计算机科学
COMPUTER SCIENCE

基于对齐查询的跨语言信息检索方法

李俊文, 宋雨秋, 张维彦, 阮彤, 刘井平, 朱焱

引用本文

李俊文, 宋雨秋, 张维彦, 阮彤, 刘井平, 朱焱. 基于对齐查询的跨语言信息检索方法[J]. 计算机科学, 2025, 52(8): 259-267.

LI Junwen, SONG Yuqiu, ZHANG Weiyan, RUAN Tong, LIU Jingping, ZHU Yan. Cross-lingual Information Retrieval Based on Aligned Query [J]. Computer Science, 2025, 52(8): 259-267.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241000055

基于对齐查询的跨语言信息检索方法

李俊文[1]　宋雨秋[2]　张维彦[2]　阮　彤[2]　刘井平[2]　朱　焱[1]
1 华东理工大学数学学院　上海 200237
2 华东理工大学信息科学与工程学院　上海 200237
(18602126280@163.com)

摘　要　跨语言信息检索是自然语言处理中一项重要的信息获取任务。最近,基于大语言模型的检索方法在这一任务中获得了广泛关注并取得了显著的进展。然而,现有基于提示大语言模型的无监督检索方法在效果和效率上仍有不足。对此,提出了一种全新的基于对齐查询的跨语言信息检索方法。具体而言,采用"预训练-微调"范式,基于预训练多语言模型提出了一种自适应的自指导编码器,通过同一语言内的检索学习指导跨语言检索学习。该方法引入与文档语种相同的语义对齐的查询,并设计了一种自适应的自指导机制,利用不同语种视角下的单语言检索结果的概率分布来指导跨语言检索。在22对语言组合上进行了广泛的实验来评估所提模型的有效性和效率,结果表明,所提方法的MRR指标达到了当前最先进水平。具体而言,其在高资源语种组合上相较于次优基线的平均MRR提高了15.45%,在低资源语种组合上相较于次优基线提高了18.9%。此外,相比基于大语言模型的方法,该方法在训练时间和推理时间上均更短,并且显著提升了收敛性能。相关代码已公开[1)]。

关键词:跨语言信息检索;对齐查询;自指导;自适应层级系数
中图分类号　TP391

Cross-lingual Information Retrieval Based on Aligned Query

LI Junwen[1],SONG Yuqiu[2],ZHANG Weiyan[2],RUAN Tong[2],LIU Jingping[2] and ZHU Yan[1]
1 School of Mathematics,East China University of Science and Technology,Shanghai 200237,China
2 School of Information Science and Engineering,East China University of Science and Technology,Shanghai 200237,China

Abstract　Cross-lingual Information Retrieval(CLIR) is an important information acquisition task in natural language processing. Recently,LLM-based retrieval methods have gained attention and demonstrated remarkable progress in this task. However,existing unsupervised retrieval methods based on prompting large language models still insufficient in effectiveness and efficiency. To solve this problem,this paper introduces a novel CLIR method based on aligned query. Specifically,this paper adopts the "pretrain-finetune" paradigm and proposes an adaptive self-teaching encoder based on a pretrained multilingual model to guide cross-lingual retrieval learning by mono-lingual retrieval learning. This method introduces semantically aligned queries in the same language as the documents and designs an adaptive self-teaching mechanism to guide cross-lingual retrieval by leveraging the probability distribution of mono-lingual retrieval results from different linguistic perspectives. To evaluate the effectiveness and efficiency of this method,this paper conducts extensive experiments on 22 language pairs. The results demonstrate that the proposed method achieves SOTA performance in terms of MRR. In particular,this method improves average MRR by 15.45% over the sub-optimal baseline in high-resource language pairs and 18.9% over the sub-optimal baseline in low-resource language pairs. Furthermore,the method reduces training and inference times compared to LLM-based approaches and exhibits faster convergence with enhanced stability.

Keywords　Cross-lingual Information Retrieval,Aligned query,Self-teaching,Adaptive layer-wise coefficient

1　引言

跨语言信息检索(Cross-lingual Information Retrieval,CLIR)旨在使用不同语言编写的查询来检索另一语言的相关文档,例如使用英文查询检索中文文档。因此,CLIR模型的性能依赖于两个关键因素:有效的查询-文档匹配以及跨语言间的差距弥合能力[1]。随着Google和Bing等搜索引擎服务于多语言用户群体,CLIR在全球信息交换时代逐渐成为不可或缺的工具[2]。此外,CLIR在解决跨语言任务(如开放域问答[3]、跨语言知识链接[4])方面也具有至关重要的作用[5]。

随着深度学习的发展,CLIR已经从基于翻译的方法、跨语言词嵌入方法和基于多语言预训练模型(Multilingual Pretrained Language Models,MPLMs)的密集检索方法发展起来。此外,在大语言模型(Large Language Models,LLMs)时

