Mind the Gap: Cross-Lingual Information Retrieval with Hierarchical Knowledge Enhancement

Fuwei Zhang[1, 2], **Zhao Zhang**[3,*], **Xiang Ao**[1, 2, 4], **Dehong Gao**[5], **Fuzhen Zhuang**[6, 7], **Yi Wei**[5], **Qing He**[1, 2]

[1] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[4] Institute of Intelligent Computing Technology, Suzhou, CAS [5] Alibaba Group, Hangzhou, China
[6] Institute of Artificial Intelligence, Beihang University, Beijing 100191, China
[7] SKLSDE, School of Computer Science, Beihang University, Beijing 100191, China
{zhangfuwei20g, zhangzhao2021, aoxiang, heqing}@ict.ac.cn, zhuangfuzhen@buaa.edu.cn,
{dehong.gdh, yi.weiy}@alibaba-inc.com

Abstract

Cross-Lingual Information Retrieval (CLIR) aims to rank the documents written in a language different from the user's query. The intrinsic gap between different languages is an essential challenge for CLIR. In this paper, we introduce the multilingual knowledge graph (KG) to the CLIR task due to the sufficient information of entities in multiple languages. It is regarded as a "silver bullet" to simultaneously perform explicit alignment between queries and documents and also broaden the representations of queries. And we propose a model named *CLIR with hierarchical knowledge enhancement* (HIKE) for our task. The proposed model encodes the textual information in queries, documents and the KG with multilingual BERT, and incorporates the KG information in the query-document matching process with a hierarchical information fusion mechanism. Particularly, HIKE first integrates the entities and their neighborhood in KG into query representations with a knowledge-level fusion, then combines the knowledge from both source and target languages to further mitigate the linguistic gap with a language-level fusion. Finally, experimental results demonstrate that HIKE achieves substantial improvements over state-of-the-art competitors.

Introduction

The escalation of globalization burgeons the great demand for Cross-Lingual Information Retrieval (CLIR), which has broad applications such as cross-border e-commerce, cross-lingual question answering, and so on (Li et al. 2020; Rücklé, Swarnkar, and Gurevych 2019; Xu et al. 2021). Informally, given a query in one language, CLIR is a document retrieval task that aims to rank the candidate documents in another language according to the relevance between the search query and the documents.

Most existing solutions to tackle the CLIR task are built upon machine translation (Dwivedi and Chandra 2016) systems (also known as MT systems). One technical route is to translate either the query or the document to the same language as the other side (McCarley 1999; Aljlayl and Frieder 2001; Picchi and Peters 1998; Croft, Turtle, and Lewis 1991). The other is to translate both the query and the document to the same intermediate language (Kishida and Kando 2003), e.g. English. After aligning the language of the query and documents, monolingual retrieval is performed to accomplish the task. Hence, the performance of the MT systems and the error accumulations may render them inefficient in CLIR.

Recent studies strive to model CLIR with deep neural networks that encode both query and document into a shared space rather than using MT systems (Zhang et al. 2019; Sasaki et al. 2018; Hui et al. 2018a; Li et al. 2020). Though these approaches achieve some remarkable successes, the intrinsic differences between different languages still exist due to the implicit alignment of these methods. Meanwhile, the query is not very long, leading the lack of information while matching with candidate documents.

To tackle these issues, we aim to find a "silver bullet" to simultaneously perform *explicit alignment* between queries and documents and *broaden* the information of queries. The multilingual knowledge graph (KG), e.g. Wikidata (Vrandečić and Krötzsch 2014), is our answer. As a representative multilingual KG, Wikidata[1] includes more than 94 million entities and 2 thousand kinds of relations, and most of the entities in Wikidata have multilingual aligned names and descriptions[2]. With such an external source of knowledge, we can build an explicit bridge between the source language and target language on the premise of the given query information. For example, Figure 1 exhibits a query "新冠病毒" in Chinese ("COVID-19" in English)

[*] Corresponding author: Zhao Zhang

[1] https://www.wikidata.org/wiki/Wikidata:Main_Page

[2] More than 260 languages are supported now.

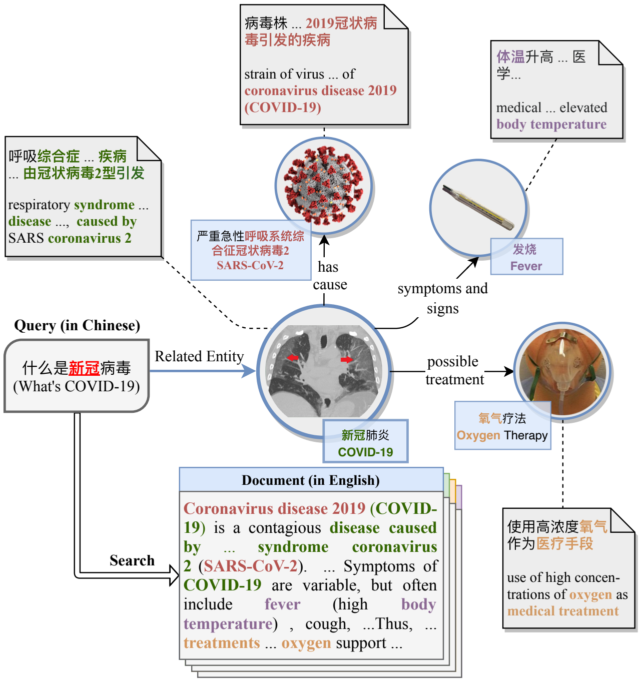
病毒株 ... 2019冠状病毒引发的疾病
strain of virus ... of coronavirus disease 2019 (COVID-19)
体温升高 ... 医学...
medical ... elevated body temperature
呼吸综合症 ... 疾病 ... 由冠状病毒2型引发
respiratory syndrome ... disease ..., caused by SARS coronavirus 2
严重急性呼吸系统综合征冠状病毒2
SARS-CoV-2
has cause
发烧
Fever
symptoms and signs
Query (in Chinese)
什么是新冠病毒
(What's COVID-19)
Related Entity
possible treatment
新冠肺炎
COVID-19
氧气疗法
Oxygen Therapy
Document (in English)
Coronavirus disease 2019 (COVID-19) is a contagious disease caused by ... syndrome coronavirus 2 (SARS-CoV-2). ... Symptoms of COVID-19 are variable, but often include fever (high body temperature) , cough, ...Thus, ... treatments ... oxygen support ...
Search
使用高浓度氧气作为医疗手段
use of high concentrations of oxygen as medical treatment

