A Study of Neural Matching Models for Cross-lingual IR

Puxuan Yu and James Allan
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
{pxyu,allan}@cs.umass.edu

ABSTRACT

In this study, we investigate interaction-based neural matching models for ad-hoc cross-lingual information retrieval (CLIR) using cross-lingual word embeddings (CLEs). With experiments conducted on the CLEF collection over four language pairs, we evaluate and provide insight into different neural model architectures, different ways to represent query-document interactions, word-pair similarity distribution and the vocabulary mismatch problem in CLIR. This study paves the way for learning an end-to-end CLIR system using CLEs.

KEYWORDS

Cross-lingual Information Retrieval; Cross-lingual Word Embeddings; Neural Information Retrieval Model

ACM Reference Format:
Puxuan Yu and James Allan. 2020. A Study of Neural Matching Models for Cross-lingual IR. In *Proceedings of the 43nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 21–25, 2020, Xi'an, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

CLIR is the task of retrieving documents in target language L_t with queries written in source language L_s. The increasing popularity of projection-based weakly-supervised [4, 6, 16] and unsupervised [1, 2] cross-lingual word embeddings has spurred unsupervised frameworks [8] for CLIR, while in the realm of mono-lingual IR, interaction-based neural matching models [5, 9, 11, 17] that utilize semantics contained in word embeddings have been the dominant force. This study fills the gap of utilizing CLEs in neural IR models for CLIR.

Traditional CLIR approaches translate either document or query using off-the-shelf SMT system such that query and document are in the same language. Later on, a lot of literature [13–15] investigates utilizing translation table to build probabilistic structured query [3] in target language. Recently, Litschko et al. show that CLEs are good translation resources by experimenting an CLIR method (dubbed TbT-QT) that translates each query term in source language to the nearest target language term in the CLE space [8]. CLEs are obtained by aligning two separately trained embeddings for two languages in the same latent space, where a term in L_s is proximate to its synonyms in L_s and its translations in L_t, and vice versa. TbT-QT takes only the top-1 translation of a query term and uses the query likelihood model [12] for retrieval. The overall retrieval performance can be damaged by vocabulary mismatch magnified with translation error. Using closeness measurement between query and document terms in the shared CLE space as matching signal for relevance can alleviate the problem, but this area has not been extensively studied.

The reasons for the success of neural IR models for mono-lingual retrieval can be grouped into two categories:

- **Pattern learning**: the construction of word-level query-document interaction representations enables learning of various matching patterns (e.g., proximity, paragraph match, exact match, semantic match) via different neural network architectures.
- **Representation learning**: models in which interactions are built with differentiable operations (e.g., kernel pooling [17]) allow customizing word embeddings via end-to-end learning from large-scale training data.

Although representation learning is capable of further improving overall retrieval performance [17], it was shown in the same study that updating word embeddings requires large-scale training data to work well (more than 100k search sessions in their case). In CLIR, however, datasets are usually in the size of less than 200 queries per available language pair and can only support training neural models with smaller capacity. Therefore, we focus on the *pattern learning* aspect of neural models.

In this study, we formulate the following research questions:

- **RQ1**: how is a neural model for CLIR different from mono-lingual IR?
- **RQ2**: how do neural models compare with each other and with unsupervised models for CLIR?

We answer these two main research questions with analysis (§ 2) and experiments (§ 3) in the rest of the paper.

2 ANALYSIS

2.1 Unsupervised CLIR Methods with CLEs

Two unsupervised CLIR approaches using CLEs are proposed by Litschko et al. [8]. **BWE-Agg** ranks documents with respect to a query using the cosine similarity of query and document embeddings, obtained by aggregating the CLEs of their constituent terms. The simpler version, namely BWE-Agg-Add, takes the average embeddings of all terms for queries and documents, while the more advanced version BWE-Agg-IDF builds document embeddings by weighting terms with their inverse document frequencies. **TbT-QT**, as described in § 1, first translates each query term to its nearest cross-lingual neighbor term and then adopts mono-lingual retrieval.

SIGIR '20, July 21–25, 2020, Xi'an, China

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

These two approaches represent different perspectives towards CLIR using CLEs. BWE-Agg builds query and document representations out of CLEs but completely neglects exact matching signals, which play important roles in IR. Also, although query and document terms are weighted based on IDF, using only one representation for a long document can fail the emphasize the truly relevant section to the query. TbT-QT only uses CLEs as query translation resources and adopts exact matching in mono-lingual setting, therefore its performance is heavily dependent on the translation accuracy (precision@1) of CLEs. Analytically, an interaction-based neural matching model that starts with word level query-document interactions and considers both exact and similar matching can make up for the shortcomings of the above two methods.

