Proceedings of the Cross-Language Search and Summarization of Text and Speech Workshop, pages 26–31
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Cross-lingual Information Retrieval with BERT

Zhuolin Jiang[†], Amro El-Jaroudi[†‡], William Hartmann[†], Damianos Karakos[†], Lingjun Zhao[†]

[†]Raytheon BBN Technologies, Cambridge, MA, 02138
[‡]University of Pittsburgh, Pittsburgh, PA, 15261
{zhuolin.jiang, amro.a.eljaroudi-nr, william.hartmann, damianos.karakos, lingjun.zhao}@raytheon.com

Abstract

Multiple neural language models have been developed recently, *e.g.*, BERT and XLNet, and achieved impressive results in various NLP tasks including sentence classification, question answering and document ranking. In this paper, we explore the use of the popular bidirectional language model, BERT, to model and learn the relevance between English queries and foreign-language documents in the task of cross-lingual information retrieval. A deep relevance matching model based on BERT is introduced and trained by finetuning a pretrained multilingual BERT model with weak supervision, using home-made CLIR training data derived from parallel corpora. Experimental results of the retrieval of Lithuanian documents against short English queries show that our model is effective and outperforms the competitive baseline approaches.

Keywords: Cross-lingual Information Retrieval; Neural Network Models; Relevance Matching; Weak Supervision

1. Introduction

A traditional cross-lingual information retrieval (CLIR) system consists of two components: machine translation and monolingual information retrieval (Nie, 2010). The idea is to solve the translation problem first, then the cross-lingual IR problem become monolingual IR. However, the performance of translation-based approaches is limited by the quality of the machine translation and it needs to handle to translation ambiguity (Zhou et al., 2012). One possible solution is to consider the translation alternatives of individual words of queries or documents as in (Zbib et al., 2019; Xu and Weischedel, 2000), which provides more possibilities for matching query words in relevant documents compared to using single translations. But the alignment information is necessarily required in the training stage of the CLIR system to extract target-source word pairs from parallel data and this is not a trivial task.

To achieve good performance in IR, deep neural networks have been widely used in this task. These approaches can be roughly divided into two categories. The first class of approaches uses pretrained word representations or embeddings, such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), directly to improve IR models. Usually these word embeddings are pretrained on large scale text corpora using co-occurrence statistics, so they have modeled the underlying data distribution implicitly and should be helpful for building discriminative models. (Vulic and Moens, 2015) and (Litschko et al., 2018) used pretrained bilingual embeddings to represent queries and foreign documents, and then ranked documents by cosine similarity. (Zheng and Callan, 2015) used word2vec embeddings to learn query term weights. However, their training objectives of trained neural embeddings are different from the objective of IR.

The second set of approaches design and train deep neural networks based on IR objectives. These methods have shown impressive results on monolingual IR datasets (Xiong et al., 2017; Guo et al., 2016; Dehghani et al., 2017). They usually rely on large amounts of query-document relevance annotated data that are expensive to obtain, especially for low-resource language pairs in cross-lingual IR tasks. Moreover, it is not clear whether they generalize well when documents and queries are in different languages.

Recently multiple pretrained language models have been developed such as BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019), that model the underlying data distribution and learn the linguistic patterns or features in language. These models have outperformed traditional word embeddings on various NLP tasks (Yang et al., 2019; Devlin et al., 2019; Peters et al., 2018; Lan et al., 2019). These pretrained models also provided new opportunities for IR. Therefore, several recent works have successfully applied BERT pretrained models for monolingual IR (Dai and Callan, 2019; Akkalyoncu Yilmaz et al., 2019) and passage re-ranking (Nogueira and Cho, 2019).

