Cross-Lingual Training of Neural Models for Document Ranking

Peng Shi, He Bai, and **Jimmy Lin**
David R. Cheriton School of Computer Science
University of Waterloo
{peng.shi, he.bai, jimmylin}@uwaterloo.ca

Abstract

We tackle the challenge of cross-lingual training of neural document ranking models for mono-lingual retrieval, specifically leveraging relevance judgments in English to improve search in non-English languages. Our work successfully applies multi-lingual BERT (mBERT) to document ranking and additionally compares against a number of alternatives: translating the training data, translating documents, multi-stage hybrids, and ensembles. Experiments on test collections in six different languages from diverse language families reveal many interesting findings: model-based relevance transfer using mBERT can significantly improve search quality in (non-English) mono-lingual retrieval, but other "low resource" approaches are competitive as well.

1 Introduction

This work proposes techniques for leveraging relevance judgments in a source language (English) to train neural models for mono-lingual document retrieval in multiple target (non-English) languages, what we refer to as cross-lingual training. Success in this task would make it easier to develop effective search engines in multiple (potentially low-resource) languages, without gathering expensive relevance judgments in *each* language. A blog post by Google suggests that the company is exploring this approach to improving web search across a number of languages.[1]

We are inspired by the work of Wu and Dredze (2019), who explored the cross-lingual potential of multi-lingual BERT as a zero-shot language transfer model for NLP tasks such as named-entity recognition and parsing. Mono-lingual BERT models (Devlin et al., 2019) have also proven effective in document retrieval (Dai and Callan, 2019; MacAvaney et al., 2019; Li et al., 2020). In particular, Akkalyoncu Yilmaz et al. (2019) demonstrated that BERT models fine-tuned with passage-level relevance data can transfer across domains: surprisingly, fine-tuning on social media data is effective for relevance classification on newswire documents without any additional modifications. Building on these results, we wondered if multi-lingual BERT could enable cross-lingual training of neural document ranking models as well.

The contribution of this work is to explore diverse methods to train neural document ranking models cross-lingually. While we are aware of two previous papers along these lines (Shi and Lin, 2019; MacAvaney et al., 2020), this work explores a far broader range of techniques and adds more nuance to previous findings. Beyond the basic approach proposed by these two papers, which we refer to as model-based transfer, we investigate additional approaches involving the translation of the training data, the translation of documents, hybrid models, as well as ensembles—which we broadly characterize into "high resource" and "low resource" settings. We show that various methods alone and in combination can yield robust increases in effectiveness across diverse languages with minimal resources, and that model-based cross-lingual transfer isn't the only way.

2 Approach

This work adopts the standard formulation of document ranking: given a user query Q, the task is to produce a ranking of documents from a collection that maximizes some ranking metric—in our case, average precision (AP). Given *source* language relevance judgments (in English), our task is to train a mono-lingual document ranking model for a *target* (non-English) language; that is, the queries and the documents are both in, for example, Bengali.

[1] https://www.blog.google/products/search/search-language-understanding-bert/

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2768–2773
November 16 - 20, 2020.

2.1 Preliminaries

Recent work on neural document ranking (Akkalyoncu Yilmaz et al., 2019; Dai and Callan, 2019) provides a general method for fine-tuning BERT: The input to the model comprises [`[CLS]`, Q `[SEP]` S `[SEP]`], which is the concatenation of the query Q and a sentence S, with the special tokens `[CLS]` and `[SEP]`. The final hidden state of the `[CLS]` token is passed to a single layer neural network with a softmax, obtaining the probability that sentence S is relevant to the query Q. Following Akkalyoncu Yilmaz et al. (2019), BERT is fine-tuned with data from the TREC Microblog Tracks (Lin et al., 2014) (MB for short). The authors showed that such a relevance matching model can be directly applied to effectively rank newswire documents, despite the mismatch in domains between training and test data; cf. Rücklé et al. (2020).

