SPECIAL ISSUE ON ECIR 2021

On cross-lingual retrieval with multilingual text encoders

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

Abstract

Pretrained multilingual text encoders based on neural *transformer architectures*, such as multilingual BERT (mBERT) and XLM, have recently become a default paradigm for cross-lingual transfer of natural language processing models, rendering cross-lingual word embedding spaces (CLWEs) effectively obsolete. 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 number of diverse language pairs. We first treat these models as multilingual text encoders and benchmark their performance in unsupervised ad-hoc sentence- and document-level CLIR. 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 multilingual encoders on average fail to significantly outperform earlier models based on CLWEs. For sentence-level retrieval, we do obtain state-of-the-art performance: the peak scores, however, are met by multilingual encoders that have been further specialized, in a supervised fashion, for sentence understanding tasks, rather than using their vanilla 'off-the-shelf' variants. Following these results, we introduce localized relevance matching for document-level CLIR, where we independently score a query against document sections. In the second part, we evaluate multilingual encoders fine-tuned in a supervised fashion (i.e., we *learn to rank*) on English relevance data in a series of zero-shot language and domain transfer CLIR experiments. Our results show that, despite the supervision, and due to the domain and language shift, supervised re-ranking rarely improves the performance of multilingual transformers as unsupervised base rankers. Finally, only with in-domain contrastive fine-tuning (i.e., same domain, only language transfer), we manage to improve the ranking quality. We uncover substantial empirical differences between cross-lingual retrieval results and results of (zero-shot) cross-lingual transfer for monolingual retrieval in target languages, which point to "monolingual overfitting" of retrieval models trained on monolingual (English) data, even if they are based on multilingual transformers.

Keywords Cross-lingual IR · Multilingual text encoders · Learning to Rank

✉ Goran Glavaš
  goran@informatik.uni-mannheim.de

Extended author information available on the last page of the article

🕲 Springer

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 reliable in-domain supervised models (e.g., neural matching models for pairwise retrieval Yu and Allan 2020; Jiang et al. 2020). This motivates the need for robust, resource-lean CLIR approaches: (1) unsupervised CLIR models and/or (2) transfer of supervised rankers across domains and languages, i.e., from resource-rich to resource-lean setups.

In previous work, Litschko et al. (2019) have shown that language transfer by means of 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 cross-lingual (in this case, bilingual) supervision (i.e., only a bilingual dictionary spanning 1–5 K word translation pairs), or even no bilingual supervision at all, in order to learn a mapping that aligns two monolingual word embedding spaces (Glavaš et al. 2019; Vulić et al. 2019). 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, and captures only "static" word-level semantics. *Contextual text representation models* alleviate this issue (Liu et al. 2020) because they encode occurrences of the same word differently depending on its context.

Such contextual dynamic representations are obtained via deep neural models pretrained on large text collections through general objectives such as (masked) language modeling (Devlin et al. 2019; Liu et al. 2019b). Multilingual text encoders pretrained on 100+ languages, such as multilingual BERT (mBERT) (Devlin et al. 2019) or XLM(-R) (Conneau and Lample 2019; Conneau et al. 2020a), 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 (Ponti et al. 2020; Liang et al. 2020): the general-purpose language knowledge obtained during pretraining is successfully specialized using task-specific training (i.e., fine-tuning). Multilingual transformers have been rendered especially effective in zero-shot transfer 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. The effectiveness of cross-lingual transfer with multilingual transformers, however, has more recently been shown to highly depend on the typological proximity between languages as well as the size of the pretraining corpora in the target language (Hu et al. 2020; Lauscher et al. 2020; Zhao et al. 2021a).

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., cross-lingual relevance judgments). Further, can they outperform unsupervised CLIR approaches based on static CLWEs (Litschko et al. 2019)? 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 Sect. 3) yield different retrieval results? Are there performance differences in unsupervised sentence-level versus document-level CLIR tasks? Can we boost performance by relying on sentence encoders that

are specialized towards dealing with sentence-level understanding in particular? Finally, can we improve ad-hoc CLIR in our target setups by fine-tuning multilingual encoders on relevance judgments from different document collections (i.e., domains) and languages (e.g., by exploiting available monolingual English relevance judgments from other collections)?

In order to address all 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, across unsupervised, supervised, and transfer setups. We evaluate state-of-the-art general-purpose pretrained multilingual encoders (mBERT Devlin et al. 2019 and XLM Conneau and Lample 2019) with a range of encoding variants, and also compare them to provenly robust CLIR approaches based on static CLWEs, as well as to specialized variants of multilingual encoders fine-tuned to encode sentence semantics (Artetxe et al. 2019; Feng et al. 2020; Reimers and Gurevych 2020, *inter alia*). Finally, we compare the unsupervised CLIR approaches based on these multilingual transformers with their counterparts fine-tuned on English relevance signal from different domains/collections. Our key contributions and findings are summarized as follows:

(1) We empirically validate (Sect. 4.2) 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).

(2) We show that multilingual sentence encoders, fine-tuned on labeled data from sentence pair tasks like natural language inference or semantic text similarity as well as using parallel sentences, substantially outperform general-purpose models (mBERT and XLM) in sentence-level CLIR (Sect. 4.3); further, they can be leveraged for localized relevance matching and in such a pooling setup improve the performance of unsupervised document-level CLIR (Sect. 4.4).

(3) Supervised neural rankers (also based on multilingual transformers like mBERT) trained on English relevance judgments from different collections (i.e., zero-shot language and domain transfer) do not surpass the best-performing unsupervised CLIR approach based on multilingual sentence encoders, either as standalone rankers or as re-rankers of the initial ranking produced by the unsupervised CLIR model based on multilingual sentence encoders (Sect. 5.1).

(4) In-domain fine-tuning of the best-performing unsupervised transformer (Reimers and Gurevych 2020) (i.e., zero-shot language transfer, no domain transfer)—yields considerable gains over the original unsupervised ranker (Sect. 5.2). This renders fine-tuning with little in-domain data more beneficial than transferring models trained on large-scale out-of-domain datasets.

(5) Finally, we show that fine-tuning supervised CLIR models based on multilingual transformers on monolingual (English) data leads to a type of "overfitting" to monolingual retrieval (Sect. 5.3): We empirically show that language transfer in IR is more difficult in true cross-lingual IR settings, in which query and documents are in different languages, as opposed to monolingual IR in a different (target) language.

This manuscript is an extension of the article *"Evaluating Multilingual Text Encoders for Unsupervised Cross-Lingual Retrieval"* published in the Proceedings of the 43rd European Conference on Information Retrieval (ECIR) (Litschko et al. 2021), where we evaluated multilingual encoders exclusively in unsupervised CLIR. In this work we, first and foremost, extend the scope of the work to supervised IR settings, and investigate how (English, in-domain or out-of-domain) relevance annotations can be leveraged to fine-tune

supervised rankers based on multilingual text encoders (e.g., multilingual BERT). To this end, we evaluate document-level CLIR performance of (1) two standard pointwise learning-to-rank (L2R) models based on multilingual BERT and trained on large-scale English corpora and (2) a multilingual encoder fine-tuned via contrastive metric-based learning on small in-domain relevance dataset; we demonstrate that only the latter offers consistent performance gains over unsupervised CLIR with the same multilingual encoders. Pointwise L2R and contrastive fine-tuning models are described in Sect. 3.4. Section 5 provides detailed experimental evaluation of those models on several document-level CLIR tasks.

We believe that this extensive empirical study offers plenty of valuable new insights for researchers and practitioners who work in the challenging landscape of cross-lingual information retrieval tasks.

2 Related work

Self-Supervised Pretraining and Transfer Learning Recently, research on universal sentence representations and transfer learning has gained much traction. InferSent (Conneau et al. 2017) transfers the encoder of a model trained on natural language inference to other tasks, while USE (Cer et al. 2018) extends this idea to a multi-task learning setting. More recent work explores self-supervised neural Transformer-based (Vaswani et al. 2017) models, all based on (causal or masked) language modeling (LM) objectives, such as BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019b), GPT (Radford et al. 2019; Brown et al. 2020), and XLM (Conneau and Lample 2019).[1] Results on benchmarks such as GLUE (Wang et al. 2019) and SentEval (Conneau and Kiela 2018) 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 (MacAvaney et al. 2019), or fine-tuned as pointwise and pairwise rankers (Nogueira et al. 2019).

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 (Conneau and Lample 2019) extends this idea and proposes natively cross-lingual LM pretraining, combining causal language modeling (CLM) and translation language modeling (TLM).[2] Strong performance of these models in supervised settings is confirmed across a range of tasks on multilingual benchmarks such as XGLUE (Liang et al. 2020) and XTREME (Hu et al. 2020). However, recent work Reimers and Gurevych (2020) and Cao et al. (2020) has indicated that these general-purpose models do not yield strong results when used as out-of-the-box text encoders in an unsupervised

---

[1] Note that self-supervised learning can come in different flavors depending on the training objective (Clark et al. 2020), but language modeling objectives still seem to be the most popular choice.

[2] 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.

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. (2020) use mBERT as a matching model by feeding pairs of English queries and foreign language documents. MacAvaney et al. (2020b) 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.

Specialized Multilingual Sentence Encoders An extensive body of work focuses on inducing multilingual encoders that capture sentence meaning. In Artetxe et al. (2019), the multilingual encoder of a sequence-to-sequence model is shared across languages and optimized to be language-agnostic, whereas Guo et al. (2018) rely on a dual Transformer-based encoder architecture 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 Yang et al. (2019a), this approach is extended by using a bidirectional dual encoder and adding an additive margin softmax function, which serves to push away non-translation-pairs in the shared embedding space. The dual-encoder approach is now widely adopted (Guo et al. 2018; Yang et al. 2020; Feng et al. 2020; Reimers and Gurevych 2020; Zhao et al. 2021b), and yields state-of-the-art multilingual sentence encoders which excel in sentence-level NLU tasks.

