Journal of Intelligent & Fuzzy Systems xx (20xx) x–xx
DOI:10.3233/JIFS-213070
IOS Press

Improving cross-lingual text matching with dual-level collaborative coarse-to-fine filter alignment network

Yan Li[a,b], Junjun Guo[a,b,*], Zhengtao Yu[a,b] and Shengxiang Gao[a,b]
[a]*Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Yunnan, Kunming, China*
[b]*Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Yunnan, Kunming, China*

Abstract. Semantic alignment is a key component in Cross-Language Text Matching (CLTM) to facilitate matching (*e.g.,* query-document matching) between two languages. The current solutions for semantic alignment mainly perform word-level translation directly, without considering the contextual information for the whole query and documents. To this end, we propose a Dual-Level Collaborative Rough-to-Fine Filter Alignment Network (DLCCFA) to achieve better cross-language semantic alignment and document matching. DLCCFA is devised with both a coarse-grained filter in word-level and a fine-grained filter in sentence-level. Concretely, for the query in word-level, we firstly extract top-k translation candidates for each token in the query through a probabilistic bilingual lexicon. Then, a Translation Probability Attention (TPA) mechanism is proposed to obtain coarse-grained word alignment, which generates the corresponding query auxiliary sentence. Afterwards, we further propose a Bilingual Cross Attention and utilize Self-Attention to achieve fine-grained sentence-level filtering, resulting in the cross-language representation of the query. The idea is that each token in the query works as an anchor to filter the semantic noise in the query auxiliary sentence and accurately align semantics of different languages. Extensive experiments on four real-world datasets of six languages demostrate that our method can outperform the mainstream alternatives of CLTM.

Keywords: Cross-language text matching , Alignment , Probabilistic bilingual lexicon , Translation probability attention , Bilingual cross attention

1. Introduction

Cross-lingual text matching (CLTM) is a valuable but challenging task in information retrieval and natural language processing areas, which can be considered as a primary form of information searching across language boundaries [1–3]. Given a query-document pair, CLTM aims to predict their semantic relations. Recently, due to requriment of information processing and management, as well as the explosive growth of the online resources, CLTM is becoming increasingly more important. Apart from this, CLTM can be very useful in some scenarios. For example, imagine a journalist who wishes to monitor the latest news of Corona Virus Disease 2019 (*i.e.,* COVID-19) around the world in real time. He/She might issue a query in Chinese, and desire to search all relevant news in any language to obtain different perspectives for his/her press release. The cross-lingual alignment is of primary importance to cross the language barrier in CLTM models.

There are several ways to bridge sematic gap in CLTM models. Traditionally, the most effective way to cross the language barrier is to utilize *query translation* approach [4–6], *document translation*

*Corresponding author. Junjun Guo, E-mail: guojjgb@163.com.

approach [7, 8] or by using both *query and document translation* approach [9]. All these approaches involve a pipeline of two components: *query translation* and *monolingual matching model*. Firstly, the query in the source language is translated into the target language by using a machine translation (MT) or a bilingual lexicon. Then, the translated query is used to match relevant documents in the target language. Rahimi et al. [10] built an effective translation model from comparable corpora for CLTM. Qin Ying et al. [11] proposed to utilize bilingual dictionaries to tackle the ambiguity and multiple matching problems. Wang et al. [12] leveraged statistical estimation of translation probabilities for CLTM. However, the performance of this pipeline workflow is fundamentally limited by the quality of machine translation, especially for low-resource languages. That is, due to the accumulation of translation errors, it will have a great impact on subsequent matching task and even directly lead to the correlation prediction failure.

Inspired by great success achieved by pre-trained word embedding techniques(*e.g.,* Word2vec or GloVe [13, 14]), many solutions utilizing word embeddings have been proposed to enhance CLTM. These works often project different languages into the same hidden space. On the basis of Word2Vec, Ivan Vulić et al. [15] utilized a comprehensive typology to generate cross-language word embeddings (CLE) over a randomly shuffle parallel corpus. Litschko et al. [16] introduced a completely unsupervised cross-language information retrieval framework that leverages off-the-shelf pre-trained CLEs to combine query translation with semantic space ranking. Some efforts hase also extended the pre-trained language models for jointly cross-language embedding learning, *i.e.,* encoding over 100 languages into a shared CLE space (*e.g.,* mBERT [17]). However, these word embeddings usually need to be pre-trained with high-quality bilingual dictionaries, which is infeasible for many low-resource languages.

Despite the encouraging improvements have achieved by the aforementioned methods, the alignment made with the help of CLE is far beyond the expectation since the contextual information of the query could be easily overlooked. To alleviate the problems caused by the above methods, we observe some examples extracted from the Zh-Vi CLTM dataset. As shown in Fig. 1, the query and the document belong to different languages. As observed from Fig. 1, for low-resource language pairs, the results of MT completely distort the meaning of the query. Therefore, the core difficulty of CLTM is to achieve context-aware semantic alignment between the query and the translated counterpart in different languages. Unfortunately, the alignment based on MT and CLEs may distort the semantics of the query in the source side, leading to the loss of valuable information. This is more useful for low-resource languages since the scarcity of the parallel multi-lingual corpus could easily hurt the semantic alignment.

2. Research objective and contribution

In order to using the contextual information for the whole query and documents to achieve better semantic alignment and cross-language text matching. In this paper, we explore an efficient method with fine-grained context-aware semantic alignment for better CLTM. Specifically, we propose a cross-language deep matching model based on Dual-Level Collaborative Rough-to-Fine Filter Alignment Network (namely DLC-CFA) to achieve better CLTM. This network, is divided into *word-level coarse-grained filter* and *sentence-level fine-grained filter*. Firstly, the shared Transformer-encoder is adopted to extract contextual representations for the query and documents. Then, in word-level coarse-grained filter, top-k translation candidates of each token in the query are firstly retrieved based on a probabilistic bilingual lexicon. Then, we propose a translation probability attention (TPA) mechanism by utilizing the Gumbel-Softmax to identify the most appropriate translation candidates, which results in a translated query in target language (namely query auxiliary sentence).

Although the query auxiliary sentence generated by word-level coarse-grained filter could lose the semantic to some extent, it can retain the keyword information of the query. Note that the keyword information has well been validated as the most important signals for query-document matching. In addition, in order to achieve fine filtering and cross-language representation of the query, we introduce a Bilingual Cross Attention and Self-Attention mechanism to form our sentence-level fine-grained filter, where each token in the query works as an anchor to identify the important information from the query auxiliary sentence. Finally, a Bilingual Reranker is further introduced for better cross-language text matching.

The main contributions of this paper are as follows:

– In this paper, we propose a dual-level collaborative coarse-to-fine filter alignment network to

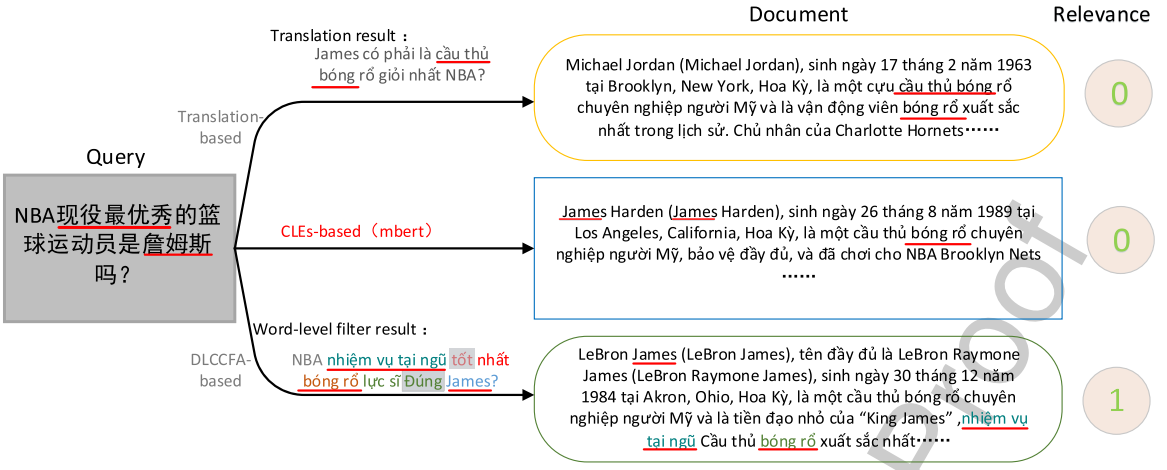

Fig. 1. The top-1 retrieval results of different methods under the same query in Zh-Vi dataset. The words underlined in red color represent relevant information in the corresponding the query and the documents. The part highlighted in grey color represents the semantic noise after word-level coarse-grained filtering. The number in the most right side indicates the relevance ranking of different methods for the query.

enhance cross-lingual text matching. To the best of our knowledge, our model is the first to utilize both word-level and sentence-level semantic alignments for the precise cross-language query representation learning, leading to better query-document matching.

