Przetwarzanie języka naturalnego (NLP)

Paweł Rychlikowski

Instytut Informatyki UWr

17 grudnia 2025

Dygresja (do pracowni)

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

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.

Dostrajanie jest relatywnie tanie – modele typu BERT są (w porównaniu do modeli GPT) dość małe.

Rekonstrukcja interpunkcji. Mniej typowe zadanie klasyfikacji tokenów

Input

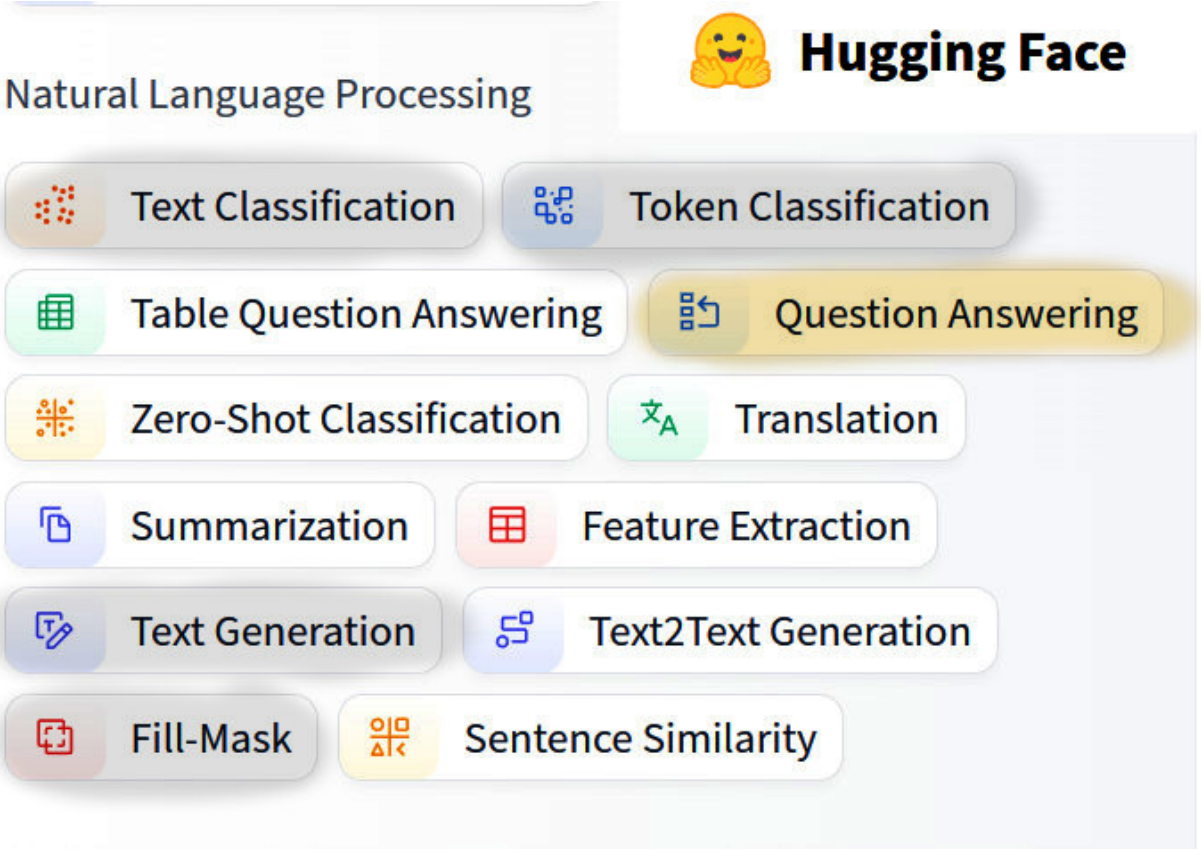
scottish actor gatwa who was born in rwanda is best known for starring in netflix's sitcom sex education he told bbc news it feels really amazing it's a true honour this role is an institution and it is so iconic

