Learning a Sparse Representation Model for Neural CLIR

Suraj Nair[1,2], Eugene Yang[2], Dawn Lawrie[2], James Mayfield[2] and Douglas W. Oard[1,2]

[1]University of Maryland, College Park MD 20742, USA
[2]HLTCOE. Johns Hopkins University, Baltimore MD 21211, USA

Abstract

In monolingual retrieval, sparse representations learned atop BERT-style models offer a complementary approach to the unsupervised BM25 model. Inspired by this line of work, we explore adapting such models to the Cross-Language Information Retrieval (CLIR) setting, in which queries and documents are in different languages. The lack of lexical match between queries and documents inhibits a naive replication of these monolingual models for CLIR. We propose SPLADE-X, a cross-language expansion model for CLIR that performs complementary to a strong PSQ baseline. We further identify the challenges in developing such models, make connections to existing CLIR models and highlight future directions that pave the way for learning sparse representations feasible for CLIR.

Keywords

Sparse Representation, Neural CLIR, Multilingual Language Models,

1. Introduction

Learning sparse representation models using pretrained language models (e.g., BERT [1]) has risen in prominence in monolingual information retrieval applications, particularly those that access English content. The main idea is to represent queries and documents in a high-dimensional space spanning BERT's vocabulary, where only a few dimensions (which correspond to vocabulary terms) are non-zero. The non-zero document term weights can then be stored in a standard inverted index. This allows us to exploit the efficiency of traditional sparse vector space "bag of words" retrieval approaches. This framework also enables query or document expansion by generating weights for terms that do not, but plausibly could have, appeared in either the queries or the documents. This partially mitigates the vocabulary mismatch faced by bag of words models such as BM25 [2, 3]. In this paper, our goal is to build a sparse representation model for Cross-Language Information Retrieval (CLIR), in which the queries and documents are expressed in different languages.

With the availability of large scale training collections such as MS MARCO [4] that have been translated into multiple languages [5, 6] and an increasing variety of sparse representation models for monolingual retrieval [7, 8, 9, 10, 11, 12, 13, 14], a natural question is whether

DESIRES 2022 – 3rd International Conference on Design of Experimental Search & Information REtrieval Systems, 30-31 August 2022, San Jose, CA, USA

srnair@cs.umd.edu (S. Nair); eugene.yang@jhu.edu (E. Yang); lawrie@jhu.edu (D. Lawrie); mayfield@jhu.edu (J. Mayfield); oard@umd.edu (D. W. Oard)

extending these ideas to CLIR involves anything more than simply replacing a monolingual pretrained model (e.g, BERT) with a multilingual model (e.g., mBERT [1] or XLM-R [15])? To answer this question, we need to look at how the existing models generate sparse term weights for queries or documents. We can group the existing models into two categories: a) exact-match, where weights are changed for terms that occur in queries or documents but no nonzero weights are added for any additional terms; or b) lexical expansion, in which the number of terms with nonzero weights is still limited in some way, but some terms that did not appear in the original query or the original document can be given non-zero weights. In the case of CLIR, the exact-match approach will not work (or at least it will not work very well!) because the queries and documents are expressed in different languages, generally using different words. Thus our natural point of comparison should be lexical expansion.

For monolingual lexical expansion, existing approaches rely on either applying a document expansion model (e.g., doc2query [16, 17] or TILDE [11]), or utilizing a pretrained language model head to expand a low-dimensional dense representation back out to a sparse representation that spans the full vocabulary space (e.g., SPLADE [13]). Building a document expansion model that generalizes well beyond English is already a challenging problem [18, 19], one that becomes even more challenging given the explosive growth in the vocabulary size of multilingual pretrained models such as mBERT (110k) and XLM-R (250k). These are 3 to 7 times the size of the monolingual BERT vocabulary (35k). It is these two factors, the need to generalize across languages and the potential benefits of limiting the vocabulary size, that distinguish CLIR applications of lexical expansion methods from their monolingual cousins.

