Warszawa, 1.06.2021

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

Recenzja rozprawy doktorskiej mgr. Tomasza Dwojaka

„Efficient algorithms for hybryd neural machine translation" przygotowanej pod kierunkiem promotora dr. hab. Krzysztofa Jassema i promotora pomocniczego dr. Marcina Junczys-Dowmunta

Praca doktorska „Efficient algorithms for hybryd neural machine translation" Tomasza Dwojaka została wykonana pod kierunkiem prof. UAM dr. hab. Krzysztofa Jassema na Wydziale Matematyki i Informatyki Uniwersytetu Adama Mickiewicza w Poznaniu. Rozprawę stanowi sześć artykułów napisanych we współautorstwie i opublikowanych w latach 2016-2019. Cztery z nich to materiały konferencyjne z międzynarodowych konferencji i warsztatów poświęconych problematyce tłumaczenia maszynowego (MT), jeden to opis systemu prezentowanego na najważniejszej konferencji z dziedziny lingwistyki komputerowej – ACL, a ostatni jest artykułem z czasopisma z listy JCR.

1. M. Junczys-Dowmunt, T. Dwojak, H. Hoang Is neural machine translation ready for deployment? A case study on 30 translation directions, *The 13th International Workshop on Spoken Language Translation* (IWSLT), 2016.
2. M. Junczys-Dowmunt, T. Dwojak, R. Sennrich, The AMU-UEDIN submission to the WMT16 news translation task: Attention-based NMT models as feature functions in phrase-based SMT, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 2016
3. M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, ... Marian: Fast neural machine translation in C++, *Proceedings of ACL 2018, System Demonstrations*
4. M. Nadejde, S. Reddy, R. Sennrich, T. Dwojak, M. Junczys-Dowmunt, ..., Predicting target language CCG supertags improves neural machine translation, *Proceedings of the Second Conference on Machine Translation, 2017*
5. Hieu Hoang, Tomasz Dwojak, Rihards Krislauk, Daniel Torregrosa, Kenneth Heafield, Fast NeuralMachine Translation Implementation, *2ndWorkshop on Neural Machine Translation and Generation, Melbourne,* 2018.
6. Tomasz Dwojak, Krzysztof Jassem, Statistical versus Neural Machine Translation – a Case Study for a Medium Size Domain-Specific Bilingual Corpus, *Poznań Studies of Contemporary Linguistics.*

Pierwszy artykuł *rozprawy, Is neural machine translation ready for deployment? A case study on 30 translation directions,* przedstawia porównanie różnych systemów MT wykorzystujących podejście statystyczne (frazowe) i sieci neuronowe (ang. *attention based encoder-decoder approach*) w zadaniu tłumaczenia między sześcioma oficjalnymi językami Organizacji Narodów Zjednoczonych. Jako dane wykorzystano korpus United Nations Parallel Corpus v1.0 z lat 1990-2014. W pracy wykazano, że

architektury sieci neuronowych mogą być najmniej tak dobre lub znacznie lepsze niż metody statystyczne. Największy postęp przy zmianie metody odnotowano dla tłumaczeń z i na język chiński. Dla sprawdzenia, czy tak dobre wyniki można uzyskać ulepszając rozwiązania metodami statystycznymi, dla części z par języków zaproponowano hierarchiczną wersję frazowego modelu tłumaczenia statystycznego. Wyniki, aczkolwiek lepsze niż pierwotne, nie okazały się tak dobre jak dla sieci neuronowej. Autorzy badali także szybkość tłumaczenia i zaprezentowali wyniki eksperymentów pokazujące efektywność implementacji modeli statystycznych i neuronowych z wykorzystaniem procesorów CPU i GPU i własnego narzędzia – AmuNMT – napisanego w C++ środowiska do implementacji neuronowego dekodera tłumaczeniowego. W opisanych eksperymentach wykazano możliwość znacznego przyspieszenia procesu tłumaczenia przy ograniczeniu zbioru podsłów dla tłumaczonego zdania do 1200 i ograniczenia szerokości wiązki przy przeszukiwaniu przestrzeni rozwiązań do 5.