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

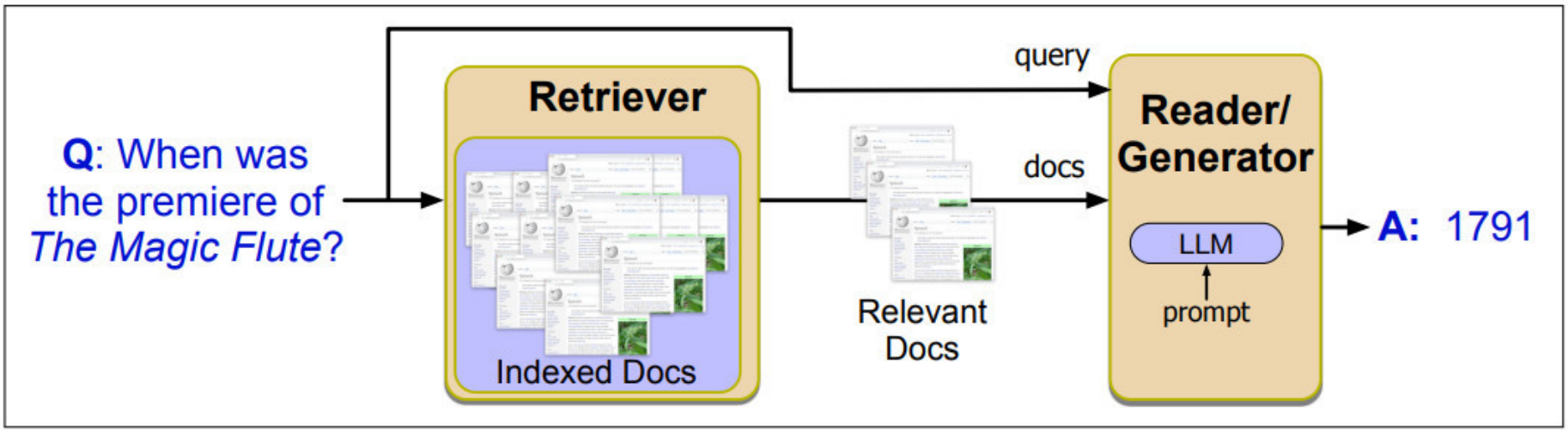
Generator

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

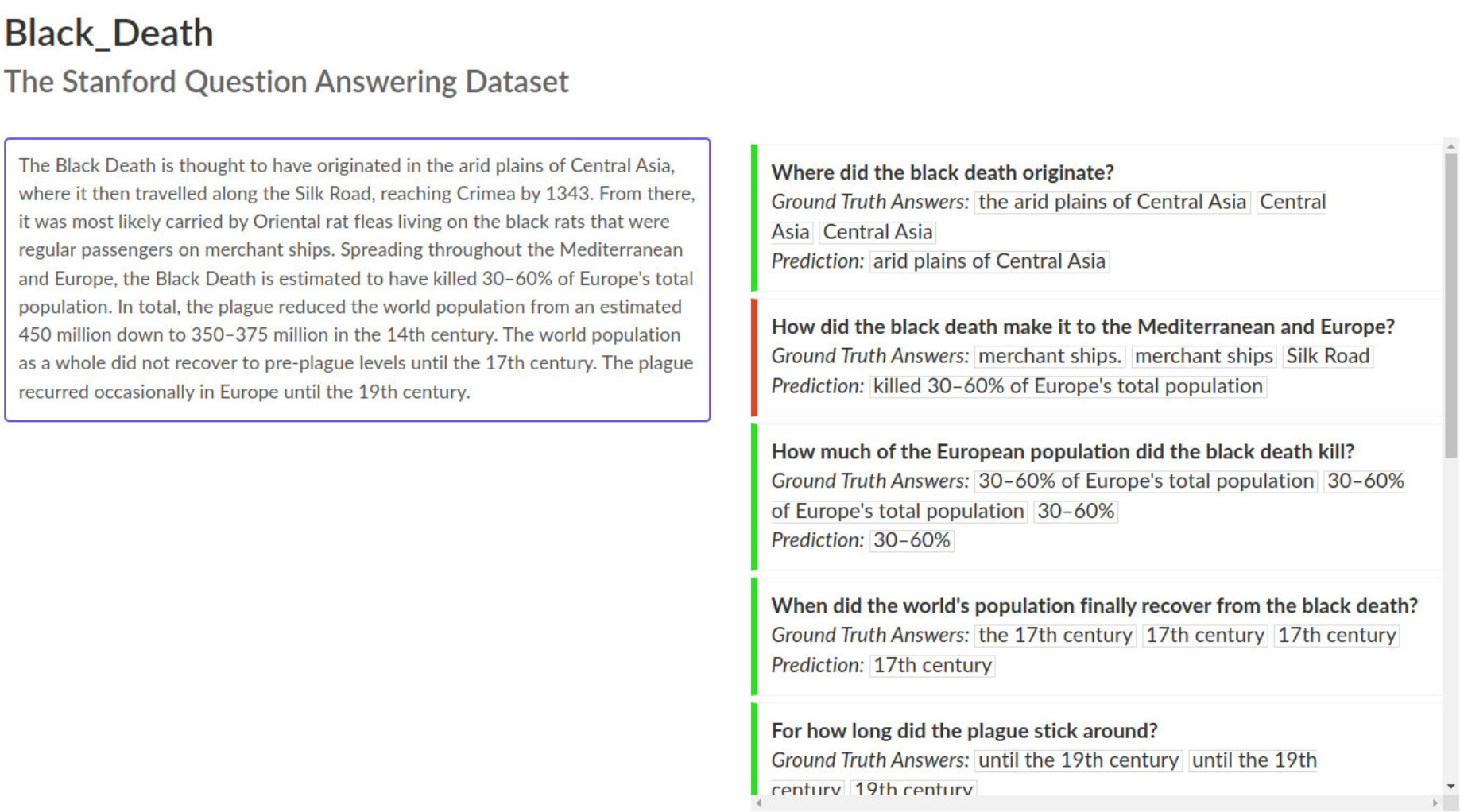
- Odwrotny indeks: odzworowanie term $\rightarrow$ zbiór-dokumentów-zawierających-term
- **termem** może być słowo, ale dla języka polskiego lepszy jest **lemat**.
- Heurystycznie ograniczamy liczbę obliczonych cosinusów (tylko dokumenty zawierające Ważne Termy z Zapytania (wszystkie? co najmniej 1?)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

