

Recenzja rozprawy doktorskiej**Katarzyny Krasnowskiej-Kieraś**

zatytułowanej:

**Automatyczne tworzenie składnikowo-zależnościowych rozbiorów wypowiedzeń z
wykorzystaniem głębokich sieci neuronowych****1. Problem badawczy i jego znaczenie**

Praca doktorska pt. "Automatyczne tworzenie składnikowo-zależnościowych rozbiorów wypowiedzeń z wykorzystaniem głębokich sieci neuronowych" porusza kluczowe zagadnienie w dziedzinie przetwarzania języka naturalnego (NLP): automatyczną analizę składniową (parsowanie). Problem ten polega na przypisaniu tekstowi formalnej reprezentacji opisującej jego budowę gramatyczną, co jest niezbędne lub przydatne w różnych zaawansowanych zastosowaniach.

Jaki jest najważniejszy problem rozważany w rozprawie?

Najważniejszym problemem badawczym jest opracowanie nowej metody automatycznej analizy składniowej, która systematycznie i spójnie łączy dwa powszechne formalizmy reprezentowania rozbioru – drzewa składnikowe i zależnościowe – w jedno, hybrydowe drzewo składnikowo-zależnościowe. Kluczowa jest propozycja nowego algorytmu, który działa w oparciu o "kręgosłupy składniowe" traktowane jako znaczniki przypisywane segmentom. Budowa kompletnej struktury hybrydowej następuje przy wykorzystaniu grafowej metody parsowania zależnościowego. Co istotne, praca weryfikuje możliwość produkowania rozbiorów uwzględniających składniki nieciągłe.

Czy ma on charakter naukowy?

Problem ma bezsprzecznie charakter naukowy. Nowy algorytm parsowania hybrydowego, łączący formalizmy składnikowe i zależnościowe w sposób spójny i systematyczny, stanowi istotne teoretyczne ujęcie dość wąsko zbadanego obszaru, zwłaszcza w kontekście przetwarzania języka polskiego. Ponadto, praca bada zaawansowane architektonicznie rozwiązania, takie jak zastosowanie głębokiej sieci

neuronowej z pretrenowanym modelem językowym typu BERT (np. HerBERT), co wpisuje się w najnowsze nurty badawcze NLP. Autorka wykazuje, że ten wyspecjalizowany model, dostrojony na specjalistycznych danych, jest skuteczniejszy lub zbliżony (w wypadku COMBO) w parsowaniu niż inne parsery dla języka polskiego. Według mojej wiedzy lista porównawcza jest wyczerpująca i rzetelnie sporządzona.

Czy ma on znaczenie praktyczne?

Zagadnienie posiada wymiar praktyczny, aczkolwiek ograniczony przez olbrzymią skuteczność i stosowalność Dużych Modeli Językowych w ostatnich 2-3 latach. Wysoka jakość automatycznej analizy składniowej osiągnięta przez parser Hydra (np. 96% UAS i 90.7% LAS dla polskiej struktury zależnościowej oraz 97.6% F1 dla składników) może być przydatna dla wybranych zastosowań. Ponadto, Autorka potwierdza praktyczne zastosowania Hydry, w tym automatyczną anotację dużego korpusu języka polskiego (KWJP), wsparcie rozbudowy ręcznie znakowanego banku drzew (Składnica) oraz wykrywanie neologizmów w systemie NeoN.

2. Wkład autora

Autorka niniejszej pracy doktorskiej wniosła znaczący wkład w rozwój metod automatycznej analizy składniowej, opierając się na głębokich sieciach neuronowych.

1. Opracowanie nowego algorytmu i architektury. Kluczowym wkładem jest opracowanie i wdrożenie nowej techniki parsowania do postaci hybrydowej, opartej na kręgosłupach składniowych. System Hydra, zaimplementowany jako model statystyczny w postaci głębokiej sieci neuronowej, osiąga stan wiedzy (*state-of-the-art*) w zakresie automatycznej analizy składniowej współczesnego języka polskiego.
2. Multitasking i elastyczność. Zaprojektowana architektura umożliwia jednoczesną realizację wielu zadań, takich jak lematyzacja, analiza fleksyjna oraz rozpoznawanie jednostek nazewniczych, z wykorzystaniem wspólnego modelu BERT.
3. Wydajność i skuteczność. Autorka porównała Hydę z innymi systemami, takimi jak COMBO oraz Stanza i Berkeley Neural Parser (BeNePar). Wykazała, że Hydra przewyższa konkurencyjne systemy w parsowaniu zależnościowym, a w parsowaniu składnikowym osiąga wyniki porównywalne z najlepszymi. Ponadto, badania udowodniły, że Hydra jest szybsza w treningu niż systemy oparte na architekturze potokowej, jak Stanza.
4. Analiza złożonych fenomenów. Wkładem jest efektywne włączenie do analizy zagadnienia składników nieciągłych, co jest szczególnie istotne w kontekście przetwarzania tekstów historycznych, gdzie zagęszczenie takich konstrukcji jest kilkakrotnie większe niż w materiale współczesnym.

