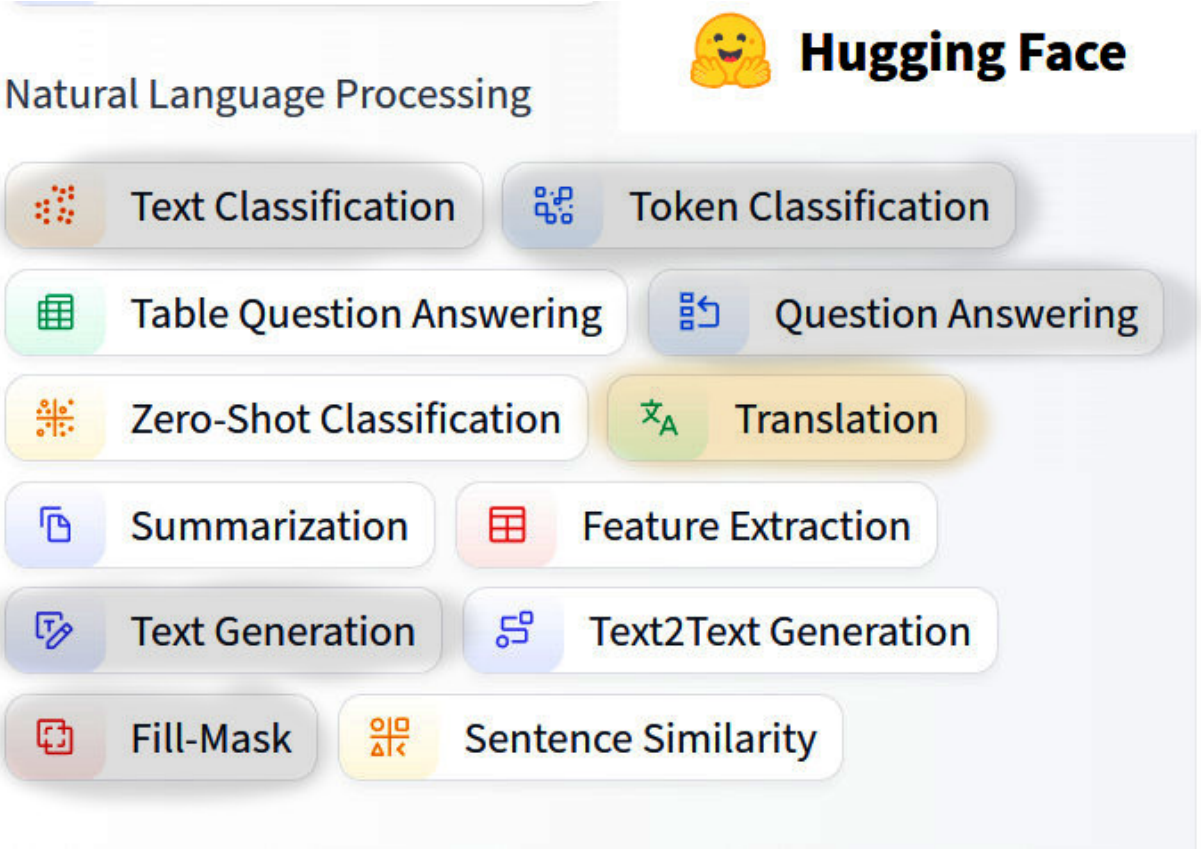
Tłumaczenie, streszczanie, klejenie oraz transformery

Paweł Rychlikowski

Instytut Informatyki UWr

4 grudnia 2024

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.

