Dr hab. inż. Maciej Piasecki
Politechnika Wrocławska
maciej.piasecki@pwr.edu.pl
ul. Wybrzeże Wyspiańskiego 27
50-370 Wrocław

10 maja 2022

Recenzja rozprawy doktorskiej

Sławomira Dadasa,

zatytułowanej:

Iteracyjne metody wykrywania hierarchicznych struktur jednostek nazewniczych

1. Problem badawczy i jego znaczenie

Rozprawa dotyczy zagadnienia automatycznego wykrywania i klasyfikacji wystąpień jednostek nazewniczych w tekście, które jest jednym najistotniejszych zadań w obrębie wydobywania informacji z tekstów (ang. *Information Extraction*). Jest to również zadanie bardzo istotne w kontekście całości obszaru inżynierii języka naturalnego (inaczej przetwarzania języka naturalnego, ang. Natural Language Processing). Zagadnienie będące przedmiotem rozprawy jest najczęściej określone angielskim terminem *Named Entity Recognition*, które ma rodowód inżynierski jest wewnętrznie sprzeczne, dlatego należy docenić Doktoranta, że konsekwentnie posługuje się, począwszy od tytułu, poprawnym polskim terminem jednostek nazewniczych.

Jaki jest najważniejszy problem rozważany w rozprawie?

Rozpoznawanie wystąpień jednostek nazewniczych w tekstach było przedmiotem bardzo wielu badań. W ogromnej większości przypadków przyjmuje się uproszczenie polegające na traktowaniu wystąpień jednostek nazewniczych jako ciągłych wyrażeń językowych i rozłącznych, tzn. poszczególne wystąpienia nie zachodzą na siebie w żaden sposób. W rozprawie Doktorant zmierzył się z trudniejszą wersją tego problemu i znacznie rzadziej podejmowana, w której nie zakłada się powyższego uproszczenia, to znaczny wystąpienia jednostek nazewniczych mogą zawierać wystąpienia innych jednostek i może powstawać w ten sposób hierarchiczna (inaczej zagnieżdżona struktura), np. *Uniwersytet im. Adama Mickiewicza*. Ponieważ stosunkowo nieliczne rozwiązania dla tej wersji problemu były obarczone przyjmowaniem znaczących ograniczeń (np. liczby poziomów zagłębień lub koniecznością reprezentowania w danych treningowych wszystkich kombinacji zagnieżdżeń), jako główny cel pracy przyjęto opracowanie modelu iteracyjnego wykrywania wystąpień jednostek nazewniczych w tekstach, który byłby otwarty na dowolnie głęboką strukturę zagnieżdżeń (liczbę poziomów hierarchii). Jednocześnie wykryte wystąpienia jednostek nazewniczych są klasyfikowana do przyjętego z góry zbioru kategorii semantycznych (obiektów lub pojęć reprezentowanych przez jednostki). Przyjęcie z góry zdefiniowanego zbioru kategorii nie należy postrzegać jako ograniczenia problemu lub proponowanej metody, ponieważ jest typowych zabiegiem w tej dziedzinie i nie ogranicza znacząco przyszłych zastosowań.

Czy ma on charakter naukowy?

Rozpoznawanie i klasyfikacja wystąpień hierarchicznych struktur jednostek nazewniczych jest problemem badawczym, który nie został wcześniej satysfakcjonująco rozwiązany. Opublikowano bardzo dużo prac na temat rozpoznawania i klasyfikacji wystąpień nazw własnych w tekstach, w tym

także dla języka polskiego, ale ogromna większość z nich opiera się na upraszczającym założeniu, że wystąpienie nazwy własnej nie pokrywa się (nie obejmuje) wystąpienia żadnej innej nazwy. W praktyce znacząca liczba nazw własnych zawiera jako składowe wystąpienia innych nazw własnych. Z pełnym problemem rozpoznawania i klasyfikacji struktury wystąpień nazw własnych zmierzyło się stosunkowo niewielu autorów prac naukowych w literaturze, szczególnie dla języka polskiego. Doktorant również trafnie określił szereg ograniczeń i uproszczeń w przedstawionych do tej pory podejściach do rozpoznawania wystąpień hierarchicznych jednostek nazewniczych i zaproponował własne, znacznie bardziej ogólne rozwiązanie.