2.2 Neural IR Models

2.2.1 Background. For interaction-based matching models, we select three representative models (MatchPyramid [10, 11], DRMM [5] and KNRM [17]) from the literature for analysis and experiments.

MatchPyramid: The MatchPyramid [10, 11] (MP for short) is one of the earliest models that start with capturing word-level matching patterns for retrieval. It casts the ad-hoc retrieval task as a series of image recognition problems, where the "image" is the matching matrix of a query-document pair (q, d), and each "pixel" is the interaction value of a query term q_i and a document term d_j. Typical interaction functions are cosine similarity, dot product, Gaussian kernel, and indicator function (for exact match). The intuition behind hierarchical convolutions and pooling is to model phrase, sentence and even paragraph level matching patterns.

DRMM: The DRMM [5] model uses a matching histogram to capture the interactions of a query term with the whole document. The valid interval of cosine similarity (i.e., $[-1, 1]$) is discretized into a fixed number of bins such that a matching histogram is essentially a fixed-length integer vector. Features from different histograms are weighted based on attention calculated on query terms. DRMM is not position-preserving, as the authors claim that relevance matching is not position related.

K-NRM: The KNRM [17] model takes matrix representation for query-document interaction (similar to MP), but "categorizes" interactions into different levels of cosine similarities (similar to DRMM), using Gaussian kernels with different mean value μ. The distinct advantage of KNRM over DRMM is that the former allows gradient to pass through Gaussian kernels, and therefore supports end-to-end embeddings learning.

2.2.2 Mono-lingual to Cross-lingual. According to results reported in respective studies [5, 11, 17], the relative performance of three models for mono-lingual IR should be KNRM > DRMM > MP, even when embedding learning is turned off with KNRM. Tweaking a neural model for support of CLIR is trivial: instead of considering interaction value as two terms' similarity in a mono-lingual embedded space, we consider the proximity of their representations in the shared cross-lingual embedded space. However, there are several matters to consider while making the transition:

Exact matching signals: The significant difference between cross-lingual and mono-lingual IR is that the former (almost) never encounters exact match of terms in different languages. However, neglecting such factors can be costly for models like MP, the disadvantage of which when compared to the other two models is the inability to capture exact and similarity matching signals at the same time. To this end, we first define in CLIR the exact matching of two terms (in different languages) as their cosine similarity in the CLE space exceeding a certain threshold value η. We then implement a hybrid version, namely MP-Hybrid, that joins exact and similar matching signals extracted from interaction matrices built with indicator function and cosine similarity function such that ranking features from dual channels are concatenated for an MLP to predict a ranking score.

Table 1: Cosine similarities of the top-5 closest words to "telephone" in an English embedding space (EN) and in an aligned English-Spanish embedding space (ES).

EN	phone	telephones	Telephone	landline	rotary-dial
	0.818	0.761	0.720	0.694	0.669
ES	telefónicos	teléfono	telefónica	telefónia	telefóno
	0.535	0.522	0.522	0.520	0.520

Word-pair similarity distribution: The cosine similarities of two terms with close meanings but in different languages are distributed differently than those in the same language. Specifically, top word-pair similarity distribution of CLEs tends to have smaller mean and variance. In an example shown in Table 1, the cosine similarity of the five closest words to "telephone" in English embedded space[1] ranges from 0.818 to 0.669, while in aligned English-Spanish embedded space[2], it ranges from 0.535 to 0.520. The similarity distribution affects histogram construction of DRMM and similarly for the kernel pooling of KNRM. The distribution also affects the exact matching threshold value η for related variants of MP. Since the cosine similarity of a query term and its perfectly correct translation can be less than 0.6, setting η too high can lead to failure of capturing positive matching signals.

Vocabulary mismatch and translation error: Query translation based CLIR methods (e.g., TbT-QT [8]) first translate queries from L_s to L_t, then does mono-lingual retrieval in L_t. Apart from the inherent vocabulary mismatch problem within L_t, the translation error from L_s to L_t has to be also counted. Looking at the example in Table 1, TbT-QT would look for occurrence of "telefónicos" in the collection, and documents containing only the correct translation ("teléfono") would be overlooked. Interaction-based neural matching models alleviate this issue by giving partial credits to sub-optimal nearest neighbors, which in many cases are the correct translations. To demonstrate the necessity of directly using cross-lingual word embedding similarity as interaction for neural models, we conduct comparative experiments where queries are first translated term-by-term like TbT-QT using CLEs, then used for retrieval in mono-lingual setting. Such models are referred to as {MP,DRMM,K-NRM}-TbT-QT, respectively.

