infodr hab. Maciej Ogrodniczuk, prof. IPI PAN

Instytut Podstaw Informatyki
Polskiej Akademii Nauk

Jana Kazimierza 5
01-248 Warszawa

e-mail: maciej.ogrodniczuk@ipipan.waw.pl
tel. 533 675 675

26 maja 2023 r.

Recenzja rozprawy doktorskiej

Artura Nowakowskiego

zatytułowanej:

Quality Optimization Methods in Neural Machine Translation Systems
(Metody optymalizacji jakości w neuronowych systemach tłumaczenia maszynowego)

1. Problem badawczy i jego znaczenie

Celem rozprawy jest zaproponowanie nowych metod optymalizacji jakości w systemach tłumaczenia maszynowego oraz zademonstrowanie ich zastosowania w praktyce. Cel ten uznaję za jasno i szczegółowo sprecyzowany, a zadanie przedstawione przez Doktorantem odpowiednie dla doktoratu wdrożeniowego. Postawiony problem badawczy ma zasadnicze znaczenie praktyczne dla rozwoju systemów tłumaczenia maszynowego, która to dziedzina przeżywa w obecnym czasie ogromny rozkwit za przyczyną metod neuronowych.

2. Zawartość pracy

Praca przedłożona do oceny ma postać zbioru siedmiu jednotematycznych artykułów w języku angielskim opublikowanych w recenzowanych materiałach z konferencji międzynarodowych. Doktorant – Autor rozprawy – jest pierwszym autorem wszystkich artykułów stanowiących jej zasadniczą treść.

Rozdział 1 stanowi podsumowanie celów rozprawy i powiązanie jej elementów, co jest warunkiem koniecznym dla ukazania przewodniego motywu pracy złożonej z kilku artykułów mających stanowić pewną zamkniętą całość. Rozdziały 2–5 zawierają opis prac badawczych nad nowymi metodami optymalizacji jakości w neuronowych systemach tłumaczenia maszynowego, zaś rozdziały 6–8 – opis prac rozwojowych opracowanych w ramach wdrożeniowej części doktoratu, czyli zastosowanie opracowanych przez Doktoranta metod w rzeczywistych systemach tłumaczenia maszynowego, wdrożonych w rozwiązaniach komercyjnych w ramach programu doktoratu wdrożeniowego realizowanego pomiędzy Uniwersytetem Adama Mickiewicza a firmą Poleng.

3. Zasadnicza treść pracy

Rozdział 2 opisuje eksperymenty, których celem było znalezienie efektywnego rozwiązania umożliwiającego zastosowanie ograniczeń leksykalnych w tłumaczeniu maszynowym dla języków

fleksyjnych takich jak polski. Zastosowano algorytm dekodowania z ograniczeniami leksykalnymi, który nie wymaga modyfikacji danych treningowych. Ocena jakości tłumaczenia opierała się na metryce BLEU oraz wskaźnikach dotyczących obecności, umiejscowienia, duplikacji i poprawności odmiany. Wykorzystano leksykon zawierający odmienione formy terminów wielowyrazowych oraz narzędzia do wyszukiwania i filtrowania. Przeprowadzono eksperymenty dla języka ogólnego i dziedzinowego, porównując rozwiązanie z tłumaczeniem bazowym pod kątem jakości i efektywności. Wyniki wykazały poprawę jakości tłumaczenia dla języka dziedzinowego kosztem czasu tłumaczenia.

Niezwykle krótki rozdział 3 nie zawiera propozycji nowej metody optymalizacji jakości tłumaczenia maszynowego, ale stanowi formę refleksji nt. istniejących metod oceny jakości w kontekście jednego z zadań konkursu PolEval 2021. Zaproponowane rozwiązania opierały się na modelu Comet.

