Warszawa, 3.06.2023

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

Recenzja rozprawy doktorskiej mgr. Artura Nowakowskiego

Quality Optimization Methods in Neural Machine Translation Systems

Recenzja rozprawy doktorskiej mgr. Artura Nowakowskiego, zrealizowanej pod opieką prof. dr hab. inż. Krzysztofa Jassema, oraz opiekuna pomocniczego dr. Macieja Lisonia, wykonana została na zlecenie Rady Naukowej dyscypliny informatyka Uniwersytetu Adama Mickiewicza w Poznaniu.

Przedstawiona rozprawa jest doktoratem wdrożeniowym zrealizowanym we współpracy z firmą Poleng, gdzie wdrożono opracowane w pracy rozwiązania.

Praca składa się z dwóch części - badawczej i wdrożeniowej. W pierwszej części umieszczone zostały 4 artykuły konferencyjne. Dwa z nich prezentowane były w 2021 i 2022 na najbardziej znanej konferencji dotyczącej tłumaczenia maszynowego – Conference on Machin Translation (WMT), która odbywa się łącznie z konferencją EMNLP. Część drugą stanowią trzy prace z konferencji międzynarodowych opisujące wykorzystanie opracowanych metod.

Spis artykułów:

1. Neural Machine Translation with Inflected Lexicon A. Nowakowski, K. Jassem MT Summit 2021
2. Approaching English-Polish Machine Translation Quality Assessment with Neural-based Methods A. Nowakowski PolEval 2021
3. Adam Mickiewicz University's English- Hausa Submissions to the WMT 2021 News Translation Task A. Nowakowski, T. Dwojak, WMT 2021 (EMNLP 2021)
4. Adam Mickiewicz University at WMT 2022: NER-Assisted and Quality-Aware Neural Machine Translation, A. Nowakowski, G. Pałka, K. Guttmann, M. Pokrywka, WMT 2022 (EMNLP 2022)
5. A Neural Translator Designed to Protect the Eastern Border of the European Union, A. Nowakowski, K. Jassem, MT Summit 2021
6. nEYron: Implementation and Deployment of an MT System for a Large Audit & Consulting Corporation A. Nowakowski, K. Jassem, M. Lison, R. Jaworski, T. Dwojak, K. Wiater, O. Posesor, EAMT 2022
7. POLENG MT: An Adaptive MT Platform, A. Nowakowski, K. Jassem, M. Lison, K. Guttmann, M. Pokrywka, EAMT 2022

Tematem, którym zajął się autor rozprawy było opracowanie metod pozwalających na podniesienie jakości tłumaczenia maszynowego dokonywanego za pomocą sieci neuronowych typu transformer. W szczególności badał on możliwość i efekty wprowadzania do sieci dodatkowych informacji.

Pierwszy artykuł rozprawy opisuje eksperymenty polegające na próbie polepszenia wyboru konkretnego tłumaczenia sekwencji słów w przypadku terminów dziedzinowych, czy innych utartych zwrotów, które w innym niż standardowe brzmieniu są niezrozumiałe lub

niezręczne. Narzucanie konkretnych elementów wyjściowych nie jest jednak proste w przypadku sieci neuronowych, zwłaszcza dla języków fleksyjnych, gdyż słowo takie musi zostać nie tylko umieszczone w dobrym miejscu, ale także w odpowiedniej formie. Zaproponowana w pracy metoda, zastosowana do tłumaczenia z angielskiego na polski, wymaga, po pierwsze, zebrania odmienionych form wybranych sekwencji słów z korpusu zrównoleglonego. Następnie, dla tłumaczeń poniżej pewnego progu log-likehood, dodawane są słowa-tłumaczenia ze zbioru wcześniej ustalonych par. Jeśli dla uzyskanych zdań wynik przekracza ustaloną granicę, dokonywany jest wybór nowego tłumaczenia. Przytoczone przykłady pokazują na dobre zstąpienie złych tłumaczeń, ale w ilościowej ewaluacji osiągnięta została wartość miary BLEU tylko nieznacznie wyższa, a zwiększony został dość istotnie czas tłumaczenia. W pracy zaproponowano też nowe, analogiczne do WER miary błędów popełnianych przez system przy sterowaniu procesem tłumaczenia poprzez narzucanie wymaganych sekwencji.

