Osadzenia kontekstowe i ich wpływ na przetwarzanie języka naturalnego

Paweł Rychlikowski

Instytut Informatyki UWr

17 grudnia 2025

Bezkontekstowe vs kontekstowe osadzenia

Bezkontekstowe (word2vec)

- **PLUS**: jakaś część znaczenia słowa jest niezależna od kontekstu (bo inaczej nie moglibyśmy się porozumiewać
- **MINUS**: ale word2vec **zawsze** pomija kontekst!

Słowa wieloznaczne lub uzyskujące znaczenie

- Słowa wieloznaczne: zamek, żabka, stan, dół, szczyt
- ta potrawa, ta autorka, nasz bohater, ...
- Polisemia: bank (budynek, instytucja, firma)
- Te zjawiska zachodzą w różnych językach (bank, lead, bass, content, ...)

BERT

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

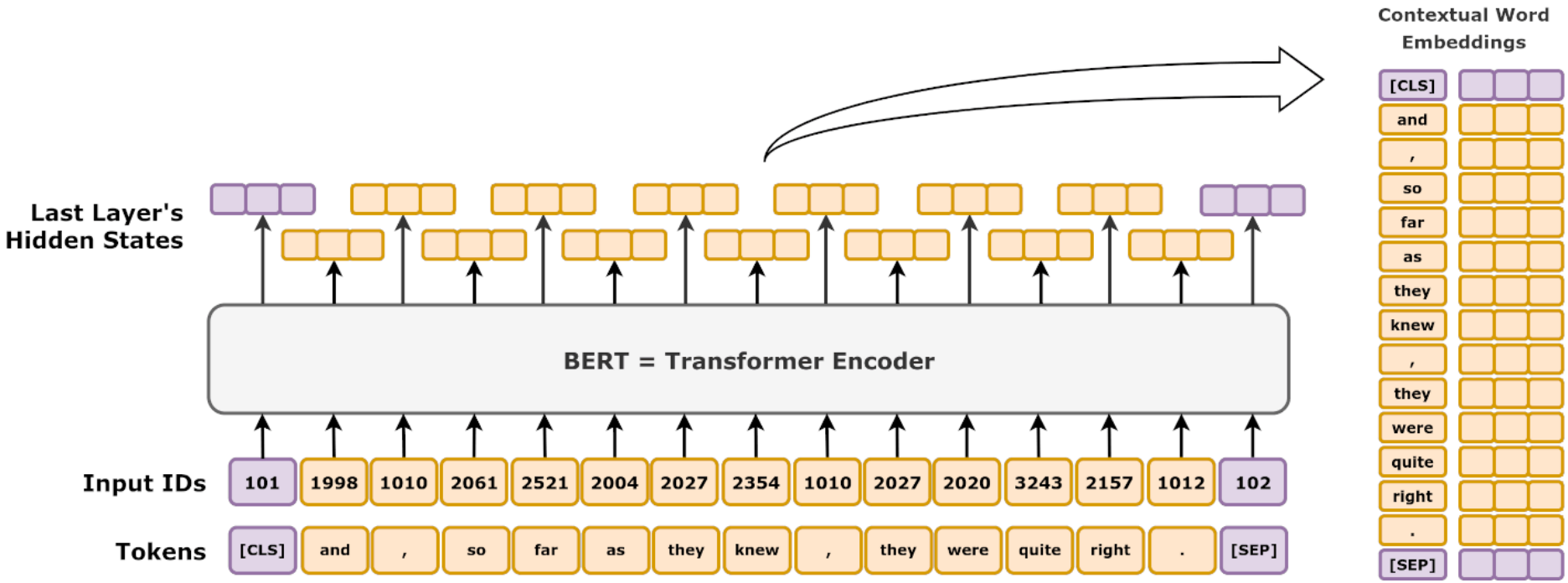
Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language repre-

