Mixed Attention Transformer for Leveraging Word-Level Knowledge to Neural Cross-Lingual Information Retrieval

Zhiqi Huang, Hamed Bonab, Sheikh Muhammad Sarwar, Razieh Rahimi, and James Allan
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
{zhiqihuang,bonab,smsarwar,rahimi,allan}@cs.umass.edu

ABSTRACT

Pre-trained contextualized representations offer great success for many downstream tasks, including document ranking. The multilingual versions of such pre-trained representations provide a possibility of jointly learning many languages with the same model. Although it is expected to gain big with such joint training, in the case of cross-lingual information retrieval (CLIR), the models under a multilingual setting are not achieving the same level of performance as those under a monolingual setting. We hypothesize that the performance drop is due to the *translation gap* between query and documents. In the monolingual retrieval task, because of the same lexical inputs, it is easier for model to identify the query terms that occurred in documents. However, in the multilingual pre-trained models that the words in different languages are projected into the same hyperspace, the model tends to "translate" query terms into related terms – i.e., terms that appear in a similar context – in addition to or sometimes rather than synonyms in the target language. This property is creating difficulties for the model to connect terms that co-occur in both query and document. To address this issue, we propose a novel Mixed Attention Transformer (MAT) that incorporates external word-level knowledge, such as a dictionary or translation table. We design a sandwich-like architecture to embed MAT into the recent transformer-based deep neural models. By encoding the translation knowledge into an attention matrix, the model with MAT is able to focus on the mutually translated words in the input sequence. Experimental results demonstrate the effectiveness of the external knowledge and the significant improvement of MAT-embedded neural reranking model on CLIR task.

CCS CONCEPTS

• **Information systems → Information retrieval**; **Multilingual and cross-lingual retrieval**; *Retrieval models and ranking*.

KEYWORDS

Cross-lingual information retrieval; Attention mechanism; Neural network

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482452

ACM Reference Format:
Zhiqi Huang, Hamed Bonab, Sheikh Muhammad Sarwar, Razieh Rahimi, and James Allan. 2021. Mixed Attention Transformer for Leveraging Word-Level Knowledge to Neural Cross-Lingual Information Retrieval. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3459637.3482452

1 INTRODUCTION

We study the problem of Cross-Lingual Information Retrieval (CLIR) in which the desired information is written in a language different than that of the user's query. From the modeling perspective, in the CLIR setting some form of language translation is needed to map the vocabulary of the query language to that of the documents' language in addition to the ranking component. This translation gap can be bridged with simple dictionaries, translation tables, machine translation, or more recently cross-language distributional representations [2, 37, 45].

Embedding the translation component in the fine-tuning stage along with the ranking makes the training of deep neural models for the CLIR more challenging, particularly when dealing with resource-lean languages [1, 23]. Pre-trained language models such as BERT [12] have shown promising performance gains for monolingual information retrieval [15, 34, 46, 49]. This success is mainly due to the unsupervised pre-training of context-aware transformer architectures with an enormous number of parameters over large corpora. To achieve success in the learning-to-rank task such models are often fine-tuned with a relatively large collection of relevance judgments such as the MS MARCO passage ranking dataset [25]. However, it is not feasible to obtain data in the scale of MS MARCO across different languages. Thus, some studies leverage different training data (e.g., weak-supervised data, cross-lingual Wikipedia-based data [38, 39]) and techniques (e.g., domain adaptation, few-shot learning) in order to adapt the model for the target task and language, reporting improvements.

The multilingual versions of pre-trained Transformer-based language models, such as mBERT [12] and XLM-R [7], provide the possibility of jointly learning representations for multiple languages with the same model. Fine-tuning these pre-trained multilingual language models for ranking, similar to the monolingual setting, enables cross-language information retrieval. In the multilingual pre-trained models that words in different languages are projected into the same hyperspace, the model tends to map query terms into target language's related terms – i.e., terms that appear in a similar context – in addition to or sometimes rather than synonyms [2, 34]. We hypothesize that this phenomena creates difficulties for the

model to connect terms that match between the query and document. It has been shown that the translation gap plays a significant role in the suboptimal success of neural CLIR models and addressing that can significantly boost the retrieval performance [2]. Therefore, the multilingual language models for the CLIR task have not yet achieved the performance gain observed with the use of pre-trained language models for monolingual information retrieval [22, 48]. This can happen in the CLIR task because the vocabulary size is almost doubled, the possibility of exact match between query and document is limited, and training data (e.g., bilingual query log or click data) is scarce. Most of the existing CLIR systems are thus deployed along with a query translation component to reduce the problem into monolingual retrieval. However, it is important to note that having a translation component as a black-box limits the retrieval component due to translation errors.

We inject word-level translation knowledge into a model at the time of fine-tuning it with relevance data. More specifically, we leverage the external knowledge in the form of a translation table, which is a look-up table that provides translation probabilities for a pair of words in two different languages. We use the translation table to create an attention matrix and use it in parallel with the Transformer's multi-head attention – both in our training and inference phase – to improve the model's cross-lingual understanding. We refer to our extended component as Mixed Attention Transformer (MAT) and create MART, a sandwich-like architecture to embed MAT into the multilingual BERT (mBERT) model. By encoding the translation knowledge into an attention matrix, we enable the overall architecture to focus on the mutually translated words in the input sequence. Our experiments explore the effectiveness of a variety of external knowledge sources and show the significant gain that we get from MART on CLIR task. MAT is a generalized architecture capable to capture any form of lexical mapping and it can be integrated with any transformer-based architecture.

We performed extensive experiments on ten different language pairs for CLIR training and evaluation, three different resources to obtain the translation knowledge, and different qualities of translations based on available translation resources for language pairs. Our experimental results demonstrate the varied effectiveness of different external knowledge sources and the significant improvement of MAT-embedded neural re-ranking model over strong baselines on the CLIR task. In terms of mean average precision (MAP), our proposed model outperforms the neural baseline by 8% on high-resource languages and 12% on low-resource languages.

The rest of this paper is organized as follows. In Section 2 we provide a review of related works. Section 3 presents our MAT architecture for injecting external translation knowledge directly into model. Section 4 and 5 provides our experimental design and results with discussions and further analyses. We conclude our study in Section 6.

2 RELATED WORK

We first provide a summary of existing CLIR models trained from both word-embedding based representations as well as representations from unsupervised language models based on the transformer architecture. We discuss the importance of the knowledge from sentence-level parallel data and how they enhance the performance of neural retrieval models. Finally, we also elaborate on the transformer-based architectures that incorporate external knowledge for a variety of tasks and compare them to MAT.

