GIIM: A Graph Information Integration Method for Chinese-Kazakh CLIR

Ping Hu HUPING@XJU.EDU.CN
He Yang ANDYANGHE@GMAIL.COM
Changle Yin 107552304193@STU.XJU.EDU.CN
Tao Wang TWANG@XJU.EDU.CN
Yuchao Chen 107552403885@STU.XJU.EDU.CN
School of Computer Science and Technology (School of Cyberspace Security), Xinjiang University
Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Chinese-Kazakh cross-lingual information retrieval (CLIR) aims to search relevant content from a collection of Kazakh documents using Chinese query statements. The intrinsic differences in grammar, vocabulary, and semantic expression between languages pose significant challenges for semantic alignment in CLIR. Existing CLIR methods that incorporate multilingual knowledge graph (MLKG) typically use simple vector stacking approaches to integrate entity information, failing to leverage deeper entity relationships and semantic connections. To address these challenges, we propose GIIM, a graph information integration method for Chinese-Kazakh CLIR that leverages the rich multilingual entity information embedded in MLKG as semantic bridges to narrow the linguistic gap during query-document matching process. Unlike previous methods, GIIM unifies query-document pairs and entity information into a graph structure and employs Graph Convolutional Network to aggregate both direct and multi-hop relations among entities, effectively modeling complex semantic paths and hierarchical knowledge propagation. To comprehensively evaluate GIIM, we construct CKIRD, a Chinese-Kazakh information retrieval dataset containing approximately 11,820 annotated query-paragraph pairs, and conduct experiments on both CKIRD and the public CLIRMatrix datasets. Experimental results show that GIIM outperforms existing baseline models across multiple ranking metrics, demonstrating its effectiveness on the Chinese-Kazakh CLIR task.

Keywords: Chinese-Kazakh information retrieval, Multilingual knowledge graph, Graph information integration

1. Introduction

Chinese-Kazakh cross-lingual information retrieval (CLIR) aims to enable users to retrieve relevant documents from a Kazakh document collection using Chinese queries. In the context of global informatization, Chinese-Kazakh CLIR plays a crucial role in promoting information exchange between China and Kazakhstan in cultural, technological, and other domains. In a typical semantic-based CLIR pipeline Nogueira et al. (2019), query-document pairs from different languages obtain unified representations through cross-encoder-like model architectures. These representations are then projected to a scalar relevance score via a linear layer, reflecting the degree of matching between queries and documents. The main challenge in Chinese-Kazakh CLIR lies in the significant differences between the two

languages in grammatical structures, vocabulary systems, and semantic expressions Diana and Assem (2019), which poses a semantic gap challenge for CLIR systems. As illustrated in Figure 1, the current mainstream solution is to introduce multilingual knowledge graph (MLKG) into CLIR to bridge semantic differences between different languages through entity information. Although incorporating MLKG has improved the ranking performance of CLIR systems, traditional methods often overlook the direct associations and multi-hop connections between entities in knowledge graph (KG), failing to fully utilize their structured semantic information.

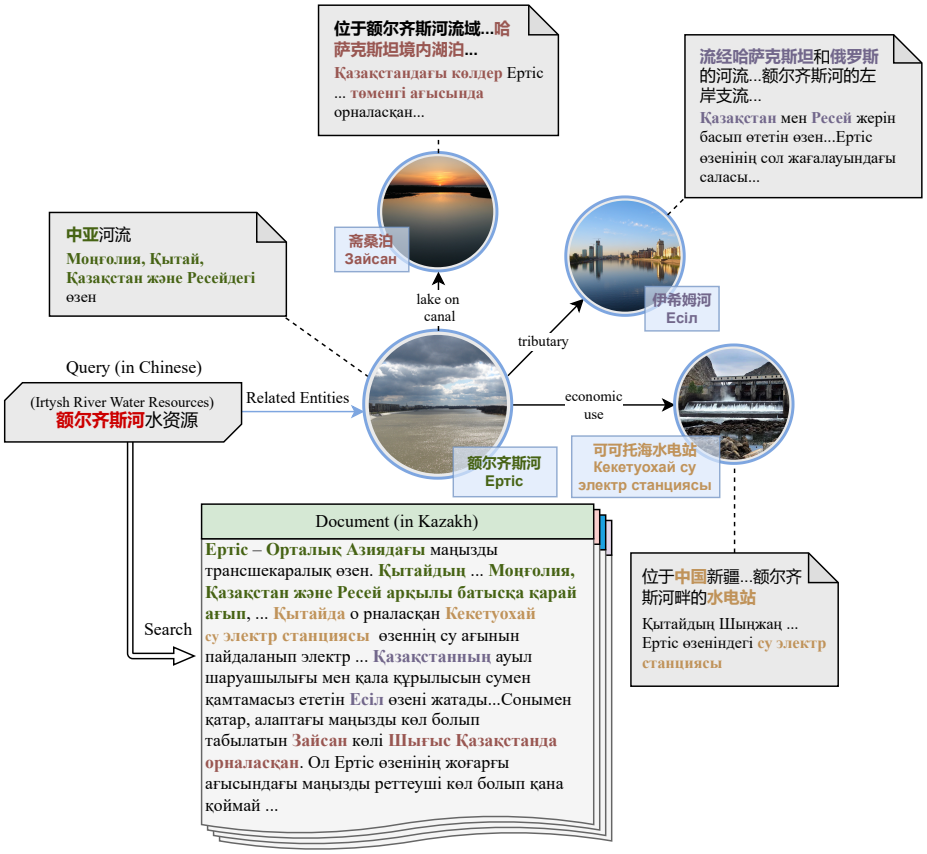

Figure 1: An example of using a MLKG for CLIR. The query is in Chinese and the document is in Kazakh. In the query, we give an English translation for better understanding. The entities are denoted in circles. The dotted black line presents the descriptions of an entity. The solid black arrow presents relations between entities.