Drugi artykuł poświęcony jest metodom ewaluowania jakości tłumaczenia maszynowego w kontekście tłumaczenia z angielskiego na polski. Praca wykonana była na danych stanowiących element zadania w konkursie Poleval. Opracowane miary dotyczyły dwóch konfiguracji danych. W jednej znane były teksty tłumaczone, w drugiej tylko same tłumaczenia. Ponadto znane były oceny dokonane niezależnie przez kilka osób. W przypadku, gdy znane były teksty tłumaczone, przeprowadzono eksperymenty z kilkoma miarami wykorzystującymi schemat uczenia na danych pochodzących z rzeczywistych ocen dokonanych przez człowieka-czytelnika. Przede wszystkim wytrenowano uznawaną za najlepszą obecnie miarę COMET. Sieć wykorzystującą polski model typu Bert (HerBERT i XLM-Roberta) dotrenowywano początkowo przy zamrożonych wagach wewnętrznych, a następnie z niewielkim krokiem, zmieniane były już wszystkie wagi. Ponadto, przetestowano metody Bleurt, TransQuest i BERTScore. Najlepsze wyniki zgodności z ocenami dokonywanymi przez ludzi uzyskano dla miary COMET przy użyciu modelu HerBERT, niewiele gorsze były wyniki dla Bleurt.

Dla sytuacji, w której nieznane były dane oryginalne, ostatecznym wyborem okazała się też być miara COMET. Aby uzupełnić dane dokonano wstecznej translacji tłumaczeń na język oryginału (angielski) używając ogólnie dostępnego systemu OPUS-MT. Dane te połączono z tymi, dla których znane były oryginalne tłumaczenia i jako zbiór uczący wylosowano 100 par dokumentów. Poza COMET wytrenowano też miarę TransQuest, ale podobnie jak poprzednio najlepsze wyniki uzyskano dla COMET w wersji z modelem HerBERT.

Trzeci artykuł z pierwszej części pracy zawiera opis rozwiązania zgłoszonego na konkurs tłumaczenia wiadomości zorganizowany przy WMT 2021. Dotyczyło ono tłumaczenia z angielskiego na hausa (i odwrotnie). Hausa to język, dla którego zasoby elektroniczne są znacznie mniejsze i który jest znacząco odmienny od angielskiego. Opracowane zostały dwa rodzaje modeli – model statystyczny wykorzystujący frazy i model neuronowy. Ten drugi oparty był na architekturze transformer. Jako rozwiązanie początkowe wybrano model statystyczny, gdyż osiągnął lepsze wyniki. By polepszyć wyniki sięgnięto do metody transfer-learning i wykorzystano dane angielsko-niemieckie. Jako dodatkowe rozszerzenie danych treningowych wykorzystano też jednojęzyczne zbiory w hausa i 5mln zbiór angielski, przetłumaczone w sposób automatyczny na drugi język. Procedurę trenowania i tłumaczenia automatycznego powtarzano iteracyjnie.

Czwarty artykuł rozprawy to także opis rozwiązania zgłoszonego na konkurs, na kolejną konferencję WMT w roku 2022 . Badanym zadaniem było tu uzyskanie tłumaczenia między ukraińskim i czeskim. Zadanie dotyczyło zatem języków dość bliskich, ale o niezbyt obszernych zasobach. Tradycyjny model tłumaczenia, wykorzystujący architekturę typu Bert, został wzbogacony poprzez użycie różnych metod, takich jak: uczenie transferowe, tłumaczenie wsteczne, tłumaczenie wspomagane NER, tłumaczenie na poziomie dokumentu, ważone kombinowanie rozwiązań i adaptacja domeny w locie. Użycie kombinacji tych me-

tod na zbiorze testowym doprowadziło do wzrostu wyników COMET o 0,22 (26,13%) przy tłumaczeniu z ukraińskiego na czeski i o 0,19 (24,18%) przy tłumaczeniu z czeskiego na ukraiński. Zaproponowane modele zajęły pierwsze miejsce.