Figure 1: A toy example for utilizing the multilingual KG for CLIR. The query is in Chinese and the documents are in English. In the query, we give an English translation for better understanding. The entities are denoted in circles. The dotted black line presents the descriptions of an entity. The solid black arrow presents relations between entities. The solid blue arrow shows the related entity of the given query. The hollow arrow presents the documents of the query. The entities and corresponding descriptions in KG are bilingual.

and candidate documents in English. Through the multilingual KG, we could link "新冠病毒" to its aligned entity in English, i.e. "COVID-19", and then extend to some related neighbors, such as "Fever", "SARS-CoV-2" and "Oxygen Therapy". Both the aligned entity and the local neighborhood might contribute to extend the insufficient query and fill in the linguistic gap between the query and documents.

Along this line, we adopt the multilingual KG as an external source to facilitate CLIR and propose a **HI**erarchical **K**nowledge **E**nhancement (HIKE for short) mechanism to fully integrate the relevant knowledge. Indeed, queries are usually short but rich in entities. HIKE establishes a link between queries and multilingual KG through the entities mentioned in queries, and makes full use of the semantic information of entities and their neighborhood in KG with a hierarchical information fusion mechanism. Specifically, a knowledge-level fusion integrates the information in each individual language in the KG, and a language-level fusion combines the integrated information from different languages. The multilingual KG provides valuable information, which helps to reduce the disparity between different languages and is beneficial to the matching process over queries and documents.

To summarize, the contributions are as follows.

- We adopt the external multilingual KG not only as an enhancement for sparse queries but also as an explicit bridge mitigating the gap between the query and the document in CLIR. To the best of our knowledge, this is the first work that utilizes multilingual KG for the neural CLIR task.
- We propose HIKE that makes full use of the entities mentioned in queries as well as the local neighborhoods in the multilingual KG for improving the performance in CLIR. HIKE contains a hierarchical information fusion mechanism to resolve the sparsity in queries and perform easier matching over the query-document pairs.
- Extensive experiments on a number of benchmark datasets in four languages (English, Spanish, French, Chinese) validate the effectiveness of HIKE against state-of-the-art baselines.

Related Work

Current information retrieval models for cross-lingual tasks can be categorized into two groups: (i) translation-based approaches (Nie 2010; Zbib et al. 2019) and (ii) semantic alignment approaches (Bai et al. 2010; Sokolov et al. 2013).

Early works mainly focus on translation-based models. One way is to translate queries to the target language of documents (Oard, He, and Wang 2008), or to translate the documents or corpus to the same language as queries (Miller, Leek, and Schwartz 1999; Xu and Weischedel 2000). The other is to translate both queries and documents to the same intermediate language, e.g. English (Kishida and Kando 2003). In both cases, they aim to simplify the process and use the monolingual information retrieval methods to solve the CLIR problem.

Recently, with the development of deep neural networks, semantic alignment approaches, which directly tackle the CLIR tasks without the translation process, have gained much attention. These methods align queries and documents into the same space with probabilistic or neural network methods and perform query-document matching in the aligned space. Sokolov et al. (2013) proposed a method about learning bilingual n-gram correspondences from relevance rankings. Sasaki et al. (2018) presented a simple yet effective method using shared representations across CLIR models trained in different language pairs. The release of BERT (Devlin et al. 2019) leads to breakthroughs in various NLP tasks (Jiang et al. 2020), including document ranking tasks. Thus Contextualized Embeddings for Document Ranking (CEDR) (MacAvaney et al. 2019) is an effective method for using BERT to enhance the current prevalent neural ranking models, such as KNRM (Xiong et al. 2017b), PACRR (Hui et al. 2018b) and DRMM (Guo et al. 2016). Sun and Duh (2020) utilized a multilingual version of BERT (a.k.a multilingual BERT or mBERT) to conduct the CLIR task. These BERT-based neural ranking models achieve the state-of-the-art results compared with other models.

Besides, due to the fast-growing scale of KGs such as Wikidata (Vrandečić and Krötzsch 2014) and DBpedia (Auer et al. 2007), some researches focus on using high-quality KGs as extra knowledge to perform the information retrieval task. Xiong et al. (2017a) presented a word-entity duet framework for utilizing KGs in ad-hoc retrieval. Entity-Duet Neural Ranking Model (EDRM) (Liu et al. 2018), which introduces KGs to neural search systems, represents queries and documents by their word and entity annotations. Despite the popularity of KG for information retrieval, the works on the topic of KG for CLIR are rarely found. Zhang, Färber, and Rettinger (2016) introduced KG to CLIR systems using the standard similarity measures for document ranking. However, this work does not use neural network models. To the best of our knowledge, our work is the first work that incorporates multilingual KG information for the neural CLIR task.

Methodology

In this section, we illustrate the overall framework of our HIKE model, including the model architecture and the detailed description of model components.

