Evaluating Multilingual Text Encoders for Unsupervised Cross-Lingual Retrieval

Robert Litschko[1], Ivan Vulić[2], Simone Paolo Ponzetto[1], and Goran Glavaš[1]

[1] Data and Web Science Group, University of Mannheim, Mannheim, Germany
{litschko,simone,goran}@informatik.uni-mannheim.de
[2] Language Technology Lab, University of Cambridge, UK
iv250@cam.ac.uk

Abstract. Pretrained multilingual text encoders based on neural Transformer architectures, such as multilingual BERT (mBERT) and XLM, have achieved strong performance on a myriad of language understanding tasks. Consequently, they have been adopted as a go-to paradigm for multilingual and cross-lingual representation learning and transfer, rendering cross-lingual word embeddings (CLWEs) effectively obsolete. However, questions remain to which extent this finding generalizes 1) to unsupervised settings and 2) for ad-hoc cross-lingual IR (CLIR) tasks. Therefore, in this work we present a systematic empirical study focused on the suitability of the state-of-the-art multilingual encoders for cross-lingual document and sentence retrieval tasks across a large number of language pairs. In contrast to supervised language understanding, our results indicate that for unsupervised document-level CLIR – a setup with no relevance judgments for IR-specific fine-tuning – pretrained encoders fail to significantly outperform models based on CLWEs. For sentence-level CLIR, we demonstrate that state-of-the-art performance can be achieved. However, the peak performance is not met using the general-purpose multilingual text encoders 'off-the-shelf', but rather relying on their variants that have been further specialized for sentence understanding tasks.

Keywords: Cross-lingual IR · Multilingual text encoders · Unsupervised IR.

1 Introduction

Cross-lingual information retrieval (CLIR) systems respond to queries in a source language by retrieving relevant documents in another, target language. Their success is typically hindered by data scarcity: they operate in challenging low-resource settings without sufficient labeled training data, i.e., human relevance judgments, to build supervised models (e.g., neural matching models for pair-wise retrieval [53, 22]). This motivates the need for robust, resource-lean and unsupervised CLIR approaches.

In previous work, Litschko et al. [27] have shown that language transfer through cross-lingual embedding spaces (CLWEs) can be used to yield state-of-the-art performance in a range of unsupervised ad-hoc CLIR setups. This approach uses very weak supervision (i.e., only a bilingual dictionary spanning 1K-5K word translation pairs), or even no supervision at all, in order to learn a mapping that aligns two monolingual word embedding spaces [19, 45]. Put simply, this enables casting CLIR tasks as 'monolingual tasks in the shared (CLWE) space': at retrieval time both queries and documents are represented as simple aggregates of their constituent CLWEs. However, this approach, by limitations of static CLWEs, cannot capture and handle polysemy in the underlying text representations. *Contextual text representation models* alleviate this issue [28]. They encode occurrences of the same word differently depending on its surrounding context.

Such contextual representations are obtained via large models pretrained on large text collections through general objectives such as (masked) language modeling [16, 30]. Multilingual text encoders pretrained on 100+ languages, such as mBERT [16] or XLM [14], have become a *de facto* standard for multilingual representation learning and cross-lingual transfer in natural language processing (NLP). These models demonstrate state-of-the-art performance in a wide range of supervised language understanding and language generation tasks [36, 26], especially in zero-shot settings: a typical *modus operandi* is fine-tuning a pretrained multilingual encoder with task-specific data of a source language (typically English) and then using it directly in a target language.

It is unclear, however, whether these general-purpose multilingual text encoders can be used directly for ad-hoc CLIR without any additional supervision (i.e., relevance judgments). Further, can they outperform unsupervised CLIR approaches based on static CLWEs [27]? How do they perform depending on the (properties of the) language pair at hand? How can we encode useful semantic information using these models, and do different "encoding variants" (see later §3) yield different retrieval results? Are there performance differences in unsupervised sentence-level versus document-level CLIR tasks? Finally, can we boost performance by relying on sentence encoders that are specialized towards dealing with sentence-level understanding in particular? In order to address these questions, we present a systematic empirical study and profile the suitability of state-of-the-art pretrained multilingual encoders for different CLIR tasks and diverse language pairs. We evaluate two state-of-the-art general-purpose pretrained multilingual encoders, mBERT [16] and XLM [14] with a range of encoding variants, and also compare them to CLIR approaches based on static CLWEs, and specialized multilingual sentence encoders. Our key contributions can be summarized as follows:

(1) We empirically validate that, without any task-specific fine-tuning, multilingual encoders such as mBERT and XLM fail to outperform CLIR approaches based on static CLWEs. Their performance also crucially depends on how one encodes semantic information with the models (e.g., treating them as sentence/document encoders directly versus averaging over constituent words and/or subwords). We also show that there is no "one-size-fits-all" approach, and the results are task- and language-pair-dependent.

(2) We provide a first large-scale comparative evaluation of state-of-the art pretrained multilingual encoders on unsupervised document-level CLIR. We also empirically show that encoder models specialized for sentence-level understanding substantially outperform general-purpose models (mBERT and XLM) on sentence-level CLIR tasks.

2 Related Work

Self-Supervised Pretraining and Transfer Learning. Recently, research on universal sentence representations and transfer learning has gained much traction. InferSent [13] transfers the encoder of a model trained on natural language inference to other tasks, while USE [8] extends this idea to a multi-task learning setting. More recent work explores self-supervised neural Transformer-based [44] models based on (causal or masked) language modeling (LM) objectives such as BERT [16], RoBERTa [30], GPT [37, 5], and XLM [14].[3] Results on benchmarks such as GLUE [47] and SentEval [12] indicate that these models can yield impressive (sometimes human-level) performance in supervised Natural Language Understanding (NLU) and Generation (NLG) tasks. These models have become *de facto* standard and omnipresent text representation models in NLP. In supervised monolingual IR, self-supervised LMs have been employed as contextualized word encoders [32], or fine-tuned as pointwise and pairwise rankers [33].