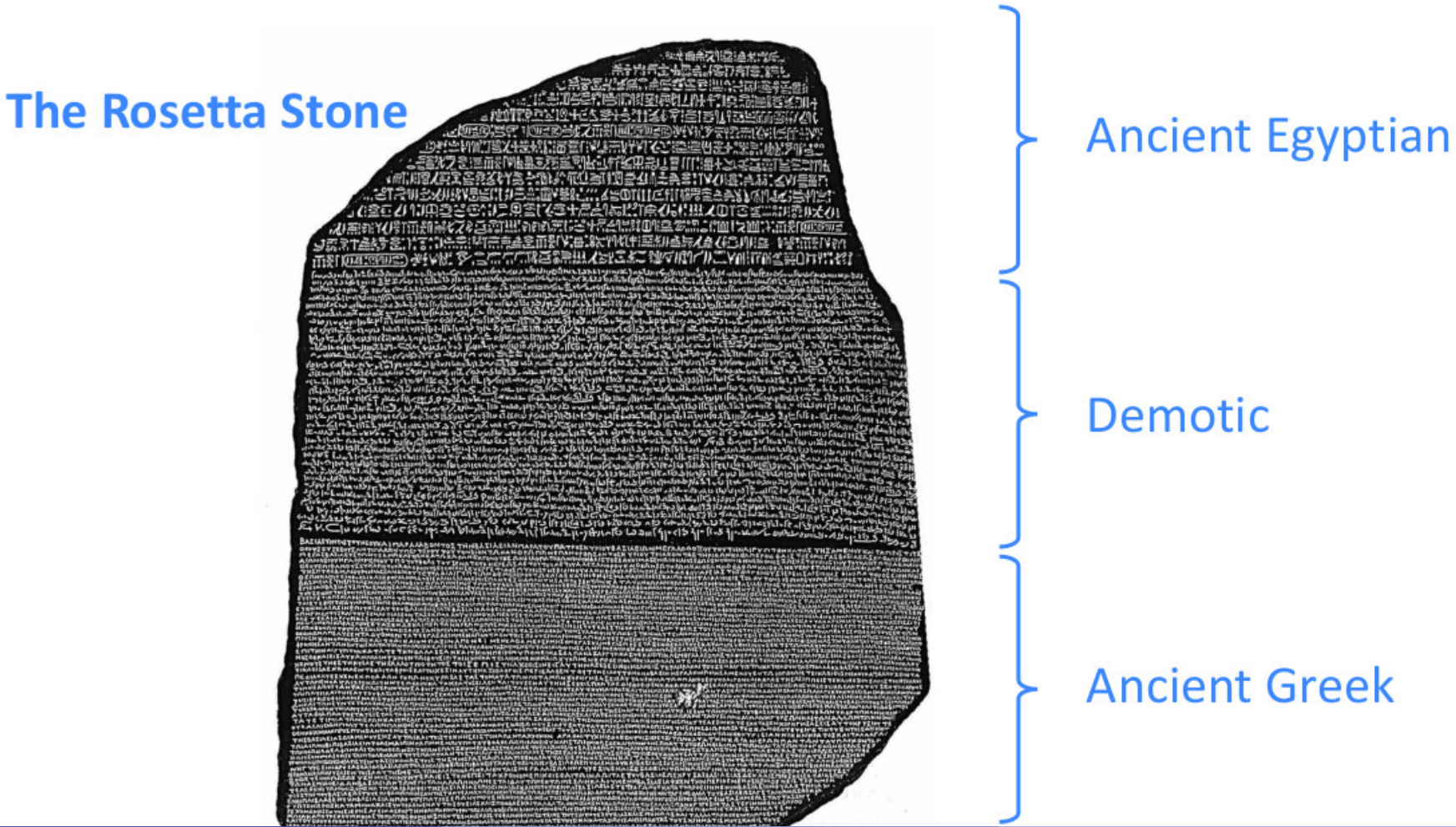
x: *L'homme est né libre, et partout il est dans les fers*