Czy ma on znaczenie praktyczne?

Rozpoznawanie i klasyfikacja wystąpień jednostek nazewniczych jest jednym z podstawowych etapów wydobywania informacji z tekstu. Ma kluczowe znaczenie dla odniesienia tekstu do baz wiedzy i występuje w bardzo wielu praktycznych zastosowaniach przetwarzania języka naturalnego. Rozpoznanie składowych wystąpień hierarchicznych jednostek nazewniczych umożliwia poprawę kompletności rozpoznawania.

Problemy szczegółowe rozważane w ramach rozprawy.

Kluczowym problem szczegółowym podjętym przez Doktoranta było opracowanie mechanizmu iteracyjnego rozpoznawania wystąpień jednostek nazewniczych umożliwiającego stopniowe rozpoznanie wszystkich wystąpień nazw zagnieżdżonych w ramach wystąpienia nazwy złożonej. Spośród kilku zagadnień szczegółowych rozważanych w pracy, warto jeszcze wyróżnić problem reprezentacji stanu procesu iteracyjnego rozpoznawania w sposób umożliwiający zastosowanie nadzorowanego uczenia maszynowego. Doktorant zaproponował tu relatywnie proste, ale przynoszące pozytywne rezultaty rozwiązanie oparte na binarnych flagach.

2. Wkład autora

Kluczowym elementami rozwiązania zaproponowanego przez Doktoranta są „neuronowy model iteracyjny" rozpoznawania elementów wystąpień hierarchicznych jednostek nazewniczych oraz struktura „wektora stanu" jako reprezentacji stanu procesu rozpoznawania. Stanowią one łączne rozwiązanie, ale każde wnosi ciekawą innowację.
Model iteracyjny stanowi proste rozwiązanie problemu nieograniczonej głębokości zagnieżdżeń jednostek nazewniczych – rozpoznawanie jest kontynuowane, wstępująco lub zstępująco, tak długo, jak długo są wykrywane kolejne jednostki. Stanowi to nowum w stosunku do wcześniejszych podejść, które albo przyjmowały jakieś ograniczenia (większość), albo były bardzo kosztowne obliczeniowo. Iteracyjność oczywiście daje narzut obliczeniowy w stosunku, np., do podejścia sekwencyjnego, jednak warto podkreślić, że Doktorant przywiązuje dużo uwagi kwestii złożoności obliczeniowej, co również wyróżnia pozytywnie jego pracę na tle prac opartych na głębokim uczeniu maszynowych w podobnej klasie problemów.
Drugi z kluczowych elementów, zaproponowany model reprezentacji stanu wykrywania jednostek w tekście, pozwala na przypisanie wielu etykiet (tagów) dla tej samej sekwencji. Opiera się na stosunkowo prostej idei binarnych flag (sygnalizujących wystąpienie określonego tagu dla określonego słowa w tekście), która była wcześniej stosowana, np., w przypadku ujednoznaczniania morfosyntaktycznego (tzw. tagowania), ale tutaj jej zastosowanie jako podstawy iteracyjnego procesu rozpoznawania jest bardzo ciekawe. Sam model reprezentacji może być stosowany w ramach innych zadań przetwarzania na poziomie wyrazowym. Pewnym jego ograniczeniem jest fakt, iż zbiór wszystkich kategorii jednostek nazewniczych musi być znany z góry, jednak jest to typowa praktyka, więc nie jest to dotkliwe ograniczenie.