1)https://github.com/juneli6/CLIR

到稿日期:2024-10-12　返修日期:2025-01-25
通信作者:朱焱(zhuygraph@ecust.edu.cn)

代，基于 LLMs 的方法也被应用于 CLIR 任务中。这些方法通常采用生成范式，可以分为 3 种类型：第一类直接生成查询和单一文档之间的相关性[6]；第二类将查询和文档列表插入提示中，并引导 LLMs 输出重排后的文档标识符[7]；第三类向 LLMs 提供一个查询和一个文档对，然后引导模型输出相关度更高的文档标识符[8]，随后使用一系列聚合的方法来重排所有候选文档。

尽管这些方法都取得了一定效果，但是基于 LLMs 的 CLIR 方法仍然存在不足之处。1）LLMs 在许多下游任务中已经表现出语言偏差，这主要是预训练数据在不同语言间的不平衡分布所致[1]。例如，在 ChatGPT[9]，LLAMA[10] 和 LLAMA2[11] 等 LLMs 中，大约 90%的预训练语料是英文，而其他语言仅占很小一部分。因此，基于这些 LLMs 构建的 CLIR 模型可能会对低资源语言产生次优结果。2）部分 LLMs 在不同语言之间实现语义对齐的能力有限。这种局限性源于它们的预训练数据不包含跨语言对齐语料，或缺乏面向 CLIR 的预训练任务语料。因此，基于这些 LLMs 的方法可能会导致跨语言检索结果不理想。3）与 LLMs 相关的硬件和时间成本相对较高。大语言模型的参数量庞大，模型加载需要占用大量的 GPU 内存。此外，这些方法大多是生成式的，导致长文本的处理速度较慢。

为了解决上述问题，本文提出了一种基于对齐查询的跨语言信息检索方法。该方法的核心思想是利用同一语言内检索结果的概率分布来指导通过引入对齐查询、从不同语言视角获得的跨语言检索结果的概率分布。具体而言，为了减轻由语言偏差带来的性能下降，设计了一种自适应的自指导编码器，通过引入与源查询在语义上对齐的查询，利用单语言检索结果的概率分布来指导跨语言检索结果的概率分布。同时，考虑到 Transformer 的不同层所蕴含的句法知识和语义知识侧重不同，本文探索了自适应层级系数，以帮助模型对低资源语言学习进行更好的表示。在训练阶段，每层的损失值通过自适应层级系数加权相加，权重系数在训练过程中自适应调整。此外，本文方法在单语言检索数据和跨语言检索数据中应用对比学习，引导模型捕捉不同候选文档与查询之间的相关性差异，从而改善模型在跨语言信息检索任务中的整体表现。最后，为了减少硬件和时间成本，本文方法不使用 LLMs 作为模型的主干。

本文的贡献总结如下：

1）提出了一种基于对齐查询的跨语言信息检索方法，设计了一种跨语言对齐自指导机制，利用语义对齐的查询增强模型感知不同语言之间对齐知识的能力。

2）为了充分利用自指导机制的指导功能，本文在每个 Transformer 层引入了自适应系数，这些系数在训练阶段自主调整，从而自适应地引导模型进行自指导学习。

3）在公开的跨语言信息检索数据集上进行了广泛的实验，结果表明，该方法在高资源语言对中相比次优基线的平均 MRR 提高了 15.45 个百分点，在低资源语言对中提高了 18.9 个百分点。与基于 LLMs 的方法相比，本文方法更加高效，并且收敛速度更快、更稳定。

2 相关工作

当前用于 CLIR 的方法可分为 3 类：基于翻译的方法、基于 MPLMs 的密集检索方法和基于 LLMs 的方法。

1）基于翻译的方法

基于翻译的 CLIR 方法通常通过翻译将 CLIR 任务转换为单语言信息检索任务。翻译可以通过统计机器翻译或神经机器翻译来完成[1,12]。现有方法可以大致分为 3 类：翻译查询、翻译文档以及翻译查询和文档。翻译查询将查询翻译成与候选文档相同的语言[13]。翻译文档将候选文档翻译成与查询相同的语言[14]，然后应用单语言匹配模型确定相关性。翻译查询和文档将查询和候选文档都翻译成相同的中间语言[15]。然而，基于翻译的 CLIR 方法严重依赖于翻译系统的性能以及翻译系统所支持的语言数量的多样性，这极大地影响了检索结果的准确性，尤其是在法律等对精确性要求高的领域[16]。