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that

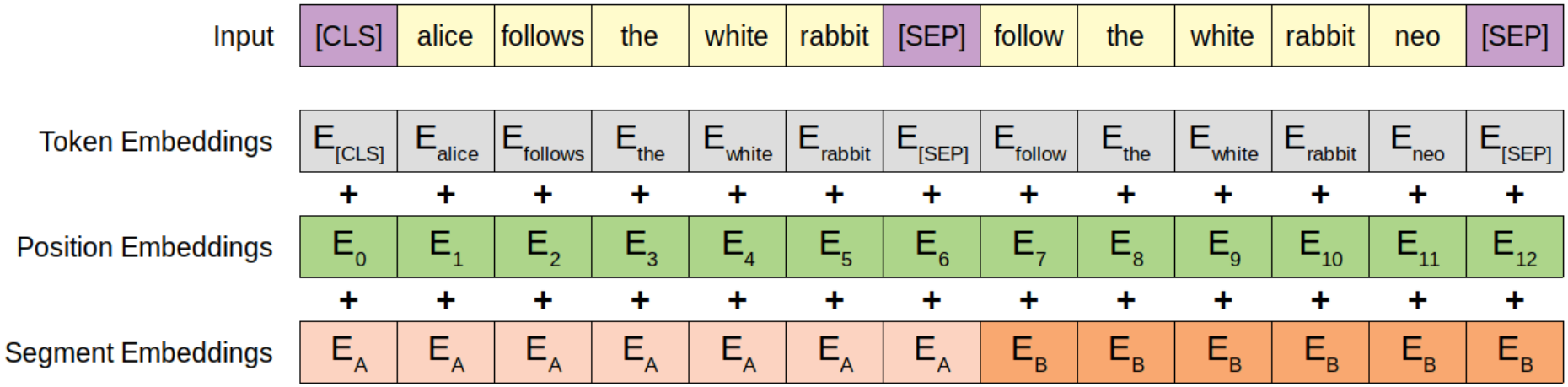
- Najbardziej udany koder transformerowy!
- Praca, która zmieniła definitywnie NLP!

Ogólny schemat BERT-a



Kodowanie wejścia BERT-a

- Sieci transformer przetwarzają wejście równolegle, traktując każdy wektor tak samo.
- Oczywiście kolejność słów ma znaczenie, trzeba ją jakoś zakodować



O dodawaniu wektorów (część 2)

- Dodawanie wektorów tego samego typu możemy traktować jako przybliżone obliczanie sumy mnogościowej
- Dodawanie wektorów innych typów (osadzenie pozycji i osadzenie sementyczne) daje dodatkowe możliwości:
  informacje różnych typów mogą być w innych mniejscach wektora (zob. tablica)