Based on the above fact, in this paper, we aim to address two problems in Chinese-Kazakh CLIR. First is the difficulty in aligning semantic information between queries and documents due to imbalanced corpus resources between the two languages. Second is the loss of structural semantic information in KG when introducing MLKG into Chinese-Kazakh CLIR. We believe that modeling query-document pairs and related entities as graph structures and aggregating them using Graph Neural Networks (GNNs) can leverage graph topol-

ogy to handle relationships between query-document pairs and entities in MLKG. However, existing research applying MLKG to CLIR Zhang et al. (2016); Xiong et al. (2017a); Liu et al. (2018); Zhang et al. (2022); Cao et al. (2024); Zhang et al. (2024a) remains limited, mainly facing the following challenges. First of all, early methods integrate entities, relations, and descriptions in MLKG based on rule-based or shallow feature modeling, without utilizing multilingual pre-trained language models (mPLMs) Pires et al. (2019); Jiang et al. (2022) that excel in multilingual representation tasks. Second, most models that encode queries, documents, and MLKG entity information based on mPLMs only adopt simple representation stacking strategies, treating entity nodes as independent sequential elements while ignoring the inherent graph structural connections between entities. Last but not least, research on Chinese-Kazakh CLIR is relatively scarce, which is partly related to insufficiency of Chinese-Kazakh CLIR datasets. Therefore, it is necessary to design a novel MLKG information aggregation method based on mPLMs. Such an approach can utilize the structural semantic information among entities in MLKG and effectively bridge the semantic gap between Chinese queries and Kazakh documents.

To address the aforementioned challenges, we propose GIIM, a graph information integration method for Chinese-Kazakh CLIR. As shown in Figure 2, we first use mBERT to separately encode Chinese query-Kazakh document pairs and the corresponding entities along with their neighboring entities from MLKG in both languages. Next, we construct a graph structure with query-document pairs and their associated entities as nodes, and aggregate node features through GNNs (e.g., GCN Kipf and Welling (2017)) to obtain graph information representations of query-document pairs. Unlike existing methods that use simple stacking strategies to combine knowledge graph information, GIIM achieves direct associations between query-document pairs and entities through graph structure modeling, and supports deep propagation and fusion of multi-hop path semantics. Additionally, we employ contrastive learning to align representations of the same entities across different languages, reducing the representation gap of identical entities between languages and thereby better aggregating entity information from different languages. Finally, by combining the original query-document representations with graph information representations, we directly compute matching scores between queries and documents. The combination of entity alignment and graph information integration effectively utilizes the structured semantic information in MLKG and enhances the matching effectiveness between queries and documents.

Furthermore, as mentioned earlier, available datasets for Chinese-Kazakh CLIR are relatively limited. To our knowledge, currently only the public CLIR dataset CLIRMatrix Sun and Duh (2020) contains both Chinese queries and annotated Kazakh documents. To promote Chinese-Kazakh CLIR research and better validate GIIM's generalization ability across different datasets, we construct a new Chinese-Kazakh information retrieval dataset named CKIRD based on the Kazakh open-domain question answering dataset KazQAD Yeshpanov et al. (2024). Specifically, we translate the Kazakh queries in KazQAD into Chinese, constructing a dataset containing 5,963 Chinese queries and 825,309 Kazakh passages. The training, validation, and test sets of this dataset contain 4,663, 300, and 1,000 queries respectively, forming a total of 11,820 query-passage pairs with binary labels.

The main contributions are as follows:

– We construct a new dataset named CKIRD for Chinese-Kazakh CLIR, which provides crucial data support for Chinese-Kazakh CLIR research.

– We propose a new framework, namely GIIM, for Chinese-Kazakh CLIR task. GIIM unifies query-document pairs and entity information from the MLKG into a graph structure, incorporating structured semantic information into the retrieval process through GNNs.

– We evaluate GIIM on both our constructed dataset CKIRD and the public dataset CLIRMatrix. The experimental results clearly demonstrate that GIIM significantly outperforms existing methods on Chinese-Kazakh CLIR task.

2. Related Work

2.1. CLIR Based on mPLMs

With the success of multilingual pre-trained language models (mPLMs) Pires et al. (2019); Jiang et al. (2022), dense retrieval based on semantic representations has become a major research direction for CLIR Zhao et al. (2024). Most existing methods follow a "retrieval-then-rerank" pipeline Nogueira et al. (2019); Karpukhin et al. (2020), in which a dual-encoder retrieves candidate documents in a shared semantic space, followed by a cross-encoder to optimize ranking. A representative framework is CEDR MacAvaney et al. (2019), which integrates BERT with neural ranking architectures such as DRMM Guo et al. (2016), KNRM Xiong et al. (2017b), and PACRR Hui et al. (2017) to enhance query-document interaction modeling. CEDR includes several variants, where VanillaBERT serves as the simplest form that directly applies BERT's classification capability for document ranking. In addition, OPTICAL Huang et al. (2023) introduces optimal transport distillation to further improve CLIR performance, especially for low-resource languages. However, most existing studies focus on high-resource language pairs such as Chinese-English or English-Russian, while limited work has explored the Chinese-Kazakh scenario. To our knowledge, this work is the first systematic study that focuses on the reranking stage of Chinese-Kazakh CLIR.