For document retrieval (i.e., at inference time), Akkalyoncu Yilmaz et al. (2019) first apply "bag of words" exact term matching to retrieve a candidate set of documents. Each document is split into sentences, and inference is applied on each sentence separately with BERT. The relevance score of each document is determined by combining the top k (by default, $k = 3$) scoring sentences with the document term-matching score as follows: $S_{doc} = \alpha \cdot S_r + (1 - \alpha) \cdot \sum_{i=1}^{k} w_i \cdot S_i$, where S_i is the i-th top sentence score according to BERT and S_r is the document level term-matching score. The parameters α and w_i's can be tuned via cross-validation. All candidate documents are sorted by the above score S_{doc} to produce the final output.

2.2 Cross-Lingual Relevance Transfer

Our main research question is as follows: Given English (source) training data, how can we bootstrap a good document ranking model in non-English (target) languages? We discuss a number of approaches below, which we characterize as "high" or "low" resource in terms of *annotation effort*.

Model-based transfer. Following Wu and Dredze (2019), the most obvious approach is to fine-tune mBERT using data in the source language, and apply inference directly on input in the target language. In essence, we follow the same setup as Akkalyoncu Yilmaz et al. (2019), with the exception that we use mBERT instead of (English) BERT. Note that this is essentially the approach explored in previous work (Shi and Lin, 2019; MacAvaney et al., 2020). We characterize this approach as "low resource" given that mBERT is pretrained in a self-supervised manner.

Training data translation. Instead of relying on mBERT to transfer models of relevance matching across languages, we can translate the English training data into the target language, and then fine-tune mBERT with the translated data.[2] At inference time, we directly apply the model on target-language documents. We considered two translation methods: Google Translate (MB_{gt}) and a simple embedding-based token-by-token translation approach (MB_{wt}). We characterize the first as "high resource" given the amount of bitext that is typically necessary to train a high-quality translation system, whereas the second as "low resource" since bilingual lexicons and aligned word embeddings are far easier to create.

Our token-based translation approach is inspired by Huang et al. (2019). The basic idea is to find the best token translation based on the cosine similarity between the token in the source language and candidate tokens in the target language. Specifically, for each token in the source language, the surface form is used for lookup in a bilingual dictionary. If the token has a unique translation, we use the translation directly. If it has multiple translations, we use an empirical scoring function $F(w, w_{t,i})$ to select the best translation. This scoring function calculates the cosine similarity between a candidate translation $w_{t,i}$ and the source token w based on its contextual tokens $w_{c,j}$ (in this work, we consider two words in the left context and two words in the right context), as follows:

$$F(w, w_{t,i}) = \gamma \cdot \cos(\text{E}(w), \text{E}(w_{t,i})) + (1 - \gamma) \cdot \sum_{j=1}^{m} \frac{\cos(\text{E}(w_{t,i}), \text{E}(w_{c,j}))}{(d_j + 1)^2} \quad (1)$$

where $\text{E}(w)$ is the bilingual embedding of the token w, d_j is the positional distance between the token w and its contextual token $w_{c,j}$, and γ is a hyperparameter for balancing the effects of the translation pair and the contextual tokens. Following previous work, we set γ to 0.5. If the source language token has no translations, the original surface form is kept unchanged.

Note that model-based transfer uses the *same* model across all languages, whereas this approach requires a separate model for *each* language.

[2] Note that here we are using mBERT in a purely monolingual manner since mono-lingual BERT models are not widely available for all target languages.

Doc Language	Source	# Topics	# Docs
Chinese	NTCIR 8	73	308,832
Arabic	TREC 2002	50	383,872
French	CLEF 2006	49	171,109
Hindi	FIRE 2012	50	331,599
Bengali	FIRE 2012	50	500,122
Spanish	TREC 3	25	57,868

Table 1: Dataset Statistics.