Rzut oka na kodowania pozycji w Papudze

Popatrzmy ponownie na notatnik embeddings.ipynb

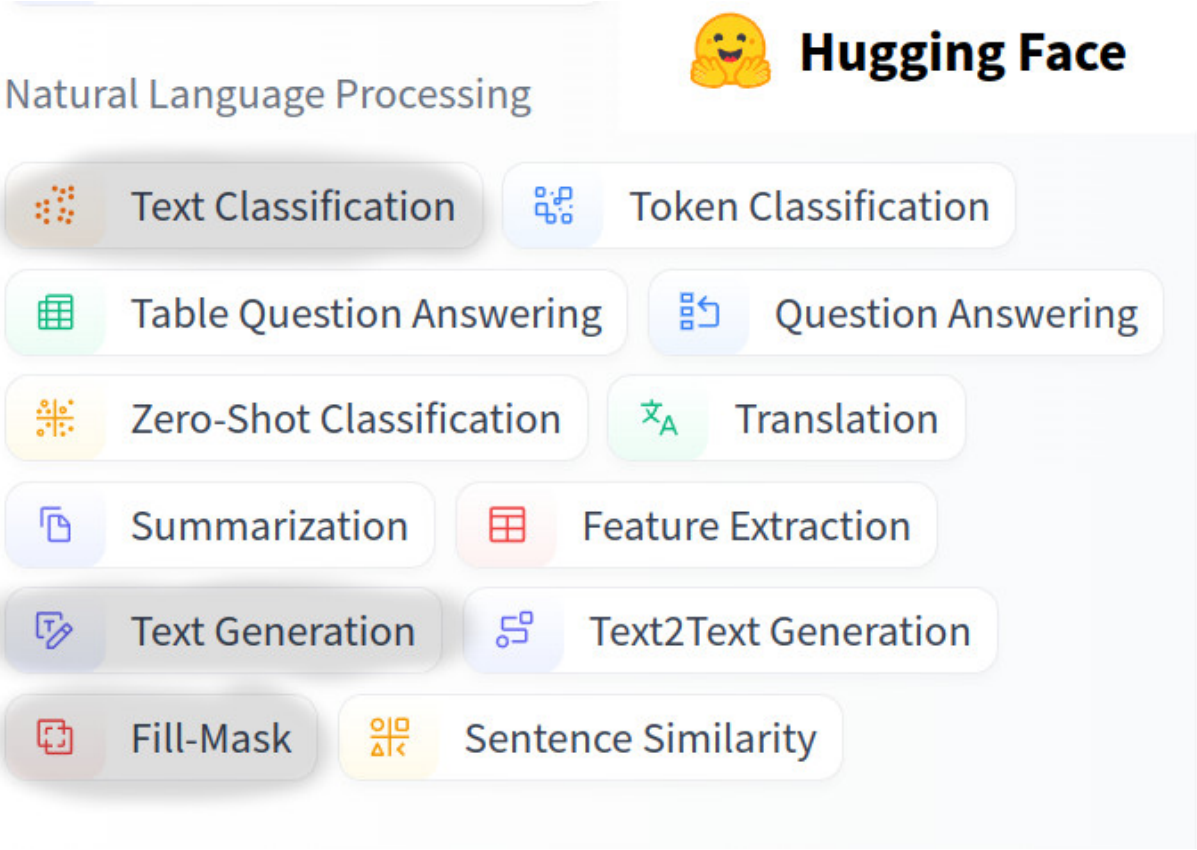
BERT jako ekstraktor cech

- Możemy potraktować wektor wynik obliczony przez BERT-a, jako reprezentację zdania.
  - Albo token dla pozycji |<CLS>|
  - Albo średnią wektorów dla tokenów

