Mózg vs Komputer
w kontekście tłumaczenia tekstu naturalnego

Jan Siemieniec s22596

Tłumaczenie tekstów między językami to potrzeba znana ludzkości od wieków. Aby sprostać tym wyzwaniom, wykształcił się zawód tłumacza, który umożliwia przekład treści z jednego języka na inny. Wraz z rozwojem technologii pojawia się jednak pytanie: jak dobrze z tym zadaniem radzą sobie komputery? Choć maszyny potrafią przetwarzać ogromne ilości danych w krótkim czasie, tłumaczenie języka naturalnego pozostaje szczególnym wyzwaniem. Wymaga ono nie tylko znajomości słów i struktur gramatycznych, lecz także zrozumienia kontekstu, intencji oraz tego, jak ludzie postrzegają świat. W niniejszym raporcie przeprowadzono analizę mającą na celu odpowiedź na pytanie: czy komputery są dziś w stanie skutecznie tłumaczyć teksty, czy też wciąż pozostaje w tym zakresie istotne pole do poprawy?

Jednak ocena tłumaczenia nie jest zadaniem prostym ani jednoznacznym. Rozważmy następujący przykład:

Zdanie do przetłumaczenia**: *I need to go to the doctor***.
Tłumaczenie 1**: *Muszę iść do lekarza***.
Tłumaczenie 2: ***Muszę udać się do doktora***.
Tłumaczenie 3: ***Muszę udać się do lekarza***.

Już w tak prostym przypadku widać, że możliwych tłumaczeń jest kilka — wszystkie poprawne, ale różniące się stylem lub doborem słów. Jest to spowodowane tym, że człowiek ocenia, które z tłumaczeń jest według niego najdokładniejsze w sposób indywidualny, uzależniony od preferencji danej osoby. Zatem pytanie brzmi, jak można uogólnić ten problem i zbudować automatyczną miarę jakości tłumaczenia?

Jednym z intuicyjnych podejść do oceny jakości tłumaczeń generowanych przez komputer jest bezpośrednie porównanie dwóch wersji — jednej stworzonej przez

człowieka, a drugiej przez maszynę. Jednak taka metoda, choć pozornie prosta, okazuje się niewystarczająca. Rozważmy dwa zdania:

Zdanie 1: *Przedmiot blokowy na Polsko Japońskiej akademii technik komputerowych o nazwie „Neuroplastyczność a neurodegeneracja: Mózg a komputer" prowadzony przez Pana Andrzeja Przybyszewskiego był prowadzony w godzinach* ***popołudniowych***.

Zdanie 2: *Przedmiot blokowy na Polsko Japońskiej akademii technik komputerowych o nazwie „Neuroplastyczność a neurodegeneracja: Mózg a komputer" prowadzony przez Pana Andrzeja Przybyszewskiego był prowadzony w godzinach* ***porannych***.

Choć oba zdania są niemal identyczne — 25 z 26 słów (czyli około 96%) pokrywa się dosłownie — różnią się kluczową informacją. Zmiana jednej tylko frazy całkowicie odwraca znaczenie wypowiedzi. Pokazuje to, że podobieństwo leksykalne (czyli zgodność słów) nie zawsze przekłada się na podobieństwo semantyczne (czyli zgodność znaczenia). Z tego powodu proste metody porównywania tekstów, oparte jedynie na liczbie wspólnych słów czy strukturze zdań, mogą prowadzić do błędnych wniosków przy ocenie tłumaczeń.

Dlatego też w niniejszej analizie skupiono się na podobieństwie semantycznym zdań. W takim przypadku nie patrzy się na podobieństwo powstałych tłumaczeń, ale na informacje, które zawierają. W tym celu wykorzystałem bibliotekę **spaCy** wraz z metodą porównywania cosinusowego między przetłumaczonymi zdaniami. Jako dane wejściowe użyłem zbiór prawie 2000 przetłumaczonych zdań z języka angielskiego na polski przez komputer i przez profesjonalnych tłumaczy. Dodatkowo, zdania przetłumaczone przez komputer zostały ocenione przez profesjonalnych tłumaczy w skali od 1 do 5, gdzie:

5 – tłumaczenie jest bardzo dobre językowo i zawiera całą treść oryginalnego zdania, bez dodawania czegokolwiek;

4,5 – ocena pomiędzy 4 a 5;

4 – tłumaczenie zawiera drobne błędy;

3,5 – ocena na pograniczu 3 i 4;

3 – tłumaczenie zawiera poważne błędy, ale może być użyte w praktyce;

2,5 – ocena na pograniczu 2 i 3;

2 – błędy uniemożliwiają użycie tłumaczenia, ale zawiera ono dobrze przetłumaczone fragmenty;

1,5 – ocena pomiędzy 1 a 2;

1 – tłumaczenie jest całkowicie błędne.

Dane wejściowe były zapisane w pliku tekstowym, który następnie załadowałem do obiektu *DataFrame* w bibliotece *pandas*. Struktura danych obejmowała następujące kolumny:

ML – tekst przetłumaczony przez komputer

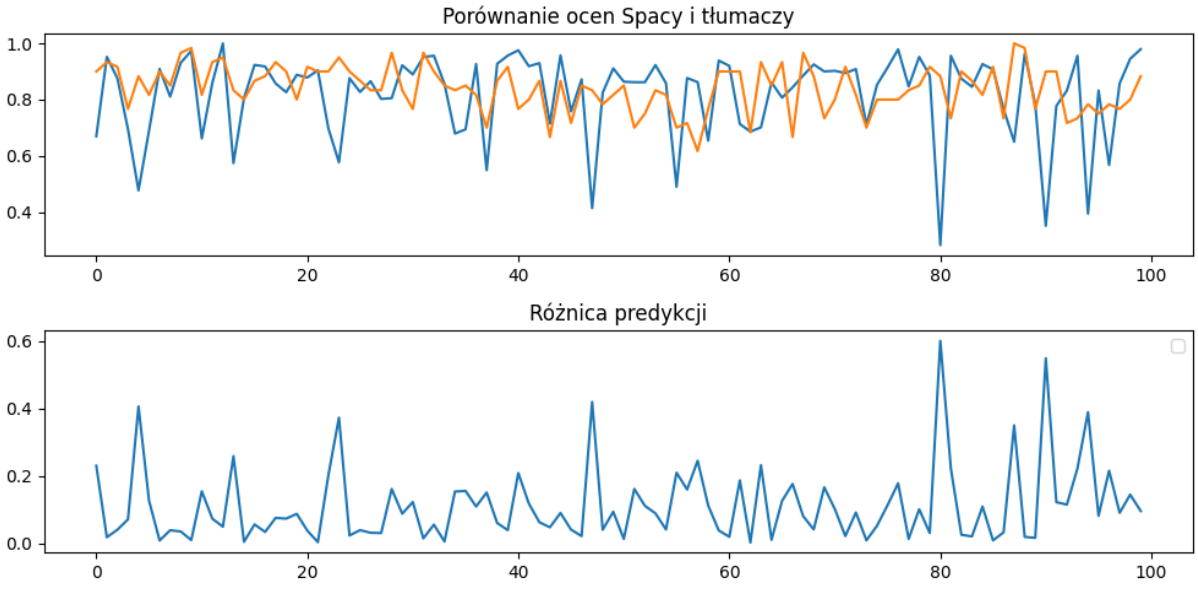
EN – oryginalne zdanie w języku angielskim

HUMAN – tekst przetłumaczony przez profesjonalnych tłumaczy

PREDICTION – ocena tłumaczy przetłumaczonego przez komputer zdania

Wszystkie użyte dane pochodzą z publicznie dostępnego repozytorium na GitHub: https://github.com/poleval/2021-quality-estimation-nonblind/tree/main/test-A

Po zastosowaniu metody similarity() z biblioteki spaCy, możliwe było porównanie podobieństwa semantycznego zdań z ocenami przyznanymi przez profesjonalnych tłumaczy. Na podstawie tych danych wygenerowano odpowiedni wykres:

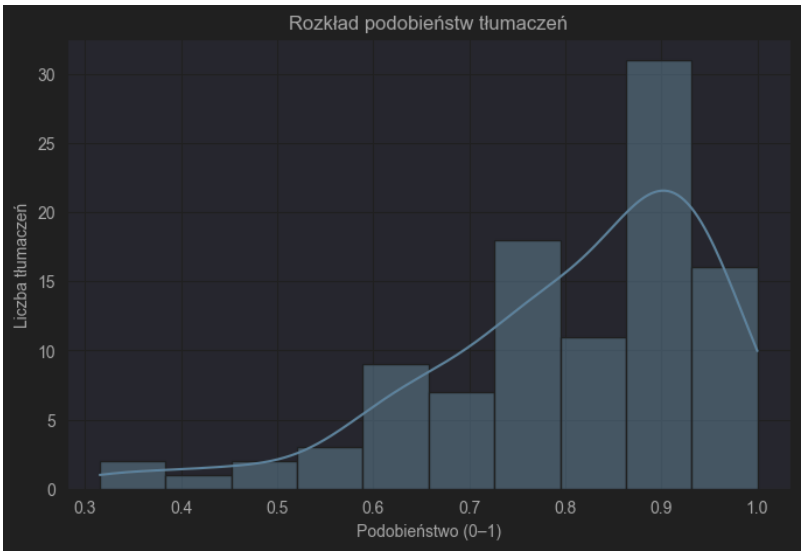


Na podstawie danych obliczono, że współczynnik korelacji wynosi 0,14, a odległość euklidesowa między oceną semantyczną spaCy, a ocenami tłumaczy to 1,6.

Podsumowując przeprowadzoną analizę, można stwierdzić, że komputery radzą sobie z tłumaczeniem tekstu znacznie lepiej, niż mogłoby się wydawać, zwłaszcza biorąc pod uwagę, że jest to zadanie silnie powiązane z ludzkim poznaniem, intuicją i doświadczeniem językowym. Pomimo że tłumaczenie wymaga zrozumienia znaczenia, kontekstu oraz niuansów językowych — elementów charakterystycznych dla ludzkiego mózgu — wyniki wskazują, że różnica między tłumaczeniami komputerowymi a ludzkimi wyniosła średnio zaledwie 15%. To niewielkie odchylenie sugeruje, że algorytmy potrafią skutecznie uchwycić sens zdań i tworzyć tłumaczenia, które pod względem semantycznym są bardzo zbliżone do tych wykonanych przez profesjonalnych tłumaczy. Choć pewne rozbieżności nadal się pojawiają, są one nieliczne i nie zmieniają ogólnego obrazu: komputery potrafią dziś tłumaczyć teksty z dużą trafnością.

W ostatnim czasie coraz większą popularnością cieszy się wykorzystywanie dużych modeli językowych do różnorodnych zadań, w tym również do tłumaczenia tekstów. W związku z tym postanowiłem sprawdzić, jak z tym zadaniem radzą sobie najnowsze i najpopularniejsze modele językowe, takie jak ChatGPT, Grok oraz Gemini. Przeprowadziłem analogiczną analizę na podstawie zestawu 100 pytań, zlecając każdemu z modeli wygenerowanie tłumaczenia. Poniżej przedstawiam uzyskane wyniki:

ChatGPT

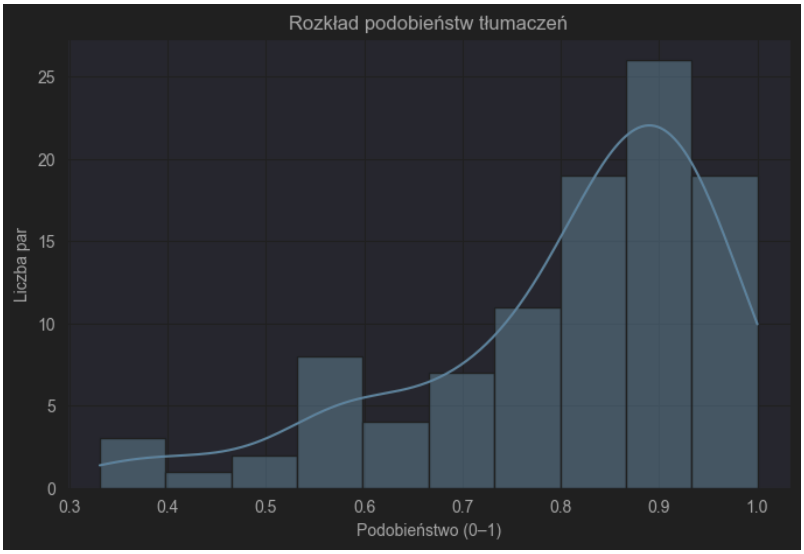


Średnie podobieństwo: 0.8065

Mediana podobieństwa: 0.8545

Odchylenie standardowe: 0.1458

Grok

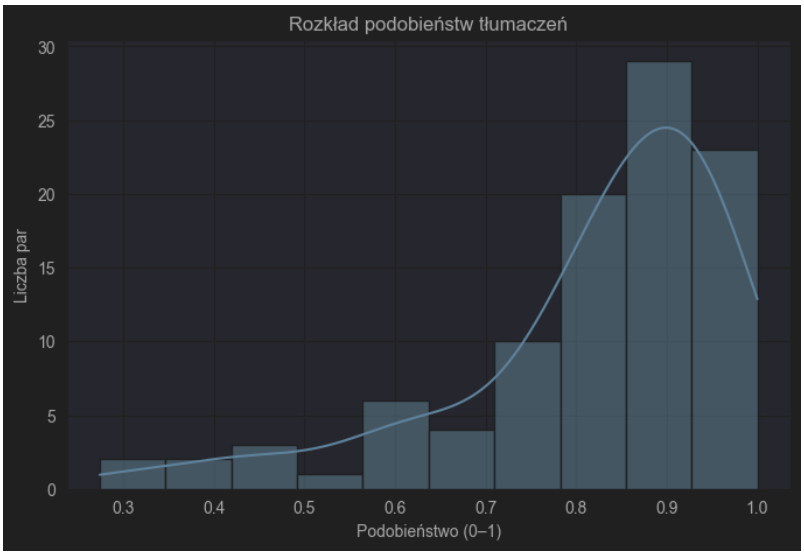


Średnie podobieństwo: 0.8060

Mediana podobieństwa: 0.8550

Odchylenie standardowe: 0.1540

Gemini



Średnie podobieństwo: 0.8202

Mediana podobieństwa: 0.8647

Odchylenie standardowe: 0.1572

Jak wynika z przeprowadzonych analiz, wszystkie trzy modele językowe osiągnęły zbliżone wyniki, z przeciętnym poziomem podobieństwa semantycznego w zakresie około **80%** względem tłumaczeń wykonanych przez profesjonalnych tłumaczy. Gemini nieznacznie wyprzedził pozostałe modele zarówno pod względem średniego, jak i medianowego podobieństwa. Warto zwrócić uwagę, że rozkłady podobieństwa dla wszystkich modeli wskazują najwyższą koncentrację wyników w okolicach **90%**, co oznacza, że znaczna część wygenerowanych tłumaczeń była niemal tożsama semantycznie z tłumaczeniami ludzkimi. Choć warto zauważyć, że odchylenie standardowe wskazuje na pewien rozrzut wyników, lecz jego wartość pozostaje relatywnie niska.

Przeprowadzone analizy pokazują, że komputery za pomocą zarówno klasycznych algorytmów, jak i przy użyciu dużych modeli językowych, potrafią tłumaczyć teksty z dużą skutecznością. Podsumowując, mimo że tłumaczenie wciąż jest zadaniem zakorzenionym w ludzkim mózgu, współczesne komputery nie tylko nadążają za człowiekiem, ale w wielu przypadkach są już w stanie go skutecznie wspierać, a czasem nawet wyręczać. Przyszłość tłumaczeń, podobnie jak wiele innych dziedzin, będzie prawdopodobnie coraz bardziej oparta na synergii między sztuczną inteligencją a ludzkim doświadczeniem.

Cały kod związany z analizami tego projektu znajduje się na publicznym repozytorium GitHub: https://github.com/s22596/MKR_Jan_Siemieniec_s22596 (główna logika projektu znajduje się w pliku *project.ipynb*).