Weakly Supervised Attentional Model for Low Resource Ad-hoc Cross-lingual Information Retrieval

Lingjun Zhao[†], **Rabih Zbib**[†], **Zhuolin Jiang**[†], **Damianos Karakos**[†], **Zhongqiang Huang**[‡*]
[†]Raytheon BBN Technologies, Cambridge, MA, USA
[‡]Alibaba Technologies, Hangzhou, China
{lingjun.zhao, rabih.zbib, zhuolin.jiang, damianos.karakos}@raytheon.com,
z.huang@alibaba-inc.com

Abstract

We propose a weakly supervised neural model for Ad-hoc Cross-lingual Information Retrieval (CLIR) from low-resource languages. Low resource languages often lack relevance annotations for CLIR, and when available the training data usually has limited coverage for possible queries. In this paper, we design a model which does not require relevance annotations, instead it is trained on samples extracted from translation corpora as weak supervision. This model relies on an attention mechanism to learn spans in the foreign sentence that are relevant to the query. We report experiments on two low resource languages: Swahili and Tagalog, trained on less than 100k parallel sentences each. The proposed model achieves 19 MAP points improvement compared to using CNNs for feature extraction, 12 points improvement from machine translation-based CLIR, and up to 6 points improvement compared to probabilistic CLIR models.

1 Introduction and Previous Work

Neural models for Information Retrieval (IR) have received a fair amount of attention in recent years (Zhang et al., 2016; Zamani and Croft, 2017; Dehghani et al., 2017; Mitra et al., 2018). This includes Cross-Lingual Information Retrieval (CLIR), where the task is to retrieve documents in a language different from that of the query. Neural models for CLIR can learn relevance ranking without directly relying on translations, but they typically require large amounts of training data annotated for relevance (cross-lingual query-document pairs), which are often not available, especially for low resource languages. When available, the annotated data usually has limited coverage of the large space of possible ad-hoc queries.

In this paper, we propose a novel neural model for Ad-hoc CLIR against short queries using weak supervision instead of annotated CLIR corpora. The model computes the probability of relevance of each sentence in a foreign document to an input query. These probabilities are then combined to compute a relevance score for a query-document pair. Our model does not rely on relevance-annotated data, but is trained on samples extracted from parallel machine translation data as weak supervision. Compared to CLIR annotated data, sentence translations are often easier to obtain and have better coverage for short queries. The main challenge is designing the model to effectively identify relevant spans in the possibly long foreign sentence. We address that by using an attention mechanism (Bahdanau et al., 2015; Vaswani et al., 2017), thus allowing the model to learn what parts of the sentence to focus on without explicit supervision (e.g. word alignments). To bridge the gap across languages, we pre-train and further optimize bilingual embeddings. We also investigate element-wise interaction between the query and sentence representations to further improve relevance matching.

In contrast, previous methods that directly model CLIR rely on large amounts of relevance-annotated data (Sasaki et al., 2018; Lavrenko et al., 2002; Bai et al., 2009; Sokolov et al., 2013). Other approaches use bilingual embeddings to represent text cross-lingually, but are not specifically optimized for CLIR (Vulic and Moens, 2015; Litschko et al., 2018). (Li and Cheng, 2018) designed a model to learn task-specific text representation using CLIR-annotated data. (Franco-Salvador et al., 2014; Sorg and Cimiano, 2012) crossed the lexical gap using external knowledge sources (Wikipedia), which are limited for low resource languages. An alternative approach translates the queries or documents, and reduces CLIR to monolingual IR (Gupta et al., 2017; Levow et al., 2005; Nie, 2003). But machine translation is not an ideal solution for CLIR either (Zhou et al., 2012), one

[*] Work was done while the author was at Raytheon BBN Technologies.

Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo), pages 259–264
Hong Kong, China, November 3, 2019.
https://doi.org/10.18653/v1/P17

reason is it often produces hallucinated sentences that has little relevance with the source side for low resource languages. On the other hand, (Zbib et al., 2019; Xu and Weischedel, 2000) model the CLIR problem using generative probabilistic model with lexical translation dictionary, while this assumes independence between query words and ignores the underlying semantic connection.