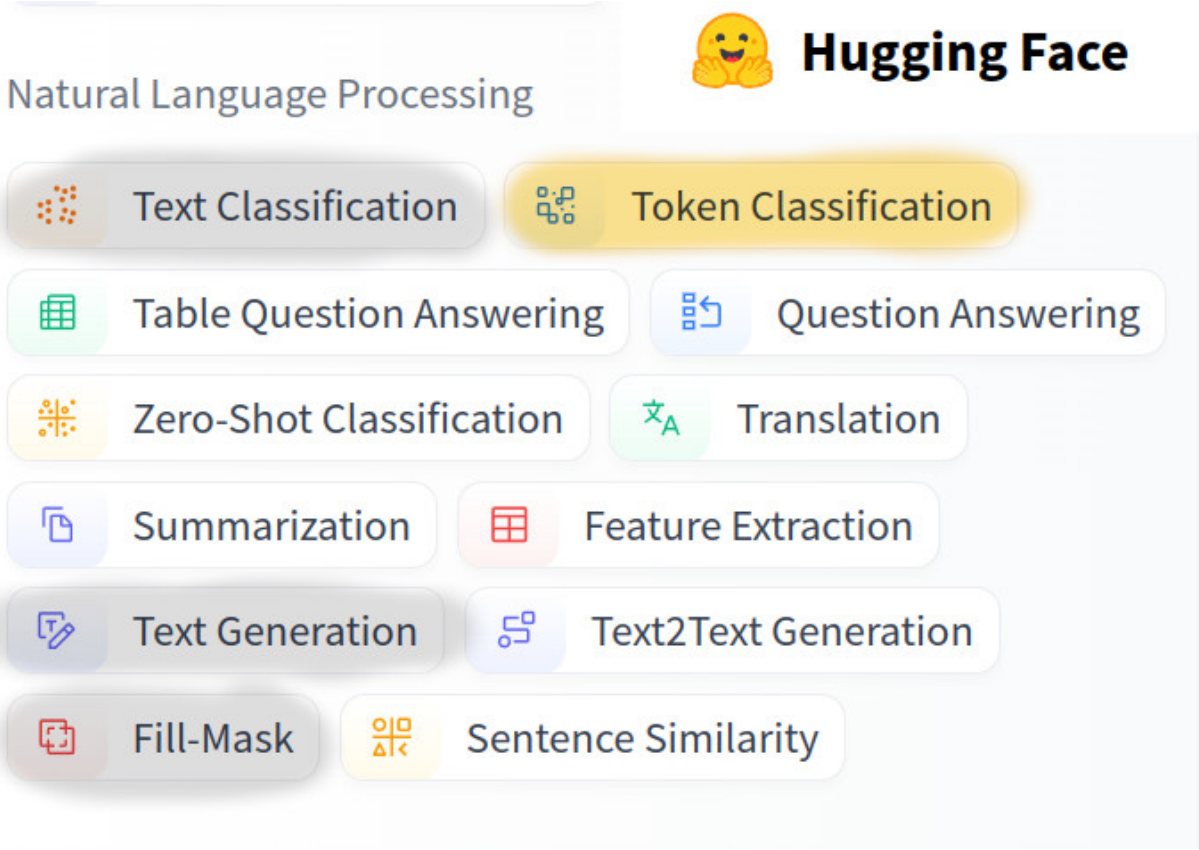
BERT po polsku

- Istnieje kilka powszechnie dostępnych modeli typu BERT dla języka polskiego (np. HerBERT, PolBERT)
- Zobaczymy jak Herbert działa w najprostszym scenariuszu
  - Przerwa na demonstracje (`herbert.ipynb`)
- Na liście 3 zbadamy bardziej złożone warianty:
  - Różne inne algorytmy ML (+regularyzacja)
  - Augmentacja danych
  - Łączenie Herberta z Papugą

