
Recenzja rozprawy doktorskiej

Tytuł rozprawy: Enhancing Natural Language Understanding Systems and Development of Reliable and Efficient Methods of Their Quality Measurements

Autor: Mgr Jakub Hościłowicz

Promotor: Dr hab. inż. Artur Janicki, prof. PW

1 Tematyka rozprawy

Recenzowana rozprawa doktorska Pana mgr. Jakuba Hościłowicza dotyczy projektowania, implementacji, testowania i wdrażania systemów NLU (*natural language understanding*), a szczególny nacisk został położony w pracy na aspekty związane z ich praktycznym (przemysłowym) wykorzystaniem. Uważam, że tematyka rozprawy jest bardzo ciekawa i aktualna, a opracowane metody i algorytmy są bezpośrednią odpowiedzią na wyzwania związane z tworzeniem takich systemów. Doktorant bardzo świadomie odnosi się również do aspektów praktycznych (np. do minimalizacji kosztu, związanego z tworzeniem systemów wielojęzycznych, czy do przekrojowej i rzetelnej walidacji i diagnostyki systemów NLU, pozwalającej zweryfikować ich bezpieczeństwo oraz odporność na niepożądane i szkodliwe działanie, np. manipulację promptem), które bezpośrednio wpływają na możliwości wdrożenia i wykorzystania systemów NLU w praktycznych zastosowaniach.

2 Ocena strony merytorycznej rozprawy doktorskiej

Rozprawa doktorska jest napisana w języku angielskim i składa się z ośmiu numerowanych rozdziałów. **Rozdział 1.** jest wprowadzeniem do tematyki rozprawy, w którym Autor przedstawia kontekst naukowy i przemysłowy, w którym będzie się poruszał w swojej pracy oraz jasno motywuje istotność tematyki, którą się zajmuje.

Rozdział 2. jest krótkim (aczkolwiek treściwym) przeglądem literatury. Doktorant omawia podstawowe elementy i wyzwania związane z projektowaniem i wdrażaniem systemów rozumienia języka naturalnego (NLU) i zgrabnie umiejscawia swoje prace w trendach związanych z tworzeniem systemów NLU w latach 2022–2026. W kolejnych podrozdziałach przedstawiony został obecny stan wiedzy w zakresie oceny systemów NLU, również z wykorzystaniem metod opartych na probingu,

oraz tworzenia, oceny i analizy wielojęzycznych systemów NLU i multimodalnych agentów. Rozdział jest zwięźzony podsumowaniem poprawnie zidentyfikowanych luk badawczych, którymi Doktorant będzie zajmował się w ramach prac nad rozprawą. Warto podkreślić, że z tej części rozprawy (oraz z innych) emanuje głębokie zrozumienie Doktoranta aspektów związanych z wdrażaniem oraz praktycznym wykorzystaniem i oceną metod NLU w przemyśle.

W **rozdziale 3.**, Autor podsumował swoje najistotniejsze osiągnięcia w kontekście rozprawy doktorskiej. Rozdział rozpoczyna się od omówienia i umotywowania trzech hipotez badawczych, dotyczących kolejno tworzenia wielojęzycznych agentów (zwłaszcza zmniejszenia kosztu podejść, w których niezbędne jest wytworzenie oetykietowanych zbiorów treningowych w przypadku nowych języków), poprawy jakości systemów NLU bez konieczności ich dodatkowego trenowania oraz wszechstronnej oceny i analizy systemów NLU. Wyniki eksperymentów prowadzonych przez Doktoranta zostały opublikowane w szeregu artykułów (m.in. na bardzo dobrej konferencji *Interspeech*) – Autor krótko omawia publikacje, w których zawarto takie wyniki. Warto podkreślić, że Doktorant jest współtwórcą dwóch patentów związanych z tematyką rozprawy.

W kolejnych **rozdziałach 4–7**, Doktorant prezentuje najistotniejsze wyniki swoich badań. W **rozdziale 4.**, mgr Jakub Hościłowicz omawia procedurę *translate-train*, która umożliwia efektywniejsze tworzenie wielojęzycznych systemów SLU (z ang. *spoken language understanding*). Autor przedstawia wyzwania związane z tworzeniem takich rozwiązań, zwłaszcza dotyczące kosztu wdrażania nowych języków – ten proces może wiązać się z bardzo kosztownym krokiem przygotowania oetykietowanych zbiorów treningowych (dla każdego nowego języka). W opracowanej przez Doktoranta procedurze, duży model językowy (LLM, *large language model*) został wykorzystany do generowania takich oetykietowanych zbiorów treningowych (z j. angielskiego). Taka translacja maszynowa umożliwia również ocenę ekspercką nowych zbiorów treningowych, których jakość bezpośrednio wpływa na jakość systemu SLU. W rozdziale szczegółowo przebadano i omówiono również aspekty związane z praktycznym wdrożeniem takiego podejścia, np. na urządzeniach brzegowych.