Rozdział 4 opisuje rozwiązanie tłumaczenia maszynowego między odległą parą języków – hausa i angielskim, zgłoszone do zadania ewaluacyjnego na konferencji WMT 2021. Zaprezentowane metody łączą czyszczenie danych, transfer learning, trening iteracyjny i tłumaczenie odwrotne. Mimo że zgłoszone rozwiązania uplasowały się w połowie stawki systemów biorących udział w konkursie, potwierdziły biegłość autorów, w szczególności Doktoranta, w posługiwaniu się nowoczesnymi narzędziami badawczymi w ich dziedzinie oraz wielojęzycznymi zasobami tłumaczeniowymi, które mogły przydać się w zadaniu (takimi jak model niemiecko-angielski).

Rozdział 5 opisuje najważniejsze (w mojej opinii) osiągnięcie Doktoranta – implementację systemu tłumaczenia maszynowego z ukraińskiego na czeski i z czeskiego na ukraiński, które to rozwiązanie zajęło pierwsze miejsce w jednym z zadań ewaluacyjnych na konferencji WMT 2022. Zwycięski system powstały na bazie pakietu Marian uwzględniał czyszczenie i filtrowanie danych, a następnie połączenie różnych metod podnoszenia jakości tłumaczenia maszynowego, takich jak transfer learning z pary języków o dużej liczbie zasobów, tłumaczenie odwrotne z szumem, tłumaczenie wspomagane rozpoznawaniem nazw własnych czy łączenie modeli. Niezależnie od samego sukcesu rozwiązania w zadaniu ewaluacyjnym na uwagę zasługuje efektywność połączenia wielu komponentów mających wpływ na końcowy wynik tłumaczenia.

Rozdział 6 dotyczy implementacji systemu tłumaczenia maszynowego na potrzeby realizowanego przez Polską Straż Graniczną projektu AI Search („Zaawansowana analiza Internetu wspomagająca wykrywanie grup przestępczych") dot. przeszukiwania stron internetowych w języku polskim, rosyjskim, ukraińskim i białoruskim w celu odnalezienia treści o charakterze kryminalnym. System wytrenowany na danych ogólnych zostań następnie dostosowany z wykorzystaniem słowników terminologii „kryminalnej" oraz nazw własnych (osób, lokalizacji geograficznych, marek przemycanych towarów itp.). Osiągnięte wyniki dla wielu par języków przewyższyły ówczesną jakość systemu Google Translate.

W rozdziale 7 omówiono proces wdrożenia systemu tłumaczenia maszynowego dla korporacji EY oraz jego elementy. Rozdział 8 skrótowo opisuje adaptacyjną platformę tłumaczenia maszynowego POLENG MT.

4. Poprawność rozwiązania

Zaproponowane rozwiązania i prezentacja wyników są zgodne z regułami sztuki i potwierdzają biegłość Doktoranta w najnowszych technikach przetwarzania języka naturalnego i przede wszystkim tłumaczenia maszynowego. Docenić należy tak dobór tematyki badawczej, zgodny z profilem badań prowadzonych na Wydziale Doktoranta, jak i znajomość najnowszej literatury i rozwiązań technicznych. Artykuły prezentujące wyniki badań zawierają oryginalne pomysły na optymalizację

jakości tłumaczenia maszynowego w różnych elementach tego procesu – tłumaczenia dla języków o mniejszej dostępności zasobów cyfrowych, wykorzystania modeli szacowania jakości tłumaczenia, tłumaczenia całych dokumentów oraz tłumaczenia jednostek nazewniczych. Artykuły prezentujące wyniki wdrożeń opisują praktyczne zastosowanie wyników Doktoranta w systemach opracowywanych przez firmę Poleng oraz w projekcie badawczo-rozwojowym. W kontekście doktoratu wdrożeniowego takie rozwiązanie należy uznać za optymalne.

Niezależnie od wartości badawczej i praktycznej uzyskanych wyników warto też zwrócić uwagę na aktualność poruszanych problemów, zarówno w zgłoszeniach do zadań ewaluacyjnych (tłumacz na i z języka ukraińskiego już po rozpoczęciu rosyjskiej agresji), jak i wdrożeniach (języki wschodnich sąsiadów Polski w ówczesnej sytuacji migracyjnej).