Multilingual Text Encoders based on the (masked) LM objectives have also been massively adopted in multilingual and cross-lingual NLP and IR applications. A multilingual extension of BERT (mBERT) is trained with a shared subword vocabulary on a single multilingual corpus obtained as concatenation of large monolingual data in 104 languages. The XLM model [14] extends this idea and proposes natively cross-lingual LM pretraining, combining causal language modeling (CLM) and translation language modeling (TLM).[4] Strong performance of these models in supervised settings is confirmed across a range of tasks on multilingual benchmarks such as XGLUE [26] and XNLI [15]. However, recent work [39, 6] has indicated that these general-purpose models do not yield strong results when used as out-of-the-box text encoders in an unsupervised transfer learning setup. We further investigate these preliminaries, and confirm this finding also for unsupervised ad-hoc CLIR tasks.

Multilingual text encoders have already found applications in document-level CLIR. Jiang et al. [22] use mBERT as a matching model by feeding pairs of English queries and foreign language documents. MacAvaney et al. [31] use mBERT in a zero-shot setting, where they train a retrieval model on top of mBERT on English relevance data and apply it on a different language. However, prior work has not investigated unsupervised CLIR setups, and a systematic comparative study focused on the suitability of the multilingual text encoders for diverse ad-hoc CLIR tasks and language pairs is still lacking.

Specialized Multilingual Sentence Encoders. An extensive body of work focuses on inducing multilingual encoders that capture sentence meaning. In [2], the multilingual encoder of a sequence-to-sequence model is shared across languages and optimized to be language-agnostic, whereas Guo et al. [20] rely on a dual Transformer-based encoder architectures instead (with tied/shared parameters) to represent parallel sentences. Rather than optimizing for translation performance directly, their approach minimizes the cosine distance between parallel sentences. A ranking softmax loss is used to classify the correct (i.e., aligned) sentence in the other language from negative samples (i.e., non-aligned sentences). In [50], this approach is extended by using a bidirectional dual encoder and adding an additive margin softmax function, which serves to push away non-translation-

[3] Note that self-supervised learning can come in different flavors depending on the training objective [10], but language modeling objectives still seem to be the most popular choice.

[4] In CLM, the model is trained to predict the probability of a word given the previous words in a sentence. TLM is a cross-lingual variant of standard masked LM (MLM), with the core difference that the model is given pairs of parallel sentences and allowed to attend to the aligned sentence when reconstructing a word in the current sentence.

pairs in the shared embedding space. The dual-encoder approach is now widely adopted [20, 51, 18, 39, 56], and yields state-of-the-art multilingual sentence encoders which excel in sentence-level NLU tasks.

Other recent approaches propose input space normalization, and re-aligning mBERT and XLM with parallel data [56, 6], or using a teacher-student framework where a student model is trained to imitate the output of the teacher network while preserving high similarity of translation pairs [39]. In [51], authors combine multi-task learning with a translation bridging task to train a universal sentence encoder. We benchmark a series of representative sentence encoders; their brief descriptions are provided in §3.3.

CLIR Evaluation and Application. The cross-lingual ability of mBERT and XLM has been investigated by probing and analyzing their internals [23], as well as in terms of downstream performance [34, 49]. In CLIR, these models as well as dedicated multilingual sentence encoders have been evaluated on tasks such as cross-lingual question-answer retrieval [51], bitext mining [58, 59], and semantic textual similarity (STS) [21, 25]. Yet, the models have been primarily evaluated on sentence-level retrieval, while classic ad-hoc (unsupervised) document-level CLIR has not been in focus. Further, previous work has not provided a large-scale comparative study across diverse language pairs and with different model variants, nor has tried to understand and analyze the differences between sentence-level and document-level tasks. In this work, we aim to fill these gaps.

3 Multilingual Text Encoders

We provide an overview of all multilingual models in our evaluation. We discuss general-purpose multilingual text encoders (§3.2), as well as specialized multilingual sentence encoders in §3.3. For completeness, we first briefly describe static CLWEs (§3.1).

3.1 CLIR with (Static) Cross-lingual Word Embeddings

We assume a query Q_{L_1} issued in a source language L_1, and a document collection of N documents $D_{i,L_2}, i = 1, \ldots, N$ in a target language L_2. Let $d = \{t_1, t_2, \ldots, t_{|D|}\} \in D$ be a document with $|D|$ terms t_i. CLIR with static CLWEs represents queries and documents as vectors $\overrightarrow{Q}, \overrightarrow{D} \in \mathbb{R}^d$ in a d-dimensional shared embedding space [46, 27]. Each term is represented independently with a pre-computed static embedding vector $\overrightarrow{t_i} = emb\,(t_i)$. There exist a range of methods for inducing shared embedding spaces with different levels of supervision, such as parallel sentences, comparable documents, small bilingual dictionaries, or even methods without any supervision [41]. Given the shared CLWE space, both query and document representations are obtained as aggregations of their term embeddings. We follow Litschko et al. [27] and represent documents as the weighted sum of their terms' vectors, where each term's weight corresponds to its inverse document frequency (idf) : $\overrightarrow{d} = \sum_{i=1}^{N_d} idf\,(t_i^d) \cdot \overrightarrow{t_i^d}$. During retrieval documents are ranked according to the cosine similarity to the query.