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

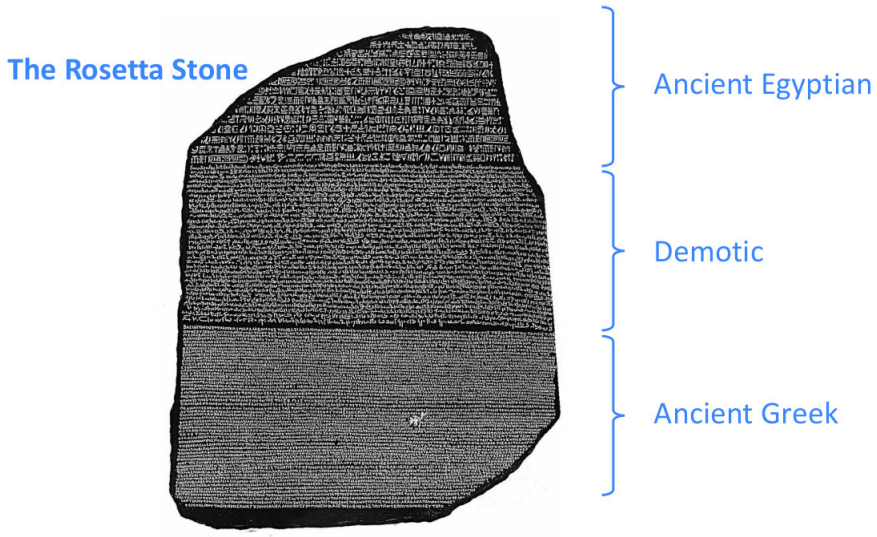
Pierwszy korpus równoległy

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**	$2.3 \cdot 10^{19}$	

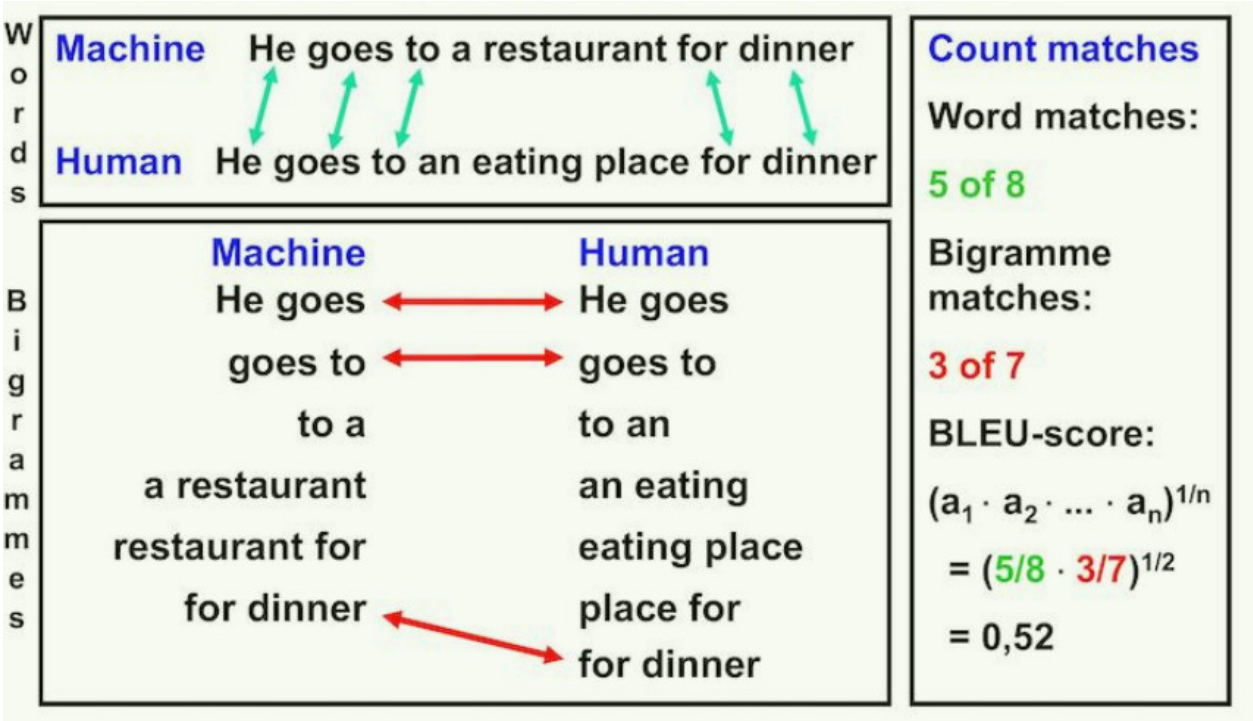
Duże modele językowe (ChatGPT, 2022) umieją tłumaczyć teksty, mimo braku dostępu do dużych korpusów równoległych. Zastanówny się dlaczego (bez wsparcia w slajdach).