Notations

CLIR is a retrieval task in which search queries and candidate documents are written in different languages. Since search queries are usually short but rich in entities, HIKE establishes a connection between CLIR and the multilingual KG via the entities mentioned in queries, and leverage the KG information through these entities and their local neighborhood in KG. Specifically, for each entity, we obtain the following information from the multilingual KG: (i) entity label[3], (ii) entity description, (iii) labels of neighboring entities, and (iv) descriptions of neighboring entities. It is worth noting that all the information in the KG is multilingual, and the information in different languages is aligned. We leverage the above information to facilitate the CLIR task. Given a query q and a document d. We present an entity $e_q \in \mathcal{E}$ and the i-th neighboring entity $n_{ei} \in \mathcal{E}$, where $\mathcal{E}$ is the entity set in KG. Both the entity and neighboring entities have two information for incorporating: labels and descriptions. Furthermore, for a specific bilingual information retrieval task, the label and description of e_q can be described as $l^r_{e_q}$ and $p^r_{e_q}$, respectively. The label and the description of n_{ei} can be descried as $l^r_{n_{ei}}$ and $p^r_{n_{ei}}$, where $r \in \{s, t\}$ indicates the source language or target language. All these information, including q, d, $l^r_{e_q}$, $p^r_{e_q}$, $l^r_{n_{ei}}$ and $p^r_{n_{ei}}$, is composed of a sequence of tokens.

Model Architecture

HIKE incorporates the multilingual semantic information of the entities and their local neighborhoods from KG into the current CLIR model. The overall architecture of HIKE is shown in Figure 2. HIKE consists of three modules: an encoder module, a hierarchical information fusion module and a query-document matching module. Specifically, in the encoder module, HIKE utilizes multilingual BERT to embed the queries, documents, and semantic information from KG into low-dimensional vectors. Thus the encoder outputs the embeddings to the hierarchical information fusion module, and the latter combines the information from KG into queries and expedites the matching with documents. Particularly, the knowledge-level (first-level) fusion integrates the information in KG, using the multi-head attention mechanism (Vaswani et al. 2017). We use two individual knowledge-level fusion modules to extract features from source and target languages. And then, the language-level (second-level) fusion integrates two representations of an entity in source and target languages through a multilayer perceptron. After the hierarchical information fusion mechanisms, we utilize a matching model to get the relevance score of the query-document pair. The higher the score, the more relevant the query and the document are.

Encoder

The encoder aims to embed the tokens from queries, documents, entities and neighboring entities. It consists of two parts: Query and Document Duet Encoder (QD-Duet-Encoder) and Knowledge Encoder (K-Encoder). QD-Duet-Encoder embeds a query-document pair to a d-dimensional vector. And K-Encoder transforms the label and description of an entity into another d-dimension vector.

QD-Duet-Encoder concatenates the tokens from queries and documents into one sequence, using [CLS] and [SEP] as meta-tokens. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (Devlin et al. 2019). And then the encoder sums the token embedding, segment embedding, positional embedding for each token to get the input embedding and computes the output embedding that represents the semantic and matching information of a query-document pair. Embedding query and document together can make the ranking model benefit from deep semantic information from BERT in addition to individual contextualized token matching (MacAvaney et al. 2019). For a given query q and document d, we have an output from QD-Duet-Encoder as shown in Equation (1). $\boldsymbol{v}_{qd}$ is the [CLS] embedding of the output.

$$\boldsymbol{v}_{qd} = \text{QD-Duet-Encoder}(\{[\text{CLS}], q, [\text{SEP}], d\}), \tag{1}$$

where QD-Duet-Encoder$(\cdot)$ is a multilingual BERT model[4] and $\{\cdot, \cdot\}$ means concatenating two sequences of tokens to one sequence.

K-Encoder aims to embed the knowledge information from entities or neighboring entities in two languages to a feature vector. Inspired by the advantages of embedding the query and document together, we use [CLS] and [SEP] to concatenate the label and the description of an entity to obtain the embedding. Suppose there are k neighboring entities, we denote the set of neighboring entity labels as $\mathcal{N}^r_l = \{l^r_{n_{e1}}, l^r_{n_{e2}}, \cdots, l^r_{n_{ek}}\}$ and the descriptions as $\mathcal{N}^r_p = \{p^r_{n_{e1}}, p^r_{n_{e2}}, \cdots, p^r_{n_{ek}}\}$. All these entities are fed into K-Encoder to compute a feature embedding of the entity as

$$\begin{aligned} \boldsymbol{v}^r_{e_q} &= \text{K-Encoder}(\{[\text{CLS}], l^r_{e_q}, [\text{SEP}], p^r_{e_q}\}), \\ \boldsymbol{v}^r_{n_{ei}} &= \text{K-Encoder}(\{[\text{CLS}], l^r_{n_{ei}}, [\text{SEP}], p^r_{n_{ei}}\}), \end{aligned} \tag{2}$$

[3] In some large-scale KGs like Wikidata (Vrandečić and Krötzsch 2014) and DBpedia (Auer et al. 2007), the name of an entity is denoted as its label.

[4] We used BERT-base, multilingual cased.