2.1 Neural Cross-lingual Representation Spaces

CLIR tackles two sub-tasks: query translation and query-document matching, and neural models are applicable to both the tasks. One approach is to translate the query to the language of the corpus by using a Statistical Machine Translation (SMT) or Neural Machine Translation (NMT) model and then apply a mono-lingual matching model to determine the relevance. While the translate-then-retrieve approach is a popular one, neural bilingual word representations creates the opportunity to skip the translation step. As a result, query-document matching can performed in a shared vector space for two languages, where similar words in two different languages are mapped close to each other. The assumption is Cross-linual Word Embeddings (CLWE) are capable to bridge the translation gap between two languages.

One of the earliest works in this direction is from Vulić and Moens [42], and they proposed a model to learn bilingual word embeddings using a document-aligned comparable data. Once all the words in both languages are represented in a shared space, they computed query and document representations using the compositional distributional semantics model and calculated their matching score based on cosine similarity metric. Litschko et al. [22] used the same matching technique but created the shared space using only monolingual data in two languages. Bonab et al. [2] assessed the effectiveness of several bilingual word embeddings under cosine similarity-based scoring framework for retrieval and found that all the existing word embeddings lack the capacity to translate a source language word into the target language word – they refer to this phenomenon as the *translation gap*. The authors showed that a bilingual word embedding brings similar pair of words in two languages close together, but often keeps the words that are translation of each other far than expected. This is because cross-lingual word embeddings are learned from the contextual information around a word but not from the translation of that word. The authors proposed a smart shuffling approach to include translation knowledge into word embeddings and created a state-of-the-art cross-lingual word embedding for retrieval. While it is clear that translation knowledge brings significance gain in retrieval, there is no study on how to incorporate this knowledge in the modern transformer based query-document matching frameworks.

Unsupervised multilingual language models based on the transformer architecture (also referred to as multilingual transformers) brought a major advancement over the cross-lingual word embeddings. There are two major realization of such models: mBERT (Multilingual BERT) [12] and XLM-R (XLM RoBERTa) [8]. These models offer a shared representation space for a large number of languages and the representation of a token is contextualized based on the other tokens in a sequence. Thus these approaches capture higher-level semantics compared to CLWE and once fine-tuned, they have been shown to be effective across a wide variety of tasks, including CLIR [23, 36, 47]. However, we assume that the *translation*

gap still exists in the multilingual transformers and it is important to inject translation knowledge into such architectures.

2.2 Neural Matching Models for CLIR

Whether we use cross-lingual word embeddings or multilingual transformers for representing query and documents, we need to provide relevance knowledge to these models for effective matching. Thus, we need to further train these language representation spaces using with relevance judgments from human [2, 21, 38, 52].

Sasaki et al. [38] constructed a large-scale weakly supervised CLIR collection by using the first sentence of a Wikipedia page as the query and all the linked foreign language articles as documents. They proposed a shallow learing-to-rank method and did not use a shared language representation space. Thus, their approach does not explicitly close the language gap between the query and document. Zhao et al. [52] leverages the sentence-aligned parallel data to create weakly supervised relevance judgments. They use a sentence from a language as a document and randomly select a word from the translation of that sentence as query. Even though they close the language gap using parallel data, they do not use relevance judgments explicitly. We use both parallel data and relevance judgments and improve the architecture of a multilingual transformer to adapt these sources.

Rather than considering relevance and translation in isolation, Li and Cheng [21] took an adversarial learning approach to jointly learn language alignment through translation knowledge and cross-lingual matching using relevance judgments. They created a weakly-supervised collection of parallel data by translating AOL queries using Google Translate. They use a Long Short Term Memory (LSTM) network to learn matching in contrast to the multilingual transformer proposed in this work. Moreover, they use weak parallel data to close the language gap, whereas we use word-level alignment learned from the parallel data or obtained from a dictionary in the fine-tuning stage. Bonab et al. [2] achieved state-of-the-art performance when they used their translation-oriented bilingual representations with DRMM matching model [13] and trained the architecture using relevance data. They showed that dictionary-oriented word embeddings can improve the performance of a DRMM model when fine-tuned with relevance data. We propose a novel multilingual transformer architecture, MAT, which learns jointly from relevance judgments and translation knowledge in the form of a dictionary or a translation table.

2.3 Knowledge Injection into Transformers

There has been a number of efforts to inject structured world knowledge into unsupervised pretraining and contextualized representations [14, 19, 20, 32, 44, 51]. Most of these works focus on integrating knowledge-graphs information such as type of an entity or relatedness between a pair of entities. Lauscher et al. [19] incorporated lexical semantics into BERT by injecting word pairs that are synonyms or hold hyponym-hypernym relations in WordNet. Levine et al. [20] injected word-supersense knowledge by predicting the supersense of a masked word in the input and the ground truth is obtained from Wikipedia. All these works augment an extra knowledge-driven loss with the standard language modeling loss in the language model pre-training stage. We augment translation knowledge in the form of attention in the fine-tuning stage. Our approach is flexible as we can adapt new knowledge as more data for fine-tuning becomes available.

A recent work from Xia et al. [43] used attention-based approach to integrate lexical knowledge for the semantic textual matching task. They created an attention matrix from WordNet and computed the Hadamard product of the attention matrix with BERT's attention matrix. They investigated this approach for computing sentence similarity in a monolingual setting.

3 METHODOLOGY

Our goal is to incorporate additional knowledge from statistical machine translation models or human-constructed dictionaries into a transformer architecture to enable it to connect query and document tokens – not only based on relevance, but also based on translations. In this section, we first define the translation attention matrix given an input query and a candidate document. Then we introduce the translation attention head and the Mixed Attention Transformer (MAT) layer. Finally, we design a sandwich-like architecture to embed MAT into the existing transformer-based neural ranking model.

3.1 Translation Attention Matrix

We define translation reference as a large structural dataset containing knowledge to translate words from one language to another e.g, a human-constructed dictionary, or a translation table trained on parallel corpora. In the CLIR task, the translation knowledge is dependent on the query and document. Therefore, we first design an algorithm that distill the translation knowledge based on tokens in the query and document.

Suppose there exists a word-level translation reference T. Given word w_s in the source language and w_t in the target language, $T(w_t, w_s)$ returns the probability of w_s being translated to w_t: $T(w_t, w_s) = P(w_t|w_s, T)$.

We assume the query is in the source language with length of m_q words and the document is in the target language with length of m_d words. Therefore, the concatenation of query and document $[q, d]$ has length $m = m_q + m_d$. Then we construct a $m \times m$ translation attention matrix M^{tr} based on $[q, d]$ and $T(\cdot, \cdot)$ by symmetrically assigning translation probabilities between query tokens and document tokens. We provide detailed instructions for constructing M^{tr} in Algorithm 1.

Note that the k^{th} row of M^{tr} represents the attention weights of k^{th} token in the input assigned across all the input tokens. In Algorithm 1, lines 2-4 guarantee each token, including out-of-vocabulary word, is assigned a weight to itself and the self weight is the upper bound of all of its translation probabilities. If q_i and d_j are mutually translated words, they get their translation probabilities to each other from lines 5-9. Finally, the row normalization ensures that the attention weights for each input token sum up to 1.