2）基于 MPLMs 的密集检索

MPLMs 的发展显著推动了跨语言密集检索的进展。密集检索利用 MPLMs 将查询和文档编码为在相同向量空间内的低维嵌入表示[17-18]。许多跨语言预训练模型已被提出用于 CLIR，如 mBERT[19]，XLM-R[20] 和 VECO[21]。Litschko 等[22]对 mBERT 和 XLM 进行微调并用于 CLIR，强调了微调对于实现高效和有效的文档级结果的重要性。C3[23] 从对齐的双语文档中随机选择一个连续的片段并使用对比学习来预训练 MPLM，使语义上相似的片段更加接近，而非来自同一语言的不同片段的嵌入更接近。XPR[24] 使用包含查询短语的文本作为样本训练了一个查询短语提取器，以获取短语表征，并进一步使用双语短语对进行对比学习。与基于翻译的方法相比，基于 MPLMs 的密集检索能够更好地提取语义信息，尤其是那些传统方法无法捕捉到的更深层次信息。然而，现有方法并没有充分解决 MPLMs 感知不同语言之间的对齐语义能力不足的问题，特别是在低资源语言中。

3）基于 LLMs 的方法

最近提出的方法利用提示引导 LLMs 以一种无监督的方式增强 CLIR。这些提示策略可以分为 3 大类：第一类方法中，LLMs 在手工制作的提示指导下生成单一查询与单一文档之间的相关性[6]。这些方法主要依赖于提示，而手工设计的提示并不总是最优的。第二类方法中，提示包含单个查询和一个文档列表，用于指导 LLMs 生成重排后的文档标识符[7,25]。这类方法的性能对提示中文档列表的顺序高度敏感，并且依赖于更大参数量的 LLMs，导致硬件成本增加[26]。第三类方法中，LLMs 接收包含一个查询和一个文档对的提示词，并被引导来识别文档对中相关性更高的一个文档，随后使用各种聚合方法对所有候选文档进行重新排序[8]。上述方法均存在时间复杂度高的问题。

3 任务定义和整体框架

3.1 任务定义

给定一个查询 e 和一组候选文档 $D=\{d_1,d_2,\cdots,d_m\}$，CLIR 的任务是识别并对 D 中与 e 相关的文档进行排序，其中 m 代表候选文档的数量。值得注意的是，查询 e 的语言与候选文档 D 的语言不同。

3.2 整体框架

为了解决多语言预训练模型对不同语言之间的对齐知识敏感性不足导致的 CLIR 性能不佳的问题，本文提出了一种基于对齐查询的 CLIR 方法。该方法的总体框架如图 1 所

示。给定一个源语言查询e^s，首先将其与目标语言文档d^t连接，其中s代表源语言，t代表目标语言。其次，引入一个与e^s在语义上对齐的目标语言查询e^t，并将其与目标语言文档连接。这些连接后的文本被依次输入嵌入层并转换为嵌入表示，之后输入自适应的自指导编码器中。核心思想是利用单语言检索的概率分布来引导跨语言检索的概率分布。具体来说，使用KL散度将跨语言检索分布与单语言检索分布对齐。此外，本文引入了自适应系数，以基于KL损失值独立调整特定层的权重，从而充分利用不同Transformer层中的各类语义信息。最后，引入对比学习来帮助模型区分相关的候选文档。

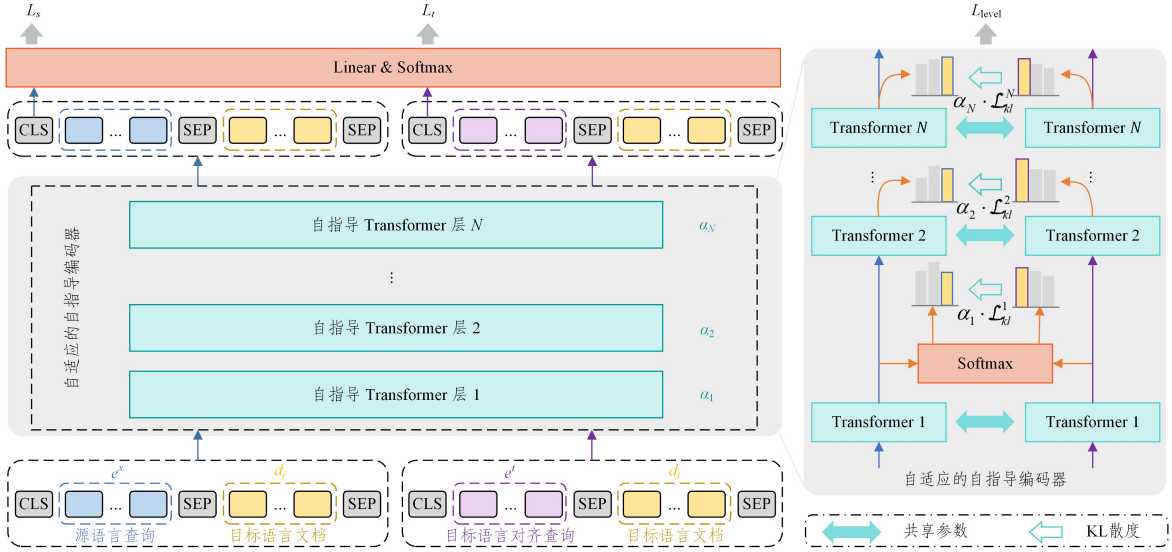

图1 基于对齐查询的跨语言信息检索模型框架

Fig. 1 Framework of cross-lingual information retrieval model based on aligned query

