The Cross-Lingual Arabic Information REtrieval (CLAIRE) System

Zhizhong Chen
Brown University
Providence, RI, USA

Carsten Eickhoff
Brown University
Providence, RI, USA

ABSTRACT
Despite advances in neural machine translation, cross-lingual retrieval tasks in which queries and documents live in different natural language spaces remain challenging. Although neural translation models may provide an intuitive approach to tackle the cross-lingual problem, their resource-consuming training and advanced model structures may complicate the overall retrieval pipeline and reduce users engagement. In this paper, we build our end-to-end Cross-Lingual Arabic Information REtrieval (CLAIRE) system based on the cross-lingual word embedding where searchers are assumed to have a passable passive understanding of Arabic and various supporting information in English is provided to aid retrieval experience. The proposed system has three major advantages: (1) The usage of English-Arabic word embedding simplifies the overall pipeline and avoids the potential mistakes caused by machine translation. (2) Our CLAIRE system can incorporate arbitrary word embedding-based neural retrieval models without structural modification. (3) Early empirical results on an Arabic news collection show promising performance.

CCS CONCEPTS
• Information systems → Retrieval models and ranking; Evaluation of retrieval results; Search engine architectures and scalability.

KEYWORDS
Cross-Lingual Retrieval, Arabic Information Retrieval, Fusion, Neural Retrieval Models

1 INTRODUCTION
In the era of big data, information retrieval systems offer a convenient and automatic approach to retrieve target information for users on a large scale. Usually, queries and documents stem from the same language. This single-language assumption stands in direct conflict with the ambition of connecting users with as broad as possible a range of information. To bridge the language gap, some works rely on automatic translation of queries or documents or consider altogether multilingual users to begin with [10, 15, 19, 24]. Many multilingual users and new language learners are not equally proficient in both languages and prefer phrasing queries in the dominant language while being able to passively consume material in the second language.

In this paper, we present late-breaking results of cross-lingual query-by-example retrieval. We assume our users are English speakers with only partial proficiency in Arabic who want to explore an Arabic news wire collection. As a consequence, the majority of the topic information is phrased in English, while a few limited Arabic documents are given as examples to aid model learning and to provide supporting retrieval information. To cater for this situation, we propose an end-to-end retrieval pipeline, Cross-Lingual Arabic Information REtrieval (CLAIRE) system, which 1) pre-selects potentially relevant documents via an exact-match step, 2) re-ranks them using a neural retrieval model and 3) fuses the result lists for each partial query into an overall ranking.

The remainder of this paper is structured as follows: Section 2 concisely discusses related work on cross-lingual information retrieval, query-by-example schemes, multilingual embeddings, and neural machine translation. Section 3 provides the details of the proposed CLAIRE retrieval system. Section 4 describes the experimental setup and Section 5 its results, before Section 6 closes with an outlook on future directions of inquiry.

2 BACKGROUND
Query-by-example (QBE) is a popular retrieval setting that offers more information to specify user’s interest. It has been applied to retrieve documents or entities from the target corpus [25–27], music information retrieval [13] and tuples from relational databases [9]. In many real-world applications, the given query and the target corpus may be written in different languages. Cross-lingual information retrieval (CLIR) is essential to bridge the language gap and to ensure reliable ranking performance on this challenging task. There are two major trends: (1) translation-based approach and (2) modelling based on cross-lingual word embedding.

For translation-based approach, Douglas Oard et al. explore methods of translating documents into the query language [19]. And Jianfeng Gao et al. choose to translate queries for efficiency [10]. Also, there is a so-called interlingual approach by mapping both queries and documents into a third common language [15, 24]. Due to the effective deep representations learned from neural networks, deep learning has revolutionized the domain of machine translation significantly. Some powerful models are LSTM-Autoencoder-based Google’s Neural Machine Translation system (GNMT) [30] and BERT-based translation models [6, 32]. However, training translation models can be expensive and time-consuming. Also, how to transfer pre-trained translation models to the new dataset properly and how to incorporate it into the information retrieval system effectively still remain challenging open tasks.
The proposed cross-lingual query by example retrieval system, CLAIRE, comprises three major sub-systems: (1) BM25 pre-selection, (2) neural re-ranking, and (3) fusion system. The overall workflow is illustrated in Figure 2.