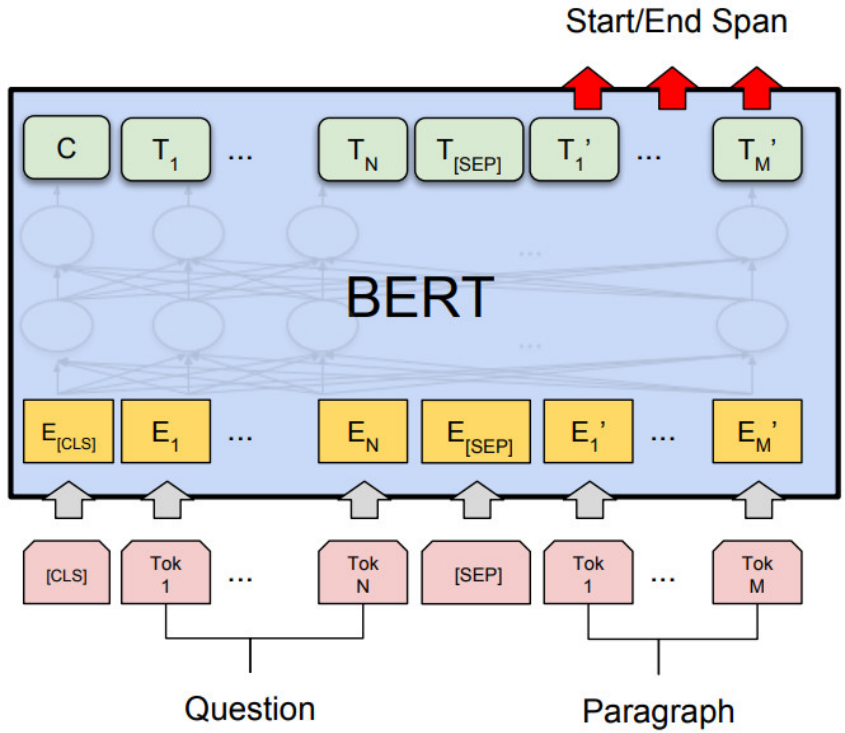
How much of the European population did the black death kill?
Ground Truth Answers: 30–60% of Europe's total population | 30–60% of Europe's total population | 30–60%
Prediction: 30–60%