Część obejmującą prace wdrożeniowe otwiera artykuł opisujący system tłumaczący włączony do wyszukiwarki tekstów poskich, ukraińskich, rosyjskich i białoruskich, które mogą zawierać informacje o potencjalnych czynach zabronionych. Program powstał na potrzeby polskiej straży granicznej. System został wytrenowany na tekstach ogólnych. Adaptacja na potrzeby specjalistycznej domeny polegała na zapewnieniu odpowiednich tłumaczeń terminów kryminalnych i nazw własnych, takich jak imiona, nazwiska i obiekty geograficzne. Wykorzystano tu metodę opisaną w pracy [2]. Ze względu na bardzo mają ilość danych polsko-białoruskich, opracowany został wielojęzyczny system polsko-rosyjsko-ukraińsko-białoruski. Niestety jakość tłumaczeń z i na białoruski nadal była gorsza niż Google Translate, podczas gdy dla rosyjskiego i ukraińskiego wyniki były lepsze. Metody sterowania tłumaczeniem wyrażeń specjalistycznych i nazw własnych przynoszą zatem korzyści dla tekstów z ograniczonego tematycznie zakresu.

Druga praca przedstawia implementację i wdrożenie systemu MT w polskim oddziale EY Global Limited. Na całość rozwiązania składają się: aplikacje webowa, pamięć tłumaczeń i Machine Translation Service. System tłumaczący obsługuje standardowe funkcje CAT i MT, takie jak wyszukiwanie rozmyte w pamięci tłumaczeniowej, tłumaczenie dokumentów i post-edycję, a także spełnia mniej typowe, specyficzne oczekiwania klientów. Potrzeba opracowania systemu związana była z wprowadzeniem prawa nakładającego na firmę obowiązek tłumaczenia znacznej części teksów na polski. Projekt realizowany był przez firmę Poleng i trwał trzy lata. Artykuł opisuje poszczególne elementy systemu, do którego budowy wykorzystano gotowe, reprezentujące aktualny stan wiedzy na temat MT metody i programy, oraz drogę jaką trzeba było przejść od prototypu do rozwiązania końcowego i problemy związane z jego faktycznym uruchomieniem. Wykonano też ewaluację systemu porównując wyniki osiągane na danych ogólnych i system dotrenowywany na danych dziedzinowych.

Trzeci artykuł drugiej części rozprawy przedstawia POLENG MT, platformę MT, która może być używana jako jako aplikacja internetowa w chmurze lub jako rozwiązanie lokalne. Platforma zapewnia tłumaczenie dokumentów, w tym transfer formatowania. Główną cechą wersji lokalnej jest dedykowana adaptacja do potrzeb klienta, która polega na dotrenowaniu modelu na tekstach specjalistycznych i i zastosowanie wymuszonego tłumaczenia terminologii zgodnie z potrzebami użytkownika. Program działa dla tłumaczeń między parami: polski-angielski, polski-ukraiński i polski-rosyjski.

Zgodnie ze sformułowanym celem, zawarte w rozprawie artykuły dotyczą metod ulepszenia automatycznego tłumaczenia tekstów. Standardowym obecnie punktem wyjściowym prac związanych z MT jest trenowanie na dużych zbiorach danych sieci typu transformer. Jednak w wielu przypadkach osiągane rezultaty nie są jeszcze zadowalające. I takimi właśnie sytuacjami zajmował się doktorant. Prace doktoranta w szczególności dotyczyły tłumaczenia między językami odległymi, różniącymi się przykładowo fleksją czy wykorzystaniem rodzajów gramatycznych oraz takimi, dla których ilość danych jedno i wielojęzycznych jest niezbyt duża. Poza tymi problemami doktorant zajął się też poprawianiem tłumaczenia terminów specjalistycznych, dla których istotne jest by były tłumaczone dokładnie tak, w jakiej formie pojawiają się w danym języku. Wszystkie te problemy mają duże praktyczne znaczenie, w szczególności dla tłumaczenia z i na język polski. Zaproponowane rozwiązania przynosiły na ogół niezbyt duże, ale widoczne poprawienie wyników.