– We introduce a translation probability attention (TPA) mechanism and a bilingual cross attention mechanism to generate cross-language query representation in a context-aware manner.
– Extensive experiments are conducted on four real-world CLTM datasets covering three high-resource languages (*i.e.,* English, Chinese and French) and three low-resource languages (*i.e.,* Swahili, Tagalog, Vietnamese). The experimental results suggest that the proposed DLCCFA achieves promising performance gain against mainstream CLTM methods. Further analysis is also conducted to illustrate the effectiveness of each design choice.

The remainder of this paper is organized as follows. Section 3 is devoted to related work. Section 4 introduces the details of our proposed method. In Section 5, we elaborate the experimental setup, dataset setup, quantitative analysis, and qualitative analysis. In Section 6, we draw our conclusion finally.

3. Related work

Cross-language text matching has become an increasingly important task in cross-language information retrieval (CLIR) and cross-lingual answer selection. Recently, many deep neural models have been widely used in TM and shown promising results on monolingual datasets. However, since CLTM does not have a large amount of annotated data like monolingual matching model, previous methods that directly model end-to-end CLTM are expensive [18]. Scholars conducted a series of studies and discussions on how to build a communication bridge between a language pair for implementing cross-language sequence matching.

To date, the task of CLTM has achieved remarkable progress. Previous works in this task could be classified into three dimensions, that is, (1) Translation-based CLTM models, (2) Cross-Lingual Embeddings (CLEs)-based CLTM models and (3) Other Existing CLTM Models.

3.1. Translation-based CLTM models

A traditional cross-lingual text matching algorithm involves a pipeline of two components: machine translation(MT) and monolingual matching model [9, 19, 20]. These approaches may be further divided into the query translation [4–6, 21], document translation [7, 8], or both query and document translation approach [9]. In order to get the selection of proper translation words, different translation techniques can be used. Among them, word-alignment based probabilistic translation algorithm, is still the most reliably used translation techniques. Zbib et al. [22] presented an effective Neural Network Lexical Translation model for low-resources CLTM that uses source

context and character-level encodings of the input. A relevance-based NMT model was designed by [23] using a multi-task learning frame for the CLTM task, and the structure with NMT has found to deliver reasonable performance.

Translation-based CLTM model is a pipeline structure, which is susceptible to the accumulation of translation errors, especially for resource-lean CLTM settings. The accumulation of MT errors will have a greater impact on subsequent matching, and even lead to relevance prediction failures. Therefore, for CLTM in low-resource languages, the above-mentioned MT-based methods are far beyond the perfect [24].

3.2. Cross-Lingual embeddings (CLEs)-based CLTM models

Following a broad use of monolingual dense representation pre-training methods for text matching (*e.g.,* the tasks rely on word2vec and GloVe [13, 14] to pre-train word embedding, and directly improve the performance of the relevance matching models [14, 25]), extensive efforts have been made toward developing dense representations to support cross- and multi-languages. Ivan Vulić et al. [15] presented a comprehensive typology of generating cross-language word embedding (CLE) based on randomly shuffle parallel corpus and word2vec. Through the above-mentioned process, various languages representations were designed to learn CLEs in a shared space such that cross-language semantic alignment can be solved regardless of the language. Bonab et al. [26] proposed a cross-lingual embedding (CLE) method, called Smart Shuffling, which draws from statistical word alignment approaches to leverage dictionaries, deriving a novel and effective cross-language word embedding for CLTM.

In addition, there are also many applications of pre-trained cross-language embeddings with the above methods. Vulić et al. [15] proposed an unsupervised semantic ranking method based on cosine similarity. This method was the earliest application of cross-language embeddings in CLTM task. Litschko et al. [16] extended the framework and presented a completely unsupervised cross-language information retrieval framework that does not need to use any bilingual data. It leveraged off-the-shelf pre-trained CLEs to combine query translation and semantic space rankings. Zhao et al. [27] designed a weakly supervised neural model which does not require relevance annotations, instead it is trained on parallel machine translation data as weak supervision.

More recently, the use of pre-trained language models (LMs) based on transformer neural networks (*e.g.,* BERT) significantly improves the accuracy of matching [28–30]. Building on that idea, Ruder et al. [31] surveyed several other recent studies on pre-trained cross-lingual representation learning, some of these even extended the LMs for encoding many (over 100) languages into a shared multilingual semantics space via jointly learning (*e.g.,* mBERT [17], XLM [32], and XLM-R [33]). XLM-R [33] improved upon XLM by incorporating more training data and languages, including more low-resource languages. In CLTM experiments [26], the XLM-R does not perform better than XLM. These neural text representation approaches was treated as translation resources for performing CLTM query translation. Surprisingly, these approaches fail to match the performance of a statistical machine translation (SMT) system for query translation [26].

However, these cross-lingual word embeddings generally need to be pre-trained on large scale corpora using co-occurrence statistics. In addition, in the case of low resources, the lack of high-quality bilingual dictionaries often results in poor word embedding [34].

3.3. Other existing CLTM models

Vilares et al. [35] analyzed the impact of misspelled queries on CLTM model and presented a Tolerant CLTM method that is able to operate with such queries. Li and Cheng [36] presented to learn task-specific text representation based on adversarial learning, which seeks a language-invariant and task-specific representation in the embedding space. To support cross-lingual query-document matching research, Sasaki et al. [18] constructed a large-scale dataset derived from Wikipedia comparable corpora, and presented a simple neural learning-to-rank model. However, we found that these neural CLTM models did not consider how to bridge the semantic gap across languages, apparently not the ideal solution.

4. The proposed DLCCFA model

In this section, details of the proposed DLCCFA are presented. Figure 2 illustrates the architecture of **DLCCFA** . Specifically, it consists of

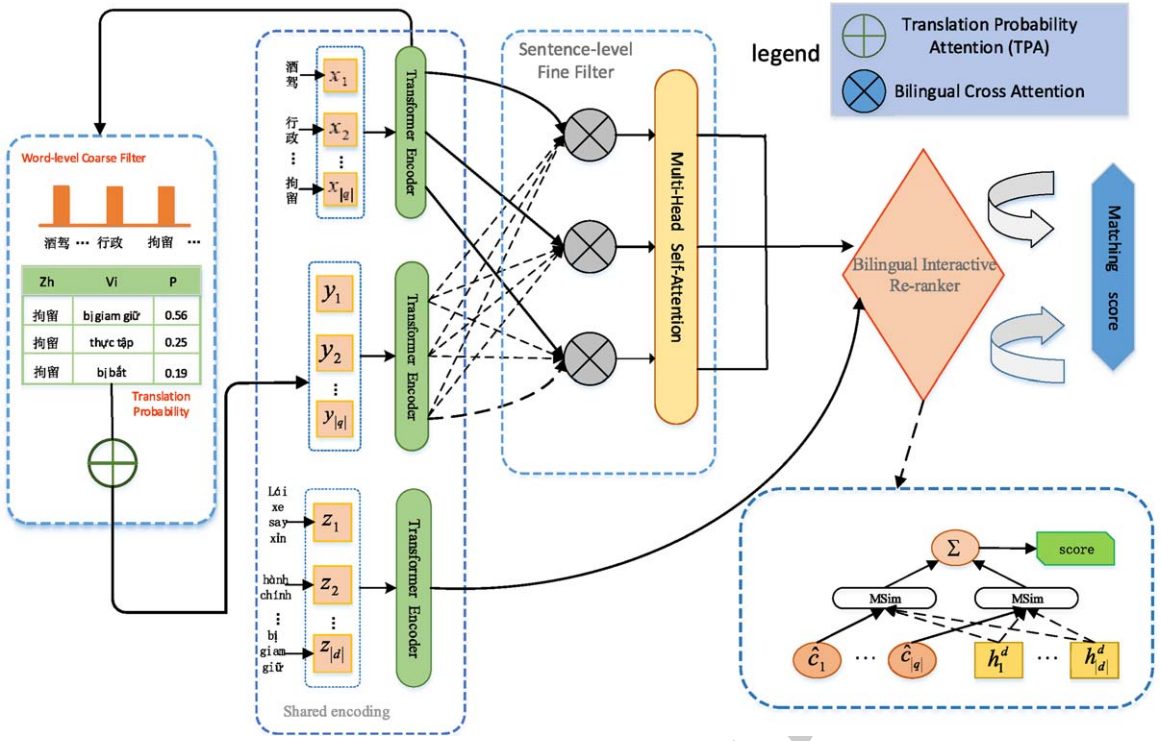

Fig. 2. The framework of our proposed cross-language deep matching model based on *Dual-Level Collaborative Coarse-to-Fine Filter Alignment Network*(**DLCCFA**), which consists of four parts: Word-level Coarse-grained Filter, Shared Encoder, Sentence-level Fine-grained Filter, and Bilingual Interactive Re-ranker.

four major components: (a) Shared Encoder, (b) Word-level Coarse-grained Filter, (c) Sentence-level Fine-grained Filter, and (d) Bilingual Interactive Re-ranker. These four parts are sequentially performed so they are dependent until the whole model achieves its optimal state.