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

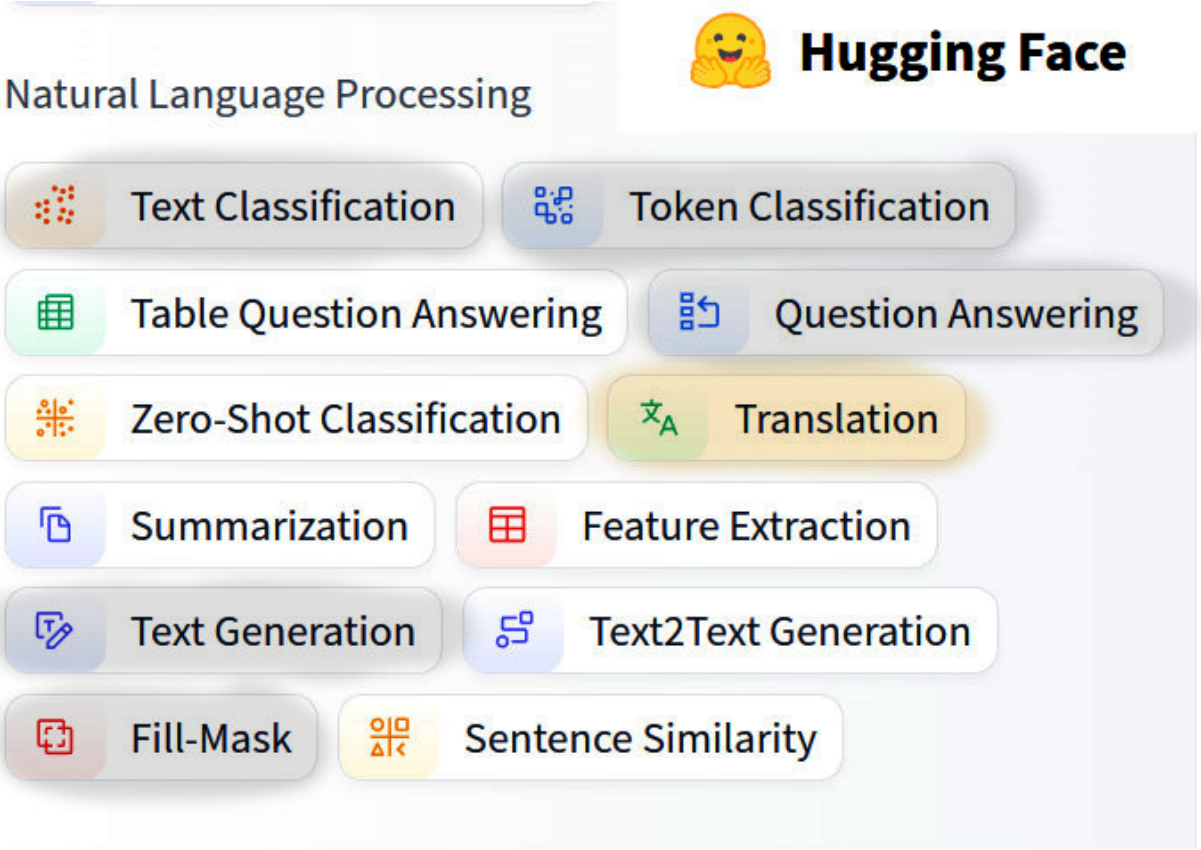
Reader (cd)

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.

NLP w HuggingFace



Tłumaczenie maszynowe

Zadanie

Dla tekstu **x** z języka źródłowego znajdź tekst **y** z języka docelowego, jak najdokładniej oddający jego znaczenie, styl, etc.

