Laboratorium 2-5

Przetwarzanie języka naturalnego

Zadanie 1: Budowa i ocena wzorcowego korpusu tekstów

Opracowanie: dr inż. Jan Kocoń, dr hab. inż. Maciej Piasecki (prof. PWr)

[Cel](#)

[Opis problemu](#)

[Realizacja (punkty)](#)
- [Wybór i opisanie problemu (5)](#)
  - [Problemy](#)
    - [Rozpoznawanie dziedzinowe wydźwięku w korpusie PolEmo 2.0](#)
    - [Wykrywanie cyberprzemocy w Tweetach](#)
    - [Ocena streszczeń artykułów prasowych](#)
    - [Ocena wydźwięku komentarzy z Allegro](#)
- [Ostateczny podział na zespoły robocze (5)](#)
- [Wybór i opisanie źródeł danych tekstowych (10)](#)
- [Budowa korpusu pełnego oraz wzorcowego (20)](#)
- [Ocena pod względem jakości i reprezentatywności (40)](#)
- [Ocena jakości anotacji danego korpusu (20)](#)
- [Oddanie zadania](#)

Cel

Zapoznanie się z metodami budowy i analizy korpusów tekstowych danych językowych, będących podstawowym źródłem wiedzy w lingwistyce informatycznej i inżynierii języka naturalnego.

Opis problemu

Korpus językowy jest zbiorem tekstów, który stanowi podstawowy zasób praktycznie w każdym zadaniu związanym z przetwarzaniem języka naturalnego. Przy pomocy wielkich korpusów można budować ogólne reprezentacje językowe, szukać kontekstów występowania słów, a także analizować specyficzne konstrukcje składniowe i semantyczne w danym języku (lub językach - w przypadku korpusu wielojęzycznego). Mniejsze korpusy często są także ręcznie wzbogacane o dodatkowe metadane, dzięki czemu możliwe jest wykorzystanie takich danych w uczeniu maszynowym do zadań klasyfikacji całych dokumentów, zdań, a także fraz, czy też pojedynczych słów. Elektroniczne wersje korpusów tekstowych służą do prowadzenia badań nad językiem, tworzenia słowników, wyszukiwarek, a także systemów tłumaczenia maszynowego, określania stylu literackiego, rozpoznawania wydźwięku i emocji, ujednoznaczniania znaczeń słów, wydobywania nazw własnych i wielu innych zastosowań.

Często wyzwaniem dla badaczy jest określenie składu korpusu, czyli takiego doboru tekstów, by były one reprezentatywne z perspektywy konkretnego problemu do rozwiązania. W kontekście budowania wielkich, ogólnych modeli językowych, istotne jest zebranie jak największej liczby tekstów z różnych źródeł, dziedzin oraz stylów wypowiedzi (książki, artykuły, blogi, komentarze, recenzje, itp.). Często jednak problemy są bardzo specyficzne i ukierunkowane dziedzinowo, np.:

- detekcja fałszywych wiadomości dotyczących COVID-19,
- rozpoznawanie wydźwięku opinii pacjentów o lekarzach,
- określanie autorstwa tekstów literackich,
- wydobywanie nazw hoteli w ofertach turystycznych,
- predykcja kursów giełdowych na podstawie tekstów ekonomicznych,
- budowa systemu dialogowego dla użytkowników operatorów telefonicznych,
- konstrukcja chatbota dla petentów Urzędu Miejskiego Wrocławia,
- utworzenie wirtualnej sekretarki lekarza pierwszego kontaktu,
- wykrywanie mowy obraźliwej w komentarzach na Twitterze,
- streszczanie artykułów prasowych.

W każdym z wyżej wymienionych przypadków należy utworzyć tzw. korpus reprezentatywny dla danej dziedziny. Modele językowe powstałe z wielkich i ogólnych korpusów językowych można stosować do tego typu zadań, lecz często wcześniej *dostraja się* je (ang. *fine-tuning*) na *korpusach wzorcowych*. Z założenia korpusy wzorcowe mają pomóc w realizacji zadania, gdyż zawierają teksty z konkretnej dziedziny, zawierają specyficzne dla danego problemu słownictwo, określone konstrukcje leksykalno-składniowo-semantyczne i dzięki temu spodziewamy się poprawy jakości działania metody dostrojonej na takim korpusie.

Realizacja (punkty)

Wybór i opisanie problemu (5)

W pierwszej kolejności należy wybrać problem, dla jakiego budowany będzie korpus wzorcowy, z listy problemów przygotowanych przez prowadzących kurs. Możliwa jest realizacja problemów spoza listy, jednak należy zgłosić temat i uzyskać akceptację przed zajęciami nr 3. Przy innym temacie należy też uzasadnić, że wpisuje się on w plan kursu określony we wprowadzeniu (Laboratorium 1). Lista problemów jest wspólna dla wszystkich grup zajęciowych, jednak możliwe jest wykluczenie części tematów przez prowadzącego. Zespoły w ramach grupy zajęciowej powinny realizować różne tematy, ewentualnie ten sam ale z uwzględnieniem innych podzbiorów danych. O przydziale tematu do zespołu decyduje kolejność zgłoszeń. Każdy wybór tematu powinien być uprzednio przedyskutowany z prowadzącym i zaakceptowany przez niego.

Uwaga! Proszę nie wybierać jako problemu rozpoznawania polaryzacji wydźwięku opinii, gdyż jest to wspólne zadanie dla wszystkich grup w ramach zadania 4 (Laboratorium 12-14).