4.1. Word embedding

Given a source language query q and a target language document d, where $|q|$ and $|d|$ are the number of words in q and d respectively. Each word in q or d is represented as a n-dimensional vector. This representation way (*i.e.,* word embedding [13, 14]) can be formulated as:

$$\mathbf{Q} = \left[\mathbf{x}_1 \cdots \mathbf{x}_{|q|}\right] = \left[E_q\left(q_1\right); E_q\left(q_2\right); \cdots; E_q\left(q_{|q|}\right)\right]$$
$$\mathbf{D} = \left[\mathbf{z}_1 \cdots \mathbf{z}_{|d|}\right] = \left[E_d\left(d_1\right); E_d\left(d_2\right); \cdots; E_d\left(d_{|d|}\right)\right] \quad (1)$$

where $\mathbf{Q} \in \mathbb{R}^{n\times|q|}$ and $\mathbf{D} \in \mathbb{R}^{n\times|d|}$ are query and document feature matrix respectively, q_i and d_i refer to i-th word in q and d respectively, $E_q(\cdot)$ and $E_d(\cdot)$ are word embedding function which transform each word in q and d to a dense n-dimensional vector respectively.

4.2. Shared encoder

To bridge the semantic gap between two different languages, we choose to share a single text encoder for both source language query and target language documents. Specifically, a Transformer Encoder [37] with a stack of N identical layers is used to generate the word contextual representations. Each identical layer is divided into two sub-layers. The first sub-layer is a multi-head self-attention mechanism, and the other is a fully connected feedforward network. A residual connection is added outside the each of sub-layers, followed by layer normalization. According to the formula:

$$\mathbf{H}^q = Transformer\_Encoder(\mathbf{Q})$$
$$\mathbf{H}^d = Transformer\_Encoder(\mathbf{D}) \quad (2)$$

where $\mathbf{H}^q = [\mathbf{h}_1^q \cdots \mathbf{h}_{|q|}^q]$ and $\mathbf{H}^d = [\mathbf{h}_1^d \cdots \mathbf{h}_{|d|}^d]$, $\mathbf{h}_i^q$ and $\mathbf{h}_i^d$ are contextual word representation for i-th word in q and d respectively. Then, we add a normalization layer such that $\|\mathbf{h}_i^q\|_2 = 1$ and $\|\mathbf{h}_i^d\|_2 = 1$. In this way, when inner-product of any two hidden representations is performed, the value falls in the range of $[-1, 1]$ (*i.e.,* equivalent to their cosine similarity).

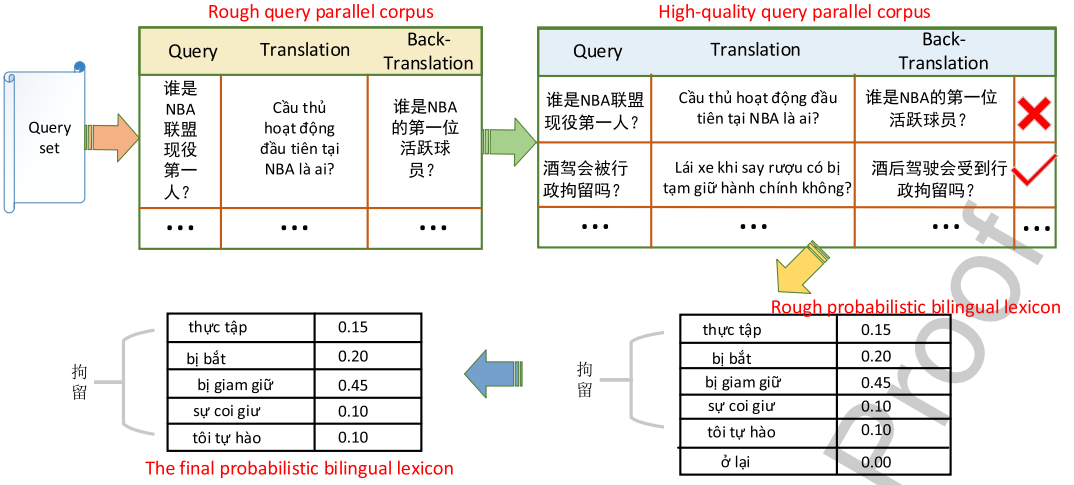

Fig. 3. The process of constructing a Chinese-Vietnamese probability bilingual dictionary (take the Zh-Vi dataset as an example). Where, *the orange arrow* indicates the process of translation and back translation. *The green arrow* indicates the operation of manual screening and correction. *The yellow arrow* indicates the operation of extracting word alignment using the fast-align tool. *The blue arrows* indicate the process of filtering word pairs and use the maximum likelihood to find the probability. *The red check mark* indicates the parallel sentence pair we think is correct, otherwise it will be discarded.

4.3. Word-level coarse-grained filter

We firstly perform word alignment for each token in the query via a word-level coarse-grained filter. Specifically, we utilize a *Probabilistic Bilingual Lexicon* and a *Translation Probability Attention (TPA)* mechanism here.

Probabilistic Bilingual Lexicon. As being widely adopted by many translation-based approaches, a bilingual lexicon would be used to derive relevant words in the target language to express the query. There have been various solutions which can be used to get the Probabilistic Bilingual Lexicon. For instance, the word alignment probabilities can be learned from either bilingual corpora [38] or monolingual corpora [39]. As for high-resource languages, we can apply the method described in [40] to construct high-quality probabilistic bilingual lexicon. This method is generally divided into two steps: 1) we firstly extract word alignments on the bilingual parallel corpus using the fast-align tool provided in [40]. Through maximum likelihood estimation, this step results in the translation probability for translating word t_1 in source language into word t_2 in target language ($P(t_1 \Rightarrow t_2)$), and vice versa; 2) Both source-to-target and target-to-source translation directions are considered for better alignment. Specifically, we calculate the average of probabilities in both directions (*i.e.,* the lexicon translation probability $P(t_1 \Rightarrow t_2)$ and $P(t_2 \Rightarrow t_1)$).

On the contrary, due to lack of sufficient number of parallel documents, the above method is difficult for low-resource languages. In this scenario, we resort to using Google translator[1] to help us construct some reliable parallel corpora. Specifically, we first use the Google translator to translate each query in the CLTM dataset into the target language. Then, this target language translation is back translated again into the source language. Lastly, high-quality sentence pairs are selected from the resultant parallel query sentence pairs. Finally, we apply the fast-align tool to construct high-quality probabilistic bilingual lexicon. Furthermore, we remove the word alignment pairs (t_1, t_2) such that $P(t_1 \Rightarrow t_2) \leq 0.05$, and obtain the final probabilistic bilingual lexicon [41]. This process is illustrated in Fig. 3.

Translation Probability Attention (TPA). With the constructed probabilistic bilingual lexicon, we can choose the most probable word alignment for each token in the query. Note that the meaning of a word is largely influenced by the contextual information [41]. Therefore, the word-level translation described above may not fully preserve the semantics of the query. In other words, the translation candidate should be selected by taking the contextual semantics of the whole query into account. Hence, we introduce *Translation Probability Attention (TPA)* mechanism to adaptively choose the correct translation in a context-aware fashion.

In particular, for each word q_i in the query, we can extract top-k candidate translations based on the probabilistic bilingual lexicon. Let $\mathbf{K}_i \in \mathbb{R}^{k \times n}$ denote the corresponding embedding matrix of these k candidate