Other recent approaches propose input space normalization, and using parallel data to re-align mBERT and XLM (Zhao et al. 2021b; Cao et al. 2020), 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 (Reimers and Gurevych 2020). In Yang et al. (2020), authors combine multi-task learning with a translation bridging task to train a universal sentence encoder. We benchmark a series of representative sentence encoders in this article; their brief descriptions are provided in Sect. 3.3.

Neural Learning-to-Rank In the context of neural retrieval the vast majority of rankers can be broadly classified into the two paradigms of (1) Cross-Encoders (2) and Bi-Encoders (Humeau et al. 2020; Thakur et al. 2021; Qu et al. 2021). Cross-Encoders compute the full interaction between pairs of queries and documents and induce a joint representation for a query-document pair by means of cross-attention. Transformed representation of the query-document pair is then fed to a relevance classifier; the encoder and classifier parameters are updated jointly in an end-to-end fashion (Nogueira et al. 2019; MacAvaney et al. 2020b; Khattab and Zaharia 2020). This paradigm is usually impractical for end-to-end ranking due to slow matching and retrieval. Recent work addresses this challenge by performing late interaction and by precomputing token-level representations (Khattab and Zaharia 2020; Gao et al. 2020). Nonetheless, neural rankers are still predominantly used for re-ranking the top-ranked results returned by some base ranker. The alternative paradigm— the so-called Bi-Encoders—computes vector representations of documents and queries independently; it then relies on fast similarity computations in the vector space of precomputed query and document embeddings. All similarity-specialized multilingual encoders described in Sect. 3.3 belong to this category of Bi-Encoders. Contrary to most NLP tasks, document-level ad-hoc IR deals with much longer text sequences. For instance, one notable approach computes document scores as an interpolation between a pre-ranking score and a weighed sum of scores of the top-k highest scoring sentences (Akkalyoncu Yilmaz et al. 2019). Our approach scores local regions of documents independently (Sect. 4.4); this is most similar to the BERT-MaxP model which encodes and scores individual passages of

a document (Dai and Callan 2019). For further discussion on long document matching we refer the reader to Chapter 3.3 of Lin et al.'s handbook (Lin et al. 2021).

A related recent line of research targets *cross-lingual transfer of (monolingual) rankers*, where such rankers are typically trained on English data and then applied in a monolingual non-English setting (Shi et al. 2020, 2021; Zhang et al. 2021). This is different from our *cross-lingual retrieval* evaluation setting where queries and documents are in different languages. 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.

CLIR Evaluation and Application The cross-lingual ability of mBERT and XLM has been investigated by probing and analyzing their internals (Karthikeyan et al. 2020), as well as in terms of downstream performance (Pires et al. 2019; Wu and Dredze 2019). In CLIR, these models as well as dedicated multilingual sentence encoders have been evaluated on tasks such as cross-lingual question-answer retrieval (Yang et al. 2020), bitext mining (Ziemski et al. 2016; Zweigenbaum et al. 2018), and semantic textual similarity (STS) (Hoogeveen et al. 2015; Lei et al. 2016). 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, or the impact of domain versus language transfer. In this work, we aim to fill these gaps.

3 Multilingual text encoders

We first provide an overview of all pretrained multilingual models in our evaluation. We discuss general-purpose multilingual text encoders (Sect. 3.2), as well as specialized multilingual sentence encoders in Sect. 3.3. Finally, we describe the supervised rankers based on multilingual encoders (Sect. 3.4). For completeness, we first briefly describe the baseline CLIR model based on CLWEs (Sect. 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 $\vec{Q}, \vec{D} \in \mathbb{R}^d$ in a d-dimensional shared embedding space (Vulić and Moens 2015; Litschko et al. 2019). Each term is represented independently with a pre-computed static embedding vector $\vec{t}_i = emb(t_i)$. 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 (Ruder et al. 2019). Given the shared CLWE space, both query and document representations are obtained as aggregations of their term embeddings. We follow Litschko et al. (2019) and represent documents as the weighted sum of their terms' vectors, where each term's weight corresponds to its inverse document frequency (idf):[3]

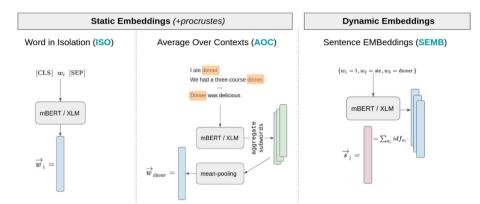[3] Only document term embeddings are idf-scaled.

Fig. 1 CLIR Models based on Multilingual Transformers. **Left**: Induce a static embedding space by encoding each vocabulary term in isolation; then refine the bilingual space for a specific language pair using the standard Procrustes projection. **Middle**: Aggregate different contextual representations of the same vocabulary term to induce static embedding space; then refine the bilingual space for a specific language pair using the standard Procrustes projection. **Right**: Direct encoding of a query-document pair with the multilingual encoder

$$\vec{d} = \sum_{i=1}^{N_d} idf(t_i^d) \cdot \vec{t_i^d} \qquad (1)$$

Documents are then ranked in decreasing order of the cosine similarity between their embeddings and the query embedding.

3.2 Multilingual (transformer-based) language models: mBERT and XLM

Massively multilingual pretrained neural language models such as mBERT and XLM(-R) 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 ($K \geq 1$; 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: $\vec{t_i} = \psi \left(\{ \overrightarrow{wp_{i,k}} \}_{k=1}^{K} \right)$, where the function $\psi()$ is the aggregation function over the K constituent subword embeddings. Once these term embeddings $\vec{t_i}$ are obtained, we follow the same CLIR setup as with CLWEs in Sect. 3.1. We illustrate three different approaches for obtaining word and sentence representations from pretrained transformers in Fig. 1 and describe them in more detail in what follows.

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 Sect. 3.1, and allows the CLIR model (Sect. 3.1) to operate at a non-contextual word level. An empirical CLIR comparison between ISO and CLIR operating on traditionally induced CLWEs (Litschko

et al. 2019) then effectively quantifies how well multilingual encoders (mBERT and XLM) capture word-level representations (Vulić et al. 2020).

In the second, 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 the term t occurs. We use the full set of Wikipedia sentences $\mathcal{S}$ to sample sets of contexts $\mathcal{S}_t$ for each vocabulary term t. 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, $\psi\left(\{\overrightarrow{wp_{j,k}}\}_{k=1}^{K}\right) = 1/K \cdot \sum_{k=1}^{K} \overrightarrow{wp_{j,k}}$. Here, the function $\psi()$ is implemented as *mean-pooling*, i.e., we obtain the contextualized representation of the term as the average of contextualized vectors of its constituent subwords. 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 lexically well-aligned across languages (Liu et al. 2019a; Cao et al. 2020). 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 (Glavaš et al. 2019). We observed poor BLI performance, suggesting that further projection-based alignment of respective monolingual ISO and AOC spaces is warranted. To this end, we adopted the standard Procrustes method (Smith et al. 2017; Artetxe et al. 2018) for learning an orthogonal linear projection from the embedding (sub)space of one language to the embedding space of the other language (Glavaš et al. 2019). Let $D = \{(w_{L1}^k, w_{L2}^k)\}_{k=1}^{K}$ be the word translation dictionary between the two languages $L1$ and $L2$, containing K word translation pairs. Let $\mathbf{X}_S = \{\mathbf{x}_{L1}^k\}_{k=1}^{K}$ and $\mathbf{X}_T = \{\mathbf{x}_{L2}^k\}_{k=1}^{K}$ be row-aligned matrices containing stacked embeddings of $\{w_{L1}^k\}_{k=1}^{K}$ and $\{w_{L2}^k\}_{k=1}^{K}$, respectively. We then obtain the projection matrix $\mathbf{W}$ by minimizing the Euclidean distance between the projection of $\mathbf{X}_S$ and the target matrix $\mathbf{X}_T$ (Mikolov et al. 2013): $\mathbf{W} = \arg\min_{\mathbf{W}} \|\mathbf{X}_{L1}\mathbf{W} - \mathbf{X}_{L2}\|_2$. If we constrain $\mathbf{W}$ to be orthogonal, the above optimization problem becomes the famous Procrustes problem, with the following closed-form solution (Schönemann 1966):

$$\mathbf{W} = \mathbf{U}\mathbf{V}^{\top}, \text{ with}$$
$$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top} = SVD(\mathbf{X}_T\mathbf{X}_S^{\top}). \tag{2}$$

In our experiments, for each language pair, we always project the AOC (ISO) embeddings of the query language to the AOC (ISO) embedding space of the document collection language, using the learned projection matrix $\mathbf{W}$.

Direct Text Embedding with Multilingual Transformers In both AOC and ISO, we exploit the multilingual (contextual) encoders to obtain the static embeddings for word types (i.e., terms): we can then leverage these static word embeddings obtained from contextualized encoders in exactly the same ad-hoc CLIR setup (Sect. 3.1) in which CLWEs had previously been evaluated (Litschko et al. 2019). In an arguably more straightforward approach, we also use pretrained multilingual Transformers (i.e., mBERT or XLM) to directly semantically encode the whole input text similar to encoding sentences into *Sentence EMBeddings* (**SEMB**). To this end, 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 Sect. 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., $\vec{t_i} = \psi\left(\{\overrightarrow{wp_{i,k}}\}_{k=1}^{K}\right) = \overrightarrow{wp_{i,1}}$.[4]

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 (Reimers and Gurevych 2020; Cao et al. 2020; Zhao et al. 2020). 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 off-the-shelf 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) Artetxe et al. (2019) 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 the encoder's last layer. The decoder then takes the sentence embedding as additional input at 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 (Yang et al. 2020). It relies on a standard dual-encoder neural framework (Chidambaram et al. 2019; Yang et al. 2019b) 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) Feng et al. (2020) is another neural dual-encoder framework, also trained with parallel data. Unlike LASER and m-USE, where the encoders are trained from scratch on parallel data, LaBSE starts its training 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 modeling (MLM and TLM, see Sect. 2). For further details, we refer the reader to the original work.