Artykuł [2] przedstawia system tłumaczeń wiadomości prasowych, który brał udział w konkursie ogłoszonym w ramach konferencji WMT w 2016 roku. W systemie przetestowano możliwość wykorzystania informacji pochodzącej z modelu z atencją z frazowym modelem statystycznym, co było pierwszym rozwiązanie tego typu na świecie. W zadaniu tłumaczenia z rosyjskiego na angielski system ten był najlepszy zarówno w sensie miary BLEU jak i w ocenie dokonywanej przez ludzi.

W artykule [3] opisano opracowany przez autorów system Marian – efektywne środowisko do tworzenia systemów neuronowego maszynowego tłumaczenia. System ten został wdrożony w Biurze Międzynarodowym WIPO (ang. *World Intellectual Property Organization*). Zaimplementowano w nim kilka wiodących architektur neuronowych. W pracy pokazano wyniki wykorzystania systemu Marian do znalezienia rozwiązania kilku problemów NLP. Poza tłumaczeniem maszynowym była to poprawa pisowni i problem edycji wyników tłumaczenia maszynowego. W każdym przypadku osiągnięto rozwiązania zarówno efektywne obliczeniowo jak i dające wysokiej jakości wyniki.

W kolejnej pracy [4] autorzy zbadali, wydaje się, że jako pierwsi, na ile informacje o budowie syntaktycznej zdań języka docelowego wpływają na jakość tłumaczenia neuronowego. W eksperymentach na danych z konkursów WMT dostarczenie tych informacji (zapisanych jako tagi CCG, *combinatory categorial grammar*) poprawiło jakość tłumaczenia zarówno dla par języków o dużej liczbie zasobów tekstowych, jak i dla języków o niewielkich zasobach. Poprawa nie była wielka, ale świadczy o tym, że podanie informacji składniowej dla sieci neuronowej może być pomocne.

Badania przedstawione w artykule [5] poświecone były jest problemowi zwiększenia efektywności neuronowego tłumaczenia maszynowego. O ile jakość tak uzyskiwanych tłumaczeń jest lepsza niż tych osiąganych metodami statystycznymi, to czas tłumaczenia znacząco się zwiększył. Wydajność proponowanych rozwiązań często nie leży w zakresie zainteresowania grup walczących o jak najlepszy wynik jakościowy, a w niektórych sytuacjach zbyt długie oczekiwanie na tłumaczenie może być istotnym mankamentem. Autorzy dokonali analizy czasu, jaki poświęcany jest na poszczególne fazy dekodowania i skupili się na poprawieniu efektywności najdłużej trwających operacji. Eksperymenty przeprowadzono przy wykorzystaniu narzędzia AmuNMT, w którym wprowadzono dwa usprawnienia: algorytm mini-batching oraz połączenie metod softmax i k-best search. Implementacja algorytmu mini-batching oraz sam algorytm połączenia metody k-best serach z warstwą softmax są autorstwa doktoranta.

Ostatnia ze stanowiących rozprawę prac [6] opisuje wykonane wspólnie z promotorem eksperymenty polegającego na wykorzystaniu opracowanych metod maszynowego tłumaczenia do stosunkowo niewielkiego ograniczonego tematycznie podzbioru symulującego dane, jakimi może dysponować jedna niezbyt wielka firma komercyjna kolekcjonująca teksty związane z własną działalnością. Na takich danych (dotyczących budowy statków) wytrenowano zarówno modele statystyczne jak i neuronowe. Te pierwsze zostały ocenione wyżej przez ludzi, ale według miary BLEU nieco wyższą ocenę uzyskały modele statystyczne.

Większość stanowiących treść rozprawy prac została opublikowana w materiałach najlepszych konferencji w zakresie lingwistyki komputerowej i szczegółowej – maszynowego tłumaczenia (nie odnoszę się tu do liczby punktów podanych w autoreferacie, która według mnie jest w przypadku jednej pracy inna niż podana, ale nie zmienia to mojej wysokiej oceny merytorycznej miejsc publikacji). Dwie z nich są licznie cytowane w Google Scholar. Oznacza, że zostały zauważone przez szerszą grupę czytelników, co wobec ogromnej liczby publikowanych obecnie prac nie jest łatwe do osiągnięcia.