[1] https://translate.google.cn/

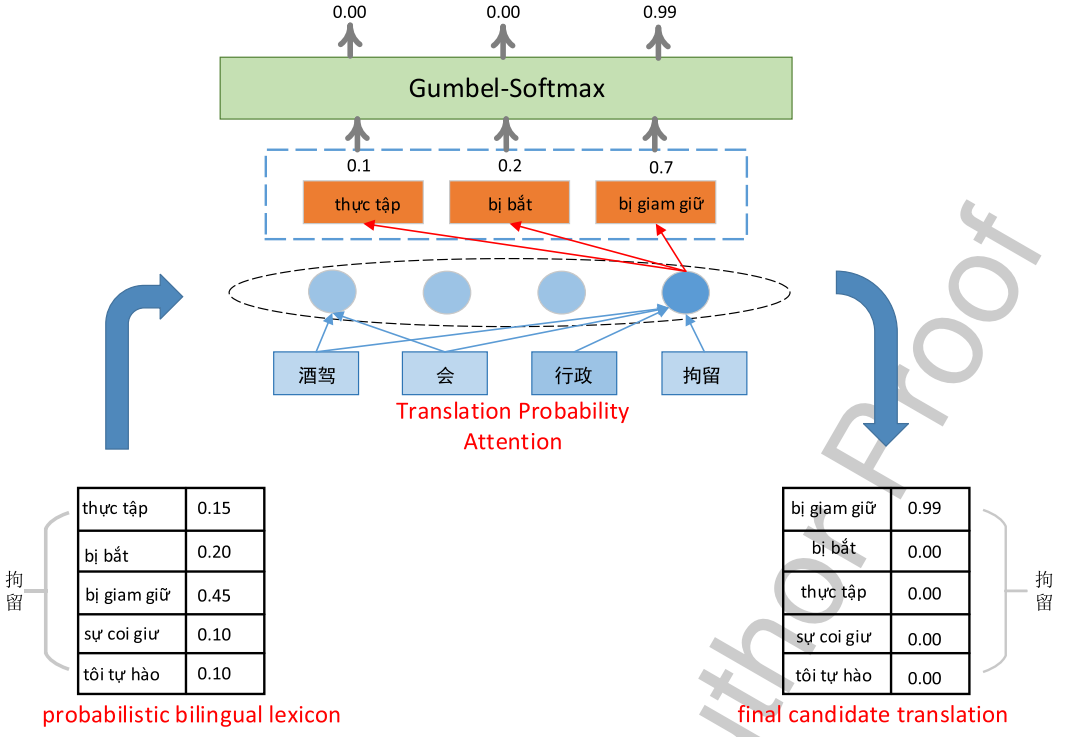

Fig. 4. The framework of our proposed Translation Probability Attention (TPA).

words, we can simply calculate their translation probability as follows:

$$\mathbf{p}_i = \text{softmax}\left(\mathbf{h}_i^q \mathbf{K}_i^T\right) \tag{3}$$

where $\mathbf{p}_i \in \mathbb{R}^k$ is translation probabilities of these k candidate words with respect to word q_i, $\mathbf{h}_i^q$ is the hidden state vector derived by the encoder. Then, we can perform word alignment by choosing the candidate with the largest probability value. It is obvious that this process neglects the semantics of the whole query. To better select candidate words based on the semantics of the query, we further choose multi-head self-attention to enrich this probability calculation. For each head l with projection matrices $\mathbf{W}_j^1, \mathbf{W}_j^2 \in \mathbb{R}^{n \times s}$, we can calculate the salience scores for these candidate words by considering each query word q_j as follows:

$$\mathbf{e}_{jl}^i = \text{softmax}\left(\frac{\mathbf{h}_j^q \mathbf{W}_l^1 (\mathbf{K}_i \mathbf{W}_l^2)^T}{\sqrt{n}}\right) \tag{4}$$

where n is the scaling factor and $\mathbf{e}_{jl}^i \in \mathbb{R}^k$ is the translation probability vector derived based on word q_j. By using m attentions with different projection matrices, we can derive the final translation probability as follows:

$$\mathbf{p}_i = TPA(q_i) = \frac{1}{m \times |q|} \sum_{j=1}^{|q|} \sum_{l=1}^{m} \mathbf{e}_{jl}^i \tag{5}$$

As shown in Equation 5, the translation probability depends not only on the target query word, but also on the other words of the query. That is, empowered by the multi-head mechanism, we can enable word alignment in a context-aware fashion.

Afterwards, it is straightforward to take the translation word with the largest probability for each query word as the alignment. However, this process is discrete in nature, which is infeasible to enable model training via back-propagation. To tackle this issue, we integrate a Gumbel-Softmax [42] layer into TPA to ensure model training. Specifically, given i-th word of query q, the translation probability is calculated as follows:

$$v_{i,j} = \frac{\exp\left(\left(\log\left(\mathbf{p}_i[j]\right) + g_j\right)/\tau\right)}{\sum_{y=1}^{k} \exp\left(\left(\log\left(\mathbf{p}_i[y]\right) + g_y\right)/\tau\right)} \tag{6}$$

where $\mathbf{v}_i \in \mathbb{R}^k$ is analogous to the one-hot vector, τ is the temperature hyperparameter. When τ approaches 0, $\mathbf{v}_i$ approximates a one-hot vector. At last, we can obtain the representation of the word alignment for i-th query word as follows:

$$\mathbf{q}_i^a = \mathbf{v}_i \mathbf{K}_i. \tag{7}$$

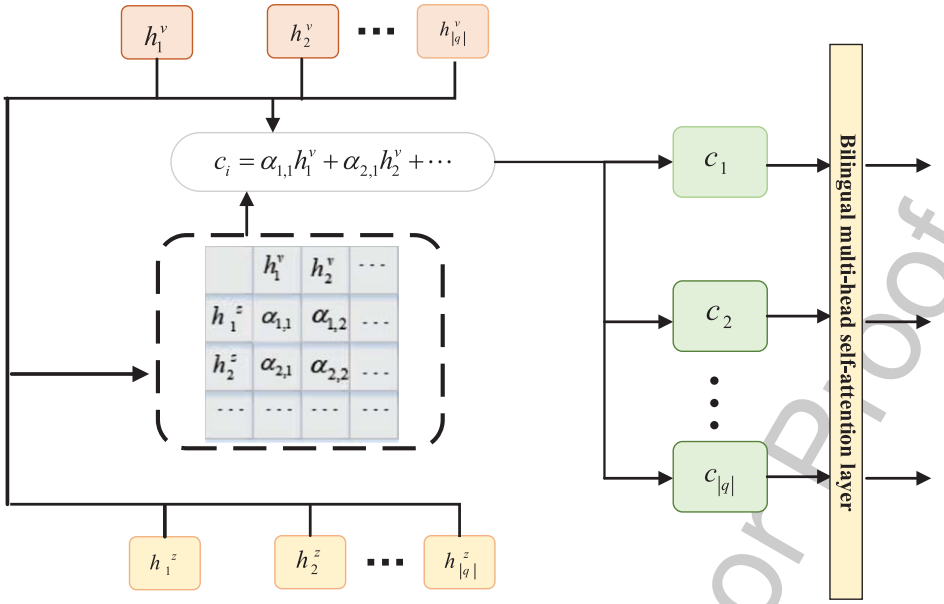

Fig. 5. The overall framework of our proposed Sentence-level Fine Filter.

Then, by forming all $\mathbf{q}_i^a$ as a matrix $\mathbf{Q}^a = [\mathbf{q}_1^a \cdots \mathbf{q}_{|q|}^a]$, we consider it as the query auxiliary sentence. By applying the same encoder shared in Equation 2, we generate the contextual representation of each word in the query auxiliary sentence as follows:

$$\mathbf{H}^a = Transformer\_Encoder\,(\mathbf{Q}^a) \tag{8}$$

where $\mathbf{H}^a = [\mathbf{h}_1^a \cdots \mathbf{h}_{|q|}^a]$ are the contextual representations. Similarly, a L2 normalization layer is utilized over $\mathbf{H}^a$.

4.4. Sentence-level fine-grained filter

There would be potentially many semantic noises in the representation matrix $\mathbf{H}^a$ via the word-level coarse filter. As illustrated in Fig. 5, we further introduce a *Sentence-level Fine-Grained Filter* to achieve fine filtering and cross-language representation of the query. The proposed fine-grained filter consists of two stages: 1) Bilingual Cross Attention Layer to filter the semantic noise in the query auxiliary sentence; and 2) Self-Attention Layer to generate cross-language representation of the query.

Bilingual Cross Attention Layer. Here, our aim is to represent each query word with the contextual represenations of the query auxiliary sentence. Specifically, for each query word, an attention mechanism is utilized to calculate the relevance of each word in the query auxiliary sentence:

$$m_{i,j} = \tanh\left(\mathbf{h}_i^q \mathbf{W} \mathbf{h}_j^a + b\right) \tag{9}$$

$$\alpha_{i,j} = \frac{e^{m_{i,j}}}{\sum_{k=1}^{|q|} e^{m_{i,k}}} \tag{10}$$

where $\mathbf{W} \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}$ are the parameters to be learned. Then, the representation of each query word is derived as follows:

$$\mathbf{c}_i = \sum_{j=1}^{|q|} \alpha_{i,j} \bullet \mathbf{h}_j^a \tag{11}$$

It is obvious that we represent each word in the query by focusing more on parts of the query auxiliary sentence that are most relevant, while reducing irrelevant semantic noise.

Self-Attention Layer. Similarly, we can form a matrix $\mathbf{C} = [\mathbf{c}_1 \cdots \mathbf{c}_{|q|}]$ by following the above procedure. We further perform feature extraction by using the multi-head self-attention mechanism [37] as follows:

$$\hat{\mathbf{C}} = Self\_Attention\,(\mathbf{C}) \tag{12}$$

where $\hat{\mathbf{C}} = [\hat{\mathbf{c}}_1 \cdots \hat{\mathbf{c}}_{|q|}]$ and $\hat{\mathbf{c}}_i$ is the cross-lingual contextual representation of the i-th token of queries.

