Wartość danych dla rozwoju technologii języka na przykładzie CLARIN-PL

Tomasz Walkowiak

CLARIN-PL
Politechnika Wrocławska
tomasz.walkowiak@pwr.edu.pl

CLARIN

CLARIN-PL
Common Language Resources and Technology Infrastructure

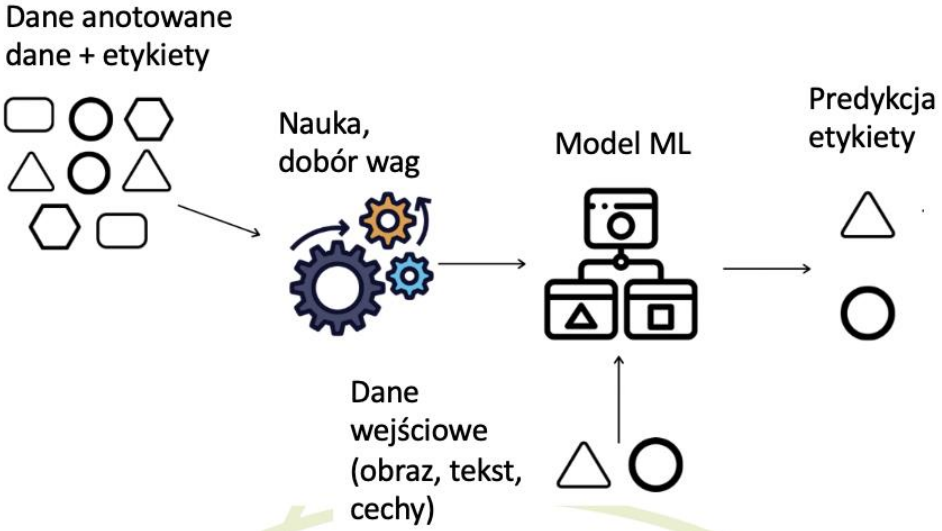
Wrocław University of Science and Technology

Dane językowe – podstawa działania CLARIN-PL

- wytwórca danych (zasobów) językowych
- repozytorium danych językowych
- informacja wejściowa infrastruktury CLARIN-PL
  - tekst, mowa => https://ws.clarin-pl.eu
  - np. wydobywanie informacji z tekstu: temat, emocje
- podstawa do budowy narzędzi językowych
  - w dużej części opartych o uczenie maszynowe
  - najprostsze zadanie: zliczanie słów w tekście
    - może wymagać sprowadzenia słów tekstu do postaci słownikowej (tzw. lematyzacji)
    - potrzebny jest tager morfo-syntaktyczny
      - tokenizacja, przypisanie tokenom części mowy i lematów
    - jak budujemy tager (uczenie maszynowe)
      - potrzebujemy danych językowych
        - tekst + meta-informacja (tokeny, lematy, części mowy)
        - zbiór treningowy, walidacyjny, testowy

Uczenie (maszynowe) na danych (anotowanych)

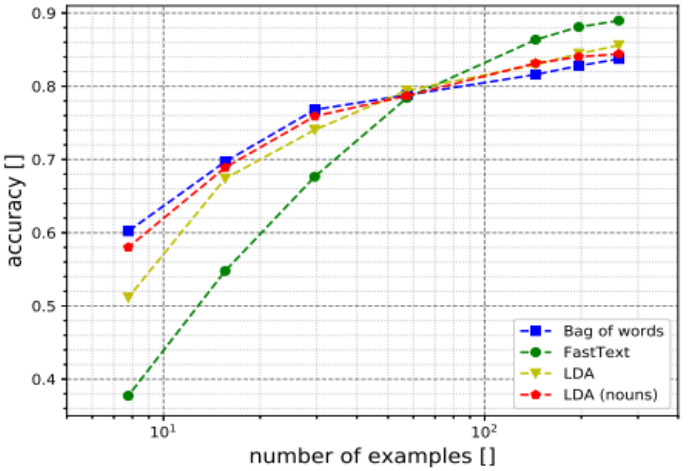
- Uczenie z nauczycielem (z ang. supervised learning)
  - dane wejściowe => model ML => wyjście
  - budujemy model na podstawie par (tekst, etykieta)



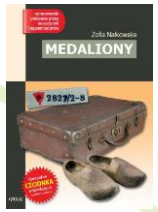
  - mamy nadzieję, że model będzie generalizował
  - nauczymy system tylko tego co było w danych (anotowanych)
  - modele są coraz większe
    (np. BERT – 340 mln. liczb do wyznaczenia)

