Improving Low-Resource Cross-lingual Document Retrieval by Reranking with Deep Bilingual Representations

Rui Zhang **Caitlin Westerfield** **Sungrok Shim**
Garrett Bingham **Alexander Fabbri** **William Hu** **Neha Verma** **Dragomir Radev**
Department of Computer Science, Yale University
{r.zhang, dragomir.radev}@yale.edu

Abstract

In this paper, we propose to boost low-resource cross-lingual document retrieval performance with deep bilingual query-document representations. We match queries and documents in both source and target languages with four components, each of which is implemented as a term interaction-based deep neural network with cross-lingual word embeddings as input. By including query likelihood scores as extra features, our model effectively learns to rerank the retrieved documents by using a small number of relevance labels for low-resource language pairs. Due to the shared cross-lingual word embedding space, the model can also be directly applied to another language pair without any training label. Experimental results on the MATERIAL dataset show that our model outperforms the competitive translation-based baselines on English-Swahili, English-Tagalog, and English-Somali cross-lingual information retrieval tasks.

1 Introduction

Cross-lingual relevance ranking, or Cross-Lingual Information Retrieval (CLIR), is the task of ranking foreign documents against a user query (Hull and Grefenstette, 1996; Ballesteros and Croft, 1996; Oard and Hackett, 1997; Darwish and Oard, 2003). As multilingual documents are more accessible, CLIR is increasingly more important whenever the relevant information is in other languages.

Traditional CLIR systems consist of two components: machine translation and monolingual information retrieval. Based on the translation direction, it can be further categorized into the document translation and the query translation approaches (Nie, 2010). In both cases, we first solve the translation problem, and the task is transformed to the monolingual setting. However, while conceptually simple, the performance of this modular approach is fundamentally limited by the quality of machine translation.

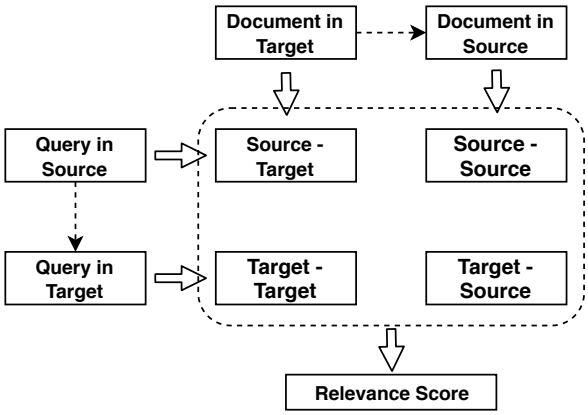
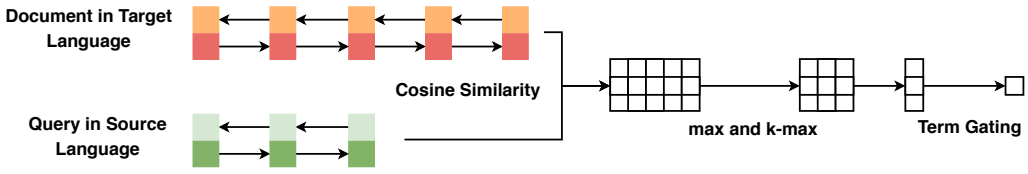


Figure 1: Cross-lingual Relevance Ranking with Bilingual Query and Document Representation.