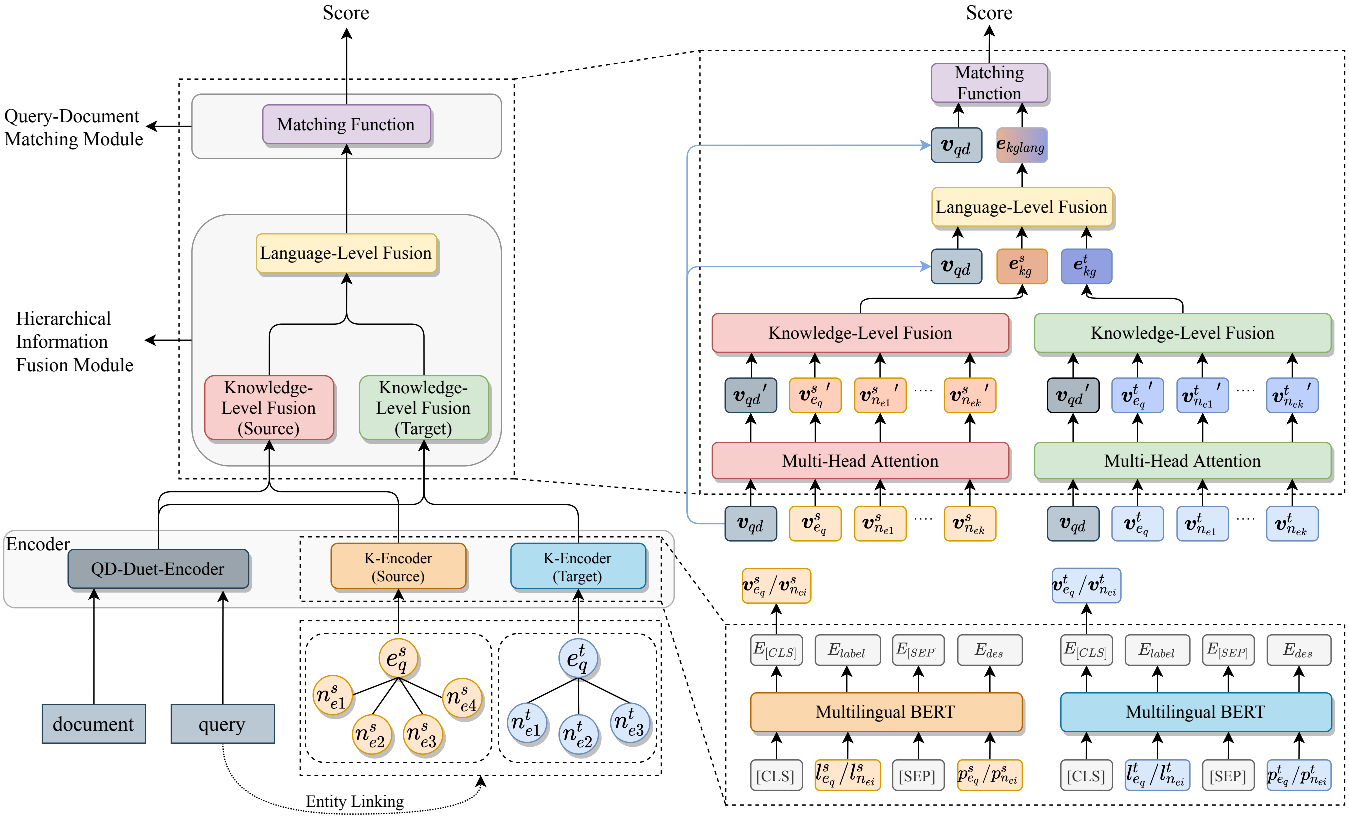

Figure 2: The overall framework of HIKE. The left part is the general architecture, and the right part is the detailed illustration.

where $i = 1, 2, \cdots, k$. K-Encoder$(\cdot)$ is also a multilingual BERT. $r \in \{s, t\}$ denotes that the parameter is for source and target languages, respectively. We sort the neighboring entities in descending order according to their relevance to the central entity and select top k neighboring entities to obtain $\boldsymbol{v}_{n_{ei}}^r$, where k is a hyper-parameter. Specifically, we first run the popular KG embedding model TransE (Bordes et al. 2013) to get the embeddings of entities, and then calculate the cosine similarity between two entities as the relevance score. $\boldsymbol{v}_{e_q}^r$ and $\boldsymbol{v}_{n_{ei}}^r$ are the [CLS] embedding of the entity and the i-th neighboring entity, respectively. The set of feature vectors of neighboring entities is $\mathcal{N}^r = \{\boldsymbol{v}_{n_{e1}}^r, \boldsymbol{v}_{n_{e2}}^r, \cdots, \boldsymbol{v}_{n_{ek}}^r\}$. $\boldsymbol{v}_{qd}$, $\boldsymbol{v}_{e_q}^r$ and $\mathcal{N}^r$ will be treated as the inputs of the fusion module in the next subsection.

Hierarchical Information Fusion

In this section, we detail the hierarchical information fusion module, which is a two-level fusion mechanism, comprising knowledge-level fusion and language-level fusion.

Knowledge-Level Fusion contains two modules: a multi-head self-attention mechanism and an information aggregator. With the help of both two modules, our model can learn a wealth of similar semantic information among the entity, neighboring entities and query-doc pair. In the self-attention mechanism, $\boldsymbol{v}_{qd}$, $\boldsymbol{v}_{e_q}^r$ and $\mathcal{N}^r$ are gathered together and fed into the attention module to calculate the attention values. The input matrix $\boldsymbol{E}^r$ is denoted as:

$$\boldsymbol{E}^r = (\boldsymbol{v}_{qd} \odot \boldsymbol{v}_{e_q}^r \odot \boldsymbol{v}_{n_{e1}}^r \odot \boldsymbol{v}_{n_{e2}}^r \odot \cdots \odot \boldsymbol{v}_{n_{ek}}^r), \quad (3)$$

where $\odot$ is an operation that stacks row vectors into a matrix.

$\boldsymbol{E}^r$ contains the embeddings from query, document, entity and the local neighborhood of the entity. To encapsulate more valuable information, we utilize the multi-head attention mechanism (Vaswani et al. 2017) to learn better latent semantic information. The self-attention module takes three inputs (the query, the key, and the value), which are denoted as $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V} \in \mathbb{R}^{(2+k)\times d}$ (d is the embedding size) respectively. To be specific, we only discuss the j-th head of the multi-head attention mechanism. First, the self-attention model uses each embedding in $\boldsymbol{E}^r$ to get the query $\boldsymbol{Q}^j$, key $\boldsymbol{K}^j$ and value $\boldsymbol{V}^j$ through a linear transformation layer. Then the model goes on using each embedding in the query to attend each embedding in the key through the scaled dot-product attention mechanism (Vaswani et al. 2017), and gets the attention score. Finally, the obtained attention score is applied upon the value $\boldsymbol{V}^j$ to calculate a new representation of Att$(\boldsymbol{Q}^j, \boldsymbol{K}^j, \boldsymbol{V}^j)$, which is formulated as:

$$\text{Att}(\boldsymbol{Q}^j, \boldsymbol{K}^j, \boldsymbol{V}^j) = \text{softmax}(\frac{\boldsymbol{Q}^j \cdot (\boldsymbol{K}^j)^T}{\sqrt{d}}) \cdot \boldsymbol{V}^j. \quad (4)$$

