Recenzja pracy doktorskiej mgr. Marcina Sowańskiego

„Multilingual Machine Translation System for Dialogue Agents"

Krzysztof Jassem

20 grudnia 2023

1 Wstęp

Celem niniejszej recenzji jest stwierdzenie, czy rozprawa doktorska mgr. Marcina Sowanskiego „Multilingual Machine Translation System for Dialogue Agents" spełnia wymagania Ustawy o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. 2003 Nr 65, poz. 595, ze zm.).

2 Ocena merytorycznej strony rozprawy

Recenzowana rozprawa jest wynikiem doktoratu wdrożeniowego, który z założenia powinien dotyczyć badań związanych z wdrożeniem w podmiocie gospodarczym. W opisywanym przypadku wdrożeniem, któremu podporządkowana jest rozprawa, jest wielojęzyczny system komunikacji między człowiekiem a wirtualnym asystentem, realizowany w firmie Samsung. Charakterystyczną cechą dialogów w tymże systemie jest lakoniczność wypowiedzi oraz ich ukierunkowanie na wykonanie określonej akcji przez system informatyczny. Dodatkowym założeniem jest wielojęzyczność – w rozprawie Autor koncentruje się na trzech językach: polskim, angielskim i hiszpańskim, natomiast wyniki badań mają na celu ich uogólnienie na większą liczbę języków.

Rozwiązanie przyjęte w rozprawie zakłada przetwarzanie sekwencyjne wypowiedzi użytkownika: rozpoznawanie mowy, rozumienie wypowiedzi, wygenerowanie reakcji i wykonanie akcji (lub synteza reakcji głosowej systemu). Aby ułatwić realizację najbardziej złożonego modułu – rozumienia wypowiedzi – Autor proponuje metodę języka pośrednika, do którego tłumaczone są wypowiedzi. W tej sytuacji kluczowym zagadnieniem staje się poprawne tłumaczenie wypowiedzi z oraz na język pośredni – angielski.

W dziedzinie tłumaczenia automatycznego notuje się w ostatnich latach stały postęp nie tylko w jakości translacji (co dokumentowane jest coraz lepszymi wynikami rozwiązań na organizowanych konkursach), lecz również w uniwersalności rozwiązań, co realizowane jest przez publiczny dostęp do modeli umożliwiających tłumaczenie między dziesiątkami języków. Niszą w tym zagadnieniu, której wypełnienie jest intencją rozprawy, jest tłumaczenie krótkich wypowiedzi, które mają na celu przekazanie instrukcji postępowania.

Autor realizuje tę intencję poprzez realizację trzech zadań badawczych: wytrenowanie systemu translacji dopasowanego do języka asystentów głosowych, dostosowanie systemu do transferu slotów (czyli elementów wypowiedzi najbardziej istotnych dla zrozumienia instrukcji postępowania) oraz

opracowanie metody augmentacji danych uczących poprzez tłumaczenie z wariantami czasowników.

W kolejnych częściach recenzji ocenię realizację poszczególnych, wyżej wymienionych zadań.

2.1 Wytrenowanie systemu translacji dostosowanego do języka asystentów głosowych

Standardową metodą dostosowania (dostrojenia) modelu języka do konkretnej dziedziny tłumaczenia automatycznego jest dostarczenie korpusu specyficznych przykładów tłumaczenia. Celem Autora było opracowanie korpusu wypowiedzi, który byłby zrównoważony pod względem wyszczególnionych domen dialogów oraz zdefiniowanego zbioru intencji. Autor postanowił wygenerować taki korpus automatycznie na podstawie manualnie opracowanych gramatyk.

W tym celu Autor opracował 20 wzorców gramatycznych dla generowania wypowiedzi języka angielskiego, z których każdy stał się podstawą do wygenerowania zestawu zdań o różnych stopniach naturalności i zawierających różne intencje.

Podobną pracę tę wykonano dla języków hiszpańskiego i polskiego, uzyskując korpus wypowiedzi w trzech językach.

Łza się w oku kręci. Przypominają się lata 80., kiedy za pomocą ręcznie opracowanych gramatyk próbowano ogarnąć ogrom konstrukcji języka polskiego. Na podstawie reguł tego typu autor niniejszej recenzji zmagał się z problemem tłumaczenia z języka polskiego na język angielski...