Hybrid transfer. Both approaches above can be combined in a stage-wise fashion: We can first fine-tune mBERT on the English data, and then fine-tune again on the translated training data (we refer to this as the en→gt direction). Alternatively, we can switch the order of fine-tuning (the gt→en direction). In these experiments, we used the output of Google Translate (and hence these are "high resource" approaches).

Document translation. Another way to leverage existing translation capabilities is to translate the documents at search time from the target language into the source language (English), and directly apply the mBERT model that is trained on MB_{en}. We used Google Translate in this method, and thus it is "high resource".

Ensembles. Ensembles of the above approaches can exploit multiple signal and resources. One approach is to interpolate scores from multiple sources, on a per-document basis: $S_{\text{agg}} = \beta \cdot S_{\text{model-transfer}} + (1-\beta) \cdot S_{\text{doc-translation}}$. This method is denoted ENS_{INT}, which combines model-based transfer and document translation (from the results, the two most promising techniques). Alternatively, we also experimented with Reciprocal Rank Fusion (Cormack et al., 2009) to aggregate two separate ranked lists, which is denoted ENS_{RRF}. These methods are "high resource".

For "low resource" ensembles, we aggregated signals from model-based transfer and the token-based approach for translating training data. These signals are either combined by per-document score interpolation or RRF, as per above.

3 Experimental Setup

We experimented with six test collections (in Chinese, Arabic, French, Hindi, Bengali and Spanish) from diverse language families (Sino-Tibetan, Semitic, Romance, and Indo-Aryan). Dataset statistics are shown in Table 1. Following standard practice in information retrieval, average precision (AP) up to rank 1000 and precision at rank 20 (P@20) were adopted as the evaluation metrics, computed with the `trec_eval` tool.

For the token-based translation method, we used the MUSE bilingual dictionary (Lample et al., 2018) and the aligned word embeddings from fastText (Joulin et al., 2018). For fine-tuning mBERT, we followed the same experimental setup as Akkalyoncu Yilmaz et al. (2019). We used data from the Microblog (MB) Tracks from TREC 2011–2014 (Lin et al., 2014) or its translated counterparts, setting aside 75% of the total data for training and the rest for validation, which was used for selecting the best model parameters. We trained each model using cross-entropy loss with a batch size of 16; the Adam optimizer was applied with an initial learning rate of 1×10^{-5}. During fine-tuning, the embeddings were fixed. The model with the highest AP on the validation set was chosen. We ran all experiments on an NVIDIA Tesla V100 16GB with PyTorch version 1.3.0. Each model was trained for up to 15 epochs, with an average running time of approximately two hours.

For retrieval, we used the open-source Anserini IR toolkit (Yang et al., 2018) with minor modifications based on version 0.6.0 to swap in Lucene Analyzers for different languages. Fortunately, Lucene provides analyzers for all the languages in our test collections. The query was used to retrieve the top 1000 hits from the corpus using BM25 or BM25+RM3 query expansion; default Anserini settings were used in both cases. Reranking with mBERT (see Section 2.1) used the approach with higher AP (either BM25 or BM25+RM3); the top three sentences were considered in aggregating sentence-level evidence. We applied five-fold cross-validation on all datasets and the parameters α, the w_i's, and β were obtained by grid search, choosing the parameters that yielded the highest AP.

4 Results

Our results are shown in Table 2. Models (0) and (1) show the effectiveness of BM25 and BM25 with RM3 query expansion. We see that with the exception of the French and Spanish collections, RM3 actually decreases effectiveness. This interesting finding was not further investigated, as our goal was simply to establish a strong baseline; however, these results are consistent with MacAvaney et al. (2020). For each language, we selected the higher of the two models as the starting point of reranking (see Section 2.1) as well as the baseline for compar-