3.2 Multilingual (Transformer-Based) Language Models: mBERT and XLM

Massively multilingual pretrained neural language models such as mBERT and XLM can be used as a dynamic embedding layer to produce contextualized word representations, since they share a common input space on the subword level (e.g. word-pieces, byte-pair-encodings) across all languages. Let us assume that a term (i.e., a word-level token) is tokenized into a sequence of K subword tokens ($K \geqslant 1$; for simplicity, we assume that the subwords are word-pieces (wp)): $t_i = \left\{wp_{i,k}\right\}_{k=1}^{K}$. The multilingual encoder then produces contextualized subword embeddings for the term's K constituent subwords $\overrightarrow{wp_{i,k}}$, $k = 1, \ldots, K$, and we can aggregate these subword embeddings to obtain the representation of the term t_i: $\overrightarrow{t_i} = \psi\left(\{\overrightarrow{wp_{i,k}}\}_{k=1}^{K}\right)$, where the function $\psi()$ is the aggregation function over the K constituent subword embeddings. Once these term embeddings $\overrightarrow{t_i}$ are obtained, we follow the same CLIR setup as with CLWEs in §3.1.

Static Word Embeddings from Multilingual Transformers. We first use multilingual transformers (mBERT and XLM) in two different ways to induce static word embedding spaces for all languages. In a simpler variant, we feed terms into the encoders *in isolation* (**ISO**), that is, without providing any surrounding context for the terms. This effectively constructs a static word embedding table similar to what is done in §3.1, and allows the CLIR model (§3.1) to operate at a non-contextual word level. An empirical CLIR comparison between ISO and CLIR operating on CLWEs [27] then effectively quantifies how well multilingual encoders (mBERT and XLM) encode word-level representations.

In a more elaborate variant we do leverage the contexts in which the terms appear, constructing *average-over-contexts* embeddings (**AOC**). For each term t we collect a set of sentences $s_i \in \mathcal{S}_t$ in which it occurs. We use the full set of Wikipedia sentences $\mathcal{S}$ to sample sets of contexts $\mathcal{S}_t$ for vocabulary terms. For a given sentence s_i let j denote the position of t's first occurrence. We then transform s_i with mBERT or XLM as the encoder, $enc(s_i)$, and extract the contextualized embedding of t via *mean-pooling*, i.e., by averaging embeddings of its constituent subwords, $\psi\left(\{\overrightarrow{wp_{j,k}}\}_{k=1}^{K}\right) = 1/K \cdot \sum_{k=1}^{K} \overrightarrow{wp_{j,k}}$. For each vocabulary term, we obtain $N_t = min(|\mathcal{S}_t|, \tau)$ contextualized vectors, with $|\mathcal{S}_t|$ as the number of Wikipedia sentences containing t and τ as the maximal number of sentence samples for a term. The final static embedding of t is then simply the average over the N_t contextualized vectors.

The obtained static AOC and ISO embeddings, despite being induced with multilingual encoders, however, did not appear to be well-aligned across languages [29, 6]. We evaluated the static ISO and AOC embeddings induced for different languages with multilingual encoders (mBERT and XLM), on the bilingual lexicon induction (BLI) task [19]. We observed poor BLI performance, suggesting that further projection-based alignment of respective monolingual ISO and AOC spaces is required. To this end, we use the the standard Procrustes method [43, 1] to align the embedding spaces of two languages, with bilingual dictionaries from [19] as the supervision guiding the alignment. Concretely, for each language pair in our experiments we project the AOC (ISO) embeddings of the source language to the AOC (ISO) space of the target language.

Direct Text Embedding with Multilingual Transformers. In both AOC and ISO, we use the multilingual (contextual) encoders to obtain the static embeddings for word types (i.e., terms): we can then leverage in exactly the same ad-hoc retrieval setup (§3.1) in which CLWEs had previously been evaluated [27]. In an arguably more straightforward approach, we also use pretrained multilingual Transformers (i.e., mBERT or XLM) to directly encode the whole input text (**SEMB**). We encode the input text by averaging the contextualized representations of all terms in the text (we again compute the weighted average, where the terms' IDF scores are used as weights, see §3.1). For SEMB, we take the contextualized representation of each term t_i to be the contextualized representation of its first subword token, i.e., $\overrightarrow{t_i} = \psi\left(\{\overrightarrow{wp_{i,k}}\}_{k=1}^{K}\right) = \overrightarrow{wp_{i,1}}$.[5]

3.3 Specialized Multilingual Sentence Encoders

Off-the-shelf multilingual Transformers (mBERT and XLM) have been shown to yield sub-par performance in unsupervised text similarity tasks; therefore, in order to be successful in semantic text (sentences or paragraph) comparisons, they first need to be fine-tuned on text matching (typically sentence matching) datasets [39, 6, 57]. Such encoders *specialized for semantic similarity* are supposed to encode sentence meaning more accurately, supporting tasks that require unsupervised (ad-hoc) semantic text matching. In contrast to mBERT and XLM, which contextualize (sub)word representations, these models directly produce a semantic embedding of the input text. We provide a brief overview of the models included in our comparative evaluation.

Language Agnostic SEntence Representations (LASER) [2] adopts a standard sequence-to-sequence architecture typical for neural machine translation (MT). It is trained on 223M parallel sentences covering 93 languages. The encoder is a multi-layered bidirectional LSTM and the decoder is a single-layer unidirectional LSTM. The 1024-dimensional sentence embedding is produced by max-pooling over the outputs of encoder's last layer. The decoder then takes the sentence embedding as additional input as each decoding step. The decoder-to-encoder attention and language identifiers on the encoder side are deliberately omitted, so that all relevant information gets 'crammed' into the fixed-sized sentence embedding produced by the encoder. In our experiments, we directly use the output of the encoder to represent both queries and documents.

Multilingual Universal Sentence Encoder (m-USE) is a general purpose sentence embedding model for transfer learning and semantic text retrieval tasks [51]. It relies on a standard dual-encoder neural framework [9, 52] with shared weights, trained in a multi-task setting with an additional translation bridging task. For more details, we refer the reader to the original work. There are two pretrained m-USE instances available – we opt for the 3-layer Transformer encoder with average-pooling.

