Przetwarzanie języka naturalnego (NLP)

Paweł Rychlikowski

Instytut Informatyki UWr

2 grudnia 2024

Cele na dzisiejszy i kolejne wykłady

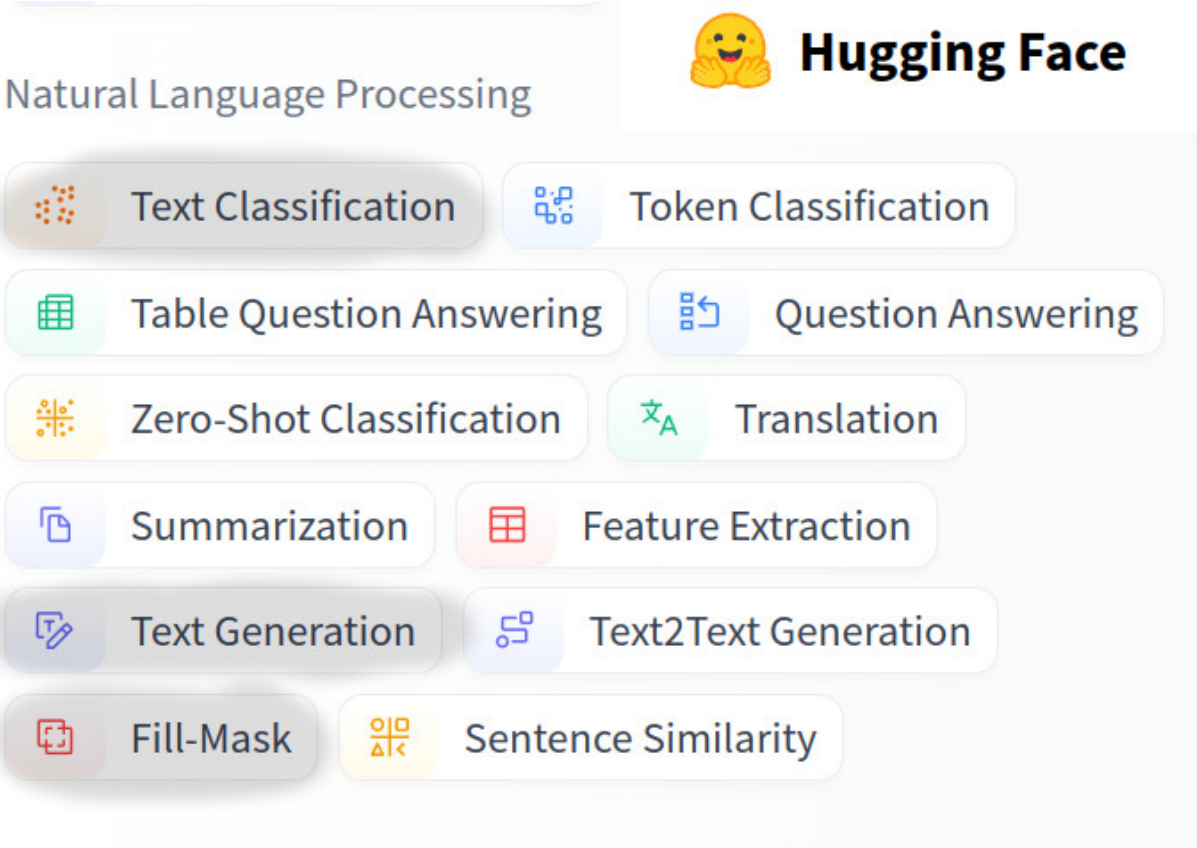
NLP z transformerami

- Zadania klasyfikacji tokenów
- Odpowiadanie na pytania (i wstępny wstęp do łączenia modeli z tradycyjnym wyszukiwaniem, czyli RAG)
- O systemech tłumaczących i streszczających (i augmentacji danych)