Model	Train	Test	R	AP	P@20	AP	P@20	AP	P@20
				NTCIR8-zh		TREC2002-ar		CLEF2006-fr	
(**0**) BM25				0.4014	0.3849	0.2932	0.3610	0.3111	0.3184
(**1**) +RM3				0.3384	0.3616	0.2783	0.3490	0.3421	0.3408
(**2**) mBERT	MB_{en}	doc	l	0.4488$^{\blacktriangle}$	0.4507$^{\blacktriangle}$	0.3081	0.4050$^{\blacktriangle}$	0.3631$^{\blacktriangle}$	0.3633$^{\blacktriangle}$
(**3**) mBERT	MB_{gt}	doc	h	0.4618	0.4616	0.3148	0.4120	0.3596	0.3531
(**4**) mBERT	MB_{wt}	doc	l	0.4220	0.4322	0.3022	0.3950	0.3557	0.3551
(**5**) mBERT	MB_{en}	doc_{gt}	h	0.4513	0.4534	0.3272¶	0.4020	0.3800¶	0.3745
(**6**) Hybrid	$MB_{en\rightarrow gt}$	doc	h	0.4525§	0.4534§	0.3209§	0.4140§	0.3706	0.3694§
(**7**) Hybrid	$MB_{gt\rightarrow en}$	doc	h	0.4423§	0.4438§	0.3075	0.4120§	0.3490	0.3459
(**8**) ENS_{INT}	MB_{en}	$+doc_{gt}$	h	0.4561	0.4521	0.3269	0.4060	0.3818	0.3694
(**9**) ENS_{RRF}	MB_{en}	$+doc_{gt}$	h	0.4582	0.4562	0.3237	0.4060	0.3767	0.3694
(**10**) ENS_{INT}	MB_{en+wt}	doc	l	0.4490	0.4507	0.3086	0.4030	0.3628	0.3622
(**11**) ENS_{RRF}	MB_{en+wt}	doc	l	0.4404	0.4486	0.3074	0.4010	0.3613	0.3500
				FIRE2012-hi		FIRE2012-bn		TREC3-es	
(**0**) BM25				0.3867	0.4470	0.2881	0.3740	0.4197	0.6660
(**1**) +RM3				0.3660	0.4430	0.2833	0.3830	0.4912	0.7040
(**2**) mBERT	MB_{en}	doc	l	0.4207$^{\blacktriangle}$	0.4800$^{\blacktriangle}$	0.3101$^{\blacktriangle}$	0.4060$^{\blacktriangle}$	0.5056$^{\blacktriangle}$	0.7240
(**3**) mBERT	MB_{gt}	doc	h	0.4150	0.4710	0.2975	0.3890	0.5051	0.7400
(**4**) mBERT	MB_{wt}	doc	l	0.4289	0.4860	0.3050	0.4070	0.5032	0.7300
(**5**) mBERT	MB_{en}	doc_{gt}	h	0.4240	0.4810	0.3419¶	0.4470	0.5238¶	0.7700¶
(**6**) Hybrid	$MB_{en\rightarrow gt}$	doc	h	0.4218§	0.4850§	0.3078§	0.4020	0.4996	0.7140
(**7**) Hybrid	$MB_{gt\rightarrow en}$	doc	h	0.4181§	0.4780	0.3030	0.3950§	0.5058	0.7220
(**8**) ENS_{INT}	MB_{en}	$+doc_{gt}$	h	0.4320	0.4910	0.3479	0.4530	0.5215	0.7660
(**9**) ENS_{RRF}	MB_{en}	$+doc_{gt}$	h	0.4283	0.4890	0.3406	0.4320	0.5209	0.7560
(**10**) ENS_{INT}	MB_{en+wt}	doc	l	0.4377	0.4860	0.3112	0.4020	0.5077	0.7260
(**11**) ENS_{RRF}	MB_{en+wt}	doc	l	0.4340	0.4900	0.3127	0.4090	0.5082	0.7240

Table 2: Ranking effectiveness of different cross-lingual training methods. "R" = Resource: high or low.

isons below. We organize results into five findings below. Unless otherwise stated, Fisher's two-sided, paired randomization test (Smucker et al., 2007) at $p < 0.05$ was applied to test for statistical significance, with Bonferroni corrections as appropriate.