Doktorant zaproponował również łączne stosowanie podejść w obu kierunkach, tj. zstępującym i wstępującym. Uzyskane rezultaty są w większości przypadków lepsze niż podejścia jednokierunkowego, ale jednak zaobserwowane różnice są dość małe i wyniki badań nie są przekonywujące. Aspekt łączenia różnych podejść, rodzaj kombinacji klasyfikatorów, wydaje się być potraktowany dość technicznie, bez jego podgłębienia, np., w zakresie architektur neuronalnych łącznych klasyfikatorów, które mogłyby przynieść dalsze ciekawe spostrzeżenia.

3. Poprawność

Czy stwierdzenia zawarte w rozprawie są godne zaufania?

Praca ma bardzo dobrą strukturę. Doktorant w systematyczny i przekonywujący sposób przedstawia intuicje kryjące się za proponowanymi rozwiązaniami. Bardzo dobrze też je odnosi do zidentyfikowanych ograniczeń podejść znanych z literatury przedmiotu. Koncepcja rozwiązania jest bardzo przekonywująca.
Ponieważ proces iteracyjny przebiega etapami, konieczne jest wytrenowanie klasyfikatora dla poszczególnych kroków. Doktorant zaproponował dość ryzykowną strategię etapowego trenowania modelu: “Dlatego też w zaproponowanej w tej pracy metodzie stosujemy uproszczone podejście, którego założeniem jest traktowanie każdej iteracji modelu jako osobnej próbki uczącej.” Oznacza to, że w każdym kroku modelowi prezentowany jest stan idealny wygenerowany na podstawie próbki treningowej. Podejście takie, nazywane ostatnimy laty “teacher forcing”, było stosowane przy trenowaniu tagerów morfosyntaktycznych dla języka polskiego. W tamtych pracach było jednak porównywane ze strategią, w ramach której następny krok modelu widzi wyniki działania poprzedniego kroku, czyli uczy się w warunkach identycznych z tymi, jakie są podczas jego stosowania. Badania nad tagerami nie dały jednoznacznej odpowiedzi, która ze strategii jest lepsza.
Przyjęta w rozdziale 6.2 miara pewności predykcji dla określonego kierunku predykcji jest tylko jednym możliwych sposobów jej zdefiniowania. Z treści pracy nie jest jasne czy były analizowane inne sformułowania takiej miary. Podobnie, dalej w tym samym rozdziale 6.2, łączenie obu kierunków predykcji przy pomocy wytrenowanego klasyfikatora sprowadza się do dość prostego modelu opartego na regresji logistycznej. Prostota modelu może być jego wielką zaletą, tylko brakuje porównania z innymi możliwymi podejściami.
Jak już było to wspomniane należy bardzo docenić zwracanie przez Doktoranta uwagi na kwestie złożoności obliczeniowej rozwiązań. Na str. 68, słusznie jest podniesiona kwestia złożoności generowania reprezentacji jako części złożoności całego podejścia. Jednak w porównaniu złożoności na str. 69, nie bierze się jednak pod uwagę złożoności wykonywania poszczególnych kroków obliczeń, który w przypadku sieci głębokiej jest bardzo wysoki.
Odnośnie wykorzystanych neuronalnych modeli językowych i ich wpływu na zaproponowane rozwiązanie, brakuje dyskusji oraz badań dotyczących potencjalnej pamięci leksykalnej modeli opartych na głębokich sieciach neuronowych. Modele ELMo i Flair są modelami znakowymi, istnieje więc niebezpieczeństwo, że model wytrenowany na takiej reprezentacji tekstu `przechowa’ w swojej strukturze informację o kształcie samych rozpoznawanych nazw własnych. Jeżeli dane testowe zawierają podobne nazwy własne co treningowe, to pomimo oczywistej różnicy pomiędzy danymi treningowymi i testowymi, model rozpoznawania może kierować się samym kształtem nazwy, co obciąża pozytywnie wynik. Problemowi temu poświęca się ostatnio co raz więcej uwagi w literaturze przedmiotu.

Czy uzasadnienia są poprawne?