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

Zadania z GLUE (1)

Zadania klasyfikacji pojedynczego zdania

- **CoLA** – The Corpus of Linguistic Acceptability (Warstadt et al., 2018
  - Czy zdanie jest poprawnym angielskiem zdaniem, czy nie (złe zdania z książek lingwistycznych/do nauki języka)
- **SST-2** – The Stanford Sentiment Treebank (Socher et al., 2013)
  - Recenzje filmów (pozytywne/negatywne)

Zadania z GLUE (2)

Zadania z GLUE działające dla pary zdań

- **MRPC** The Microsoft Research Paraphrase Corpus (Dolan & Brockett, 2005)
  - Czy dwa zdania są sobie równoważne (przykłady negatywne mają podobne słownictwo, TF-IDF nie wystarcza)
- **QQP** The Quora Question Pairs
  - Czy dwa pytania są równoważne
- **STS-B** The Semantic Textual Similarity Benchmark (Cer et al., 2017)
  - Czy zdania są podobne, ludzie oceniali od 1 do 5, zadanie regresji
- **MNLI** The Multi-Genre Natural Language Inference Corpus
  - Relacje między parą zdań: wynikanie, sprzeczność, brak związku
- **RTE** The Recognizing Textual Entailment (RTE)
  - Coś jak MNLI

Zadania z GLUE (3)

Zadania z GLUE działające dla pary zdań

- **QNLI** – nowe zadanie bazujące na SQUAD
  - Czy zdanie zawiera odpowiedź na pytania?
- **WNLI** The Winograd Schema Challenge (Levesque et al., 2011)
  - Ciekawe zadanie, które zasługuje na osobny slajd

Winograd Schema

Mamy zdanie z zaimkiem (it, on, she, ...) i należy podać, co w tym kontekście oznacza zaimek.

Przykłady

The city councilmen refused the demonstrators a permit because they **[feared/advocated]** violence.
Puchar nie zmieści się w czerwonym kufrze, ponieważ jest on za **[duży/mały]**

Oczywiście odpowiedź na pytanie, kim są **they** (i co jest za duże bądź za małe), zależy od wyboru wyróżnionego słowa