4 本文方法

4.1 跨语言对比学习

给定一个源语言查询和一些目标语言候选文档，跨语言信息检索(CLIR)的目标是建立查询与文档之间的语义相关性，从而提高检索结果的准确性。本文方法利用对比学习来建模语义相似性，是实现CLIR的一种高度有效机制[27-29]。

具体而言，给定一个源语言查询e^s和目标语言候选文档$D^t=\{d_1^t,d_2^t,\cdots,d_m^t\}$，首先将$d_j^t(1\leq j\leq m)$与$e^s$分别连接起来，然后输入多语言预训练语言模型(如mBERT)中，获得它们的特征表示，记为：

$$H_j^s=MPLM([\langle \mathrm{CLS}\rangle,e^s,\langle \mathrm{SEP}\rangle,d_j^t,\langle \mathrm{SEP}\rangle]) \tag{1}$$

其中，$\langle \mathrm{CLS}\rangle$和$\langle \mathrm{SEP}\rangle$是MPLM中的特殊标记。给定$H_j^s=\{h_{\langle \mathrm{CLS}\rangle},h_{e^s},h_{\langle \mathrm{SEP}\rangle},h_{d_j^t},h_{\langle \mathrm{SEP}\rangle}\}$，本文使用$\langle \mathrm{CLS}\rangle$的特征表示作为文本向量表示。为了便于阅读，如果文档与查询相关，则将其表示为正样本$h^+_{\langle \mathrm{cls}\rangle}$，否则将其表示为负样本$h^-_{\langle \mathrm{CLS}\rangle}$。然后，将正负样本对通过一个Linear&Softmax层获得相似性概率sim^+和sim^-，计算如下：

$$sim^+,sim^-=\mathrm{Softmax}([\mathrm{Linear}(h^+_{\langle \mathrm{CLS}\rangle}),\mathrm{Linear}(h^-_{\langle \mathrm{CLS}\rangle})]) \tag{2}$$

本文使用Pairwise Hinge Loss作为对比学习的损失函数。正负样本对的损失值计算式如下：

$$loss^s=\max(0,\Delta-sim^++sim^-) \tag{3}$$

其中，Δ是预设的边界值，表示正样本得分必须比负样本得分高出此边界值才满足要求。当有多个正负样本对时，记正负样本对的数量为Z。跨语言对比学习的损失计算式如下：

$$L_s=\frac{1}{Z}\sum_{z=1}^{Z}loss_z^s \tag{4}$$

4.2 自适应自指导编码器

本模块旨在额外利用同一语言内的相似性学习来指导跨语言相似性学习，因为在同一语言中捕捉相似语义更容易。为此，本模块设计了两个组件：单语言数据的对比学习以及自指导学习。

首先，通过翻译引入一个与e^s语义对齐的目标语言查询e^t。其次，与式(1)类似，将e^t和d_j^t连接起来，并将其输入相同的MPLM中，以获得单语言数据的向量表示$\boldsymbol{H}_j^t$。然后，使用与式(3)相同的计算方法，获得单语言数据上的对比学习损失值L_t。除此之外，本模块还引入了自指导学习来指导跨语言表示的学习，包括自指导机制和自适应系数。

4.2.1 自指导机制

给定单语言检索文本和跨语言检索文本的向量表示，这些向量通过参数共享的Transformer块获得，自指导机制旨在利用单语言学习结果的概率分布来指导跨语言向量表示的学习。

具体而言，给定跨语言文本的向量表示$\boldsymbol{H}_j^s$和单语言文本的向量表示$\boldsymbol{H}_j^t$，本文方法使用$\langle \mathrm{CLS}\rangle$作为单语言和跨语言文本的特征表示。为了便于理解，本文将$\boldsymbol{H}_j^s$中的$h^s_{\langle \mathrm{CLS}\rangle}$和$\boldsymbol{H}_j^t$中的$h^t_{\langle \mathrm{CLS}\rangle}$分别记作$\boldsymbol{v}_j^s$和$\boldsymbol{v}_j^t$。由于$\boldsymbol{v}_j^s$和$\boldsymbol{v}_j^t$的值不一定满足概率分布的性质，因此需要将它们通过Softmax层，以获得概率分布$P(\boldsymbol{v}_j^s)$和$P(\boldsymbol{v}_j^t)$：