We propose SPLADE-X, a cross-language generalization of the lexical expansion framework of SPLADE [13], built on top of mBERT. To manage the large vocabulary of mBERT, we use a vocabulary reduction technique in which we choose dimensions corresponding to vocabulary terms (i.e., subwords) belonging to only the query language (which in our case is English). This choice forces the model to learn cross-language lexical expansions for the non-English documents, which roughly corresponds to an encoder-only translation task. To train SPLADE-X, we explore several cross-lingual transfer learning strategies, including zero-shot [20], translate-train [21] and bilingual training. In the zero-shot setup, we train the model on pairs consisting of English queries and English passages from MS MARCO, and then we simply use that trained system with our test collections that contain English queries and documents in some language other than English. In the translate-train setup, we train the model on pairs consisting of English queries from MS MARCO and translated MS MARCO passages, where we use machine translation to produce those translated passages.

For bilingual training, we train on triples of English queries and English passages from MS MARCO, together with a translation of the MS MARCO passage. Inspired by Reimers and Gurevych [22], we propose a bilingual alignment loss that encourages the sparse representation of the English passage and its translation to be similar. Furthermore, we extend the monolingual distillation loss proposed by Yang et al. [23] to the cross-language setting. Specifically, we distill the similarity matrix produced by the monolingual SPLADE (teacher) model to the multilingual SPLADE-X (student) model.

2. Background and Related Work

In this section, we describe the preliminaries involved in sparse representation learning, and we describe existing CLIR models that use different techniques to generate similar representations.

2.1. Sparse representation learning

Using the notation defined in Lin [24], given a query q and document d, we can generate fixed-length vectors $\eta_q(q)$ and $\eta_d(d)$ for the query and document respectively. Here, the η_q and η_d are the query and document encoder initialized using a pretrained language model.[1] Using the generated vector, we can compute a relevance score for query q and document d using a custom scoring function ϕ as follows

$$s(q, d) = \phi(\eta_q(q), \eta_d(d)) \tag{1}$$

The most common approach to train these models is to define a ranking loss such that the relevance score of a query q and a relevant document d^+ is higher than the relevance score of the same query q and a non-relevant document d^- (i.e., $s(q, d^+) > s(q, d^-)$). Specifically, the ranking loss is defined as follows:

$$\mathscr{L}_{\text{rank}} = -\log \frac{e^{s(q,d^+)}}{e^{s(q,d^+)} + \sum_{d^-} e^{s(q,d^-)}} \tag{2}$$

There are several ways of sampling non-relevant ("negative") documents, including in-batch negative samples [25] or using a large queue of negatives [26, 27]. In the case of a sparse retrieval model, the fixed-length of the query and document representations corresponds to the vocabulary size of the pretrained model. However, training with ranking loss does not ensure the resulting representations are sparse. Existing work enforces sparsity in these representations either by choosing specific terms that can receive non-zero weights (e.g., the terms with the highest weights in the query or document representations [23]) or by optimizing for some form of regularization loss (such as L1 regularization [7] or that of the FLOPS optimizer [28] used in SPLADE [13]). Among the existing sparse models, SPLADE has the best performance on the monolingual BEIR benchmark [29]. Hence, we have chosen to generalize SPLADE's basic approach for application to CLIR tasks in this paper.

2.2. Similarly Structured CLIR models

Our application of SPLADE to CLIR takes a sequence of terms from a non-English document as input, and it outputs a corresponding set of (unordered) term weights for (only) English terms. This is essentially the same behavior that we would expect from a statistical machine translation system that lacks a language model for the generated English, or a neural translation model that decodes isolated terms. The first approach, modeled on statistical machine translation, has been called Probabilistic Structured Queries (PSQ) [30]. PSQ maps term frequency vectors from the document language to the query language using a matrix of translation probabilities (normalized to sum to one in the document language to query language direction) as a simple matrix-vector

[1] In this paper, we set the query and document encoder to be the same.

product. This results in a sparse document representation, which contains nonzero term weights only for plausible translations of terms that appear in the document. Any traditional term weighting function that can accept partial term counts (e.g., BM25 or query likelihood) can then be computed on the resulting term frequency vector.

