A General-Purpose Multilingual Document Encoder

Onur Galoğlu[1], Robert Litschko[2], Goran Glavaš[3]
[1]Independent Researcher
[2]MaiNLP, CIS, LMU Munich, Germany
[3]CAIDAS, University of Würzburg

Problem Statement

1) Massively multilingual Transformers(MMTs) have drastically pushed the state-of-the-art in multilingual NLP...
2) …but standard transformer-based models process text **linearly**, which leads to suboptimal performance on document-level multilingual tasks.
   a) They do not correspond to the **hierarchical nature** of document organization, i.e. sequences of paragraphs and of sequences of sentences.
   b) Representing documents longer than the MMTs **maximal input length** requires either document trimming or segmentation.

Hierarchical Multilingual Document Encoder (HDME)

d_4: Random Wikipedia Article

Rand

Text: [BOS] w1,1 w1,2 … [EOS]
[BOS] w2,1 w2,2 … [EOS]
[BOS] w3,1 w3,2 … [EOS]
…

Category: Wikipedia category

d_1: en.wikipedia.org/wiki/Car

EN

Text: A car, or an automobile, is a motor vehicle with wheels. Most definitions of cars state that they run primarily roads, [...]

Category: Private Transport

d_2: de.wikipedia.org/wiki/Automobil

DE

Text: Ein Automobil, kurz Auto (in Deutschland amtlich Kraftfahrzeug, in der Schweiz amtlich Motorwagen), ist ein mehrspuriges [...]

Category: Automobile

d_3: en.wikipedia.org/wiki/Bicycle

EN

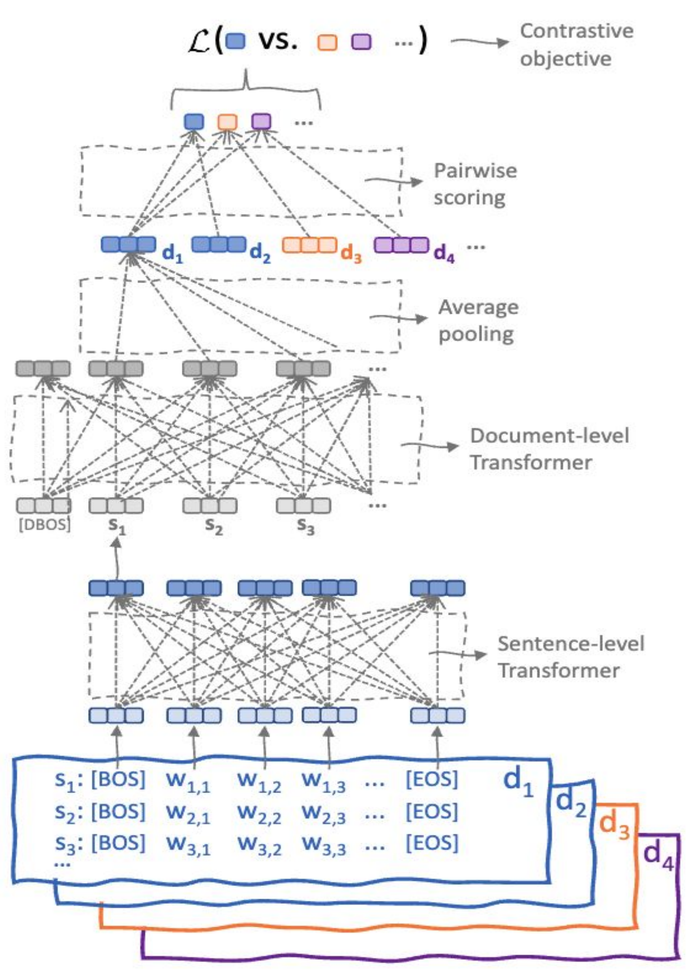
Text: A bicycle, also called a pedal cycle, bike, push-bike or cycle, is a human-powered or motor-powered assisted, pedal-driven [...]