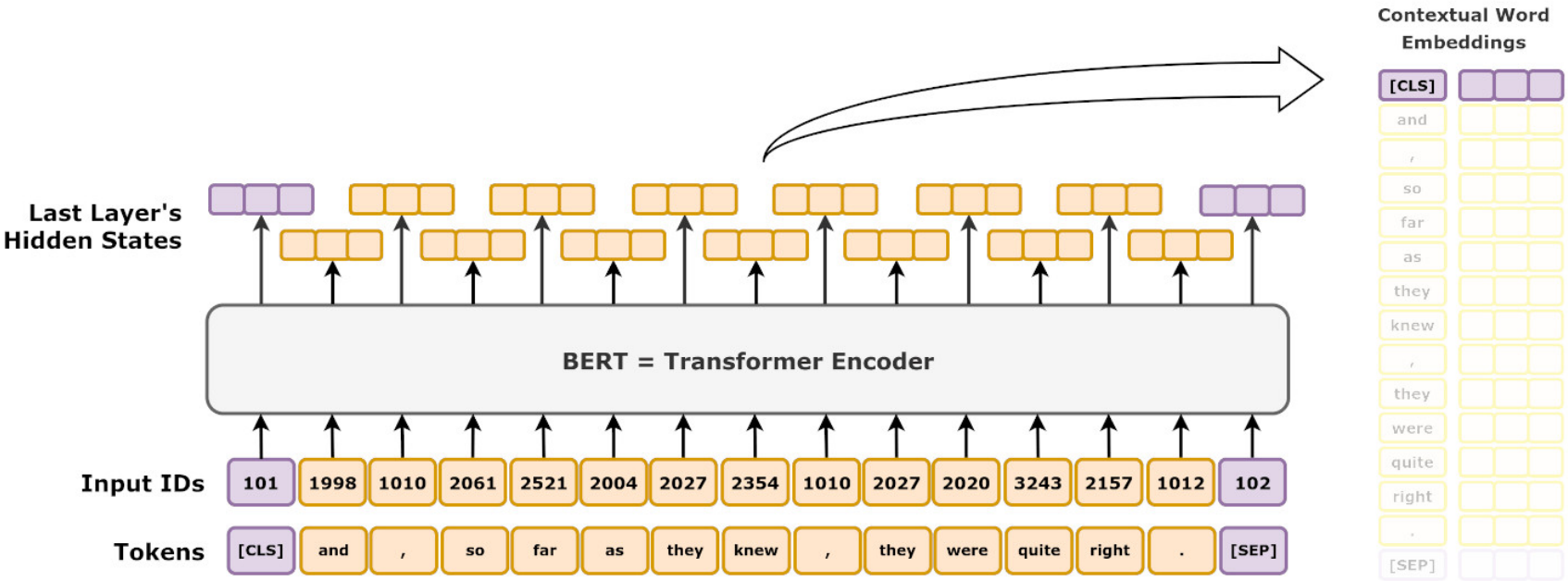
NLP w HuggingFace



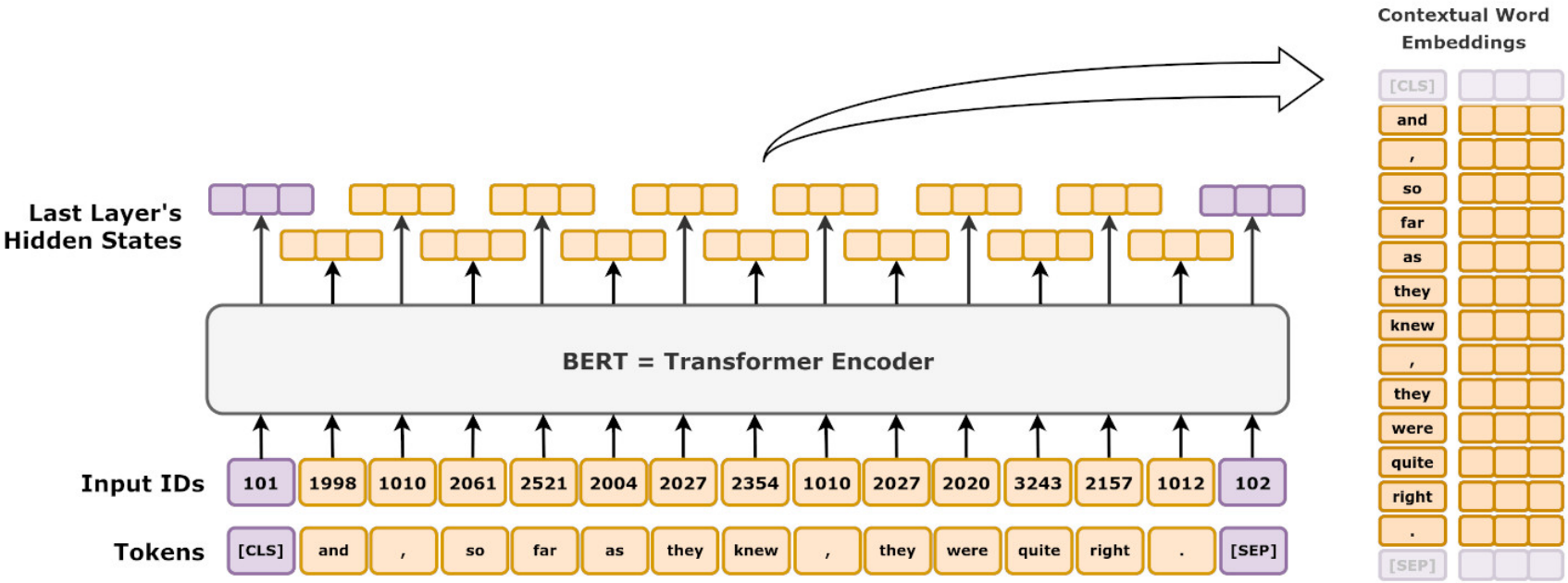
NLP w HuggingFace



Klasyfikacja dokumentów



Klasyfikacja tokenów



Klasyfikacja tokenów

- Klasyczne zadania NLP: POS-tagging oraz Named Entity Recognition
- Znajdywanie istotnych fragmentów tekstu (0/1 dla każdego tokenu)
- Rekonstrukcja interpunkcji
- ...

A teraz trochę koniecznej lingwistyki

Czego uczyli nas w szkole?

A teraz trochę koniecznej lingwistyki

Czego uczyli nas w szkole?