One limitation of PSQ is that it pays no attention to the terms that precede or follow the term being translated; the translation probabilities for a term are constant regardless of context. But the same result, a weighted mapping of each term to its plausible translations, can also be achieved in ways that leverage context. Perhaps the most straightforward such approach is simply to use a weighted n-best decoder in a neural MT system to generate weighted alternatives for each translated term. Neural translation systems are data hungry at training time, however, and translation sequences that are rare in the training data may not be well modeled. Rare terms are particularly useful in information retrieval tasks (because of their specificity), so techniques that leverage context on the source (document) side, but not the target (English) side have also been explored. Two such examples are Searcher [31], which leverages pretrained language models, and the Neural Network Lexical Translation Model (NNLTM) [32], built using a character-level Convolutional Neural Network (CNN) model.

The key difference between our approach and those summarized in this section is the training objective. In the CLIR techniques we have described, the first step leverages parallel text to generate translation probabilities; those probabilities are then used in a modular way with some (traditional or neural) ranked retrieval method. Our approach also learns from parallel text, but with two key differences: a) the parallel texts in our case are the English and non-English version of the documents for which we have training judgments; and b) we jointly optimize translation and retrieval by balancing multiple loss functions. Because these techniques generate conformal representations, we can experiment with either early fusion, in which we combine alternative ways of estimating term weights in the query language [33], or late fusion, in which we combine ranked retrieval results. In this paper we experiment with Reciprocal Rank Fusion (RRF), a late fusion technique [34], finding that our approach yields results complementary to those obtained using PSQ.

3. Sparse Model for CLIR

Here we describe the SPLADE framework and introduce the training procedure for our CLIR extension, which we call SPLADE-X.

3.1. SPLADE → SPLADE-X

SPLADE is a sparse lexical expansion model that generates $|V|$-dimensional vectors for queries and documents, where the weights represents the importance of the underlying terms.

To generalize the model to a cross-language setting, let $V_{\mathcal{S}}$ be the vocabulary space of the input text and $V_{\mathcal{T}}$ for the output, where $V = V_{\mathcal{S}} = V_{\mathcal{T}}$ in the original monolingual SPLADE model. Given an input (either query or document) text sequence $t \in V_{\mathcal{S}}^N$ of length N, SPLADE uses a BERT Masked Language Model (MLM) head to get the term weights for every query subword. Specifically, for a query subword $t_i \in V_{\mathcal{S}}$, the model generates the term weights w_{ij} for

candidate output subword $t_j \in V_{\mathcal{T}}$ as

$$w_{ij} = \psi(h_i)^T e_j + b_j \tag{3}$$

where the ψ is a composition of linear layer with GeLU activation and LayerNorm applied to the contextual BERT embedding h_i of t_i. Here e_j denotes the j-th row of the BERT MLM decoder learnable matrix and b_j stands for the token-level bias.

To produce an aggregate score for each candidate output token $t_j \in V_{\mathcal{T}}$, SPLADEv2 [14] proposes a max pooling over the input vocabulary dimension as follows:

$$w_j = \max_i \log\left(1 + \text{ReLU}(w_{ij})\right) \tag{4}$$

While the original implementation of SPLADE used the FLOPS optimizer to sparsify the representation, more recent work has proposed simply selecting the top-k dimensions from the final $|V_{\mathcal{T}}|$-sized vectors [23]. Owing to the simplicity of the approach, we use that top-k masking approach in this paper.

Generalizing SPLADE to the CLIR setting, we create SPLADE-X by replacing the BERT encoder with multilingual BERT (mBERT). Since the size of the mBERT vocabulary (110k) is roughly 3x that of BERT (35k), we create a vocabulary mask that filters out any subword that does not belong to the query language (in our case, English). This forms an output vocabulary space $V_{\mathcal{T}}$ containing only English subwords. To do so, we tokenize the MS MARCO corpus (in English) using the mBERT tokenizer and select only those subwords that contain alphanumeric or punctuation characters.[2] This gives us a list of 33k unique subwords that we use for SPLADE-X modeling. In addition to constraining vocabulary size, this also forces the model to learn cross-language expansions for non-English documents.

3.2. Bilingual Training

To train SPLADE-X, we propose a bilingual training recipe that utilizes the monolingual MS MARCO [4] queries and passages along with translations of the MS MARCO passages [5] to document language that we plan to search. Here, we refer to English as the source language $\mathcal{S}$ and non-English as the target language $\mathcal{T}$, since that is the direction in which MS MARCO was translated. Our training loss is composed of three components, source-only loss ($L_{\mathcal{S}}$), target-only loss ($L_{\mathcal{T}}$) and source-target loss ($L_{\mathcal{ST}}$).