2.2. Applications of MLKG in CLIR

To enhance retrieval accuracy, MLKG has been increasingly used in CLIR as external semantic resources. Early works such as XKnowSearch Zhang et al. (2016), AttR-Duet Xiong et al. (2017a), and EDRM Liu et al. (2018) integrated entities, relations, and descriptions to alleviate limitations of keyword-based retrieval. More recent studies combine MLKG with mPLMs. For example, Cao et al. Cao et al. (2024) proposed a dual-encoder architecture that integrates knowledge graphs with text-level structures to improve semantic disambiguation in word sense discrimination tasks. In CLIR, HIKE Zhang et al. (2022) adopted hierarchical fusion of MLKG information, and KEPT Zhang et al. (2024a) further incorporated contrastive learning to enhance cross-lingual dense retrieval. While these models leverage entity information effectively, most still rely on simple feature stacking and overlook deeper entity relationships and multi-hop semantic propagation, limiting the full potential of knowledge graphs.

2.3. Application of GNNs in NLP and Information Retrieval

Compared to approaches that integrate entity information through simple vector stacking, GNNs demonstrate stronger capabilities in modeling complex semantic relationships. GNNs can aggregate information from neighboring nodes in graph-structured data, thereby optimizing node representations. This makes them widely applicable on tasks such as text classification and question-answering systems Wang et al. (2024); Vo (2022). BertGCN Lin et al. (2021) is a typical example, which combines BERT with GCN to perform label propagation and semantic enhancement on document graph structures, significantly improving text classification performance.

In the field of information retrieval, Li et al. Li et al. (2020) proposed a neural information retrieval model based on user behavior graphs such as click graphs and conversation graphs. By leveraging graph structures to uncover latent semantic associations, this approach enhances retrieval effectiveness. Liu et al. Liu et al. (2022) introduced GNN-Encoder, constructing graph structures between queries and documents and fusing interaction information, which notably improves the performance of dense paragraph retrieval. Although GNNs have shown excellent performance on monolingual tasks, their application in CLIR remains in its early stages, particularly for low-resource language pairs like Chinese-Kazakh, where exploration is still lacking.

3. Methodology

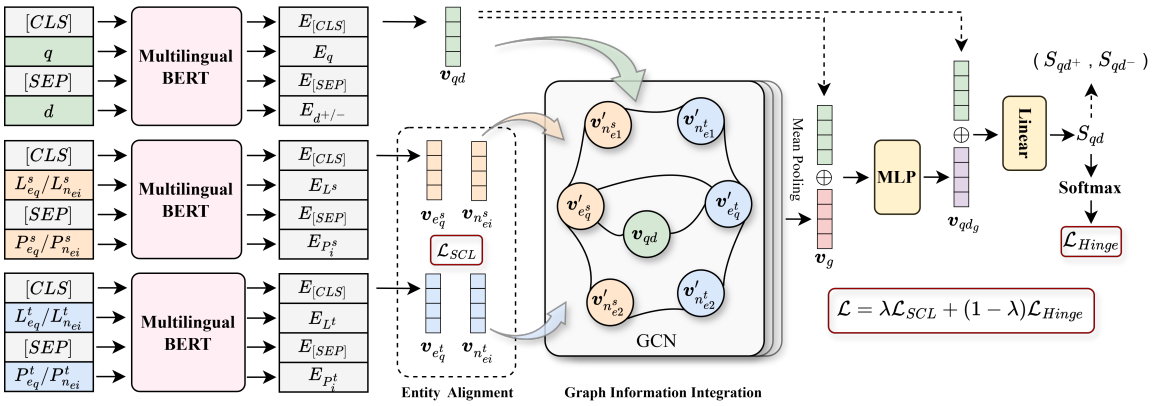

Figure 2: The overall framework of GIIM.

This section introduces the proposed GIIM framework for Chinese-Kazakh CLIR. The core idea is to leverage MLKG to enhance cross-lingual semantic alignment by modeling entity information as graph structures and integrating it into the query-document matching process. GIIM consists of four main components: multilingual encoding of queries, documents, and entities; cross-lingual entity alignment; graph-based information integration; and the final matching and optimization procedure. The details of each component are described below.

3.1. Multilingual Encoder

The objective of the multilingual encoder is to encode query-document pairs, and the corresponding entities along with their neighboring entities in the MLKG into vector representations.

For a given Chinese query q and a Kazakh document d, we first link q to its corresponding entity e_q and neighboring entities n_{ei}, obtaining their labels L and descriptions P. The query-document pair and the entity label-description pairs are then encoded separately using a multilingual encoder to obtain their vector representations, as follows:

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

$$\begin{aligned} \boldsymbol{v}_{e_q^l} &= \mathit{Multi\text{-}Encoder}([[CLS], L_{e_q}^l, [SEP], P_{e_q}^l]) \\ \boldsymbol{v}_{n_{ei}^l} &= \mathit{Multi\text{-}Encoder}([[CLS], L_{n_{ei}}^l, [SEP], P_{n_{ei}}^l]) \end{aligned} \tag{2}$$

We employ mBERT as the $\mathit{Multi\text{-}Encoder}$ and extract the embedding of the $[CLS]$ token from the final layer as the vector representation $\boldsymbol{v}_{qd}$, $\boldsymbol{v}_{e_q^l}$, $\boldsymbol{v}_{n_{ei}^l}$ for the query-document pair and the entity label-description pairs. Where $l \in \{s, t\}$ indicates the source language (Chinese) or target language (Kazakh). $L_{e_q}^l$ and $P_{e_q}^l$ represent the label and description of e_q, while $L_{n_{ei}}^l$ and $P_{n_{ei}}^l$ represent the label and description of n_{ei}, $i = \{1, 2, 3, ..., N\}$. For each query, we use the BGE[1] embedding model to compute the cosine similarity between the query and all neighboring entities of the corresponding entity, selecting the top N similar neighboring entities. Here, $\boldsymbol{v}_{qd}, \boldsymbol{v}_{e_q^l}, \boldsymbol{v}_{n_{ei}^l} \in \mathbb{R}^{1\times dim}$. We define $\boldsymbol{v}_{e^l} = (\boldsymbol{v}_{e_q^l} \oplus \boldsymbol{v}_{n_{e1}^l} \oplus ... \oplus \boldsymbol{v}_{n_{eN}^l}) \in \mathbb{R}^{(N+1)\times dim}$, where $\oplus$ denotes the horizontal concatenation operation of vectors. $\boldsymbol{v}_{e^l}$ provides the key input for entity alignment.