Finding #1: Model-based transfer, model (2), improves upon the baseline, with significant gains (denoted by $^{\blacktriangle}$) everywhere except for AP in Arabic and P@20 in Spanish. Since mBERT is widely available, mono-lingual retrieval improvements can be obtained "for free" with microblog relevance judgments in English. These results indicate that mBERT effectively transfers relevance matching across languages. This finding confirms previous work (Shi and Lin, 2019; MacAvaney et al., 2020), but see additional discussion below.

Finding #2: Comparing model-based transfer and the two approaches to translating training data, models (3) and (4), it is difficult to spot trends or reach definitive conclusions. Model-based transfer does not consistently beat simply translating the training data. In terms of AP, Google Translate, model (3), outperforms model-based transfer for Chinese and Arabic; token-based translation, model (4), beats model-based transfer in Hindi and achieves comparable scores in Arabic and Spanish. Interestingly, it is not always the case that Google Translate ("high resource") is better than token-based translation ("low resource"); the latter achieves higher AP for Hindi and Bengali. A Tukey's HSD test across models (2–4) showed no significant differences.

These results suggest that model-based transfer is not the only effective approach, and that simply translating the training data is at least competitive; neither Shi and Lin (2019) nor MacAvaney et al. (2020) explored this obvious baseline.

Finding #3: Results show that hybrid two stage training in the en→gt direction, model (6), can further improve over model-based transfer alone or translating training data with Google Translate alone, but the gains are not consistent; lower AP than either models (2) or (3) in Chinese, Bengali, and Spanish. When compared to the baseline, model (6) yields significant improvement on Chinese, Arabic, and Hindi (denoted by §). In the opposite direction, gt→en, while the hybrid model (7) significantly outperforms the baseline in

a few cases, it doesn't seem to be consistently more effective than either models (2) or (3). Note that both hybrid approaches are "high resource" since they require Google Translate.

Finding #4: Document translation, model (5), generally beats model transfer, but it requires substantial resources, such as large amounts of parallel text for training a translation system. Because all our documents are in the newswire domain, the output of Google Translate is quite reasonable. Since this approach avoids language mismatch between training and test, it can outperform the model-based transfer approach: these improvements are significant (denoted by ¶) for the Spanish collection on both metrics, and for the Arabic, Bengali, and French collections on AP.

Finding #5: In general, ensembles outperform model transfer alone, with the "high resource" approaches beating the "low resource" approaches (as expected). Comparing the interpolation and RRF methods, we see no consistent trends. A Tukey's HSD test showed no significant differences between the four ensemble methods.

5 Discussion

Given the effectiveness of model transfer, we additionally investigated a research question focused on model (2): How much contextual information does mBERT rely on besides term matching?

Inspired by the query-centric assumption (Wu et al., 2007) that relevance information is localized in the contexts around query terms, we conducted the following experiments: For each query term, we only kept the texts around the matched tokens in each sentence within a window size, and used only those contexts for reranking. We tried window size 1 (only the matched query terms are kept), 3 (the matched query terms with their left and right tokens), 5, 7, 11, and "sentence" (the entire sentence is kept if at least one query token matched). If the segments are from the same sentence, they are concatenated to form a new "sentence".

