Warszawa, 9.12.2025

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

Recenzja rozprawy doktorskiej mgr. Mateusza Klimaszewskiego

Multi-objective modularity in Natural Language Processing

Recenzja rozprawy doktorskiej mgr. Mateusza Klimaszewskiego, zrealizowanej pod opieką prof. Tomasza Gambina i promotora pomocniczego dr. Piotra Andruszkiewicza, wykonana została na zlecenie Rady Naukowej Politechniki Warszawskiej dyscypliny informatyka techniczna i telekomunikacja.

Na zasadniczą treść pracy doktorskiej składają się 4 artykuły opublikowane na bardzo dobrych międzynarodowych konferencjach. Praca ma postać spójnego tekstu o ujednolicone j formie edycyjnej, w którym artykuły poprzedzone zostały wprowadzeniem zwierającym przedstawienie ogólnego zamysłu pracy i jej kontekst. Pracę zamyka uwspólniona bibliografia. Ze względu na fakt, że zasadniczą treścią pracy są artykuły opublikowane na znaczących konferencjach, co oznacza, że przeszły one przez proces recenzyjny, w swojej recenzji skupię się na ogólnej charakterystyce rozprawy, nie wchodząc w analizę każdego detalu poszczególnych artykułów.

Zawartość pracy

Przedstawionym we wstępie tematem spajającym składające się na rozprawę artykuły jest sprawdzenie efektywności modeli językowych w konfiguracjach innych niż realizacja jednego zadnia. Testowane warianty uogólnień to wiele-zadań, wiele-dziedzin, wiele-języków i wiele modeli. Tak postawiony problem jest bardzo aktualny gdyż wielość możliwych zastosowań sprawia, że z jednej strony naturalnym zdaje się budowanie konfiguracji zdolnych rozwiązywać więcej niż jeden wariant postawionego zadania, a z drugiej strony wykorzystanie danych przeznaczonych dla innych zadań może pomóc rozwiązać takie, dla którego mamy mało danych treningowych i/lub testowych. W ostatnich latach zyskał on odrębną nazwę – modularne uczenie głębokie (MDL) – i spore zainteresowanie.

Pytania sformułowane przez doktoranta to:

1. Jak zbudować wielozadaniowy system NLP, w którym użytkownik może wybrać odpowiednie zadania końcowe?
2. Czy można wykorzystać zewnętrzny wielodomenowy model nauczyciela ?
3. Czy możemy wykorzystać wiedzę z wielu modułów specyficznych dla danego języka bez zwiększania kosztów wnioskowania?
4. Czy możemy ponownie wykorzystać wstępnie wytrenowane moduły z różnych modeli podstawowych (*Foundation Models*)? Jakie są ograniczenia obecnych metod modułowych podczas przenoszenia między modelami?

Pierwszy artykuł, z roku 2021, gdy nie mówiono jeszcze o pojęciu uczenia modularnego, opisuje opracowany pod kierunkiem dr Aliny Wróblewskiej, ale w dużej mierze samodzielnie, system COMBO, implementujący podstawowe narzędzia pozwalające na wstępna analizę zdań w języku polskim. Jest on tagerem, analizatorem morfologicznym, lematyzatorem i

parserem, budującym reprezentujące strukturę syntaktyczną zdania drzewa zależnościowe. Można przy tym wybrać, które z wymienionych zadań mają być realizowane.

Zaproponowana architektura wykorzystuje szereg cech słów, takich jak embedding słowa, część mowy oraz wartości cech morfologicznych. Cechy te są uwzględniane przez model wtedy, gdy wybrane zadanie wymaga ich do trenowania. Autor zaproponował metodę kodowania wartości cech w sposób kontekstowy poprzez przepuszczenie wartości dla poszczególnych słów przez sieć BILSTM. Dla rozwiązywania poszczególnych zadań zdefiniowane są odrębne architektury korzystające z kontekstowej reprezentacji cech i zewnętrznych wektorów reprezentacji słów. Przykładowo, predykcja łuków zależnościowych odbywa się w sposób następujący: Dwie pojedyncze warstwy FC przekształcają globalne wektory cech dla elementów głównych i zależnych. Reprezentacją zależności jest macierz sąsiedztwa, której elementy są iloczynami skalarnymi wszystkich par osadzeń głównych i zależnych (iloczyn skalarny określa pewność krawędzi między dwoma słowami). Funkcja softmax zastosowana do każdego wiersza macierzy przewiduje sąsiednie pary głównych i zależnych. Takie podejście nie gwarantuje jednak, że wynikowa macierz sąsiedztwa jest prawidłowo zbudowanym drzewem zależności. W ostatnim etapie przewidywania stosuje się zatem algorytm Chu-Liu-Edmonds [39, 56].

Wartością zaproponowanego rozwiązania jest dostarczenie narzędzi do trenowania własnych modeli, co jest możliwe dzięki przyjęciu standardowego modelu wejścia i formalizmu Universal Dependencies jako formy reprezentowania składni języka. Modele wytrenowane przez autora wykazały się bardzo dobrą jakością w porównaniu do najlepszych w danym momencie rozwiązań (Stanza i spaCy) i bardzo dobrą efektywnością na etapie trenowania, przy wolniejszym przebiegu predykcji.