4.5. Bilingual interactive re-ranker

With both $\hat{\mathbf{C}}, \mathbf{H}^d$, the relevance score of d to q, denoted as $s_{q,d}$, is estimated via a bilingual interactive

re-ranker. Specifically, we calculate the sum of maximum similarity against each query word [30] as follows:

$$s_{q,d} = \sum_{i \in |q|} \max_{j \in |d|} \left(\hat{\mathbf{c}}_i^T \cdot \mathbf{h}_j^d \right) \tag{13}$$

4.6. Model optimization

The model parameters θ of DLCCFA include the feature embeddings and $\{Transformer\_Encoder(\cdot), \mathbf{W}_*^1, \mathbf{W}_*^2, \mathbf{W}, \mathbf{b}, Self\_Attention(\cdot)\}$. In the training phase, we minimize pairwise ranking loss to guide parameter learning, which is widely used for learning-to-rank [18], defined as follows:

$$L\left(q, d^+, d^-; \theta\right) = \max\left\{0, 1 - \left(s\left(q, d^+\right) - s\left(q, d^-\right)\right)\right\} \tag{14}$$

where d^+ and d^- are relevant and non-relevant document respectively.

5. Experimental results and discussion

5.1. Zh-Vi CLTM datasets construction

We construct a Zh-Vi (*i.e.,* Chinese-Vietnamese) CLTM dataset from Wikipedia. The core idea is to extract an English sentence as query, and label foreign-document pages via Wikidata links as relevant. We apply the same techniques [18] to create Zh-Vi CLTM dataset. Figure 1 illustrates the whole construction process.

Specifically, we first download the English Wikipedia dump and then extract the first sentence of each article as English queries. Then the cross-lingual linked Vietnamese documents from the same article are taken as the relevant documents. From a practical point of view, the first sentence of an English article is usually a summary of the article. To prevent the simplification of the task, the core subject terms from the queries are removed. Afterwards, we use the Google translator to translate the query sentences into Chinese, and further generate new English query sentences with the back-translation. Finally, based on the translation results in both directions, we manually select high-quality Chinese queries.

We then truncate each document to retain only the first 200 tokens of each article. The triples of the Zh-Vi CLTM dataset will be obtained via a series of pre-processing. Here each triplet is in the form of (q, d, r), where q is Chinese query and d is denoted as Vietnamese document. r is denoted as relevance judgment, in which $r \in \{0, 1\}$ represents relevance groundtruth.

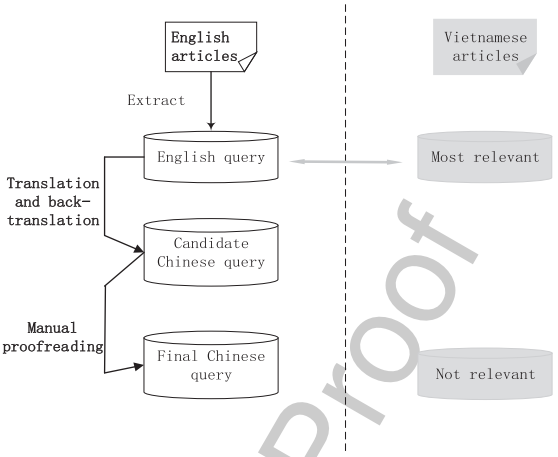

Fig. 6. Zh-Vi CLTM dataset construction process: we first extract the English query from an English article. Then by using the inter-language link, we obtain the most relevant Vietnamese document. All other articles are not relevant. Finally, based on the translation results in both directions, we manually selected high-quality Chinese queries.

5.2. CLTM datasets and evaluation metrics

To validate the proposed DLCCFA , we conduct experiments on four CLTM datasets: En-Fr (*i.e.,* English-French) dataset, En-Tl (*i.e.,* English-Tagalog) dataset, En-Sw (*i.e.,* English-Swahili) dataset, and Zh-Vi(*i.e.,* Chinese-Vietnamese) dataset built by ourselves. En-Fr, En-Tl and En-Sw datasets is derived from a large-scale CLTM data from Wikipedia [18]. The first language is the query-side language and the second is the language of the document collection. The length of all queries and documents is limited to 30 and 200, respectively. For each query in the training set, we pick candidate documents in which only one document is positive. As for testing set, the number of candidate documents for each query is in range of $15 - 40$ for a query, except for Zh-Vi CLTM dataset where the number of test documents corresponding to a query is in the range of $200 - 350$. As mentioned above, each instance (*e.g.,* Figure 1) in a CLTM dataset is formed as (q, d, r), where q is the query and d is denoted as targe-language document. r is the relevance judgment, in which $r \in \{0, 1\}$ represents relevance groundtruth. The statistics of the four datasets after pre-processing are shown in Table 1.

To perform a fair evaluation, five widely used metrics ([43, 44]), *i.e., MRR*(Mean Reciprocal Rank), $P@k$ (Precision at k), $R@k$ (Recall at k), MAP (Mean Average Precision) and $NDCG@k$ (Normalized Discounted Cumulative Gain), are applied to measure the performance for cross-lingual match-

Table 1
The statistics of the four CLTM datasets

Language pair	Train size		Test size	
	Query size	Document size	Query size	Document size
Zh-Vi	110k	220k	5.4k	1,350k
En-Fr	370k	1,850k	54k	1,512k
En-Tl	16.3k	81.5k	2.3k	64.4k
En-Sw	7.2k	36k	1.1k	38.5k

ing. In the following, we describe these metrics in detail.

For each query q, Let the number of relevant documents be G_q, and the number of irrelevant documents be Y_q. Both $P@k$ and $R@k$ are calculated as follows:

$$\begin{aligned} P@k &= \frac{Y_q}{k} \\ R@k &= \frac{Y_q}{G_q} \end{aligned} \tag{15}$$

Let the ranking positions of the real relevant documents be $k_1, k_2, \cdots, k_r$, where r is the number of all positive documents in the entire list. The MAP score can be calculated as follow:,

$$MAP = \frac{\sum_{i=1}^{r} P@k_i}{r} \tag{16}$$

When we only consider the top ground-truth positive document k_1, MRR can be derived as follows:

$$MRR = P@k_1 \tag{17}$$

We also employ the widely used $NDCG$ (normalized discounted cumulative gain) as evaluation metrics. Given a query q, let δ_i be the relevance groundtruth indicator of the i-th document ranked by the model. The $NDCG@k$ is calculated as follows:

$$\begin{aligned} DCG@k &= \sum_{i=1}^{k} \frac{2^{\beta_i} - 1}{\log_2 i + 1} \\ NDCG@k &= \frac{1}{Z} DCG@k \end{aligned} \tag{18}$$

where Z is the normalization factor being equivalent to $NCG@k$ with the ideal ranking.

5.3. Implementation details

We consider Chinese [Zh], French [Fr] as high-resource languages and Tagalog [Tl], Swahili [Sw], Vietnamese [Vi] as low-resource languages. The dimension of word embedding is set to be 300 for each language. We train all DLCCFA model with learning rate $5e-5$ with a batch size 32. We train our model using transformer encoder with 6 layers and 512 as the hidden dimension. The number of candidate translations is set to 5 (*i.e.,* $k = 5$). We adopt Adam [45] for optimization. The maximum number of training epochs is set to be 5 on En-Fr dataset and 80 epochs on other datsets respectively. The temperature τ in Gumbel_Softmax is initialized as 0.5, and updated according to exponential decay:

$$new_\tau = initial_\tau \times \gamma^{epoch} \tag{19}$$

where γ denotes the decay rate, which is set to 0.9.

5.4. Baseline methods

In this section, we compare the proposed DLCCFA with uptodate SOTA alternatives:

Query Translation based CLTM (CLTM-TQ): It firstly translates a query into target language via a Transformer-based model [37] and then performs monolingual information retrieval. A similar apporach is also utilized in [9].

Document Translation based CLTM (CLTM-TD): Similar to CLTM-TQ, this method first uses Transformer [37] to translate documents into source language and performs monolingual retrieval.

Cosine Model based CLTM (CLTM-S-COS): It refers to the deep learning method [18], which simultaneously builds a CNN model to extract features from the query and the document. Here CLTM-S-COS uses the cosine similarity for calculating the matching score.

Deep Model based CLTM (CLTM-DEEP): In this model, it also utilizes CNN to extract contextual representations for both the query and the document. Then, a deep network is devised to calculate the matching score between the query representation $\hat{\mathbf{q}}$ and the document representation $\hat{\mathbf{d}}$ as follows:

$$S = \tanh(\mathbf{O} \bullet relu(\mathbf{W}_d \bullet [\hat{\mathbf{q}}, \hat{\mathbf{d}}])) \tag{20}$$

where $\mathbf{W}_d \in \mathbb{R}^{h \times 2n}$ and $\mathbf{O} \in \mathbb{R}^{1 \times h}$ are learnable parameters. The deep model is divided into CLTM-S-DEEP300, CLTM-S-DEEP400, CLTM-S-DEEP500 according to depth and the size of hidden layers. The hyper-parameter settings of the entire model refer to [18].

