

Metódy transformérov a ich aplikácia v oblasti textovej analýzy

Obhajoba diplomovej práce
Martin Bača

Vedúci práce: RNDr. Šimon Horvát, PhD.
Konzultant práce: doc. RNDr. Ľubomír Antoni, PhD.



Prírodovedecká fakulta
Univerzita Pavla Jozefa Šafárika v Košiciach

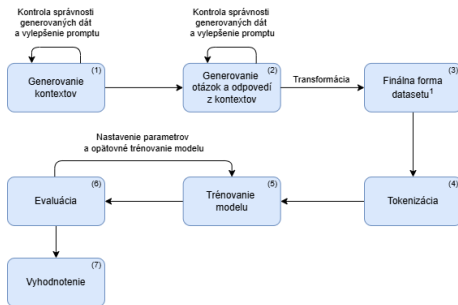
Košice, 24. apríl 2025

- Question Answering sa stáva kľúčovou súčasťou moderných AI systémov
- Prudké zlepšenie a vývoj jazykových modelov
- Väčšina výskumu v oblasti QA sa opiera o veľké existujúce datasety, ktoré majú často len jednojazyčný charakter alebo riešia len jeden typ otázok
- Skúmanie schopností QA systémov v odpovedaní na boolean otázky

- 1 Vygenerovať syntetický QA dataset z JSON štrukturovaných náhodne vygenerovaných údajov.
- 2 Implementovať tokenizáciu a predspracovanie textov.
- 3 Trénovať a vyhodnotiť extraktívne a boolean QA modely.
- 4 Porovnať výsledky monolingválnej a multilingválnej úlohy s existujúcimi štúdiami.

Postup riešenia

- Generovanie datasetu (kontext – otázky – odpovede)
- Predspracovanie a tokenizácia
- Trénovanie QA modelov (BERTa based a XLM-R based)
- Vyhodnotenie úspešnosti modelov pre rôzne typy otázok



Obr. 1: Diagram postupu riešenia (vlastný obrázok)

- Náhodne vygenerované informácie o objednávkach na prepravu štrktúrovaných vo forme JSON.

```
"pickup": {
  "company_name": "COMPANY_PICKUP",
  "street_name": "STREET_PICKUP",
  "city": "CITY_PICKUP",
  "postal_code": "ZIP_PICKUP",
  "country": "COUNTRY_PICKUP",
  "datetime": "between 2021-03-04 13:30:00 and 2021-03-13 05:15:00"
},
"delivery": {
  "company_name": "COMPANY_DELIVERY",
  "street_name": "STREET_DELIVERY",
  "city": "CITY_DELIVERY",
  "postal_code": "ZIP_DELIVERY",
  "country": "COUNTRY_DELIVERY",
  "datetime": "between 2021-03-17 13:45:00 and 2021-03-24 18:45:00"
},
"goods": {
  "weight": "92 kg",
  "dimensions": [
    {
      "length": "96m",
      "width": "85m",
      "height": "12m",
      "weight": "11 kg"
    },
    {
      "length": "86m",
      "width": "48m",
      "height": "78m",
      "weight": "12 kg"
    },
    {
      "length": "46m",
      "width": "58m",
      "height": "19m",
      "weight": "69 kg"
    }
  ]
},
"stackable": "no",
"required vehicle": "Small lorry",
"special (vehicle) request": "Sideboards required"
```

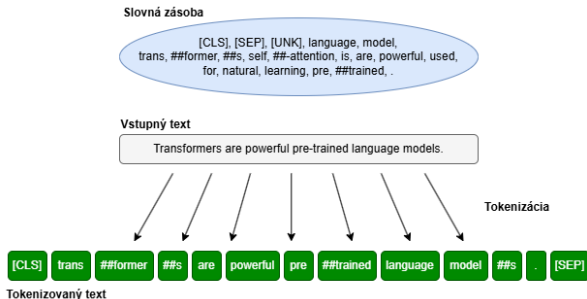
Tvorba datasetu pomocou jazykových modelov