To encode rare words with limited vocabulary size, Byte Pair Encoding (BPE) is often used by pre-trained language models, which splits words into sub-word units. There is evidence that self-attention treats split words differently than non-split ones [10]. Therefore, we use tokens before BPE to query the translation reference and then assign the same attention weight to all parts of the

Algorithm 1: Generate translation attention matrix

Input: $[q, d]$ and $T(\cdot, \cdot)$
Output: M^{tr}

Initialize M^{tr} as a $m \times m$ zero matrix.
for *each token* w_k *in the input sequence* **do**
  $M^{tr}_{kk} = 1$
end
for *each query token* w_i **do**
  for *each document token* w_j **do**
    $M^{tr}_{ij} = M^{tr}_{ji} = T(w_j, w_i)$
  end
end
$M^{tr} \leftarrow \text{RowNorm}(M^{tr})$
return: M^{tr}

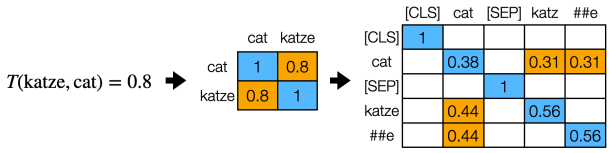


Figure 1: A toy example for generating M^{tr}.

same word. The dimension m of M^{tr} is the same as the length of sequence of $[q, d]$ tokenized by a pre-trained language model. A simplified example for generating M^{tr} with query "cat" and document "katze" (German translation of cat) is shown in Figure 1.

3.2 Mixed Attention Transformer

In order to inject M^{tr} into a transformer-based model, we propose a novel transformer network, named Mixed Attention Transformer (MAT) by combining the multi-head attention with translation-based attention.

The multi-head attention [41] is the core of the transformer architecture which consists of n different attention heads. Given the vector representations as the hidden states $\mathbf{h}$, each head computes the dot-product attention:

$$\text{Attention}_i(\mathbf{h}) = \text{softmax}\Big(\frac{W_i^q\mathbf{h} \cdot W_i^k\mathbf{h}}{\sqrt{d/n}}\Big)W_i^v\mathbf{h}$$

where $\mathbf{h}$ is a d dimensional hidden vector for an input sequence. In BERT, the W_i^q, W_i^k and W_i^v are matrices with size $d/n \times d$. Thus, each head projects to a different subspace of size d/n, learning different information.

Then the outputs of the multi-head attention, $\text{MH}(\cdot)$, are concatenated n heads together and linearly transformed:

$$\text{MH}(\mathbf{h}) = W^o[\text{Attention}_1, \ldots, \text{Attention}_n]$$

In parallel to multi-head attention, we introduce the translation attention head denoted as $\text{TH}(\cdot)$. Inspired by the scaled dot-product attention, we replace the attention weights learned from matrices W_i^q and W_i^k by the fixed attention weights in M^{tr}. Then, the multi-head attention becomes a single fixed attention head as follows

$$\text{TH}(\mathbf{h}) = W^o_{\text{TH}}\big(M^{tr}(W^v_{\text{TH}}\mathbf{h})\big),$$

where both W^o_{TH} and W^v_{TH} are trainable matrices in $\text{TH}(\cdot)$ with dimension $d \times d$. By matrix multiplying with M^{tr}, the translation attention head is capable to reduce the distance between mutually translated tokens in the token representation hyperspace. We prove the effect of M^{tr} on hidden states in a simplified scenario.

Lemma 1. Let convex combinations of vectors A and B be $\alpha A + \beta B$ and $\beta A + \alpha B$ where $\alpha + \beta = 1$. Then, the cosine similarity between $\alpha A + \beta B$ and $\beta A + \alpha B$ is greater or equal to the cosine similarity between A and B.

Proof.

$$\begin{aligned} Sim(\alpha A + \beta B, \beta A + \alpha B) &= \frac{(\alpha A + \beta B) \cdot (\beta A + \alpha B)}{\|\alpha A + \beta B\|\|\beta A + \alpha B\|} \\ &\geq \frac{(\alpha^2 + \beta^2)A \cdot B + \alpha\beta(\|A\|^2 + \|B\|^2)}{(\alpha^2 + \beta^2)\|A\|\|B\| + \alpha\beta(\|A\|^2 + \|B\|^2)} \\ &\geq \frac{A \cdot B}{\|A\|\|B\|}. \end{aligned}$$

Therefore, $Sim(\alpha A + \beta B, \beta A + \alpha B) \geq Sim(A, B)$.

Suppose query word w_i and document word w_j are the translations of each other with probability $p > 0$, and words other than w_j in documents all have zero translation probability with w_i. Then, the only two non-zero weights in the i^{th} row of M^{tr} are self attention (M^{tr}_{ii}) and attention on w_j (M^{tr}_{ij}):

$$M^{tr}_{ii} = \frac{1}{(1+p)};\ M^{tr}_{ij} = \frac{p}{(1+p)}$$

Similarly for w_j, the non-zero weights in the j^{th} row are $M^{tr}_{jj} = 1/(1+p)$ and $M^{tr}_{ji} = p/(1+p)$. If we ignore the trainable matrices in $\text{TH}(\cdot)$ and directly multiply M^{tr} with hidden states $\mathbf{h}$, the translation attention output of w_i and w_j are a convex combination of each other's hidden representations:

$$\text{TH}(\mathbf{h}_{w_i}) = \frac{1}{1+p}\mathbf{h}_{w_i} + \frac{p}{1+p}\mathbf{h}_{w_j}$$

$$\text{TH}(\mathbf{h}_{w_j}) = \frac{1}{1+p}\mathbf{h}_{w_j} + \frac{p}{1+p}\mathbf{h}_{w_i}$$

According to **Lemma 1**, because $p > 0$,

$$Sim\big(\text{TH}(\mathbf{h}_{w_i}), \text{TH}(\mathbf{h}_{w_j})\big) > Sim(\mathbf{h}_{w_i}, \mathbf{h}_{w_j})$$

Thus, when p is large, the words in query and document are likely to be translation to each other. The attention matrix M^{tr} "pays attention" to all these pair of words and $\text{TH}(\cdot)$ tends to "pull" their hidden representations closer in the hyperspace.