*DISTIL*Reimers and Gurevych (2020) is a teacher-student framework for injecting the knowledge obtained through specialization for semantic similarity from a specialized monolingual transformer (e.g., BERT) into a non-specialized multilingual transformer (e.g., mBERT). It first specializes for semantic similarity a monolingual (English) teacher

---

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

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 (Reimers and Gurevych 2019), BERT specialized for embedding sentence meaning on semantic text similarity (Cer et al. 2017) and natural language inference (Williams et al. 2018) 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.

3.4 Learning to (re-)rank with multilingual encoders

Finally, we consider another common setup, in which some relevance judgments (typically in English) are available and can be leveraged as supervision for fine-tuning multilingual encoders for ad-hoc retrieval. We consider two common scenarios: (1) an abundance of relevance annotations from other retrieval tasks and collections (but none for the target collection on which we want to perform ad-hoc retrieval) and (2) a small number of relevance judgments for the target collection. As an example of the former, we apply pointwise rankers pretrained on large-scale data (and based on multilingual encoders) in document-level CLIR on the CLEF benchmark. For the latter, we use a small number of CLEF relevance judgments to fine-tune, via contrastive metric-based learning, the representation space of the multilingual encoder. These two fine-tuning approaches are described in what follows.

Pointwise Ranking with Multilingual Transformers A common learning-to-rank (L2R) approach with pretrained neural text encoders is the pointwise classification of query-document pairs (Nogueira et al. 2019; MacAvaney et al. 2020b). In this so-called Cross-Encoder approach, the input to the pretrained encoder is a query-document concatenation. More specifically, let query q consist of the query (subword) tokens $t_1^q, \ldots t_n^q$ and document d consist of the document (subword) tokens $t_1^d, \ldots t_m^d$. The input to the pretrained encoder is then [CLS] $t_1^q, \ldots t_n^q$ [SEP] $t_1^d, \ldots t_m^d$ [SEP], with [CLS] and [SEP] being the special sequence start and segment separation tokens of the corresponding pretrained encoder, e.g., BERT (Devlin et al. 2019). When needed, the documents are truncated in order to meet the maximum input length constraint of the respective pretrained transformer. This setup—i.e., concatenation of two texts—is common for various sentence-pair classification tasks in natural language processing (e.g., natural language inference or semantic text similarity). The encoded representation of the sequence start token ([CLS]), taken from the last layer of the Transformer-based encoder is then fed into a feed-forward classifier with a single hidden layer, which outputs the probability of the document being relevant for the query. The parameters of the feed-forward classifier are (fine-)tuned together with the encoder's parameters in an end-to-end fashion, by means of minimizing the standard cross-entropy loss. The positive training instances are simply the available relevance judgments (i.e., queries paired with documents indicated as relevant); the non-trivial negative instances are commonly created by pairing queries with irrelevant documents that are ranked highly by some baseline ranker (e.g., BM25) (Nogueira et al. 2019).

Pointwise neural rankers have been shown both ineffective (many false positive) and inefficient (at inference, one has to feed the query paired with each document through the classifier) when used to rank the entire document collection from scratch. In contrast, they have been very successful in re-ranking the top of the ranking produced by some baseline ranker, such as BM25. In CLIR, however, due to the very limited lexical overlap between languages, one cannot use base rankers based on lexical overlap such as BM25 or the vector space model (VSM). In our re-ranking experiments (see Sect. 5.1) we thus employ our unsupervised CLIR rankers based on multilingual encoders from Sect. 3.3 as base rankers.

Contrastive Metric-Based Learning The above pointwise approach which *cross-encodes* each query-document pair (by concatenating the query with each document and passing them jointly to the encoder) is computationally heavy. Therefore, as mentioned before, it is primarily used for re-ranking. Further, it introduces additional trainable parameters of the classifier: their reliable estimation requires a large amount of training instances. In contrast, in most ad-hoc retrieval setups, one at best has a handful of relevance judgments for the test collection of interest. An alternative approach in such low-supervision settings is to use the few available relevance judgments to reshape the representation space of the (multilingual) text encoder, without training a dedicated relevance classifier (i.e., no additional trainable parameters). In this so called Bi-Encoder paradigm, the objective is to bring representations of queries, produced independently by the pretrained encoder, closer to the representations of their relevant documents (produced again independently by the same encoder) than to the representations of irrelevant documents. The objectives of contrastive metric-based learning push the instances that stand in a particular relation (e.g., query and *relevant* document) closer together according to a predefined similarity or distance metric (e.g., cosine similarity) than corresponding pairs that do not stand in the relation of interest (e.g., the same query and some irrelevant document). It is precisely the approach used for obtaining multilingual encoders specialized for sentence similarity tasks covered in Sect. 3.3 (Reimers and Gurevych 2019; Feng et al. 2020; Yang et al. 2020).

We propose to use contrastive metric-based learning to fine-tune the representation space for the concrete ad-hoc retrieval task, using a limited amount of relevance judgments available for the target collection. To this end, we employ a popular contrastive learning objective referred to as Multiple Negative Ranking Loss (MNRL) (Thakur et al. 2021). Given a query vector q_i, a relevant document d_i^+ and a set of in-batch negatives $\{d_{i,j}^-\}_{j=1}^m$ we fine-tune the parameters of a pretrained multilingual encoder by minimizing MNRL, given as:

$$\mathcal{L}\left(q_i, d_i^+, \{d_{i,j}^-\}_{j=1}^m\right) = -\log \frac{e^{\lambda \cdot \text{sim}(q_i, d_i^+)}}{e^{\lambda \cdot \text{sim}(q_i, d_i^+)} + \sum_{j=1}^m e^{\lambda \cdot \text{sim}(q_i, d_{i,j}^-)}}$$

Each document, the relevant d_j^+ and each of the irrelevant $d_{i,j}^-$, receives a score that reflects their similarity to the query q_i: for this, we rely on cosine similarity, i.e. $\text{sim}(q_i, d_j) = \cos(q_i, d_j)$. Document scores, scaled with a temperature factor λ, are then converted into a probability distribution with a softmax function. The loss is then, intuitively, the negative log likelihood of the relevant document d_j^+. In Sect. 5.2, we fine-tune in this manner the best-performing multilingual encoder (see Sect. 4.2).

Table 1 Basic statistics of CLEF 2003 and Europarl test collections: number of documents (#doc); average number of tokens produced by the XLM/mBERT tokenizer (#xlm, #mbert); average number of relevant documents per query (#rel)

Lang.	CLEF 2003				Europarl			
	#doc	#rel	#mbert	#xlm	#doc	#rel	#mbert	#xlm
EN	169k	18.6	700.4	746.6	–	–	–	–
DE	295k	32.6	490.9	518.7	100k	1	35.6	38.3
IT	158k	15.9	482.8	491.8	100k	1	41.4	38.2
FI	55k	10.7	648.7	623.7	100k	1	37.6	38.1
RU	17k	5.4	557.8	536.3	–	–	–	–

4 Unsupervised CLIR

We first present the experiments demonstrating the suitability of pretrained multilingual models as text encoders for ad-hoc unsupervised CLIR (i.e., we evaluate models described in Sects. 3.2 and 3.3).

4.1 Experimental setup

Evaluation Data We follow the experimental setup of Litschko et al. (2019), and compare the models from Sect. 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 (Braschler 2003),[5] with 60 queries per language pair. For sentence-level retrieval, also following Litschko et al. (2019), for each language pair we sample from Europarl (Koehn 2005) 1K source language sentences as queries and 100K target language sentences as the "document collection". We refer the reader to Table 1 for summarystatistics.[6]

 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 (Litschko et al. 2019), dubbed Proc-B. Proc-B induces a bilingual CLWE space from pretrained monolingual FASTTEXT embeddings[7] 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 Sect. 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 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

---

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

[6] 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.

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

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 (Ponte and Croft 1998) with Dirichlet smoothing (Zhai and Lafferty 2004).

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 Sect. 4.5. In all cases (SEMB, ISO, AOC), we surround the input with the special sequence start and end tokens of respective pretrained models: [*CLS*] and [*SEP*] for BERT-based models and $\langle s \rangle$ and $\langle /s \rangle$ for XLM-based models. For vanilla multilingual encoders (mBERT and XLM) and all three variants (SEMB, ISO, AOC), we independently evaluate representations from different Transformer layers (cf. Sect. 4.5). 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) DISTIL$_{\text{XLM-R}}$ initializes the student model with the pretrained XLM-R transformer (Conneau et al. 2020b); DISTIL$_{\text{USE}}$ instantiates the student as the pretrained m-USE instance (Yang et al. 2020); whereas DISTIL$_{\text{DistilmBERT}}$ distils the knowledge from the Sentence-BERT teacher into a multilingual version of DistilBERT (Sanh et al. 2019), a 6-layer transformer pre-distilled from mBERT.[8] 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.

4.2 Document-level CLIR results

We show the performance (MAP) of multilingual encoders on document-level CLIR tasks in Table 2. 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 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 Proc-B$_{\text{LEN}}$. This suggests that the document-level retrieval could benefit from encoders able to encode longer portions of text, e.g., Beltagy et al. (2020) and Zaheer et al. (2020). 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 DISTIL$_{\text{DistilmBERT}}$, dubbed 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 and AOC demonstrates the importance of contextual information and seemingly limited usability of

---

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

Table 2 Document-level CLIR results (Mean Average Precision, MAP)

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

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

off-the-shelf multilingual encoders as word encoders, if no context is available, and if they are not further specialized to encode word-level information (Liu et al. 2021).