3.2. Entity Alignment

Entity alignment aims to reduce the distance between entity representations of different languages in the semantic space, thereby providing higher-quality entity node representations for graph information integration. We employ contrastive learning to perform entity alignment, bridging the semantic gap between Chinese and Kazakh entity representations.

For each entity, we define the source language and target language representations of the same entity as positive pairs, while representations of different entities in the target language serve as negative samples within a batch. We optimize the entity information alignment using the InfoNCE He et al. (2020) loss function, defined as follows:

$$\mathcal{L}_{SCL} = -\frac{1}{N+1} \sum_{i=0}^{N+1} \log \frac{\exp(sim(\boldsymbol{v}_{e_i^s}, \boldsymbol{v}_{e_i^t}^+)/\tau)}{\sum_{j=0}^{N} \exp(sim(\boldsymbol{v}_{e_i^s}, \boldsymbol{v}_{e_j^t})/\tau)} \tag{3}$$

Where $\boldsymbol{v}_{e_i^s}$ represents the i-th row vector of $\boldsymbol{v}_{e^s}$. When $i = 0$, $\boldsymbol{v}_{e_0^s}$ is the vector representation of the query's corresponding source language entity information $\boldsymbol{v}_{e_q^s}$. When $i = \{1, 2, 3, ..., N\}$, $\boldsymbol{v}_{e_i^s}$ represents $\boldsymbol{v}_{n_{ei}^s}$. The same applies to $\boldsymbol{v}_{e_i^t}$. $\boldsymbol{v}_{e_i^t}^+$ represents the positive sample corresponding to the source language entity vector, i.e., the vector representation

1. https://huggingface.co/BAAI/bge-base-zh-v1.5

of the same entity in the target language. The function $sim(,)$ denotes the cosine similarity between samples, and τ is the temperature coefficient, serving as a hyperparameter to control the smoothness of the similarity distribution. The aligned entity information representations, denoted as $\boldsymbol{v}'_{e_i^s}$ and $\boldsymbol{v}'_{e_i^t}$, along with $\boldsymbol{v}_{qd}$, serve as inputs for the subsequent graph information integration.

3.3. Graph Information Integration

In the graph information integration process, query-document pairs and query-related entities are modeled as a graph structure, which is then processed through GCN Kipf and Welling (2017) to integrate structured knowledge.

For each query-document pair, a graph $G = (\boldsymbol{X}, \boldsymbol{A})$ is constructed, where nodes correspond to the vectors $\boldsymbol{v}_{qd}$, $\boldsymbol{v}'_{e_i^s}$, and $\boldsymbol{v}'_{e_i^t}$. The total number of nodes is $n = 2(N+1)+1$. These node representations are stacked to form the node feature matrix $\boldsymbol{X} \in \mathbb{R}^{n \times dim}$, with each row representing a node. The adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ encodes the node connections: the query-document node connects to its corresponding entities in both languages; each entity connects to its cross-lingual counterpart; and within each language, the query's entity connects to all neighboring entities. This design results in a star-shaped topology that bridges textual and knowledge graph information while preserving cross-lingual alignment. Self-loops are added to form $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}_n$. Finally, we employ an l-layer GCN to aggregate node information, as follows:

$$\boldsymbol{X}^{(l)} = \sigma(\tilde{\boldsymbol{D}}^{-\frac{1}{2}} \tilde{\boldsymbol{A}} \tilde{\boldsymbol{D}}^{-\frac{1}{2}} \boldsymbol{X}^{(l-1)} \boldsymbol{W}^{(l)}) \tag{4}$$

Where $l = \{1, 2, 3...\}$. $\tilde{\boldsymbol{D}}$ is the degree matrix of $\tilde{\boldsymbol{A}}$, $\boldsymbol{X}^0 = \boldsymbol{X} = [\boldsymbol{v}_{qd}{}^T \oplus \boldsymbol{v}'_{e_i^s}{}^T \oplus \boldsymbol{v}'_{e_i^t}{}^T]^T$ with the superscript T denoting the transpose operation, $\oplus$ denotes the horizontal concatenation operation of vectors and $\boldsymbol{W}^{(l)}$ is the weight matrix of the l-th layer. The function $\sigma(\cdot)$ denotes the ReLU activation function.

After the GCN operations, we obtain the final node representations $\boldsymbol{X}^{(l)}$. As shown in Equation (5), to generate the graph representation $\boldsymbol{v}_g$ that captures the aggregated semantic information from all nodes, we apply mean pooling across all node representations.

$$\boldsymbol{v}_g = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^{(l)} \tag{5}$$

Where $\boldsymbol{X}_i^{(l)}$ represents the i-th node representation from the final GCN layer, and $\boldsymbol{v}_g \in \mathbb{R}^{1 \times dim}$.