The complete attention mechanism in MAT is a combination of the attention outputs from both $\text{MH}(\cdot)$ and $\text{TH}(\cdot)$. We first employ a residual connection around each type of attention output, followed by layer normalization, denoted as $\text{LN}(\cdot)$, resulting two sub-layer outputs. Then we sum two sub-layer outputs:

$$\text{Sublayer}_{\text{MH}}(\mathbf{h}) = \text{LN}(\mathbf{h} + \text{MH}(\mathbf{h}))$$

$$\text{Sublayer}_{\text{TH}}(\mathbf{h}) = \text{LN}(\mathbf{h} + \text{TH}(\mathbf{h}))$$

$$\mathbf{h}' = \text{Sublayer}_{\text{MH}}(\mathbf{h}) + \text{Sublayer}_{\text{TH}}(\mathbf{h})$$

And apply the summed result to the position-wise feed-forward networks (FFN),

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

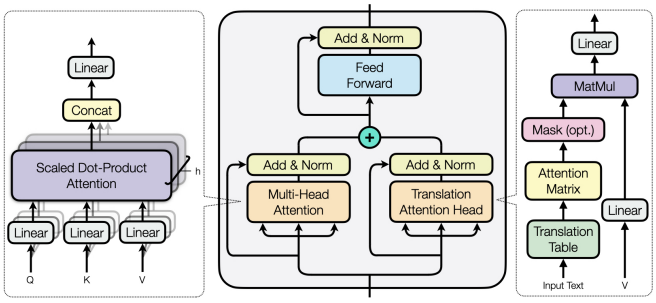


Figure 2: (left) Multi-Head Attention. (right) Translation Attention Head. (middle) Mixed Attention Transformer Layer.

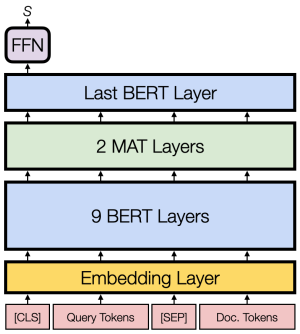


Figure 3: Use MAT layers in BERT ranking model

The final output of MAT is another residual connection around the output of FFN:

$$\text{MAT}(\mathbf{h}) = \text{LN}(\mathbf{h}' + \text{FFN}(\mathbf{h}'))$$

The complete MAT architecture is depicted in Figure 2 (middle). The left and right of Figure 2 are two types of attention component in MAT. The benefits of this network architecture are that the MAT can attend to both contextual information from multi-head attention and translation knowledge from translation attention head during training. Because we keep the multi-head attention mechanism and share the FFN sublayer, MAT contains a vanilla transformer network. This design allows MAT to be easily embedded into recent transformer-based pre-trained models and fully leverage the pre-trained weights.

3.3 Embed MAT into Pre-trained Model

The transformer-based models usually have the following architecture: First, the embedding layer encodes the input tokens, segments, and positions into hidden representations. The representation of each input token is then updated by a stack of encoder layers based on the attention mechanism. Finally, a specialized add-on network maps the hidden representations to an output based on the task.

Qiao et al. [34] analyzed different ranking models based on BERT and found that the `Last-Int` approach which applies BERT on the concatenated $[q, d]$ sequence and uses the last layer's representation of the [CLS] token as the matching feature gives the best performance. In this section, we use the same BERT (`Last-Int`) as a re-ranker to discuss how to embed MAT into a transformer-based pre-trained language model.

MART (**MA**T+BE**RT**), the new model architecture we propose is to keep the embedding layer and add-on network while replacing some of the transformer layers in the middle by MAT.

During fine-tuning, the BERT layers close to the output (higher layers) are more sensitive than the lower layers [53]. Also, another study on BERT [40] has shown that most local syntactic phenomena are encoded in lower layers while higher layers capture more complex semantics. Consider the fine-tuning efficiency and semantic quality of the token representations, the layer replacement should start from the higher layers of BERT. Moreover, in the `Last-Int` ranking approach, the output score is only based on the [CLS] token in the last BERT layer. Therefore, we keep the last BERT (`Base`) layer as the output layer and start to embed MAT from the 11^{th} layer. Figure 3 shows an examples of the sandwich-like architecture based on a BERT-based ranking model where MAT layers are embedded into 10^{th} and 11^{th} layers of BERT. Using the same hidden dimension as BERT, each MAT layer only introduces about 1.18M new parameters comparing to the BERT layer. At initialization, MAT is able to use pre-trained weights of its corresponding BERT layer. This compatibility increases the fine-tuning efficiency and reduces the training data requirement.

4 EXPERIMENTAL SETUP

4.1 Dataset

CLIR dataset. We create our training and evaluation data from the Cross-Language Evaluation Forum (CLEF) 2000-2008 campaign for bilingual ad-hoc retrieval tracks [3–6, 27–31]. We use the text fields of the documents to construct our retrieval corpus and discard other meta data. We concatenate the title and description fields of a topic and consider it as our query. We consider all the topics and relevance judgments from all the tracks to show the consistent effectiveness of MAT across several cross-language retrieval settings on both high- and low-resource languages.

Translation Resources. Our goal is to leverage translation resources as external knowledge into the query-document matching process and we compare the effectiveness of three types of resources: sentence-level parallel data, dictionary, and bi-lingual word embeddings. We use sentence-level parallel data with GIZA++ toolkit [26] to construct a translation table, which we use to generate M^{tr}. Translation tables for European languages are based on the Europarl v7 sentence-aligned corpora [18]. For our limited-resource (both in terms of parallel data and relevance judgments) setting based on Somali and Swahili languages, we use the translation tables provided by Zhang et al. [50].

As the dictionary-based translation resource we use Panlex, a dictionary [16] whose data acquisition strategy emphasizes high-quality lexical mapping and broad language coverage. Finally, we also explore the a multilingual word embedding as a translation resource following Bonab et al. [2]. Given a pair of words we use their representations from a multilingual word embedding model and compute cosine similarity to model relatedness of the pair of words. In our experiments, we use MUSE, an unsupervised multilingual word embedding from [9] as translation resource.

Text Pre-processing. In order to have consistent pieces of text across different resources, we normalize characters by mapping

Table 1: Summary of CLIR setting. First four rows indicate the backward and the last row indicates the forward setting.

CLIR Setting	Collection Source	Collection Size	Query Size
Eng-Fre	Le Monde, Sda French	129,689	185
Eng-Ita	La Stampa, Sda Italian	144,040	176
Eng-Deu	Der Spiegel, Frankfurter Rundschau	153,496	184
Eng-Spa	EFE News 94-95	452,027	156
Xxx-Eng	Los Angeles Times 94	113,005	246

diacritic characters to the corresponding unmarked characters and then lower-casing text. For initial step of retrieval and translation table extraction from parallel corpora, we remove non-alphabetic, non-printable, and punctuation characters. We use NLTK library to tokenize and remove stop-words, but do not stem the tokens.

4.2 CLIR Settings

Forward: Non-English Query and English Documents. In this setting, we use non-English queries against an English document collection. To evaluate cross-lingual matching performance, we use human translation of a fixed query set to obtain queries in different languages. While we have translations of queries in different languages, we keep the content and language of the retrieval corpus fixed. We have both high-resource and low-resource CLIR settings in our experiments. In a high-resource setting, for example, French-English, we have higher amount of sentence-level parallel data and relevance judgments compared to a low-resource setting.