3. Poprawność

Czy stwierdzenia zawarte w rozprawie są godne zaufania?

Badania eksperymentalne zostały przeprowadzone z dużą rzetelnością. Autorka stosuje **powszechnie uznane miary ewaluacji** w NLP (UAS, LAS, F1) oraz porównuje wyniki Hydry z wynikami osiąganymi przez czołowe, dedykowane systemy (np. COMBO, Stanza, Trankit, BeNePar). Skuteczność modeli jest potwierdzona na zróżnicowanych korpusach polskojęzycznych (Zależnica, NKJP 1M, dane historyczne, PolEval) oraz na danych niemieckojęzycznych (TIGER).

Czy uzasadnienia są poprawne?

Uzasadnienia są poprawne i poparte szczegółową analizą, która wykracza poza samą prezentację wyników końcowych. Autorka szczegółowo uzasadnia wybór architektury grafowego klasyfikatora zależnościowego oraz przyjęcie atomowej reprezentacji kręgosłupów składniowych. Przedstawia również warianty alternatywne i wyniki eksperymentów porównawczych, np. analizując wpływ różnych modeli językowych (HerBERT, XLM-RoBERTa, plT5) na jakość parsowania oraz konieczność ważenia danych treningowych pochodzących z niepełnych źródeł. Uzasadnienia te potwierdzają naukową skrupulatność i solidność podjętych decyzji projektowych.

4. Wiedza kandydata

Autorka wykazała się zaawansowaną i kompleksową wiedzą w zakresie nowoczesnego NLP, zwłaszcza w obszarze parsowania i głębokiego uczenia.

1. Głęboka znajomość formalizmów. Autorka sprawnie porusza się w zakresie reprezentacji składniowej (składnikowej, zależnościowej i hybrydowej), w tym specyfiki problemów z nieciągłością w obu tradycyjnych formalizmach.
2. Znajomość architektur neuronowych. Wykorzystanie i modyfikacja architektury BERT/Transformer oraz zastosowanie grafowych parserów świadczą o doskonałej orientacji w architekturach State-Of-The-Art.
3. Multitasking i NLP. Praca demonstruje szczegółową znajomość zasobów i specyfiki języka polskiego, w tym tagsetów fleksyjnych i słowników, a także problemów związanych z przetwarzaniem danych historycznych.

5. Podsumowanie

Rozprawa doktorska Katarzyny Krasnowskiej-Kieraś pt. "Automatyczne tworzenie składnikowo-zależnościowych rozbiorów wypowiedzi z wykorzystaniem głębokich sieci neuronowych" stanowi istotne i oryginalne osiągnięcie naukowe. Opracowana metoda parsowania hybrydowego, zaimplementowana w narzędziu Hydra, rozwiązuje problem braku systematycznej i spójnej analizy składnikowo-zależnościowej dla języka polskiego, jednocześnie osiągając stan wiedzy w zakresie jakości predykcji. Wyniki są

rzetelnie zweryfikowane, a dowiedzione praktyczne zastosowania potwierdzają znaczenie pracy.

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez Prawo o szkolnictwie wyższym i nauce (Dz.U.2023 poz. 742), moja ocena rozprawy pod względem podstawowych kryteriów jest następująca:

- rozprawa zawiera oryginalne rozwiązanie problemu naukowego, które może mieć zastosowanie w sferze gospodarczej i społecznej.
- Kandydatka posiada ogólną wiedzę teoretyczną w dyscyplinie.
- Kandydatka posiada umiejętność samodzielnego prowadzenia pracy naukowej.

Podsumowując, praca mieści się w dyscyplinie Informatyka Techniczna i Telekomunikacja i spełnia w dostatecznym stopniu formalne i zwyczajowe wymagania stawiane rozprawom doktorskim. Wnoszę o dopuszczenie Doktorantki do dalszych faz przewodu doktorskiego.



Dr hab. inż. Bartosz Ziółko