- Generovanie human-like kontextov z JSON dát
- Generovanie otázok a odpovedí ku kontextom
- Výstup: multilingválny dataset obsahujúci dve kategórie otázok - extraktívne a boolean

```
"context": "Sure! Here's the natural language text based on the provided JSON  
-structured transportation order:\n\nWe have a transportation order for  
you! We need to pick up some goods from J.C. Penney at Maple 8904 in  
Chicago, 38250, Bahrain between 2021-03-04 13:30:00 and 2021-03-13  
05:15:00. We'll be using a Small lorry to transport these goods, which  
include three packages with the following dimensions:\n\nPackage 1:  
Length 96m, Width 85m, Height 12m, Weight 11kg\n\nPackage 2: Length 86m,  
Width 48m, Height 78m, Weight 12kg\n\nPackage 3: Length 46m, Width 58m,  
Height 19m, Weight 69kg\n\nWe'll deliver these goods to Qantas at Redwood  
3525 in Memphis, 95769, Georgia between 2021-03-17 13:45:00 and  
2021-03-24 18:45:00. We've been requested to provide sideboards for the  
delivery, so please make sure to include those in the transportation plan  
\n\nThe total weight of these goods is 92kg, and we'll need a Small  
lorry to transport them safely. Please let us know if you have any  
questions or concerns about this order.\n\nI hope this text meets your  
needs! Let me know if there's anything else I can do for you.",  
"qas": [  
  {  
    "id": 2,  
    "question": "What type of vehicle is needed to transport the goods  
safely?",  
    "lang": "en",  
    "answers": [  
      {  
        "text": "a Small lorry",  
        "answer_type": "extractive",  
        "answer_start": 297  
      }  
    ],  
  },  
  {  
    "id": 8,  
    "question": "Where do we need to pick up the goods?",  
    "lang": "en",  
    "answers": [  
      {  
        "text": "J.C. Penney at Maple 8904 in Chicago, 38250, Bahrain",  
        "answer_type": "extractive",  
        "answer_start": 176  
      }  
    ],  
  },  
]
```

Tokenizácia

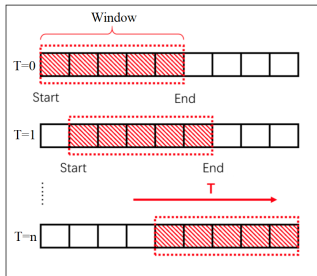
- Proces rozdeľovania textu na menšie podjednotky nazývané tokeny
- Tokeny môžu byť znaky, podslová, slová alebo celé vety.
- Väčšina modelov používa tokenizáciu na báze slov alebo podslov



Obr. 2: Proces tokenizácie (vlastný obrázok)

Tokenizácia

- Pri dlhých textoch sme narazili na limit vstupnej dĺžky
- Použili sme techniku sliding window, ktorá vytvára prekrývajúce sa segmenty
- Tento prístup maximalizuje pravdepodobnosť, že správna odpoveď sa bude nachádzať aspoň v jednom segmente celá



Obr. 3: Sliding window prístup

Rozdelenie na tréningovú a testovaciu vzorku

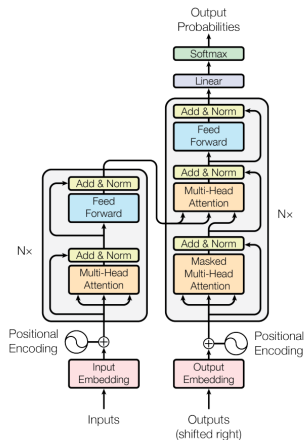
- Dataset sme rozdelili v pomere 80/20
- Typy otázok sú náhodne rozdelené medzi tréningovú a testovaciu množinu

- Model na vstup dostáva tokenizovanú otázku a kontext oddelené špeciálnymi tokenmi

[CLS] Otázka [SEP] Kontext [SEP]

- [CLS] token sa pridáva na začiatok vstupu, jeho výstupná reprezentácia sa používa na klasifikačné úlohy
- [SEP] token slúži na oddelenie rôznych častí vstupu

Transformer



Výpočet Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V.$$

Hlavy pozronosti

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W^O,$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V).$$

Pozičné kódovanie

$$PE_{(pos, 2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right),$$

$$PE_{(pos, 2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right),$$

Obr. 4: Transformer

- Extrahujeme anglický subset vygenerovaného datasetu
- Len otázky extraktívneho charakteru
- cca 4200 otázok distribuovaných medzi 500 kontextov

Použili sme modely založené na architektúre BERT s rôznou veľkosťou

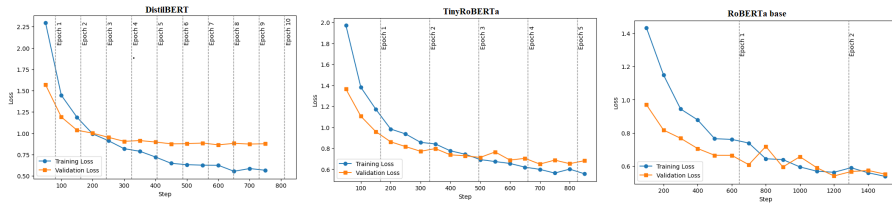
- DistilBERT - 66,4M parametrov
- TinyRoBERTa - 81,5M parametrov
- RoBERTa base - 125M parametrov

Zvolili sme vhodné tréningové parametre pre malý dataset - aby sme zabránili overfittingu

- Malý batch-size: 8-16
- Menšie hodnoty učiaceho pomeru: $2e-5$ – $5e-5$
- Regularizácia váh: 0.01 - 0.1
- Warmup ratio: 10% - stabilizuje učenie na začiatku
- Vyšší dropout: 0.1 - 0.3 (default 0.1)
- Zmrazené vrstvy: 1 - 3 (v závislosti od veľkosti modelu)

Model	EM	F1	Precision	Recall
DistilBERT	0.6282	0.7582	0.8296	0.7985
TinyRoBERTa	0.6628	0.7903	0.8435	0.8086
RoBERTa base	0.6842	0.8021	0.8525	0.8200