3 EXPERIMENTS

Datasets: We evaluate the models on the CLEF test suite for the CLEF 2000-2003 campaigns. We select four language pairs: English (EN) queries to {Dutch (NL), Italian (IT), Finnish (FI), Spanish (ES)} documents. All documents for the four languages are used

[1] https://dl.fbaipublicfiles.com/fasttext/vectors-english/wiki-news-300d-1M.vec.zip
[2] https://dl.fbaipublicfiles.com/fasttext/vectors-aligned/wiki.es.align.vec

Table 2: Basic statistics of CLEF data for evaluation: number of queries (#queries), number of documents (#docs), average number of relevant documents per query (#rel), and average number of labeled documents per query (#label).

Lang. Pair	EN→ NL	EN→ IT	EN→ FI	EN→ ES
#queries	160	160	90	160
#docs	42,734	40,320	16,351	46,540
#rel	29.1	19.5	10.9	49.5
#label	375.4	338.3	282.6	372.7

for evaluation, and are truncated to preserve the first 500 tokens for computational efficiency [10]. The statistics of the evaluation datasets are shown in Table 2. The titles of CLEF topics are used as English queries. All queries and documents are lower-cased, with stopwords, punctuation marks and one-character token removed.

Cross-lingual word embeddings: We adopt the pre-aligned fastText CLEs[3]. Mono-lingual fastText embeddings are trained on Wikipedia corpus in respective languages, and aligned using weak supervision from a small bilingual lexicon with the RCSLS loss as the optimization objective [6].

Model specifications: We implemented two CLEs based unsupervised CLIR algorithms BWE-Agg and TbT-QT as baselines [8]. In addition to the query likelihood model in the original study, we pair TbT-QT with BM25 to investigate the influence of retrieval models to queries translated using CLEs.

We experiment with five variants of the MP model, two for the DRMM model and two for the KNRM model. As the interaction value of query term q_i and document term d_j, {MP,DRMM,KNRM}-Cosine uses the cosine similarity $\cos(q_i, d_j) = \vec{q_i}^\top \vec{d_j}/(||\vec{q_i}|| \cdot ||\vec{d_j}||)$, MP-Gaussian uses $e^{-||\vec{q_i}-\vec{d_j}||^2}$, and MP-Exact takes $\mathbb{1}_{\{\cos(q_i,d_j)\geq\eta\}}$, where η is a pre-defined threshold value (set to .3 for Table 3). MP-Hybrid concatenates the flattened features after dynamic pooling layer from MP-Cosine and MP-Exact into one vector, and uses an MLP to predict a final score. {MP,DRMM,KNRM}-TbT-QT is equal to first translating query q to target language query $\text{tr}(q)$, and running $\text{tr}(q)$ with {MP,DRMM,KNRM}-Cosine model.

For the MP model, we adopt one layer convolution with kernel size set to 3×3, dynamic pooling size set to 5×1, and kernel count set to 64. For the DRMM model, we adopt the log-count-based histogram (applying logarithm over the count value in each bin) with bin size set to 30. For the KNRM model, kernel count is set to 20 and sigma (standard deviation) of each Gaussian kernel is set to 0.1. All decisions made above are based on extensive hyper-parameter tuning that first prioritizes generalizable retrieval performance then computational efficiency and model simplicity.

Model training: All neural models in the experiments are trained with the pairwise hinge loss. Given a triple $(q, d+, d-)$, where document $d+$ is relevant and document $d-$ is irrelevant with respect to query q, the loss function is defined as:

$$L(q, d+, d-; \Theta) = \max\{0, 1 - s(q, d+) + s(q, d-)\}$$

where $s(q, d)$ denotes the predicted matching score for (q, d), and Θ represents the learnable parameters in the neural network. Note that we randomly select documents that are explicitly labeled irrelevant (-1) as negative samples for training. Five negative (q, d) pair are sampled for each positive pair. We apply stochastic gradient descent method Adam [7] (learning rate=1e-3) in mini-batches (64 in size) for optimization. Maximum number of training epochs allowed is 50.