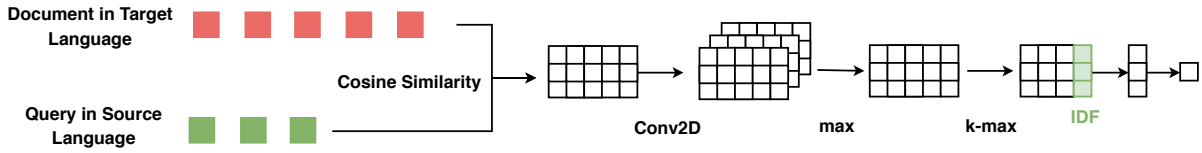
Recently, many deep neural IR models have shown promising results on monolingual data sets (Huang et al., 2013; Guo et al., 2016; Pang et al., 2016; Mitra et al., 2016, 2017; Xiong et al., 2017; Hui et al., 2017, 2018; McDonald et al., 2018). They learn a scoring function directly from the relevance label of query-document pairs. However, it is not clear how to use them when documents and queries are not in the same language. Furthermore, those deep neural networks need a large amount of training data. This is expensive to get for low-resource language pairs in our cross-lingual case.

In this paper, we propose a cross-lingual deep relevance ranking architecture based on a bilingual view of queries and documents. As shown in Figure 1, our model first translates queries and documents and then uses four components to match them in both the source and target language. Each component is implemented as a deep neural network, and the final relevance score combines all components which are jointly trained given the relevance label. We implement this based on state-

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3173–3179
Florence, Italy, July 28 - August 2, 2019.



(a) Bilingual POSIT-DRMM. The colored box represents hidden states in bidirectional LSTMs.



(b) Bilingual PACRR-DRMM. The colored box represents cross-lingual word embeddings. Bilingual PACRR is the same except it uses a single MLP at the final stage.

Figure 2: Model architecture. We only show the component of the source query with the target document.

of-the-art term interaction models because they enable us to make use of cross-lingual embeddings to explicitly encode terms of queries and documents even if they are in different languages. To deal with the small amount of training data, we first perform query likelihood retrieval and include the score as an extra feature in our model. In this way, the model effectively learns to rerank from a small number of relevance labels. Furthermore, since the word embeddings are aligned in the same space, our model can directly transfer to another language pair with no additional training data.

We evaluate our model on the MATERIAL CLIR dataset with three language pairs including English to Swahili, English to Tagalog, and English to Somali. Experimental results demonstrate that our model outperforms other translation-based query likelihood retrieval and monolingual deep relevance ranking approaches.

2 Our Method

In cross-lingual document retrieval, given a user query in the source language Q and a document in the target language D, the system computes a relevance score $s(Q, D)$. As shown in Figure 1, our model first translates the document as $\hat{D}$ or the query as $\hat{Q}$, and then it uses four separate components to match: (1) source query with target document, (2) source query with source document, (3) target query with source document, (4) target query with target document. The final relevance score combines all components:

$$s(Q, D) = s(Q, D) + s(Q, \hat{D}) + s(\hat{Q}, \hat{D}) + s(\hat{Q}, D)$$

To implement each component, we extend three state-of-the-art *term interaction models*: PACRR (Position-Aware Convolutional Recurrent Relevance Matching) proposed by Hui et al. (2017), POSIT-DRMM (POoled SImilariTy DRMM) and PACRR-DRMM proposed by McDonald et al. (2018). In term interaction models, each query term is scored to a document's terms from the interaction encodings, and scores for different query terms are aggregated to produce the query-document relevance score.

2.1 Bilingual POSIT-DRMM

This model is illustrated in Figure 2a. We first use bidirectional LSTMs (Hochreiter and Schmidhuber, 1997) to produce the context-sensitive encoding of each query and document term. We also add residual connection to combine the pre-trained term embedding and the LSTM hidden states. For the source query and document term, we can use the pre-trained word embedding in the source language. For the target query and document term, we first align the pre-trained embedding in the target language to the source language and then use this cross-lingual word embedding as the input to LSTM. Thereafter, we produce the document-aware query term encoding by applying max pooling and k-max pooling over the cosine similarity matrix of query and document terms. We then use an MLP to produce term scores, and the relevance score is a weighted sum over all terms in the query with a term gating mechanism.