In this paper, we extend and apply BERT as a ranker for CLIR. We introduce a cross-lingual deep relevance matching model for CLIR based on BERT. We finetune a pretrained multilingual model with home-made CLIR data and obtain very promising results. In order to finetune the model, we construct a large amount of training data from parallel data, which is mainly used for machine translation and is much easier to obtain compared to the relevance labels of query-document pairs. In addition, we don't require the source-target alignment information to construct training samples and avoid the quality issues of machine translation in traditional CLIR. The entire model is specifically optimized using a CLIR objective. Our main contributions are:

- We introduce a cross-lingual deep relevance architecture with BERT, where a pretrained multilingual BERT model is adapted for cross-lingual IR.
- We define a proxy CLIR task which can be used to easily construct CLIR training data from bitext data, without requiring any amount of relevance labels of query-document pairs in different languages.

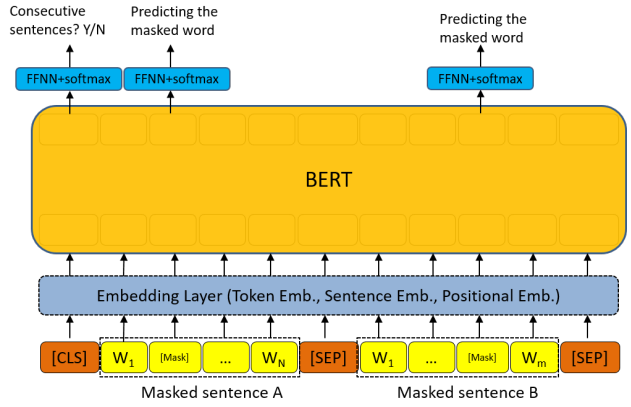

Figure 1: BERT pretraining architecture (Devlin et al., 2019). FFNN denotes feed-forward neural network.

2. Our approach

2.1. Motivation

BERT (Devlin et al., 2019) is the first bidirectional language model, which makes use of left and right word contexts simultaneously to predict word tokens. It is trained by optimizing two objectives: masked word prediction and next sentence prediction. As shown in Figure 1, the inputs are a pair of masked sentences in the same language, where some tokens in the both sentences are replaced by symbol '[Mask]'. The BERT model is trained to predict these masked tokens, by capturing within or across sentence meaning (or context), which is important for IR. The second objective aims to judge whether the sentences are consecutive or not. It encourages the BERT model to model the relationship between two sentences. The self-attention mechanism in BERT models the local interactions of words in sentence A with words in sentence B, so it can learn pairwise sentence or word-token relevance patterns. The entire BERT model is pretrained on large scale text corpora and learns linguistic patterns in language. So search tasks with little training data can still benefit from the pretrained model.

Finetuning BERT on search task makes it learn IR specific features. It can capture query-document exact term matching, bi-gram features for monolingual IR as introduced in (Dai and Callan, 2019). Local matchings of words and n-grams have proven to be strong neural IR features. Bigram modeling is important, because it can learn the meaning of word compounds (bi-grams) from the meanings of individual words. Motivated by this work, we aim to finetune the pretrained BERT model for cross-lingual IR.

2.2. Finetuning BERT for CLIR

Figure 2 shows the proposed CLIR model architecture with BERT. The inputs are pairs of single-word queries q in English and foreign-language sentences s. This is different from the pretraining model in Figure 1, where the model is fed with pairs of sentences in the same language. We concatenate the query q and the foreign-language sentence s into a text sequence '[[CLS], q, [SEP], s, [SEP]]'. The output embedding of the first token '[CLS]' is used as a representation of the entire query-sentence pair. Then it is fed into a single layer feed-forward neural network to predict the relevance score, which is the probability, $p(q|s)$, of query q occurring in sentence s.

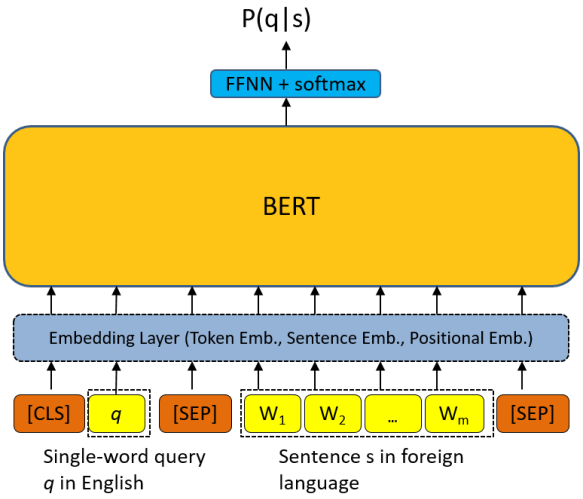

Figure 2: Fine-tuned CLIR BERT model architecture.