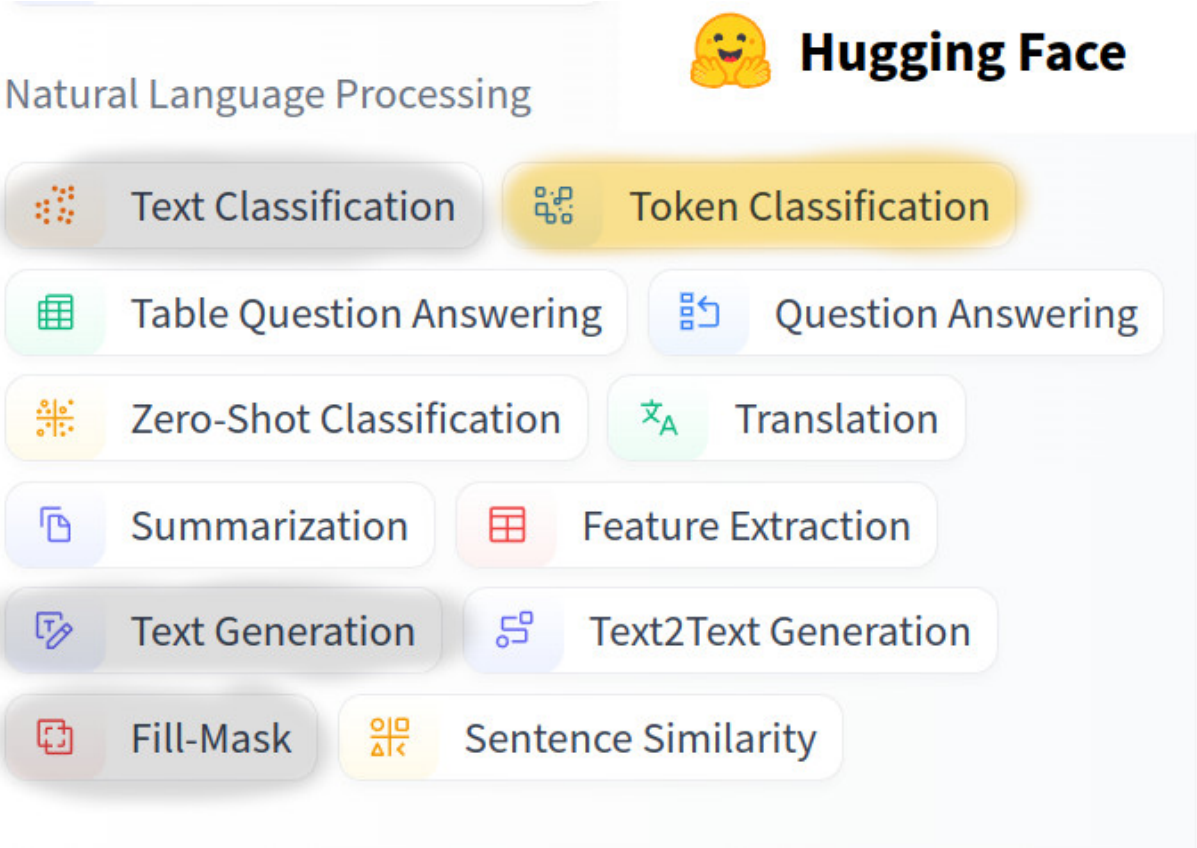
Anatomia transformerów

- Mechanizm uwagi (**Attention is all you need**)
- ... i cała reszta architektury
- Trening i dostrajanie transformerów