$$\begin{aligned}P(\boldsymbol{v}_j^s)&=\mathrm{Softmax}(\boldsymbol{v}_j^s)=\frac{e^{\boldsymbol{v}_j^s}}{\sum_{r=1}^{R}e^{\boldsymbol{v}_{j(r)}^s}}\\ P(\boldsymbol{v}_j^t)&=\mathrm{Softmax}(\boldsymbol{v}_j^t)=\frac{e^{\boldsymbol{v}_j^t}}{\sum_{r=1}^{R}e^{\boldsymbol{v}_{j(r)}^t}}\end{aligned} \tag{5}$$

不同语言尽管字符不同，但它们在表达相同语义时，在MPLM向量空间中的分布应保持一致。因此本文使用KL散度作为向量分布一致性的衡量，以实现自指导机制[30]。故这一部分的损失函数可以表述为：

$$L_{KL}=D_{KL}(P(\boldsymbol{v}_j^s)\parallel P(\boldsymbol{v}_j^t))$$
$$=\sum_{r=1}^{R}P(\boldsymbol{v}_j^s(r)\log\left(\frac{P(\boldsymbol{v}_j^s(r)}{P(\boldsymbol{v}_j^t(r)}\right) \tag{6}$$

其中，R 是向量 $\boldsymbol{v}_j^s$ 和 $\boldsymbol{v}_j^t$ 的维度。

4.2.2 自适应系数学习

MPLM 包含多个 Transformer 层，不同层专注于不同的语言学知识。研究表明，较低层主要关注句法知识，而较高层关注语义知识[31]。因此，本文认为各层分布的一致性不同，并提出自适应系数学习，以充分利用自指导机制的指导功能。

具体来说，本文使用每层〈CLS〉的表示向量作为该层的特征表示，然后通过式(6)获取每层在自指导机制下的损失值 L_{KL}^i，其中 i 表示第 i 个 Transformer 层。此外，定义每层的权重系数为 α_i，并在训练过程中调整和更新。最终，损失函数可以表示为：

$$L_{level}=\sum_{i=1}^{N}\alpha_i L_{KL}^i \tag{7}$$

其中，N 是 MPLM 中 Transformer 层的数量。

初始时，$\alpha_i^{(q)}$ 进行随机初始化。在训练过程中，该系数通过 Softmax 层进行归一化，然后对每层的损失值进行加权。随着训练的进行，参数使用梯度下降算法进行更新，从而允许模型调整和更新系数。格式化表达如下：

$$\alpha_i^{(q)}=\frac{\exp(\alpha_i^{(q-1)})}{\sum_{j=1}^{N}\exp(\alpha_j^{(q-1)})} \tag{8}$$

其中，$\alpha_i^{(q)}$ 表示第 q 轮训练迭代中的模型第 i 层损失值的权重系数，$1\leqslant q\leqslant Q$。

4.3 训练和排序

在训练阶段，本文采用联合训练的方式来合并上述目标函数。最终的损失函数定义为：

$$L=L_s+L_t+L_{level} \tag{9}$$

其中，L_s 是跨语言对比学习的目标函数，L_t 是单语言对比学习的目标函数，L_{level} 是自适应系数学习的目标函数。

在排序阶段，本文首先按照式(1)中的方法将源语言查询 e^s 和目标语言文档 d_j^t 连接起来。其次，将连接后的结果 I_j^s 输入训练好的模型中。经过 N 层的 Transformer 块后输出特征表示 $\boldsymbol{H}_j^s=Transformer^{1\sim N}(I_j^s)$。然后，通过线性层将 $\boldsymbol{H}_j^s$ 中〈CLS〉的表示映射为标量，以获得模型预测 e^s 和 d_j^t 之间的相似性得分。最后，将所有预测的相似性得分按降序排序，排名列表作为 CLIR 的结果。

5 实验

本章进行了广泛的实验以验证所提出的 CLIR 方法的有效性。此外，还通过一系列消融实验分析了该方法的性能。

5.1 实验设置

5.1.1 数据集

本文使用来自 CLIRMatrix[32] 的 MULTI-8 作为实验数据集，其中查询包含 8 种语言的对齐版本。本文选择了 6 种语言：英语(EN)、中文(ZH)、法语(FR)、西班牙语(ES)、阿拉伯语(AR)和日语(JA)。具体来说，本文将 EN,ES,FR 和 ZH 分类为高资源语言，将 AR 和 JA 分类为低资源语言。高资源语言按对组合，形成 12 个语言对，低资源语言与 EN 和 ZH 配对，形成 12 个语言对。排除重复的语言对后，得到 22 个独特的语言对进行对比实验。每种语言对的训练集包含 10 000 个查询，每个查询有 100 个候选文档。本文随机采样 1 600 个查询用于训练。对于每个查询，随机采样一个正样本文档和一个负样本文档。验证集和测试集各包含 1 000 个查询。