The main contributions of this paper are:

- We design a weakly supervised neural model for CLIR using parallel machine translation data for training, rather than using annotated CLIR corpora.
- To the best of our knowledge, this is the first application of attention mechanisms to CLIR.
- We further propose and demonstrate the importance of an interaction-based relevance matching layer.

We report experiments on two low-resource languages: Swahili and Tagalog, using data from the MATERIAL (MAT, 2017) program. The proposed model obtains scores 19 MAP points higher than neural models that use CNNs for feature extraction. Compared to the machine translation-based CLIR, this model has about 12 MAP points better performance. The model also has better performance than the probabilistic CLIR models with up to 6 MAP points improvement. Additionally, the proposed interaction-based relevance matching layer is usually effective for the QRANN model.

2 Query Relevance Attentional Neural Network Model (QRANN) for CLIR

Direct modeling of CLIR is not practical for low resource languages, as annotated query-document pairs are usually not available. English queries and foreign sentences extracted from parallel translation corpora can serve as weakly supervised training data to learn a model that estimates relevance between short queries and foreign sentences, which can then be applied to computing the query-document relevance scores.

2.1 QRANN Model

Our goal is to design a model that measures the relevance between an English[1] query and a foreign sentence. Formally, given a query $q = [q_1, q_2, ..., q_m]$ and a foreign sentence $s = [s_1, s_2, \ldots s_n]$, the model estimates relevance probability $P(rel|q, s) \in [0, 1]$. We first describe the model architecture, shown in Figure 1, and later explain how this model output is used to compute document-level relevance scores. Each word in q and s is represented as a d-dimensional embedding using lookup table $E(\cdot)$:

$$Q = [Q_1, ..., Q_m] = [E(q_1), ..., E(q_m)]$$
$$S = [S_1, .., S_n] = [E(s_1), ..., E(s_n)] \quad (1)$$

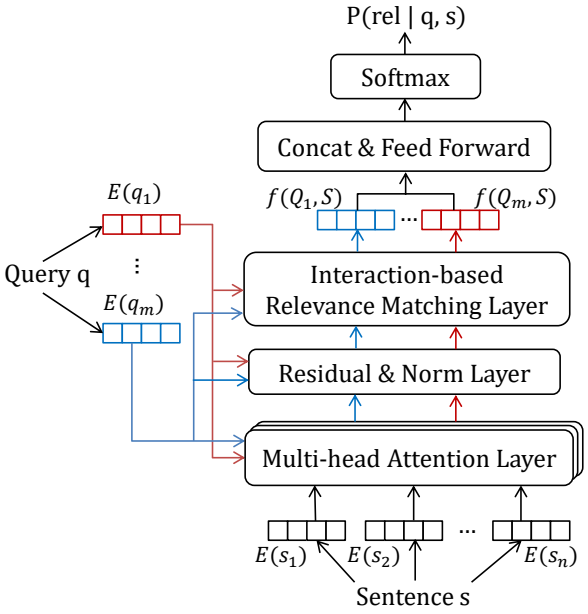

Figure 1: Query relevance attentional neural network (QRANN) model architecture. Each word in the query has an attention mechanism with the sentence to identify relevant spans, followed by a residual connection and layer normalization. After relevance matching, the outputs are fed to a feedforward layer to obtain relevance features of the entire query, which are used for final relevance estimation.

Multi-head Attention Layer A key feature of our approach is to avoid an explicit alignment of the query words with the relevant foreign sentence spans. Instead, we use an attention mechanism over S, which allows the model to learn which spans contain relevance evidence, as well as how to weight that evidence. We compute a context vector C_j for word q_j as a weighted sum of S_i:

$$C_j = \sum_{i=1}^{n} \alpha_i^j S_i \quad (2)$$

the attention weight α_i^j for each word s_i in sen-

[1]The discussion and experiments are in terms of English queries, but the model is language-independent.

tence scores how well the sentence word s_i and q_j match: $\alpha_i^j = softmax(e_i^j)$, where $e_i^j = v_\alpha tanh(U_\alpha Q_j + W_\alpha S_i)$. U_α, W_α and v_α are shared across each q_j in q for better generalization. Inspired by recent work in machine translation (Vaswani et al., 2017), we found it beneficial to use a multi-head attention to allow the model to jointly attend to information at different positions. We compute multi-head attention M_j by concatenating k different attention context vectors and applying a linear transformation:

$$M_j = [C_j^1; C_j^2; ...; C_j^k]W_M \quad (3)$$

Residual & Norm Layer We use a residual connection to the attention outputs, followed by layer normalization: $N_j = LayerNorm(Q_j + M_j))$.