4 EXPERIMENTS

We conduct our retrieval experiments on a proprietary collection of 730k Arabic news articles covering a broad range of subjects. We construct 35 search topics as input to our retrieval pipeline. On average, each topic identifies 2.5 known relevant input example documents.

4.1 BM25 Pre-selection

The number of documents pre-selected by BM25 provides a critical foundation for our overall CLAIRE system. Too small thresholds can ignore some reasonably relevant candidates due to the simplicity of BM25. If the threshold is too large, many non-relevant documents may introduce noise or slow down the overall pipeline. Hence, we empirically investigate thresholds of 1,000, 5,000, 10,000 most highly ranked BM25 documents to be included in neural re-ranking.
4.2 Neural Retrieval Re-ranking

Our system trains neural retrieval models via the ListNet ranking loss [2] due to its superior ranking performance, fast convergence and scalable reliability. As a listwise approach, ListNet compares a list of candidates simultaneously and extracts the inter-document ranking signals to feed neural retrieval models. It first applies the classic softmax function to transform both candidate prediction list and relevance label list into two corresponding probability distributions, then utilises cross entropy loss to finalize optimization signals. Its simple and effective formulation often outperforms other listwise ranking schemes, such as ListMLE [31] and ApproxNDCG [21]. We choose the list size of ListNet as 50 concurrently evaluated documents. For each BM25 pre-selected candidate pool, we treat non-example documents as non-relevant and construct all possible candidate lists by randomly selecting non-example documents. Each candidate list contains only a single example document as the relevant candidate.

We consider three neural retrieval models: (1) KNRM [3], a kernel based neural retrieval model that uses Gaussian kernel pooling techniques on the word-level query-document similarity matrix to extract multi-level soft-match features. (2) ConvKNRM [33], the convolutional kernel-based extension of KNRM via Convolutional Neural Networks to compose adjacent words’ n-gram embeddings and (3) MatchPyramid [207] that regards the constructed word-level similarity matrix from query and document embeddings as a special ‘image’ and utilises several stacked CNN layers to extract matching signals for the final ranking score computation.

4.3 Model Implementation

All neural retrieval models are implemented in Tensorflow 1. ArabEngVec [16], the cross-lingual word embedding, is used for vocabulary initialization. We use an Adam optimizer to train all models for a fixed 21 epochs and save models every three epochs. Then models with the best performance are selected. The learning rate is set to 0.0001. The maximum length of queries and documents are set as 150 and 400, respectively. Regarding model-specific parameters, we apply 11 Gaussian kernels with one harvesting exact matches (mean value $\mu_0 = 1.0$ and standard deviation $\sigma_0 = 10^{-3}$) and ten capturing soft matches (mean values spaced evenly from -0.9 to 0.9, and standard deviation $\sigma_1 = 0.1$) for both KNRM and ConvKNRM. The number of CNN filters in ConvKNRM is set to 128. For MatchPyramid, we choose 3 convolution layers, each with kernel size $3 \times 3$ and 16 convolutional channels.

5 RESULTS

For each BM25 pre-selection threshold and neural retrieval model, we further translate the English query parts, (1) topic title, (2) topic background and (3) event/knowledge descriptions, to Arabic directly using the simple word-level English-Arabic lookup table for our neural retrieval re-ranking system as the baseline. During model evaluation, each topic query on average has around 50 relevant documents identified by a large pool of in-house expert annotators. We evaluate and compare different system settings in terms of Precision@10, Recall@10 and Normalized Discounted Cumulative Gain at the threshold of 10 (nDCG@10). Test performance is shown in Table 1.

---

1 https://www.tensorflow.org/
BM25 pre-selection provides a solid foundation for the overall CLAIRE system. Due to the range of number of relevant documents for each query, a relatively low BM25 threshold of 1,000 pre-retrieved documents yields most reliable results. When there are more unnecessary documents pre-selected from BM25, the ranking performance of neural retrieval models and the final fusion results deteriorates gradually. This is especially significant for MatchPyramid model. Compared to the fully translated systems, the direct retrieval from English query components usually produces better performance in most cases, which shows the effectiveness and robustness of the joint cross-lingual word embeddings. Moreover, KNRM models provide competitive ranking performance, even better than the more sophisticated ConvKNRM in terms of nDCG@10. Since there are only 35 topics and only two or three relevant document examples for model training, simpler models can be a reasonable choice due to their weaker dependency on an abundance of training information. Furthermore, the RRF-based rank fusion improves the overall performance of the system by integrating the neural retrieval models and the final fusion results.