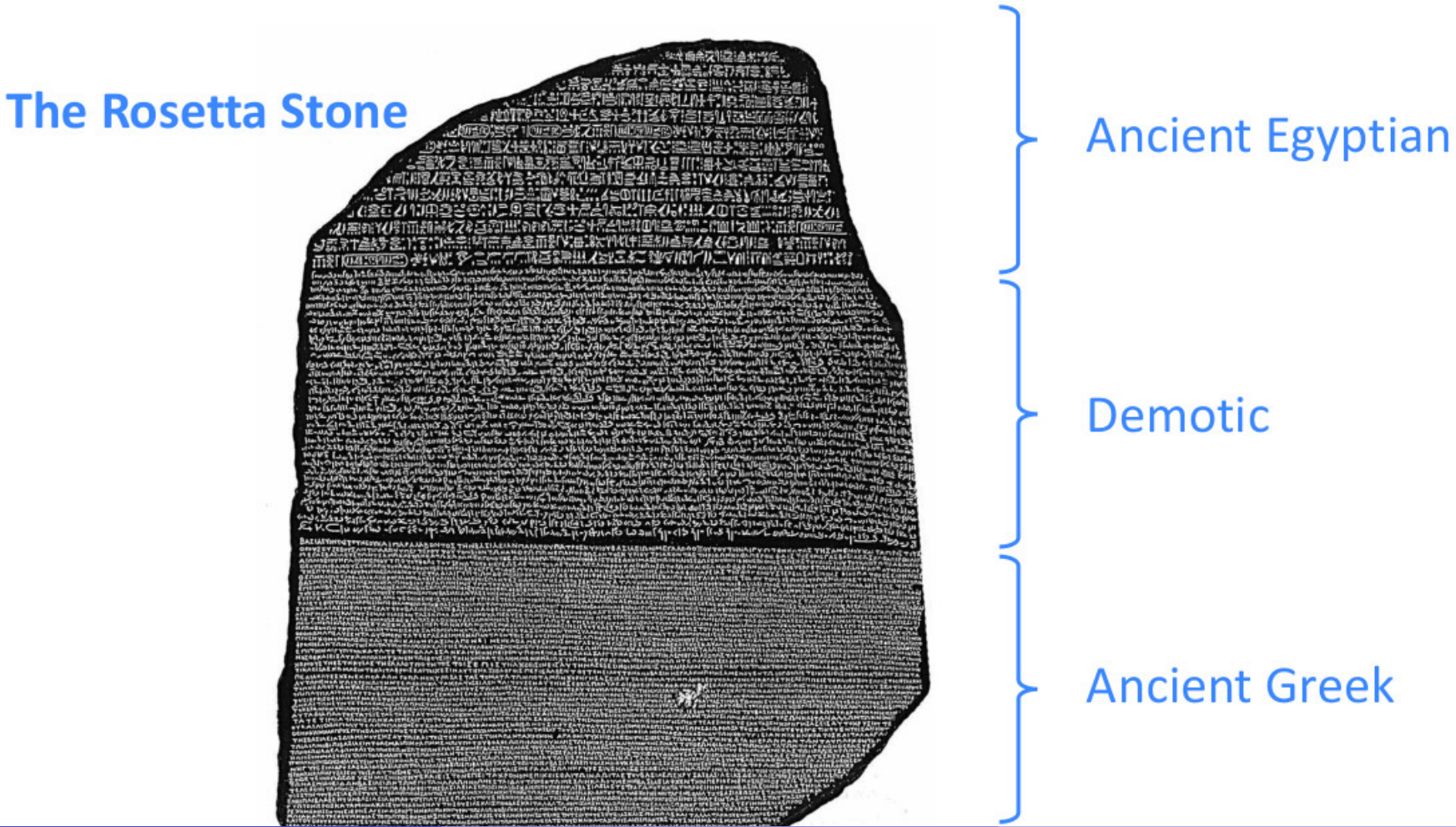
x: *L'homme est né libre, et partout il est dans les fers*

y: *Man is born free, but everywhere he is in chains*

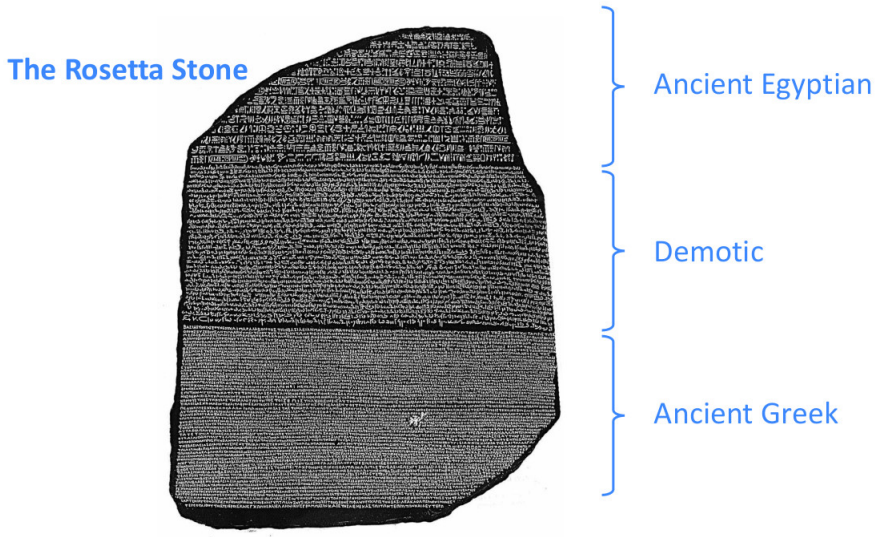
Pierwszy korpus równoległy

Definicja

Korpus równoległy zawiera fragmenty tekstu w dwóch językach, będące swoimi tłumaczeniami