3.4. Query-Document Matching and Loss Function

As shown in Equation (6), before computing the query-document score, we integrate the graph-based semantic information by passing v_{qd} and v_g through a k-layer multi-layer perceptron (MLP) to obtain the query-document representation that aggregates graph information v_{qdg}.

$$\boldsymbol{v}_{qd_g} = \boldsymbol{h}^{(k)} = \tanh(\boldsymbol{h}^{(k-1)} \boldsymbol{W}^{(k)} + \boldsymbol{b}^{(k)}) \tag{6}$$

Where $\boldsymbol{h}^{(0)} = [\boldsymbol{v}_{qd} \oplus \boldsymbol{v}_g] \in \mathbb{R}^{1\times 2dim}$, $\boldsymbol{h}^{(k)}$ represents the hidden representation at layer k, and $\boldsymbol{W}^{(k)}$ and $\boldsymbol{b}^{(k)}$ are the weight matrix and bias vector for layer k, respectively. The tanh function serves as the activation function for each layer. We take the output of the final MLP layer to obtain $\boldsymbol{v}_{qd_g}$.

The final relevance score between the query and document is computed as follows:

$$f(q,d) = [\boldsymbol{v}_{qd} \oplus \boldsymbol{v}_{qd_g}]\boldsymbol{W} + \boldsymbol{b} \tag{7}$$

Where $f(q,d)$ represents the relevance score between query q and document d, $[\boldsymbol{v}_{qd} \oplus \boldsymbol{v}_{qd_g}] \in \mathbb{R}^{1\times 2dim}$ denotes the concatenation of input vectors,, $\boldsymbol{W} \in \mathbb{R}^{2dim\times 1}$ is the weight matrix, and $\boldsymbol{b} \in \mathbb{R}^{1\times 1}$ is the bias vector.

During training, for the positive document d^+ and negative document d^- of the current query q, we can obtain their relevance scores $f(q,d^+)$ and $f(q,d^-)$. As shown in Equation (8), we apply the softmax operation to map the scores to the $(0,1)$ interval, obtaining the normalized scores S_{qd^+}/S_{qd^-} between the query and positive/negative documents.

$$S_{qd^+}, S_{qd^-} = softmax(f(q,d^+), f(q,d^-)) \tag{8}$$

Where $d^+ \in D_q^+$ and $d^- \in D_q^-$, and the sets D_q^+ and D_q^- represent relevant and irrelevant documents for query q, respectively.

We adopt the standard pairwise hinge loss to optimize the model's ranking capability, defined as follows:

$$\mathcal{L}_{Hinge} = \sum_{d^+\in D_q^+} \sum_{d^-\in D_q^-} max(0, 1 - S_{qd^+} + S_{qd^-}) \tag{9}$$

The overall training objective combines the InfoNCE loss for entity alignment and the pairwise hinge loss for ranking optimization, as follows:

$$\mathcal{L} = \lambda\mathcal{L}_{SCL} + (1-\lambda)\mathcal{L}_{Hinge} \tag{10}$$

Where the hyperparameter $\lambda \in [0,1)$ controls the trade-off between these two objectives, allowing GIIM to balance the emphasis on cross-lingual entity alignment and document ranking optimization during training.

4. Experiments

In this section, we present a comprehensive evaluation of the proposed GIIM framework on Chinese-Kazakh CLIR tasks. We describe the datasets and evaluation metrics used, detail the implementation settings, and report results in terms of overall ranking performance, ablation studies, and parameter sensitivity analysis. These experiments are designed to assess the effectiveness, robustness, and generalizability of GIIM across different retrieval scenarios.

4.1. Datasets and Evaluations

4.1.1. Datasets

We evaluate GIIM on two datasets: (1) **CLIRMatrix** Sun and Duh (2020), a large-scale CLIR benchmark that includes the CLIR subset BI-139 and the multilingual retrieval subset MULTI-8. In this work, the BI-139 subset is used for Chinese-Kazakh CLIR, where the query language is Chinese and the document language is Kazakh. The training set contains 10,000 queries, with 1,000 queries each in the validation and test sets. Each query is associated with 100 candidate documents, and labels are multi-level relevance scores in $\{0, 1, 2, 3, 4, 5, 6\}$, where higher scores indicate stronger relevance. (2) **CKIRD**, a Chinese-Kazakh information retrieval dataset constructed based on the Kazakh open-domain question answering dataset KazQAD Yeshpanov et al. (2024). The original Kazakh queries were translated into Chinese using *Google Translate*[2], and the dataset was re-split into 4,663 queries for training, 300 for validation, and 1,000 for testing. Similar to CLIRMatrix, 100 candidate passages were constructed for each query, with binary relevance labels $\{0, 1\}$ for reranking.

Table 1: Statistical Comparison of CLIRMatrix and CKIRD Datasets

Attribute \ Dataset	CLIRMatrix			CKIRD		
	Train	Test	Dev	Train	Test	Dev
Chinese Queries	5,087	1,088	387	4,653	1,000	298
Annotated query-document pairs	508,700	108,800	38,700	9,531	1,748	518
Avg. Relevant Docs/Passages	10.57	10.74	11.31	1.19	1.41	1.41
Candidate Docs/Passages	-	100	100	-	100	100
Relevance Label	Multi-level			Binary		
Task Type	Document Retrieval			Passage Retrieval		

Wikidata Vrandečić and Krötzsch (2014) is used as the MLKG in this study. We manually annotated all Chinese queries in both CLIRMatrix and CKIRD with corresponding entities and retrieved their Chinese, Kazakh, and English labels and descriptions via the *Wikidata API*[3]. Missing information in Chinese or Kazakh was supplemented through English translation. Queries without matching entities were removed. Table 1 presents the statistics of the two datasets after preprocessing. It is worth noting that the number of Chinese queries in the table is lower than that in the original datasets due to the removal of queries without matched entities.

