TULeCZech: Jazykový embedovací model pro český jazyk

Autor: *Denis Tauchman*

28. dubna 2025

Abstrakt

Tento článek popisuje vývoj a trénování embedovacího jazykového modelu optimalizovaného pro český jazyk, založeného na architektuře **xlm-RoBERTa-base**. Cílem projektu bylo vytvořit model vhodný pro vektorizaci textů v českém jazyce za účelem vyhledávání a porovnávání sémantické podobnosti.

1 Úvod

V rámci bakalářské práce na Technické univerzitě v Liberci byl vytvořen embedovací model TULeCZech, zaměřený na reprezentaci českých textů ve formě vektorů.

2 Cíl projektu

Cílem bylo vytvořit jazykový model optimalizovaný pro český jazyk, vhodný pro vyhledávání relevantních odpovědí a porovnávání sémantické podobnosti za pomocí kosinové podobnosti

3 Použité metriky

Pro objektivní vyhodnocení výkonu jazykového modelu pro navrácení relevantních dokumentů byly použity standardní metriky přesnosti známé jako Top@k accuracy, konkrétně ve variantách acc@1, acc@3 a acc@10. Tyto metriky měří, zda se očekávaný (správný) výstup nachází mezi prvními k výsledky navrácenými modelem. Hodnota metriky se pohybuje v rozmezí od 0 do 1 (nebo 0–100 %), kde vyšší číslo indikuje lepší výkon modelu.

- acc@1 (Top1 accuracy): Měří, zda je správný výsledek hned na první pozici. Jedná se o velmi přísné kritérium, protože model musí přesně odhadnout nejlepší odpověď.
- acc@3 (Top3 accuracy): Umožňuje modelu chybu na první pozici, ale stále hodnotí úspěch, pokud je správná odpověď mezi prvními třemi výsledky.
- acc@10 (Top10 accuracy): Hodnotí, zda se správná odpověď nachází mezi prvními deseti návrhy. Tato metrika lépe reflektuje praktické použití v technice RAG, jelikož se z vektorové databáze vrací více částí dokumentů najednou, pro lepší přenost a kontext.

4 Výběr základního modelu

Byly porovnány dostupné open-source modely založené na architekturách **RoBERTa**, **BERT** a **ELECTRA**. Výsledky shrnuje Tabulka 1.

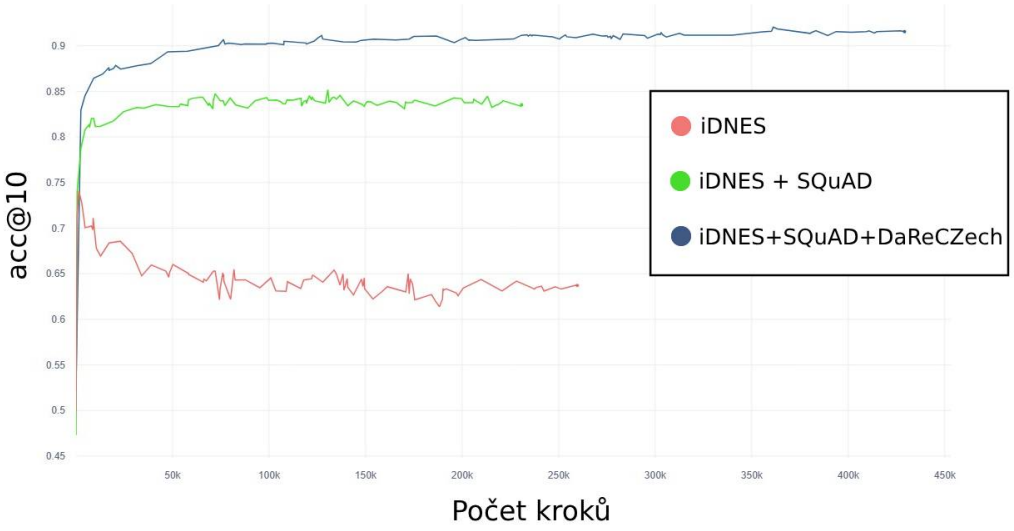
Model	**acc@1 (%)**	**acc@3 (%)**	**acc@10 (%)**
multilingual-e5-base	**65.4**	**86.8**	**91.7**
multilingual-e5-small	63.1	85.1	90.6
sentence-transformers-multilingual-e5-small	63.1	85.1	90.6
paraphrase-multilingual-mpnet-base-v2	44.1	69.6	79.1
distiluse-base-multilingual-cased-v2	42.6	68.6	77.3
LaBSE	42.3	68.3	77.0
Seznam/simcse-dist-mpnet-czeng-cs-en	36.1	62.1	71.8
Seznam/retromae-small-cs	27.1	49.9	59.9
Seznam/simcse-small-e-czech	3.4	9.9	14.8