Problemy z danymi w uczeniu maszynowym

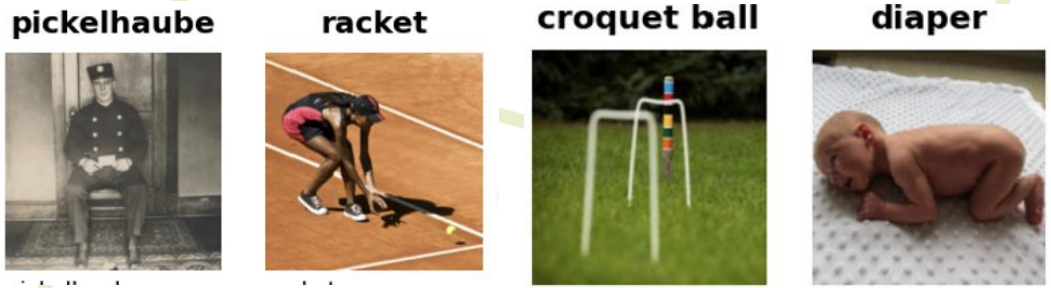
- przygotowanie danych anotowanych jest kosztowne
- potrzebujemy dużo danych
  - klasyfikacja tematyczna, 32 klasy



- system może błędnie działać dla danych odległych od danych uczących
- pozorne korelacje
  - ImageNet

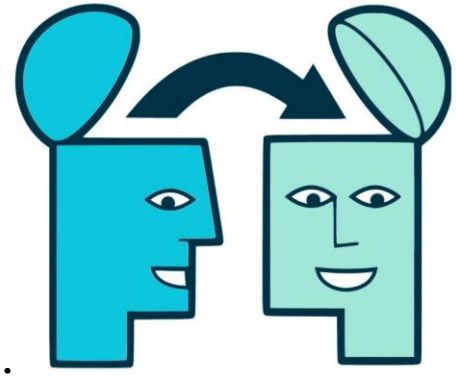


NKJP



- dane mogą być błędnie oznaczone

Transfer wiedzy jako remedium na część problemów



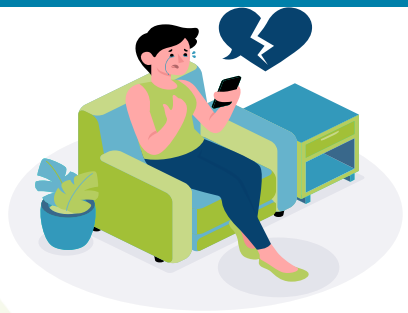
- *Transfer learning*
  - wykorzystanie modelu nauczonego na innym zdaniu
  - douczenie go w nowym zadaniu
- Modele językowe
  - gigantyczne modele na bazie bardzo dużych zbiorów tekstów nie wymagających znakowania
  - uczenie nadzorowane
    - mimo formalnego braku etykiet
    - predykcja słowa (maskowanie), predykcja następnego zdania
  - ogólnodostępne modele

fastText

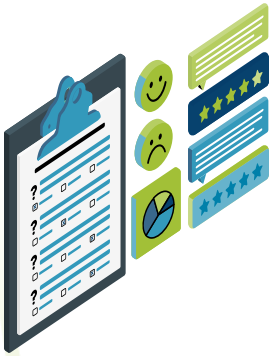
- SOTA w wielu zdaniach klasyfikacji tekstu
  - bierzemy model językowy
  - douczamy go do konkretnego zadania
    - dane potrzebne, ale mniej, albo lepsze wyniki na tych samych danych
- Transfer międzyjęzykowy

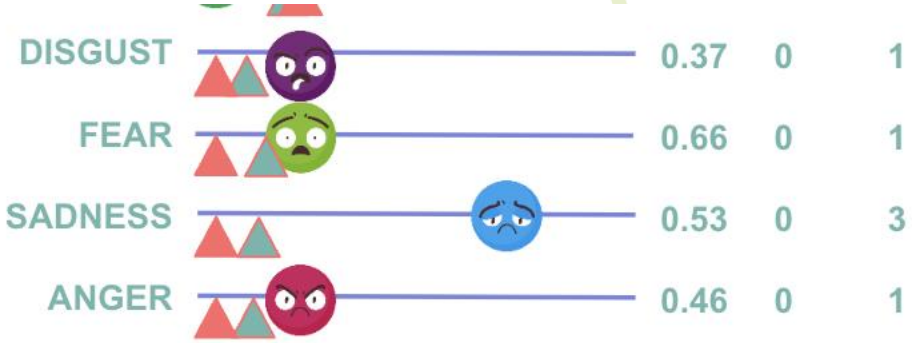
HUGGING FACE

Spersonalizowane NLP



- Analiza emocji (wiele modelu, różne wymiary)
  - bardzo gorący temat
  - uczenie maszynowe: tekst -> poziomy różnych emocji
  - ludzie są różni i mają różne oceny emocji tego samego tekstu