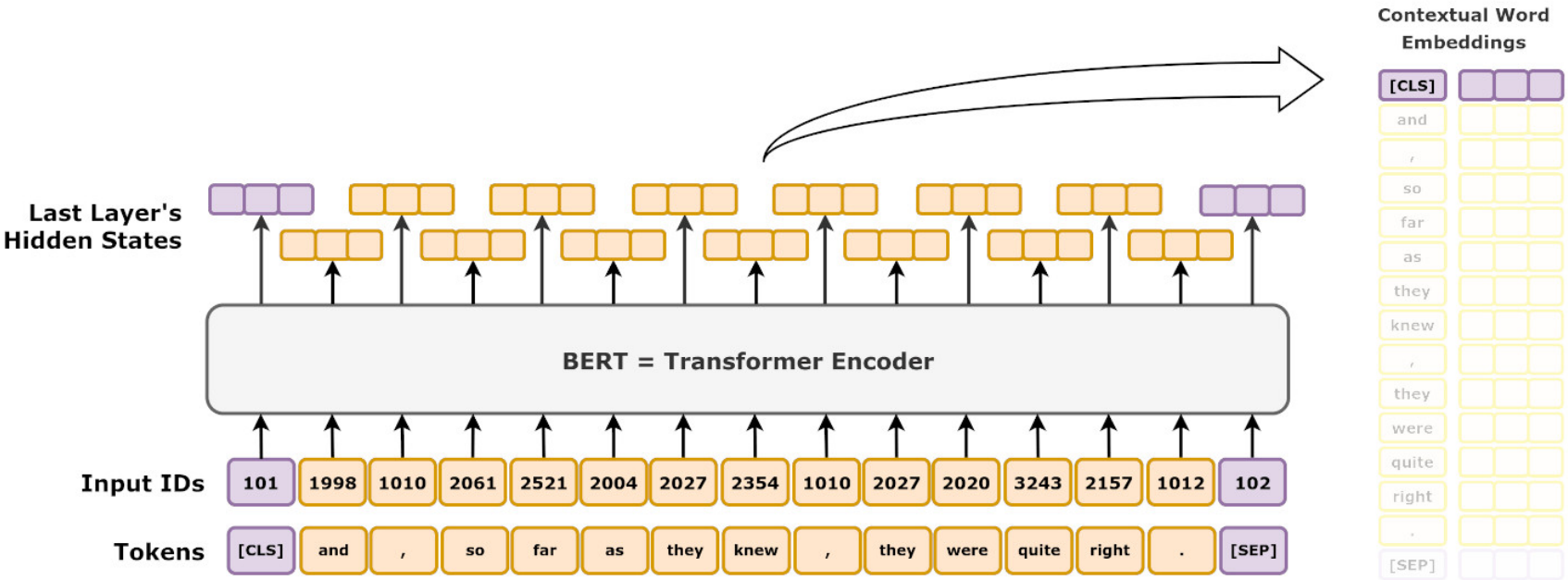
NLP w HuggingFace



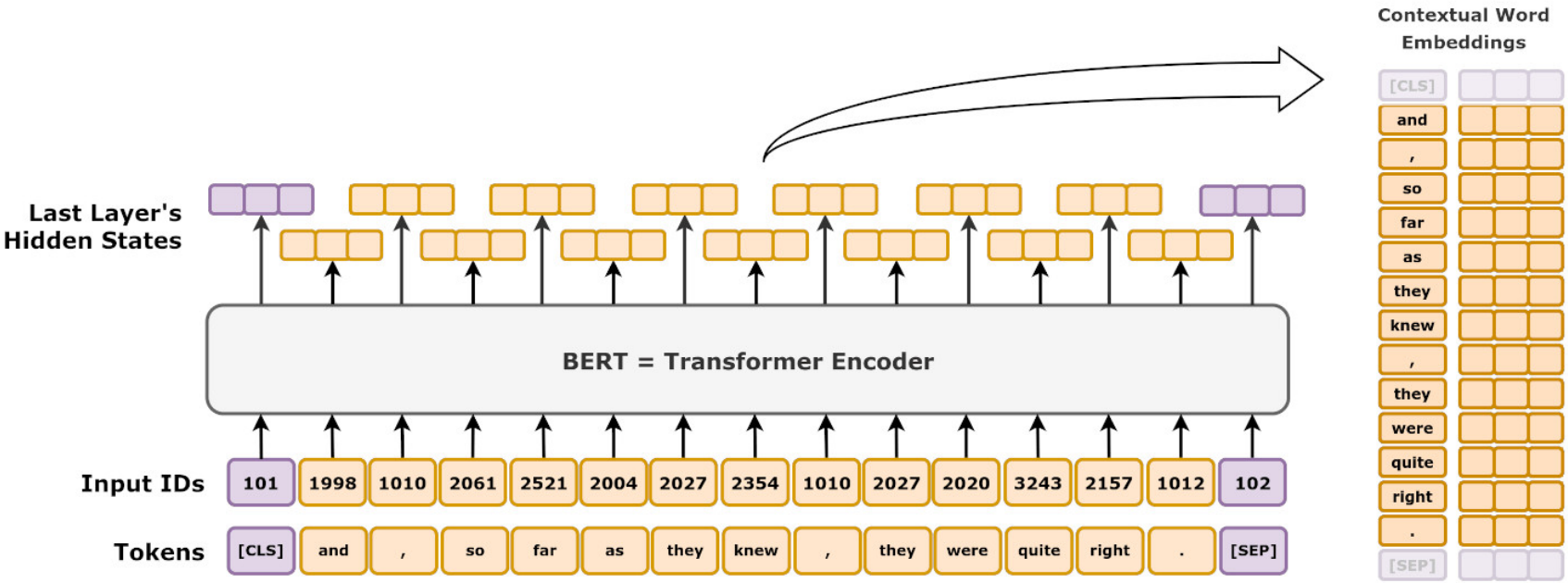
NLP w HuggingFace



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?

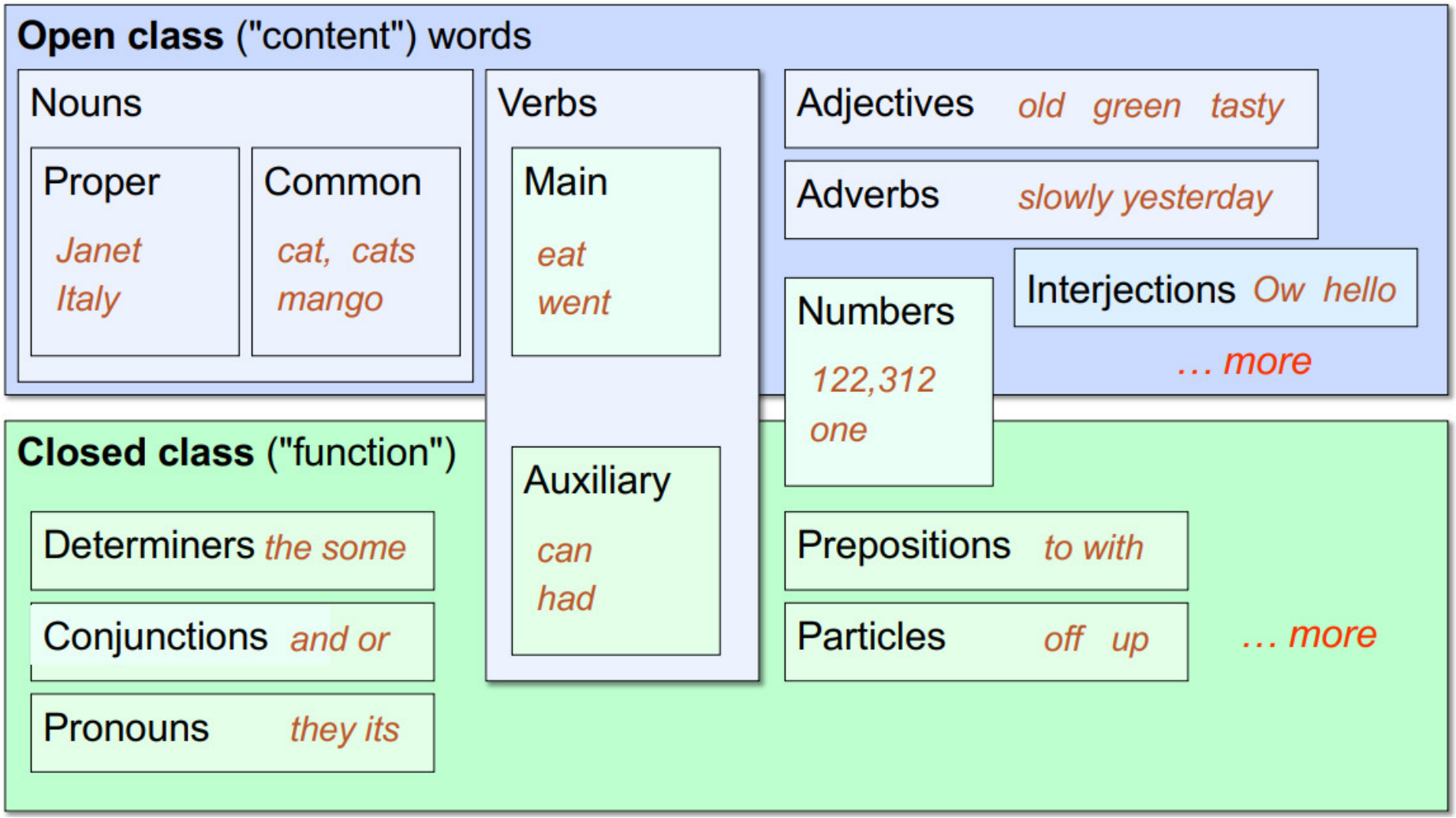
- Każdy wyraz jest jakąś częścią mowy.
- Główne części mowy to rzeczownik, czasownik, przymiotnik, przysłówek.
- Istnieją też inne części mowy, takie jak przyimek, spójnik, zaimek, partykuła.