5.1.2 基线

当前研究表明，大语言模型(LLMs)的表现通常优于小模型。因此，本文主要关注基于 LLMs 的方法，比较了 3 种 LLMs：LLaMA2-7b[11]，Falcon-7b[33] 和 Vicuna-7b[34]。具体来说，在训练阶段，本文利用 FL-tuning 策略[35]对 LLMs 进行微调，每层新添加的隐藏单元数量设置为 10。训练使用的指令数据格式如图 2 所示。

Below is an instruction that describes a task. Please make output that meets the requirements.

Instruction: Given a query entity and a document, determine whether the document is related to the query entity.
Query Entity: [source language query entity]
Document: [target language document]
If relevant, output 1, if not output 0.
Output:

图 2 基于 LLMs 的方法的训练数据格式

Fig. 2 Training data format for LLM-based method

在排序阶段，LLMs 预测输出“1”的概率被作为文档排序的相似性分数。此外，本文还比较了 mBERT 的性能。在 mBERT 方法中，将查询和候选文档进行连接并输入 mBERT 模型中，随后使用对比学习进行训练。mBERT 的输出作为二者的相似性分数进行排序。

5.1.3 评估指标

与大多数现有研究一致，本文采用 nDCG@k 和 MRR 作为评估指标。

nDCG@k(考虑前 k 个结果的归一化折损累计增益)：该指标用于衡量前 k 个排名结果的质量，值的范围为[0,1]，值越高代表排名质量越好。具体来说，当 $k=1$ 时，每个查询的计算式如下：

$$DCG@k=\sum_{i=1}^{k}\frac{2^{rel_i}-1}{\log_2(i+1)}$$
$$IDCG@k=\sum_{i=1}^{k}\frac{2^{label_i}-1}{\log_2(i+1)} \tag{10}$$
$$nDCG@k=\frac{DCG@k}{IDCG@k}$$

其中，rel_i 代表在预测结果的排序中，第 i 个候选文档对查询的相关性得分；$label_i$ 表示在最优排名结果中，第 i 个候选文档对查询的相关性标签。具体来说，本节的实验数据取每个测试集中的 1 000 个查询的 nDCG@k 值的平均值。

MRR(平均倒数排名)：该指标用于衡量结果中最相关的文档所出现位置的排名质量。具体来说，对于每个查询，在预测结果排序中找到最相关的候选文档的排名，然后计算其倒数 $\frac{1}{rank_i}$。最后，对每个查询的 MRR 值进行平均，计算式如下：

$$MRR=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{rank_i} \tag{11}$$

其中，n 是查询的总数。

5.1.4 参数设置

本文 CLIR 方法使用多语言预训练模型作为文本编码器。具体来说，本文使用 bert-base-multilingual-cased 模型。在训练过程中，使用 AdamW 优化器，学习率为 1×10^{-5}，训练的 epoch 为 15，每个 epoch 的批量大小为 4，每个 epoch 结束

后进行验证。实验设备是一台带有 24 GB 显存的 NVIDIA GeForce RTX 4090 GPU 服务器。

5.2 主要结果

为了充分验证所提出的 CLIR 方法的有效性，本文不仅在高资源语言（EN，ES，FR，ZH）上进行实验，还在低资源语言（AR，JA）上进行实验，一共 22 对语言组合。这些语言组合的结果表明，本文方法在 MRR 指标上优于所有基线方法。