Opisane w artykułach badania dotyczyły maszynowego tłumaczenia w bardzo ciekawym momencie jego rozwoju, czyli wtedy gdy nowe pomysły na architektury sieciowe i techniki uczenia sieci w połączeniu z możliwościami sprzętu pozwoliły na wykorzystanie z sukcesem sieci neuronowych do MT. W ramach swoich prac doktorant implementował i modyfikował uznawane za najbardziej skuteczne architektury sieciowe, testował różne wpływające na jakość i czas tłumaczenia parametry, badał skuteczność poszczególnych konfiguracji dla wielu par języków i wielu typów danych, w tym danych ogólnych i specjalistycznych danych o ograniczonej objętości. Osiągane wyniki porównywane były z tymi dla z systemów statystycznego frazowego tłumaczenia, w tym zaproponowanej metody hierarchicznego tłumaczenia frazowego. Dużą uwagę w pracach, w których uczestniczył doktorant zwracano na efektywność proponowanych rozwiązań. Tomasz Dwojak brał udział w implementacji efektywnych narzędzi NLP: systemu przeznaczonego do budowania systemów NLP wykorzystujących sieci neuronowe (Marian) i narzędzia do budowy dekodera tłumaczeniowego AmuNMT. Ciekawym wątkiem prac jest też próba wzbogacenia informacji o zdaniach w języku docelowym o tagi CCG, a zatem próba wykorzystania wiedzy dotyczącej parsowania składniowego podanej w sposób bezpośredni jako uzupełnienie danych tekstowych dla języka, na który tekst jest tłumaczony.

Celem prowadzonych przez doktoranta prac było przetestowanie zarówno mocy różnorodnych metod maszynowego tłumaczenia, zbadania wpływu na osiągane wyniki i efektywność przyjęcia różnych wartości ważnych dla procesu uczenia i dekodowania parametrów, porównania systemów neuronowych z frazowym modelem statystycznym, a także dokonanie prób połączenia tych dwóch podejść. Jakkolwiek obecnie systemy NMT zdecydowanie dominują, w rozprawie dowodzi się, że w pewnych okolicznościach (na przykład małe dane dziedzinowe) systemy SMT mogą mieć jakość (mierzoną automatycznie) nawet wyższą niż systemy NMT. Jednak nawet wtedy teksty generowane przez systemy NMT, jako bardziej płynne, są zwykle lepiej odbierane przez ludzi. W pracy [2] udało się także autorom zaproponować model hybrydowy łączący cechy modelu neuronowego i statystycznego i wykazać, że może on skutecznie rywalizować w systemami czysto NMT.

Przedstawione prace doktorant wykonywał w dużym stopniu we współpracy z jednym z najlepszych ośrodków zajmujących się maszynowym tłumaczeniem - Uniwersytetem w Edynburgu, na którym do 2017 zatrudniony był promotor pomocniczy, sam będący uznanym autorytetem w dziedzinie MT

(obecnie Microsoft Translator). Trwająca kilka lat stała współpraca także z innymi badaczami i stanowiące treść rozprawy wspólne publikacje są pośrednim dowodem na ich wysoką ocenę jakości pracy doktoranta. Opracowywane implementacje współzawodniczyły w ramach konkursów międzynarodowych z innymi rozwiązaniami czołowych grup zajmujących się MT i zajmowały wysokie, a nawet pierwsze miejsca. Dowodzi to, że artykuły zawarte w doktoracie opisują rozwiązania, które stanowiły czołowe osiągnięcia w tramach MT w momencie ich publikacji, a doktorant uczestniczył w rozwoju tej dziedziny jako członek jednej z najbardziej znanych i odnoszących duże sukcesy naukowych grup badawczych wnosząc do niej swój własny wkład.

Podsumowując, stwierdzam, że recenzowana rozprawa doktorska mgra Tomasza Dwojaka „Efficient algorithms for hybryd neural machine translation" spełnia warunki stawiane rozprawom doktorskim i wnoszę o jej publiczną obronę.

A. Mykowiecka

Agnieszka Mykowiecka