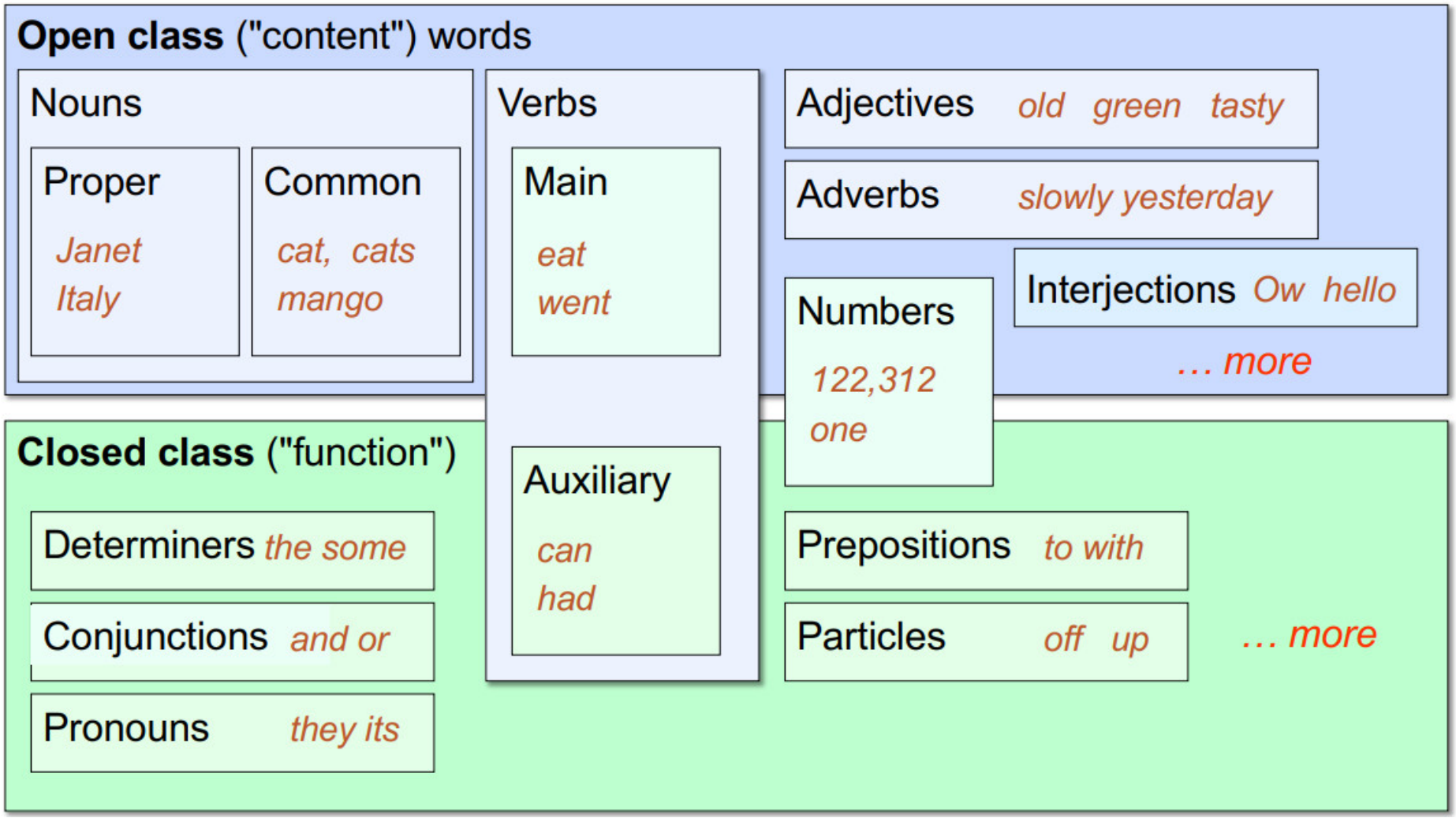
- Każdy wyraz jest jakąś częścią mowy.
- Główne części mowy to rzeczownik, czasownik, przymiotnik, przysłówek.
- Istnieją też inne części mowy, takie jak przyimek, spójnik, zaimek, partykuła.