Pierwszy korpus równoległy



Jego treść stanowi dekret wydany 27 marca roku 196 p.n.e. przez kapłanów egipskich w Memfis dla uczczenia faraona Ptolemeusza V z okazji pierwszej rocznicy koronacji, w związku z doznanymi od niego dobrodziejstwami. Faraon po wstąpieniu na tron ogłosił amnestię, obniżył podatki i podniósł dochody kapłanów[2][5].

Trzy języki, w tym egipskie hieroglify!

Historia tłumaczenia maszynowego

- 1950+: Systemy regułowe, gramatyki, automaty, prawdopodobieństwo. Działały różnie (najsłynniejsza anegdota poniżej):
  - Angielski: the spirit is strong but the flesh is weak
  - Polski (w oryginalnej anegdocie rosyjski): wódka jest mocna, ale mięso się zepsuło
- 1990+: Systemy bazujące na statystyce (z korpusów), ukryte łańcuchy Markowa
- 2014+: Dominacja sieci neuronowych
  - Najpierw modele seq2seq, głównie LSTM
  - Potem z dodanym mechanizmem uwagi

Historia tłumaczenia maszynowego

- 2017 Word2Vec (i inne osadzenia) wielojęzyczne (czyli vec('queen') leży blisko vec('królowa'))
  - Praca: Word translation without parallel data
- 2018: Próby tłumaczenia bez nadzoru, całkiem udane. Trening uwzględniający pięć więzów (lub ich podzbiór):
  1. Złożenie pol-ang i ang-pol to identyczność (w drugą stronę również)
  2. pol-ang produkuje sensowne angielskie teksty
  3. ang-pol produkuje sensowne polskie teksty
  4. Zarówno pol-ang, jak i ang-pol zachowują słownictwo (co możemy sprawdzić dzięki dwujęzycznym osadzeniom)
- Praca: Unsupervised Machine Translation Using Monolingual Corpora Only, ICLR 2018

Historia tłumaczenia maszynowego

Transformery (2017) pojawiły się jako rozwiązanie zadania tłumaczenia:

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

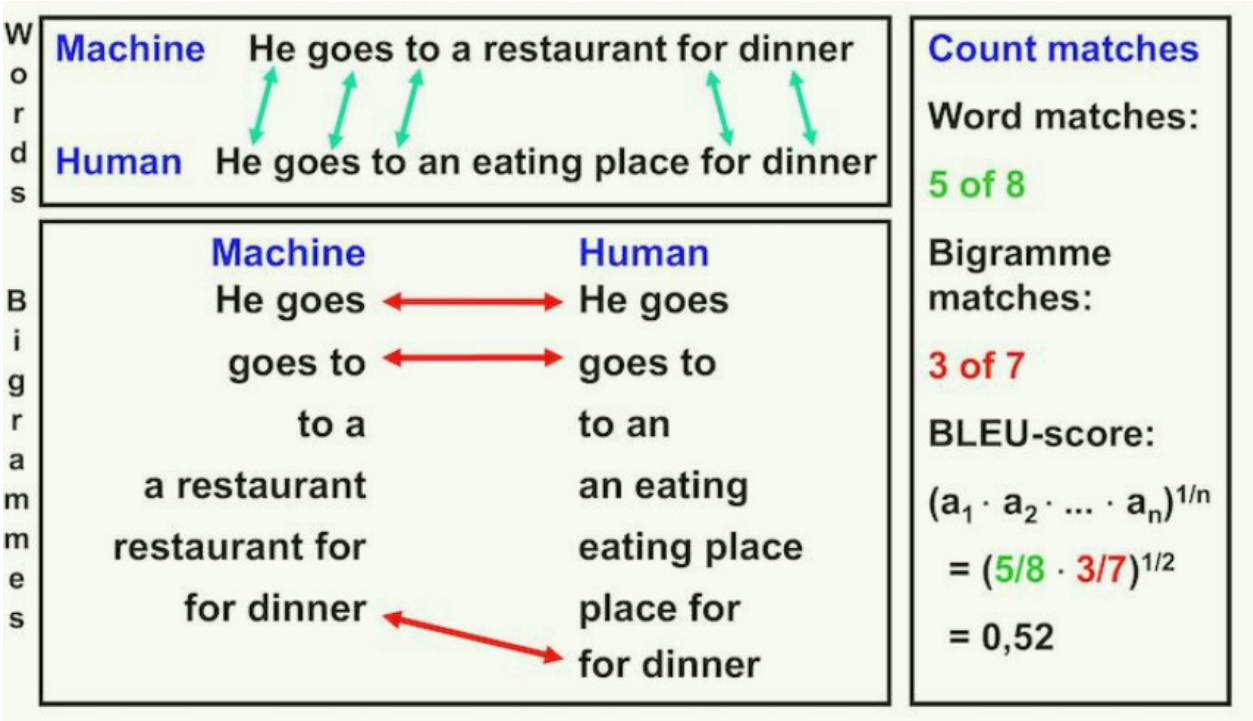
Duże modele językowe (ChatGPT, 2022) umieją tłumaczyć teksty, mimo braku dostępu do dużych korpusów równoległych. Zastanówny się dlaczego (bez wsparcia w slajdach).