For source-only loss ($L_{\mathcal{S}}$), we use the ranking loss in Eqn. 2, using the English queries and English passages, and an additional distillation loss using the monolingual SPLADE model, as proposed in Yang et al. [23]. The distillation loss enforces knowledge transfer from the monolingual teacher to the multilingual student model. We compute the target-only loss ($L_{\mathcal{T}}$) in a similar way as the source-only loss, except we replace the English passages with their translated versions. We further introduce an alignment loss between source and target ($L_{\mathcal{ST}}$) that brings the representations of English passages and their translated version closer, using a MSE loss.

[2] If we had wanted to experiment with using non-English queries, we would have instead used the translated MS MARCO corpora.

Our final training loss is summarized as follows,

$$L = L_{\mathcal{S}} + L_{\mathcal{T}} + L_{\mathcal{ST}} \tag{5}$$

We also experiment with a zero-shot variant with only the source loss, and a translate-train variant with only the target loss.

4. Experiments

Here we describe test collections, training, evaluation and results.

Test collection

For evaluation we use CLIR test collections from the CLEF 2003 multilingual ad-hoc retrieval task, with queries in English and news documents in either German, French, Italian or Spanish. [35].[3] Each collection includes 60 English topics, for which we use the title field as the query.

Training & Inference setup

For training SPLADE-X, we use the Tevatron toolkit[4] that supports the HuggingFace Transformers [36] library. We initialize the encoder using mBERT and train three variants of SPLADE-X: zero-shot (ZS), translate-train (TT), and our proposed bilingual training (BI). We train the ZS and TT variants using the Adam optimizer with a learning rate of 1e-5 and a batch size of 32 using 4 V100 GPUs for 100k steps. For our new BI variant, we train on 8 V100 GPUs for 60k steps, keeping the rest of the parameters same.

We fix the query and the passage length to be 32 and 128 respectively and choose k to be 1% of the total mBERT vocabulary size. For distillation, we use the publicly available DistilSPLADE-max checkpoint,[5] initialized using coCondenser-medium [37].

For inference, we split the CLEF documents into overlapping passages of length 128 with a stride of 42 tokens. We index the SPLADE-X passage term weights using Anserini [38]. To rank documents, we first run a passage-retrieval task using the SPLADE-X queries and the indexed passages using Anserini. The output is then passed through a score aggregation function that selects the maximum passage score as the corresponding document score.

Baseline

As a traditional alternative to our SPLADE-X variants, we also report a PSQ baseline. Specifically, we implement a PSQ-based Query Likelihood model [39] to estimate the relevance of a document in some other language given a query in English To obtain the translation probabilities to be used in PSQ, we rely on the word alignment output from the GIZA++ [40] aligner. We train GIZA++ using a combination of parallel sentences from Europarl [41] and the Panlex [42] dictionaries. For each language pair, we have 2.5 to 3 million sentence pairs for training. Translation probabilities that are less than 1e-5 are filtered out.

[3] We omit evaluation on translated MS MARCO collections due to their synthetic nature.

[4] https://github.com/texttron/tevatron

[5] http://download-de.europe.naverlabs.com/Splade_Release_Jan22/splade_distil_CoCodenser_medium.tar.gz

Table 1
CLEF MAP. Bold is best CLIR system, individually or with fusion. * is significant improvement over PSQ.

	Language			
System	German	French	Italian	Spanish
Mono. BM25	0.296	0.406	0.387	0.431
QMT BM25	0.260	0.370	0.306	0.400
PSQ	**0.322**	0.363	0.307	**0.347**
SPLADE-X (ZS)	0.213	0.277	0.210	0.253
SPLADE-X (TT)	0.300	0.383	**0.318**	0.310
SPLADE-X (BI)	0.317	**0.393**	0.314	0.331
PSQ+SPLADE-X (ZS)	0.339	0.367	0.296	0.351
PSQ+SPLADE-X (TT)	0.398*	0.451*	**0.376***	0.386*
PSQ+SPLADE-X (BI)	**0.405***	**0.458***	0.367*	**0.392***

Evaluation

To evaluate our CLIR models and baseline we compute Mean Average Precision (MAP) using trec_eval.[6] For system combination, we perform Reciprocal Rank Fusion using TrecTools [43]. Differences in means are tested for significance using a two-tailed paired *t*-test ($p<0.05$) with Bonferroni-Holm correction.