Similarity-specialized multilingual encoders, which rely on pretraining with parallel data, yield mixed results. Three models, DISTIL$_{DistilmBERT}$, DISTIL$_{USE}$ and m-USE, generally outperform the Proc-B baseline.[9] 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 subjected to any self-supervised pretraining like DISTIL models. Even the best-performing CLIR model based on a multilingual encoder (DISTIL$_{DistilmBERT}$) overall falls behind the MT-based baseline (MT-IR). However, it is very important to note that the performance of MT-IR critically 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$_{DistilmBERT}$ can substantially outperform MT-IR (e.g., 9 MAP points for FI-RU and DE-RU). In contrast, the gap in favor of MT-IR is, as expected, largest for the pairs of large typologically similar languages, 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 languages.

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 (Cer et al. 2018), it is important to carefully choose a sentence encoder in document-level retrieval, and (b) there is an inherent mismatch between the granularity of information encoded by the current state-of-the-art text representation models and the document-level CLIR task.

4.3 Sentence-level cross-lingual retrieval

We show the sentence-level CLIR performance in Table 3. 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 sentence 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 (i.e., sentence translations), 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, being also trained on parallel data from Europarl, effectively subjected 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.

---

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

Table 3 Sentence-level CLIR results (MAP)

	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
SEMB$_{mBERT}$	.199*	.570	.355	.231*	.481*	.353*	.365	.469
AOC$_{XLM}$	.099	.527	.274*	.102*	.282	.070*	.226	.361
AOC$_{mBERT}$	.095*	.433*	.274*	.088*	.230*	.059*	.197	.312
ISO$_{XLM}$	.016*	.178*	.053*	.006*	.017*	.002*	.045	.082
ISO$_{mBERT}$	.010*	.141*	.087*	.005*	.017*	.000*	.043	.082
Similarity-specialized sentence encoders (with parallel data supervision)								
DISTIL$_{XLM-R}$	.935*	.944*	.943*	.911*	.919*	.914*	.928	.935
DISTIL$_{USE}$	.084*	.960*	.952*	.137	.920*	.072*	.521	.944
DISTIL$_{DistilmBERT}$	.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

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

The fact that similarity-specialized multilingual encoders perform much better in sentence-level than in document-level CLIR suggests viability of a different approach to document-level retrieval: instead of obtaining a single encoding for the document, one may (independently) encode its sentences (or larger windows of content) and (independently) measure their semantic correspondence to the query. We investigate this *localized relevance matching* approach to document-level CLIR with similarity-specialized multilingual encoders in the next section (Sect. 4.4).

4.4 Localized relevance matching

Contrary to most NLP tasks, in ad-hoc document retrieval we face the challenge of semantically representing long documents. According to Robertson et al. (1994), documents can be viewed either as a concatenation of topically heterogeneous short sub-documents (*"Scope Hypothesis"*) or as a more verbose version of a short document on the same topic (*"Verbosity Hypothesis"*). Under both hypotheses, the source of relevance of the document for the query is localized, i.e., there should exist (at least one) segment (relatively short w.r.t. the length of the whole document) that is the source of relevance of the document for the query. Furthermore, a query may represent an information need on a specific aspect of a topic that is simply not discussed at the beginning, but rather somewhere later in the document: the maximum input sequence length imposed by neural text encoders directly limits the retrieval effectiveness in such cases. Even if we assume that we can encode the complete document with our multilingual encoders, these document representations would

likely become semantically less precise (i.e., fuzzier) as they would aggregate contextualized representations of many more tokens; in Sect. 4.5 we validate this empirically and show that simply increasing the maximum sequence length of multilingual encoders does not improve their retrieval performance.

Recent work proposed pretraining procedures for encoding long documents (Zaheer et al. 2020; Dai et al. 2019; Beltagy et al. 2020). These models have been pretrained only for English. Pretraining their multilingual counterparts, however, would require extremely large and massively multilingual corpora and computational resources of the scale that we do not have at our disposal. In the following, we instead experiment with two resource-lean alternatives: we represent documents either as (1) sets of overlapping text *segments* obtained by running a sliding window over the document or (2) a collection of document *sentences*, which we then encode independently similar to Akkalyoncu Yilmaz et al. (2019). For a single document, we now need to store multiple semantic representations (i.e., embeddings), one for each text segment or sentence. While these approaches clearly increase the index size as well as the retrieval latency (as the query representation needs to be compared against embeddings of all document segments or sentences), sufficiently fast ad-hoc retrieval for most use cases can still be achieved with highly efficient approximate search libraries such as FAISS (Johnson et al. 2017). Representing documents as multiple segments or sentences allows for fine-grained local matching against the query: a setting in which sentence-specialized multilingual encoders are supposed to excel, see Table 3.

Localized Relevance Matching: Segments. In this approach, we slide a window of size 128 word tokens over the document with a stride of 42 tokens, creating multiple overlapping 128-word segments from the input document. Each segment is then encoded separately, leveraging the encoders from Sect. 3. We then score for relevance each segment by comparing its respective embedding with the query embedding. We then compute the final relevance score by averaging the relevance scores of the top-k highest-scoring segments.

Table 4 displays the results of all multilingual encoders in our comparison, for $k \in \{1, 2, 3, 4\}$.[10] For most encoders (with the exception of LaBSE and the Proc-B baseline) we observe gains from segment-based localized relevance matching: we observe the largest average gain of 3.25 MAP points for $DISTIL_{XLM-R}$ (from 0.177 for document encoding to 0.209 for segment-based localized relevance matching). Most importantly, we observe gains for our best-performing multilingual encoder $DISTIL_{DmBERT}$: localized relevance matching (for $k = 2$) pushes its performance by 1.6 MAP points (the base performance of 0.28 is shown in Table 2). We suspect that applying IDF-Sum in Proc-B (see Sect. 3.1) has a similar (albeit query-independent) soft filtering effect to localized relevance matching and that this is why localized relevance matching does not yield any gains for this competitive baseline.

For all five multilingual encoders for which we observe gains from localized relevance matching, these gains are the largest for $k = 2$, i.e., when we average the relevance scores of the two highest-scoring segments. In 63.7% of the cases, the two highest-scoring segments are mutually consecutive, overlapping segments: we speculate that in those cases it is the span of text in which they overlap that contains the signal that makes the document relevant for the query. These findings are in line with similar observations from previous work Akkalyoncu Yilmaz et al. (2019) and Dai and Callan (2019): aggregating local relevance signals yields strong retrieval results. Matching queries with the most similar

---

[10] For $k = 1$, the relevance of the document is exactly the score of the highest scoring segment.

Table 4 Document-level CLIR results for *localized relevance matching* against document *segments* (overlapping 128-token segments)

	k	EN-FI	EN-IT	EN-RU	EN-DE	DE-FI	DE-IT	DE-RU	FI-IT	FI-RU	AVG	Δ AVG
Proc-B	1	.242	.253	.182	.286	.280	.217	.158	.147	.166	.215	−0.86
	2	.241	.244	.153	.287	.282	.207	.116	.147	.115	.199	−2.40
	3	.234	.235	.150	.277	.269	.194	.113	.153	.109	.193	−3.04
	4	.228	.217	.135	.255	.276	.171	.105	.167	.098	.184	−3.95
DISTIL$_{\text{DmBERT}}$	1	.330	.327	.248	.365	.324	.293	.244	.268	.236	.293	+1.32
	2	**.349**	.315	**.269**	.382	**.347**	.287	.216	**.272**	**.226**	**.296**	+1.61
	3	.323	.291	.261	.353	.335	.268	.226	.248	.208	.279	−0.03
	4	.299	.263	.207	.330	.316	.236	.189	.217	.181	.249	−3.10
DISTIL$_{\text{XLM-R}}$	1	.284	.218	.160	.233	.267	.195	.162	.181	.156	.206	+2.92
	2	.279	.208	.164	.253	.264	.194	.179	.187	.157	.209	+3.25
	3	.264	.191	.141	.228	.253	.188	.145	.171	.157	.193	+1.60
	4	.236	.169	.105	.203	.237	.167	.114	.153	.113	.166	−1.07
DISTIL$_{\text{USE}}$	1	.149	.355	.202	.363	.138	.332	.199	.074	.118	.214	+1.64
	2	.162	**.377**	.192	.416	.136	**.344**	.197	.081	.095	.222	+2.42
	3	.150	.344	.180	.391	.137	.319	.181	.079	.091	.208	+1.00
	4	.135	.313	.163	.364	.128	.280	.158	.064	.086	.188	−1.03
LaBSE	2	.212	.118	.102	.189	.199	.103	.060	.085	.083	.128	−1.13
	1	.221	.108	.124	.141	.198	.093	.077	.063	.143	.130	−1.32
	3	.198	.104	.080	.153	.190	.089	.052	.076	.066	.112	−2.90
	4	.186	.088	.065	.128	.176	.075	.036	.069	.049	.097	−4.42
mUSE	1	.073	.345	.215	.361	.082	.331	.210	.053	.084	.195	+1.19
	2	.102	.370	.213	**.404**	.085	**.344**	.209	.056	.085	.208	+2.46
	3	.083	.333	.198	.376	.074	.296	.186	.053	.082	.187	+0.38
	4	.075	.291	.178	.348	.067	.257	.178	.047	.077	.169	−1.43

Table 4 (continued)

	k	EN-FI	EN-IT	EN-RU	EN-DE	DE-FI	DE-IT	DE-RU	FI-IT	FI-RU	AVG	Δ AVG
LASER	1	.135	.058	.049	.075	.155	.054	.070	.082	.061	.082	+1.40
	2	.150	.069	.071	.099	.161	.055	.060	.088	.062	.091	+2.26
	3	.136	.054	.053	.074	.142	.044	.052	.072	.049	.075	+0.71
	4	.113	.037	.038	.057	.118	.032	.045	.052	.038	.059	−0.91