Rozdział 5. skupia się na omówieniu metod poprawy agentów opartych na wykorzystaniu LLM-ów: Autor rozpoczyna ten rozdział od omówienia techniki *Non-Linear Inference-Time Intervention* (NL-ITI), w której wewnętrzne reprezentacje są modyfikowane, aby zwiększyć prawdziwość odpowiedzi modelu (metoda ta stanowi rozszerzenie techniki ITI). Wyniki eksperymentalne, otrzymane z wykorzystaniem zbioru benchmarkowego *TruthfulQA*, zaprezentowane w pracy pozwalają stwierdzić, że NL-ITI przewyższa metodę bazową ITI. Warto podkreślić, że Doktorant zweryfikował możliwości generalizacyjne opracowanych metod dla innych zbiorów danych. W kolejnej części tego rozdziału, mgr Jakub Hościłowicz przedstawia system *ClickAgent*, będący modułową architekturą, w której Doktorant odseparował lokalizację UI oraz wysokopoziomowe rozumowanie i refleksję. Takie podejście pozwala na tworzenie i wykorzystanie modeli specjalizowanych dla wybranych kroków pełnego przetwarzania w systemach wykorzystujących agentów LLM-owych.

W kolejnym **rozdziale 6.**, Doktorant skupia się na ocenie i diagnostyce systemów NLU z wykorzystaniem algorytmów probingowych, wykazując że takie podejścia mogą być niewystarczające do

wszechstronnej oceny systemów NLU w praktycznych wdrożeniach, zwłaszcza w kontekście zrozumienia i wydobycia informacji, która została zaszyta w reprezentacjach wewnętrznych analizowanego modelu. Autor jasno motywuje istotność podjęcia badań, które miały na celu opracowanie procedur oceny gotowości takich systemów do ich bezpiecznego, praktycznego wdrożenia.

Rozdział 7. skupia się na praktycznych problemach, które są bezpośrednio związane z komercyjnym wdrożeniem i oceną systemów NLU. Autor przedstawił podejścia, które mogą służyć do diagnostyki interfejsów, przez które systemy NLU mogą być manipulowane, sterowane i potencjalnie zakłócanie. Doktorant przeanalizował sterowalność modeli (zamierzoną i szkodliwą) z wykorzystaniem manipulacji promptem, wydobywanie ukrytych danych i informacji z wag LLM-ów oraz odporność modeli wielomodalnych. Mgr Jakub Hościłowicz opracował szereg metod, które pozwalają na wydobywanie ukrytych informacji z LLM-ów (*Unconditional Token Forcing* oraz *Unconditional Token Forcing Confusion*; obie metody zostały upublicznione na platformie **GitHub**) czy weryfikację odporności wielomodalnych LLM-ów (*Adversarial Confusion Attack*). Eksperymenty wykazały skuteczność opracowanych technik.

Rozdział 8. zawiera podsumowanie rozprawy – Doktorant odnosi się do postawionych hipotez badawczych i je omawia. W rozdziale tym przedstawione zostały także kierunki dalszych badań, które – według Autora – powinny być dalej eksplorowane.

Opracowane metody oraz otrzymane wyniki eksperymentalne przedstawione w rozprawie pozwalają mi z pełnym przekonaniem stwierdzić, że Doktorant jest świetnie przygotowana do tego, żeby prowadzić dalsze badania związane z tworzeniem, oceną i rzetelną weryfikacją systemów NLU. Rozprawa doktorska mgra Jakuba Hościłowicza, a także artykuły i patenty, których jest współautorem prezentują dojrzałość badawczą oraz wiedzę teoretyczną i praktyczną Doktoranta (oraz Jego świadomość dotyczącą aspektów związanych z praktycznym wdrożeniem systemów NLU) w dyscyplinie *Informatyka Techniczna i Telekomunikacja*. Stronę merytoryczną rozprawy oceniam wysoko.

2.1 Pytania i uwagi

W trakcie lektury rozprawy nasunęły mi się następujące pytania i uwagi:

1. Na str. 24 czytamy, że „*In the reported generation run, 82% of utterances passed the tag-count check after the first translation attempt, while a further 10% required two or three attempts.*” – czy to oznacza, że w 8% przypadków potrzebne były co najmniej 4 próby?
2. W niektórych miejscach pracy, dotyczących opracowanych metod i algorytmów, Autor mógłby – zamiast odnieść do literatury – umieścić dokładniejszy opis wykorzystanych metod oraz architektur sieci neuronowych (np. SLU na str. 24, czy BiLSTM na str. 25). Pozwoliłoby to na ułatwienie czytelnikowi pełnego zrozumienia rozprawy bez konieczności analizy dodatkowych artykułów. Podobną uwagę można sformułować w przypadku metryk wykorzystywanych w pracy, np. MC1 i MC2 (str. 35), które można byłoby formalnie zdefiniować w rozprawie.