Results

Table 1 shows results for different SPLADE-X variants and the PSQ baseline which performs on par with BM25 on human-translated queries (Mono. BM25) and machine-translated queries (QMT BM25) using a Marian MT model.[7] Contrary to findings in monolingual retrieval (where comparisons are to a traditional monolingual retrieval baseline), we observe that none of the SPLADE-X variants consistently outperform the PSQ baseline. Comparing among the variants, ZS performs the worst, as expected, illustrating the challenges of relying on mBERT alone to CLIRize a monolingual IR method. This aligns with the finding that tasks involving cross-language input sequences are harder to generalize [44]. BI generally performs numerically better than the TT variant, except in Italian, but none of those apparent differences are statistically significant.

When using Reciprocal Rank Fusion to combine each variant with PSQ baseline, we observe statistically significant improvements over PSQ alone for both TT and BI. Looking at the pattern of results, we consistently see smaller improvements from Reciprocal Rank Fusion with PSQ and BI (28% from 0.317 to 0.405 in German) than fusing PSQ and ZS (59% from 0.213 to 0.339 in German), suggesting that BI may be learning some lexical translations from its more powerful training scheme, thus receiving less advantage from PSQ when fusing. In the future, we plan to explore early fusion between of PSQ and SPLADE-X (BI) so as to also be able to take advantage of their synergies during training.

[6] https://github.com/usnistgov/trec_eval
[7] https://huggingface.co/Helsinki-NLP

Original	après une réunion de spécialistes durant trois jours en autriche importante réforme de l'orthographe allemande à partir de fin 1995. les experts de ces trois pays ont tenue une conférence de trois jours sur la réforme de l'orthographe de la langue allemande. ils ont décidé que celle-ci pourra être appliquée dès la fin de l'année 1995 et sera contraignante à l'issue d'une période de transition de six ans, à partir de l'an 2001. l'application de la réforme devra être surveillée par une ̈commission sur l'orthographe ̈.
PSQ	year own german important in first be as germany will reform austria written letter 2001 end 1995 conference three experts
SPLADE-X (BI)	aut deutschen allemand language years script year application reform important allemande spelling deutsche grammar transition conference languages ##rap ort phone

Figure 1: Example token weights produced by PSQ and SPLADE-X (BI). The example exhibits a French relevant document for the query "german spelling reform".

Cross-language expansions.

Figure 1 shows a relevant French document and the top-20 terms in the PSQ and SPLADE-X document vectors. It is interesting to see that both systems rank this document highly, despite very different term weights. Among the top terms, both the systems cover two query terms apiece, whereas a union covers all of them. This might shed some light on why the combination of the two systems improves over the individual systems, and perhaps can open up ways to generate a low indexing footprint as the union of the top-k entries from the two lists.

The Curious Case of XLM-R.

Our choice of mBERT as a multilingual encoder for SPLADE-X was empirical, despite the fact that XLM-R has been shown to be outperforming mBERT on several tasks [45]. Figure 2 shows Kernel Density Estimation (KDE) plots generated using the mBERT and XLM-R MLM decoders for a specific query term. We observe that a XLM-R has a relatively higher mean than mBERT, which affects the SPLADE-X model as fewer terms are being sparsified by the ReLU. To counteract this effect, we used an additional LayerNorm that normalizes weights following Eqn. 3. We then trained the monolingual SPLADE models with both mBERT and XLM-R embeddings on English MS MARCO and tested it on the English MS MARCO dev set. We observed an MRR@10 of 0.347 for mBERT, compared to a 0.295 for XLM-R. Investigating the output generated by XLM-R, we see the distribution of weights is quite uniform relative to mBERT. For future work, we intend to investigate ways to get sharper distribution of weights