Tabuľka 1: Porovnanie metrík modelov



Obr. 5: Grafy loss funkcie

- Berieme celý dataset (anglické, nemecké a francúzske)
- Jazyky kontextov a otázok sú rôzne kombinované (napr. ang. kontext - nem. otázka)
- cca 80 000 otázok rovnomerne distribuovaných do dvoch kategórií: extraktívne a boolean

Použili sme modely založené na architektúre X-LMR s rôznou veľkosťou

- XtremeDistill - 33M parametrov
- XLM-RoBERTa Base - 279M parametrov
- XLM-RoBERTa Large - 560M parametrov

Zvolili sme vhodné tréningové parametre pre veľký dataset

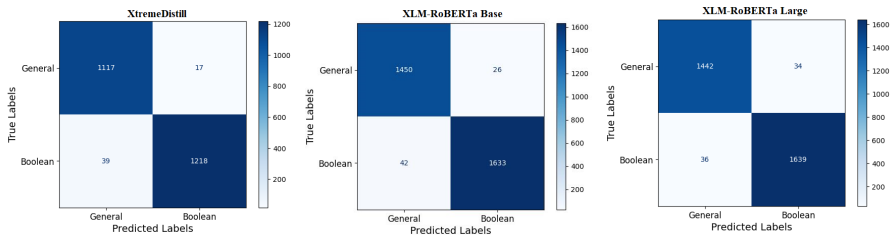
- Malý batch-size: 8-16
- Menšie hodnoty učiaceho pomeru: $4e-6$ – $1e-5$
- Regularizácia váh: 0.01 - 0.05
- Warmup ratio: 10% - stabilizuje učenie na začiatku
- Vyšší dropout: 0.1 - 0.15 (default 0.1)
- Zmrazené vrstvy: žiadne

Model	EM	F1	Prec.	Rec.
XtremeDistill	0.5727	0.7294	0.7419	0.8013
XLM-RoBERTa Base	0.7869	0.8734	0.8847	0.8942
XLM-RoBERTa Large	0.7889	0.8775	0.8931	0.8971

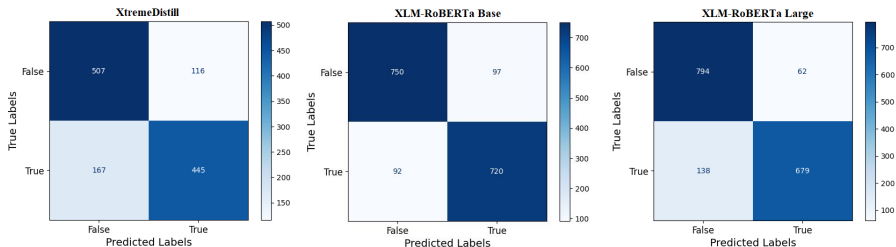
Tabuľka 2: Výsledky modelov na extraktívnych QA otázkach

Model	Acc.	F1	Prec.	Rec.
XtremeDistill	0.7709	0.7703	0.7727	0.7705
XLM-RoBERTa Base	0.8861	0.8860	0.8860	0.8861
XLM-RoBERTa Large	0.8805	0.8799	0.8841	0.8794

Tabuľka 3: Výsledky modelov na boolean QA otázkach



Obr. 6: Matice zámien pre kategorizáciu boolean a extractive otázok



Obr. 7: Matice zámeny pre boolean QA

Monolingválna úloha:

- Videli sme, že distilovaný model polovičnej veľkosti mal len o 6% horšie výsledky ako plný base model.
- TinyRoBERTa mala len o 2% horšie výsledky ako RoBERTa base, aj napriek tretinovej redukcii počtu parametrov.

Multilingválna úloha:

- Vyskúšali sme extrémne distilovanú verziu X-LMR, ktorá si s úlohou poradila celkom dobre, vzhľadom na zlomok počtu parametrov z väčších modelov.
- RoBERTa base a RoBERTa large mali skoro totožné výsledky - väčší model sa nedokázal naučiť viac vzťahov, ani s dvojnásobným počtom parametrov

Aplikácie metód spracovania prirodzeného jazyka a využitie Transformerov má veľký potenciál, hlavne vzhľadom na prudký vzostup jazykových modelov.

Ďalšia práca:

- Experimentovať s predspracovaním dát
- Zahnúť do generovania textov a otázok viac jazykov
- Porovnať jazykové rodiny (románske, germánske,...) - ako dobre si modely dokážu poradiť s ktorou jazykovou rodinou

- <https://api.time.com/wp-content/uploads/2022/05/Google-Search-Personal-Information.jpg>
- <https://www.webharvy.com/images/web%20scraping.png>
- <https://learnopencv.com/wp-content/uploads/2023/01/tensorflow-keras-cnn-vgg-architecture-1024x611.png>
- <https://cdn1.byjus.com/wp-content/uploads/2020/10/Bayes-Theorem.png>

Ďakujem za pozornosť