Document relevance is the average of relevance scores of *k* highest-scoring segments. Results (for 9 language pairs from CLEF) shown for the Proc-B baseline and all similarity-specialized encoders. **Bold**: Best model for each language-pair. Δ AVG denotes relative performance increases/decreases w.r.t. the respective base performances from Table 2
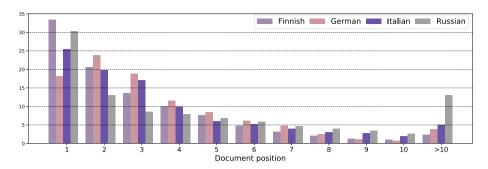
Fig. 2 Comparison of within-document positions of top-ranked segments in segment-based localized relevance matching for different *collection languages*. Proportions aggregated across all multilingual CLIR models from Table 4
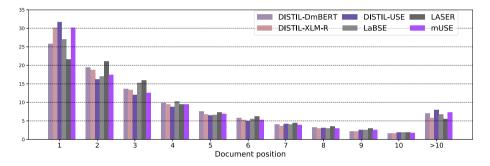

Fig. 3 Comparison of within-document positions of top-ranked segments in segment-based localized relevance matching for different *multilingual text encoders*. Proportions aggregated across all multilingual CLIR models from Table 4

segment embedding effectively filters out the rest of the document. Our results suggest that improvements are mostly consistent across language pairs: we only fail to observe gains when Russian is the language of the target document collection. Localized relevance matching can in principle decrease the performance if segmentation produces (many) false positives (i.e., irrelevant segments with high semantic similarity with the query). We suspect this to more often be the case for Russian than for the other languages. We further investigate this by comparing positions of high-scoring segments across document collection languages. We look at the distributions of document positions among the top-ranked 100 segments (gathered from all collection documents): the distributions of top-ranked segment results per positions in respective documents (i.e., 1 indicates the first segment of the document, 2 the second, etc.) are shown for each of the four collection languages (aggregated across all multilingual encoders from Table 4) in Fig. 2.

The distributions of positions of high-scoring segments confirms our suspicion that something is different for Russian compared to other languages: we observe a much larger presence of high-scoring segments that appear later in the documents, i.e., at positions larger than 10 (>10): while there is between 2% and 5% of such "late" high-scoring segments in Italian, German, and Finnish collections, in the Russian collection there is 13% of such segments. Our manual inspection confirmed that these late segments are indeed most

often false positives (i.e., irrelevant for the query, yet with representations highly similar to those of the queries): this presumably causes the lower performance on *-RU benchmarks.

Figure 3 compares the individual multilingual encoders along the same dimension: document positions of the segments they rank the highest. Unlike for collection languages, we do not observe major differences across multilingual encoders—for all of them, the top-ranked segments seem to have similar within-document position distributions, with "early" segments (positions 1 and 2) having the highest relative participation at the top of the ranking. In general, the analysis of positions of high-scoring segments empirically validates the intuition that the most relevant content is often localized at the beginning of the target documents within the newswire CLEF corpora, which in turn reflects the writing style of the news domain.

Localized Relevance Matching: Sentences. The selection of the segmentation strategy can have a profound effect on the effectiveness of localized relevance matching. Instead of (overlapping 128-token) segments, one could, for example, measure the relevance of each document sentence for the query and (max-)pool the sentence relevance scores. Sentence-level segmentation and relevance pooling is particularly interesting when considering multilingual encoders that have been specialized precisely for sentence-level semantics (i.e., produce accurate sentence-level representations; see Sect. 3.3). In Table 5 we show the results of sentence-level localized relevance matching for all multilingual encoders. Unlike with segment-based localized relevance matching (see Table 4), here we see improvements for all multilingual encoders: what is more important, improvements over the baseline performance of the same encoders (see Table 2) are substantially larger than for segment-based localized relevance matching (e.g., 10 and 3.8 MAP-point improvements from sentence matching for LASER and LaBSE, respectively, compared to 2-point improvement for LASER and an 1-point MAP drop for LaBSE from segment matching). Sentence-level matching with the best-performing base multilingual encoder $\text{DISTIL}_{\text{DmBERT}}$ and pooling over two highest-ranking sentences (i.e., $k = 2$) yields the best unsupervised CLIR score that we observed overall (31.4 MAP points). For all encoders, averaging the scores of $k = 2$ or $k = 3$ highest-scoring sentences gives better results than considering only the single best sentence (i.e., $k = 1$)—this would indicate that the query-relevant content is still not overly localized within documents (i.e., not confined to a single sentence).

Finally, it is important to note that the gains in retrieval effectiveness (i.e., MAP gains) obtained with localized relevance matching (segment-level and sentence-level) come at the expense of reduced retrieval efficiency (i.e., increased retrieval time): the query representation now needs to be compared with each of the segment or sentence representations, instead of with only one aggregate representation for the whole document. The slowdown factor is proportional to the average number of segments/sentences per document in the document collection. Table 6 summarizes the approximate slowdown factors (i.e., average numbers of segments and sentences) for CLEF document collections in different languages.

4.5 Further analysis

We now 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.

Layer Selection All multilingual encoders have multiple layers and one may in principle choose to take (sub)word representations for CLIR at the output of any of them. Figure 4 shows the impact of taking subword representations after each layer for
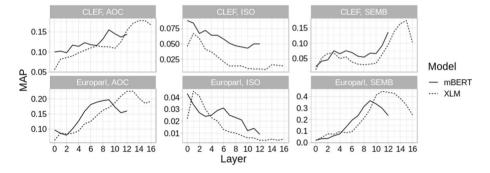
Table 5 Document-level CLIR results for *localized relevance matching* against document *sentences*

	k	EN-FI	EN-IT	EN-RU	EN-DE	DE-FI	DE-IT	DE-RU	FI-IT	FI-RU	AVG	Δ AVG
Proc-B	1	.219	.207	.136	.191	.235	.203	.138	.089	.126	.171	−5.16
	2	.216	.273	.158	.238	.267	.247	.176	.142	.122	.204	−1.90
	3	.229	.267	.165	.245	.284	.231	.168	.153	.120	.207	−1.61
	4	.231	.247	.173	.235	.286	.215	.166	.150	.120	.202	−2.07
DISTIL$_{\text{DmBERT}}$	1	**.381**	.288	.249	.332	.338	.248	.234	.234	.234	.282	+0.24
	2	.371	.313	**.303**	.399	.343	.285	.286	.246	**.280**	**.314**	+3.44
	3	.360	.308	.288	.407	**.359**	.274	**.288**	.247	.279	.312	+3.26
	4	.345	.298	.264	.382	.352	.262	.263	**.248**	.271	.298	+1.87
DISTIL$_{\text{XLM-R}}$	1	.323	.220	.144	.239	.316	.215	.148	.200	.149	.217	+4.00
	2	.339	.250	.199	.306	.305	.246	.200	.229	.196	.252	+7.51
	3	.328	.260	.205	.311	.318	.237	.209	.222	.208	.255	+7.81
	4	.311	.263	.188	.298	.319	.225	.178	.220	.179	.242	+6.52
DISTIL$_{\text{USE}}$	1	.131	.270	.181	.332	.121	.244	.200	.070	.054	.178	−2.01
	2	.139	.331	.226	.408	.134	.321	.240	.076	.132	.223	+2.50
	3	.131	.329	.220	**.433**	.129	.334	.235	.074	.129	.224	+2.56
	4	.134	.340	.212	.428	.122	.329	.225	.068	.124	.220	+2.21
LaBSE	1	.188	.182	.126	.167	.185	.147	.101	.112	.112	.147	+0.57
	2	.225	.197	.182	.213	.227	.180	.108	.138	.139	.179	+3.77
	3	.245	.186	.157	.234	.255	.163	.089	.136	.110	.175	+3.39
	4	.249	.192	.117	.235	.248	.139	.077	.145	.106	.167	+2.65
mUSE	1	.123	.270	.147	.317	.112	.256	.124	.070	.034	.161	−2.17
	2	.139	.368	.212	.395	.127	.334	.187	.079	.069	.212	+2.92
	3	.142	**.369**	.230	.428	.122	**.341**	.189	.083	.077	.220	+3.72
	4	.138	.357	.220	.429	.116	.331	.172	.081	.086	.214	+3.13

Table 5 (continued)

	k	EN-FI	EN-IT	EN-RU	EN-DE	DE-FI	DE-IT	DE-RU	FI-IT	FI-RU	AVG	Δ AVG
LASER	1	.207	.130	.096	.147	.206	.123	.107	.141	.112	.141	+7.30
	2	.175	.172	.127	.184	.206	.138	.133	.165	.129	.159	+9.07
	3	.191	.177	.153	.185	.197	.141	.154	.172	.136	.167	+9.94
	4	.175	.172	.133	.179	.184	.131	.125	.166	.123	.154	+8.60

Document relevance is the average of relevance scores of *k* highest-scoring sentences. Results (for 9 language pairs from CLEF) shown for the Proc-B baseline and all multilingual encoders specialized for encoding sentence-level semantics. **Bold**: The best performance for each language pair. Δ AVG denotes relative performance increases/decreases w.r.t. the respective base performances from Table 2

Table 6 Increase in computational complexity (i.e., decrease in retrieval efficiency) due to localized relevance matching via segments and sentences		#Documents	Segmentation		Sentence Splitting	
			#Segments	Factor	#Sentences	Factor
	DE	294,809	1,281,993	4.35	5,385,103	18.27
	IT	157,558	749,855	4.76	2,225,069	14.12
	FI	55,344	224,390	4.05	1,286,702	23.25
	RU	16,715	72,102	4.31	289,740	17.33

Fig. 4 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)

Fig. 5 CLIR performance of AOC variants (mBERT and XLM) w.r.t. the number of contexts used to obtain the term embeddings

self-supervised mBERT and XLM variants. We find that the optimal layer differs across the encoding strategies (AOC, ISO, and SEMB; cf. Sect. 3.2) 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 (Ethayarajh 2019; Rogers et al. 2020). 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 L11/L12 for XLM).