	EN->SW			EN->TL			EN->SO
# Document	813			844			695
# Document Token (Min/Avg/Max)	34/341/1724			32/404/2501			69/370/2671
Query Set	Q1	Q2	Q3	Q1	Q2	Q3	Q1
# Query	300	400	600	300	400	600	300
# Relevant Pairs	411	489	828	236	576	1018	496

Table 1: The MATERIAL dataset statistics. For SW and TL, we use the ANALYSIS document set with Q1 for training, Q2 for dev, and Q3 for test. For transfer learning to SO, we use the DEV document set with Q1. Q1 contains open queries where performers can conduct any automatic or manual exploration while Q2 and Q3 are closed queries where results must be generated with fully automatic systems with no human in the loop.

2.2 Bilingual PACRR and Bilingual PACRR-DRMM

These models are shown in Figure 2b. We first align the word embeddings in the target language to the source language and build a query-document similarity matrix that encodes the similarity between the query and document term. Depending on the query language and document language, we construct four matrices, $\text{SIM}_{Q,D}$, $\text{SIM}_{Q,\hat{D}}$, $\text{SIM}_{\hat{Q},\hat{D}}$, $\text{SIM}_{\hat{Q},D}$, for each of the four components. Then, we use convolutional neural networks over the similarity matrix to extract n-gram matching features. We then use max-pooling and k-max-pooling to produce the feature matrix where each row is a document-aware encoding of a query term. The final step computes the relevance score: Bilingual PACRR uses an MLP on the whole feature matrix to get the relevance score, while Bilingual PACRR-DRMM first uses an MLP on individual rows to get query term scores and then use a second layer to combine them.

3 Related Work

Cross-lingual Information Retrieval. Traditional CLIR approaches include document translation and query translation, and more research efforts are on the latter (Oard and Hackett, 1997; Oard, 1998; McCarley, 1999; Franz et al., 1999). Early methods use the dictionary to translate the user query (Hull and Grefenstette, 1996; Ballesteros and Croft, 1996; Pirkola, 1998). Other methods include the single best SMT query translation (Chin et al., 2014) and the weighted SMT translation alternatives known as the probabilistic structured query (PSQ) (Darwish and Oard, 2003; Ture et al., 2012). Recently, Bai et al. (2010) and Sokolov et al. (2013) propose methods to learn the sparse query-document associations from supervised ranking signals on cross-lingual Wikipedia and patent data, respectively. Furthermore, Vulić and Moens (2015) and Litschko et al. (2018) use cross-lingual word embeddings to represent both queries and documents as vectors and perform IR by computing the cosine similarity. Schamoni et al. (2014) and Sasaki et al. (2018) also use an automatic process to build CLIR datasets from Wikipeida articles.

Neural Learning to Rank. Most of neural learning to rank models can be categorized in two groups: representation based (Huang et al., 2013; Shen et al., 2014) and interaction based (Pang et al., 2016; Guo et al., 2016; Hui et al., 2017; Xiong et al., 2017; McDonald et al., 2018). The former builds representations of query and documents independently, and the matching is performed at the final stage. The latter explicitly encodes the interaction between terms to direct capture word-level interaction patterns. For example, the DRMM (Guo et al., 2016) first compares the term embeddings of each pair of terms within the query and the document and then generates fixed-length matching histograms.

4 Experiments

Training and Inference. We first use the Indri[1] system which uses query likelihood with Dirichlet Smoothing (Zhai and Lafferty, 2004) to preselect the documents from the collection. To build the training dataset, for each positive example in the returned list, we randomly sample one negative example from the documents returned by Indri. The model is then trained with a binary cross-entropy loss. On validation or testing set, we use our prediction scores to rerank the documents returned by Indri.