...z efektem podobnym do wyniku raportowanego w rozprawie. W stosunku do modelu bazowego (publicznego modelu M2M100) uzyskano w ramach rozprawy co prawda znaczący postęp wyrażony w metryce BLUE i nieco mniejszy w metryce BLEURT, jednak zanotowano bardzo wyraźny spadek jakości na zdaniach spoza dziedziny. Fakt ten stawia pod znakiem zapytania odporność metody na wypowiedzi wykraczające poza zdefiniowaną gramatykę.

Podsumowując tę część recenzji, stwierdzam, że uzyskane wyniki nie przekonują mnie o wyższości zastosowanej metody pozyskiwania danych nad alternatywną metodą – crowd sourcingu – której efektem byłoby pozyskanie danych bardziej zbliżonych do autentycznych.

2.2 Dostosowanie systemu do transferu slotów

W zadaniu transferu slotów Autor dostrajał model bazowy (M2M100) z wykorzystaniem nie tylko korpusu autorskiego, ale również innych dostępnych korpusów dialogów z wirtualnym asystentem. Wynik eksperymentu wydaje się pozytywny: nastąpił prawie dwukrotny wzrost jakości tłumaczenia w stosunku do modelu bazowego, a miara F1 określająca poprawność lokalizacji slotów w tekście docelowym wynosiła ponad 65%.

Z drugiej strony pozostają pewne niedopowiedzenia:

1. Dlaczego w pomiarze jakości tłumaczenia zastosowano wyłącznie wychodzącą z użycia metrykę BLEU?

2. Do jakiego rozwiązania można porównać jakość lokalizacji slotów?

Podsumowując tę część recenzji, chciałbym podkreślić, że metoda zastosowana przez Autora pozwoliła na osiągnięcie wyraźnego postępu w stosunku do metody bazowej. Podane przez Autora dane nie pozwalają jednak czytelnikowi na wyciągnięcie wniosków dotyczących odporności metody oraz porównawczej oceny jakości lokalizacji slotów w tekście docelowym.

2.3 Opracowanie metody augmentacji danych trenujących poprzez tłumaczenie z wariantami czasowników

W mojej ocenie trzeci podrozdział pracy jest najciekawszy, gdyż łączy kompetencje ludzkie ze stosunkowo nowymi metodami tłumaczenia leksykalnego przez system neuronowy. Kompetencje ludzkie niezbędne są do opracowania klasyfikacji czasowników stosowanych w instrukcjach podawanych do asystentów wirtualnych – do jednej klasy mają należeć czasowniki o podobnym znaczeniu w określonym kontekście. Klasyfikacja ta pomaga wygenerować kilka tłumaczeń o podobnym znaczeniu różniących się wyłącznie tłumaczeniem czasownika. Warto zauważyć, że metodologia ingerowania w rezultaty tłumaczenia neuronowego za pomocą leksykonów jest ciekawym wyzwaniem, szczególnie dla języków fleksyjnych.

Niezwykle trudne do analizy są wyniki ewaluacji tych eksperymentów. Okazuje się bowiem, że tłumaczenie z leksykonem okazuje się gorsze niż bez niego. Jest to jednak fakt ogólnie znany i potwierdzony w wielu eksperymentach (z tego powodu leksykony stosuje się tylko w bardzo specyficznych przypadkach tłumaczenia). Optymizmem napawa fakt, że tłumaczenie z trzema wariantami daje lepsze wyniki niż tłumaczenie z jednym wariantem.

Pozytywnie oceniam pomysł przedstawiony w tej części rozprawy. Jedyne zastrzeżenie mam do tego, że nie wyjaśniono, dlaczego wynik niższy od bazowego można uznać w tym wypadku za pozytywny.

2.4 Zastosowania

W mojej ocenie najwyższą wartością rozprawy jest pozytywne wdrożenie jej wyników do gospodarki. Jeżeli przyjmiemy, że najbardziej obiektywną miarą jakości rozwiązania informatycznego jest ewaluacja zewnętrzna, czyli "money in the bank" lub wysoki stopień zadowolenia użytkowników, to szósty rozdział pracy wykazuje, że cel ten został spełniony. Rozwiązanie zostało wdrożone w systemie Bixby, stosowanym przez miliony użytkowników, którzy pozytywnie oceniają swoje doświadczenia.