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



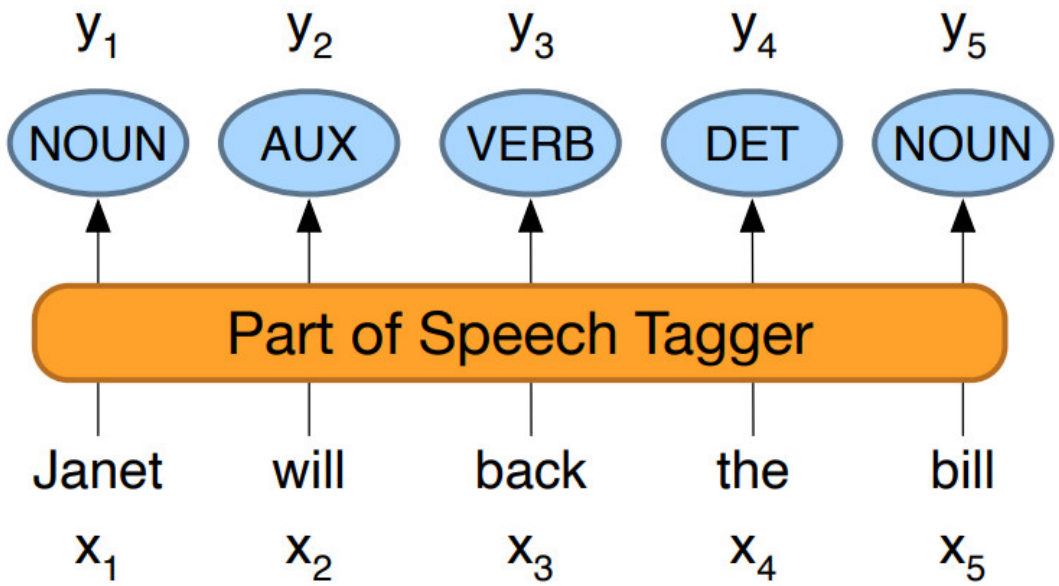
Przykłady po polsku

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**
- 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
Już dawno po **kolacji**, a ja myślę wciąż 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. biblioteka spaCy)

Jutro

Być może umieszczanie tagów pomaga transformerom modelować język (hipoteza)

Dygresja (do pracowni)

Obejrzmy plik tags.txt i zastanówmy się, jak można go wykorzystać wraz z word2vec do augmentacji tekstu

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!

Zamrożony transformer vs dostrajanie

- W naszej demonstracji transformer był zamrożony (zakładaliśmy, że osadzenia są na tyle uniwersalne, że zadziałają do konkretnego zadania)
- Alternatywą jest dołożenie części klasyfikującej i trenowanie takiej całości na zadaniu docelowym (nieco bardziej kosztowne, większe ryzyko przetrenowania, ale ogólnie – raczej dominująca taktyka)
- Można też chwilę trenować całość, po czym zamrozić większość sieci i wytrenować mały klasyfikator na tak zaadaptowanych zanurzeniach kontekstowych.