3. Na str. 25, Autor podkreśla, że „*In the presented experiments, separate models are trained for each target language, as a single mBERT model trained jointly on all eight translated languages performed substantially worse.*” – co w tym wypadku dokładnie oznacza „*substantially worse*”?
4. Jak dokładnie wyglądała architektura MLP użyta w metodzie NL-ITI (str. 34)? Czy ta architektura była w jakiś sposób (pół)automatycznie optymalizowana, czy została ręcznie zaprojektowana przez Doktoranta?
5. Czy, według Doktoranta, dałoby się automatycznie zoptymalizować wartości hiperparametrów τ i ρ w metodzie NL-ITI? Jeśli tak, to w jaki sposób? Analiza czułościowa przedstawiona na str. 36 wykazała, że wybór wartości tych hiperparametrów ma istotny wpływ na jakość działania systemu. Czy według Autora dałoby się ją jeszcze bardziej poprawić optymalizując te hiperparametry?
6. W rozprawie Autor wspomina o zjawisku *catastrophic forgetting*, np. na str. 44. Obserwacje Doktoranta są generalnie prawidłowe, ale czy Autor próbował przeprowadzić eksperymenty, które jednoznacznie wskazałyby, że to zjawisko występuje w praktyce?
7. Na str. 79, Doktorant podkreśla, że „*An assistant may achieve strong benchmark results while still being vulnerable to prompt-level manipulation, hidden trigger-payload behavior, or adversarial visual disruption.*” – warto byłoby to uzasadnić również ilościowo (zwłaszcza w kontekście wspomnianych wyników benchmarkowych).

Powyższe uwagi i komentarze nie wpływają negatywnie na moją bardzo wysoką ocenę jakości pracy.

3 Ocena strony formalnej rozprawy doktorskiej

3.1 Ocena układu pracy

Układ rozprawy jest logiczny, a kolejne rozdziały wprowadzają czytelnika w najistotniejsze aspekty prac prowadzonych przez Doktoranta.

3.2 Ocena zastosowanego piśmiennictwa

Bibliografia jest bardzo dobrze skomponowana, zawiera 167 pozycji i nie pozostawia wątpliwości, że Autor świetnie orientuje się w tematyce rozprawy. Zauważyłem tylko kilka drobnych uchybień:

- Niektóre wpisy w bibliografii są niepełne (brakuje stron, wydawcy, itp.), np. w [13], [19], [33].
- Nazwy konferencji i czasopism są czasami pisane niespójnie (tj. w wersji pełnej i skróconej), np. [25], [16], [131], [135], itd.
- W niektórych wpisach w bibliografii wykorzystanie wielkich/małych liter jest niepoprawne (np. w [49]).

3.3 Uwagi redakcyjne

Rozprawa jest napisana niezwykle starannie, a tekst jest poprawny pod względem językowym, stylistycznym i interpunkcyjnym. W trakcie lektury zauważyłem bardzo drobne uchybienia:

- W tekście nie znalazłem odniesienia do rysunku 1.1.
- Unikałbym akapitów, które zawierają tylko jedno zdanie (np. na str. 7).
- Warto byłoby uspoźnić strategię pogrubiania najlepszych wyników w przedstawianych tabelach (np. w Tabeli 4.4 można byłoby pogrubić najlepsze wyniki dla każdego analizowanego języka).
- Na osi X na rys. 5.5, zamiast liczby parametrów, lepiej byłoby przedstawić nazwę konkretnego modelu – mogłoby to ułatwić analizę tego rysunku.

4 Konkluzja

Z pełnym przekonaniem stwierdzam, że przedmiotem rozprawy doktorskiej jest oryginalne rozwiązanie poprawnie zdefiniowanego problemu naukowego, a recenzowana dysertacja Pana mgr. Jakuba Hościłowicza **spełnia** wymagania stawiane rozprawom doktorskim przez Ustawę *Prawo o szkolnictwie wyższym i nauce*, Dz. U. 2023, poz. 742 z późn. zm. – praca prezentuje ogólną wiedzę teoretyczną i praktyczną Doktoranta w dyscyplinie *Informatyka Techniczna i Telekomunikacja* oraz umiejętność samodzielnego prowadzenia pracy naukowej.

W związku z powyższym, **wnioskuje o przyjęcie rozprawy doktorskiej oraz o dopuszczenie mgr. Jakuba Hościłowicza do publicznej obrony.**

Jakub Nalepa