5.1 Topic Component Analysis

To further analyse the relative importance of each topic component, we run our system for different neural retrieval models on each component individually. We again choose 1,000 as the BM25 threshold. The test performance of the final fusion is shown in Table 2. Across different neural retrieval models, the general relative importance among the four major components is consistent and clear. The Arabic example documents play an essential role and provide reliable ranking performance. Although there are no specific retrieval targets indicated in the example documents, our system can still exploit the main ideas and perform query by example retrieval effectively. On the other hand, the topic title shows relatively weaker contribution due to its limited content. This is especially difficult when there are limited amounts of data for model learning. Topic background and descriptions yield reasonable performance according to their content size. It is worth noting that even though the full fusion system performs on par with example document components alone in terms of Precision@10 and Recall@10, its nDCG@10 is significantly better.
Table 3: Test Performance of Different Fusion Methods.

| Fusion Type          | BM25 Pre-Selection | Neural Retrieval | Final Fusion |
|----------------------|--------------------|------------------|--------------|
|                      | Precision@10       | Recall@10        | nDCG@10      |
|                      |                    |                  |              |
| KNRM_RRF             | 0.66000            | 0.16332          | 0.52567      |
| KNRM_CombSUM         | 0.61714            | 0.14313          | 0.47995      |
| KNRM_CombMNZ         | 0.58571            | 0.13492          | 0.46652      |
| KNRM_ISR             | 0.51429            | 0.13028          | 0.44940      |
| ConvKNRM_RRF        | 0.66000            | 0.16332          | 0.52567      |
| ConvKNRM_CombSUM    | 0.61714            | 0.14313          | 0.47995      |
| ConvKNRM_CombMNZ    | 0.58571            | 0.13492          | 0.46652      |
| ConvKNRM_ISR        | 0.51429            | 0.13028          | 0.44940      |
| MatchPyramid_RRF    | 0.66000            | 0.16332          | 0.52567      |
| MatchPyramid_CombSUM| 0.61714            | 0.14313          | 0.47995      |
| MatchPyramid_CombMNZ| 0.58571            | 0.13492          | 0.46652      |
| MatchPyramid_ISR     | 0.51429            | 0.13028          | 0.44940      |

5.2 Effect of the Fusion System

Besides Reciprocal Rank Fusion (RRF) [4] there are various other unsupervised rank fusion algorithms, such as CombSUM [28], CombMNZ [28] and Inverse Square Rank (ISR) [18]. Table 3 explores the effectiveness of such alternative fusion schemes. RRF-based systems outperform other fusion methods significantly across three neural retrieval models and different sub-systems. Although the difference between RRF and CombSUM is relatively small for BM25 pre-selection, CombSUM yields much worse results on neural re-ranking systems and the final fusion pipeline. In general, simple methods such as RRF and CombSUM perform better than CombMNZ and the more involved ISR. Due to the unsupervised nature of these rank fusion methods, simpler methods with fewer a priori assumptions are more robust and generalizable to diverse scenarios.

6 CONCLUSION

This chapter describes our late-breaking investigation results of cross-lingual query by example retrieval models. Our end-to-end Cross-Lingual Arabic Information REtrieval (CLAIRE) system yields promising performance without incorporating the complicated neural translation models into our neural retrieval system. To simulate how non-Arabic speakers conduct Arabic retrieval tasks in realistic scenarios, we design 35 structural topic queries to reflect various knowledge information users obtained from their experience. We use three fundamental retrieval steps: (1) BM25 pre-selection, (2) neural re-ranking, and (3) retrieval fusion. It is interesting to observe that the simpler KNRM model provides very competitive performance under the circumstances of the limited training information, compared to ConvKNRM and MatchPyramid with higher complexities. Moreover, we present the effect of different BM25 thresholds and various fusion methods. As a classic retrieval approach, BM25 provides a solid foundation of our CLAIRE system. Compared to other rank fusion methods, Reciprocal Rank Fusion (RRF) is robust and generalizable to different retrieval systems. Its reliable performance further boosts our CLAIRE system. For future work, we would like to explore more general approaches to build multi-lingual retrieval systems to aid users in exploring various languages simultaneously.

ACKNOWLEDGMENTS

Acknowledgements blinded for anonymous review.