There are three types of parameterized layers in this model: (1) an embedding layer including token embedding, sentence embedding and positional embedding (Devlin et al., 2019); (2) BERT layers which are 12 layers of transformer blocks; (3) a feed-forward neural network (FFNN) which is a single layer neural network in our implementation. The embedding layer and BERT layer are initialized with the pretrained BERT model [1], while the FFNN is learned from scratch. During finetuning, the entire model is tuned to learn more CLIR-specific features. We only train the model using single-word queries since the queries in MATERIAL dataset are typically short and keyword based, but our approach can be easily extended to be multi-word queries or query phrases. After finetuning, this model produces a sentence-level relevance score for a pair of input query and foreign language sentence.

For the CLIR task, given a user-issued query Q, the foreign-language document Doc is ranked by its relevance score with respect to Q. The document-level relevance score $P(Doc \text{ is } R|Q)$ is calculated by aggregating the sentence-level scores with a Noisy-OR model:

$$\begin{aligned} P(Doc \text{ is } R|Q) &= P(Q \text{ occurs at least in one sentence in } Doc) \\ &= 1 - \prod_{s \in Doc} (1 - P(Q|s)) \\ &= 1 - \prod_{s \in Doc} (1 - \prod_{q \in Q} p(q|s)) \end{aligned} \quad (1)$$

Note that a multi-word query will be split into multiple single-word queries when computing document-level relevance scores. The individual query terms $q \in Q$ are modeled independently.

[1] We used the pretrained multi-lingual BERT model, which is trained on the concatenation of monolingual Wikipedia corpora from 104 languages. It has 12 layers, 768 hidden dimensions, 12 self-attention heads and 110 million parameters.

Query in English	Foreign-language sentence	Relevant
doctors	medikų teigimu dabar veikianti sistema efektyvi	Yes
allege	medikų teigimu dabar veikianti sistema efektyvi	Yes
controller	medikų teigimu dabar veikianti sistema efektyvi	No
leisure	medikų teigimu dabar veikianti sistema efektyvi	No

Table 1: Four training examples derived from a bitext: *Source-Lithuanian*: medikų teigimu dabar veikianti sistema efektyvi; *Target-English*: doctors allege that the system currently in operation is effective.

2.3. Finetuning using Weak Supervision

To finetune the BERT CLIR model, we start with bitext data in English and the desired foreign-language. We then define a proxy CLIR task to construct training samples: Given a foreign-language sentence s and an English query term q, sentence s is relevant to q if q occurs in one plausible translation of s. Any non-stop English word in the bitext can serve as a single-word query. The English word and its the corresponding foreign-language sentence constitute a positive example. Similarly, we randomly select other words from the English vocabulary, which are not in the English sentence, to be query words to construct negative examples. Table 1 shows an illustration of constructing four training examples from a bitext in Lithuanian and English. We select 'doctors' and 'allege' in the English sentence as two single-word queries and use the Lithuanian sentence to construct two positive examples, and pick another two words "controller" and "leisure" in the English vocabulary, which are not in the English sentence, to construct negative examples. In this way, we can construct a large-scale training corpus for CLIR using parallel data only, which are much easier to obtain compared to query-document relevance annotated data.

3. Experiments

We report experimental results on the retrieval of Lithuanian text and speech documents against short English queries. We use queries and retrieval corpora provided by the IARPA MATERIAL program. The retrieval corpora have two datasets: an analysis set (about 800 documents) and a development set (about 400 documents). The query set $Q1$ contains 300 queries.