Historia GLUE

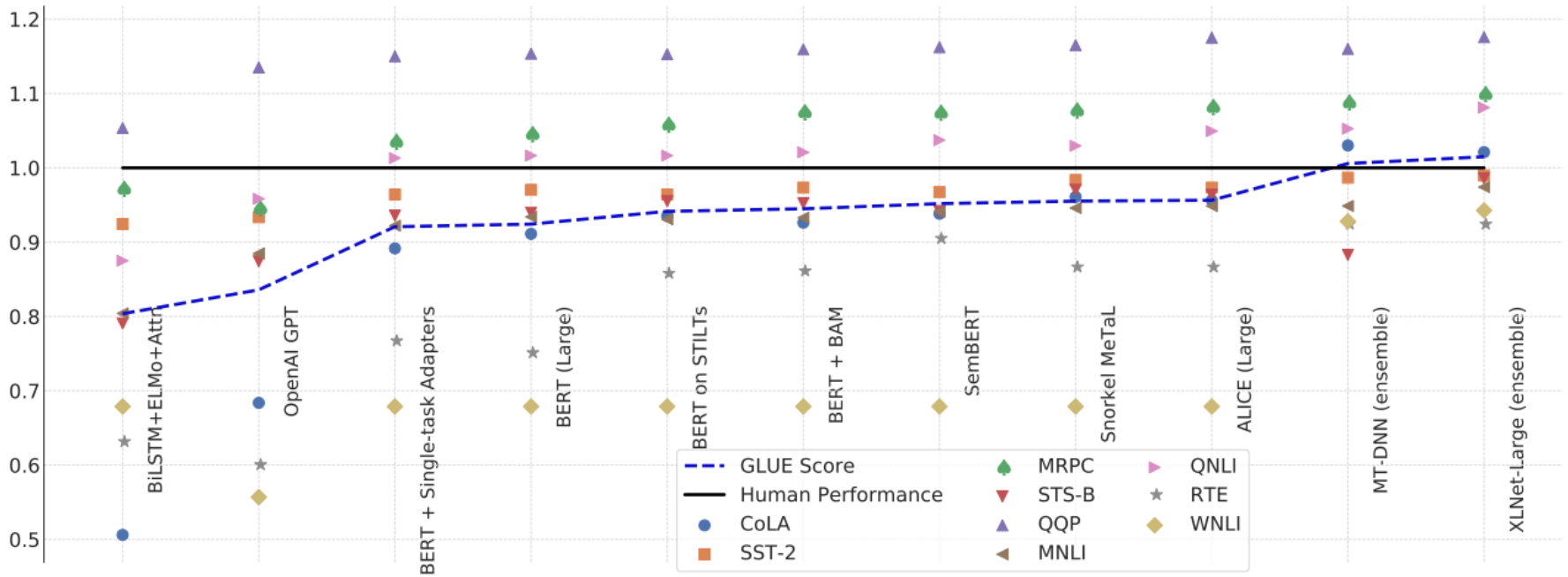


Figure 1: GLUE benchmark performance for submitted systems, rescaled to set human performance to 1.0, shown as a single number score, and broken down into the nine constituent task performances. For tasks with multiple metrics, we use an average of the metrics. More information on the tasks included in GLUE can be found in Wang et al. (2019a) and in Warstadt et al. (2018, CoLA), Socher et al. (2013, SST-2), Dolan and Brockett (2005, MRPC), Cer et al. (2017, STS-B), and Williams et al. (2018, MNLI), and Rajpurkar et al. (2016, the original data source for QNLI).

Notka filozoficzna

Pytanie

Czy model językowy rozumie to co pisze? Jak to sprawdzić?

- Model językowy może stwarzać wrażenie, że rozumie text (jak on dobrze uchwycił sedno tego artykułu!)
- Oczywiście każda kompromitująca pomyłka pokazuje, że nie jest to prawdziwe rozumienie


Atttenttion
is All YouNeed

Dodawanie wektorów zanurzeń (slajd do ewentualnego usunięcia)

Przeanalizujmy działanie zapytania:

jabłko + *porucznik* + *matematyka* + *łyżka*

Wyniki iloczynów skalarnych z wektorami kąpiel, **algebra**, **gruszka**, wątroba.

Dodawanie wektorów działa trochę jak obliczanie zbioru.

Mechanizm uwagi (level: novice)

- Rozważamy zdanie:

  Grażyna i Janusz swoim samochodem z wysokoprężnym silnikiem jechali koło słoni, żyraf i bawołów w Parku Narodowym Serengeti.

- Wyrazy o tym samym kolorze są powiązane:

  Grażyna i Janusz swoim samochodem z wysokoprężnym silnikiem jechali koło słoni, żyraf i bawołów w Parku Narodowym Serengeti.

- Wyznaczamy podobieństwa każdy z każdym, do każdego słowa *domieszowujemy* słowa podobne (tak naprawdę to dodajemy wszystkie, ale z wagą zależną od podobieństwa)

Mechanizm uwagi (level: beginner)

- Mamy dwa rodzaje wektorów: słowa i konteksty
- To jak dobrze dane słowo czuje się w czyimś towarzystwie zależy od iloczynu osadzenia tego słowa z osadzeniem słowa kontekstowego.
- Być może do ustalania poziomu domieszek warto użyć różnych rodzajów wektorów?

Uwaga

Moglibyśmy trzymać dwa rodzaje osadzeń w jednym wektorze i *wyjmować* je za pomocą odpowiedniej projekcji.