Table 2
Results of three CLMT datasets.

Method	Dataset(En-Tl)			Dataset(En-Sw)			Dataset(En-Fr)		
	P@1	P@2	MAP	P@1	P@2	MAP	P@1	P@2	MAP
CLTM-TQ	0.426	0.292	0.557	0.411	0.281	0.524	0.523	0.358	0.635
CLTM-TD	0.412	0.277	0.521	0.394	0.267	0.503	0.517	0.352	0.613
CLTM-S-COS	0.510	0.351	0.680	0.510	0.346	0.670	0.550	0.370	0.700
CLTM-S-DEEP300	0.420	0.286	0.570	0.500	0.328	0.650	0.740	0.446	0.840
CLTM-S-DEEP400	0.490	0.319	0.630	0.510	0.337	0.660	0.750	0.449	0.850
CLTM-S-DEEP500	0.510	0.330	0.640	0.530	0.349	0.680	0.760	0.454	0.850
mBERT	0.515	0.349	0.661	**0.538**	0.349	0.679	**0.764**	0.451	**0.855**
XLM	0.155	0.159	0.357	0.115	0.119	0.307	0.249	0.240	0.403
XLMR	0.202	0.182	0.401	0.264	0.209	0.451	0.642	0.416	0.778
LASER	0.361	0.224	0.464	0.369	0.235	0.496	0.366	0.237	0.493
DLCCA	**0.553**	**0.382**	**0.711**	0.536	**0.376**	**0.697**	0.761	**0.455**	**0.855**

mBERT-based CLTM (mBERT): [44] extend the ranking model via using multilingual BERT[2]. This solution encodes a query-document pair with mBERT, and documents are reranked based on the output scores from the mBERT-based re-ranker model.

XLM-based CLTM (XLM): We use the multilingual pre-trained model XLM[3] to encode the query-document pairs, and build a cross-language reranker model based on XLM.

XLMR-based CLTM (XLMR): XLMR[4] model proposed in [39] for cross-language sentence retrieval. It is an upgraded pre-trained language model of XLM, including more than 100 languages. On Tatoeba [39], XLMR-based sentence retrieval accuracy are the previous SOTA.

LARSE-based CLTM (LASER): It is a state-of-the-art supervised approach that exploits LASER[5] to achieve cross-language sentence retrieval [46].

5.4.1. Performance comparison

Effectiveness analysis on low-resource language. For two low-resource languages, a summary of results of all models are reported in Table 2. Here, we have the following observations:

– The conventional MT-based pipeline techniques like CLTM-TQ and CLTM-TD generally perform worse than deep learning methods on both datasets. It is clear to see that both CLTM-TQ and CLTM-TD are essentially limited by the quality of machine translation as the pipeline architecture is easy to accumulate translation errors. However, the two traditional translation-based methods have better accuracy than some pre-trained language models. This phenomenon is more intense in low-resource languages. It may be that the pre-trained low-resource language data is insufficient.
– It is clear that deep learning methods perform better than pipeline techniques on both low-resource datasets. This explains the effectiveness of applying deep neural network to capture the latent semantics of different texts. Among these models, we observe the performance of the cosine model and the deep models are very close. The reason might be that deep models with larger parameters may require more sufficient training data. The MAP value of CLTM-S-DEEP500 model is about $2.5-15\%$ higher than the conventional MT-based pipeline techniques.
– Since multilingual pre-training models can project different languages into the same hidden space, it has gradually become the dominating technique for multilingual semantic alignment. It is worthwhile to highlight that the performance order is: mBERT > LASER > XLMR > XLM. This suggests that the alignment effect of mBERT is better than other LMs in the case of low resource settings. In addition, XLMR and LASER have achieved the state-of-the-art results for cross-language sentence retrieval. However, they are not very effective on the two low-resource datasets. This is reasonable since there is a big gap in length between query and document and the alignment performance of the pre-trained language model on low-resource languages is largely affected by this discrepancy.
– All in all, our proposed DLCCFA achieves the best performance against all the methods on En-Tl dataset. For En-Sw CLTM dataset, DLCCFA outperforms CLTM-TQ and CLTM-TD on all the metrics. Compared to CLTM-S-COS and CLTM-S-DEEP500, DLCCFA outperforms or achieves comparable performance on most metrics. Especially on MAP, DLCCFA has 2.7% and 1.7% improvements, respectively. As for the baselines

[2]We used BERT-Base, Multilingual Cased.
[3]We used xlm-mlm-100-1280.
[4]We used xlm-roberta-base.
[5]https://github.com/facebookresearch/LASER.

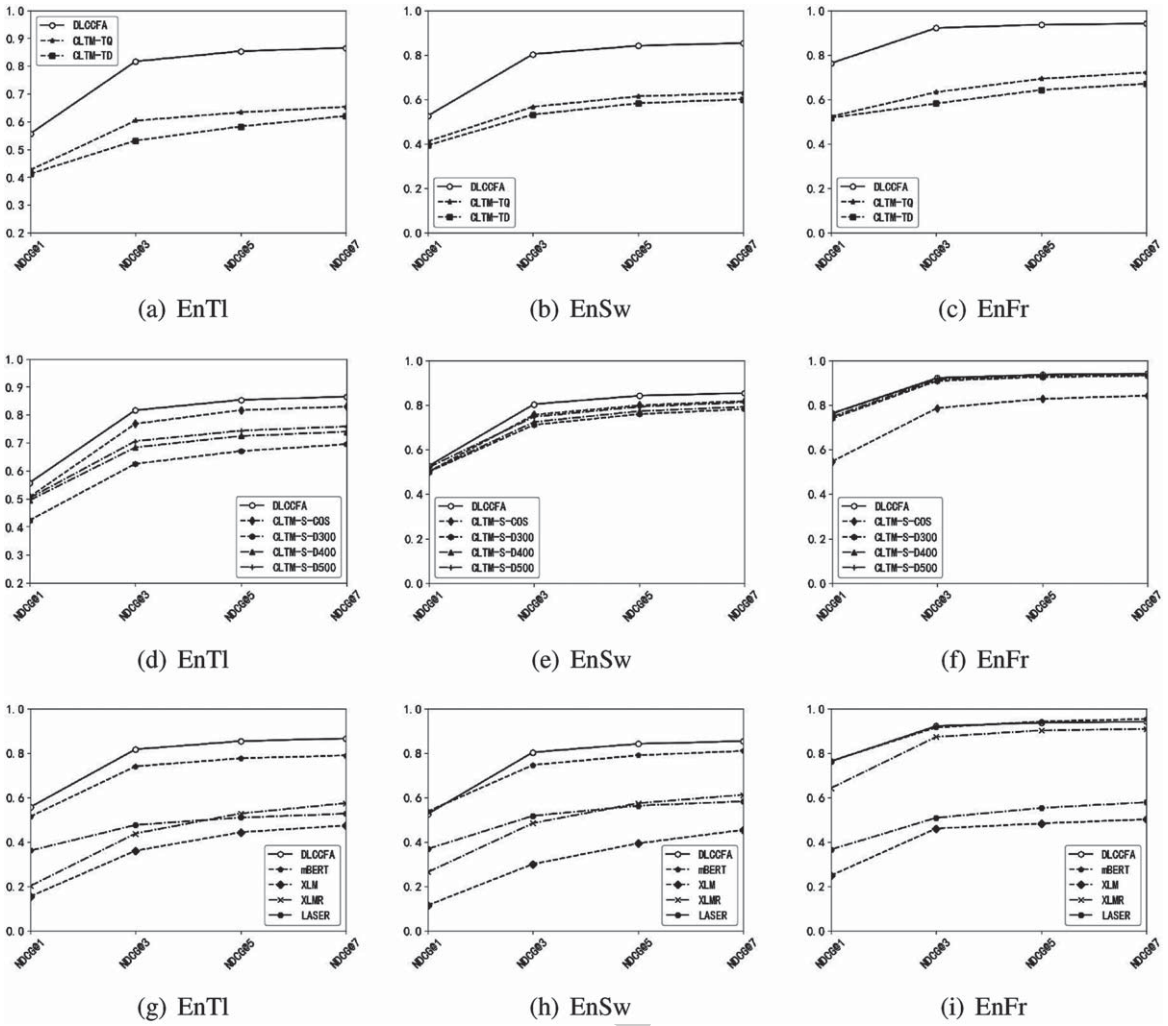

Fig. 7. The $NDCG@K$ curves on three datasets. The first row shows the comparison result with the traditional method. The second row represents the comparison result with the CNN-based deep learning methods. The third row represents the comparison result with pre-trained-based the language model.

based on pre-training CLTM, our model achieves comparable or surpass mBERT-based metheds on all the metrics. In sum, the experimental results demonstrate the effectiveness of utilizing both word-level and sentence-level semantic alignments.