Table 7 Document-level unsupervised CLIR results w.r.t. the input text length

Length	SEMB$_{mBERT}$	SEMB$_{XLM}$	DIST$_{use}$	DIST$_{XLM-R}$	DIST$_{DmBERT}$	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

Scores averaged over all language pairs not involving Finnish

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 5 shows the AOC results depending on the number of contexts used to induce the term vectors (cf. τ in Sect. 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. This limitation was, in part, also the motivation for localized relevance matching approaches from the previous section. In our main experiments we truncated the documents to first 128 word pieces. Now we quantify (Table 7) 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 to semantically accurately encode a longer portion of text, which leads 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. These results, combined with gains obtained from localized relevance matching in the previous section render localized matching (i.e., document relevance pooled from segment- or sentence-level relevance scores) as a more promising strategy for retrieving long documents than attempts to increase the input length of multilingual transformers. Our findings from localized relevance matching seem to indicate that the relevance signal is highly localized: in such a setting, aggregating representations of very many tokens (i.e., across the whole document), e.g., with long-input transformers (Beltagy et al. 2020; Zaheer et al. 2020), is poised to produce semantically fuzzier (i.e., less precise) representations, from which it is harder to judge the document relevance for the query.

5 Supervised (re-)ranking

We next evaluate, on the same document-level collection from CLEF, the CLIR effectiveness of the multilingual encoders that have been exposed to some amount of supervision, i.e., fine-tuned using certain amount of relevance judgments, described in Sect. 3.4. We first discuss in Sect. 5.1 the performance of pointwise (re-)rankers based on mBERT

trained on large-scale out-of-domain collections; we then analyse (Sect. 5.2) how contrastive in-domain fine-tuning affects CLIR performance. In both cases, we exploit annotated English data for model fine-tuning: the transfer to other languages is directly enabled by the multilingual nature of the encoders.

5.1 Re-ranking with pointwise rankers

Transferring (re-)rankers across domains and/or languages is a promising method when in-language and in-domain fine-tuning data is scarce (MacAvaney et al. 2019). We experimented with two pointwise rankers, both based on mBERT, pretrained on English relevance data. The first model[11] was trained on the large-scale MS MARCO passage retrieval dataset (Nguyen et al. 2016), consisting of approx. 400M tuples, each consisting of a query, a relevant passage and a non-relevant passage. Transferring rankers trained on MS MARCO to various ad-hoc IR settings (i.e., domains) has been shown successful (Li et al. 2020; MacAvaney et al. 2020a; Craswell et al. 2021). Here, we investigate the performance of this supervised ranker trained on MS MARCO in simultaneous domain and language transfer. The second multilingual pointwise ranker (MacAvaney et al. 2020b) is trained on TREC 2004 Robust dataset (Voorhees 2005). Although TREC 2004 Robust is substantially smaller than MS MARCO (528K documents and 311K relevance judgments), by covering newswire documents it is domain-wise closer to our target CLEF test collection. As discussed in Sect. 3.4, pointwise neural rankers are typically used to re-rank the top of the ranking produced by some base ranker, rather than to rank the whole collection from scratch. Accordingly, we use the two above-described mBERT-based pointwise re-rankers to re-rank the top 100 documents from the initial rankings produced by each of the similarity-specialized multilingual encoders from Sect. 3.3).[12]
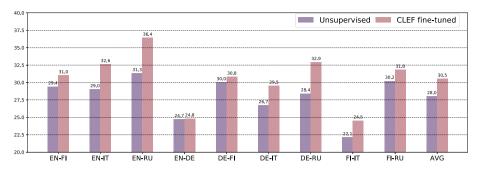
  Table 8 summarizes the results of our domain and language transfer experiments with the two pointwise mBERT-based re-rankers. For clarity, at the top of the table, we repeat the reference unsupervised CLIR performance of the similarity-specialized multilingual encoders (i.e., without any re-ranking) from Table 2. Intuitively, re-ranking—both with the MS MARCO-trained model and TREC-trained model—brings the largest gains for the weakest unsupervised rankers: mUSE, LaBSE, and LASER. The gains are somewhat larger when transferring the model trained on MS MARCO. However, re-ranking the results of the best-performing unsupervised ranker—DISTIL$_{DmBERT}$—brings no performance gains; in fact, re-ranking with the TREC-trained model reduces the quality of the base ranking by 7 MAP points. The transfer performance of the better-performing MS MARCO re-ranker in our CLIR benchmarks from CLEF depends on (1) the performance of the base ranker and (2) the target language pair. MS MARCO re-ranker improves the performance of our best-performing initial ranker, DISTIL$_{DmBERT}$, only for EN-DE and EN-IT, two language pairs in our evaluation for which the query language (EN) and collection language (DE, IT) are the closest to the source language of MS MARCO (EN) on which the re-ranker was trained; conversely, the MS MARCO re-ranking yields the largest performance drop for FI-RU, i.e., the pair of languages in our evaluation that are typologically most distant from

---

[11] https://huggingface.co/amberoad/bert-multilingual-passage-reranking-msmarco.

[12] We also experimented with re-ranking top 1000 documents, but the results were slightly worse for all base multilingual encoders than when re-ranking only the top 100 results.

Table 8 Document-level CLIR results on the CLEF collection obtained by language and domain transfer of supervised re-ranking models

	EN-FI	EN-IT	EN-RU	EN-DE	DE-FI	DE-IT	DE-RU	FI-IT	FI-RU	AVG	Δ
No re-ranking (reference)											
Proc-B	.258	.265	.166	.288	.294	.230	.155	.136	.216	.223	–
DISTIL$_{DmBERT}$	.294	.290	**.313**	.247	.300	.267	**.284**	**.221**	**.302**	**.280**	–
DISTIL$_{XLM-R}$	.219	.191	.149	.148	.215	.179	.142	.167	.125	.170	–
DISTIL$_{USE}$	.141	.346	.182	.258	.139	**.324**	.179	.104	.111	.198	–
mUSE	.077	.313	.186	.262	.077	.293	.183	.053	.092	.171	–
LaBSE	.191	.163	.136	.087	.172	.136	.103	.117	.140	.138	–
LASER	.146	.092	.060	.039	.153	.089	.062	.117	.076	.093	–
Re-ranker trained on MS MARCO											
Proc-B	.327	.303	.191	.321	.321	.230	.212	.160	.149	.246	+2.30
DISTIL$_{DmBERT}$	**.340**	.335	.219	.288	**.339**	.284	.245	.217	.160	.270	-1.02
DISTIL$_{XLM-R}$	.310	.252	.137	.232	.307	.219	.165	.183	.062	.207	+3.74
DISTIL$_{USE}$	.215	**.354**	.224	.295	.219	.310	.236	.133	.075	.229	+3.10
mUSE	.170	.348	.235	**.314**	.162	.301	.253	.110	.093	.220	+4.95
LaBSE	.300	.275	.169	.170	.306	.204	.138	.166	.109	.204	+6.60
LASER	.258	.166	.089	.092	.228	.151	.114	.127	.106	.148	+5.49
Re-ranker trained on TREC ROBUST											
Proc-B	.290	.292	.141	.310	.278	.214	.148	.108	.103	.209	-1.38
DISTIL$_{DmBERT}$	.284	.283	.153	.274	.252	.246	.130	.147	.119	.210	-6.98
DISTIL$_{XLM-R}$	.270	.227	.093	.242	.226	.200	.079	.129	.069	.170	+0.00
DISTIL$_{USE}$	.195	.321	.119	.309	.194	.287	.113	.113	.117	.196	-0.19
mUSE	.143	.330	.129	.313	.139	.261	.131	.086	.079	.179	+0.82
LaBSE	.275	.234	.086	.158	.245	.180	.076	.115	.077	.161	+2.25
LASER	.201	.164	.121	.095	.171	.137	.118	.111	.093	.135	+4.19

For each query, we re-rank the top 100 results produced by the base multilingual ranker with two mBERT-based L2R models trained on English data: MS MARCO Nguyen et al. (2016) (middle part of the table) and TREC ROBUST (bottom third of the table) Voorhees (2005); MacAvaney et al. (2020b). **Bold**: The best performance in each column (for each language-pair and the average)

Fig. 6 The effects of "in-domain" fine-tuning: comparison of CLIR performance with DISTIL$_{DmBERT}$ on the CLEF CLIR collections: (a) without any fine-tuning (i.e., an unsupervised CLIR approach; see Sect. 4.2) and (b) after in-domain fine-tuning on English CLEF data via contrastive metric-based learning (see Sect. 3.4): here we have only zero-shot language transfer, but no domain transfer (as was the case with L2R models from the previous section)

EN. These results suggest that, assuming a strong multilingual encoder as the base ranker, supervised re-ranking does not transfer well to distant language pairs. Overall, our results are in line with the most recent findings from Craswell et al. (2021), which also suggests that a ranker trained only on the large dataset like MS MARCO (i.e., without any fine-tuning on the target collection) yields mixed ad-hoc retrieval results.