[3]https://fasttext.cc/docs/en/aligned-vectors.html

Table 3: MAP performance of all CLIR methods. Boldfaced is the best performer in each language pair. Underlined is the best MP variant.

Lang. Pair	EN→ NL	EN→ IT	EN→ FI	EN→ ES
BWE-Agg-Add	.237	.173	.170	.297
BWE-Agg-IDF	.246	.178	.180	.298
TbT-QT-BM25	.240	.231	.122	.341
TbT-QT-QL	.297	.268	.126	.387
MP-Cosine	<u>.348</u>	<u>.331</u>	<u>.254</u>	.423
MP-Gaussian	.322	.319	.203	.405
MP-Exact	.327	.295	.202	.415
MP-Hybrid	.343	.326	.243	<u>.427</u>
MP-TbT-QT	.327	.300	.195	.409
DRMM-Cosine	**.374**	**.352**	**.304**	**.462**
DRMM-TbT-QT	.345	.324	.193	.450
KNRM-Cosine	.368	.313	.286	.423
KNRM-TbT-QT	.329	.288	.200	.405

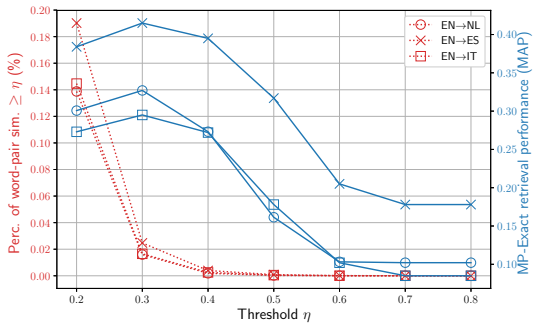
Metric: As the CLEF dataset uses binary relevance judgement, we adopt mean average precision (MAP) as the evaluation metric.

Cross-validation for neural models: In order to conduct evaluation on more queries such that drawn conclusions are statistically more significant, we adopt 5-fold cross-validation with validation and test set. For each language pair, the collection is split into 5 sets based on queries. In each run, one set is selected as test set, one as validation set, and the other three as training set. The recorded performance on test set is generated by model when MAP on validation set is the highest. By rotating the sets after each run, five runs generate evaluation results on all queries in the dataset.

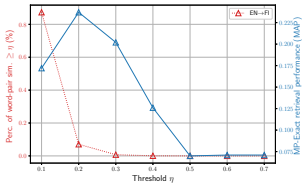
4 DISCUSSION AND CONCLUSION

4.1 Parsing Results

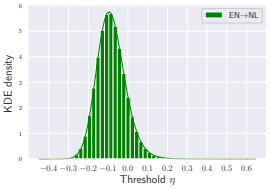
The experimental results of CLIR on four language pairs are reported in Table 3. TbT-QT generally works better than BWE-Agg except for EN→FI. This might indicate that the English-Finnish CLEs are not aligned well to provide quality top-1 query term translation. The larger gaps between {MP,DRMM,K-NRM}-Cosine and {MP,DRMM,K-NRM}-TbT-QT for EN-FI than the other three language pairs reinforce this argument. All neural models achieve statistically significant improvement over heuristic baselines. DRMM-Cosine consistently achieves the best performance for all language pairs. Although DRMM and KNRM are conceptually similar, the former performs significantly better, with KNRM's embedding layer kept frozen. The attention mechanism applied to query terms for DRMM can be a factor. On EN→{IT,ES}, the MP model performs on par with or better than KNRM. This finding indicates that the convolution plus dynamic pooling architecture can also be an option for learning an end-to-end CLIR model. Comparing different approaches to build query-document interaction matrices for MP,



(a) EN→{NL,ES,IT}



(b) EN→FI



(c) EN→NL

Figure 1: (a,b) – Red: percentage of cross-lingual word pair with similarity ≥ η; Blue: MP-Exact retrieval performance with different similarity threshold value η. (c): Similarity distribution of word-pairs in the EN→NL collection.

it is clear that cosine similarity of source language query term and target language document term in the CLE space is the best choice, which contradicts the conclusions in the study of mono-lingual IR [10] where Gaussian kernel and indicator function are found to work better. The exact matching variant MP-Exact we proposed works reasonably well, indicating that most decisions of relevance are influenced by top similarity matching signals. The hybrid variant MP-Hybrid we propose improves upon MP-Exact but doe not outperform MP-Cosine (except for EN→ES). This is expected because matching signals from MP-Exact are not from truly exact matches of terms, but are derived from cosine similarity matrices as in MP-Cosine. Combination of two models results in redundant information. The fact that {MP,DRMM,K-NRM}-TbT-QT outperform baseline approaches but are not as good as respective cosine variants demonstrates (1) the effectiveness of pattern learning of neural models; and (2) the necessity to directly build cross-lingual interactions of query and document in two languages, rather than building interactions after translation.