  - rozwijamy metody, które dodatkowo modelują ludzi (i nie tylko)
  - model człowieka (tendencyjność) można dostosować do osoby na 5-6 przykładach
- Dane językowe w obszarach subiektywnych (emocje, mowa nienawiści, …) powinny zawierać informacje o oceniających
  - przynajmniej identyfikator
  - dane demograficzne (wiek, płeć, …)

Wydobywanie relacji między fragmentami tekstów - semantyka i podobieństwo zdań

- Uchwycenie charakteru relacji (jakiego rodzaju jest to podobieństwo) wymaga analizy tekstów na wielu poziomach (leksykalnym, składniowym, semantycznym)
- Zdania pozornie podobne mogą być sprzeczne np. ze względu na wystąpienie w nich odmiennych jednostek nazewniczych
- Konstruowanie zbioru (testowo-treningowego) wymaga zebrania **reprezentatywnej** próbki tekstów:
  - różnorodność tematyczna / dziedzinowa
  - różnorodność zjawisk językowych (jednostki nazewnicze, wyrażenia czasowe i przestrzenne itd.)
  - różnorodność relacji (parafraza, zawieranie, krzyżowanie, sprzeczność, wynikanie itd.)
  - praca ręczna minimum 2+1; dla pewnych typów relacji trudno osiągnąć zgodność

Czyszczenie danych
(wejściowych czy do uczenia modelu)

- Z danych potrzeba usunąć dane osobowe
  - https://ws.clarin-pl.eu/anonymizer
- Pozyskiwane dane tekstowe zawierają wiele zanieczyszczeń (specyficznych dla źródła)
  - ASR: brak interpunkcji, duże litery, podział na zdania
    - https://ws.clarin-pl.eu/punctuator
  - OCR/PDF
    - podział na kolumny i ciągłość tekstu, kroje czcionki, obrazy
    - niejęzykowe wtrącenia
    - literówki, sklejenia wyrazów, przeniesienia, podział na akapity
  - portale informacyjne, społecznościowe
    - brak dostępu
    - wtrącenia z reklam czy odsyłaczy do powiązanych treści
    - materiały zbiorcze

Zanieczyszczenia danych - przykład

- Modelowanie tematyczne
  - https://ws.clarin-pl.eu/topic
- Analiza dyskursu publicznego
- Dane na zadany temat (słowa klucze) z kilku portali informacyjnych (15 tys. dokumentów)
- Detekcja tematyki i jej zmienności w czasie

Akapit tekstu
REDAKCJA POLECA
Stwórz swoją umowę - szybko i profesjonalnie!
3 linie tekstu
Komentarz dnia
Newsletter Zapisz się i bądź na bieżąco
Niepełnosprawni
Tekst
Nie masz jeszcze konta w serwisie? Dołącz do nas
Pamiętaj mnie

Podsumowanie

DBAJ O...
JĘZYK POLSKI

- Dane językowe są niezbędne do budowy i działania narzędzi językowych
- Pozyskanie danych, zapewnienie ich reprezentatywności, ich anotacja, ale też i czyszczenie jest procesem kosztownym
- Przy opracowaniu danych (korpusów) należy dbać o zachowanie
  - jakości: spójności/reprezentatywności/….
  - jak największej ilości informacji
- Przy automatycznym pozyskiwaniu danych należy zwrócić uwagę na minimalizowanie zanieczyszczeń
- Do opracowania i efektywnego użycia technologii językowych potrzeba trzech rzeczy
  - danych, danych, danych….
  - dobrej jakości i poprawnie anotowanych
- Dużo, dobrej jakości, otwartych zbiorów danych językowych

CLARIN-PL
Common Language Resources and Technology Infrastructure

Dziękuję za uwagę!

tomasz.walkowiak@pwr.edu.pl

https://ws.clarin-pl.eu/

Wartość danych dla rozwoju technologii języka na przykładzie CLARIN-PL

CLARIN