Experimental results are shown in Figure 1 for two representative collections. For comparison, we also repeat results of the baseline, either model (0) or (1), denoted *bm25* in the figure, and the results of model (2), denoted *full* in the figure. We can see that as the window size increases, AP tends to rise as well. This seems intuitive, as context is needed for relevance matching. Furthermore, results show that some words critical for determining relevance

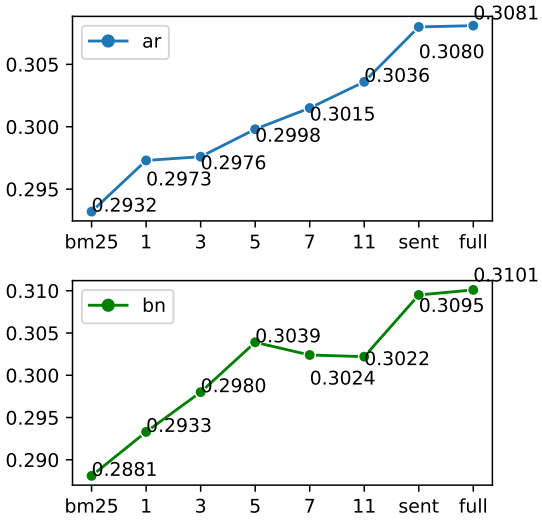

Figure 1: AP results on TREC02-ar and FIRE12-bn.

are located quite far from the query terms; these are discarded when the window size is too small, leading to lower AP scores. However, if we only keep sentences that have at least one query term, the ranking effectiveness is already comparable to using all sentences (0.3080 vs. 0.3081 in Arabic, 0.3095 vs. 0.3101 in Bengali). This simple filter can decrease the inference time needed for ranking 60% to 80% depending on different characteristics of the collections.

6 Conclusion

As a high-level summary, our experiments confirm that mBERT can enable cross-lingual training of document ranking models. However, mBERT's "multi-lingual capacity" for direct model-based transfer does not appear to be consistently better than other approaches of bridging language gaps. For example, simple approaches such as token-based translation of the training data also work well. However, model-based transfer requires only a single model, whereas the latter requires a model for each language. Overall, our work contributes to a better understanding of how relevance judgments in high-resource languages can be leveraged to improve search in low(er)-resources languages. Our code is available on GitHub.[3]

Acknowledgments

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

[3] https://github.com/Impavidity/cross-lingual-doc-ranking

References

Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-domain modeling of sentence-level evidence for document retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3490–3496, Hong Kong, China.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, pages 758–759, Boston, Massachusetts.

Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*, pages 985–988, Paris, France.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Xiaolei Huang, Jonathan May, and Nanyun Peng. 2019. What matters for neural cross-lingual named entity recognition: An empirical analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6395–6401, Hong Kong, China.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2020. PARADE: Passage representation aggregation for document reranking. *arXiv:2008.09093*.

Jimmy Lin, Miles Efron, Yulu Wang, and Garrick Sherman. 2014. Overview of the TREC-2014 Microblog Track. In *Proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014)*, Gaithersburg, Maryland.

Sean MacAvaney, Luca Soldaini, and Nazli Goharian. 2020. Teaching a new dog old tricks: Resurrecting multilingual retrieval using zero-shot learning. In *Proceedings of the 42nd European Conference on Information Retrieval, Part II (ECIR 2020)*, pages 246–254.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. CEDR: Contextualized embeddings for document ranking. In *Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*, pages 1101–1104, Paris, France.

Andreas Rücklé, Jonas Pfeiffer, and Iryna Gurevych. 2020. MultiCQA: Zero-shot transfer of self-supervised text matching models on a massive scale. *arXiv:2010.00980*.

Peng Shi and Jimmy Lin. 2019. Cross-lingual relevance transfer for document retrieval. *arXiv:1911.02989*.

Mark D. Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the Sixteenth International Conference on Information and Knowledge Management (CIKM 2007)*, pages 623–632, Lisbon, Portugal.

Ho Chung Wu, Robert W. P. Luk, Kam-Fai Wong, and K. L. Kwok. 2007. A retrospective study of a hybrid document-context based retrieval model. *Information Processing & Management*, 43(5):1308–1331.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China.

Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: reproducible ranking baselines using Lucene. *Journal of Data and Information Quality*, 10(4):Article 16.