Therefore, each row of Att$(\boldsymbol{Q}^j, \boldsymbol{K}^j, \boldsymbol{V}^j)$ is capable of incorporating the semantic information from the rows in $\boldsymbol{V}^j$. Furthermore, a layer normalization operation (Ba, Kiros, and Hinton 2016) is applied to the output of attention model to obtain the representation of the j-th head $\boldsymbol{H}^j = \text{LayerNorm}(\text{Att}(\boldsymbol{Q}^j, \boldsymbol{K}^j, \boldsymbol{V}^j))$. Next, we pack the multi-head information using the following operation:

$$\text{Multi-Head}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = (\boldsymbol{H}^1 || \boldsymbol{H}^2 || \cdots || \boldsymbol{H}^m)\boldsymbol{W}_H, \quad (5)$$

where $\boldsymbol{W}_H \in \mathbb{R}^{md\times d}$ is a parameter matrix and m is the number of heads.

Accordingly, we obtain the representation after the multi-head attention $\boldsymbol{M}^r = (\boldsymbol{v}_{qd}{}' \odot \boldsymbol{v}^r_{e_q}{}' \odot \boldsymbol{v}^r_{n_{e1}}{}' \odot \boldsymbol{v}^r_{n_{e2}}{}' \odot \cdots \odot \boldsymbol{v}^r_{n_{ek}}{}') = \text{Multi-Head}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \in \mathbb{R}^{(2+k)\times d}$, where $r \in \{s, t\}$ denotes that the parameter is for source and target languages respectively. $\boldsymbol{v}_{qd}{}'$, $\boldsymbol{v}^r_{e_q}{}'$ and $\boldsymbol{v}^r_{n_{ei}}{}'(i = 1, 2, \ldots, k)$ represent the output vectors of multi-head self attention. Finally, we use an information aggregator which consists of a linear transformation layer as Equation (6) to compute the final representation of the knowledge-level features.

$$\boldsymbol{e}^r_{kg} = \text{Tanh}(\boldsymbol{W}_K \cdot \text{vec}(\boldsymbol{M}^r) + \boldsymbol{b}_K), \tag{6}$$

where $\text{vec}(\cdot)$ is a vectorization function that concatenates each row of a matrix as a long vector. $\boldsymbol{W}_K \in \mathbb{R}^{d\times(2+k)d}$ is a parameter matrix and $\boldsymbol{b}_K$ is a d-dimension vector. $\boldsymbol{e}^r_{kg}$ incorporates the deep semantic information from the KG.

Language-Level Fusion combines the query-document pair information with $\boldsymbol{e}^s_{kg}$ and $\boldsymbol{e}^t_{kg}$, which are obtained from the knowledge-level fusion. We use the $\boldsymbol{v}_{qd}$ as guidance in the fusion processing, which is donated in blue arrow in Figure 2. And then, these embeddings are combined by a linear transformation layer which uses Tanh as the activation function to generate a unified representation as:

$$\boldsymbol{e}_{kglang} = \text{Tanh}[\boldsymbol{W}_L(\boldsymbol{v}_{qd}||\boldsymbol{e}^s_{kg}||\boldsymbol{e}^t_{kg}) + \boldsymbol{b}_L], \tag{7}$$

where s and t represent the source and target languages. $\boldsymbol{W}_L \in \mathbb{R}^{d\times 3d}$ and $\boldsymbol{b}_L \in \mathbb{R}^d$ are parameters. $\boldsymbol{e}_{kglang}$ is the unified embedding that incorporates the information from queries, documents, and the multilingual KG.

Matching Function

Finally, HIKE uses the matching function to obtain the score of a query-document pair. Particularly, $\boldsymbol{v}_{qd}$ and $\boldsymbol{e}_{kglang}$ will be concatenated and fed into another linear layer to obtain the relevant ranking score of the query-document pair:

$$f(q, d) = \text{Softmax}[\boldsymbol{W}_S(\boldsymbol{v}_{qd}||\boldsymbol{e}_{kglang}) + b_S], \tag{8}$$

where $f(q, d)$ is the ranking score between the query and document. $\boldsymbol{W}_S \in \mathbb{R}^{1\times 2d}$ and $b_S \in \mathbb{R}^1$ are parameters. And Softmax is an activate function to convert the results into the probability over different classes.

In the training stage, we use standard pairwise hinge loss to train the model as shown in Equation (9).

$$\mathcal{L} = \sum_{d\in D^+_q} \sum_{d'\in D^-_q} [1 - f(q, d) + f(q, d')]_+. \tag{9}$$

D^+_q and D^-_q are the set of relevant documents and irrelevant documents of the query q , and $[\cdot]_+ = \max(0, \cdot)$.

Experiment Methodology

In this section, we describe the details of our experiments, including the dataset, the multilingual KG, baselines, evaluation metrics and implementation details.

Dataset

We evaluate the HIKE model in a public CLIR dataset CLIRMatrix (Sun and Duh 2020). Specifically, we use the MULTI-8 set in CLIRMatrix, in which queries and documents are jointly aligned in 8 different languages. The dataset is mined from 49 million unique queries and 34 billion (query, document, relevance label) triplets. The relevance label $\in \{0, 1, 2, 3, 4, 5, 6\}$ indicates the relevance of the query-document pair. The higher the value, the more relevant the query-document pair is. In MULTI-8, queries remain the same no matter what the language of documents is. For instance, three language pairs English-Spanish, English-French and English-Chinese in MULTI-8 share the same queries. Furthermore, we choose four widely used languages in the world to conduct the bilingual information retrieval tasks, including English (EN), French (FR), Spanish (ES) and Chinese (ZH). Thus there are 12 language pairs in the dataset for training, validation and testing. The training sets of every language pair contain 10,000 queries, while the validation and the test sets contain 1,000 queries. Meanwhile, the number of candidate documents for each query is 100. We use the test1 set in MULTI-8 as our test set to verify the model performance. The statistics of the datasets are summarized in Table 1.