5.2.1 高资源语言组合结果

将所提出的 CLIR 方法与基线方法在高资源语言组合（EN，ES，FR 和 ZH）上进行比较，结果如表 1 所列。

表 1 高资源语言对的 nDCG@k 和 MRR 结果

Table 1 Results of nDCG@k and MRR for high-resource language pairs

（%）

Language Pair	Metric	mBERT	LLaMA2	Falcon	Vicuna	Ours
EN-ES	nDCG@1	65.73	56.98	62.57	57.90	**85.89**
	nDCG@10	76.68	78.08	78.51	77.57	**84.98**
	nDCG@20	78.28	80.12	80.60	80.43	**85.35**
	MRR	70.33	34.32	41.77	34.80	**87.46**
EN-FR	nDCG@1	68.86	62.12	64.17	62.01	**84.60**
	nDCG@10	76.02	79.59	79.13	78.51	**83.80**
	nDCG@20	78.23	81.18	80.72	80.98	**84.57**
	MRR	73.43	36.97	41.73	37.08	**86.70**
EN-ZH	nDCG@1	69.05	48.62	50.95	50.67	**81.31**
	nDCG@10	78.67	77.39	77.69	75.61	**84.37**
	nDCG@20	80.14	79.38	80.76	79.92	**85.98**
	MRR	73.36	30.08	31.81	32.45	**83.95**
ES-EN	nDCG@1	61.33	66.87	69.50	67.30	**84.18**
	nDCG@10	65.74	82.98	**84.66**	83.69	82.79
	nDCG@20	67.52	85.04	**85.42**	84.56	84.23
	MRR	64.45	35.84	43.00	36.27	**84.00**
ES-FR	nDCG@1	62.61	61.70	62.83	61.50	**71.12**
	nDCG@10	72.05	80.33	80.11	79.24	**83.90**
	nDCG@20	75.34	82.84	82.86	82.63	**85.71**
	MRR	62.09	33.15	34.84	30.54	**70.60**
ES-ZH	nDCG@1	59.23	54.67	55.38	53.87	**78.61**
	nDCG@10	69.22	75.12	76.39	75.86	**82.81**
	nDCG@20	72.77	78.35	80.08	78.76	**84.93**
	MRR	62.44	27.98	31.60	29.27	**79.24**
FR-EN	nDCG@1	59.52	66.42	70.82	67.07	**82.63**
	nDCG@10	65.70	83.67	**84.44**	82.69	81.60
	nDCG@20	67.65	84.72	**85.11**	83.75	83.88
	MRR	61.83	35.21	44.97	36.37	**80.84**
FR-ES	nDCG@1	60.11	60.73	64.05	61.35	**71.83**
	nDCG@10	69.69	77.91	80.76	78.28	**82.48**
	nDCG@20	71.90	80.64	81.49	80.95	**82.89**
	MRR	63.34	32.44	42.04	36.82	**73.49**
FR-ZH	nDCG@1	51.72	51.62	53.55	55.03	**75.24**
	nDCG@10	66.17	74.80	74.24	75.78	**79.87**
	nDCG@20	69.32	77.59	77.32	78.54	**83.46**
	MRR	56.63	26.61	30.22	32.36	**76.94**
ZH-EN	nDCG@1	46.97	67.20	**68.35**	66.55	66.08
	nDCG@10	57.61	82.04	**82.95**	82.13	72.83
	nDCG@20	60.06	83.25	**83.77**	82.14	75.90
	MRR	52.61	33.14	36.04	31.19	**70.81**
ZH-ES	nDCG@1	41.69	60.97	**64.13**	60.83	59.46
	nDCG@10	59.91	79.11	**80.19**	79.58	70.43
	nDCG@20	65.57	80.56	81.23	**81.44**	73.28
	MRR	48.06	31.20	35.53	30.59	**64.85**
ZH-FR	nDCG@1	44.32	61.70	**63.43**	62.75	59.90
	nDCG@10	59.43	79.17	79.62	**80.42**	69.76
	nDCG@20	64.71	80.25	80.58	**81.84**	73.49
	MRR	49.46	30.87	32.54	30.47	**64.57**
Avg.	nDCG@1	57.60	59.97	62.48	60.57	**75.07**
	nDCG@10	68.07	79.18	79.89	79.11	**79.97**
	nDCG@20	70.96	81.16	81.66	81.33	**81.97**
	MRR	61.50	32.32	37.17	33.18	**76.95**

观察表 1 中的结果，可以得出以下结论：1）本文方法在 MRR 指标上优于所有基线方法，具体而言，本文方法比次优基线的平均 MRR 高出 15.45 个百分点，表明了本文方法的有效性；2）在所有基线方法中，基于 LLMs 的方法比 mBERT 表现更好，但本文方法在大多数高资源语言组合上的表现（nDCG@1 和 MRR）仍然优于基于 LLMs 的方法，表明本文方法在最优结果的排名和整体排名质量方面表现更好，原因在于本文采用了联合训练方法，不仅结合了单语言和跨语言对比学习，还引入了自适应分层自指导机制。