5.2 Contrastive in-domain fine-tuning

We now empirically investigate the second common scenario in ad-hoc retrieval: a limited amount of "in-domain" relevance judgments that can be leveraged for fine-tuning of text encoders (as opposed to a large amount of "out-of-domain" training data sufficient to train full-blown learning-to-rank classifiers, covered in the previous subsection). To this end, we use the relevance judgments in the English portion of the CLEF collection to fine-tune our best-performing multilingual encoder (DISTIL$_{DmBERT}$), using the contrastive metric-based learning objective (see Sect. 3.4) to refine the representation space of the encoder. We carry out fine-tuning and evaluation in a 10-fold cross-validation setup (i.e., we carry out fine-tuning 10 different times, each time training on different nine-tenths of the queries and evaluating on the remaining one-tenth) in order to prevent any information leakage between languages: in the CLEF collection, queries in languages other than English are simply translations of the English queries. This resulted (in each fold) with a fine-tuning training set consisting of merely 800–900 positive instances (in English). We trained in batches of 16 positive instances and for each of them created all possible in-batch negatives[13] for the Multiple Negative Ranking Loss objective (see Sect. 3.4). With cross-validation in place, for each language pair, we obtain predictions for all queries without any information leakage, which makes the results of contrastive fine-tuning fully comparable with all previous results.

The CLIR results of the ranking with the contrastively fine-tuned DISTIL$_{DmBERT}$ are shown in Fig. 6. Unlike re-ranking with full-blown pointwise learning to rank models

---

[13] This means at most 15 in-batch negatives created from the other query-document pairs in the batch; there is less than 15 negatives only if there are other positive instances for the same query in the batch.

from the previous section, contrastive in-domain reshaping of the representation space of the multilingual encoder yields performance gains for all language pairs (2.5 MAP points on average). It is important to emphasize again that—because contrastive metric-based fine-tuning only updates the parameters of the original multilingual transformer (DISTIL$_{DmBERT}$) and introduces no additional parameters (i.e., no classification head on top of the encoder, as in the case of L2R models trained on MS MARCO and TREC ROBUST from the previous section)—we can, in exactly the same manner as with the base model before fine-tuning, fully rank the entire document collection for a given query, instead of restricting ourselves to re-ranking the top results of the base ranker.

Summarizing the results from this section and the previous one, it appears that—at least when it comes to zero-shot language transfer for cross-lingual document retrieval—specializing the representation space of a multilingual encoder with few(er) in-domain relevance judgments is more effective than employing a neural L2R ranker trained on large amounts of "out-of-domain" data.

5.3 Cross-lingual retrieval or cross-lingual transfer for monolingual retrieval?

At first glance, our negative CLIR results for the mBERT-based pointwise L2R rankers (Sect. 5.1)—i.e., the fact that using them for re-ranking does not improve the performance of our best-performing unsupervised ranker (DISTIL$_{DmBERT}$)—seem at odds with their solid cross-lingual transfer results reported in previous work MacAvaney et al. (2020b). It is, however, important to notice the fundamental difference between two evaluation settings: what was previously evaluated (MacAvaney et al. 2020b) was the effectiveness of (zero-shot) *cross-lingual transfer* of a *monolingual retrieval* model, trained on English data and transferred to a set of target languages. In other words, both in training and at inference time the models deal with queries and documents written in the same language. Our work here, instead, focuses on a fundamentally different scenario of cross-lingual retrieval, where the language of the query is different from the language of document collection. We argue that, in a supervised setting, in which one trains on monolingual English data only, the latter (i.e., CLIR) represents a more difficult transfer setup.

To validate the above assumption, we additionally evaluate the two mBERT-based re-rankers from Sect. 5.1 trained on MS MARCO and TREC ROBUST, respectively, on monolingual portions of the CLEF collection. We use them to re-rank two strong monolingual baselines: (1) Query Likelihood Model (QLM, based on unigrams) Ponte and Croft (1998) with Dirichlet smoothing Zhai and Lafferty (2004), which we also used for the machine-translation baseline (MT-IR) in our base evaluation (see Sect. 4.1); and (2) a retrieval model based on aggregation of IDF-scaled static word embeddings (Sect. 3.1; Eq. (1)).[14] For the latter, we used the monolingual FastText embeddings trained on Wikipedias of respective languages,[15] with vocabularies limited to the 200K most frequent terms.

The results of mBERT-based re-rankers in cross-lingual transfer for monolingual retrieval are summarized in Table 9. We see that, unlike in CLIR (see Table 8), mBERT-based re-rankers do substantially improve the performance of the base retrieval models, even despite the fact that the base performance of the monolingual baselines (QLM and

---

[14] This corresponds to the Proc-B baseline in CLIR evaluations; only here we use monolingual embeddings of the target language (instead of a bilingual word embedding space, as in CLIR).

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

Table 9 Cross-lingual zero-shot transfer for monolingual retrieval: results on the monolingual CLEF portions.

	EN-EN	FI-FI	DE-DE	IT-IT	RU-RU	AVG	Δ AVG
No re-ranking (reference)							
QLM	.471	.376	.400	.463	.325	.407	–
FastText	.310	.327	.314	.314	.214	.296	–
Re-ranker trained on MS MARCO							
QLM	.520	.469	.424	.488	.359	.452	+4.53
FastText	.434	.430	.384	.468	.359	.415	+11.90
Re-ranker trained on TREC ROBUST							
QLM	.481	.520	.420	.454	.303	.436	+1.98
FastText	.375	.462	.367	.429	.299	.386	+8.76

Base rankers (top third of the table)—QLM with Dirichlet Smoothing and aggregation of static monolingual word embeddings (FastText) and re-ranking with pointwise mBERT-based models trained on English MS MARCO (middle third) and TREC ROBUST data (bottom third), respectively

FastText) is significantly above the best CLIR performance we observed with unsupervised rankers (see DISTIL$_{DmBERT}$ in Table 2). This is in line with the findings from MacAvaney et al. (2020b): multilingual encoders (e.g., mBERT) do seem to be a viable solution for (zero-shot) cross-lingual transfer of learning-to-rank models for monolingual retrieval. But why are they not as effective when transferred to CLIR settings (as shown in 5.1)? We hypothesize that monolingual English training on large-scale datasets like MS MARCO or TREC ROBUST leads to a sort of "overfitting" to monolingual retrieval (e.g., the model may implicitly learn to assign a lot of importance to exact term matches)—such (latent) features will, in principle, transfer reasonably well to other monolingual retrieval settings, regardless of the target language; with queries in different language from documents, however, CLIR instances are likely to generate out-of-training-distribution values for these latent features (e.g., if the model learned to value exact matches during training, at predict time in CLIR settings, it would need to recognize word-level translations between the two languages), confusing the pointwise classifier.

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 diverse languages.

We first profiled the popular self-supervised multilingual encoders (mBERT and XLM) as well as the multilingual encoders specialized for semantic text matching on semantic similarity datasets and parallel data as text encoders for unsupervised CLIR. Our empirical results show that self-supervised multilingual encoders (mBERT and XLM), without exposure to task supervision, generally fail to outperform CLIR models based on static cross-lingual word embeddings (CLWEs). Semantically-specialized multilingual sentence

encoders, on the other hand, do outperform CLWEs; the gains, however, are pronounced only in sentence retrieval, while being much more modest in document retrieval.

Acknowledging that sentence-specialized multilingual encoders are not designed for encoding long documents, we proposed to exploit their strength—precise semantic encoding of short texts—in document retrieval too, by means of localized relevance matching, where we compare the query with individual document segments or sentences and max-pool the relevance scores; we showed that such localized relevance matching with sentence-specialized multilingual encoders yields substantial document-level CLIR gains.

Finally, we investigated how successful supervised (re-)rankers based on multilingual encoders are in ad-hoc CLIR evaluation settings. We show that, while rankers trained monolingually on large-scale English datasets (e.g., MS-MARCO) can be successfully transferred to monolingual retrieval tasks in other languages by means of multilingual encoders, their transfer to CLIR setups, in which the query language differs from the language of the document collection, is much less successful. Furthermore, we introduced an alternative supervised approach, based on contrastive metric-based learning, designed for fine-tuning the representation space of a multilingual encoder when only a limited amount of "in-domain" relevance judgments is available. We show that such small-scale in-domain fine-tuning of multilingual encoders yields better CLIR performance than rankers trained on large external collections (i.e., out-of-domain).

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 believe that our systematic comparative evaluation of a multitude of multilingual encoders (as both unsupervised and supervised rankers) offers a multitude of insights for practitioners dealing with (ad-hoc) cross-lingual retrieval task. While there are scenarios in which multilingual encoders can substantially improve CLIR performance, our work identifies potential pitfalls and emphasizes conditions needed for solid CLIR performance with multilingual text encoders. We make our code and resources available at https://github.com/rlitschk/EncoderCLIR.

References

Akkalyoncu Yilmaz, Z., Yang, W., Zhang, H., & Lin, J. (2019). Cross-domain modeling of sentence-level evidence for document retrieval. In Proceedings of EMNLP, pp. 3490–3496.

Artetxe, M., Labaka, G., & Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL*, pp. 789–798.

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the ACL pp. 597–610.

Beltagy, I., Peters, M.E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150

Braschler, M. (2003). CLEF 2003–Overview of results. In *Workshop of the cross-language evaluation forum for european languages*, pp. 44–63.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In Proceedings of NeurIPS, pp. 1877–1901.

Cao, S., Kitaev, N., & Klein, D. (2020). Multilingual alignment of contextual word representations. In Proceedings of ICLR.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Proceedings of SemEval, pp. 1–14.

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. (2018). Universal sentence encoder for English. In Proceedings of EMNLP, pp. 169–174.

Chidambaram, M., Yang, Y., Cer, D., Yuan, S., Sung, Y., Strope, B., & Kurzweil, R. (2019). Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of ACL: Workshop on representation learning for NLP*, pp. 250–259.

Clark, K., Luong, M., Le, Q.V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In Proceedings of ICLR.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of ACL, pp. 8440–8451.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of ACL, pp. 8440–8451.

Conneau, A., & Kiela, D. (2018). SentEval: An evaluation toolkit for universal sentence representations. In Proceedings of LREC, pp. 1699–1704.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In Proceedings of EMNLP, pp. 670–680.

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In Proceedings of NeurIPS, pp. 7059–7069.