There are four high-resource language pairs in our experiments: French (Fre-Eng), Italian (Ita-Eng), German (Deu-Eng), and Spanish (Spa-Eng). Queries are selected from CLEF C001 – C350 topic set for each language. We take the intersection of the topic ID and remove topics without any relevant document, resulting in 246 overlapped queries across four languages. For cross-language information retrieval involving low-resource languages, we experiment on Somali (Som-Eng) and Swahili (Swa-Eng). Bonab et al. [1] provided Somali and Swahili translations of 151 English queries from the CLEF C001 – C200 topic set and we use those queries in our setting. The collection of English documents is the Los Angeles Times corpus comprised of 113k news articles.

Backward: English Query and Non-English Documents In this setting, we use English queries against document collections in four languages: French (Eng-Fre), Italian (Eng-Ita), German (Eng-Deu) and Spanish (Eng-Spa). For each language, we create a retrieval corpus from a combination of sources which we report in Table 1.As the retrieval corpus varies for each language, relevance judgments are not available for all the English topics from CLEF C001 – C350 topic set. Thus, for each CLIR setting we have a different number of queries in the backward setting compared to the forward setting. Table 1 provides information about query sets and document collections in both the settings.

4.3 Implementation Details

Pre-trained passage re-ranker Nogueira and Cho [25] fine-tuned the `Base`, `Uncased` multilingual BERT (mBERT) on MS MARCO document retrieval dataset to create a passage ranking model. We refer to this pre-trained model as m^{2}BERT and further fine-tune it with cross-lingual relevance judgments. To prepare the input sequence for m^{2}BERT we concatenate a query and a document separated by a special [SEP] token from mBERT's vocabulary. We prefix the concatenated sequence with the special [CLS] token from mBERT's vocabulary. We obtain the last layer representation of this sequence from m^{2}BERT, but only use the representation of the [CLS] token, and pass it through a linear combination layer to obtain the probability of the document being relevant to the query. At test time, given a query, m^{2}BERT computes the probability for each document independently and obtains a document ranking after sorting with these probability scores. Because the mBERT input sequence is limited to 512 tokens, longer documents are split evenly and [CLS] representations from all document segments are averaged to obtain a representation for fine-tuning. MacAvaney et al. [24] used the same approach for monolingual retrieval.

Evaluation. For evaluating retrieval effectiveness, we follow prior work on CLEF dataset [2, 23] and report mean average precision (MAP) of the top 100 ranked documents and precision of the top 10 retrieved documents (P@10). We determine statistical significance using the two-tailed paired t-test with p-value less than 0.05 (i.e., 95% confidence level).

Model training. We train all neural re-ranking models using pairwise cross-entropy loss [11]. We use all the positive document from the query relevance judgments and randomly sample negative documents to form training pairs. We truncate document contents to the first 800 tokens and create two passages to represent a document if the sum of the query length and document length is over the 512 tokens, which is the limit of mBERT. We pass a two query-document pairs in each forward pass but use gradient accumulation to make our effective batch size to 16. We train all the models for 100 epochs with an early stopping strategy with patience value of 20. All models are trained using Adam's optimization algorithm [17] with a learning rate of 2e-5.

Given the limited number of queries in each language, we use 5-fold cross-validation for robust evaluation. For each fold, the training, validation, and test data are 60%, 20%, and 20% of the query set, respectively. The reported evaluation metrics are averaged across 5 folds. We also fix the random seed is set to guarantee that all models receive the same training data. For the validation queries, we re-rank the top 100 documents and use MAP to select the best-performing model.

4.4 Compared Methods.

We compare the proposed model with the methods in the following

- **SMT**: We first use the GIZA++ toolkit [26] to build translation tables from parallel corpora. We select top-10 translations from the translation table for each query term and apply Galago[1]'s weighted *#combine* operator to form a translated query. Then we use the Galago's implementation of Okapi BM25 [35] with default parameters. This setting is taken from Bonab et al. [2], and we call this method statistical machine translation (SMT). It serves as one of our baselines. Moreover, the training data for neural re-ranking models are sampled based on the top 500 retrieved documents by the SMT model.
- **m^{2}BERT**: To create the m^{2}BERT baseline we begin with the pre-trained checkpoint provided by [25]. This checkpoint is a

[1] https://www.lemurproject.org/galago.php/

Table 2: Model performance on forward and backward settings for high-resource languages. The highest value for each column is marked with bold text. Statistically significant improvements are marked by † (over SMT) and ‡ (over BERT).

	Model	Fre-Eng		Ita-Eng		Deu-Eng		Spa-Eng	
		MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
Forward Setting	Human Translation	0.4569	0.3940	0.4569	0.3940	0.4569	0.3940	0.4569	0.3940
	SMT	0.3618	0.3492	0.3561	0.3431	0.3588	0.3354	0.3624	0.3317
	m^2BERT	0.3802†	0.3799†	0.3652	0.3545	0.3582	0.3335	0.3819†	0.3693†
	MART-PLB	0.3859†	0.3666†	0.3701	0.3689†	0.3593	0.3501†	0.3824†	0.3676†
	MART	**0.4126**†‡	**0.3935**†‡	**0.3944**†‡	**0.3732**†‡	**0.3862**†‡	**0.3770**†‡	**0.3953**†‡	**0.3830**†‡
	Model	Eng-Fre		Eng-Ita		Eng-Deu		Eng-Spa	
		MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
Backward Setting	Human Translation	0.2955	0.3054	0.2629	0.2892	0.2970	0.3060	0.2518	0.2436
	SMT	0.2258	0.2319	0.1883	0.1852	0.2614	0.2424	0.1985	0.2088
	m^2BERT	0.2841†	0.2875†	0.2635†	0.2605†	0.3241†	0.3246†	0.2355†	0.2285†
	MART-PLB	0.2807†	0.2823†	0.2713†	0.2771†	0.3262†	0.3230†	0.2389†	0.2351†
	MART	**0.3002**†‡	**0.3108**†‡	**0.2823**†‡	**0.2846**†‡	**0.3433**†‡	**0.3414**†‡	**0.2558**†‡	**0.2439**†‡

result of fine-tuning the multilingual BERT (mBERT) architecture with MSMARCO passage ranking dataset. We further fine-tune it with training data from a specific CLIR setting. We use the same fine-tuning approach described in section 4.3 for this baseline and our proposed model to ensure fair comparison.

- **MART-PLB**: This is a variant of MART. In order to evaluate the effect of external knowledge on MAT, we replace M_{tr} by an identity matrix so that each token is only paying attention to itself. Therefore, instead of injecting translation knowledge into the model, we design a "placebo" attention matrix for MAT. Using MART-PLB as a controlled experiment, we are able to evaluate the effect of external knowledge.

