Streszczenie

Systemy rozumienia języka naturalnego (RJN) są kluczowym komponentem asystentów głosowych i powinny działać niezawodnie w warunkach rzeczywistego użycia, gdzie nawet niewielkie błędy mogą prowadzić do niepoprawnego wykonania komendy użytkownika. Niniejsza rozprawa analizuje praktyczne metody ulepszania takich systemów oraz ich automatycznej diagnostyki ukierunkowanej na problemy istotne z perspektywy produkcyjnej. Główną praktyczną obserwacją rozprawy jest to, że jakość RJN często może zostać znacząco poprawiona poprzez identyfikację dominujących problemów i zastosowanie wyspecjalizowanych metod ukierunkowanych na ich rozwiązanie. Rozprawa powstała w ramach doktoratu wdrożeniowego realizowanego w Samsung R&D Institute Poland i odzwierciedla ewolucję komercyjnych systemów RJN w latach 2022–2026: od klasycznych systemów SLU (ang. *spoken language understanding*), przez asystenty oparte na modelach LLM (ang. *large language models*), aż po agenty GUI.

W przypadku SLU jednym z głównych problemów wdrożeniowych jest wysoki koszt rozszerzania systemu o nowe języki. Rozprawa przedstawia procedurę *translate–train*, w której relatywnie mały model LLM został zaadaptowany do generowania zaanotowanych danych treningowych dla nowych języków. Kluczowym wnioskiem jest to, że dzięki translacji maszynowej koszty ręcznego tworzenia danych mogą zostać znacząco zredukowane. Na publicznym benchmarku MultiATIS++ metoda poprawia średnią skuteczność o 7,1 punktu procentowego względem najlepszej metody bazowej oraz uzyskuje wynik slot F1 na poziomie 89,68% dla sześciu docelowych języków.

Rozprawa pokazuje również, że znaczące poprawy jakości systemów RJN można uzyskać bez kosztownego dotrenowywania modelu bazowego. W kontekście asystentów opartych na modelach LLM rozprawa przedstawia metodę Non-Linear Inference-Time Intervention (NL-ITI), która modyfikuje wybrane reprezentacje wewnętrzne, aby zwiększyć prawdziwość odpowiedzi zwracanych przez model. Na benchmarku TruthfulQA NL-ITI poprawia wynik głównej metryki MC1 z 36,35% do 50,19%, wyraźnie przewyższając bazową metodę ITI. Dla multimodalnych agentów GUI rozprawa wprowadza architekturę *ClickAgent*, która oddziela wysokopoziomowe rozumowanie i refleksję od precyzyjnej lokalizacji UI. Na benchmarku AITW ClickAgent osiąga 73,5% skuteczności realizacji zadań. Wyniki te pokazują, że ulepszanie systemów RJN może być często realizowane efektywniej poprzez ukierunkowaną interwencję w reprezentacje wewnętrzne lub modularną przebudowę systemu niż przez kosztowne i ryzykowne dotrenowywanie modelu bazowego.

Rozprawa pokazuje także, że sama skuteczność zadaniowa nie wystarcza do oceny gotowości systemów RJN do komercyjnego wdrożenia. Taka ocena powinna być uzupełniona diagnostyką skupioną na problemach, które mogą wystąpić w środowisku produkcyjnym. W pracy wskazano również istotne ograniczenie klasyfikatorów probingowych: nie uwzględniają one tego, jak łatwo wydobyć informację z reprezentacji wewnętrznych modelu, dla-

tego same w sobie nie są wystarczające do oceny systemów RJN. W rozprawie odniesiono się do tych ograniczeń, proponując metody diagnostyczne interfejsów, przez które systemy RJN mogą być sterowane, manipulowane lub zakłócane. W kontekście interfejsu wejściowego modelu rozprawa definiuje autoryzowaną i nieautoryzowaną sterowalność oraz pokazuje, że obecne modele LLM mogą być silnie ukierunkowane w stronę niepożądanych zachowań za pomocą podstawowych manipulacji promptem. Rozprawa omawia także możliwe zabezpieczenia, wskazując wymazywanie umiejętności i wiedzy modelu jako obiecujący kierunek. W części dotyczącej interfejsu dekodowania rozprawa wprowadza metodę Unconditional Token Forcing (UTF), która pozwala na wydobywanie ukrytych danych, takich jak znaczniki identyfikacyjne, ukryte wiadomości lub wyzwalacze behawioralne, z wag modelu LLM. Przedstawia także wariant defensywny, Unconditional Token Forcing Confusion (UTFC), dla którego jak dotąd nie zidentyfikowano skutecznej metody ekstrakcji. Dla interfejsu wizualnego rozprawa wprowadza metodę Adversarial Confusion Attack (ACA), która testuje odporność multimodalnych modeli LLM przez wprowadzenie ich w stan wysokiej niepewności dekodowania, prowadzący do halucynacji odpowiedzi. W rozprawie pokazano, że ta metoda ataku jest skuteczna zarówno w przypadku modeli open-source, jak i komercyjnych, takich jak GPT-5. Może być także wykorzystana w scenariuszu adversarial CAPTCHA jako mechanizm blokujący działanie agentów GUI.

Słowa kluczowe: rozumienie języka naturalnego, duże modele językowe, agenty GUI, transfer wielojęzyczny, multimodalne duże modele językowe, inżynieria reprezentacji, probing, ataki adwersarialne, ocena odporności