Kluczową weryfikacją dla zaproponowanego podejścia jest ewaluacja przeprowadzona na zbiorach wzorcowych. W swojej ogólnej strukturze, a także sposobie realizacji proces ewaluacji został

zaprojektowany i przeprowadzony prawidłowo. Uzyskane wyniki pokazują wyraźną przewagę zaproponowanego podejścia iteracyjnego. Aczkolwiek zaobserwowane różnice pomiędzy poszczególnymi sposobami jego realizacji nie są już tak wyraźne, jak to sugeruje Doktorant we wnioskach.

Przy porównaniu wyników z literaturą warto byłoby zaznaczyć, które rezultaty są cytowane z prac, a które są wynikiem własnych eksperymentów autora, np. uruchomienia kodów lub ich re-implementacji. Zwykle jednak wyniki w publikacjach okazują się wyższe, niż te, które uzyskuje się przy skrupulatnej re-implementacji podejść, więc przewaga własnych rozwiązań Doktoranta nie budzi wątpliwości.

Do procesu ewaluacji i dyskusji uzyskanych wyników można mieć kilka poniższych uwag szczegółowych.

- W większości eksperymentów nie zbadano, czy zmiany obserwowane w ramach studiów ablacyjnych dotyczyły jednostek zagnieżdżonych, czyli głównego celu – jest to bardzo mało prawdopodobne, ale mogłoby się okazać, że zaproponowane podejście przynosi ogólnie poprawę, ale niekoniecznie w ramach podzbioru jednostek zagnieżdżonych.
- W ramach testów na zbiorze GENIA:
    - w Tab. 7.3 - wynik samego wariantu "Tylko outside-in" tak niewiele odbiega od wyniku całego modelu, że powstaje pytanie, czy ta różnica jest istotna statystycznie?
    - na str. 82 "natomiast w przypadku zbioru GENIA nie jest on znaczny." - wyniki nie zostały odniesione do potencjalnych możliwości poprawy, czyli podzbioru jednostek zagnieżdżonych, który jest niewielki w tym zbiorze;
    - na str. 82: "Biomedyczny Word2Vec został natomiast zastąpiony 300 wymiarowym modelem GloVe" - to są różne modele (sic!), co oznacza wprowadzanie dwóch zmian jednocześnie.
- Zbiór NNE:
    - "Wynik ten nie jest zaskakujący, bowiem struktury jednostek nazewniczych w tym zbiorze są głębsze i bardziej złożone, a wiele jednostek nazewniczych znajduje się w wewnętrznych warstwach, które w wersji bez iteracji są pomijane." - można to było dobrze odnieść do danych i sprawdzić zakres wpływu.
- Zbiór PolEval:
    - "100 wymiarowego modelu Word2Vec" – nie jest jasne dlaczego zastosowano tak mały rozmiar wektora (i tylko taki), skoro dla innych zbiorów wybierano dłuższe?
- Zbiór GermEval
    - różnice pomiędzy pojedynczymi modelami i podejściami łączonymi są bardzo niewielkie na całości zbioru;
    - "Dużą przewagę modeli iteracyjnych nad pozostałymi metodami w tym przypadku można tłumaczyć wykorzystaniem bardziej efektywnej reprezentacji tekstu," – słuszne spostrzeżenie, warto byłoby poddać je głębszej analizie.
- Propagacja krzyżowa
    - ponownie różnice pomiędzy jednokierunkowymi modelami, a dwukierunkowymi nie są za duże.
- Na str. 90 pojawiają się w końcu wyniki oceny statystycznej istotności różnic. Szkoda, że nie zostały przedstawione od razu w poszczególnych w tabelach z porównaniami. Analiza ta pojawia się późno w pracy, nie jest zapowiadana wcześniej i jest ograniczona jedynie do rozwiązań własnych autora.