Transformery. Wersja autoregresywna

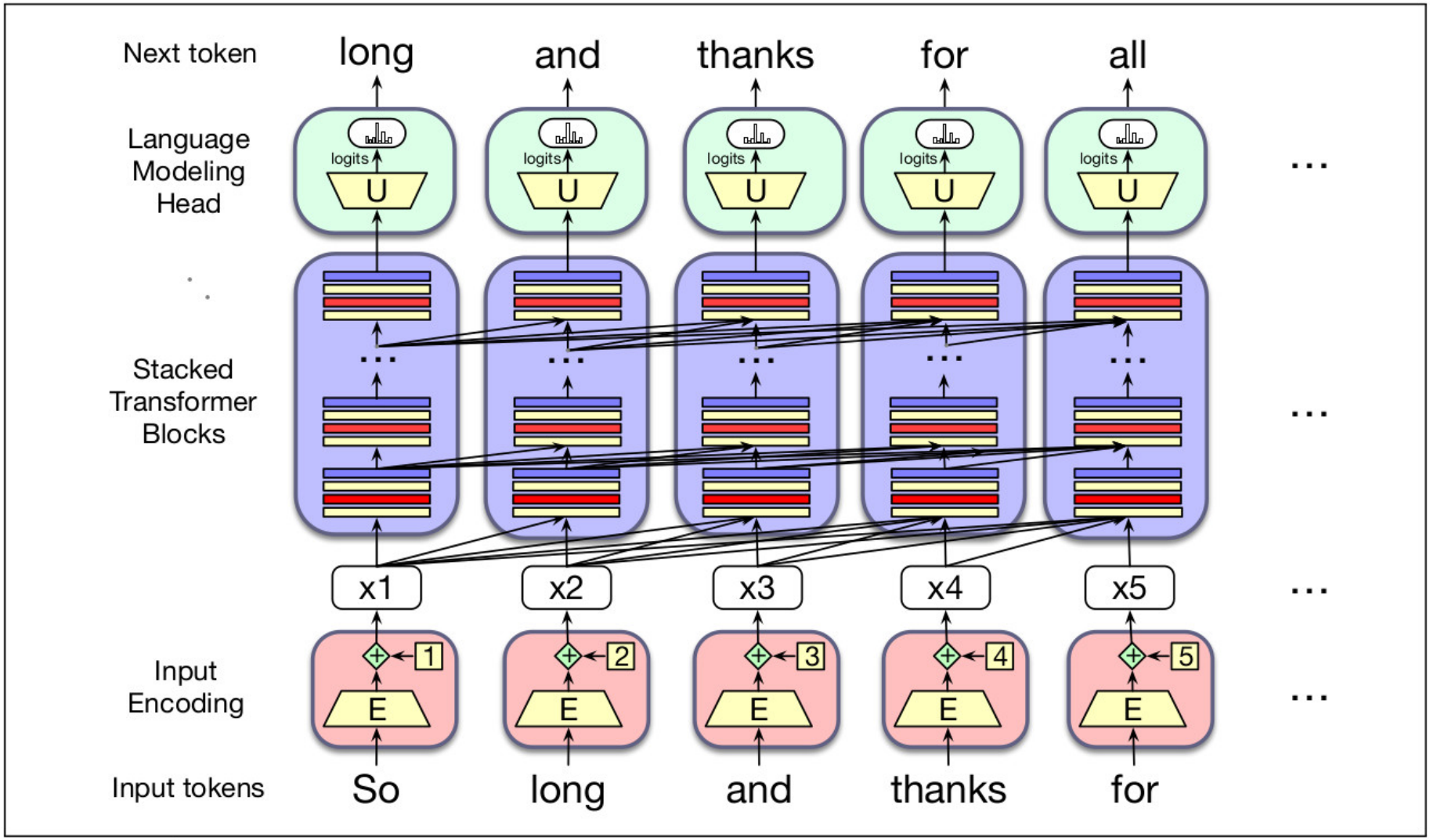
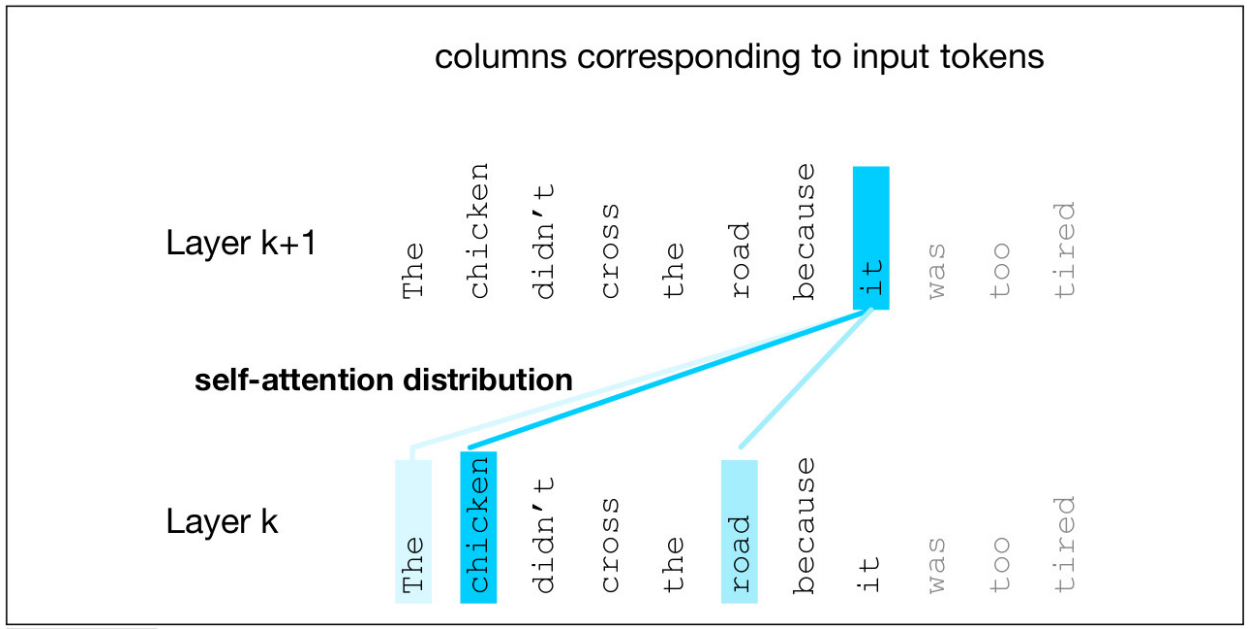