Extra Features. Following the previous work (Severyn and Moschitti, 2015; Mohan et al., 2017; McDonald et al., 2018), we compute the final relevance score by a linear model to combine the model output with the following set of extra fea-

[1] www.lemurproject.org/indri.php

	EN->SW				EN->TL			
	MAP	P@20	NDCG@20	AQWV	MAP	P@20	NDCG@20	AQWV
Query Translation and Document Translation with Indri								
Dictionary-Based Query Translation (DBQT)	20.93	4.86	28.65	6.50	20.01	5.42	27.01	5.93
Probabilistic Structured Query (PSQ)	27.16	5.81	36.03	12.56	35.20	8.18	44.04	19.81
Statistical MT (SMT)	26.30	5.28	34.60	13.77	37.31	8.77	46.77	21.90
Neural MT (NMT)	26.54	5.26	34.83	15.70	33.83	8.20	43.17	18.56
Deep Relevance Ranking								
PACRR	24.69	5.24	32.85	11.73	32.53	8.42	41.75	17.48
PACRR-DRMM	22.15	5.14	30.28	8.50	32.59	8.60	42.17	16.59
POSIT-DRMM	23.91	6.04	33.83	12.06	25.16	8.15	34.80	9.28
Deep Relevance Ranking with Extra Features in Section 4								
PACRR	27.03	5.34	35.36	14.18	41.43	8.98	49.96	27.46
PACRR-DRMM	25.46	5.50	34.15	12.18	35.61	8.69	45.34	22.70
POSIT-DRMM	26.10	5.26	34.27	14.11	39.35	9.24	48.41	25.01
Ours with Extra Features in Section 4: In-Language Training								
Bilingual PACRR	29.64	5.75	38.27	17.87	43.02	9.63	52.27	29.12
Bilingual PACRR-DRMM	26.15	5.84	35.54	12.92	38.29	9.21	47.60	22.94
Bilingual POSIT-DRMM	**30.13**	**6.28**	**39.68**	**18.69**	**43.67**	**9.73**	**52.80**	**29.12**
Bilingual POSIT-DRMM (3-model ensemble)	**31.60**	**6.37**	**41.25**	**20.19**	**45.35**	**9.84**	**54.26**	**31.08**

Table 2: Test set result on English to Swahili and English to Tagalog. We report the TREC ad-hoc retrieval evaluation metrics (MAP, P@20, NDCG@20) and the Actual Query Weighted Value (AQWV).

Train: EN->SW + EN->TL, Test: EN->SO			
	MAP	P@20	AQWV
PSQ	17.52	5.45	2.35
SMT	19.04	6.12	4.62
Bilingual POSIT-DRMM	**20.58**	**6.51**	**5.71**
+3-model ensemble	**21.25**	**6.68**	**5.89**

Table 3: Zero-shot transfer learning on English to Somali test set.

tures: (1) the Indri score with the language modeling approach to information retrieval. (2) the percentage of query terms with an exact match in the document, including the regular percentage and IDF weighted percentage. (3) the percentage of query term bigrams matches in the document.

Cross-lingual Word Embeddings. We apply the supervised iterative Procrustes approach (Xing et al., 2015; Conneau et al., 2018) to align two pre-trained mono-lingual fastText (Bojanowski et al., 2016) word embeddings using the MUSE implementation[2]. To build the bilingual dictionary, we use the translation pages of Wiktionary[3]. For Swahili, we build a training dictionary for 5301 words and a testing dictionary for 1326 words. For Tagalog, the training dictionary and testing dictionary contains 7088 and 1773 words, respectively. For Somali, the corresponding number is 7633 and 1909. We then learn the cross-lingual word embeddings from Swahili to English, from Tagalog to English, and from Somali to English. Therefore, all three languages are in the same word embedding space.

Data Sets and Evaluation Metrics. Our experiments are evaluated on the MATERIAL[4] program as summarized in Table 1. It consists of three language pairs with English queries on Swahili (EN->SW), Tagalog (EN->TL), Somali documents (EN->SO).

We use the TREC ad-hoc retrieval evaluation script[5] to compute Precision@20, Mean Average Precision (MAP), Normalized Discounted Cumulative Gain@20 (NDCG@20). We also report the Actual Query Weighted Value (AQWV) (NIST, 2017), a set-based metric with penalty for both missing relevant and returning irrelevant documents. We use $\beta = 40.0$ and find the best global fixed cutoff over all queries.