Craswell, N., Mitra, B., Yilmaz, E., Campos, D., & Lin, J. (2021). Ms marco: Benchmarking ranking models in the large-data regime. In Proceedings of SIGIR, pp. 1566–1576.

Dai, Z., & Callan, J. (2019). Deeper text understanding for IR with contextual neural language modeling. In Proceedings of SIGIR, pp. 985–988

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., & Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. In Proceedings of ACL, pp. 2978–2988.

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL, pp. 4171–4186.

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Proceedings of EMNLP, pp. 55–65.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic BERT sentence embedding. arXiv preprint arXiv:2007.01852

Gao, L., Dai, Z., & Callan, J. (2020). Modularized transfomer-based ranking framework. In Proceedings of EMNLP, pp. 4180–4190.

Glavaš, G., Litschko, R., Ruder, S., & Vulić, I. (2019). How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In Proceedings of ACL, pp. 710–721.

Guo, M., Shen, Q., Yang, Y., Ge, H., Cer, D., Hernandez Abrego, G., Stevens, K., Constant, N., Sung, Y.H., Strope, B., & Kurzweil, R. (2018). Effective parallel corpus mining using bilingual sentence embeddings. In Proceedings of WMT, pp. 165–176.

Hoogeveen, D., Verspoor, K.M., & Baldwin, T. (2015). CQADupStack: A benchmark data set for community question-answering research. In Proceedings of ADCS, pp. 3:1–3:8.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Proceedings of ICML, pp. 4411–4421.

Humeau, S., Shuster, K., Lachaux, M.A., & Weston, J. (2020). Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In Proceedings of ICLR.

Jiang, Z., El-Jaroudi, A., Hartmann, W., Karakos, D., & Zhao, L. (2020). Cross-lingual information retrieval with BERT. In Proceedings of LREC, p. 26.

Johnson, J., Douze, M., & Jégou, H. (2017). Billion-scale similarity search with gpus. arXiv preprint arXiv: 1702.08734

Karthikeyan, K., Wang, Z., Mayhew, S., & Roth, D. (2020). Cross-lingual ability of multilingual BERT: An empirical study. In Proceedings of ICLR.

Khattab, O., & Zaharia, M. (2020). Colbert: Efficient and effective passage search via contextualized late interaction over bert. In Proceedings of SIGIR, pp. 39–48.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th machine translation summit (MT SUMMIT)*, pp. 79–86.

Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020). From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In Proceedings of EMNLP, pp. 4483–4499.

Lei, T., Joshi, H., Barzilay, R., Jaakkola, T., Tymoshenko, K., Moschitti, A., & Màrquez, L. (2016). Semi-supervised question retrieval with gated convolutions. In Proceedings of NAACL, pp. 1279–1289.

Li, C., Yates, A., MacAvaney, S., He, B., & Sun, Y. (2020). Parade: Passage representation aggregation for document reranking. arXiv preprint arXiv:2008.09093

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., Fan, X., Zhang, R., Agrawal, R., Cui, E., Wei, S., Bharti, T., Qiao, Y., Chen, J.H., Wu, W., Liu, S., Yang, F., Campos, D., Majumder, R., & Zhou, M. (2020). XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In Proceedings of EMNLP, pp. 6008–6018.

Lin, J., Nogueira, R., & Yates, A. (2021). *Pretrained transformers for text ranking: BERT and beyond*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Litschko, R., Glavaš, G., Vulić, I., & Dietz, L. (2019). Evaluating resource-lean cross-lingual embedding models in unsupervised retrieval. In Proceedings of SIGIR, pp. 1109–1112.

Litschko, R., Vulić, I., Ponzetto, S.P., & Glavaš, G. (2021). Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In Proceedings of ECIR, pp. 342–358.

Liu, F., Vulić, I., Korhonen, A., & Collier, N. (2021). Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In Proceedings of EMNLP, pp. 1442–1459.

Liu, Q., Kusner, M.J., & Blunsom, P. (2020). A survey on contextual embeddings. arXiv preprint arXiv: 2003.07278

Liu, Q., McCarthy, D., Vulić, I., & Korhonen, A. (2019). Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation. In Proceedings of CoNLL, pp. 33–43.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692

MacAvaney, S., Cohan, A., & Goharian, N. (2020). SLEDGE-Z: A zero-shot baseline for COVID-19 literature search. In Proceedings of EMNLP, pp. 4171–4179.

MacAvaney, S., Soldaini, L., & Goharian, N. (2020). Teaching a new dog old tricks: Resurrecting multilingual retrieval using zero-shot learning. In Proceedings of ECIR, pp. 246–254.

MacAvaney, S., Yates, A., Cohan, A., & Goharian, N. (2019). Cedr: Contextualized embeddings for document ranking. In Proceedings of SIGIR, pp. 1101–1104.

Mikolov, T., Le, Q.V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. CoRR, abs/1309.4168

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng, L. (2016). MS MARCO: A human generated machine reading comprehension dataset. *Workshop on cognitive computing at NIPS*.

Nogueira, R., Yang, W., Cho, K., & Lin, J. (2019). Multi-stage document ranking with BERT. arXiv preprint arXiv:1910.14424

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? In Proceedings of ACL, pp. 4996–5001.

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In Proceedings of SIGIR, pp. 275–281.

Ponti, E.M., Glavaš, G., Majewska, O., Liu, Q., Vulić, & I., Korhonen, A. (2020). XCOPA: A multilingual dataset for causal commonsense reasoning. In Proceedings of EMNLP, pp. 2362–2376.

Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W.X., Dong, D., Wu, H., & Wang, H. (2021). RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In Proceedings of NAACL, pp. 5835–5847.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In Proceedings of EMNLP, pp. 3973–3983.

Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In Proceedings of EMNLP, pp. 4512–4525.

Robertson, S.E., & Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In Proceedings of SIGIR, pp. 232–241.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. Transactions of the ACL pp. 842–866.

Ruder, S., Vulić, & I., Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research,* pp. 569–631.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108

Schönemann, P. H. (1966). A generalized solution of the orthogonal Procrustes problem. *Psychometrika,* pp. 1–10.

Shi, P., Bai, H., & Lin, J. (2020). Cross-lingual training of neural models for document ranking. In Proceedings of EMNLP (Findings), pp. 2768–2773.

Shi, P., Zhang, R., Bai, H., & Lin, J. (2021). Cross-lingual training with dense retrieval for document retrieval. CoRR, abs/2109.01628

Smith, S. L., Turban, D. H., Hamblin, S., & Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In Proceedings of ICLR.

Thakur, N., Reimers, N., Daxenberger, J., & Gurevych, I. (2021). Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In Proceedings of NAACL, pp. 296–310.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Proceedings of NeurIPS, pp. 5998–6008.

Voorhees, E. (2005). Overview of the trec 2004 robust retrieval track. https://doi.org/10.6028/NIST.SP.500-261

Vulić, I., Glavaš, G., Reichart, R., & Korhonen, A. (2019). Do we really need fully unsupervised cross-lingual embeddings? In Proceedings of EMNLP, pp. 4406–4417.

Vulić, I., & Moens, M. F. (2015). Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In Proceedings of SIGIR, pp. 363–372.

Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., & Korhonen, A. (2020). Probing pretrained language models for lexical semantics. In Proceedings of EMNLP, pp. 7222–7240.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of ICLR.

Williams, A., Nangia, N., & Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of NAACL, pp. 1112–1122 (2018)

Wu, S., & Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In Proceedings of EMNLP, pp. 833–844.

Yang, Y., Abrego, G.H., Yuan, S., Guo, M., Shen, Q., Cer, D., Sung, Y.H., Strope, B., & Kurzweil, R. (2019). Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In Proceedings of AAAI, pp. 5370–5378.

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. (2020). Multilingual universal sentence encoder for semantic retrieval. In Proceedings of ACL: System demonstrations, pp. 87–94.

Yang, Y., Hernandez Abrego, G., Yuan, S., Guo, M., Shen, Q., Cer, D., Sung, Y. H., Strope, B., & Kurzweil, R. (2019). Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In Proceedings of IJCAI, pp. 5370–5378.

Yu, P., & Allan, J. (2020). A study of neural matching models for cross-lingual IR. In Proceedings of SIGIR, pp. 1637–1640.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big bird: Transformers for longer sequences. In Proceedings of NeurIPS, pp. 17283–17297.

Zhai, C., & Lafferty, J. A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems pp. 179–214 (2004)

Zhang, X., Ma, X., Shi, P., & Lin, J. (2021). Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st workshop on multilingual representation learning*, pp. 127–137.

Zhao, M., Zhu, Y., Shareghi, E., Vulić, I., Reichart, R., Korhonen, A., & Schütze, H. (2021). A closer look at few-shot crosslingual transfer: The choice of shots matters. In Proceedings of ACL, pp. 5751–5767.

Zhao, W., Eger, S., Bjerva, J., & Augenstein, I. (2021). Inducing language-agnostic multilingual representations. In Proceedings of *SEM, pp. 229–240.

Zhao, W., Glavaš, G., Peyrard, M., Gao, Y., West, R., & Eger, S. (2020). On the limitations of crosslingual encoders as exposed by reference-free machine translation evaluation. In Proceedings of ACL, pp. 1656–1671.

Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (2016). The United Nations parallel corpus v1.0. In Proceedings of LREC, pp. 3530–3534.

Zweigenbaum, P., Sharoff, S., & Rapp, R. (2018). Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of 11th workshop on building and using comparable corpora*, pp. 39–42.

Authors and Affiliations

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

Robert Litschko
litschko@informatik.uni-mannheim.de

Ivan Vulić
iv250@cam.ac.uk

Simone Paolo Ponzetto
simone@informatik.uni-mannheim.de

[1]   University of Mannheim, Mannheim, Baden-Wuerttemberg, Germany

[2]   Language Technology Lab, University of Cambridge, Cambridge, UK