4.1.2. Evaluation Metrics

To evaluate GIIM's performance, we employ Mean Reciprocal Rank (MRR@n) and Normalized Discounted Cumulative Gain (NDCG@n) at cutoff points $n \in \{1, 5, 10\}$. MRR@n measures the ability to rank relevant documents higher, especially important for precision-

2. https://translate.google.com/
3. https://www.wikidata.org/w/api.php

focused retrieval scenarios, while NDCG@n evaluates the quality of ranking considering both relevance and position.

4.2. Ranking Performance Comparison

Table 2 presents the ranking performance of GIIM compared to various baseline models on the CLIRMatrix and CKIRD datasets. The results demonstrate that GIIM achieves the best performance across all MRR and NDCG metrics. On the CLIRMatrix dataset, GIIM improves MRR@1 and NDCG@1 scores by 2.39% and 2.65% respectively compared to the HIKE model, which also enhances CLIR through MLKG. These results validate that graph-based information integration approaches offer significant advantages over traditional vector stacking methods, particularly in identifying highly relevant documents. Moreover, even when compared against advanced models specifically designed for multilingual text retrieval such as BGE-reranker-v2-m3 and mGTE-reranker, GIIM maintains its leading position across all metrics. This further demonstrates that incorporating entity information from MLKG relevant to the query effectively narrows the semantic gap between Chinese and Kazakh, thereby enhancing the model's semantic understanding of queries. On the CKIRD dataset, despite the challenges of binary passage-level classification and sparse relevant passages, GIIM still maintains its leading performance across all evaluation metrics. This fully demonstrates the method's strong adaptability and robustness in low-resource and label-sparse scenarios. Considering the experimental results from both datasets, it is evident that GIIM significantly enhances semantic alignment between Chinese and Kazakh by deeply aggregating structured semantic information from MLKG, thus achieving more stable and efficient ranking performance on CLIR tasks.

Table 2: Retrieval performance comparison between GIIM and other models on CLIRMatrix and CKIRD datasets (%)

Dataset	Model	MRR@1	MRR@5	MRR@10	NDCG@1	NDCG@5	NDCG@10
CLIRMatrix	VanillaBERT MacAvaney et al. (2019)	68.38	77.92	78.57	49.40	58.91	64.63
	CEDR-DRMM MacAvaney et al. (2019)	72.06	79.78	80.35	51.48	58.91	64.00
	CEDR-KNRM MacAvaney et al. (2019)	77.02	83.48	84.10	57.23	63.05	68.15
	CEDR-PACRR MacAvaney et al. (2019)	73.35	81.37	81.81	53.75	60.48	64.62
	HIKE Zhang et al. (2022)	79.41	85.90	86.27	58.79	65.82	70.59
	BGE-reranker-v2-m3 Chen et al. (2024)	78.95	85.12	85.48	58.46	65.04	70.21
	mGTE-reranker Zhang et al. (2024b)	79.10	85.97	86.31	58.92	66.10	70.82
	GIIM(ours)	**81.80**	**87.09**	**87.43**	**61.44**	**67.25**	**71.76**
CKIRD	VanillaBERT MacAvaney et al. (2019)	66.10	78.05	78.41	66.10	80.28	81.97
	CEDR-DRMM MacAvaney et al. (2019)	59.90	71.47	72.30	59.90	72.63	75.78
	CEDR-KNRM MacAvaney et al. (2019)	78.10	86.18	86.42	78.10	87.05	88.28
	CEDR-PACRR MacAvaney et al. (2019)	66.70	78.49	79.00	66.70	81.30	83.16
	HIKE Zhang et al. (2022)	77.40	86.03	86.26	77.40	87.25	88.46
	BGE-reranker-v2-m3 Chen et al. (2024)	77.20	85.79	86.07	77.20	86.98	88.02
	mGTE-reranker Zhang et al. (2024b)	77.90	86.11	86.32	77.90	87.31	88.47
	GIIM(ours)	**78.40**	**86.20**	**86.46**	**78.40**	**87.54**	**88.71**

4.3. Ablation Study

To validate the contributions of each module in the GIIM model to retrieval performance, we conducted ablation experiments. We analyzed the impact of entity alignment (EA), graph information integration (GII), and adjacent entity information. The specific experimental

settings are as follows: (1) Removing the EA module to study its effect on GCN's integration of graph information; (2) Removing the GII module to analyze its impact on retrieval performance; (3) Deleting adjacent entity information to investigate its contribution to retrieval performance. The experimental results are shown in Table 3.

Table 3: Ablation study results of GIIM on CLIRMatrix and CKIRD datasets (%)

Dataset	Model	MRR@1	MRR@5	MRR@10	NDCG@1	NDCG@5	NDCG@10
CLIRMatrix	*w/o EA + GII*	68.38	77.92	78.57	49.40	58.91	64.63
	w/o GII	79.78	86.07	86.38	58.94	65.48	70.20
	w/o EA	79.96	85.81	86.26	58.99	65.49	70.38
	w/o neighboring entities info	81.37	86.81	87.25	60.92	66.82	71.45
	GIIM(ours)	**81.80**	**87.09**	**87.43**	**61.44**	**67.25**	**71.76**
CKIRD	*w/o EA + GII*	66.10	78.05	78.41	66.10	80.28	81.97
	w/o GII	75.60	83.97	84.37	75.60	85.07	86.74
	w/o EA	72.50	82.79	83.16	72.50	84.56	86.06
	w/o neighboring entities info	76.30	85.01	85.37	76.30	86.72	87.77
	GIIM(ours)	**78.40**	**86.20**	**86.46**	**78.40**	**87.54**	**88.71**

From Table 3, it is evident that GIIM achieves the best retrieval performance compared to the other incomplete model variants, indicating that each module contributes significantly to ranking performance. Specifically, removing the entity alignment module hinders the GCN from effectively aggregating structural information across cross-lingual entities, leading to a notable performance drop. Eliminating the integration module prevents the model from fully leveraging the semantic structure of the knowledge graph, making it rely solely on alignment information from contrastive learning, which also limits retrieval accuracy. Furthermore, the results of removing adjacent entity information suggest that, although its contribution to performance improvement is limited, it still offers certain value in enhancing the contextual semantics of queries. This is particularly evident in the CKIRD dataset—a low-resource, passage-level binary classification retrieval task—where the absence of any single module results in varying degrees of performance degradation. These findings further validate the critical role of structured semantic fusion in supporting Chinese-Kazakh CLIR tasks.

