arkanami sztuki. Dobór metryki ewaluacyjnej w postaci miary precision, recall i F1 jest również prawidłowy.

Eksperymenty pokazały, że metodyka Autora dla zbioru GENIA wybija się pod względem Recall oraz miary F1. Jednakże już na tym zbiorze widać, że projekt ewaluacji eksperymentu nie został d końca dobrze przemyślany. Podstawowym wyróżnikiem pracy Doktoranta miała być zdolność wykrywania zagnieżdżonych

(hierarchicznych) jednostek nazewniczych. Tymczasem w tabeli wynikowej 7.2 nie ma podziału na proste i zagnieżdżone jednostki nazewnicze. Biorąc pod uwagę fakt, iż mniej niż 10% jednostek nazewniczych było zagnieżdżonych, osiągnięcie recall na poziomie 80% było teoretycznie możliwe przy kompletnym nierozpoznaniu jednostek zagnieżdżonych. Osobiście nalegałbym na uzupełnienie odpowiednim zestawieniem danych dla tej jak i innych kolekcji danych, nawet jeśli dane takie nie byłoby dostępne dla metod literaturowych, a wyniki dotyczyłyby tylko metodologii Doktoranta. Byłyby te wyniki interesujące same w sobie, a także mogłyby posłużyć jako wzorcowe dla innych prac. Tym bardziej, że tabela 7.3 budzi podejrzenie, że metoda nie radzi sobie wcale z zagnieżdżonymi jednostkami nazewniczymi (por. wiersz pierwszy i czwarty).

Dla zbioru NNE metody Autora wygrywają z literaturowymi zarówno dla precision, jak i recall jak i F1 (tabela 7.4). Nadmienić należy jednakże, że to na tym zbiorze Autor dobierał hiperparametry. Wyniki w wymiarze zagnieżdżeń są tu naturalnie mnie „podejrzane", gdyż 60% jednostek nazewniczych ma zagnieżdżenia. Ale i tu przydałaby się dodatkowa statystyka rozpoznawania nazw z poszczególnych poziomów zagnieżdżeń, tym bardziej, że jest ich aż sześć.

Dla zbioru PolEval ponownie metody Autora wygrywają z literaturowymi dla F1 (tabela 7.6). Wyników dla precision, jak i recall niestety nie podano (a byłoby miejsce). 42% jednostek nazewniczych posiada zagnieżdżenia, więc jest pewność, że rozpoznawanie jednostek hierarchicznych funkcjonuje, choć dokładniejsze dane byłyby mile widziane.

W wypadku zbioru GermEval metody Autora wygrywają z literaturowymi zarówno dla precision, jak i recall jak i F1 (tabela 7.7). Niestety, podobnie jak dla GENIA, liczba jednostek zagnieżdżonych jest mniejsza niż reszta z recall (wynosi 8%).

Autor zbadał ponadto wpływ propagacji prostej versus krzyżowej na wyniki wykrywania. Nie widać jakiejś zdecydowanej przewagi żadnej z nich.

Mimo wskazanych powyżej niedociągnięć w projekcie eksperymentów należy uznać przeprowadzone badania z jednej strony za wszechstronne, z drugiej na wskazujące na szereg zalet algorytmów zaproponowanych przez Autora. Na pewno nie są one gorsze od publikowanych w literaturze, w wielu przypadkach prostsze do wytrenowania i szybsze w podejmowaniu decyzji, a zaproponowany mechanizm selekcji między wynikami predykcji od ogółu do szczegółu i od szczegółu do ogółu umożliwił uzyskanie najlepszych wyników dla wybranej metryki dla prawie każdego zbioru danych. Oznacza to zdolność dopasowania mechanizmu do będącej przedmiotem zainteresowania dziedziny zastosowania.

Wobec powyższego stwierdzam, zastrzegając konieczność uzupełnienia wyników badań o wskazane analizy zachowań specyficznie dla jednostek zagnieżdżonych, że **przedłożona praca spełnia wymogi formalne i zwyczajowe stawiane pracom doktorskim i wobec tego wnoszę o dopuszczenie Kandydata do dalszych etapów przewodu.**