Figure 9.1 The architecture of a (left-to-right) transformer, showing how each input token get encoded, passed through a set of stacked transformer blocks, and then a language model head that predicts the next token.

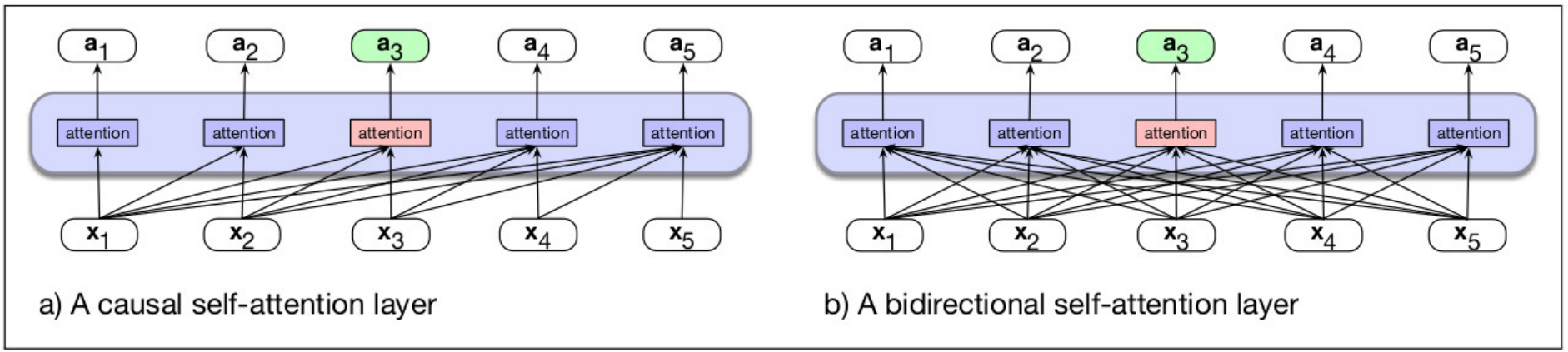
Transformery. Mechanizm uwagi (1)