4.4. Parameter Sensitivity Analysis

To further investigate the impact of hyperparameters on the performance of the GIIM model, we evaluated the effects of the number of neighboring entities N and the contrastive learning loss weight λ on ranking performance. The results are shown in Figure 3. On the CLIRMatrix dataset, a moderate increase in N provides additional semantic support, while an excessive number of neighbors may introduce noise. For λ, a balance must be struck between enhancing entity alignment and optimizing the ranking objective; overly large values of λ can weaken the model's ability to match queries with documents. On the CKIRD dataset, the observed parameter sensitivity trends closely align with those on CLIRMatrix, with optimal performance consistently achieved at $N = 4$ and $\lambda = 0.3$. This consistency across datasets demonstrates GIIM's robustness and transferability in diverse retrieval scenarios. Overall, proper configuration of these two key parameters plays a crucial role in improving model performance and offers valuable guidance for real-world deployment.

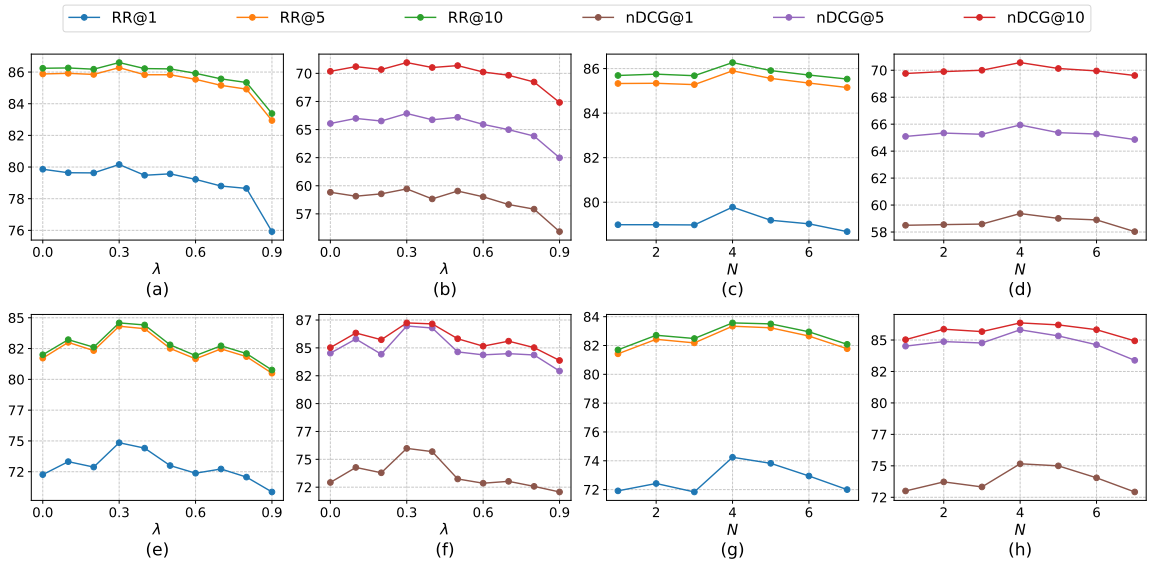

Figure 3: Performance trends of GIIM under varying numbers of neighboring entities (N) and contrastive learning weights (λ) on CLIRMatrix and CKIRD datasets. (a)–(d) show MRR@n and NDCG@n scores on CLIRMatrix under fixed λ ((a), (b)) and fixed N ((c), (d)); (e)–(h) present the corresponding results on CKIRD. The results indicate that model performance is stable within a moderate range of N and λ, with optimal performance achieved at $N = 4$ and $\lambda = 0.3$.

5. Conclusion

In this paper, we propose GIIM, which leverages structured knowledge from MLKG to integrate query-document and related entity information through graph reconstruction. It effectively overcomes the limitations of simple vector stacking approaches and enables deeper modeling of entity relationships and semantic connections, thereby enhancing semantic alignment in Chinese-Kazakh CLIR. To support evaluation, we constructed CKIRD, a new Chinese-Kazakh paragraph retrieval dataset containing 5,963 Chinese queries, 825,309 Kazakh passages, and approximately 11,820 annotated query-paragraph pairs. Comprehensive experiments on CLIRMatrix and CKIRD datasets demonstrate that GIIM consistently outperforms the strong CLIR baseline HIKE, achieving an average improvement of 1.58% in MRR and 1.75% in NDCG. These results validate the effectiveness of our graph information integration method.

The core innovation of GIIM lies in its heterogeneous graph structure that unifies query-document pairs and multilingual entities, enabling deep semantic propagation through GCNs and cross-lingual alignment via contrastive learning. Unlike traditional vector stacking approaches, GIIM models complex semantic paths and hierarchical knowledge propagation. The framework demonstrates transferability to other low-resource language pairs such as Chinese-Mongolian and Chinese-Uyghur. While depending on MLKG quality, future work will explore leveraging large language models to enhance entity information robustness.