Konstrukcja doktoratu ze zbioru artykułów stanowi pewne wyzwanie przy ocenie osiągnięcia doktoranta. Artykuły przeznaczone do prezentowania na konferencjach specjalistycznych z konieczności są krótkie i opisują rozwiązanie często niezbyt dokładnie. Trudno

też precyzyjnie ustalić na podstawie oświadczeń wkład autora w ich powstanie. Podanie jako wkładu własnego koncepcji pracy oczywiście podnosi wartość naukową tego udziału, jest to jednak chyba niezbyt precyzyjne sformułowane w przypadku, gdy współautorami jest promotor lub promotor pomocniczy. Niezależnie jednak od tego, uważam, że odnalezienie się zarówno w grupie badawczej osiągającej dobre wyniki jak i we wprowadzającej te rozwiązania do praktycznego użycia firmie komercyjnej dowodzą osobistych umiejętności i osiągnięć doktoranta.

Miejsca publikacji artykułów dowodzą, że wyniki osiągane przez doktoranta (wspólnie z zespołem) znajdują się wśród tych prac, które nadają kierunek obecnym poszukiwaniom najlepszych rozwiązań. Ze względu na to, że wstęp do przedstawionych artykułów jest bardzo krótki, a rozprawa składa się z opublikowanych, dobrze napisanych artykułów, nie ma właściwie potrzeby ani pola do oceny edytorskiego aspektu rozprawy, poza jedną ogólną uwagą. Dla pełniejszego przedstawienia wyników i umieszczenia ich w kontekście oczywiście dłuższy wstęp zawierający zarówno wprowadzenie jak i komentarz byłby jednak bardzo pożądany. Moje uwagi do pracy są także tylko bardzo ogólnej natury, gdyż przedstawione w pracy rozwiązania nie budzą żadnych moich zastrzeżeń. Po pierwsze, konstrukcja i proces uczenia sieci neuronowych mają swój formalny, matematyczny opis, jednak w artykułach konferencyjnych nie jest on (nie mógł być) przedstawiony. Po drugie, czasem brakuje przetestowania innych wariantów proponowanych rozwiązań. Przykładowo być może dodanie słowników zewnętrznych nie tylko pozyskanych z tekstów dziedzinowych spowodowałoby osiągnięcie lepszych rezultatów na tekstach ogólnych, a może też dziedzinowych (problem odmiany można było próbować rozwiązać korzystając z istniejących zasobów)? Z kolei w przypadku tłumaczenia z i na hausa i transfer learnig wykorzystano dane niemieckie. Oczywiście jest to język z dużymi zasobami, ale też dość podobny do angielskiego i zapewne mało podobny do hausa. Nie ma w pracy uzasadnienia, że lepszego wyboru nie było. Wydaje się jednak, że w kontekście doktoratu wdrożeniowego obie moje uwagi mają niezbyt istotne znaczenie i nie zmieniły one mojej pozytywnej oceny.

Wniosek końcowy

Stwierdzam, iż przedłożona mi do recenzji rozprawa, której autorem jest mgr Artur Nowakowski, zawiera opis ważnych osiągnięć w dziedzinie poprawy wyników neuronowego uczenia maszynowego dla wybranych scenariuszy, co zostało potwierdzone zajęciem wysokiego miejsca w konkursach organizowanych na wiodącej konferencji z tej dziedziny. Mgr Artur Nowakowski uczestniczył w budowaniu praktycznych systemów realizujących zadanie MT i w faktycznych wdrożeniach opracowywanych metod tłumaczenia będąc współautorem rozwiązań firmy Poleng. Doktorant wykazał się wiedzą w tematyce rozprawy oraz znajomością aktualnych osiągnięć w dziedzinie, której dotyczył doktorat i umiejętnością jej wykorzystania we własnej pracy badawczej i wdrożeniowej. Moim zdaniem recenzowana rozprawa spełnia wymagania ustawowo stawiane rozprawom doktorskim, zatem wnoszę o dopuszczenie magistra Artura Nowakowskiego do publicznej obrony.

A. Wykowska