Effectiveness analysis on high-resource language. For En-Fr CLTM dataset, a summary of results is reported in Table 2(En-Fr dataset). Unlike low-resource languages, the deep models outperform the cosine model on all the metrics, suggesting that deep networks can require sufficient data in learning more expressive representations for both query and document. Except for mBERT-based model, DLCCFA outperforms all baseline models on all the metrics. It suggests that the alignment made by our proposed DLCCFA is more accurate in the case of low resources. At the same time, this also proves the generalization of the model for high- and low-resource languages.

Additional analysis. We further vary k values of $NDCG@k$ to verify the performance patterns of DLCCFA in different language pairs. The curives are plotted in Figure 7. We can see that comparing with all baselines (*i.e.,* MT-based CLTM, CNN-based CLTM and pre-trained-based models), DLCCFA obtains a better performance. This observations on three high- and low-resource datasets are almost consistent. These experimental results further suggest that our model can not only be generalized to high-resource language pairs, but also in the case of insufficient training data for low-resource languages. It demonstrates the effective-

Table 3
Ablation study results obtained on Zh-Vi CLTM dataset.

Method	Dataset(zh-vi)			
	MRR	R@3	R@5	R@10
TE+rank	0.3497	0.3983	0.4959	0.6111
TE+TPA+rank	**0.4472**	**0.5076**	**0.6054**	**0.7184**

ness of integrating both word-level and sentence-level semantic alignments for cross-language representation learning in CLTM.

5.5. Model analysis

In this section, we examine the impact of each design choice made in DLCCFA by a series of ablation studies on the Zh-Vi CLTM dataset. For a fair comparison, all the studies take Transformer Encoder as the encoder, and use MRR(Mean Reciprocal Rank) and $R@k$(Recall at k) as evaluation metrics.

5.5.1. The influence of word-level coarse-grained filter

transformer encoder + ranking module (TE+Rank): In this method, we use transformer encoder to directly model and calculate similarity scores between queries and documents, where the relevance between two languages are merely learned through the shared encoder.

transformer encoder + word-level coarse-grained filter + ranking module (TE+TPA+Rank): According to the proposed word-level coarse-grained filter, the k translation candidates of each token in the query are first filtered by the probabilistic bilingual lexicon. Then, a TPA mechanism is also introduced to calculate a translation probability of each translation candidate and to adaptively choose the correct translation in a context-aware fashion.

The experimental results are reported in Table 3. From Table 3, we can see that, TE+WLF+Rank outperforms TE+Rank on all the metrics, with an increase of 10.93% on $R@3$. Experimental results well validate our claim that word-level coarse-grained filter is effective in selecting the translation candidate by taking the contextual semantics of the whole query into account. This also verifies the importance of word-level alignment to help bridge the semantic gap across languages.

5.5.2. The influence of shared encoder

In order to reduce the amount of parameters in DLCCFA , and bridge semantic gap across the two different languages, we share a single transformer encoder for both the query and the document. We also conduct a comparative study to prove whether the shared encoder is effective to enhance the CLTM performance for DLCCFA . The results on Zh-Vi CLTM dataset are shown in Table 4. From Table 4, we can see that, when we use a shared encoder, MRR, $R@3$, $R@5$, and $R@10$ are increased by 0.59%, 0.32%, 2.22%, 3.60%, respectively. It indicates the necessity of shared encoder to extract compatible features across language.

Table 4
Whether to share encoder the comparison test results obtained on Zh-Vi CLTM dataset. Best value in each column is highlighted in bold.

Method	Dataset(zh-vi)			
	MRR	R@3	R@5	R@10
No-shared	0.4413	0.5044	0.5832	0.6824
Shared	**0.4472**	**0.5076**	**0.6054**	**0.7184**

5.5.3. The influence of sentence-level fine-grained Filter

Sentence-level fine-grained filter is one of the most important components of our model, which is used to identify the important information and obtain the cross-language representation of the query. We propose the following three different strategies:

CLTM+MTL. It refers to the multi-task method which exploits the query representation as a constraint and add the Mean Squared Error (MSE) as the loss function. The loss function is then updated as follows:

$$Loss = (1-\lambda) * Loss\left(q, d^{+}, d^{-}; \theta\right) + \lambda * MSE\left(\mathbf{Q}, \mathbf{Q}^{*}\right) \quad (21)$$

where d^{+} and d^{-} are relevant and non-relevant document respectively, and θ denotes model parameters. $\mathbf{Q}$ is the representation matrix of the query. $\mathbf{Q}^{*}$ is the representation matrix of query auxiliary sentence. λ is weight coefficient of alignment task, and λ is tuned among $\{0.5, 0.1, 0.05, 0.01\}$, in which $\lambda = 0.05$ achieve the best results.

CLTM+concat. In this method, we directly concatenate the feature representations of the corresponding token in the query and auxiliary sentence. This enables the semantic space of the two languages to be narrowed, and finally use the concatenated features to retrieve document features.

CLTM+c_att. In this strategy, we utilize sentence-level fine-grained filter to filter the semantic noise in the query auxiliary sentence and accurately align semantics of different languages, which has been introduced in detail in Section 4.4.

The results are shown in Table 5. As we can see, three different strategies will directly lead to dif-

Table 5
Results of models on Zh-Vi CLTM dataset with different strategies to obtain the cross-language representation of queries. Best performance in each metric is highlighted in boldface.

Method	Dataset(zh-vi)			
	MRR	R@3	R@5	R@10
CLTM+MTL	0.3503	0.4006	0.4962	0.6178
CLTM+concat	0.3712	0.4246	0.5184	0.6382
CLTM+c_att	**0.4472**	**0.5076**	**0.6054**	**0.7184**

Table 6
Results of models on Zh-Vi CLTM dataset with three ranking strategies. Best performance in each metric is highlighted in boldface.

Method	Dataset(zh-vi)			
	MRR	R@3	R@5	R@10
CLTM-sent_cos	0.1951	0.2044	0.2922	0.422
CLTM-sent_deep	0.2112	0.2314	0.3067	0.4558
CLTM-bi_rank	**0.4472**	**0.5076**	**0.6054**	**0.7184**

ferent retrieval performance. Where, by comparing CLTM+c_att with CLTM+MTL, it has 9.69% improvement on MRR, 10.70% improvement on $R@3$, 10.92% improvement on $R@5$, and 10.06% improvement on $R@10$. In addition, compared to CLTM+concat, CLTM+c_att outperforms it on MRR, $R@3$, $R@5$, and $R@10$. Our conjecture is that the sentence-level fine-grained filter is effective to remove semantic noise for better CLTM.

5.5.4. The influence of bilingual interactive re-ranker

We further examine the impact of different scoring functions. Without changing the other parts of our model, we also investigate three ranking strategies here.

CLTM-sent_cos. We obtain representation vectors $\hat{\mathbf{q}}$ and $\hat{\mathbf{q}}$ by average-pooling for the query q and the document d. Then, the relevance score is calculated through the cosine similarity as follows:

$$score = \cos sim(\hat{\mathbf{q}}, \hat{\mathbf{b}}) \tag{22}$$

CLTM-sent_deep. This strategy utilizes a fully connected layer on top of the concatenation of $\hat{\mathbf{q}}$ and $\hat{\mathbf{q}}$:

$$score = \tanh(\mathbf{O} \bullet relu(\mathbf{W}_e \bullet [\hat{\mathbf{q}}, \hat{\mathbf{q}}])) \tag{23}$$

where $\mathbf{W}_e \in \mathbb{R}^{h \times 2n}$ and $\mathbf{O} \in \mathbb{R}^{1 \times h}$ denotes learnable parameters.

CLTM-bi_rank. This strategy utilize Bilingual Interactive Re-ranker we devise in Section 4.5.

Table 6 reports the retrieval performance by using the three scoring strategies. Compared to CLTM-sent_cos and CLTM-sent_deep, our CLTM-bi_rank outperforms them on the all metrics. This also verifies the importance of an expressive scoring functions to re-rank documents. Ours conjecture is that, since the length difference between the query and the document is relatively large, directly using the features of the document will greatly lose the key semantic information. Note that the bilingual interactive re-ranker of DLCCFA can start from the word-level granularity and more comprehensively calculate the similarity between short queries and long documents.

In summary, this set of experimental comparisons suggests that each design choice in DLCCFA is rational to enahnce cross language query-document matching.