5. Pytania i uwagi

W przypadku artykułu z rozdziału 2 moje pytania budzi jedynie etap konstrukcji leksykonu polegający na pobieraniu odmienionych form wielowyrazowych przy pomocy wyszukiwarki Google. Zapytania do wyszukiwarki były tworzone na podstawie form bazowych terminów, a następnie pobierano odmienione wyrażenia z pierwszych 20 stron wyników wyszukiwania. Doktorant pisze dalej: „We then limited the number of inflected variants to those that covered 95% of cases" – i tu pierwsze pytanie: czy to znaczy, że do dalszych analiz wybierano jedynie przypadki pokrywające cały wzorzec odmiany w 95%, czy może coś innego? A jeśli tak, to jak obliczano ten procent? Tak czy inaczej, bardziej przemawiałoby do mnie skorzystanie z obszernych korpusów językowych, których mamy pod dostatkiem (co podobnie zapewniłoby niezależność wyników od języka) niż opieranie się na wyszukiwarce, bo przecież pamiętamy, że „Googleology is Bad Science".

Z metodami opisanymi w rozdziale 3 trudno dyskutować, jako że zostały przedstawione w bardzo konkretnym kontekście zadania ewaluacyjnego na konkursie PolEval 2021 (dana para języków, skala ewaluacji, konkretna metryka w danych wejściowych dla wersji *non-blind*). Użycie modelu HerBERT, specyficznego dla języka polskiego, z jednej strony ogranicza stosowalność tej metody, ale z drugiej jego porównanie z modelem XLM-RoBERTa może stanowić potwierdzenie potrzeby rozwoju metod NLP dla konkretnych języków w porównaniu z metodami wielojęzycznymi.

Rozwiązanie opisane w rozdziale 5 cechuje znaczny poziom złożoności oraz duże wymagania obliczeniowe. Czy w obecnej sytuacji istnieje możliwość zmniejszenia tych wymagań? Dość dobrze widoczny jest kontrast w tym zakresie między rozwiązaniem zgłoszonym na WMT21 vs. WMT22. W tym pierwszym przypadku sami autorzy przyznają się do niezależnych od nich ograniczeń obliczeniowych, w drugim widać już w tym względzie sporą poprawę („For all model training, we used 4 x NVIDIA A100 80GB GPUs"). Niezależnie od nieporównywalności obu zadań i sytuacji, czy nie oznacza to, że dostęp do zasobów obliczeniowych staje dziś jednym z najważniejszych elementów warunkujących uprawianie inżynierii lingwistycznej?

Rozdział 6, choć zawarty w części „wdrożeniowej", jest wciąż bliższy artykułowi naukowemu niż opisowi systemu, w przeciwieństwie do rozdziału 7, który mógłby się znaleźć w ścieżce „system demonstrations". Rozdział 8 to już klasyczny dwustronicowy abstrakt, a nie artykuł, zaprezentowany zresztą na konferencji EAMT 2022 w ścieżce projektowo-produktowej. Zakładam, że taki układ pracy jest właściwy dla doktoratów wdrożeniowych.

6. Ocena wiedzy Doktoranta i osiągnięcia celu rozprawy

Wkład Doktoranta we wszystkie prace należy ocenić jako kluczowy – od ustalenia metodologii pracy badawczej przez implementację systemów i przeprowadzenie ewaluacji. Mimo że deklaracje procentowego wkładu pozostałych współautorów publikacji nie są podane, już z samych opisów da się wyczytać, że Doktorant odegrał w ich powstaniu pierwszoplanową rolę.

W związku z formą pracy, nie stanowiącej monografii, ale serię artykułów, istniejący stan wiedzy omawiany jest po części w każdym z nich. Odwołując się do bieżącego stanu wiedzy Doktorant potwierdza bardzo dobrą orientację i stan wiedzy w zakresie informatyki.