Language-agnostic BERT Sentence Embeddings (LaBSE) [18] is another neural dual-encoder framework, also trained with parallel data. Unlike in LASER and m-USE, where the encoders are trained from scratch on parallel data, LaBSE training starts from a pretrained mBERT instance (i.e., a 12-layer Transformer network pretrained on the concatenated corpora of 100+ languages). In addition to the multi-task training objective of m-USE, LaBSE additionally uses standard self-supervised objectives used in pretraining of mBERT and XLM: masked and translation language modelling (MLM and TLM, see §2). For further model details, we refer the reader to the original work.

DISTIL [39] is a teacher-student framework for injecting the knowledge obtained through specialization for semantic similarity from a specialized monolingual trans-

[5] In our initial experiments taking the vector of the first term's subword consistently outperformed averaging vectors of all its subwords.

former (e.g., BERT) into a non-specialized multilingual transformer (e.g., mBERT). It first specializes for semantic similarity a monolingual (English) teacher encoder M using the available semantic sentence-matching datasets for supervision. In the second, *knowledge distillation* step a pretrained multilingual student encoder $\widehat{M}$ is trained to mimic the output of the teacher model. For a given batch of sentence-translation pairs $\mathcal{B} = \{(s_j, t_j)\}$, the teacher-student distillation training minimizes the following loss:

$$\mathcal{J}(\mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} \left[\left(M(s_j) - \widehat{M}(s_j) \right)^2 + \left(M(s_j) - \widehat{M}(t_j) \right)^2 \right].$$

The teacher model M is Sentence-BERT [38], BERT specialized for embedding sentence meaning on semantic text similarity [7] and natural language inference [48] datasets. The teacher network only encodes English sentences s_i. The student model $\widehat{M}$ is then trained to produce for both s_j and t_j the same representation that M produces for s_j. We benchmark different DISTIL models in our CLIR experiments, with the student $\widehat{M}$ initialized with different multilingual transformers.

4 Experimental Setup

Evaluation Data. We follow the experimental setup of Litschko et al. [27], and compare the models from §3 on language pairs comprising five languages: English (EN), German (DE), Italian (IT), Finnish (FI) and Russian (RU). For document-level retrieval we run experiments for the following nine language pairs: EN-{FI, DE, IT, RU}, DE-{FI, IT, RU}, FI-{IT, RU}. We use the 2003 portion of the CLEF benchmark [4],[6] with 60 queries per language pair. The document collection sizes are 17K (RU), 55K (FI), 158K (IT), and 295K (DE). For sentence-level retrieval, also following [27], for each language pair we sample from Europarl [24] 1K source language sentences as queries and 100K target language sentences as the "document collection".[7]

Baseline Models. In order to establish whether multilingual encoders outperform CLWEs in a fair comparison, we compare their performance against the strongest CLWE-based CLIR model from the recent comparative study [27], dubbed Proc-B. Proc-B induces a bilingual CLWE space from pretrained monolingual FASTTEXT embeddings[8] using the linear projection computed as the solution of the Procrustes problem given the dictionary of word-translation pairs. Compared to simple Procrustes mapping, Proc-B iteratively (1) augments the word translation dictionary by finding mutual nearest neighbours and (2) induces a new projection matrix using the augmented dictionary. The final bilingual CLWE space is then plugged into the CLIR model from §3.1.

Our document-level retrieval SEMB models do not get to see the whole document but only the first 128 word-piece tokens. For a more direct comparison, we therefore additionally evaluate the Proc-B baseline (Proc-B_{LEN}) which is exposed to exactly the

[6] http://catalog.elra.info/en-us/repository/browse/ELRA-E0008/

[7] Russian is not included in Europarl and we therefore exclude it from sentence-level experiments. Further, since some multilingual encoders have not seen Finnish data in pretraining, we additionally report the results over a subset of language pairs that do not involve Finnish.

[8] https://fasttext.cc/docs/en/pretrained-vectors.html

same amount of document text as the multilingual XLM encoder (i.e., the leading document text corresponding to first 128 word-piece tokens) Finally, we compare CLIR models based on multilingual Transformers to a baseline relying on machine translation baseline (MT-IR). In MT-IR, 1) we translate the query to the document language using Google Translate and then 2) perform monolingual retrieval using a standard Query Likelihood Model [35] with Dirichlet smoothing [55].

Model Details. For all multilingual encoders we experiment with different input sequence lengths: 64, 128, 256 subword tokens. For AOC we collect (at most) $\tau = 60$ contexts for each vocabulary term: for a term not present at all in Wikipedia, we fall back to the ISO embedding of that term. We also investigate the impact of τ in §5.3. For purely self-supervised models (SEMB, ISO, AOC) we independently evaluate representations from different Transformer layers (cf. §5.3). For comparability, for ISO and AOC – methods that effectively induce static word embeddings using multilingual contextual encoders – we opt for exactly the same term vocabularies used by the Proc-B baseline, namely the top 100K most frequent terms from respective monolingual fastText vocabularies. We additionally experiment with three different instances of the DISTIL model: (i) $\text{DISTIL}_{\text{XLM-R}}$ initializes the student model with the pretrained XLM-R transformer [11]; $\text{DISTIL}_{\text{USE}}$ instantiates the student as the pretrained m-USE instance [51]; whereas $\text{DISTIL}_{\text{DistilmBERT}}$ distils the knowledge from the Sentence-BERT teacher into a multilingual version of DistilBERT [42], a 6-layer transformer pre-distilled from mBERT.[9] For SEMB models we scale embeddings of special tokens (sequence start and end tokens, e.g., `[CLS]` and `[SEP]` for mBERT) with the mean IDF value of input terms.