3 Ocena formalnej strony rozprawy

3.1 Język

Praca napisana jest w języku angielskim. Autor posługuje się językiem obcym bardzo swobodnie. Niezwykle trudno jest dostrzec jakiekolwiek błędy językowe. Nie ma w rozprawie zbędnych zawiłych konstrukcji zdaniowych – praca napisana jest w sposób klarowny.

3.2 Ocena układu pracy

Układ pracy jest zgodny z przyjętymi standardami. Wstęp zawiera dokładnie tyle informacji, ile jest niezbędnych dla zrozumienia dalszych części. Nie zauważyłem odwołań do niezdefiniowanych wcześniej pojęć.

3.3 Strona estetyczne

Stronę estetyczną oceniam bardzo wysoko. Format wydruku jest czytelny, praca jest ładnie złożona. Czytelność znacząco poprawiają diagramy przepływu danych, które są czytelne i nieskomplikowane.

3.4 Spójność opisu

W ocenie spójności opisu biorę pod uwagę dwa kryteria: jednorodność stosowanej terminologii oraz spójność raportowanych danych. W ramach pierwszego kryterium nie mam większych zastrzeżeń poza nazewnictwem slotów. W tytule rozdziału 4. mowa jest o tłumaczeniu i transferze jednostek (Entity Translation and Transfer), w hipotezie T2 (str. 24) mowa jest o tłumaczeniu jednostek nazwanych, a w większości pracy mowa jest o transferze slotów.

Pod względem spójności raportowanych danych rozprawa pozostawia wiele do życzenia: Liczba domen tłumaczenia, która niewątpliwie powinna być stała w całej pracy, waha się od wartości 18 (wstęp) do 20 (tabela 1) i 21 (tabela 2). Liczba intencji wynosi albo 186 (wstęp), albo 187 (tabela 2), albo nawet 193 (str. 41).

4 Podsumowanie recenzji

W podsumowaniu recenzji odwołuję się do poszczególnych ustępów ustawy o rozprawie doktorskiej.

1. ust. 1. „Rozprawa doktorska prezentuje ogólną wiedzę teoretyczną kandydata w dyscyplinie albo dyscyplinach oraz umiejętność samodzielnego prowadzenia pracy naukowej lub artystycznej."

   Autor posiada ogólną wiedzę teoretyczną w dyscyplinie, co wykazuje w rozdziale 1. rozprawy.

   Główne wyniki rozprawy, dotyczące przygotowania korpusu Leyzer, transferu slotów oraz translacji wielowariantowej zostały również opisane w trzech publikacjach międzynarodowych

– w każdej z nich Doktorant jest wymieniony jako pierwszy Autor. Ponadto, Doktorant był liderem zespołu informatycznego, który wdrażał wyniki rozprawy do obszaru gospodarczego. Fakty te niepodważalnie świadczą o umiejętności samodzielnego prowadzenia pracy badawczej przez Doktoranta.

2. ust. 2. „Przedmiotem rozprawy doktorskiej jest oryginalne rozwiązanie problemu naukowego, oryginalne rozwiązanie w zakresie zastosowania wyników własnych badań naukowych w sferze gospodarczej lub społecznej albo oryginalne dokonanie artystyczne."

   Recenzowana rozprawa nie wnosi moim zdaniem nowatorskiego rozwiązania problemu naukowego. Autor w sposób niezwykle umiejętny stosuje i modyfikuje zastany aparat badawczy do rozwiązania problemu praktycznego, czym moim zdaniem wypełnia postulat „zastosowania wyników własnych badań naukowych w sferze gospodarczej".

3. ust. 3. „Rozprawę doktorską może stanowić praca pisemna, w tym monografia naukowa, zbiór opublikowanych i powiązanych tematycznie artykułów naukowych, praca projektowa, konstrukcyjna, technologiczna, wdrożeniowa lub artystyczna, a także samodzielna i wyodrębniona część pracy zbiorowej."

   W omawianym przypadku rozprawę stanowi praca pisemna w postaci monografii naukowej.

Stwierdzam zatem, że recenzowana rozprawa spełnia wymagania stawiane pracom doktorskim. Rekomenduję dopuszczenie Doktoranta do kolejnych etapów przewodu.

Krzysztof Jassem