A teraz trochę koniecznej lingwistyki

Czego uczyli nas w szkole?

- Każdy wyraz jest jakąś częścią mowy.
- Główne części mowy to rzeczownik, czasownik, przymiotnik, przysłówek.
- Istnieją też inne części mowy, takie jak przyimek, spójnik, zaimek, partykuła.
- Podział na części mowy zawdzięczamy Dionizusowi Thraxowi z Aleksandrii (ok 100pne). Wyodrębnił on 8 wyżej wymienionych części mowy (bez partykuły, ale za to z rodzajnikiem).

Części mowy po angielsku



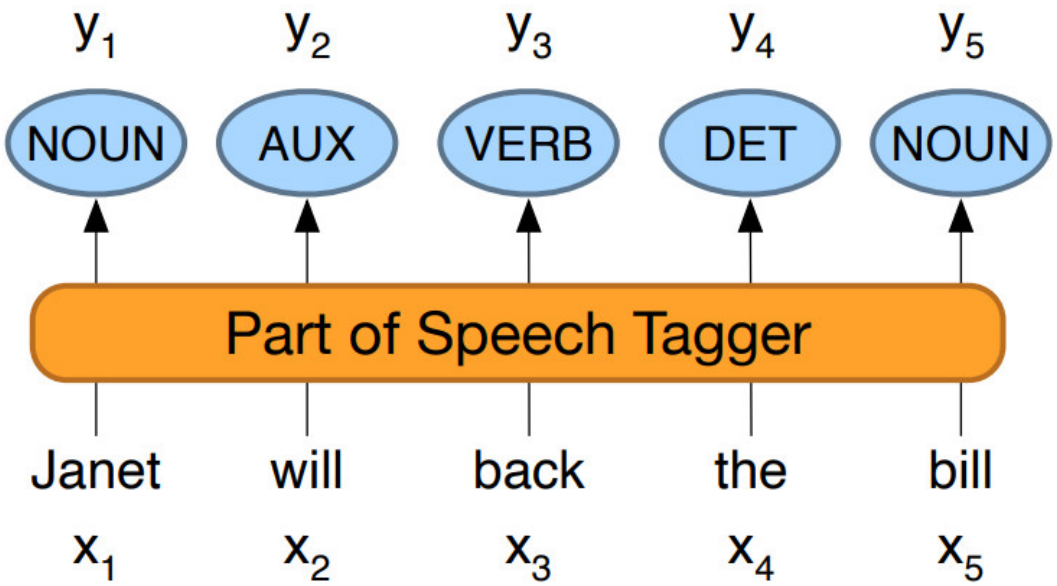
Przykłady po polsku

1. Rzeczownik: krowa, koń, sytuacja, uczucie
2. Czasownik: być, mieć, robić
3. Przymiotnik: ładny, piękny, najurodziwszy
4. Przysłówek: ładnie, pięknie, najurodziwiej, bardzo
5. Przyimek: do, poprzez, od, wokół, niczym
6. Zaimek: on, jego, mój, tak, taki, ile, gdzie
7. Imiesłów: umierając, umierający, umarłszy, umarły, zabijany (!umierany)
8. Spójnik: i, oraz, lecz, lub, że
9. Liczebnik: dwa, trzy, czwarty
10. Rodzajnik: a, the, der, die, das, eine, les
11. Inne dziwne (wykrzykniki, partykuły, kubliki, partykułoprzysłówki, ...): ha, się, nie, żesz,

Zadanie Part-of-Speech tagging

Zadanie

Dla ciągu słów $x_1, \ldots, x_n$ znajdź odpowiadający im ciąg POS-tagów $y_1, \ldots, y_n$



- Zwróćmy uwagę, że długości sekwencji **są równe** (inaczej niż w tłumaczeniu)
- Musimy umówić się na tzw. **tagset** (co nie jest oczywiste, ale nie będziemy się tym zajmować)

Trudność (?) tagowania

Tagowanie (w wielu językach, w tym polskim i angielskim) nie jest **tylko** odczytaniem tagu z wielkiej tablicy słów.