W kolejnych przedstawionych pracach widać rozwój umiejętności technicznych Doktoranta, który stosuje coraz to bardziej złożone metody dochodzenia do coraz lepszych rozwiązań. Przynosi to także mierzalne efekty w postaci zwycięstwa w konkursie WMT w roku 2022 (w porównaniu z miejscami 8–12 i 11–13 w zadaniach z roku 2021).

Stwierdzam zatem, że kandydat posiada stosowną wiedzę w dyscyplinie informatyka.

7. Osiągnięcie celu rozprawy

Ocena rozprawy złożonej z artykułów jest z jednej strony ułatwiona, gdyż prace te zostały już poddane ocenie zewnętrznych recenzentów i ocenę tę pomyślnie przeszły. Na osobną uwagę zasługuje jednak układ rozprawy i wartość artykułów w ramach proponowanego cyklu. W mojej opinii przedstawiona rozprawa posiada taką wartość – stanowi oryginalne rozwiązanie postawionego problemu optymalizacji jakości w systemach tłumaczenia maszynowego, potwierdza ogólną wiedzę teoretyczną Doktoranta w dyscyplinie informatyki, dowodzi także Jego umiejętności samodzielnego zaplanowania i przeprowadzenia badań oraz, co w tym przypadku szczególnie ważne, uczestnictwa w ich wdrożeniu. W ten sposób osiągnięte wyniki uzyskują wartościowe i przetestowane w praktyce aspekty użytkowe, co zgodne jest z ideą doktoratu wdrożeniowego.

Powiązanie części badawczej z praktyczną również nie budzi wątpliwości – metody opisywane w rozdziałach 2–5, takie jak wspomaganie tłumaczenia algorytmem rozpoznawania nazw własnych czy użycie słownika zostały wdrożone w systemach opisywanych w rozdziałach 6–8.

8. Inne uwagi

Podobnie jak w poprzednich recenzowanych przeze mnie pracach chciałbym zwrócić uwagę na kwestię punktacji prac publikowanych w materiałach z warsztatów organizowanych przy wysoko punktowanych konferencjach międzynarodowych. W tabeli 1.1 Doktorant przypisał konferencji WMT 140 pkt, co wydaje mi się wartością niewłaściwą mimo oczywistej wysokiej rangi tej konferencji i ogromnego sukcesu Doktoranta w zadaniu ewaluacyjnym (wobec ogromnej konkurencji zespołów z całego świata). To jednak oczywiście zarzut do dla twórców ministerialnej listy, a nie do Autora rozprawy.

Jednocześnie chciałbym podkreślić moje wielkie uznanie dla Doktoranta za pomyślny udział w międzynarodowych zadaniach ewaluacyjnych o światowym poziomie, co jest w mojej opinii faktem niezwykle ważnym dla rozwoju polskiego środowiska naukowego. UAM ma w dziedzinie tłumaczenia maszynowego światową markę, którą Doktorant i jego współpracownicy kolejny raz potwierdzili.

Nawiązując do wcześniejszego komentarza o dużych wymaganiach obliczeniowych współczesnych systemów chciałbym na koniec przywołać symptomatyczne zdanie z końca rozdziału 4, dość dobrze opisujące potrzeby całego polskiego środowiska NLP: „Due to a lack of computing power and time, our experiments and submissions were based on single model training.”, wyrażając nadzieję, że doczekamy się już niedługo efektywnego systemu dystrybucji taniej mocy obliczeniowej ułatwiającego polskim zespołom inżynierii lingwistycznej, także z mniejszych jednostek, konkurowanie z ich odpowiednikami na całym świecie.

9. 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) moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

A. Czy rozprawa zawiera oryginalne rozwiązanie problemu naukowego? (wybierz jedną opcję stawiając znak **X**)

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

B. Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie informatyka?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

C. Czy kandydat posiada umiejętność samodzielnego prowadzenia pracy naukowej?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

PODPIS ZAUFANY
MACIEJ
OGRODNICZUK
29.05.2023 12:06:06 [GMT+2]
Dokument podpisany elektronicznie podpisem zaufanym

Podpis