In order to provide an empirical upper-bound on retrieval performance, we use human translation of the queries and apply BM25 as the retrieval technique. The human translations of the queries are obtained from the CLEF dataset as they have a common topic ID for the same queries across different languages.

5 EXPERIMENTAL RESULTS

5.1 Performance on High-resource Languages

Table 2 lists evaluation results on both Forward (top) and Backward (bottom) settings for language pairs with high translation resources.

As a neural re-ranker, m^2BERT significantly improves upon SMT on all language pairs in backward setting and two language pairs on the forward setting while performs on par with SMT for Deu-Eng and Ita-Eng languages. While fine-tuned on English document retrieval dataset, m^2BERT can transfer to cross-lingual task with small amount of fine-tuning data. This agrees with the previous finding by Pires et al. [33] that mBERT is capable to generalize across languages.

We observed substantial improvements on the retrieval performance when translation knowledge is incorporated into MART. For all language setting combination in Table 2, MART performs significantly better than the BERT architecture (m^2BERT) in terms of both MAP and P@10. MART improves m^2BERT by 8% on the forward and 7% on the backward settings in terms of MAP. This comprehensive comparisons with vanilla BERT based ranker demonstrate the effectiveness of the MAT-embedded model.

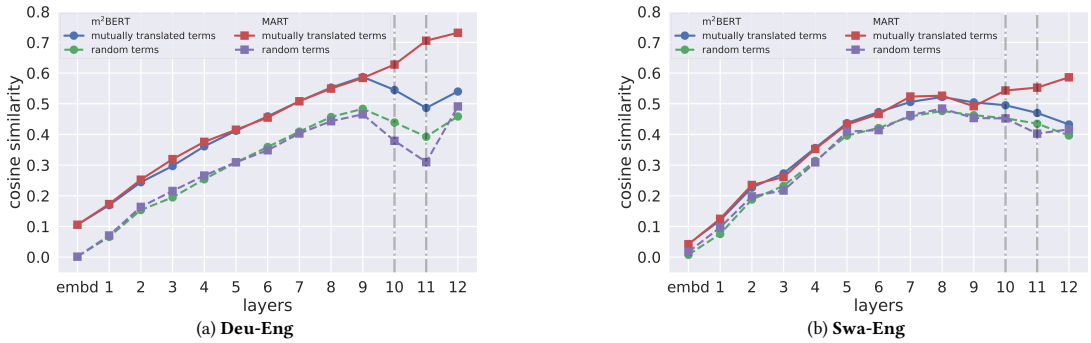
Replacing M_{tr} by the identity matrix in MART-PLB, the translation attention head degenerates to two additional feed-forward layers. MART-PLB behaves insignificantly comparing to the vanilla BERT architecture on all languages. Such results indicate that the performance gain in MART relies on injecting the external knowledge, not from adding new parameters. When M_{tr} becomes non-informative, the translation attention head is ineffectual.

Comparing MART with Human Translation, we can see that in forward setting, correct translation with basic retrieval model still lead the neural CLIR model. However, in backward setting, MART achieves relatively the same as (Eng-Fre, Eng-Spa) or better than (Eng-Ita, Eng-Deu) Human Translation. We hypothesize that in the backward setting, translation tables provide higher quality translations which enable better semantic matching between query and document tokens.

5.2 Performance on Low-resource Languages

The evaluation results for two language pairs with limited translation resources on the forward setting are shown in Table 3. We make several observations. First, m^2BERT mostly under-performs SMT for both Somali and Swahili languages. Note that Somali is not included in the pre-training of mBERT. Even if Swahili is included, there is only a small number of Swahili sentences in the pre-training data. The low performance of m^2BERT on low-resource language pairs demonstrates that absence or inadequate pre-training data on a particular language leads to poor performance on target tasks involving those languages.

On the other hand, the MART model achieves the highest MAP performance for both Somali and Swahili languages. The consistent and significant improvements in terms of MAP over compared methods make MART the best model in our experiments. Due to the lack of pre-training data, the translation gap is more critical in low-resource language pairs. The performance of MART for Somali and Swahili languages proves that leveraging the external translation



(a) **Deu-Eng** (b) **Swa-Eng**

Figure 4: The comparison of MART to m²BERT on layer-wise token representations.

Table 3: Model performance for low-resource languages on Forward setting. The highest value for each column is marked with bold text. Statistically significant improvements are marked by † (over SMT) and ‡ (over m²BERT).

Model	Som-Eng		Swa-Eng	
	MAP	P@10	MAP	P@10
Human Translation	0.4563	0.3940	0.4563	0.3940
SMT	0.1948	0.1865	0.2184	**0.2152**
m²BERT	0.1986	0.1772	0.2055	0.2089
MART-PLB	0.2049	0.1972†‡	0.2130	0.2106
MART	**0.2207**†‡	**0.2135**†‡	**0.2348**†‡	0.2151

knowledge can help to bridge the translation gap. Moreover, the experiments with the placebo setting, similar to those for the high-resource languages, have shown no significance in performance compared to m²BERT. These results strengthen the conclusion that the translation attention matrix is the key component of MAT.

Human Translation leads neural ranking models by a large margin in CLIR tasks involving low-resource languages. This is expected because, with less sentence-level parallel data, the CLIR models often suffer from low quality of translations.

5.3 Representation Analysis

To study the influence of MAT on the translation gap in neural CLIR, we compare the token representation from each layer between m²BERT and MART. Specifically, both models are fine-tuned on Deu-Eng and Swa-Eng training data. Figure 4 shows the distances between contextualized token representations in two model architectures where x-axis represents layers from low to high and y-axis is the cosine similarity. We focus on two types of word pairs (one from query and another from document) in an input sequence: (i) Mutually translated words, where all pairs of words that are translations to each other according to the external translation knowledge are selected; and (ii) Random non-translated words, where we randomly sample 10 pairs of words which are not translations of each other. We compute the average cosine similarity of the token representations at each layer for all selected word pairs in the test data of Deu-Eng (high-resource) and Swa-Eng (low-resource).

From the diagrams in Figure 4, we can see that in general, the similarity of token representations increases as the layer gets higher.

Table 4: MART performance for different external knowledge. The highest value for each column is marked with bold text. "–" if language is not supported.