Interaction-based Relevance Matching Layer The essential function of the relevance matching layer is to help the model identify the relatedness or difference between sentence and query. Previous work (Li et al., 2018) shows that element-wise interactions, like difference could capture offset between two words in an embedding space, and inner product could measure their relatedness. Hence, we propose an effective approach for relevance matching by computing both difference and relatedness between Q_j and N_j:

$$e_{N_j,Q_j}^{diff} = N_j - Q_j; e_{N_j,Q_j}^{prod} = N_j \odot Q_j \quad (4)$$

The interaction-based relevance matching layer is a hidden layer on top of the concatenation of e_{N_j,Q_j}^{diff} and e_{N_j,Q_j}^{prod}:

$$f(Q_j, S) = relu(W_C[e_{N_j,Q_j}^{diff}; e_{N_j,Q_j}^{prod}] + b_C) \quad (5)$$

W_C, b_C are shared across each q_j in q.

In order to show the effectiveness of interaction-based relevance matching, instead of using (5), we also try simply concat N_j and Q_j as an alternative relevance matching layer:

$$f(Q_j, S) = relu(W_C[N_j; Q_j] + b_C) \quad (6)$$

Concat & Feed Forward We concatenate the relevance matching outputs and pass them through another hidden layer:

$$g(Q, S) = tanh(W_h[f(Q_1, S); ...; f(Q_m, S)] + b_h) \quad (7)$$

As shown in Figure 1, the feed forward layer concatenates query-word specific features. Each neuron has connections with all query-specific features, aiming to capture semantic relationships among query words.

Softmax Output Finally, the relevance probability between q and s is computed by:

$$P(rel|q, s) = softmax(W_o g(Q, S) + b_o) \quad (8)$$

where W_h, b_h, W_o, b_o are trainable parameters.

2.2 Weakly Supervised Learning for CLIR

As mentioned, the QRANN model is not trained on relevance-annotated data. Instead, it is trained with weakly supervised data. Weak supervision has been studied in monolingual IR. For example, (Dehghani et al., 2017) used BM25 to produce weakly supervised query-document labels. Different from monolingual IR, CLIR requires the training data to bridge the language gap, thus we propose a novel weak supervision used in the QRANN model for CLIR: we construct cross-lingual query-sentence pairs from parallel data as weakly supervised labels to learn cross-lingual query-document relevance. Positive samples are constructed from a foreign sentence and a content word or noun phrase from its English translation. We generate negative samples by selecting a foreign sentence and an English word or phrase that does not appear in the sentence translation. We find using larger negative-to-positive ratio improves the model performance as this would provide more negative samples variety, and fix the ratio to 20:1 for both model performance and training speed. We avoid using stop words for both types of samples.

We use bilingual embeddings pre-trained on the same parallel data for both low resource languages, using the method in (Gouws and Søgaard, 2015), and optimize them further during model training.

The CLIR end task requires an estimation of relevance of the whole foreign document to the query, using relevance outputs between query and sentence from the QRANN model. We experimented with different methods for combining sentence relevance scores, including average and maximum, and found the most effective method to be the probability of relevance to at least one sentence in the document:

$$P(rel|q, D) \approx 1 - \prod_{s \in D} \Big(1 - P(rel|q, s)\Big) \quad (9)$$

3 Experiments

3.1 Datasets and Experimental Setup

We report experimental results on CLIR datasets provided by the MATERIAL (MAT, 2017) program for two low resource languages: Swahili and Tagalog. Each language has two datasets: Test1 (about 800 documents) and Test2 (about 500 documents). We use two query sets: 83 query phrases in Q1 and 102 query phrases in Q2 for Swahili, 140 query phrases in Q1 and 205 in Q2 for Tagalog. The CLIR performance is reported using Mean Average Precision (MAP).