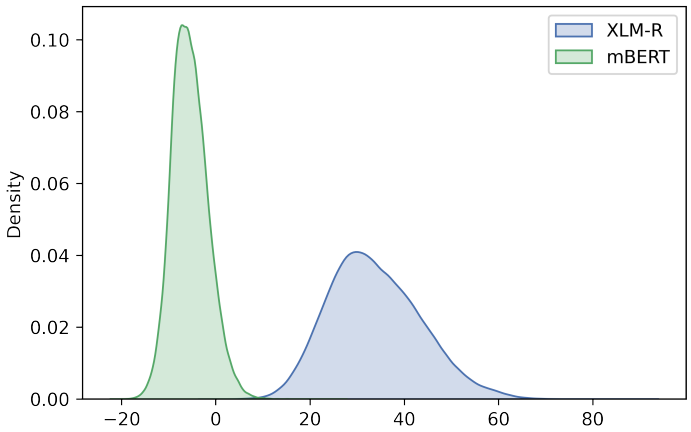

Figure 2: KDE plot for mBERT and XLM-R LM weights for "capital" in the query "What is the capital city of France"

from the XLM-R models, which might in turn lead to more effective retrieval.

5. Conclusion

We propose SPLADE-X, a sparse retrieval model that performs cross-language lexical expansion. We introduce a joint bilingual training procedure using both the monolingual and the translated MS MARCO collections with an additional alignment loss between the two. Our experiments show that our model performs on par with a strong PSQ baseline on several CLIR test collections and, when combined, performs significantly better than the individual systems. In the future, we would like to explore better ways to combine the PSQ with the SPLADE-X training. Another direction is to investigate integration of multilingual models belonging to the XLM-R family into the SPLADE-X framework.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[2] G. W. Furnas, T. K. Landauer, L. M. Gomez, S. T. Dumais, The vocabulary problem in human-system communication, Communications of the ACM 30 (1987) 964–971.

[3] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, in: TREC, 1994, pp. 109–123.

[4] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, T. Wang, MS MARCO: A human generated machine reading comprehension dataset, 2018. arXiv:1611.09268.
[5] L. H. Bonifacio, I. Campiotti, V. Jeronymo, R. Lotufo, R. Nogueira, mMARCO: A multilingual version of the MS MARCO passage ranking dataset, arXiv preprint arXiv:2108.13897 (2021).
[6] S. Nair, E. Yang, D. Lawrie, K. Duh, P. McNamee, K. Murray, J. Mayfield, D. W. Oard, Transfer learning approaches for building cross-language dense retrieval models, in: Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I, 2022, pp. 382–396.
[7] H. Zamani, M. Dehghani, W. B. Croft, E. Learned-Miller, J. Kamps, From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018, pp. 497–506.
[8] Z. Dai, J. Callan, Context-aware sentence/passage term importance estimation for first stage retrieval, arXiv preprint arXiv:1910.10687 (2019).
[9] A. Mallia, O. Khattab, T. Suel, N. Tonellotto, Learning passage impacts for inverted indexes, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1723–1727.
[10] J. Lin, X. Ma, A few brief notes on DeepImpact, COIL, and a conceptual framework for information retrieval techniques, arXiv preprint arXiv:2106.14807 (2021).
[11] S. Zhuang, G. Zuccon, Tilde: Term independent likelihood model for passage re-ranking, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1483–1492.
[12] Y. Bai, X. Li, G. Wang, C. Zhang, L. Shang, J. Xu, Z. Wang, F. Wang, Q. Liu, Sparterm: Learning term-based sparse representation for fast text retrieval, arXiv preprint arXiv:2010.00768 (2020).
[13] T. Formal, B. Piwowarski, S. Clinchant, SPLADE: Sparse lexical and expansion model for first stage ranking, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2288–2292.
[14] T. Formal, C. Lassance, B. Piwowarski, S. Clinchant, Splade v2: Sparse lexical and expansion model for information retrieval, arXiv preprint arXiv:2109.10086 (2021).
[15] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2020. arXiv:1911.02116.
[16] R. Nogueira, W. Yang, J. Lin, K. Cho, Document expansion by query prediction, arXiv preprint arXiv:1904.08375 (2019).
[17] R. Nogueira, J. Lin, From doc2query to docTTTTTquery, Technical Report, University of Waterloo, 2019. URL: https://cs.uwaterloo.ca/~jimmylin/publications/Nogueira_Lin_2019_docTTTTTquery-v2.pdf.
[18] Z. Chi, L. Dong, F. Wei, W. Wang, X.-L. Mao, H. Huang, Cross-lingual natural language generation via pre-training, Proceedings of the AAAI Conference on Artificial Intelligence 34 (2020) 7570–7577. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6256. doi:10.1609/aaai.v34i05.6256.
[19] P. Shi, R. Zhang, H. Bai, J. Lin, Cross-lingual training with dense retrieval for document