Druga praca – Gated Adapters for Multi-Domain Neural Machine Translation – dotyczy poprawy wyników tłumaczenia maszynowego poprzez dodanie informacji o dziedzinie, z której pochodzą teksty. Zaproponowana architektura pozwala na pozostawienie bez zmian początkowego modelu. Trenowane są tylko mniejsze moduły dodatkowe (Adapters). W standardowym ustawieniu neuronowego tłumaczenia maszynowego adapter (AD) przetwarza ukryty stan transformera i składa się warstwy normalizacji oraz dwóch warstw liniowych: warstwy zwężającej i warstwy rozszerzającej, z funkcją aktywacji RelU. Zaproponowane rozwiązanie, nazwane Gated Adapters, polega na możliwości łączenia informacji z wielu adapterów w miejsce wyboru jednego. Mechanizm ten jest dołączony na każdym z poziomów transformera i daje w wyniku ważoną średnią wyniku wszystkich adapterów z tego poziomu. Wynik działania każdego adaptera jest mnożony przez wartość prawdopodobieństwa dostarczoną przez zewnętrzny moduł podający stopień, w jakim zdanie należy do domeny reprezentowanej przez adapter.

Możliwe jest wykorzystanie zaproponowanej architektury bez znajomości a priori, który z adapterów powinien zostać wykorzystany, gdyż klasyfikacja odbywa się on-line z przetwarzaniem tekstu. W rozpatrywanym zadaniu uczenia maszynowego adaptery odpowiadały za dostosowanie się systemu do tematyki tekstu. Łączenie adapterów pozwala na lepsze dostosowane się do nieznanej tematyki, zbliżonej do więcej niż jednej z tych ustalonych a priori. Ewaluacji dokonano na tłumaczeniu z angielskiego na polski i z angielskiego na grecki, przy 6 zakresach tematycznych dla obu par. Ewaluacja wyników maszynowego tłumaczenia zawsze nastręcza trudności, gdyż tekst może być przetłumaczony na różne sposoby, a ocena jakości tłumaczeń jest subiektywna. Trudno zatem polegać do końca na metodach automatycznych, ale w pracy zastosowano najbardziej zaawansowane z nich. W cytowanych wynikach widać poprawę wyników, choć nie jest to poprawa bardzo widoczna. Zacytowany przykład, zapewne staranie dobrany, ale jednak rzeczywisty, pokazuje jak istotne jest w niektórych przypadkach prawidłowe rozpoznanie kontekstu tłumaczenia.

Trzecia praca – No Train but Gain: Language Arithmetic for training-free Language Ad-

apters enhancement – dotyczy ulepszenia architektury wykorzystującej adaptery językowe. Autor proponuje zamiast binarnej selekcji języka składanie wektorów reprezentujących różne języki. W środowisku wielojęzycznym pozwala to na dostosowanie modelu do jednego z wielu języków oraz uwzględnienie informacji o językach pokrewnych. Wykorzystując adaptery językowe i zadaniowe, arytmetyka językowa umożliwia przetwarzanie bez konieczności dodatkowego trenowania w dwóch przypadkach użycia: (i) zero-shot, gdzie adapter językowy dla języka docelowego nie został wytrenowany lub (ii) w celu ulepszenia istniejących adapterów językowych poprzez arytmetykę z językiem pokrewnym lub językiem, na którym został wytrenowany adapter zadania.Adaptery językowe są tu rozłączne od adapterów związanych z konkretnym zadaniem. Autorzy przeprowadzili eksperymenty z danymi w 13 językach sumarycznie w kontekście trzech zadań: rozpoznawania nazw własnych (Named Entity Recognition, NER), wnioskowania (Natural Language Inference, NLI) i odpowiadnaia na pytaia (Question Answering, QA). Wykorzystano przy tym ogólnodostępne zasoby danych. Eksperymenty przeprowadzono z użyciem dwóch wielojęzycznych modeli: XLM-R i mBERT. Uzyskana poprawa wyników to do 3 punktów F1 bez konieczności wykonywania dodatkowego trenowania. Nieco zaskakujące są generalnie niższe wyniki dla zadania NER niż NLI.

Ostatnia część rozprawy – Is Modularity Transferable? A Case Study through the Lens of Knowledge Distillation – poświęcona jest zbadaniu możliwości transferu wiedzy (rozumianego jako transfer parametrów) z jednego modelu do drugiego. Rozpatrywane są dwa schematy. W pierwszym oba modele różnią się tylko głębokością sieci, w drugim cała struktura, w tym wielkość reprezentacji, może być różna. Główną inspiracją dla tego pomysłu było umożliwienie trenowania mniejszych modeli w wykorzystaniem już dokonanego treningu modeli większych. Przeprowadzono eksperymenty z modelami mBert, DistilBert, XLM-Base i XLM-Large w kontekście trzech zadań: rozpoznawania nazw własnych, tworzenia parafraz oraz wnioskowania o zależnościach między zdaniami. Potwierdziły one wstępne założenia o możliwości transferu wiedzy w przypadku modeli o tym samym wymiarze reprezentacji. W przypadku przenoszenia informacji z modelu głębszego do posiadającego mniejszą liczbę warstw metodą skuteczniejszą od uśredniania wag okazało się kopiowanie wag co którąś warstwę. Niestety przenoszenie informacji między modelami różniącymi się wielkością reprezentacji w niektórych wariantach nie przyniosło oczekiwanych rezultatów. Ten problem wymaga dalszych badań.