Dataset	train	valid	test
$\{s \to t\}$	10000	1000	1000

Table 1: Statistic of datasets. Here $s, t \in \{\text{EN}, \text{ES}, \text{FR}, \text{ZH}\}$ and $s \neq t$.

	EN-ES	EN-FR	EN-ZH	ES-EN	ES-FR	ES-ZH
source language		7.11			6.53	
target language	6.15	6.34	4.86	7.37	6.73	5.21
	FR-EN	**FR-ES**	**FR-ZH**	**ZH-EN**	**ZH-ES**	**ZH-FR**
source language		6.41			4.95	
target language	7.11	6.19	4.93	7.02	6.13	6.33

Table 2: Average number of golden neighboring entities. "Golden" means the neighboring entities have both the description and the label in a specific language of the queries. The source language is on the left of the connector "-", while the target language is on the right.

Knowledge Graph

We use Wikidata (Vrandečić and Krötzsch 2014), a multilingual KG with entities and relations in a multitude of languages. Up until now, Wikidata contains more than 94 million entities and more than 2000 kinds of relations. And the related entities of queries are annotated by mGENRE (Cao et al. 2021), a multilingual entity linking model which has a high accuracy of entity linking on 105 languages. Table 2 shows the average number of neighboring entities in each dataset.

Baselines

To demonstrate the effectiveness of our model, we compare the performance with the following baselines.

Language Pair	Metrics	Models					
		Vanilla BERT	CEDR-DRMM	CEDR-KNRM	CEDR-PACRR	HIKE$^-$	HIKE
EN-ES	NDCG@1	75.82	73.55	75.40	77.28	80.05	**83.81***
	NDCG@5	80.08	79.19	80.30	80.69	82.63	**84.05***
	NDCG@10	83.36	82.55	83.47	83.42	85.14	**86.18***
EN-FR	NDCG@1	76.92	74.63	71.40	78.33	80.05	**82.93***
	NDCG@5	78.99	78.27	78.53	80.90	81.21	**83.43***
	NDCG@10	82.02	81.01	81.89	83.40	83.20	**85.22***
EN-ZH	NDCG@1	68.98	70.33	76.60	75.10	72.25	**78.16***
	NDCG@5	78.30	78.13	81.35	79.92	78.90	**81.86***
	NDCG@10	82.32	81.91	84.23	82.71	82.90	**84.96***
ES-EN	NDCG@1	74.88	70.73	74.05	74.55	76.38	**80.13***
	NDCG@5	75.04	72.34	74.58	75.05	75.10	**78.34***
	NDCG@10	76.09	74.60	75.99	76.44	76.20	**78.61***
ES-FR	NDCG@1	67.40	74.97	76.05	77.38	73.97	**80.21***
	NDCG@5	72.86	74.65	76.75	76.73	75.18	**78.97***
	NDCG@10	75.51	76.59	78.20	78.16	77.10	**79.88***
ES-ZH	NDCG@1	64.25	65.00	69.35	65.62	65.75	**70.70***
	NDCG@5	69.82	68.69	73.16	73.58	70.71	**74.75***
	NDCG@10	74.08	72.70	75.99	75.85	74.60	**77.06***
FR-EN	NDCG@1	71.15	71.28	70.52	76.90	76.23	**81.03***
	NDCG@5	72.99	72.82	73.99	76.58	75.37	**77.73***
	NDCG@10	75.46	75.14	75.58	78.03	76.78	**78.72***
FR-ES	NDCG@1	77.01	74.60	74.43	80.85	78.98	**83.52***
	NDCG@5	78.18	76.67	77.22	78.89	79.70	**80.57***
	NDCG@10	79.91	78.41	79.16	80.56	80.81	**81.69***
FR-ZH	NDCG@1	63.33	62.37	69.75	65.33	65.37	**70.78***
	NDCG@5	71.73	70.65	73.86	67.82	72.34	**74.42***
	NDCG@10	75.92	74.49	76.89	74.79	76.16	**77.47***
ZH-EN	NDCG@1	56.63	62.83	60.32	61.53	60.45	**68.52***
	NDCG@5	61.69	64.71	64.61	64.53	63.89	**68.43***
	NDCG@10	64.79	66.99	67.03	66.57	66.43	**69.72***
ZH-ES	NDCG@1	54.03	59.95	61.55	60.45	63.33	**67.88***
	NDCG@5	61.64	64.53	66.47	65.61	66.16	**68.95***
	NDCG@10	66.20	67.99	69.30	68.55	69.19	**71.09***
ZH-FR	NDCG@1	59.05	53.23	59.97	58.85	59.47	**65.40***
	NDCG@5	63.40	61.68	64.81	63.91	64.84	**68.07***
	NDCG@10	66.97	65.71	68.34	67.27	68.26	**70.51***

Table 3: NDCG values of baselines and our model. Numbers in the table are in percentages. * marks statistically significant improvements (t-test with p-value $<$ 0.05) compared with the best baseline.

- Vanilla BERT (MacAvaney et al. 2019; Sun and Duh 2020): a fine-tuned multilingual BERT model for CLIR.
- CEDR (MacAvaney et al. 2019): the contextualized embeddings for document ranking (CEDR) model. This model can be applied to various popular neural ranking models, including KNRM (Xiong et al. 2017b), DRMM (Guo et al. 2016) and PACRR (Hui et al. 2018b), to form CEDR-KNRM/DRMM/PACRR.
- HIKE$^-$: A variant of HIKE, which concatenates the KG information with the query directly. The difference between HIKE$^-$ and HIKE is that HIKE$^-$ does not use the hierarchical information fusion mechanism.