For training, we use parallel sentences released by the MATERIAL and LORELEI (LOR, 2015) programs (72k for Swahili, 98k for Tagalog), and parallel lexicons dowloaded from Panlex (Kamholz et al., 2014) (190k for Swahili, 65k for Tagalog). We extract 40-50M samples from the parallel corpora for each language to train the QRANN model. We use the Adam optimizer with a learning rate of 0.0005, batch size of 512, and dropout probability of 0.1. We pre-train bilingual word embeddings with size $d = 512$, use 4 attention heads, each with a size of 512. The hidden layer sizes are 512 for W_C and 1024 for W_h.

3.2 Baseline Approaches

Probabilistic CLIR Model with Statistical MT Generative probabilistic models (Miller et al., 1999; Xu and Weischedel, 2000) have been an effective approach to CLIR. We use such a model as baseline, with probabilistic lexical translations estimated from statistical machine translation alignment of the parallel training data. We use the concatenation of GIZA++ (Och and Ney, 2003) and the Berkeley Aligner (Haghighi et al., 2009) to estimate lexical translation probabilities by normalizing the alignment counts.

Occurrence Probability Variant We also use a baseline that computes the document relevance score as the probability of each of the query terms occurring at least once in the document. Using translation probabilities $p(q \mid f)$, the document score is computed as:

$$\prod_{q \in Q} \Big[1 - \prod_{f \in Doc} \big(1 - p(q \mid f) \big) \Big] \qquad (10)$$

Machine Translation (MT) based CLIR We compared our model with a MT-based CLIR, which translates foreign documents into English using Transformer (Vaswani et al., 2017), and then does monolingual information retrieval. This approach is similar to (Nie, 2003).

CNN Feature Extraction Convolutional Neural Networks (CNNs) have been found effective for extracting features from text. Here we use this model for feature extraction for comparison. Instead of using multi-head attention and normalization layers, we build a CNN model to extract features from sentence and query, which includes an embedding layer, a convolutional layer with max-pooling and a dropout layer. We use CNN kernel sizes 1 to 5, each with 100 filters. We then pass the extracted features to an interaction-based relevance matching layer, followed by softmax to obtain relevance probability output. This CNN feature extraction for CLIR is similar to (Sasaki et al., 2018).

3.3 Results and Discussion

We compare the performance of different CLIR models on the two low resource languages in Table 1. Comparing the QRANN models with the CNN baseline model, we note that the MAP scores of QRANN models are significantly higher than CNN model in all cases. While the QRANN models do not use CNNs to extract features, they perform better because of the multi-head attention mechanism, which helps the model identify spans in the foreign sentences that are relevant to the query.

We also note that the QRANN model performs better than two strong baselines: the probabilistic CLIR model as well as the probablistic CLIR occurrence variant using translation dictionary. An important feature of the QRANN model is that it jointly represents the tokens of a multi-word query, while probabilistic CLIR models impose a strong independence assumption between query words. For example, query 'New York Times' is treated as independent words, and the translations for each word are used to rank documents independently, which is problematic. The results in table show the benefit in the QRANN model of dropping the query term independence assumption that the probabilistic CLIR model and its occurrence variant use. The QRANN model is designed to model the dependency between words in a multi-word query, in order to capture compositional semantic relationship.

The MT-based CLIR model does not perform well than the QRANN models or the probablistic

Lang	Model	Test1/Q1+2	Test2/Q2
Swa	Prob. CLIR	0.375	0.376
	Prob. Occ.	0.365	0.443
	MT	0.240	0.373
	CNN	0.228	0.217
	QRANN Con.	**0.408**	0.450
	QRANN Int.	0.402	**0.457**
Tag	Prob. CLIR	0.545	0.486
	Prob. Occ.	0.488	0.510
	MT	0.309	0.424
	CNN	0.384	0.359
	QRANN Con.	0.523	0.475
	QRANN Int.	**0.545**	**0.536**

Table 1: Retrieval performance (MAP scores) of all models on Swahili and Tagalog CLIR evaluation datasets. QRANN Con. corresponds to equation (6), QRANN Int. corresponds to equation (5).