Baselines. For traditional CLIR approaches, we use query translation and document translation with the Indri system. For query translation, we use Dictionary-Based Query Translation (DBQT) and Probabilistic Structured Query (PSQ). For document translation, we use Statistical Machine Translation (SMT) and Neural Machine Translation (NMT). For SMT, we use the moses system (Koehn et al., 2007) with word alignments using mGiza and 5-gram KenLM language model (Heafield, 2011). For NMT, we use sequence-to-

[2] github.com/facebookresearch/MUSE

[3] https://www.wiktionary.org/

[4] www.iarpa.gov/index.php/research-programs/material

[5] https://trec.nist.gov/trec_eval/

sequence model with attention (Bahdanau et al., 2015; Miceli Barone et al., 2017) implemented in Marian (Junczys-Dowmunt et al., 2018).

For deep relevance ranking baselines, we investigate recent state-of-the-art models including PACRR, PACRR-DRMM, and POSIT-DRMM. These models and our methods all use an SMT-based document translation as input.

Implementation Details. For POSIT-DRMM and Bilingual POSIT-DRMM, we use the k-max-pooling with $k = 5$ and 0.3 dropout of the BiLSTM output. For PACRR, PACRR-DRMM and their bilingual counterparts, we use convolutional filter sizes with [1,2,3], and each filter size has 32 filters. We use $k = 2$ in the k-max-pooling. The loss function is minimized using the Adam optimizer (Kingma and Ba, 2014) with the training batch size as 32. We monitor the MAP performance on the development set after each epoch of training to select the model which is used on the test data.

4.1 Results and Discussion

Table 2 shows the result on EN->SW and EN->TL where we train and test on the same language pair.

Performance of Baselines. For query translation, PSQ is better than DBQT because PSQ uses a weighted alternative to translate query terms and does not limit to the fixed translation from the dictionary as in DBQT. For document translation, we find that both SMT and NMT have a similar performance which is close to PSQ. The effectiveness of different approaches depends on the language pair (PSQ for EN->SW and SMT for EN->TL), which is a similar finding with McCarley (1999) and Franz et al. (1999). In our experiments with deep relevance ranking models, we all use SMT and PSQ because they have strong performances in both language pairs and it is fair to compare.

Effect of Extra Features and Bilingual Representation. While deep relevance ranking can achieve decent performance, the extra features are critical to achieve better results. Because the extra features include the Indri score, the deep neural model essentially learns to rerank the document by effectively using a small number of training examples. Furthermore, our models with bilingual representations achieve better results in both language pairs, giving additional 1-3 MAP improvements over their counterparts. To compare language pairs, EN->TL has larger improvements over EN->SW. This is because EN->TL has better query translation, document translation, and query likelihood retrieval results from the baselines, and thus it enjoys more benefits from our model. We also found POSIT-DRMM works better than the other two, suggesting term-gating is useful especially when the query translation can provide more alternatives. We then perform ensembling of POSIT-DRMM to further improve the results.

Zero-Shot Transfer Learning. Table 3 shows the result for a zero-shot transfer learning setting where we train on EN->SW + EN->TL and directly test on EN->SO without using any Somali relevance labels. This transfer learning delivers a 1-3 MAP improvement over PSQ and SMT. This presents a promising approach to boost performance by utilizing relevance labels from other language pairs.

5 Conclusion

We propose to improve cross-lingual document retrieval by utilizing bilingual query-document interactions and learning to rerank from a small amount of training data for low-resource language pairs. By aligning word embedding spaces for multiple languages, the model can be directly applied under a zero-shot transfer setting when no training data is available for another language pair. We believe the idea of combining bilingual document representations using cross-lingual word embeddings can be generalized to other models as well.