6. Conclusion and future work

In this paper, we address a semantic alignment problem in Cross-anguage text matching, The proposed cross-language deep matching model based on Dual-Level Collaborative Coarse-to-Fine Filter Alignment Network (DLCCFA) achieves cross-language semantic alignment for CLTM. Specifically, we first extract top-k translation candidates for each token in the query through a probabilistic bilingual lexicon. Then, we devise a Translation Probability Attention (TPA) mechanism to achieve coarse-grained word alignment for generating the corresponding query auxiliary sentence. After that, a Bilingual Cross Attention and cooperate Self Attention is introduced to filter the semantic noise in the query auxiliary sentence and accurately align semantics of both langualges. Extensive comparison experiments are conducted on CLTM datasets of four different language pairs. Experimental results show that our method achieves the state-of-the-art performances.

In our future work, we consider incorporating hashing into our method to achieve fast matching under low memory. Besides, we will also explore multilingual text matching method with limited languages.

Acknowledgments

The research work described in this paper has been supported by the National Key Research and Development Program (No.2019QY1802), National Natural Science Foundation of China (No.61866020), National Natural Science Foundation of China (No.61761026), National Natural Science Foundation of China (No.619 72186), General Project of Yunnan Science and Technology Department (No.2019FB082), Natural Science of Yunnan Province Fund (No.2018FB104) and Talent

Fund for Kunming University of Science and Technology (No.KKSY201703015). We thank the anonymous reviewers for their insightful comments and suggestions.

References

[1] B. Li, E. Gaussier and D. Yang, The Dilution/Concentration conditions for cross-language information retrieval models, *Information Processing & Management* **54**(2) (2018), 291–302.

[2] H.B. Hashemi and A. Shakery, Mining a Persian–English comparable corpus for cross-language information retrieval, *Information Processing & Management* **50**(2) (2014), 384–398.

[3] K. Darwish and D.W. Oard, Probabilistic structured query methods, in: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, (2003), 338–344.

[4] D. Nguyen, A. Overwijk, C. Hauff, D.R. Trieschnigg, D. Hiemstra and F. De Jong, Wiki Translate: query translation for cross-lingual information retrieval using only Wikipedia, in: *Workshop of the Cross-Language Evaluation Forum for European Languages*, Springer, (2008), 58–65.

[5] A. Seetha, S. Das and M. Kumar, Evaluation of the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method, in: *10th International Conference on Information Technology (ICIT 2007)*, IEEE, (2007), 56–61.

[6] J. Dadashkarimi, A. Shakery, H. Faili and H. Zamani, An expectation-maximization algorithm for query translation based on pseudo-relevant documents, *Information Processing & Management* **53**(2) (2017), 371–387.

[7] M. Braschler and P. Schäuble, Using corpus-based approaches in a system for multilingual information retrieval, *Information Retrieval* **3**(3) (2000), 273–284.

[8] M. Franz, J.S. McCarley and R.T. Ward, Ad hoc, crosslanguage and spoken document information retrieval at IBM, *Nist Special Publication Sp* (2000), 391–398.

[9] J.-Y. Nie, Cross-language information retrieval, *Synthesis Lectures on Human Language Technologies* **3**(1) (2010), 1–125.

[10] R. Rahimi, A. Shakery and I. King, Extracting translations from comparable corpora for Cross-Language Information Retrieval using the language modeling framework, *Information Processing & Management* **52**(2) (2016), 299–318.

[11] Q. Ying, Applying Bilingual Lexicons to Detect Correspondences in English-Chinese Cross-lingual Plagiarism Docu ments, *Data Analysis and Knowledge Discovery* **30**(7) (2014), 114–119.

[12] J. Wang and D.W. Oard, Matching meaning for cross-language information retrieval, *Information Processing & Management* **48**(4) (2012), 631–653.

[13] F. Diaz, B. Mitra and N. Craswell, Query Expansion with Locally-Trained Word Embeddings, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2016), 367–377.

[14] B. Mitra, N. Craswell, et al., An introduction to neural information retrieval, *Now Foundations and Trends* (2018).

[15] I. Vulić and M.-F. Moens, 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*, (2015), 363–372.

[16] R. Litschko, G. Glavaš, S.P. Ponzetto and I. Vulić, Unsupervised cross-lingual information retrieval using monolingual data only, in: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (2018), 1253–1256.

[17] S. Wu and M. Dredze, Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (2019), 833–844.

[18] S. Sasaki, S. Sun, S. Schamoni, K. Duh and K. Inui, Crosslingual 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*, **2**(Short Papers) (2018), 458–463.

[19] P. Gupta, R.E. Banchs and P. Rosso, Continuous space models for CLIR, *Information Processing & Management* **53**(2) (2017), 359–370.

[20] G.-A. Levow, D.W. Oard and P. Resnik, Dictionary-based techniques for cross-language information retrieval, *Information Processing & Management* **41**(3) (2005), 523–547.

[21] D. He and D. Wu, Enhancing query translation with relevance feedback in translingual information retrieval, *Information Processing & Management* **47**(1) (2011), 1–17.

[22] R. Zbib, L. Zhao, D. Karakos, W. Hartmann, J. De Young, Z. Huang, Z. Jiang, N. Rivkin, L. Zhang, R. Schwartz, et al., Neural-network lexical translation for cross-lingual IR from text and speech, in: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, (2019), 645–654.

[23] S.M. Sarwar, H. Bonab and J. Allan, A Multi-Task Architecture on Relevance-based Neural Query Translation, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (2019), 6339–6344.

[24] D. Zhou, M. Truran, T. Brailsford, V. Wade and H. Ashman, Translation techniques in cross-language information retrieval, *ACM Computing Surveys (CSUR)* **45**(1) (2012), 1–44.

[25] H. Zamani and W.B. Croft, Embedding-based query language models, in: *Proceedings of the 2016 ACM international conference on the theory of information retrieval*, (2016), 147–156.

[26] H. Bonab, S.M. Sarwar and J. Allan, 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* (2020), 9–18.

[27] L. Zhao, R. Zbib, Z. Jiang, D. Karakos and Z. Huang, Weakly supervised attentional model for low resource ad-hoc crosslingual information retrieval, in: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, (2019), 259–264.

[28] S. MacAvaney, A. Yates, A. Cohan and N. Goharian, CEDR: Contextualized embeddings for document ranking, in: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, (2019), 1101–1104.

[29] Z.A. Yilmaz, W. Yang, H. Zhang and J. Lin, Cross-domain modeling of sentence-level evidence for document retrieval, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th Inter-*

national Joint Conference on Natural Language Processing (EMNLPIJCNLP) (2019), 3481–3487.

[30] O. Khattab and M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over bert, in: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, (2020), 39–48.

[31] S. Ruder, I. Vulić and A. Søgaard, A survey of cross-lingual word embedding models, *Journal of Artificial Intelligence Research* **65** (2019), 569–631.

[32] A. Conneau and G. Lample, Cross-lingual language model pretraining, in: *Advances in Neural Information Processing Systems* (2019), 7057–7067.

[33] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G.Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer and V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), 8440–8451.

[34] S. Ruder, I. Vulić and A. Søgaard, A survey of cross-lingual word embedding models, *Journal of Artificial Intelligence Research* **65** (2019), 569–631.

[35] J. Vilares, M.A. Alonso, Y. Doval and M. Vilares, Studying the effect and treatment of misspelled queries in Cross-Language Information Retrieval, *Information Processing & Management* **52**(4) (2016), 646–657.

[36] B. Li and P. Cheng, Learning neural representation for clir with adversarial framework, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (2018), 1861–1870.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin, Attention is All you Need, in: *NIPS*, (2017).

[38] M. Artetxe, G. Labaka and E. Agirre, Learning principled bilingual mappings of word embeddings while preserving monolingual invariance, in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, (2016), 2289–2294.

[39] M. Artetxe, G. Labaka and E. Agirre, A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (2018), 789–798.

[40] C. Dyer, V. Chahuneau and N.A. Smith, A simple, fast, and effective reparameterization of ibm model 2, in: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (2013), 644–648.

[41] J. Zhu, Y. Zhou, J. Zhang and C. Zong, Attend, translate and summarize: An efficient method for neural cross-lingual summarization, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), 1309–1321.

[42] E. Jang, S. Gu and B. Poole, Categorical Reparameterization with Gumbel-Softmax, in: *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, (2017).

[43] M. Roostaee, M.H. Sadreddini and S.M. Fakhrahmad, An effective approach to candidate retrieval for cross-language plagiarism detection: A fusion of conceptual and keyword-based schemes, *Information Processing & Management* **57**(2) (2020), 102150.

[44] S. Sun and K. Duh, CLIRMatrix: A massively large collection of bilingual and multilingual datasets for Cross-Lingual Information Retrieval, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2020), 4160–4170.

[45] D.P. Kingma and J.L. Ba, Adam: A method for stochastic optimization, in: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015).

[46] M. Artetxe and H. Schwenk, Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond, *Transactions of the Association for Computational Linguistics* **7** (2019), 597–610.