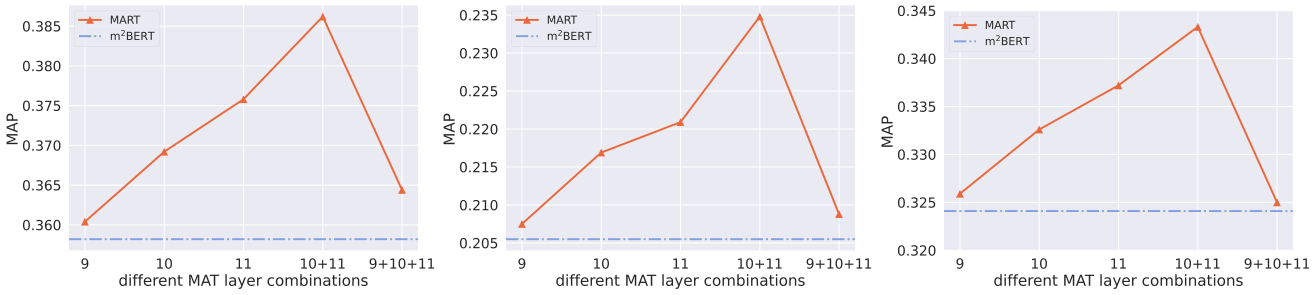
External Knowledge	Forward				Backward	
	Deu-Eng		Swa-Eng		Eng-Deu	
	MAP	P@10	MAP	P@10	MAP	P@10
Parallel Corpus	**0.3862**	**0.3770**	**0.2348**	**0.2151**	**0.3433**	**0.3414**
Panlex	0.3713	0.3612	0.2265	0.2073	0.3326	0.3360
MUSE	0.3693	0.3580	–	–	0.3335	0.3348

Also, the mutually translated words always have smaller cosine distances than non-translated words. The closer lines between two types of word pairs in Swa-Eng prove that the translation gap is more critical in resource-lean languages. We can also see that in 10th and 11th layers, the similarity of two types of words in m²BERT drops for both language pairs. According to the previous analysis [33], one hypothesis for such drop is that before fine-tuning on MS MARCO dataset, mBERT was pre-trained on surrounding contexts for language modeling, it needs more contextual information to correctly predict the missing words. Therefore, mBERT favors text sequence pairs that are closer in their semantic meanings. Such models trained on surrounding context are not as effective for ad-hoc document ranking with respect to keyword queries [34].

MART shows the same behavior as m²BERT up to the MAT layers. The representations of mutually translated words in MAT layers become similar to each other in terms of cosine distance. This matches the design purpose of MAT. Meanwhile, because MAT keeps the native multi-head attention from BERT layer, the similarity of non-translations still drops in MAT layers. The increased similarity on mutually translated words and decreased similarity on non-translated words demonstrate that model is bridging translation gap with the help of external knowledge.

5.4 Effect of Translation Resources

From the previous results, we have seen that the translation attention matrix is critical to the success of MAT. As a knowledge injection model, it is palpable that the quality of the knowledge affects the model performance. In this experiment, we study the effect of different sources of external knowledge on the MART. Besides the translation table built from parallel data, we use two



(a) **Deu-Eng** (b) **Swa-Eng** (c) **Eng-Deu**

Figure 5: The performance comparison of different MART model architectures.

different translation knowledge for M_{tr} generation: Panlex dictionary [16] and multilingual word embedding (MUSE [9]). To obtain translation probability for a single word in Panlex, we uniformly distribute weights to all possible translations. And in MUSE, we use the 5 nearest neighbors of a word in the target languages as its potential translations and assign translation probability based on their normalized cosine similarity. In order to cover different languages and retrieval settings, we select Deu-Eng (high-resource) and Swa-Eng (low-resource) from forward setting and Eng-Deu from backward setting for this experiment.

Table 4 shows the results of all compared translation knowledge. We observe a performance drop on both alternative knowledge resources. For Panlex, although the translations are more precise than those in a translation table, they do not provide a broad coverage of words. Multilingual word embeddings are learnt from the contexts of words, not their translations. Therefore, given a word, the embeddings of semantically similar words are often closer than those of its translations to the embedding of a word [2]. Thus, using multilingual word embeddings, the problem of the translation gap will not be completely resolved.

5.5 Ablation Study on Model Architecture

In this section, we empirically study the effects of different numbers and positions of MAT layers in a MART model. We further train and evaluate the MART with various combinations of MAT layers. It is worth mentioning that given the number of layers in BERT architecture, there exist exponential number of possible combinations. We only explore several representative models. Leaving the last layer as the output layer, we still focus on the higher transformer layers of BERT architecture. For models with a single MAT layer, we investigate MART with MAT embedded at 9^{th}, 10^{th}, or 11^{th} layer. For double MAT layers, we use the previous results from MAT at 10^{th} and 11^{th} layers. We also consider an architecture with three MAT layers where 9^{th}, 10^{th} and 11^{th} layers in BERT are all replaced by the MAT layer.

Figure 5 shows the performance of different MART model architectures on Deu-Eng, Swa-Eng and Eng-Deu. We can see that all model variants have the similar pattern across three selected CLIR tasks. Because higher BERT layers are more sensitive to fine-tuning [53] and their hidden representations capture complex semantic information [40], the retrieval performance for the single MAT layer increases from MAT at the 9^{th} layer to MAT at the 11^{th} layer. The double MAT layer can further boost performance from the single-layer approach. We also can see that models get less improved when 9^{th} in replaced by MAT. We hypothesize that the token representations after the 8^{th} layer (the input of the 9^{th} layer) do not contain enough semantic information [40] so it is too early to apply the translation attentions.

6 CONCLUSION

In this paper, we propose a novel Mixed Attention Transformer (MAT) network to leverage external translation knowledge for cross-lingual information retrieval tasks.

First, we build attention matrix for mutually translated words between query and document based on the translation resource. Then using the attention matrix, we design a new translation attention head and show that it is able to reduce the cosine distance between hidden representations of mutually translated words. Finally, the complete architecture of MAT is a combination of multi-head attention and translation attention head with shared feed-forward networks. As a layer component, we further design a sandwich-like architecture to embed MAT into the Transformer model. Our comprehensive experimental results demonstrate the effectiveness of external knowledge and the significant improvement of MAT-embedded neural model on CLIR task.

For future work, we are particularly interested in fine-tuning MART on a large CLIR dataset with a mix of cross-language settings to learn a language-agnostic neural ranking model. We also plan to apply MAT to other retrieval tasks, e.g., event retrieval, by incorporating information other than translation knowledge.

ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part under USC (University of Southern California) subcontract no. 124338456 under IARPA prime contract no. 2019-19051600007., and in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) via AFRL contact #FA8650-17-C-9116 under subcontract #94671240 from the University of Southern California. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the 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 thereon.

REFERENCES

[1] Hamed Bonab, James Allan, and Ramesh Sitaraman. 2019. Simulating CLIR Translation Resource Scarcity Using High-Resource Languages. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval* (Santa Clara, CA, USA) *(ICTIR '19)*. Association for Computing Machinery, New York, NY, USA, 129–136. https://doi.org/10.1145/3341981.3344236

[2] Hamed Bonab, Sheikh Muhammad Sarwar, and James Allan. 2020. Training Effective Neural CLIR by Bridging the Translation Gap. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 9–18.

