Understanding of long texts - MVI milestone

Jan Flajžík

November 2024

Link na repozitář: https://gitlab.fit.cvut.cz/flajzjan/mvi-sp

1 Zadání

Cílem mojí semestrální práce je vytvořit přehled existujících knihoven pro generování embeddingů z dlouhých textů. Tyto knihovny porovnat pomocí standartních benchmarků a nakonec demonstrovat jejich použití při klasifikci textu.

2 Přečtené články

Pro tuto práci jsem přečetl následující články:

- Language-Agnostic BERT Sentence Embedding https://arxiv.org/pdf/2007.01852,
- The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT https://aclanthology.org/2020.wmt-1.139.pdf,
- Leveraging Multi-lingual Positive Instances in Contrastive Learning to Improve Sentence Embeddinghttps://arxiv.org/pdf/2309.08929.

3 Datasety

K porovnání embeddingů využívám následující vícejazyčné datasety:

- BUCC[1] dataset zahrnující 35 tisíc textů přeložených ze 4 jazyků (de, ru, fr, zh) do angličtiny [4],
- Tatoeba[2] corpus, který obsahuje překlady do více než 400 jazyků, který bude pro naše experimenty omezen jen na 14 jazyků použitých v prvním cross-lingual sentence retrieval experimentu v paperu [1].

Porovnávání embeddingů je tedy založeno na Cross-Lingual Sentence Retrieval, kdy vytvoříme embeddingy pro stejnou větu ve více jazycích a poté spočítáme jejich cosinovou podobnost.

[1] BUCC
[2] Tatoeba

4 Experimenty

Zatím byly provedeny pouze základní experimenty s architekturami InfoXLM-base a InfoXLMlarge vytvořené Microsoftem [1] na zkráceném datasetu BUCC pro všechny dostupné jazyky omezené na maximálně 3000 datových bodů. Výsledky experimentu jsou zobrazeny v tabulce 1.

model	de-en	fr-en	ru-en	zh-en
infoxlm-base	0.95	0.95	0.94	0.90
infoxlm-large	0.94	0.89	0.91	0.88

Table 1: Porovnání bi-text retrival accuracy modelů infoxlm-base[2] a infoxlm-large[3] pro jednotlivé páry jazyků.

Všechna použitá data se nahrávájí do notebooku se základními experimenty při spuštění přímo z huggingface.

5 Pokračování práce

V této práci se dále budu zabývat experimenty na dalších knihovnách pro generování embeddingů jako je LaBSE. Dále provedu potřebné benchmarky všech knihoven na datasetu Tatoeba. Nakonec se budu zabývat využitím prozkoumaných knihoven při klasifikaci textu.

References

[1] Zewen Chi et al. "InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 3576–3588. DOI: 10.18653/v1/2021.naacl-main.280. URL: https://aclanthology.org/2021.naacl-main.280.

[2] *microsoft/infoxlm-base*. URL: https://huggingface.co/microsoft/infoxlm-base.

[3] *microsoft/infoxlm-base*. URL: https://huggingface.co/KnutJaegersberg/infoxlm-large-sentence-embeddings.

[4] Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. "Overview of the Second BUCC Shared Task: Spotting Parallel Sentences in Comparable Corpora". In: *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*. Ed. by Serge Sharoff, Pierre Zweigenbaum, and Reinhard Rapp. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 60–67. DOI: 10.18653/v1/W17-2512. URL: https://aclanthology.org/W17-2512.