Category: Private Transport

In-Batch Negative Example

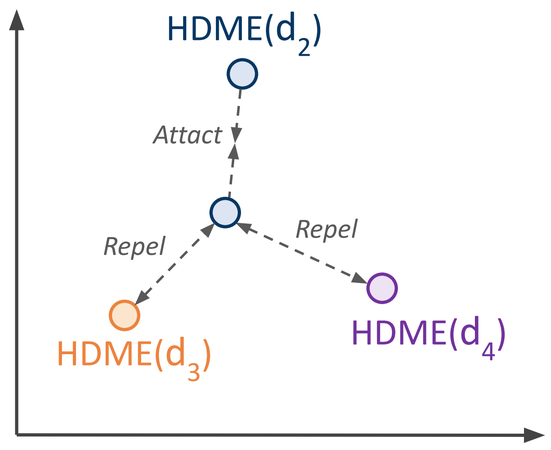
Positive Example

Hard Negative Example



Pre-training Objective

InfoNCE Loss (Oord et al. 2018)



$$\mathcal{L} = -\sum_{i=1}^{N}\left[\frac{1}{\tau}s(\mathbf{d}_1^{(i)},\mathbf{d}_2^{(i)}) - \log\left(e^{s(\mathbf{d}_1^{(i)},\mathbf{d}_{neg}^{(i)})/\tau} + \sum_{j=1}^{N} e^{s(\mathbf{d}_1^{(i)},\mathbf{d}_2^{(j)})/\tau}\right)\right]$$

s = cosine similarity

τ = temperature parameter

Main Results

Models and Baselines

Standard Multilingual Transformers

- Sentence Encoder: LaBSE (Feng et al. 2022)
- Vanilla mPLM: XLM-R (Conneau et al. 2019)
- Vanilla mPLM: mBERT (Devlin et al. 2018)

Multilingual Long Document Encoders

- Sliding Window: LaBSE-Seg (size=128, stride=42)
- Multilingual Longformer (Beltagy et al. 2020)
  - Parameters are initialized from LaBSE
  - MLM pretraining on the same corpus as HMDE

HDME (ours)

- Pre-trained on four languages (HDME, **XLW-4L**, shown in Tbl. 1 and 2): EN, DE, FR and IT.
- Pre-trained on twelve languages (HDME, **XLW-12L**): EN, FR, RU, JA, ZH, HU, FI, AR, FA, TR, GK and MS.

Model	En	Es	De	Fr	It	Ru	Ja	Zh	AVG
Standard Multilingual Transformers									
LaBSE	95.5	79.0	89.6	87.2	76.8	63.9	**80.8**	86.1	82.4
XLM-R	93.0	84.6	**92.5**	87.1	73.2	68.9	78.2	85.8	83.0
mBERT	**96.9**	81.9	88.3	83.1	74.1	72.3	74.6	84.4	82.0
Multilingual Long Document Encoders									
LaBSE-Seg	94.0	82.9	90.2	89.9	78.1	71.9	75.5	88.4	84.0
mLongformer	95.8	**87.0**	93.4	91.9	**80.6**	71.7	79.5	88.5	86.1
HMDE	95.4	85.6	91.2	**92.0**	78.5	**83.9**	76.3	**89.5**	**86.8**

Table 1: **Supervised document classification** results (MLDOC) in terms of Accuracy. Best result per language highlighted in **blue**.

Experimental Results

- Evaluate HDME as a document encoder on a **supervised classification task**, MLDoc (Schwenk and Li, 2018)
  - Apply classification (FFN) layer on top of encoder.
  - Our model and mLongformer trained on four languages (XLW-4) **outperform** all of the standard MMTs.
- We evaluate HDME representations on **unsupervised cross-lingual IR**, CLEF-2003 (Braschler et al. 2003)
  - Bi-Encoder Paradigm: Encode queries and documents independently, rank according to cosine sim.
  - Our model trained on four languages (XLW-4L) **outperforms baselines** by a large margin.
  - HDME sems to **generalize well to unseen languages**, i.e. languages that are not included in XLW-4/12L.

Model	En–Fi	En–It	En–Ru	En–De	De–Fi	De–It	De–Ru	Fi–It	Fi–Ru	AVG
Standard Multilingual Transformers										
LaBSE	.247	.224	.131	.138	.247	.214	.135	.211	.125	.186
mBERT	.145	.146	**.167**	.107	.151	.116	**.149**	.117	.128	.136
Multilingual Long Document Encoders										
LaBSE-Seg	.243	.169	.107	.194	.268	.178	.104	.153	.014	.159
mLongformer	.150	.088	.094	.082	.190	.072	.120	.097	.091	.109
HMDE	**.389**	**.282**	.141	**.326**	**.352**	**.259**	.130	**.238**	**.129**	**.249**

Table 2: **Unsupervised cross-lingual document retrieval** results (CLEF-2003) in terms of Mean Average Precision (MAP). Best results per language highlighted in **blue**.

References

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. Language-agnostic BERT Sentence Embedding. In Proceedings of ACL 2022.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. arXiv preprint arXiv:1911.02116.

J. Devlin, M.- W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL 2019.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. arXiv preprint arXiv:1901.04085.

I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The Long-Document Transformer. arXiv preprint arXiv:2004.05150.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).