Tabulka 1: Výsledky modelů na datasetu SQuAD dev2.0

5 Použité datasety

- **SQuAD (cz)** – překlad anglického datasetu upravený na dvojice *(korpus, otázka)*.
- **DaReCZech** – dataset Seznam.cz s dvojicemi *(dotaz, dokument)*.
- **iDnes dataset** – vlastní dataset vytvořený scrapingem článků.

V rámci experimentů, se testovaly různé kombinace datasetů, následný graf zobrazuje přesnost modelu za použití různých kombinací modelů



Obrázek 1: Doladění pomocí různých kombinací datasetů - acc@10

6 Architektura a trénink

Model je založen na **xlm-RoBERTa-base** a trénován na úloze návratu relevantních dokumentů

Architektura **xlm-RoBERTa-base** byla vybrána také z důvodu předtrénování na datasetu, který obsahuje více jak 100 jazyků, což výrazně pomohlo v případě dotrénování na českém jazyce

7 Hyperparametry

Nastavení tréninku:

- Maximální délka vstupu: 512 tokenů
- Batch size: 56
- Optimalizátor: Adam + AdamW
- Počet epoch: 30
- Learning rate: 2e-5

- Loss funkce: MultipleNegativesRankingLoss

8 Experimentální část

Při experimentování bylo testováno několik variant:

- **Learning rate**: [1e-6 - 6e-5]; nejlepší stabilita a konvergence byla při 2e-5.

Learning rate	**acc@1 [%]**
1e-6	49.3
2e-6	52.2
3e-6	53.3
4e-6	53.4
5e-6	54.2
6e-6	54.1
7e-6	54.4
8e-6	54.3
9e-6	54.5
1e-5	54.5
2e-5	**55.3**
3e-5	53.6
4e-5	54.1
5e-5	53.1
6e-5	52.6

Tabulka 2: Výsledná přesnost modelu při různých hodnotách learning rate

- **Batch size**: 4-64; větší batch size zrychlila konvergenci.

Batch size	**acc@1 [%]**
4	34.6
8	43.1
16	50.2
24	53.4
32	53.9
48	55.1
56	**56.3**
64	55.7

Tabulka 3: Výsledná přesnost modelu při různých hodnotách batch size

- **Počet epoch**: mezi 20 a 40 epochami; optimální byl 30 epoch.

9 Výsledky modelu

Vytvořený model dokázal překonat všechny vyhodnocené modely v metrice acc@10, která je z hlediska praktičnosti nejvíce relevantní

model	**acc@1 [%]**	**acc@3 [%]**	**acc@10 [%]**
Multilingual-E5-base	65.4	86.8	91.7
Multilingual-E5-small	63.0	85.0	90.5
distiluse-base-multilingual-cased-v2	42.6	68.5	77.3
LaBSE	42.3	68.2	76.9
Seznam/simcse-dist-mpnet-czeng-cs-en	36.1	62.1	71.8
Seznam/RetroMAE-small-cs	27.0	49.9	59.9
Seznam/simcse-small-e-czech	3.3	9.9	14.8
E5-Czech-Base	63.2	81.2	**92.0**

Tabulka 4: Porovnání s ostatními modely

10 Závěr

Projekt TULeCZech přináší kvalitní embedovací model pro český jazyk. V porovnání s existujícími modely dosahuje konkurenceschopných výsledků a je vhodný pro různé NLP účely.