Rekonstrukcja interpunkcji. Mniej typowe zadanie klasyfikacji tokenów

Input

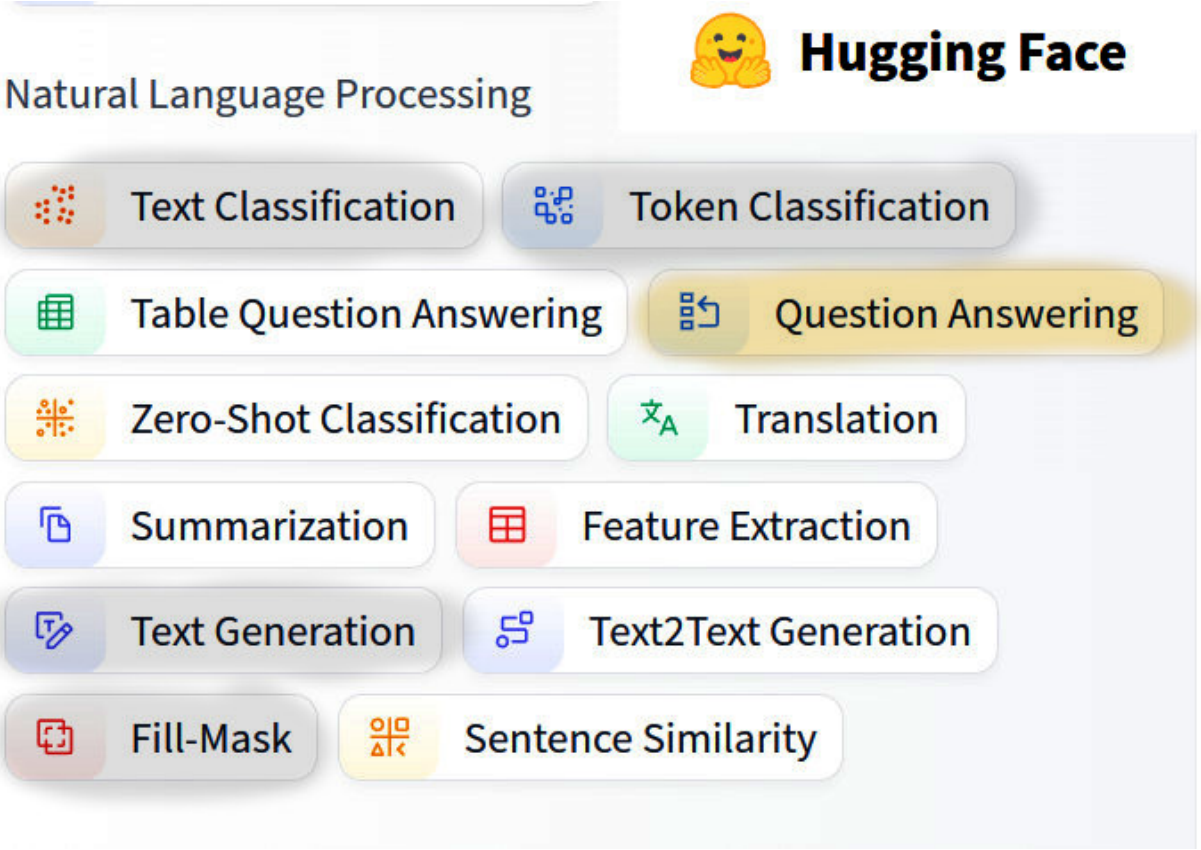
scottish actor gatwa who was born in rwanda is best known for starring in netflix's sitcom sex education he told bbc news it feels really amazing it's a true honour this role is an institution and it is so iconic

Output

Scottish actor Gatwa, who was born in Rwanda, is best known for starring in Netflix's sitcom Sex Education.
He told BBC News: "It feels really amazing. It's a true honour. This role is an institution and it's so iconic."

Klasy: normal Capital UPPER normal-comma normal-dot Capital-dot ...

NLP w HuggingFace



Odpowiadanie na pytanie

Uwaga

W zadaniu tym zakładamy, że pytanie ma prostą, jednoznaczną odpowiedź, jest raczej *encją*, niż opinią czy zdaniem.

Odpowiadanie na pytanie

Uwaga

W zadaniu tym zakładamy, że pytanie ma prostą, jednoznaczną odpowiedź, jest raczej *encją*, niż opinią czy zdaniem.

