Universal Discourse

1. Czego dotyczy projekt?

W szkole uczymy się, jak słowa i wyrażenia są ze sobą powiązane w zdaniu. Ale czy poszczególne zdania w tekście nie są ze sobą powiązane w podobny sposób? Tak właśnie jest – możemy nawet nazwać niektóre typy takich powiązań (czyli relacji): jedno zdanie może być parafrazą drugiego, stanowić jego wyjaśnienie, zawierać porównanie. Nazywamy takie relacje „dyskursywnymi" (lub po prostu relacjami dyskursu), ponieważ sprawiają one, że cały dyskurs (czyli tekst) jest spójny i zrozumiały. Relacje tego typu często łatwo zidentyfikować analizując spójniki lub wyrażenia łączące zdania, choć sprawa nie zawsze jest prosta. Jeden spójnik może być użyty w wielu znaczeniach, czasem może go w ogóle nie być.

Od wielu lat naukowcy badają takie relacje w różnych językach i oznaczają je w tekstach. Ponieważ jednak istnieją różne sposoby interpretowania fraz łączących i nazywania relacji lub ról poszczególnych zdań wchodzących w relacje, tworzone w ten sposób zbiory danych nie są ze sobą zgodne. Brak dobrych danych jest prawdopodobnie powodem, dla którego duże modele językowe, takie jak ChatGPT, wciąż mają problemy z wykrywaniem i nazywaniem relacji dyskursywnych w analizowanych tekstach.

Nasz projekt ma na celu znalezienie wspólnego mianownika dla istniejących zbiorów tekstów opisanych relacjami dyskursywnymi poprzez wykorzystanie modelu zaproponowanego przez Międzynarodową Organizację Normalizacyjną (ISO) do ponownej interpretacji 100 000 relacji tego typu i stworzenia w ten sposób standardowego zestawu danych do dalszych badań. Nazwaliśmy tę inicjatywę Universal Discourse *(Uniwersalny Dyskurs)*, ponieważ ma ona na celu połączenie różnych teorii we wspólny model metodą wielostronnej analizy zbiorów tekstów w różnych językach.

2. Jak będzie przebiegać praca?

Wspólny model relacji dyskursu zostanie stworzony w oparciu o standard ISO i przy użyciu przykładów zaczerpniętych z istniejących zbiorów danych i literatury. Następnie wybrane zbiory danych zostaną przekonwertowane na nowy model w trzech etapach: konwersji automatycznej, ręcznej korekty danych przez lingwistów i uzgodnieniu powstałych interpretacji przez doświadczonego lingwistę.

Powstały zbiór danych zostanie wykorzystany do stworzenia parserów dyskursu – programów do automatycznego wykrywania relacji dyskursywnych w tekstach. Obecnie takie programy są konstruowane przy użyciu metod sztucznej inteligencji, poprzez dostosowanie zewnętrznych modeli językowych z użyciem nowych danych wzorcowych dla realizowanego zadania.

Wszystkie wyniki projektu zostaną opisane w monografii opublikowanej pod koniec 4-letniego okresu trwania projektu. Wyniki częściowe będą prezentowane innym badaczom na najważniejszych konferencjach naukowych z dziedziny językoznawstwa i przetwarzania języka naturalnego.

Projekt będzie miał charakter interdyscyplinarny; wezmą w nim udział badacze o wykształceniu lingwistycznym i komputerowym, zarówno początkujący, jak i doświadczeni naukowcy (profesorowie, doktorzy, eksperci zewnętrzni i doktoranci) posiadający uzupełniające się umiejętności.

3. Jakie korzyści przyniesie projekt społeczności naukowej i społeczeństwu?

Nowy, ujednolicony model relacji dyskursywnych rozpocznie badania na temat powiązań między różnymi istniejącymi schematami reprezentacji dyskursu. Powstały zbiór danych udostępni językoznawcom nowe możliwości analityczne, a dla informatyków będzie nowym źródłem cennych danych gotowych do wykorzystania w procesie opracowania narzędzi i modeli językowych.

Użycie parsera dyskursu umożliwi analizę powiązań między częściami tekstu, np. argumentów w dyskusji lub wykrycie najbardziej istotnych części tekstu, które powinny znaleźć się w automatycznie wygenerowanym streszczeniu. Tego rodzaju funkcje będą mogły znaleźć się w aplikacjach dostępnych całemu społeczeństwu.