Acknowledgments

This work was supported partially by 1) Project 2022xjkk0704-1 of the Talent Base of the Ministry of Science and Technology; 2) Outstanding Graduate Student Innovation Project of Xinjiang University (XJDX2025YJS196).

References

Yukun Cao, Chengkun Jin, Yijia Tang, and ZiYue Wei. Word sense disambiguation combining knowledge graph and text hierarchical structure. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(12), November 2024. ISSN 2375-4699. doi: 10.1145/3677524. URL https://doi.org/10.1145/3677524.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Rakhimova Diana and Shormakova Assem. Problems of semantics of words of the kazakh language in the information retrieval. In Ngoc Thanh Nguyen, Richard Chbeir, Ernesto Exposito, Philippe Aniorté, and Bogdan Trawiński, editors, *Computational Collective Intelligence*, pages 70–81, Cham, 2019. Springer International Publishing. ISBN 978-3-030-28374-2.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 55–64, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340731. doi: 10.1145/2983323.2983769. URL https://doi.org/10.1145/2983323.2983769.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Zhiqi Huang, Puxuan Yu, and James Allan. Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, pages 1048–1056, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394079. doi: 10.1145/3539597.3570468. URL https://doi.org/10.1145/3539597.3570468.

Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. PACRR: A position-aware neural IR model for relevance matching. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1049–1058, Copenhagen, Denmark, September

2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1110. URL https://aclanthology.org/D17-1110/.

Xiaoze Jiang, Yaobo Liang, Weizhu Chen, and Nan Duan. Xlm-k: Improving cross-lingual language model pre-training with multilingual knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10840–10848, Jun. 2022. doi: 10.1609/aaai.v36i10.21330. URL https://ojs.aaai.org/index.php/AAAI/article/view/21330.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Xiangsheng Li, Maarten de Rijke, Yiqun Liu, Jiaxin Mao, Weizhi Ma, Min Zhang, and Shaoping Ma. Learning better representations for neural information retrieval with graph information. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 795–804, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3411957. URL https://doi.org/10.1145/3340531.3411957.

Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. BertGCN: Transductive text classification by combining GNN and BERT. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1456–1462, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.126. URL https://aclanthology.org/2021.findings-acl.126/.

Jiduan Liu, Jiahao Liu, Yang Yang, Jingang Wang, Wei Wu, Dongyan Zhao, and Rui Yan. GNN-encoder: Learning a dual-encoder architecture via graph neural networks for dense passage retrieval. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 564–575, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.39. URL https://aclanthology.org/2022.findings-emnlp.39/.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2395–2405, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1223. URL https://aclanthology.org/P18-1223/.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. Cedr: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 1101–1104, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361729. doi: 10.1145/3331184.3331317. URL https://doi.org/10.1145/3331184.3331317.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. Multi-stage document ranking with BERT. *arXiv preprint arXiv:1910.14424*, 2019. URL http://arxiv.org/abs/1910.14424.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, Jul 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL https://aclanthology.org/P19-1493/.

Shuo Sun and Kevin Duh. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.340. URL https://aclanthology.org/2020.emnlp-main.340/.

Tham Vo. An integrated topic modelling and graph neural network for improving cross-lingual text classification. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(1), November 2022. ISSN 2375-4699. doi: 10.1145/3530260. URL https://doi.org/10.1145/3530260.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, September 2014. ISSN 0001-0782. doi: 10.1145/2629489. URL https://doi.org/10.1145/2629489.

Kunze Wang, Yihao Ding, and Soyeon Caren Han. Graph neural networks for text classification: a survey. *Artificial Intelligence Review*, 57(8):190, 2024. ISSN 1573-7462. doi: 10.1007/s10462-024-10808-0. URL https://doi.org/10.1007/s10462-024-10808-0.

Chenyan Xiong, Jamie Callan, and Tie-Yan Liu. Word-entity duet representations for document ranking. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 763–772, New York, NY, USA, 2017a. Association for Computing Machinery. ISBN 9781450350228. doi: 10.1145/3077136.3080768. URL https://doi.org/10.1145/3077136.3080768.

Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 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*, SIGIR '17, pages 55–64, New York, NY, USA, 2017b. Association for Computing Machinery. ISBN 9781450350228. doi: 10.1145/3077136.3080809. URL https://doi.org/10.1145/3077136.3080809.

Rustem Yeshpanov, Pavel Efimov, Leonid Boytsov, Ardak Shalkarbayuli, and Pavel Braslavski. KazQAD: Kazakh open-domain question answering dataset. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9645–9656, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.843/.

Fuwei Zhang, Zhao Zhang, Xiang Ao, Dehong Gao, Fuzhen Zhuang, Yi Wei, and Qing He. Mind the gap: Cross-lingual information retrieval with hierarchical knowledge enhancement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):4345–4353, June 2022. doi: 10.1609/aaai.v36i4.20355. URL https://ojs.aaai.org/index.php/AAAI/article/view/20355.

Hang Zhang, Yeyun Gong, Dayiheng Liu, Shunyu Zhang, Xingwei He, Jiancheng Lv, and Jian Guo. Knowledge enhanced pre-training for cross-lingual dense retrieval. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9810–9821, Torino, Italia, May 2024a. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.857/.

Lei Zhang, Michael Färber, and Achim Rettinger. 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*, CIKM '16, pages 2425–2428, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340731. doi: 10.1145/2983323.2983324. URL https://doi.org/10.1145/2983323.2983324.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.103. URL https://aclanthology.org/2024.emnlp-industry.103/.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense text retrieval based on pretrained language models: A survey. *ACM Trans. Inf. Syst.*, 42(4), February 2024. ISSN 1046-8188. doi: 10.1145/3637870. URL https://doi.org/10.1145/3637870.