Evaluation Metrics

Normalized Discounted Cumulative Gain (NDCG) is adopted for evaluation. And we choose NDCG@1, NDCG@5 and NDCG@10 (only evaluate the top 1, 5 and 10 returned documents) as the metrics in all language pairs.

Implementation Details

In the training stage, the number of heads for the multi-head attention mechanism in knowledge-level fusion is set to 6. In order to reduce the GPU memory and training time, we save the embeddings of entity information before training. The number of all entities we extracted from KG is 376,785. And we only fine-tune the BERT model to obtain textual representations. The learning rates are divided into two parts: the BERT lr_1 and the other modules lr_2. And we set lr_1 to 1e-5 and lr_2 to 1e-3. We set the number of neighboring entities in KG as 3. For those entities without enough neighboring entities, we copy the existing neighboring entities instead. We randomly sample 1600 query-document pairs as our training data per epoch. The maximum training epochs are set to 15.

Evaluation Results

We conduct three experiments to demonstrate the effectiveness of the HIKE model.

Ranking Accuracy

Table 3 summarizes the evaluation results of different cross-lingual retrieval models. From Table 3, we have the following findings. (i) The results indicate that HIKE significantly and consistently outperforms all the baseline models on 12 language pairs w.r.t all metrics, which demonstrates

Model	**EN**			**ES**			**FR**			**ZH**		
	ES	FR	ZH	EN	FR	ZH	EN	ES	ZH	EN	ES	FR
HIKE	**86.18**	**85.22**	**85.30**	**78.61**	**79.88**	**77.06**	**78.72**	**81.69**	**77.47**	**69.72**	**71.09**	**70.51**
HIKE w/o descriptions	85.39	84.29	84.05	77.27	79.09	76.35	77.79	80.95	76.41	68.69	70.31	69.51
HIKE w/o labels	85.47	84.86	84.81	78.34	79.57	76.38	78.58	81.36	76.71	69.29	70.59	70.34
HIKE w/o neighboring entities	85.33	84.47	84.58	78.03	78.17	76.65	78.15	80.90	76.55	68.65	70.23	69.09
HIKE w/o target language information	84.68	83.98	83.84	77.70	78.39	76.22	77.79	81.18	76.25	68.59	69.94	69.09

Table 4: NDCG@10 of models in ablation study.

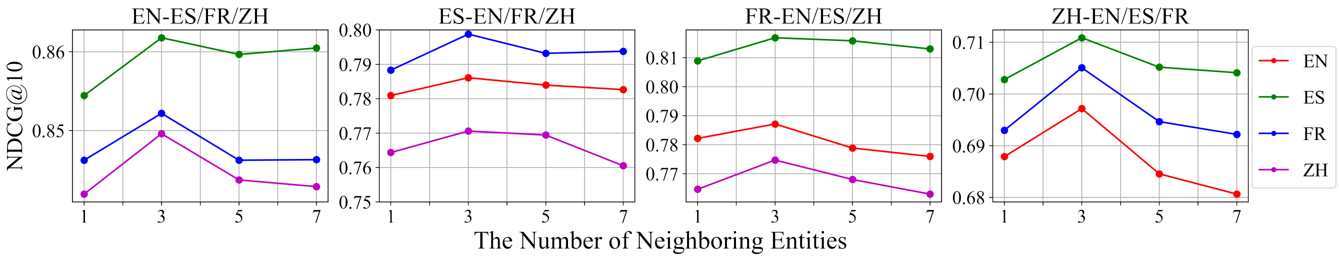

Figure 3: The change of NDCG@10 with the number of neighboring entities increasing.

the effectiveness of the proposed model HIKE. (ii) Comparing with Vanilla BERT, the improvement of HIKE$^-$ embodies the usefulness and importance of the KG. The external KG makes up for the deficiency of queries and provides accurate information while ranking the documents. Moreover, the results of HIKE perform better than HIKE$^-$, which shows the advantages of our hierarchical fusion mechanism. (iii) Specifically, HIKE achieves substantial improvements of both NDCG@1 and NDCG@5 on most datasets comparing with other models, which indicates the knowledge information learned from the entities and neighboring entities is highly related to the task. This result shows that HIKE is capable of ranking the most relevant documents to the top.

All these findings prove that KG information and the hierarchical information fusion can facilitate the CLIR task, and narrow the gap between different languages.

Ablation Study

In this section, we conduct the ablation study to testify the effectiveness of different information used in HIKE. In addition, we do the experiments as:

- Remove the labels or descriptions of entities and neighboring entities to verify the effects of them.
- Remove the information of neighboring entities to study the influence of neighboring entities.
- Remove the information of target language to learn the importance of them in document ranking.

The results are shown in Table 4. From the results, we observe that (i) HIKE obtains the best ranking performance than other incomplete models, indicating that every part of our model makes contributions to the ranking performance. (ii) The model without entity labels outperforms the one without entity description. We conjecture the reason lies in that the information from entity descriptions is more abundant than that from the labels, which is able to provide more beneficial information for the CLIR task. (iii) The model without target language information performs worst in our ablation test. It demonstrates that target language information plays a significant role in the CLIR task, which establishes an explicit connection between the query in the source language and the documents in the target language.

The Effect of Neighboring Entity Number

In this subsection, we explore the influence of the number of neighboring entities. We set the number of neighboring entities from 1 to 7 (step-size is 2) and conduct the experiments over all datasets. Figure 3 demonstrates the results, which are divided into four groups according to the different source languages. Each group contains three different target languages. From the figure, there exists an optimal number of neighbors for each language pair. The model performance first goes up as the number of neighboring entities increases. After the optimal value, the performance falls down. We conjecture the reason lies in that models with small numbers of neighbors cannot take full advantage of the local neighborhood information in KG, resulting in weak NDCG@10 values. While large numbers of neighboring entities may bring in some unrelated information, leading to unsatisfactory results as well.