To construct the training set, we use parallel sentences released under the MATERIAL (MAT, 2017) and the LORILEI (LOR, 2015) programs. We also include a parallel lexicon downloaded from Panlex (Kamholz et al., 2014). These parallel data contain about 2.6 million pairs of bitexts. We extract about 54 million training samples from these parallel data to finetune BERT. The positive-negative ratio of CLIR training data is $1:2$. To finetune BERT, we use the ADAM optimizer with an initial learning rate set to 1×10^{-5}, batch size of 32 and max sequence length of 128. We report the results from the model trained for one epoch.The training took one week using a Telsa V100 GPU.

We also extract 877K testing samples from the bitexts in MATERIAL Lithuanian analysis set to test the classification accuracy of different neural CLIR models. The positive-negative ratio of this test set is $1:1$. In addition, we evaluate our model on the MATERIAL Lithuanian analysis set and development set in terms of Mean Average Precision (MAP) and Maximum Query Weighted Value (MQWV) scores. MQWV is used in the MATERIAL program and denotes the maximum of the metric Average Query Weighted Value (AQWV): $AQWV = 1 - P_{Miss} - \beta P_{FA}$, where P_{Miss} is the average per-query miss rate, P_{FA} is the average per-query false alarm rate and β is a constant that changes the relative importance of the two types of error. We use $\beta = 40$. AQWV is the score using a single selected detection threshold. MQWV is the score that could be obtained with the optimal detection threshold. To verify the effectiveness of our BERT CLIR model, we compare against four baselines:

Probabilistic CLIR Model (Xu and Weischedel, 2000) is a generative probabilistic model which requires a probabilistic translation dictionary. The translation dictionary is generated from the word alignments of the parallel data. We used the GIZA++ (Och and Ney, 2003) and the Berkeley aligner (Haghighi et al., 2009) to estimate lexical translation probabilities.

Probabilistic Occurrence Model (Zbib et al., 2019) computes the document relevance score as the probability that each query term q occurs at least once in the document. $P(Doc \text{ is } R|Q) = \prod_{q \in Q} \Big[1 - \prod_{f \in Doc}(1 - p(q|f))\Big]$, where f is a foreign term in the document.

Query Relevance Attentional Neural Network Model (QRANN) (Zhao et al., 2019) uses an attention mechanism to compute a context vector derived from word embeddings in the foreign sentences, followed by a feed-forward layer to capture the relationship between query words. The idea is similar to a single transformer layer. The QRANN models are trained on multi-word queries, which are noun phrases in the English sentences of bitexts, and single-word queries.

Dot-product Model is a simplified version of QRANN, that computes a context vector from the word embeddings of foreign sentence using multiplicative attention, followed by the dot product of between the query embeddings and the context vector. The dot-product model is trained using single-word queries only.

3.1. Classification Accuracy of different neural CLIR models

The QRANN and Dot-product models are trained using the same CLIR training data used to train BERT model described earlier. The classification results of different neural CLIR approaches are shown in Table 2. The CLIR BERT model achieves the best result compared to other two neural models. From the confusion matrix in the table, BERT significantly improves the performance of classifying relevant query-sentence pairs (*i.e.*, true positives), while matching the performance of classifying irrelevant query-sentence

Approach	Accuracy	Confusion Matrix	
BERT	**95.3%**	0.93	0.07
		0.02	0.98
Dot-Product	84.2%	0.74	0.26
		0.07	0.93
QRANN	87.3%	0.73	0.27
		0.003	0.997

Table 2: Performance of classification accuracy on the generated query-sentence pairs from the bitexts of the MATERIAL analysis set. The first column in the confusion matrix corresponds to the positive class (*i.e.*, relevant query-sentence pair) while the second the column is the negative class.

Approach	phrase query subset	entire query set
Prob. CLIR	57.4	**61.2**
Prob. Occurrence	51.4	56.9
BERT	**61.3**	56.8
Dot-Product	50.8	39.2
QRANN	55.8	45.5

Table 3: Performance of MAP scores on the MATERIAL analysis set and Q1 queries.

pairs (*i.e.*, true negatives).