5 Results and Discussion

5.1 Document-Level Cross-lingual Retrieval

We show the performance (MAP) of multilingual encoders on document-level CLIR tasks in Table 1. The first main finding is that none of the self-supervised models (mBERT and XLM in ISO, AOC, and SEMB variants) outperforms the CLWE baseline Proc-B. However, the full Proc-B baseline has, unlike mBERT and XLM variants, been exposed to the full content of the documents. A fairer comparison, against $\text{Proc-B}_{\text{LEN}}$, which has also been exposed only to the first 128 tokens, reveals that SEMB and AOC variants come reasonably close, albeit still do not outperform $\text{Proc-B}_{\text{LEN}}$. This suggests that the document-level retrieval could benefit from encoders able to encode longer portions of text, e.g., [3, 54]. For document-level CLIR, however, these models would first have to be ported to multilingual setups. Scaling embeddings by their *idf* (Proc-B) effectively filters out high-frequent terms such as stopwords. We therefore experiment with explicit a priori stopword filtering in $\text{DISTIL}_{\text{DistilmBERT}}$, dubbed $\text{DISTIL}_{\text{FILTER}}$. Results show that performance deteriorates which indicates that stopwords provide important contextualization information. While SEMB and AOC variants exhibit similar performance, ISO variants perform much worse. The direct comparison between ISO

[9] Working with mBERT directly instead of its distilled version led to similar scores, while increasing running times.

Table 1: Document-level CLIR results (Mean Average Precision, MAP). **Bold**: best model for each language-pair. *: difference in performance w.r.t. Proc-B significant at $p = 0.05$, computed via paired two-tailed t-test with Bonferroni correction.

	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	.276	**.428**	.383	**.263**	**.332**	**.431**	.238	**.406**	.261	**.335**	**.349**
Proc-B	.258	.265	.166	.288	.294	.230	.155	.151	.136	.216	.227
Proc-B$_{\mathrm{LEN}}$	.165	.232	.176	.194	.207	.186	.192	.126	.154	.181	.196
Models based on multilingual Transformers											
SEMB$_{\mathrm{XLM}}$	.199*	.187*	.183	.126*	.156*	.166*	.228	.186*	.139	.174	.178
SEMB$_{\mathrm{mBERT}}$	.145*	.146*	.167	.107*	.151*	.116*	.149*	.117	.128*	.136	.137
AOC$_{\mathrm{XLM}}$	.168	.261	.208	.206*	.183	.190	.162	.123	.099	.178	.206
AOC$_{\mathrm{mBERT}}$	.172*	.209*	.167	.193*	.131*	.143*	.143	.104	.132	.155	.171
ISO$_{\mathrm{XLM}}$	.058*	.159*	.050*	.096*	.026*	.077*	.035*	.050*	.055*	.067	.083
ISO$_{\mathrm{mBERT}}$	.075*	.209	.096*	.157*	.061*	.107*	.025*	.051*	.014*	.088	.119
Similarity-specialized sentence encoders (with parallel data supervision)											
DISTIL$_{\mathrm{FILTER}}$	.291	.261	.278	.255	.272	.217	.237	.221	.270	.256	.250
DISTIL$_{\mathrm{XLM\text{-}R}}$	.216	.190*	.179	.114*	.237	.181	.173	.166	.138	.177	.167
DISTIL$_{\mathrm{USE}}$	.141*	.346*	.182	.258	.139*	.324*	.179	.104	.111	.198	.258
DISTIL$_{\mathrm{DistilmBERT}}$	**.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

and AOC demonstrates the importance of contextual information and seemingly limited usability of multilingual encoders as word encoders, if no context is available.

Similarity-specialized multilingual encoders, which rely on pretraining with parallel data, yield mixed results. Three models, DISTIL$_{\mathrm{DistilmBERT}}$, DISTIL$_{\mathrm{USE}}$ and m-USE, generally outperform the Proc-B baseline[10] LASER is the only encoder trained on parallel data that does not beat the Proc-B baseline. We believe this is because (a) LASER's recurrent encoder provides text embeddings of lower quality than Transformer-based encoders of m-USE and DISTIL variants and (b) it has not been subdued to any self-supervised pretraining like DISTIL models. Even the best-performing CLIR model based on a multilingual encoder (DISTIL$_{\mathrm{DistilmBERT}}$) overall falls behind the MT-based baseline (MT-IR). However, the performance of MT-IR crucially depends on the quality of MT for the concrete language pair: for language pairs with weaker MT (e.g., FI-RU, EN-FI, FI-RU, DE-RU), DISTIL$_{\mathrm{DistilmBERT}}$ can substantially outperform MT-IR (e.g., 9 MAP points for FI-RU and DE-RU); the gap in favor of MT-IR is, as expected, largest for most similar language pairs, for which also the most reliable MT systems exist (EN-IT, EN-DE). In other words, the feasibility and robustness of a strong MT-IR CLIR model seems to diminish with more distant language pairs and lower-resource language pairs. We plan to investigate this conjecture in more detail in future work.

The variation in results with similarity-specialized sentence encoders indicates that: (a) despite their seemingly similar high-level architectures typically based on dual-encoder networks [8], it is important to carefully choose a sentence encoder in document-level retrieval, and (b) there is an inherent mismatch between the granularity

[10] As expected, m-USE and DISTIL$_{\mathrm{USE}}$ perform poorly on language pairs involving Finnish, as they have not been trained on any Finnish data.

Table 2: Sentence-level CLIR results (MAP). **Bold**: best model for each language-pair. *: difference in performance with respect to Proc-B, significant at $p = 0.05$, computed via paired two-tailed t-test with Bonferroni correction.