retrieval, arXiv preprint arXiv:2109.01628 (2021).
[20] S. MacAvaney, L. Soldaini, N. Goharian, Teaching a new dog old tricks: Resurrecting multilingual retrieval using zero-shot learning, in: Proceedings of the 42nd European Conference on Information Retrieval Research, 2020, pp. 246–254. URL: https://link.springer.com/chapter/10.1007/978-3-030-45442-5_31. doi:10.1007/978-3-030-45442-5_31.
[21] P. Shi, J. Lin, Cross-lingual relevance transfer for document retrieval, arXiv preprint arXiv:1911.02989 (2019).
[22] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, arXiv preprint arXiv:2004.09813 (2020).
[23] J.-H. Yang, X. Ma, J. Lin, Sparsifying sparse representations for passage retrieval by top-k masking, arXiv preprint arXiv:2112.09628 (2021).
[24] J. Lin, A proposed conceptual framework for a representational approach to information retrieval, arXiv preprint arXiv:2110.01529 (2021).
[25] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, arXiv preprint arXiv:2004.04906 (2020).
[26] Z. Wu, Y. Xiong, S. X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3733–3742.
[27] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.
[28] B. Paria, C.-K. Yeh, I. E. Yen, N. Xu, P. Ravikumar, B. Póczos, Minimizing FLOPs to learn efficient sparse representations, arXiv preprint arXiv:2004.05665 (2020).
[29] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models, arXiv preprint arXiv:2104.08663 (2021).
[30] K. Darwish, D. W. Oard, Probabilistic structured query methods, in: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 2003, pp. 338–344.
[31] J. Barry, E. Boschee, M. Freedman, S. Miller, SEARCHER: Shared embedding architecture for effective retrieval, in: Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020), European Language Resources Association, Marseille, France, 2020, pp. 22–25. URL: https://aclanthology.org/2020.clssts-1.4.
[32] R. Zbib, L. Zhao, D. Karakos, W. Hartmann, J. DeYoung, Z. Huang, Z. Jiang, N. Rivkin, L. Zhang, R. Schwartz, et al., Neural-network lexical translation for cross-lingual IR from text and speech, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 645–654.
[33] S. Nair, P. Galuscakova, D. W. Oard, Combining contextualized and non-contextualized query translations to improve CLIR, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1581–1584.
[34] G. V. Cormack, C. L. Clarke, S. Buettcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009, pp. 758–759.

[35] M. Braschler, C. Peters, CLEF 2003 methodology and metrics, in: Comparative Evaluation of Multilingual Information Access Systems, Springer Berlin Heidelberg, 2004, pp. 7–20.

[36] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[37] L. Gao, J. Callan, Unsupervised corpus aware language model pre-training for dense passage retrieval, arXiv preprint arXiv:2108.05540 (2021).

[38] P. Yang, H. Fang, J. Lin, Anserini: Enabling the use of lucene for information retrieval research, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 1253–1256.

[39] J. Xu, R. Weischedel, Cross-lingual information retrieval using hidden Markov models, in: 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 2000, pp. 95–103.

[40] F. J. Och, H. Ney, A systematic comparison of various statistical alignment models, Computational Linguistics (2003).

[41] P. Koehn, Europarl: A parallel corpus for statistical machine translation, Machine Translation Summit, 2005 (2005) 79–86.

[42] D. Kamholz, J. Pool, S. Colowick, PanLex: Building a resource for panlingual lexical translation, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 3145–3150. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1029_Paper.pdf.

[43] J. Palotti, H. Scells, G. Zuccon, TrecTools: An open-source python library for information retrieval practitioners involved in TREC-like campaigns, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 1325–1328.

[44] Z. Wang, S. Mayhew, D. Roth, et al., Cross-lingual ability of multilingual BERT: An empirical study, arXiv preprint arXiv:1912.07840 (2019).

[45] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, M. Johnson, Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation, in: International Conference on Machine Learning, PMLR, 2020, pp. 4411–4421.