Ocena

Przedstawiona mi do oceny rozprawa zawiera cztery artykuły, których wspólnym tematem jest uelastycznienie lub ulepszanie rozwiązań pozwalające na zmniejszenie kosztów trenowania lub/i poprawienie wyników w konfiguracji multi-językowej lub/i multi-zadaniowej. Podjęty temat różnie rozumianej modularności rozwiązań jest bardzo aktualny wobec dużych kosztów trenowania modeli i bardzo różnorodnej gamy możliwych zastosowań oznaczającej konieczność wytrenowania bardzo dużej ich liczby. Obecnie powszechne wykorzystywanie LLM nieco zmieniło punkt ciężkości w tym obszarze, co autor zauważa w kończącym rozprawę podsumowaniu, jednak wiele problemów wciąż nie ma zadowalającego rozwiązania, a zatem i tej eksplorowanej w pracy ścieżki nie można uznać za zamkniętą.

Zaproponowane w pracy modyfikacje istniejących rozwiązań to typowa droga postępu w nauce, gdzie tylko od czasu do czasu mamy okazję obserwować znaczący przełom. Publikowanie wyników prac na najważniejszych konferencjach dowodzi, że podejmowane tematy były uznane za aktualne i ważne przez społeczność międzynarodową.

Oczywiście rozprawa będąca zbiorem artykułów ma standardowe niedostatki związane z tą koncepcją. Zaprezentowane rozwiązania budowane na osiągnięciach innych nie są bardzo szczegółowo opisane. Struktura sieci przedstawiona jako schematy blokowe, skądinąd dobrze skonstruowane i cenne, to trochę mało dokładny poziom opisu. W szczególności dla mnie nie jest dostatecznie dobrze opisane jak tworzone są adaptery językowe. Oczywiście można zajrzeć do repozytoriów i tam prześledzić rozwiązanie, jest to jednak już poza samą rozprawą. Podobnie praca nie zawiera opisów wykorzystywanych zbiorów danych. Jest to zrozumiałe w artykułach, ale stanowi pewien brak rozprawy.

Uzyskiwane dzięki zaproponowanym modyfikacjom rezultaty nie były na ogół znacząco różne od tych uzyskiwanych innymi metodami. Niewielka poprawa w przypadkach gdy wyniki te są dość niskie, nie wydaje się mieć dużego praktycznego znaczenia. Jednak wobec ograniczeń kosztu rozwiązania nawet niewielkie pogorszenie wyników może okazać się opłacalne. Jako osobie zajmującej się przetwarzaniem danych tekstowych brakuje mi w pracy przykładów lingwistycznych, choć rozumiem że przy skali i liczbie eksperymentów trudno w artykule umieścić sensowną ilustrację przykładami tego, co zostało poprawione (tu jest jeden przykład z tłumaczeniem) lub uległo zmianie na gorsze po wprowadzeniu zmian. W rozszerzonej wersji pracy głębsza analiza jakościowa wyników mogłaby jednak znaleźć miejsce.

Bardzo pozytywnie oceniam stronę edycyjną rozprawy. Zebranie prac w spójny tekst rozszerzony o dosyć obszerny wstęp jest bardzo dobrym pomysłem na zaprezentowanie rozprawy niemającej postaci spójnej monografii. Dobrym ruchem było też uwspólnienie bibliografii, choć jej długość (204 pozycje) budzi pewne wątpliwości, czy naprawdę została przeczytana. Język pracy jest dobry, praktycznie nie zauważa się żadnych błędów językowych czy pomyłek literowych. Jednym zauważonym przeze mnie miejscem, w którym błąd wpływa na znaczenie am jest użycie 'prove' w zdaniu (str 70): "In the following Sections 5.3.2-5.3.2, we present the framework ... and (ii) an enhancement case, where we prove existing language adapters ..." Czy to miało być 'improve'?

Wniosek końcowy

Stwierdzam, że przedłożona mi do recenzji rozprawa mgr. Mateusza Klimaszewskiego zawiera opis istotnych osiągnięć w dziedzinie badania architektur sieci rozwiązujących problemy NLP ze szczególnym uwzględnieniem modularności rozwiązań. Doktorant wykazał się dużą wiedzą w tematyce związanej z rozprawą i zaproponował oraz przetestował na ogólnodostępnych danych własne rozszerzenia istniejących metod. Propozycje te zyskały uznanie społeczności międzynarodowej skupionej wokół najważniejszych dziedzinowych konferencji. Moim zdaniem rozprawa spełnia wymagania stawiane rozprawom doktorskim i wnoszę o dopuszczenie magistra Mateusza Klimaszewskiego do publicznej obrony.

A. Tykocinski