Problemy

Wykrywanie cyberprzemocy w Tweetach
Dodatkowe informacje: https://github.com/ptaszynski/cyberbullying-Polish
Link: https://huggingface.co/datasets/poleval2019_cyberbullying
Publikacja: Ptaszynski, Michal, Agata Pieciukiewicz, and Paweł Dybała. "Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in Polish Twitter." (2019). [link]

Ocena streszczeń artykułów prasowych
Dodatkowe informacje: http://zil.ipipan.waw.pl/PolishSummariesCorpus
Link: https://klejbenchmark.com/static/data/klej_psc.zip
Publikacja: Ogrodniczuk, Maciej, and Mateusz Kopeć. "The Polish summaries corpus." Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 2014. [link]

Ocena wydźwięku komentarzy z Allegro
Dodatkowe informacje: https://github.com/allegro/klejbenchmark-allegroreviews
Link:
https://huggingface.co/datasets/allegro/summarization-polish-summaries-corpus
Publikacja: Rybak, Piotr, et al. "KLEJ: Comprehensive Benchmark for Polish Language Understanding." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. [link]

Efekt: opis problemu przez zespół roboczy.
Termin: Laboratorium 3

Ostateczny podział na zespoły robocze (5)

Wybór problemu powinien być powiązany z ostatecznym wyborem składu zespołu. Należy przedstawić ten skład prowadzącemu zajęcia. Każde z czterech oddawanych zadań w semestrze powinno zawierać krótką notatkę z informacją o podziale obowiązków w zespole, tj. kto realizował daną część zadania i w jakim zakresie.

Efekt: opis składu zespołu i wybranego tematu problemu.
Termin: Laboratorium 3

Wybór i opisanie źródeł danych tekstowych (10)

Kolejny krok związany jest z określeniem źródeł danych do budowy zarówno pełnego korpusu języka, jak i wzorcowego korpusu dla wybranego problemu. Przykładowym otwartym źródłem, zawierającym wiele tekstów, jest Wikipedia, której zrzuty dla poszczególnych języków można przekształcić do postaci tekstowej przy pomocy narzędzia WikiExtractor. Dostępne są także korpusy języka polskiego, takie jak milionowy podkorpus NKJP, Korpus Języka Polskiego Politechniki Wrocławskiej, czy też Oscar. W przypadku specyficznych problemów warto rozważyć pobieranie tekstów bezpośrednio z dedykowanych źródeł, np. komentarze z prasy, wpisy z Twittera, czy też opinie ze stron związanych z konkretną dziedziną.

Efekt: lista źródeł danych tekstowych wraz z krótkim opisem każdego źródła oraz przewidzianym zastosowaniem (pełny korpus języka, korpus wzorcowy) i uzasadnieniem wyboru.
Termin: Laboratorium 3

Budowa korpusu pełnego oraz wzorcowego (20)

Należy zgromadzić dane w ilości min. 1GB tekstu na potrzeby budowy pełnego korpusu języka oraz min. 100MB tekstu na potrzeby budowy wzorcowego korpusu dla danego problemu. Jeżeli pozyskanie korpusu wzorcowego z ogólnodostępnych danych nie jest możliwe, należy to uzasadnić, a następnie przy pomocy technik filtrowania wyodrębnić z pełnego korpusu te teksty, które powinny być reprezentatywne dla wybranego problemu. Filtrowanie wykonać z wykorzystaniem np. słów kluczowych, istotnych fraz, itp. Można także wykorzystać podobieństwo wektorowe tekstów z korpusu dla wybranego problemu do tekstów z pełnego korpusu i w ten sposób wyodrębnić korpus wzorcowy.

Efekt: przygotowanie 2 archiwów: 1) pełny korpus języka; 2) korpus wzorcowy.
Termin: Laboratorium 4

Ocena pod względem jakości i reprezentatywności (40)

Należy ocenić pełny korpus oraz korpus wzorcowy względem korpusu dla wybranego problemu, z wykorzystaniem technik omówionych na wykładzie.

Efekt: raport dotyczący oceny jakości i reprezentatywności korpusu pełnego i wzorcowego.
Termin: Laboratorium 5

Ocena jakości anotacji danego korpusu (20)

Jednym z bardzo istotnych zadań w procesie anotacji korpusu dla określonego problemu, jest ocena zgodności anotatorów. Dla dostarczonego korpusu PolEmo 1.0 [link] należy przeprowadzić ocenę zgodności anotatorów (anotatorzy: super, a, e, k), zarówno parami (z wykorzystaniem współczynnika Kappa Cohena) oraz łącznie (z wykorzystaniem współczynnika Wave Kappa lub Alfa Krippendorfa). W przypadku Kappa Cohena należy pominąć te przypadki, dla których nie istnieją dokładnie 2 anotacje dla wybranej pary anotatorów.

Efekt: raport podsumowujący analizę zgodności anotacji na danym korpusie
Termin: Laboratorium 5

Oddanie zadania

Proszę o przygotowanie paczki z raportem oraz kodem o nazwie X.zip, gdzie X jest numerem Państwa grupy. Proszę, by tylko jedna osoba z grupy wgrała plik. Proszę o przygotowanie linków do Google Drive, które umieszczą Państwo w raporcie do pobrania korpusów: wzorcowego oraz pełnego (umieszczonych na Drive w postaci paczek zip). Jeżeli obowiązuje Państwa NDA w kontekście danych, proszę ich nie wgrywać, a same dane zaprezentować na zajęciach.