- Podstawową metryką oceniającą sukces jest **Exact Match** – czyli że oczekiwany napis i zwrócony przez system są identyczne (oczekiwanych napisów może być więcej, wystarczy że 1 trafimy)
- Mamy dwa podejścia:
  1. **Closed book**: sam model ma wygenerować odpowiedź (Polka daje koło 15%, bardzo duże modele są istotnie lepsze)
  2. **Open book**: łączenie modeli językowych z mniej lub bardziej tradycyjnym wyszukiwaniem informacji (albo w kolekcjach tekstów, albo w bazach danych)

Retriever/Reader/Generator

Retriever

System wyszukiwania informacji (Google like). Dla zapytania (query) zwraca listę dokumentów (zdań, akapitów, ...) pasujących do zapytania

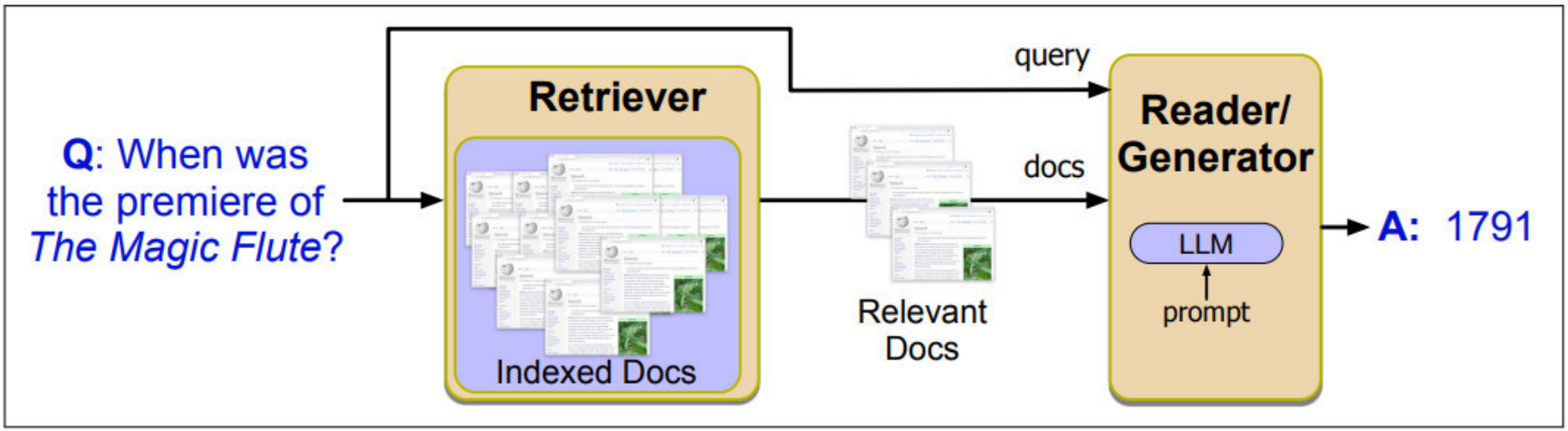
Generator

Model językowy, generujący odpowiedź token po tokenie.

Reader

Sieć neuronowa, biorąca na wejściu akapit tekstu oraz pytanie, zaznacza w akapicie te tokeny, które są odpowiedzią.

Retriever/Reader/Generator. Schemat



- Możemy mieć połączenia: Reader+Retriever lub Generator+Retriever (albo wszystkie 3)
- Można też wykorzystywać model językowy (autoregresywny) do przekształcenia pytania (question) w zapytanie (kwerendę, query).

Reader+Generator

Schematic of a RAG Prompt

```
retrieved passage 1

retrieved passage 2

...

retrieved passage n

Based on these texts, answer this question:  Q: Who wrote
the book ‘‘The Origin of Species"?  A:
```

Jak napisać retriever

Dwie opcje:

1. Klasyczna wyszukiwarka (na przykład bazująca na TF-IDF), zobacz również: **Ellasticsearch**)
2. Dense Passage Retrieval (to jak omówimy sobie architekturę transformera)

Wyszukiwanie informacji w pigułce

Ogólna zasada

Znajdź dokumenty (akapity, zdania) możliwie najbardziej podobne do zapytania. Podobieństwo mierz cosinusem rzadkich reprezentacji (TF-IDF, BM-25)

Killka użytecznych zaleceń/pomysłów/heurystyk

- Odwrotny indeks: odzworowanie term $\rightarrow$ zbiór-dokumentów-zawierających-term
- **termem** może być słowo, ale dla języka polskiego lepszy jest **lemat**.
- Heurystycznie ograniczamy liczbę obliczonych cosinusów (tylko dokumenty zawierające Ważne Termy z Zapytania (wszystkie? co najmniej 1?)
- Wagę termu możesz oceniać za pomocą IDF.