3.2. MAP scores of different CLIR models

We compare the MAP score of the BERT model with those of other CLIR models in Table 3. In the table, we report MAP scores on the phrase query subset and the entire query set separately, to see how our model trained with single-word queries performs on query phrases. In the model training stage, QRANN model is the only model that is trained with the query phrases directly, all other models (including BERT) in this experiment will split a multi-word query or query phrase into multiple single-word queries. Surprisingly, the BERT MAP scores for the phrase query subset is the best compared with the performances of other approaches. It shows that BERT model can produce better relevance model for single-word queries and foreign-language sentence.The table also shows that BERT outperforms the other neural approaches over the entire query set.

3.3. MQWV scores of different CLIR models

We compare BERT models with other CLIR models in terms of MQWV scores. The results are summarized in Table 4. The first row in the table shows the best results of non-neural CLIR models, which are probabilistic CLIR model and probabilistic occurrence model. In this table, we separate the results based on the type of source documents: text or speech. Speech documents are converted into text documents via automatic speech recognition (Povey et al., 2011). The results of the BERT model on the speech sets are the best, compared with the non-neural CLIR systems, QRANN and Dot-product models, while the results on the text sets are comparable to those from the non-neural systems, and better than the other neural systems.

Approach	Analysis Set		Development Set	
	Text	Speech	Text	Speech
Best non-neural system	**66.3**	63.3	**68.8**	64.0
BERT	65.7	**65.4**	61.8	**65.1**
Dot-Product	61.0	60.4	56.1	63.7
QRANN	62.3	58.4	57.2	65.0

Table 4: MQWV scores on the Lithuanian analysis and development sets and Q1 queries.

3.4. Analysis on attention patterns from BERT

In Figure 3, we visualize the attention patterns produced by the attention heads from a transformer layer for the input English query 'writing well' and the foreign-language sentence 'mano nuomone ši autore rašo arba gerai arba blogai arba vidutiniškai'. The query term 'writing' attends to the foreign word 'rašo' (source-target word matching), while also attends to the foreign word 'gerai' , which correspond to the next English word 'well' in the query (bigram modeling). BERT CLIR model is able to capture these local matching features, which have been proven to be strong neural IR features.

4. Conclusions

We introduce a deep relevance matching model based on BERT language modeling architecture for cross-lingual document retrieval. The self-attention based architecture models the interactions of query words with words in the foreign-language sentence. The relevance model is initialized by the pretrained multi-lingual BERT model, and then finetuned with home-made CLIR training data that are derived from parallel data. The results of the CLIR BERT model on the data released by the MATERIAL program are better than two other competitive neural baselines, and comparable to the results of the probabilistic CLIR model. Our future work will use public IR datasets in English to learn IR features with BERT and transfer them to cross-lingual IR.

Acknowledgement

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Air Force Research Laboratory contract number FA8650-17-C-9118.

5. Bibliographical References

Akkalyoncu Yilmaz, Z., Wang, S., Yang, W., Zhang, H., and Lin, J. (2019). Applying BERT to document retrieval with birch. In *Proceedings of the 2019 EMNLP-IJCNLP*.

Dai, Z. and Callan, J. (2019). Deeper text understanding for ir with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Dehghani, M., Zamani, H., Severyn, A., Kamps, J., and Croft, W. B. (2017). Neural ranking models with weak supervision. In *Proceedings of the 40th International*