(9.1) **The chicken** didn't cross the road because **it** was too tired.
(9.2) **The chicken** didn't cross the road because **it** was too wide.

Osadzenie słowa **it** powinno zależeć od wcześniejszych słów (w BERT-cie może również zależeć od późniejszych!)



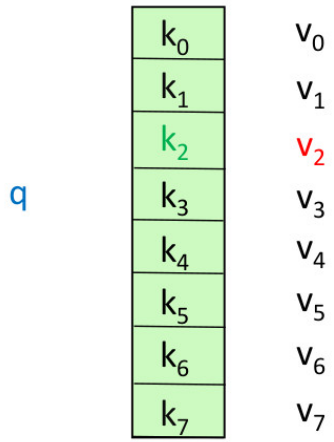
Transformery. Mechanizm uwagi (2)

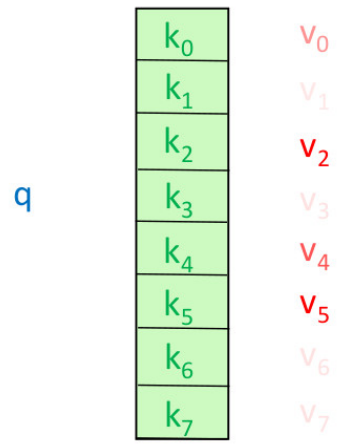


- Obrazek lewy to sposób działania atencji wg schematu **GPT**
- Obrazek prawy to sposób działania atencji wg schematu **BERT**

Transformery. Mechanizm uwagi (3)

- Z naszego wektora osadzenia będziemy wydobywać trzy aspekty (role, projekcje): zapytania, klucza i wartości.
- Nazwy zapytanie, klucz i wartość tłumaczy poniższa (obrazkowa) analogia:
  - Let's think of attention as a "fuzzy" or approximate hashtable:
    - To look up a **value**, we compare a **query** against **keys** in a table.
    - In a hashtable (shown on the bottom left):
      - Each **query** (hash) maps to exactly one **key-value** pair.
    - In (self-)attention (shown on the bottom right):
      - Each **query** matches each **key** to varying degrees.
      - We return a sum of **values** weighted by the **query-key** match.





28

Transformery. Mechanizm uwagi (4)

- Wydobywanie roli z poprzedniego slajdu to mnożenie przez macierz. Potrzebujemy trzech macierzy: W^Q, W^V i W^K.
- d jest wymiarem wektorów (wartości, osadzeń), natomiast d_k to wymiar wektorów klucza (i zapytania). Czasem $d = d_k$
- Dzielenie przez $\sqrt{d_k}$ ma związek z redukcją wariancji (z d_k do 1, to będzie bardziej jasne, jak zobaczymy całość obliczeń)

$$\mathbf{q}_i = \mathbf{x}_i\mathbf{W}^{\mathbf{Q}};\ \ \mathbf{k}_j = \mathbf{x}_j\mathbf{W}^{\mathbf{K}};\ \ \mathbf{v}_j = \mathbf{x}_j\mathbf{W}^{\mathbf{V}}$$

$$\text{score}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{q}_i \cdot \mathbf{k}_j}{\sqrt{d_k}}$$

$$\alpha_{ij} = \text{softmax}(\text{score}(\mathbf{x}_i, \mathbf{x}_j))\ \ \forall j \leq i$$

$$\mathbf{a}_i = \sum_{j \leq i} \alpha_{ij}\mathbf{v}_j$$