4. Wiedza kandydata

Doktorant wykazał się bogatą wiedzą na temat stanu badań i metod stosowanych w inżynierii języka naturalnego, jak w również w zakresie sztucznej inteligencji (np. algorytmy maszynowego uczenia), czy szerzej informatyki (np. zagadnienia złożoności obliczeniowej i optymalizacji).
Doktorant przedstawił zwięzłe, ale bardzo przejrzyste, wręcz wzorowe, wprowadzenie do zastosowania sieci do modelowania języka. Doktorant wiele razy wykazał się bardzo cenną umiejętnością zwięzłego, ale bardzo celnego wytłumaczenia pojęć, algorytmów i metod

Rozprawa zawiera dobry przegląd prac z literatury. Bibliografia jest obszerna i dobrze pokazuje stan badań w odniesieniu do problemu. Jednak niektóre prace z literatury zostały pominięte w przeglądzie (a wszystkie są łatwo dostępne w ważnej bazie ACL Anthology), np.

Xinwei Long, Shuzi Niu, and Yucheng Li. 2020. Hierarchical Region Learning for Nested Named Entity Recognition. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4788–4793, Online. Association for Computational Linguistics.

Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural Architectures for Nested NER through Linearization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5326–5331, Florence, Italy. Association for Computational Linguistics.

Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. Locate and Label: A Two-stage Identifier for Nested Named Entity Recognition. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2782–2794, Online. Association for Computational Linguistics

Yu Wang, Yun Li, Hanghang Tong, and Ziye Zhu. 2020. HIT: Nested Named Entity Recognition via Head-Tail Pair and Token Interaction. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6027–6036, Online. Association for Computational Linguistics.

Ying Luo and Hai Zhao. 2020. Bipartite Flat-Graph Network for Nested Named Entity Recognition. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6408–6418, Online. Association for Computational Linguistics.

Xinwei Long, Shuzi Niu, and Yucheng Li. 2020. Hierarchical Region Learning for Nested Named Entity Recognition. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4788–4793, Online. Association for Computational Linguistics.

5. Inne uwagi[1]

Rozprawa jest bardzo dobrze dopracowana od strony edycyjnej. Można było dostrzec jedynie nieliczne błędy, np.

[1] Opcjonalnie

s 11: “Analiza sentymentu” - chociaż zaraz obok “a badanie nacechowania emocjonalnego użytkowników”
s. 12: “opracowania dedykowanych rozwiązań. ” — anglicyzm
s. 22: ”zapomniana” — błędne cudzysłowy - notacja ang., a pierwszy w złą stronę
s. 23: “Sieci Transformer” — nieuzasadnione użycie wielkiej litery
s. 24: “sumaryzacji” — streszczaniu
s. 35: kropka jako przecinek dziesiętny, powtarzający się błąd składu
s. 37: “Jeżeli dane słowo nie występuje ze słownika V , jego wektor jest wyliczany jako suma samych wektorów n-gramowych.” — niestety w implementacji wektor zawsze jest wyliczany jako średnia, nawet jeżeli całe słowo jest w słowniku
s. 38: ”jest mapowany na dokładnie” — anglicyzm
s 47: “nie w oparciu o długość jednostki ale w oparciu o jej ” — przecinek
s 48: “różne typu wierzchołków” — typy
s 50: “typy wierzchołków ale też “ — przecinek
s. 61: “nie wykryciem” —> niewykryciem
s. 70: “korzysta w warstwy predykcyjnej” —> z
s. 74: “tzw.” - za duży odstęp po kropce, brak komendy .\
s. 80: “liczone według kolejności ich występowania w oryginalnym zbiorze” - to dość akcydentalne kryterium, które grozi obciążeniem doboru próbki
s. 90: “uzyskazno” — !

6. Podsumowanie

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez art. 187 Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (z późniejszymi zmianami)[2] moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

- rozprawa zawiera oryginalne rozwiązanie problemu naukowego,
- kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka techniczna i telekomunikacja,
- oraz kandydat posiada umiejętność samodzielnego prowadzenia pracy naukowej.

Podsumowując praca dobrze spełnia formalne i zwyczajowe wymagania stawiane rozprawom doktorskim.
Wnoszę o dopuszczenie Doktoranta do dalszych faz przewodu doktorskiego.

Podpis

[2] http://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20190000276