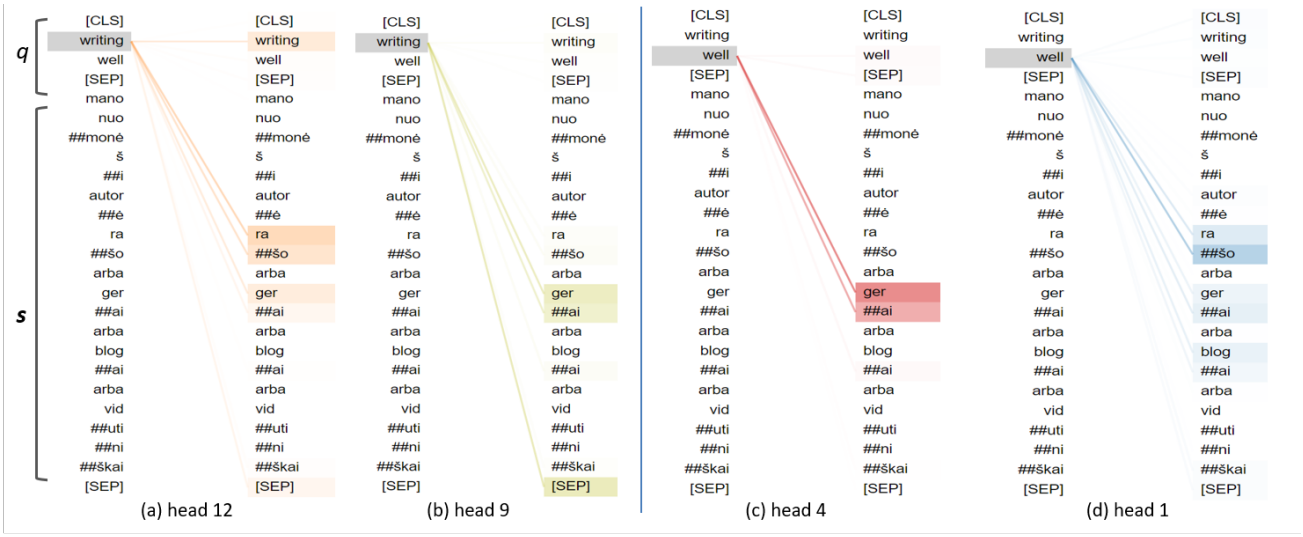

Figure 3: Visualization of CLIR BERT model. Colors identify the corresponding attention heads, while the line weight reflects the attention score. Different heads from layer 12 can capture different matching features. Word pieces‘ ra’ , ‘##šo’ in Lithuanian correspond to ‘’write’ in English while ‘ger’, ‘##ai’ are for ‘well’ in English. Head 12 and head 4 in (a)(c) can capture source-target word matching, head9 and head1 in (b)(d) could attend to its previous or next words (bigram modeling).

ACM SIGIR Conference on Research and Development in Information Retrieval.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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*.

Guo, J., Fan, Y., Ai, Q., and Croft, W. B. (2016). A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*.

Haghighi, A., Blitzer, J., DeNero, J., and Klein, D. (2009). Better word alignments with supervised itg models. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Process*.

Kamholz, D., Pool, J., and Colowick, S. (2014). PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Lan, Z.-Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *ArXiv*.

Litschko, R., Glavas, G., Ponzetto, S. P., and Vulic, I. (2018). Unsupervised cross-lingual information retrieval using monolingual data only. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*.

(2015). Darpa lorelei program - broad agency announcement (baa). `https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents`.

(2017). Iarpa material program - broad agency announcement (baa). `https://www.iarpa.gov/index.php/research-programs/material`.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Nie, J.-Y. (2010). *Cross-Language Information Retrieval*. Morgan and Claypool Publishers.

Nogueira, R. and Cho, K. (2019). Passage re-ranking with BERT. volume abs/1901.04085.

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

Pennington, J., Socher, R., and Mannin, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Vulic, I. and Moens, M.-F. (2015). Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Xiong, C., Dai, Z., Callan, J., Liu, Z., and Power, R. (2017). End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Xu, J. and Weischedel, R. (2000). Cross-lingual information retrieval using hidden markov models. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*.

Zbib, R., Zhao, L., Karakos, D., Hartmann, W., DeYoung, J., Huang, Z., Jiang, Z., Rivkin, N., Zhang, L., Schwartz, R. M., and Makhoul, J. (2019). 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*.

Zhao, L., Zbib, R., Jiang, Z., Karakos, D., and Huang, Z. (2019). Weakly supervised attentional model for low resource ad-hoc cross-lingual information retrieval. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*.

Zheng, G. and Callan, J. (2015). Learning to reweight terms with distributed representations. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zhou, D., Truran, M., Brailsford, T., Wade, V., and Ashman, H. (2012). Translation techniques in cross-language information retrieval. *ACM Comput. Surv.*, 45(1):1–44.