Evaluating Multilingual Text Encoders for Unsupervised Cross-Lingual Retrieval

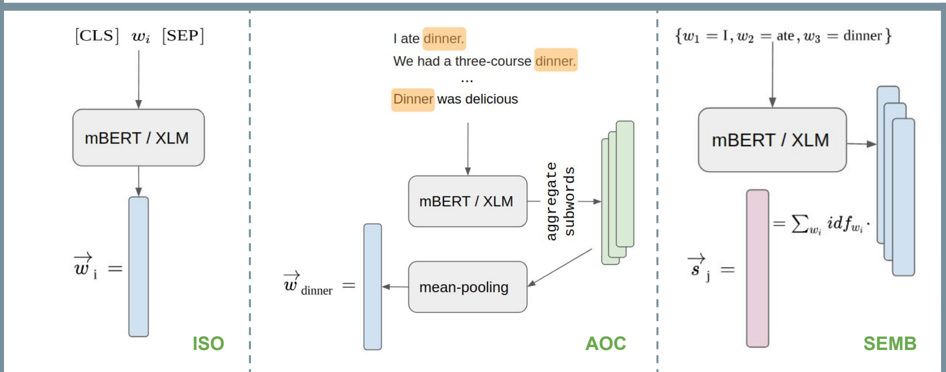
Robert Litschko*, Ivan Vulić**, Simone Paolo Ponzetto*, Goran Glavaš*
*University of Mannheim, **Cambridge University

presented at ECIR 2021

Motivation

- Pre-trained Transformers achieve strong performance in NLP and have been adopted for multilingual NLP.
- Multilingual Text Encoders render Cross-lingual Word Embeddings (CLWE) effectively obsolete.
- RQ: To which extend does this generalize to **unsupervised Cross-lingual Information Retrieval (CLIR)**?
- Unsup. CLIR: Encode queries and documents by their constituent word embeddings, rank with cosine similarity.
- In previous work we benchmarked a range of methods for inducing CLWE spaces [2].
- This work studies the efficacy of representations from multilingual encoders in the context of unsupervised CLIR.

Models



Baselines

- **MT-IR**: Translate query into the document language, retrieve documents with Query Likelihood Model.
- **Proc-B**: (1) Row-align monolingual embedding matrices with word translation pairs from bilingual dictionary.
  (2) Learn linear mapping (Procrustes [5]): $W_{L_1} = \arg\min_W ||X_{L_1} W - X_{L_2}||_2$
  (3) Boostrap new word pairs from cross-lingual nearest neighbors, repeat.
- **Proc-B$_{\text{LEN}}$**: Additional max. sequence length constraint.

Models based on multilingual Transformers

- **Average over contexts** (**AOC**): Avg. contextualized embeddings.
- Dynamic and in-place "sentence embedding encoding" (**SEMB**), *idf*-weighted token aggregation.

} Static emb. spaces refined with procrustes [5]

Similarity specialized sentence encoders

Seq2Seq NMT (**LASER**) [0], multi-task learning (**m-USE**) [6] and multi-task + self-supervision (**LaBSE**) [1].

Results

	EN-FI	EN-IT	EN-RU	EN-DE	DE-FI	DE-IT	DE-RU	FI-IT	FI-RU	AVG	w/o FI
Baselines											
MT-IR	.278	**.423**	.225	**.339**	**.340**	**.418**	.196	**.389**	.212	**.313**	**.319**
Proc-B	.258	.265	.166	.288	.294	.230	.155	.151	.136	.216	.227
Proc-B$_{\text{LEN}}$	.165	.232	.176	.194	.207	.186	.192	.126	.154	.181	.196
Models based on multilingual Transformers											
SEMB$_{\text{XLM}}$	.199*	.187*	.183	.126*	.156*	.166*	.228	.186*	.139	.174	.178
SEMB$_{\text{mBERT}}$	.145*	.146*	.167	.107*	.151*	.116*	.149*	.117	.128*	.136	.137
AOC$_{\text{XLM}}$	.168	.261	.208	.206*	.183	.190	.162	.123	.099	.178	.206
AOC$_{\text{mBERT}}$	.172*	.209*	.167	.193*	.131*	.143*	.143	.104	.132	.155	.171
ISO$_{\text{XLM}}$	.058*	.159*	.050*	.096*	.026*	.077*	.035*	.050*	.055*	.067	.083
ISO$_{\text{mBERT}}$	.075*	.209	.096*	.157*	.061*	.107*	.025*	.051*	.014*	.088	.119
Similarity specialized sentence encoders											
DISTIL$_{\text{XLM-R}}$	.216	.190*	.179	.114*	.237	.181	.173	.166	.138	.177	.167
DISTIL$_{\text{USE}}$	.141*	.346*	.182	.258	.139*	.324*	.179	.104	.111	.198	.258
DISTIL$_{\text{DistilmBERT}}$	**.294**	.290*	**.313**	.247*	.300	.267*	**.284**	.221*	**.302***	.280	.280
LaBSE	.180*	.175*	.128	.059*	.178*	.160*	.113*	.126	.149	.141	.127
LASER	.142	.134*	.076	.046*	.163*	.140*	.065*	.144	.107	.113	.094
m-USE	.109*	.328*	.214	.230*	.107*	.294*	.204	.073	.090	.183	.254

Layer Ablation (Avg.)

- Results here presented as Mean Average Precision (MAP) on **document retrieval** (CLEF 2003).
- Multilingual transformers and sentence encoders **are not universally superior** to static CLWE's in cross-lingual retrieval, upper layers performing best.
- **Sentence retrieval** experiments *(not shown here)* indicate opposing results: (1) SEMB outperforms Proc-B, similarity specialized encoders outperform Prob-B and MT-IR; (2) middle layers yield best results.

Future Work: Semantic similarity ≠ relevance matching

- Sentence similarity matching results don't translate to document retrieval.
- What model and dataset biases are necessary for successful cross-lingual transfer of IR rankers/encoders?
- Large scale and realistic CLIR dataset for supervised cross-lingual document rankers.

[0] Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot crosslingual transfer and beyond. TACL 2019
[1] Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic BERT sentence embedding. arXiv:2007.01852 (2020)
[2] Glavaš, G., Litschko, R., Ruder, S., Vulić, I.: How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In: ACL 2019
[4] Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: EMNLP 2020
[5] Smith, S.L., Turban, D.H., Hamblin, S., Hammerla, N.Y.: Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In: ICLR 2017
[6] Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G.H., Yuan, S., Tar, C., Sung, Y.h., Strope, B., Kurzweil, R.: Multilingual universal sentence encoder for semantic retrieval. In: ACL 2020