	EN-FI	EN-IT	EN-DE	DE-FI	DE-IT	FI-IT	AVG	w/o FI
Baselines								
MT-IR	.659	.803	.725	.541	.694	.698	.687	.740
Proc-B	.143	.523	.415	.162	.342	.137	.287	.427
Models based on multilingual Transformers								
SEMB_{XLM}	.309*	.677*	.465	.391*	.495*	.346*	.447	.545
$\text{SEMB}_{\text{mBERT}}$	.199*	.570	.355	.231*	.481*	.353*	.365	.469
AOC_{XLM}	.099	.527	.274*	.102*	.282	.070*	.226	.361
$\text{AOC}_{\text{mBERT}}$	.095*	.433*	.274*	.088*	.230*	.059*	.197	.312
ISO_{XLM}	.016*	.178*	.053*	.006*	.017*	.002*	.045	.082
$\text{ISO}_{\text{mBERT}}$	.010*	.141*	.087*	.005*	.017*	.000*	.043	.082
Similarity-specialized sentence encoders (with parallel data supervision)								
$\text{DISTIL}_{\text{XLM-R}}$	.935*	.944*	.943*	.911*	.919*	.914*	.928	.935
$\text{DISTIL}_{\text{USE}}$	.084*	.960*	.952*	.137	.920*	.072*	.521	.944
$\text{DISTIL}_{\text{DistilmBERT}}$	.847*	.901*	.901*	.811*	.842*	.793*	.849	.882
LaBSE	.971*	.972*	.964*	.948*	.954*	.951*	.960	.963
LASER	**.974***	**.976***	**.969***	**.967***	**.965***	**.961***	**.969**	**.970**
m-USE	.079*	.951*	.929*	.086*	.886*	.039*	.495	.922

of information encoded by the current state-of-the-art text representation models and the document-level CLIR task.

5.2 Sentence-Level Cross-Lingual Retrieval

We show the sentence-level CLIR performance in Table 2. Unlike in the document-level CLIR task, self-supervised SEMB variants here manage to outperform Proc-B. The better relative SEMB performance than in document-level retrieval is somewhat expected: sentences are much shorter than documents (i.e., typically shorter than the maximal sequence length of 128 word pieces). All purely self-supervised mBERT and XLM variants, however, perform worse than the translation-based baseline.

Multilingual encoders specialized with parallel data excel in sentence-level CLIR, all of them substantially outperforming the competitive MT-IR baseline. This however, does not come as much of a surprise, since these models (a) have been trained using parallel data, and (b) have been optimized exactly on the sentence similarity task. In other words, in the context of the cross-lingual sentence-level task, these models are effectively supervised models. The effect of supervision is most strongly pronounced for LASER, which was, by being also trained on parallel data from Europarl, effectively subdued to in-domain training. We note that at the same time LASER was the weakest model from this group on average in the document-level CLIR task.

5.3 Further Analysis

We further investigate three aspects that may impact CLIR performance of multilingual encoders: (1) layer(s) from which we take vector representations, (2) number of contexts used in AOC variants, and (3) sequence length in document-level CLIR.

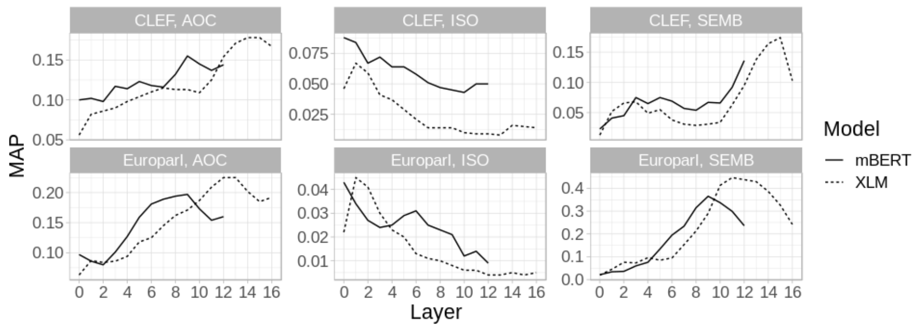

Fig. 1: CLIR performance of mBERT and XLM as a function of the Transformer layer from which we obtain the representations. Results (averaged over all language pairs) shown for all three encoding strategies (SEMB, AOC, ISO).

Layer Selection. All multilingual encoders have multiple layers and one may select (sub)word representations for CLIR at the output of any of them. Figure 1 shows the impact of taking subword representations after each layer for self-supervised mBERT and XLM variants. We find that the optimal layer differs across the encoding strategies (AOC, ISO, and SEMB) and tasks (document-level vs. sentence-level CLIR). ISO, where we feed the terms into encoders without any context, seems to do best if we take the representations from lowest layers. This makes intuitive sense, as the parameters of higher Transformer layers encode compositional rather than lexical semantics [17, 40]. For AOC and SEMB, where both models obtain representations by contextualizing (sub)words in a sentence, we get the best performance for higher layers – the optimal layers for document-level retrieval (L9/L12 for mBERT, and L15 for XLM) seem to be higher than for sentence-level retrieval (L9 for mBERT and L12/L11 for XLM).

Number of Contexts in AOC. We construct AOC term embeddings by averaging contextualized representations of the same term obtained from different Wikipedia contexts. This raises an obvious question of a sufficient number of contexts needed for a reliable (static) term embedding. Figure 2 shows the AOC results depending on the number of contexts used to induce the term vectors (cf. τ in §3). The AOC performance seems to plateau rather early – at around 30 and 40 contexts for mBERT and XLM, respectively. Encoding more than 60 contexts (as we do in our main experiments) would therefore bring only negligible performance gains.