4.2 Word-pair Similarity Distribution with CLE

The distribution of word pair similarities influences exact matching threshold η in MP-Exact, query translation strategy in TbT-QT, and embedding fine-tuning for an end-to-end model. We take source language terms in the queries and target language terms in the documents, calculate their pairwise cosine similarities in the aligned CLE space, and plot the similarity distributions. In Figure 1a and 1b, we show in red the percentage of cross-lingual word-pairs with similarity above η. Three distributions in Figure 1a are very similar at tail ($\eta \geq 0.2$), therefore the corresponding MP-Exact's performance peaks at the same $\eta = 0.3$. EN→FI is distributed differently but the pattern shown is similar (Figure 1b). The shapes of cross-lingual similarity distribution for all four language pairs are very similar, therefore we only plot EN→NL in Figure 1c for demonstration. Mono-lingual similarity distribution in Xiong et al.'s study [17] has large variance, positive mean, strong positive skewness and high density at large η. In comparison, the cross-lingual similarity distribution (Figure 1c) has small variance, negative mean, no obvious skewness to the left or right, and the density drops low and flat after $\eta = 0.4$, where word-pairs are considered highly similar (i.e., quality translations). This provides insights into why top-1 translation with CLEs is not necessarily significantly better than translations ranked at slightly lower positions.

4.3 Conclusions

Answer to RQ1: To adapt a neural model for CLIR, we first have to consider three factors: exact matching representations, cross-lingual word-pair similarity distribution, and translation error using CLEs. In specific model settings, choices of interaction representations and hyper-parameters (e.g., dynamic pooling size at document side for MP) are found to be different from mono-lingual IR.

Answer to RQ2: Neural matching models experimented in this study all outperform baselines using CLEs. The DRMM achieves the best results across the board, while MP and KNRM perform inconsistently on different language pairs.

Moving forward, a worthwhile endeavor will be to investigate an end-to-end neural model that learns from large-scale CLIR data. How to keep two embedded spaces aligned during embedding updates will be an interesting question.

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Air Force Research Laboratory (AFRL) and IARPA under contract #FA8650-17-C-9118 under subcontract #14775 from Raytheon BBN Technologies Corporation. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. *arXiv preprint arXiv:1809.01272* (2018).
[2] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087* (2017).
[3] Kareem Darwish and Douglas W Oard. 2003. Probabilistic structured query methods. In *SIGIR*. ACM, 338–344.
[4] Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. In *ACL*. 710–721.
[5] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *CIKM*. 55–64.
[6] Armand Joulin, Piotr Bojanowski, Tomáš Mikolov, Hervé Jégou, and Édouard Grave. 2018. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *EMNLP*. 2979–2984.
[7] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[8] Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised cross-lingual information retrieval using monolingual data only. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1253–1256.
[9] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations of text for web search. In *WWW*. 1291–1299.

[10] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2016. A Study of MatchPyramid Models on Ad-hoc Retrieval. CoRR abs/1606.04648 (2016). hp. *arxiv. org/abs/1606.04648* (2016).
[11] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text matching as image recognition. In *AAAI*.
[12] Jay M Ponte and W Bruce Croft. 1998. A language modeling approach to information retrieval. In *SIGIR*. 275–281.
[13] Ferhan Ture and Jimmy Lin. 2013. Flat vs. hierarchical phrase-based translation models for cross-language information retrieval. In *SIGIR*. ACM, 813–816.
[14] Ferhan Ture and Jimmy Lin. 2014. Exploiting representations from statistical machine translation for cross-language information retrieval. *ACM Transactions on Information Systems (TOIS)* 32, 4 (2014), 19.
[15] Ferhan Ture, Jimmy Lin, and Douglas W Oard. 2012. Looking inside the box: Context-sensitive translation for cross-language information retrieval. In *SIGIR*. 1105–1106.
[16] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *NAACL-HLT*. 1006–1011.
[17] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-end neural ad-hoc ranking with kernel pooling. In *SIGIR*. 55–64.