CLIR models, because it does not provide enough variation in lexical translations for matching query words to be effective for CLIR.

The same table also compares two variants of QRANN using different relevance matching layers. The interaction-based relevance matching layer usually has better performance than the simple concatenation.

We run statistical significance testing on our results, and found the difference between the QRANN Int. model and the baseline models is statistically significant with p-value less than 0.05 on more than half of the conditions.

4 Conclusion and Future Work

We propose a weakly supervised model to learn cross-lingual query document relevance for low resource languages. Rather than relying on lexical translations, the model uses a multi-head attention mechanism to learn which foreign sentence spans are important for estimating relevance to the query, and also benefits from an effective interaction-based relevance matching layer. Our future work includes using context-dependent pretrained bilingual embeddings, and using high resource languages to improve the CLIR performance of low resource languages.

References

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.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Q. Weinberger. 2009. Learning to rank with (a lot of) word features. *Information Retrieval*, 13:291–314.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural ranking models with weak supervision. In *SIGIR*.

Marc Franco-Salvador, Paolo Rosso, and Roberto Navigli. 2014. A knowledge-based representation for cross-language document retrieval and categorization. In *EACL*.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *HLT-NAACL*.

Parth Gupta, Rafael E. Banchs, and Paolo Rosso. 2017. Continuous space models for clir. *Inf. Process. Manage.*, 53(2):359–370.

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 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 Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 923–931, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Kamholz, Jonathan Pool, and Susan M. Colowick. 2014. Panlex: Building a resource for panlingual lexical translation. In *LREC*.

Victor Lavrenko, Martin Choquette, and W. Bruce Croft. 2002. Cross-lingual relevance models. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 175–182, New York, NY, USA. ACM.

Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. 2005. Dictionary-based techniques for cross-language information retrieval. *Inf. Process. Manage.*, 41:523–547.

Bo Li and Ping Cheng. 2018. Learning neural representation for clir with adversarial framework. In *EMNLP*.

Chenliang Li, Wei Zhou, Feng Ji, Yu Duan, and Haiqing Chen. 2018. A deep relevance model for zero-shot document filtering. In *ACL*.

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

David R. H. Miller, Tim Leek, and Richard M. Schwartz. 1999. A hidden markov model information retrieval system.

Bhaskar Mitra, Nick Craswell, et al. 2018. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126.

Jian-Yun Nie. 2003. Cross-language information retrieval. In *Cross-Language Information Retrieval*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

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)*, pages 458–463. Association for Computational Linguistics.

Artem Sokolov, Laura Jehl, Felix Hieber, and Stefan Riezler. 2013. Boosting cross-language retrieval by learning bilingual phrase associations from relevance rankings. In *EMNLP*.

Philipp Sorg and Philipp Cimiano. 2012. Exploiting wikipedia for cross-lingual and multilingual information retrieval. *Data Knowl. Eng.*, 74:26–45.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Ivan Vulic and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *SIGIR*.

Jinxi Xu and Ralph M. Weischedel. 2000. Cross-lingual information retrieval using hidden markov models. In *EMNLP*.

Hamed Zamani and W. Bruce Croft. 2017. Relevance-based word embedding. In *SIGIR*.

Rabih Zbib, Lingjun Zhao, Damianos Karakos, William Hartmann, Jay DeYoung, Zhongqiang Huang, Zhuolin Jiang, Noah Rivkin, Le Zhang, Richard Schwartz, et al. 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*, pages 645–654. ACM.

Yingjie Zhang, Md. Mustafizur Rahman, Alex Braylan, Brandon Dang, Heng-Lu Chang, Henna Kim, Quinten McNamara, Aaron Angert, Edward Banner, Vivek Khetan, Tyler McDonnell, An Thanh Nguyen, Dan Xu, Byron C. Wallace, and Matthew Lease. 2016. Neural information retrieval: A literature review. *CoRR*, abs/1611.06792.

Dong Zhou, Mark Truran, Tim J. Brailsford, Vincent P. Wade, and Helen Ashman. 2012. Translation techniques in cross-language information retrieval. *ACM Comput. Surv.*, 45:1:1–1:44.