BLEU

Uwaga

Takie zadania jak tłumaczenia maszynowe wymagają bardziej zaawansowanych metryk (bo jest wiele *dobrych* tłumaczeń zdania).

Jedną z metryk jest BLEU (Bilingual Evaluation Understudy)



Jak mamy więcej wzorców tłumaczenia, to bierzemy maksimum wartości BLEU dla każdego wzorca.

Wykorzystanie systemów tłumaczących do augmentacji danych

- Można tłumaczyć dane z innych języków
- Można tłumaczyć *tam-i-z-powrotem*
- Można mieszać różne systemy tłumaczące

Uwaga

Tłumaczenie nie muszą być idealne, żeby były użyteczne (inaczej niż w generowaniu na przykład systemów dialogowych). I tak ostateczne patrzymy na osadzenia kontekstowe

Streszczanie

Co do zasady: można **streszczanie** potraktować dokładnie jak tłumaczenie, z danymi tego samego typu (korpus równoległy), z tymi samymi metrykami (BLEU, ROUGE, ...)

- Całkiej dobrze działa w scenariuszu 'zero shot': do tekstu doklejamy frazę streszczającą typu tl;dr, Podsumowując w kilku słowach:, let us summarize it:,
- Tradycyjnie był podział na dwa rodzaje streszczania:
  - **Ekstraktywne**: zaznacz istotne fragmenty
  - **Generatywne**: wygeneruj streszczenie
- Drugie może robić autoregresywny model językowy, pierwsze sieć typu BERT.

Oczywista uwaga na koniec

Model streszczający może posłużyć do augmentacji danych (wszak streszczenie nie zmienia wydźwięku, tematyki, ...)


KLEJ
GUMA
ARABSKA
Poj. 50 ml


SUPER GLUE


UHU
Power
all purpose adhesive
trans-parent

SuperGLUE

GLUE

KLEJ

Sposoby ewaluacji modeli językowych

Ogólna zasada

- Mamy zbiór zadań, zwykle wcześniej istniejących związanych z NLP.
- Każde zadanie ma swój zbiór uczący i testowy.
- **Wstępnie wytrenowany model językowy** jest dostrajany na danych uczących i testowany na danych testowych.

- **GLUE** == **G**eneral **L**anguage **U**nderstanding **E**valuation
- **KLEJ** == **K**ompleksowa **L**ista **E**waluacji **J**ęzykowych
- Jest też **Super-GLUE**, wprowadzony, gdy zwykły GLUE przestał wystarczać

Inny benchmark dla języka polskiego

Polish linguistic and cultural competency benchmark

`https://huggingface.co/spaces/sdadas/plcc`

Pytań nie jest dużo, nie należy przykładać zbyt wielkiej uwagi do niewielkich różnic między modelami (bo mogą być wynikiem losowych fluktuacji)