Input Sequence Length. Multilingual encoders have a limited input length and they, unlike CLIR models operating on static embeddings (Proc-B, as well as our AOC and ISO variants), effectively truncate long documents. In our main experiments we truncated the documents to first 128 word pieces. Now we quantify (Table 3) if and to which extent this has a detrimental effect on document-level CLIR performance. Somewhat counterintuitively, encoding a longer chunk of documents (256 word pieces) yields a minor performance deterioration (compared to the length of 128) for *all* multilingual encoders. We suspect that this is a combination of two effects: (1) it is more difficult

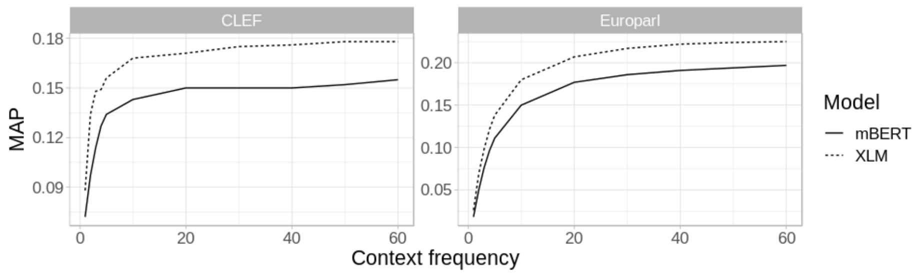

Fig. 2: CLIR performance of AOC variants (mBERT and XLM) w.r.t. the number of contexts used to obtain the term embeddings.

Table 3: Document CLIR results w.r.t. the input text length. Scores averaged over all language pairs not involving Finnish.

Length	$\text{SEMB}_{\text{mBERT}}$	SEMB_{XLM}	DIST_{use}	$\text{DIST}_{\text{XLM-R}}$	$\text{DIST}_{\text{DmBERT}}$	mUSE	LaBSE	LASER
64	.104	.128	.235	.167	.237	.254	.127	.089
128	.137	.178	.258	.162	.280	.247	.125	.068
256	.117	.158	.230	.146	.250	.197	.096	.027

to semantically accurately encode a longer portion of text, leading to semantically less precise embeddings of 256-token sequences; and (2) for documents in which the query-relevant content is not within the first 128 tokens, that content might often also appear beyond the first 256 tokens, rendering the increase in input length inconsequential to the recognition of such documents as relevant.

6 Conclusion

Pretrained multilingual encoders have been shown to be widely useful in natural language understanding (NLU) tasks, when fine-tuned in supervised settings on some task-specific data; their utility as general-purpose text encoders in unsupervised settings, such as the ad-hoc cross-lingual IR, has been less investigated. In this work, we systematically validated the suitability of a wide spectrum of cutting-edge multilingual encoders for document- and sentence-level CLIR across several language pairs. Our study invluded self-supervised multilingual encoders, mBERT and XLM, as well as the those that have been specialized for semantic text matching on semantic similarity datasets and parallel data. Opposing the findings from supervised NLU, we demonstrated that self-supervised multilingual encoders (mBERT and XLM), without exposure to task supervision, typically fail to outperform CLIR models based on cross-lingual word embeddings (CLWEs). Semantically-specialized multilingual sentence encoders, on the other hand, do outperform CLWEs, but the gains are pronounced only in the sentence retrieval task. While state-of-the-art multilingual text encoders excel in so many seemingly more complex language understanding tasks, our work renders ad-hoc CLIR in general and document-level CLIR in particular a serious challenge for these models. We make our code and resources available at `https://github.com/rlitschk/EncoderCLIR`.

Acknowledgments

The work of Ivan Vulić is supported by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no 648909). Robert Litschko and Goran Glavaš are supported by the Baden Württemberg Stiftung (Eliteprogramm, AGREE grant).

References

1. Artetxe, M., Labaka, G., Agirre, E.: A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In: Proceedings of ACL. pp. 789–798 (2018)
2. Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the ACL pp. 597–610 (2019)
3. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020)
4. Braschler, M.: CLEF 2003–Overview of results. In: Workshop of the Cross-Language Evaluation Forum for European Languages. pp. 44–63 (2003)
5. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: Proceedings of NeurIPS (2020)
6. Cao, S., Kitaev, N., Klein, D.: Multilingual alignment of contextual word representations. In: Proceedings of ICLR (2020)
7. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of SemEval. pp. 1–14 (2017)
8. Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., Kurzweil, R.: Universal sentence encoder for English. In: Proceedings of EMNLP. pp. 169–174 (2018)
9. Chidambaram, M., Yang, Y., Cer, D., Yuan, S., Sung, Y., Strope, B., Kurzweil, R.: Learning cross-lingual sentence representations via a multi-task dual-encoder model. In: Proceedings of the ACL Workshop on Representation Learning for NLP. pp. 250–259 (2019)
10. Clark, K., Luong, M., Le, Q.V., Manning, C.D.: ELECTRA: Pre-training text encoders as discriminators rather than generators. In: Proceedings of ICLR (2020)
11. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of ACL. pp. 8440–8451 (2020)
12. Conneau, A., Kiela, D.: SentEval: An evaluation toolkit for universal sentence representations. In: Proceedings of LREC. pp. 1699–1704 (2018)
13. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of EMNLP. pp. 670–680 (2017)
14. Conneau, A., Lample, G.: Cross-lingual language model pretraining. In: Proceedings of NeurIPS, pp. 7059–7069 (2019)
15. Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., Stoyanov, V.: XNLI: Evaluating cross-lingual sentence representations. In: Proceedings of EMNLP. pp. 2475–2485 (2018)
16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL. pp. 4171–4186 (2019)
17. Ethayarajh, K.: How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In: Proceedings of EMNLP-IJCNLP. pp. 55–65 (2019)

18. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic BERT sentence embedding. arXiv preprint arXiv:2007.01852 (2020)
19. 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: Proceedings of ACL. pp. 710–721 (2019)
20. Guo, M., Shen, Q., Yang, Y., Ge, H., Cer, D., Hernandez Abrego, G., Stevens, K., Constant, N., Sung, Y.H., Strope, B., Kurzweil, R.: Effective parallel corpus mining using bilingual sentence embeddings. In: Proceedings of WMT. pp. 165–176 (2018)
21. Hoogeveen, D., Verspoor, K.M., Baldwin, T.: CQADupStack: A benchmark data set for community question-answering research. In: Proceedings of ADCS. pp. 3:1–3:8 (2015)
22. Jiang, Z., El-Jaroudi, A., Hartmann, W., Karakos, D., Zhao, L.: Cross-lingual information retrieval with BERT. In: Proceedings of LREC. p. 26 (2020)
23. Karthikeyan, K., Wang, Z., Mayhew, S., Roth, D.: Cross-lingual ability of multilingual BERT: An empirical study. In: Proceedings of ICLR (2020)
24. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of the 10th Machine Translation Summit (MT SUMMIT). pp. 79–86 (2005)
25. Lei, T., Joshi, H., Barzilay, R., Jaakkola, T., Tymoshenko, K., Moschitti, A., Màrquez, L.: Semi-supervised question retrieval with gated convolutions. In: Proceedings of NAACL. pp. 1279–1289 (2016)
26. Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., et al.: XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In: Proceedings of EMNLP (2020)
27. Litschko, R., Glavaš, G., Vulić, I., Dietz, L.: Evaluating resource-lean cross-lingual embedding models in unsupervised retrieval. In: Proceedings of SIGIR. pp. 1109–1112 (2019)
28. Liu, Q., Kusner, M.J., Blunsom, P.: A survey on contextual embeddings. arXiv preprint arXiv:2003.07278 (2020)
29. Liu, Q., McCarthy, D., Vulić, I., Korhonen, A.: Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation. In: Proceedings of CoNLL. pp. 33–43 (2019)
30. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
31. MacAvaney, S., Soldaini, L., Goharian, N.: Teaching a new dog old tricks: Resurrecting multilingual retrieval using zero-shot learning. In: Proceedings of ECIR. pp. 246–254 (2020)
32. MacAvaney, S., Yates, A., Cohan, A., Goharian, N.: Cedr: Contextualized embeddings for document ranking. In: Proceedings of SIGIR. pp. 1101–1104 (2019)
33. Nogueira, R., Yang, W., Cho, K., Lin, J.: Multi-stage document ranking with BERT. arXiv preprint arXiv:1910.14424 (2019)
34. Pires, T., Schlinger, E., Garrette, D.: How multilingual is multilingual BERT? In: Proceedings of ACL. pp. 4996–5001 (2019)
35. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of SIGIR. pp. 275–281 (1998)
36. Ponti, E.M., Glavaš, G., Majewska, O., Liu, Q., Vulić, I., Korhonen, A.: XCOPA: A multilingual dataset for causal commonsense reasoning. In: Proceedings of EMNLP (2020)
37. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
38. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of EMNLP. pp. 3973–3983 (2019)
39. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: Proceedings of EMNLP (2020)

40. Rogers, A., Kovaleva, O., Rumshisky, A.: A primer in BERTology: What we know about how BERT works. Transactions of the ACL (2020)
41. Ruder, S., Vulić, I., Søgaard, A.: A survey of cross-lingual word embedding models. Journal of Artificial Intelligence Research **65**, 569–631 (2019)
42. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
43. Smith, S.L., Turban, D.H., Hamblin, S., Hammerla, N.Y.: Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In: Proceedings of ICLR (2017)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of NeurIPS. pp. 5998–6008 (2017)
45. Vulić, I., Glavas, G., Reichart, R., Korhonen, A.: Do we really need fully unsupervised cross-lingual embeddings? In: Proceedings of EMNLP. pp. 4406–4417 (2019)
46. Vulić, I., Moens, M.F.: Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In: Proceedings of SIGIR. pp. 363–372 (2015)
47. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of ICLR (2019)
48. Williams, A., Nangia, N., Bowman, S.: A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of NAACL. pp. 1112–1122 (2018)
49. Wu, S., Dredze, M.: Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In: Proceedings of EMNLP. pp. 833–844 (2019)
50. Yang, Y., Abrego, G.H., Yuan, S., Guo, M., Shen, Q., Cer, D., Sung, Y.H., Strope, B., Kurzweil, R.: Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In: Proceedings of AAAI. pp. 5370–5378 (2019)
51. 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: Proceedings of ACL: System Demonstrations. pp. 87–94 (2020)
52. Yang, Y., Hernandez Abrego, G., Yuan, S., Guo, M., Shen, Q., Cer, D., Sung, Y.h., Strope, B., Kurzweil, R.: Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In: Proceedings of IJCAI. pp. 5370–5378 (2019)
53. Yu, P., Allan, J.: A study of neural matching models for cross-lingual IR. In: Proceedings of SIGIR. p. 1637–1640 (2020)
54. Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al.: Big Bird: transformers for longer sequences. arXiv preprint arXiv:2007.14062 (2020)
55. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems (TOIS) **22**(2), 179–214 (2004)
56. Zhao, W., Eger, S., Bjerva, J., Augenstein, I.: Inducing language-agnostic multilingual representations. arXiv preprint arXiv:2008.09112 (2020)
57. Zhao, W., Glavaš, G., Peyrard, M., Gao, Y., West, R., Eger, S.: On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In: Proceedings of ACL. pp. 1656–1671 (2020)
58. Ziemski, M., Junczys-Dowmunt, M., Pouliquen, B.: The United Nations parallel corpus v1.0. In: Proceedings of LREC. pp. 3530–3534 (2016)
59. Zweigenbaum, P., Sharoff, S., Rapp, R.: Overview of the third BUCC shared task: Spotting parallel sentences in comparable corpora. In: Proceedings of LREC (2018)