Acknowledgements

We thank Petra Galuščáková, Douglas W. Oard, Efsun Kayi, Suraj Nair, Han-Chin Shing, and Joseph Barrow for their helpful discussion and feedback. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract # FA8650-17-C-9117. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Weinberger. 2010. Learning to rank with (a lot of) word features. *Information retrieval*, 13(3):291–314.

Lisa Ballesteros and Bruce Croft. 1996. Dictionary methods for cross-lingual information retrieval. In *International Conference on Database and Expert Systems Applications*, pages 791–801. Springer.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Jeffrey Chin, Maureen Heymans, Alexandre Kojoukhov, Jocelyn Lin, and Hui Tan. 2014. Cross-language information retrieval. US Patent 8,799,307.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *ICLR*.

Kareem Darwish and Douglas W Oard. 2003. Probabilistic structured query methods. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 338–344. ACM.

Martin Franz, J Scott McCarley, and Salim Roukos. 1999. Ad hoc and multilingual information retrieval at ibm. *NIST special publication SP*, pages 157–168.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 55–64. ACM.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2333–2338. ACM.

Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2017. Pacrr: A position-aware neural ir model for relevance matching. *arXiv preprint arXiv:1704.03940*.

Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2018. Co-pacrr: A context-aware neural ir model for ad-hoc retrieval. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 279–287. ACM.

David A Hull and Gregory Grefenstette. 1996. Querying across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–57. ACM.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised cross-lingual information retrieval using monolingual data only. In *SIGIR*.

J Scott McCarley. 1999. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 208–214. Association for Computational Linguistics.

Ryan McDonald, Georgios-Ioannis Brokos, and Ion Androutsopoulos. 2018. Deep relevance ranking using enhanced document-query interactions. *arXiv preprint arXiv:1809.01682*.

Antonio Valerio Miceli Barone, Jindřich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch. 2017. Deep architectures for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1291–1299. International World Wide Web Conferences Steering Committee.

Bhaskar Mitra, Eric Nalisnick, Nick Craswell, and Rich Caruana. 2016. A dual embedding space model for document ranking. *arXiv preprint arXiv:1602.01137*.

Sunil Mohan, Nicolas Fiorini, Sun Kim, and Zhiyong Lu. 2017. Deep learning for biomedical information retrieval: learning textual relevance from click logs. *BioNLP 2017*, pages 222–231.

Jian-Yun Nie. 2010. Cross-language information retrieval. *Synthesis Lectures on Human Language Technologies*, 3(1):1–125.

NIST. 2017. The Official Original Derivation of AQWV.

Douglas W Oard. 1998. A comparative study of query and document translation for cross-language information retrieval. In *Conference of the Association for Machine Translation in the Americas*, pages 472–483. Springer.

Douglas W Oard and Paul Hackett. 1997. Document translation for cross-language text retrieval at the university of maryland. In *TREC*, pages 687–696. Citeseer.

Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2016. A study of matchpyramid models on ad-hoc retrieval. *arXiv preprint arXiv:1606.04648*.

Ari Pirkola. 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 55–63. ACM.

Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. Cross-lingual learning-to-rank with shared representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 458–463.

Shigehiko Schamoni, Felix Hieber, Artem Sokolov, and Stefan Riezler. 2014. Learning translational and knowledge-based similarities from relevance rankings for cross-language retrieval. In *Proceedings of the 52 Annual Meeting of the Association for Computational Linguistics (ACL)*.

Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 373–382. ACM.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 373–374. ACM.

Artem Sokolov, Laura Jehl, Felix Hieber, and Stefan Riezler. 2013. Boosting cross-language retrieval by learning bilingual phrase associations from relevance rankings. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1688–1699.

Ferhan Ture, Jimmy Lin, and Douglas Oard. 2012. Combining statistical translation techniques for cross-language information retrieval. *Proceedings of COLING 2012*, pages 2685–2702.

Ivan Vulić and Marie-Francine Moens. 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*, pages 363–372. ACM.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011.

Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 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*, pages 55–64. ACM.

Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214.