CSI Translator

Nowoczesne metody tłumaczenia automatycznego

Mikołaj Pokrywka, Kamil Guttmann, Martyna Drumińska

Wprowadzenie

Nasz projekt badawczo-rozwojowy polega na rozwijaniu modelu tłumaczenia automatycznego na poziomie całego dokumentu. W celu zwiększenia jakości tłumaczeń proponujemy wykorzystanie metody quality-aware decoding oraz adaptacji specjalistycznego słownictwa za pomocą glosariuszy. Dzięki tym metodom jesteśmy w stanie poprawić jakość tłumaczeń automatycznych, uwzględniając szeroki kontekst.

Tłumaczenie na poziomie dokumentu

- Skupiamy się na rozwijaniu modelu tłumaczenia automatycznego, który uwzględnia kontekst dokumentu. Dotychczasowe modele tłumaczenia automatycznego opierały się głównie na kontekście pojedynczego zdania, co nie zawsze pozwalało na skuteczne rozwiązanie problemów dwuznacznych słów.
- Nasz model opiera się na poszerzeniu kontekstu o kilka/kilkanaście zdań, co pozwala na lepsze rozumienie tłumaczonych treści. Dzięki temu model jest w stanie lepiej radzić sobie z tłumaczeniem tekstów o złożonej strukturze gramatycznej i semantycznej, co przyczynia się do podwyższenia jakości tłumaczeń.

Zdanie źródłowe	Model na poziomie zdań	Model na poziomie dokumentu
This is John's favourite movie. He says it's very interesting and original	To ulubiony film Johna. Mówi, **że to** bardzo **ciekawe i oryginalne**	To jest ulubiony film Johna. Mówi, **że jest** bardzo **ciekawy** i **oryginalny**.

Quality-aware decoding

- Wykorzystujemy metodę quality-aware decoding, która dokonuje wyboru tłumaczenia o najwyższej jakości spośród kilku potencjalnych możliwości. Standardowo, decyzja o wyborze najbardziej prawdopodobnego tłumaczenia jest podejmowana przez model tłumaczenia. Badania wykazują jednak, że tłumaczenia wybrane przez metodę quality-aware decoding mają wyższą korelacje z oceną ludzką.
- Metoda quality-aware decoding polega na wyborze najlepszego tłumaczenia bazując na osobnym modelu estymacji jakości tłumaczenia. Dzięki tej metodzie jesteśmy w stanie wybrać tłumaczenie, które charakteryzuje się najlepszą jakością.

Zdanie źródłowe	Wygenerowane tłumaczenia
He did not feel fufilled until his first job.	Nie czuł się spełniony aż do rozpoczęcia swojej pierwszej pracy. Nie czuł się spełniony aż do rozpoczęcia swojej pierwszej pracy **Nie czuł się spełniony do momentu rozpoczęcia swojej pierwszej pracy.** Nie czuł się zaspokojony aż do początku swojej pierwszej pracy.

Glosariusze

- Stosujemy metodę umożliwiającą użycie glosariuszy w tłumaczeniu automatycznym. W wielu dziedzinach, takich jak medycyna czy technologie, używa się specjalistycznego słownictwa, którego model tłumaczenia może nie znać. Aby temu zaradzić, proponujemy zastosowanie metody opartej na wykorzystaniu tłumaczeń pochodzących z glosariuszy przygotowanych przez specjalistów.
- Proponowana metoda polega na dodaniu informacji o poprawnym tłumaczeniu danej frazy do wejścia modelu. Model jest w stanie samodzielnie odmienić tłumaczenie w odpowiedni sposób, tak aby pasowało do kontekstu. Stosując tę metodę, jesteśmy w stanie poprawić jakość tłumaczeń specjalistycznego słownictwa.

Glosariusz	
EN	PL
audit committee	komisja rewizyjna

Zdanie źródłowe	Model bez glosariuszy	Model z glosariuszami
The **audit committee** should be composed exclusively of non-executie or supervisory directors.	**Komitet ds. audytu** powinien składać się wyłącznie z dyrektorów niewykonawczych.	W skład **komisji rewizyjnej** powinni wchodzić wyłącznie dyrektorzy niewykonawczy.

Implementacje

Dalszy rozwój projektu

- Adaptacja dziedzinowa offline
- Adaptacja dziedzinowa w locie
- Generowanie syntetycznych korpusów

Przedmiot Projekt Badawczo-Rozwojowy realizowany w ramach projektu Akademia Innowacyjnych Zastosowań Technologii Cyfrowych
Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego Program Operacyjny Polska Cyfrowa na lata 2014-2020,
Oś Priorytetowa nr 3 "Cyfrowe kompetencje społeczeństwa" Działanie nr 3.2 "Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej"
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)” Porozumienie nr POPC.03.02.00-00-0001/20-00