Zbiór danych SQUAD

SQUAD == The Stanford Question Answering Dataset

- Zbiór danych, który wywarł duży wpływ na NLP (ciągle użyteczny)
- Wesje 1.1 oraz 2.0 (ta druga zawiera złe pytania (czyli takie, na które w akapicie nie ma odpowiedzi)

SQUAD (cd)

Black_Death

The Stanford Question Answering Dataset

The Black Death is thought to have originated in the arid plains of Central Asia, where it then travelled along the Silk Road, reaching Crimea by 1343. From there, it was most likely carried by Oriental rat fleas living on the black rats that were regular passengers on merchant ships. Spreading throughout the Mediterranean and Europe, the Black Death is estimated to have killed 30–60% of Europe's total population. In total, the plague reduced the world population from an estimated 450 million down to 350–375 million in the 14th century. The world population as a whole did not recover to pre-plague levels until the 17th century. The plague recurred occasionally in Europe until the 19th century.

Where did the black death originate?
Ground Truth Answers: the arid plains of Central Asia | Central Asia | Central Asia
Prediction: arid plains of Central Asia

How did the black death make it to the Mediterranean and Europe?
Ground Truth Answers: merchant ships. | merchant ships | Silk Road
Prediction: killed 30–60% of Europe's total population

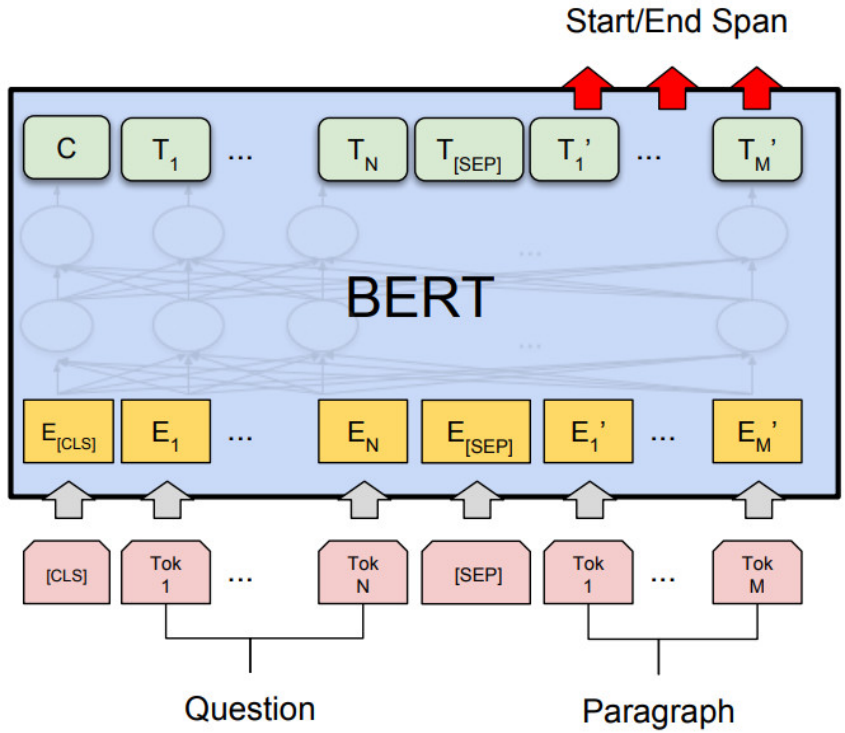
How much of the European population did the black death kill?
Ground Truth Answers: 30–60% of Europe's total population | 30–60% of Europe's total population | 30–60%
Prediction: 30–60%

When did the world's population finally recover from the black death?
Ground Truth Answers: the 17th century | 17th century | 17th century
Prediction: 17th century

For how long did the plague stick around?
Ground Truth Answers: until the 19th century | until the 19th century | 19th century

Reader (cd)

Jako reader najczęściej występuje obecnie sieć transformer typu BERT



- Tu raczej konieczne jest dostrajanie (fine-tuning)
- Ale jak najbardziej możliwy dzięki takim zbiorom danych jak SQUAD.