[3] Martin Braschler. 2001. CLEF 2000 — Overview of Results. In *Cross-Language Information Retrieval and Evaluation*, Carol Peters (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 89–101.

[4] Martin Braschler. 2002. CLEF 2001 — Overview of Results. In *Evaluation of Cross-Language Information Retrieval Systems*, Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 9–26.

[5] Martin Braschler. 2002. CLEF 2002—Overview of results. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 9–27.

[6] Martin Braschler. 2003. CLEF 2003–Overview of results. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 44–63.

[7] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8440–8451.

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[9] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087* (2017).

[10] Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. 2019. Adaptively Sparse Transformers. 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)*. Association for Computational Linguistics, Hong Kong, China, 2174–2184. https://doi.org/10.18653/v1/D19-1223

[11] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 65–74.

[12] 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[13] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A Deep Relevance Matching Model for Ad-hoc Retrieval. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (Oct 2016). https://doi.org/10.1145/2983323.2983769

[14] Bin He, Di Zhou, JingHui Xiao, X. Jiang, Qun Liu, Nicholas Jing Yuan, and T. Xu. 2020. Integrating Graph Contextualized Knowledge into Pre-trained Language Models. *ArXiv* abs/1912.00147 (2020).

[15] Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual Information Retrieval with BERT. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*. European Language Resources Association, Marseille, France, 26–31. https://www.aclweb.org/anthology/2020.clssts-1.5

[16] David Kamholz, Jonathan Pool, and Susan M Colowick. 2014. PanLex: Building a Resource for Panlingual Lexical Translation.. In *LREC*. 3145–3150.

[17] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. (2015).

[18] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, Vol. 5. Citeseer, 79–86.

[19] Anne Lauscher, Ivan Vuli'c, E. Ponti, A. Korhonen, and Goran Glavavs. 2020. Specializing Unsupervised Pretraining Models for Word-Level Semantic Similarity. In *COLING*.

[20] Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, S. Shalev-Shwartz, A. Shashua, and Y. Shoham. 2020. SenseBERT: Driving Some Sense into BERT. *ArXiv* abs/1908.05646 (2020).

[21] Bo Li and Ping Cheng. 2018. Learning neural representation for clir with adversarial framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 1861–1870.

[22] 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.

[23] Robert Litschko, Goran Glavaš, Ivan Vulic, and Laura Dietz. 2019. Evaluating Resource-Lean Cross-Lingual Embedding Models in Unsupervised Retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1109–1112. https://doi.org/10.1145/3331184.3331324

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

[25] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

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

[27] Carol Peters. 2005. What Happened in CLEF 2004?. In *Multilingual Information Access for Text, Speech and Images*, Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–9.

[28] Carol Peters. 2006. What Happened in CLEF 2005. In *Accessing Multilingual Information Repositories*, Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Müller, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–10.

[29] Carol Peters. 2007. What Happened in CLEF 2006. In *Evaluation of Multilingual and Multi-modal Information Retrieval*, Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–10.

[30] Carol Peters. 2008. What Happened in CLEF 2007. In *Advances in Multilingual and Multimodal Information Retrieval*, Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivien Petras, and Diana Santos (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–12.

[31] Carol Peters. 2009. What Happened in CLEF 2008. In *Evaluating Systems for Multilingual and Multimodal Information Access*, Carol Peters, Thomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J. F. Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Vivien Petras (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–14.

[32] Matthew E. Peters, Mark Neumann, IV RobertLLogan, Roy Schwartz, V. Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge Enhanced Contextual Word Representations. In *EMNLP/IJCNLP*.

[33] Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 4996–5001. https://doi.org/10.18653/v1/P19-1493

[34] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the Behaviors of BERT in Ranking. arXiv:1904.07531 [cs.IR]

[35] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp* 109 (1995), 109.

[36] Shadi Saleh and Pavel Pecina. 2020. Document Translation vs. Query Translation for Cross-Lingual Information Retrieval in the Medical Domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6849–6860. https://doi.org/10.18653/v1/2020.acl-main.613

[37] Sheikh Muhammad Sarwar, Hamed Bonab, and James Allan. 2019. A Multi-Task Architecture on Relevance-based Neural Query Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6339–6344. https://doi.org/10.18653/v1/P19-1639

[38] 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)*. 458–463.

[39] 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 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 488–494.

[40] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 4593–4601. https://doi.org/10.18653/v1/P19-1452

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[42] 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*. 363–372.

[43] Tingyu Xia, Yue Wang, Yuan Tian, and Yi Chang. 2021. Using Prior Knowledge to Guide BERT's Attention in Semantic Textual Matching Tasks. (2021).

[44] Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2020. Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model. *ArXiv* abs/1912.09637 (2020).

[45] Mahsa Yarmohammadi, Xutai Ma, Sorami Hisamoto, Muhammad Rahman, Yiming Wang, Hainan Xu, Daniel Povey, Philipp Koehn, and Kevin Duh. 2019. Robust Document Representations for Cross-Lingual Information Retrieval in Low-Resource Settings. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*. European Association for Machine Translation, Dublin, Ireland, 12–20. https://www.aclweb.org/anthology/W19-6602

[46] Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained Transformers for Text Ranking: BERT and Beyond. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*. Association for Computational Linguistics, Online, 1–4. https://www.aclweb.org/anthology/2021.naacl-tutorials.1

[47] Puxuan Yu and James Allan. 2020. *A Study of Neural Matching Models for Cross-Lingual IR*. Association for Computing Machinery, New York, NY, USA, 1637–1640. https://doi.org/10.1145/3397271.3401322

[48] Puxuan Yu, Hongliang Fei, and Ping Li. 2021. Cross-Lingual Language Model Pretraining for Retrieval. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1029–1039. https://doi.org/10.1145/3442381.3449830

[49] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. An Analysis of BERT in Document Ranking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1941–1944. https://doi.org/10.1145/3397271.3401325

[50] Le Zhang, Damianos Karakos, William Hartmann, Manaj Srivastava, Lee Tarlin, David Akodes, Sanjay Krishna Gouda, Numra Bathool, Lingjun Zhao, Zhuolin Jiang, Richard Schwartz, and John Makhoul. 2020. The 2019 BBN Cross-lingual Information Retrieval System. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*. European Language Resources Association, Marseille, France, 44–51. https://www.aclweb.org/anthology/2020.clssts-1.8

[51] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1441–1451. https://doi.org/10.18653/v1/P19-1139

[52] Lingjun Zhao, Rabih Zbib, Zhuolin Jiang, Damianos Karakos, and Zhongqiang Huang. 2019. Weakly supervised attentional model for low resource ad-hoc cross-lingual information retrieval. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. 259–264.

[53] Yiyun Zhao and Steven Bethard. 2020. How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4729–4747. https://doi.org/10.18653/v1/2020.acl-main.429