Conclusion

In this paper, we presented HIKE, a hierarchical knowledge-enhanced model for the CLIR task. HIKE introduces external multilingual KG into the CLIR task and is equipped with a hierarchical information fusion mechanism to take full advantage of the KG information. Specifically, the knowledge-level fusion integrates the KG information in each language. And the language-level fusion combines the information from both source and target languages. The multilingual KG is capable of providing valuable information for the CLIR task, which is beneficial to bridge the gap between queries and documents in different languages. Finally, extensive experiments on benchmark datasets clearly validated the superiority of HIKE against various state-of-the-art baselines.

Acknowledgments

This work is supported by Alibaba Group through Alibaba Innovative Research Program. The research work is supported by the National Key Research and Development Program of China under Grant No. 2017YFB1002104, the National Natural Science Foundation of China under Grant No. U1836206, 62176014, U1811461, and the China Postdoctoral Science Foundation under Grant No. 2021M703273. Xiang Ao is also supported by the Project of Youth Innovation Promotion Association CAS and Beijing Nova Program Z201100006820062.

References

Aljlayl, M.; and Frieder, O. 2001. Effective arabic-english cross-language information retrieval via machine-readable dictionaries and machine translation. In *Proceedings of the tenth international conference on Information and knowledge management*, 295–302.

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, 722–735. Springer.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bai, B.; Weston, J.; Grangier, D.; Collobert, R.; Sadamasa, K.; Qi, Y.; Chapelle, O.; and Weinberger, K. 2010. Learning to rank with (a lot of) word features. *Information retrieval*, 13(3): 291–314.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *Annual Conference on Neural Information Processing Systems*.

Cao, N. D.; Wu, L.; Popat, K.; Artetxe, M.; Goyal, N.; Plekhanov, M.; Zettlemoyer, L.; Cancedda, N.; Riedel, S.; and Petroni, F. 2021. Multilingual Autoregressive Entity Linking. In *arXiv pre-print 2103.12528*.

Croft, W. B.; Turtle, H. R.; and Lewis, D. D. 1991. The use of phrases and structured queries in information retrieval. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, 32–45.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186.

Dwivedi, S. K.; and Chandra, G. 2016. A Survey on Cross-Language Information Retrieval. *International Journal on Cybernetics & Informatics (IJCI) Vol*, 5.

Guo, J.; Fan, Y.; Ai, Q.; and Croft, W. B. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, 55–64.

Hui, K.; Yates, A.; Berberich, K.; and De Melo, G. 2018a. Co-PACRR: A context-aware neural IR model for ad-hoc retrieval. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 279–287.

Hui, K.; Yates, A.; Berberich, K.; and De Melo, G. 2018b. Co-PACRR: A context-aware neural IR model for ad-hoc retrieval. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 279–287.

Jiang, Z.; El-Jaroudi, A.; Hartmann, W.; Karakos, D.; and Zhao, L. 2020. Cross-lingual information retrieval with BERT. *arXiv preprint arXiv:2004.13005*.

Kishida, K.; and Kando, N. 2003. Two-stage refinement of query translation in a pivot language approach to cross-lingual information retrieval: An experiment at CLEF 2003. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, 253–262. Springer.

Li, J.; Liu, C.; Wang, J.; Bing, L.; Li, H.; Liu, X.; Zhao, D.; and Yan, R. 2020. Cross-Lingual Low-Resource Set-to-Description Retrieval for Global E-Commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8212–8219.

Liu, Z.; Xiong, C.; Sun, M.; and Liu, Z. 2018. Entity-Duet Neural Ranking: Understanding the Role of Knowledge Graph Semantics in Neural Information Retrieval. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2395–2405.

MacAvaney, S.; Yates, A.; Cohan, A.; and Goharian, N. 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.

McCarley, J. S. 1999. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 208–214.

Miller, D. R.; Leek, T.; and Schwartz, R. M. 1999. A hidden Markov model information retrieval system. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 214–221.

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

Oard, D. W.; He, D.; and Wang, J. 2008. User-assisted query translation for interactive cross-language information retrieval. *Information Processing & Management*, 44(1): 181–211.

Picchi, E.; and Peters, C. 1998. Cross-language information retrieval: A system for comparable corpus querying. In *Cross-language information retrieval*, 81–92. Springer.

Rücklé, A.; Swarnkar, K.; and Gurevych, I. 2019. Improved cross-lingual question retrieval for community question answering. In *The World Wide Web Conference*, 3179–3186.

Sasaki, S.; Sun, S.; Schamoni, S.; Duh, K.; and Inui, K. 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.

Sokolov, A.; Jehl, L.; Hieber, F.; and Riezler, S. 2013. Boosting cross-language retrieval by learning bilingual phrase associations from relevance rankings. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1688–1699.

Sun, S.; and Duh, K. 2020. 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)*, 4160–4170.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vrandečić, D.; and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10): 78–85.

Xiong, C.; Callan, J.; Liu, T.-Y.; Xiong, C.; and Xiong, C. 2017a. Word-entity duet representations for document ranking. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, 763–772.

Xiong, C.; Dai, Z.; Callan, J.; Liu, Z.; and Power, R. 2017b. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, 55–64.

Xu, J.; and Weischedel, R. 2000. Cross-lingual information retrieval using Hidden Markov models. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 95–103.

Xu, Y.; Wang, Q.; An, Z.; Wang, F.; Zhang, L.; Wu, Y.; Dong, F.; Qiu, C.-W.; Liu, X.; Qiu, J.; et al. 2021. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 100179.

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

Zhang, L.; Färber, M.; and Rettinger, A. 2016. XKnowSearch! exploiting knowledge bases for entity-based cross-lingual information retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2425–2428.

Zhang, R.; Westerfield, C.; Shim, S.; Bingham, G.; Fabbri, A. R.; Hu, W.; Verma, N.; and Radev, D. 2019. Improving Low-Resource Cross-lingual Document Retrieval by Reranking with Deep Bilingual Representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3173–3179.