Przykłady

Mam radę: nie **mam mam** pustymi obietnicami –

Dwie **dziewczyny** idą do trzeciej **dziewczyny** –
Patrzę na **stół**, a ten **stół** ciągle stoi. –
Już dawno po **kolacji**, a ja myślę wciąż o **kolacji**. –

Trudność (?) tagowania

Tagowanie (w wielu językach, w tym polskim i angielskim) nie jest **tylko** odczytaniem tagu z wielkiej tablicy słów.

Przykłady

Mam radę: nie **mam mam** pustymi obietnicami – [czas.], [rozkaźnik], [rzecz.]
Dwie **dziewczyny** idą do trzeciej **dziewczyny** – poj vs mnoga
Patrzę na **stół**, a ten **stół** ciągle stoi. – biernik vs mianownik
Nie jadłem jeszcze **kolacji**, więc wciąż myślę o **kolacji**. – dopełniacz vs miejscownik

POS-tagging wczoraj i dziś

Wczoraj

Podstawowe zadanie z NLP, poprzedzające wiele innych aplikacji.

Dziś

- Do analiz lingwistycznych (jakie proporcje rzeczowników do przymiotników miał Sienkiewicz)
- Może pomóc w prostych aplikacjach NLP
- (zob. biblioteki **spaCy**, **Stanza**)

Jutro

Być może umieszczanie tagów pomaga transformerom modelować język (hipoteza)

Named Entity Recognition (NER)

- Po polsku: rozpoznawanie nazwanych encji
- Identyfikacja fraz (najczęściej nazw własnych), czasem wielowyrazowych, o różnych typach.

PER (Person): “Marie Curie”
LOC (Location): “New York City”
ORG (Organization): “Stanford University”
GPE (Geo-Political Entity): "Boulder, Colorado"

Najczęściej płytkie, bez struktury, choć III Liceum Ogólnokształcące im. **Adama Mickiewicza**

NER jako zadanie tagowania

- Identyfikację fraz można potraktować jako zadanie klasyfikacji tokenów

[PER Jane Villanueva] of [ORG United] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago] route.

Words	BIO Label
Jane	B-PER
Villanueva	I-PER
of	O
United	B-ORG
Airlines	I-ORG
Holding	I-ORG
discussed	O
the	O
Chicago	B-LOC
route	O
.	O

NER jako zadanie tagowania

- Możliwych jest wiele wariantów definiowania tagów
- BIO jest najbardzień powszechny!

[PER Jane Villanueva] of [ORG United] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago] route.

Words	IO Label	BIO Label	BIOES Label
Jane	I-PER	B-PER	B-PER
Villanueva	I-PER	I-PER	E-PER
of	O	O	O
United	I-ORG	B-ORG	B-ORG
Airlines	I-ORG	I-ORG	I-ORG
Holding	I-ORG	I-ORG	E-ORG
discussed	O	O	O
the	O	O	O
Chicago	I-LOC	B-LOC	S-LOC
route	O	O	O
.	O	O	O

Dlaczego to zadanie jest istotne

- Monitorowanie mediów (wyłapywanie marek produktów w różnych kontekstach)
- Odpowiadanie na pytania (Kto? – fraza o typie [PER])
- Ekstrakcja wiedzy (faktów) z tekstu

Popularne rozwiązania (kiedyś)

- Ukryte łańcuchy Markowa
- CRF (Conditional Random Fields)
- Różne sieci neuronowe (w tym rekurencyjne)

Uwaga

- Wiele modeli zakładało „osobne" modelowanie języka znaczników: że po B-PER może być I-PER, ale nie I-LOC itd (wraz z prawdopodobieństwami).
- Teraz zakładamy raczej, że osadzenia kontekstowe zawierają wystarczająco dużo wiedzy, by na ich podstawie podejmowac niezależnie decyzję.

NER obecnie

Uczenie klasyfikatora biorącego **kontekstowe** osadzenie **bieżącego tokena**

- Prawie dokładnie ten sam kod, co w naszej demonstracji z wydźwiękiem (zamiast tokenu [CLS] bierze się wszystkie inne tokeny)
- Więcej przypadków uczących z jednego zdania!