5.2.2 低资源语言组合结果

为了展示所提出的 CLIR 方法在低资源语言上的有效性，在 AR 和 JA 上进行了实验，结果如表 2 所列。

表 2 低资源语言对的 nDCG@k 和 MRR 结果

Table 2 Results of nDCG@k and MRR for low-resource language pairs

（%）

Language Pair	Metric	mBERT	LLaMA2	Falcon	Vicuna	Ours
EN-ES	nDCG@1	64.51	54.23	50.25	47.08	**82.16**
	nDCG@10	75.38	74.51	73.57	72.44	**84.49**
	nDCG@20	77.10	77.58	76.70	76.64	**85.04**
	MRR	67.88	37.74	30.44	27.10	**85.30**
EN-JA	nDCG@1	74.65	54.02	50.32	49.38	**88.38**
	nDCG@10	79.02	73.94	74.84	72.92	**87.00**
	nDCG@20	81.53	76.26	75.51	74.73	**88.35**
	MRR	77.63	32.87	32.16	27.48	**89.97**
AR-EN	nDCG@1	44.46	**65.92**	65.73	65.90	61.90
	nDCG@10	67.92	**81.50**	81.27	81.34	77.81
	nDCG@20	70.29	82.88	**83.71**	82.04	81.09
	MRR	46.93	30.85	29.83	27.84	**63.77**
AR-JA	nDCG@1	43.33	57.08	60.55	59.30	**62.38**
	nDCG@10	56.32	78.60	79.71	79.52	75.93
	nDCG@20	61.15	80.81	81.02	80.14	77.26
	MRR	46.21	27.18	31.54	30.70	**65.46**
AR-ZH	nDCG@1	42.59	58.47	59.03	58.82	**69.83**
	nDCG@10	60.50	79.11	78.99	78.95	**80.46**
	nDCG@20	67.25	81.33	80.73	81.13	**83.61**
	MRR	47.21	29.99	28.51	29.60	**75.93**
JA-AR	nDCG@1	46.24	57.90	55.35	57.35	**64.07**
	nDCG@10	62.21	76.21	76.63	**77.66**	76.05
	nDCG@20	71.06	78.27	78.64	**79.70**	78.82
	MRR	52.91	30.27	26.10	29.79	**70.14**
JA-EN	nDCG@1	54.42	68.60	66.52	70.70	**71.18**
	nDCG@10	64.25	84.52	82.07	**86.38**	78.25
	nDCG@20	67.37	84.67	83.20	**87.11**	82.75
	MRR	60.54	35.83	30.98	40.85	**75.64**
JA-ZH	nDCG@1	55.31	64.05	61.18	61.68	**77.47**
	nDCG@10	69.40	**81.09**	80.79	80.04	78.66
	nDCG@20	75.31	**82.03**	81.58	81.79	80.34
	MRR	60.20	34.56	29.59	30.58	**80.74**
ZH-AR	nDCG@1	35.11	59.10	57.85	59.40	**60.78**
	nDCG@10	55.55	**78.46**	78.22	77.52	75.76
	nDCG@20	60.63	**80.24**	79.19	78.93	77.61
	MRR	40.78	31.46	29.48	31.08	**66.21**
ZH-JA	nDCG@1	56.63	62.25	61.63	61.02	75.63
	nDCG@10	66.68	80.22	79.97	**80.89**	75.16
	nDCG@20	70.49	81.57	81.04	**81.62**	77.23
	MRR	62.14	32.24	31.07	29.97	**78.24**
Avg.	nDCG@1	51.73	60.16	58.84	59.06	**71.38**
	nDCG@10	65.72	78.82	78.61	78.77	**78.96**
	nDCG@20	70.22	80.56	80.13	80.38	**81.21**
	MRR	56.24	32.30	29.97	30.50	**75.14**

可以观察到，与高资源语言组合类似，本文方法的 nDCG@1 和 MRR 指标优于所有基线方法，表现极为出色。这一表现为该方法的有效性提供了有力证据。具体而言，本文方法的

平均 MRR 比次优基线提高了 18.9 个百分点，平均 nDCG@1 比次优基线提高了 11.22 个百分点。

5.3 详细分析

本节将对提出的方法进行全面分析，包括消融研究、自适应系数的评估以及模型效率和收敛性的评估。

5.3.1 消融实验

对于 CLIR 任务，本文提出了一个自适应的自指导编码器，它利用单语言中的相似性学习来指导跨语言的相似性学习。为了验证该模块的有效性，进行了消融实验。具体来说，依次移除一个组件并评估其余组件组合的有效性。实验结果如表 3 所列。“wo/ level”指移除自适应系数学习，仅在第 N 层的输出分布上应用自指导机制；“wo/ KL”代表完全移除了自指导机制。在表 3 中，最佳结果用粗体表示，次优结果用下划线标出。

表 3 基于对齐查询的 CLIR 方法的消融实验

Table 3 Ablation study for CLIR method based on aligned queries

(%)

Language Pair	Metric	wo/KL	wo/level	Ours	Language Pair	Metric	wo/KL	wo/level	Ours
EN-AR	nDCG@1	69.83	79.25	**82.16**	ZH-AR	nDCG@1	53.28	53.82	**60.78**
	MRR	75.93	83.04	**5.30**		MRR	60.72	60.45	**66.21**
EN-ES	nDCG@1	75.34	79.12	**85.89**	ZH-EN	nDCG@1	57.33	64.74	**66.08**
	MRR	80.23	81.66	**87.46**		MRR	63.76	68.84	**70.81**
EN-FR	nDCG@1	74.70	83.42	**85.60**	ZH-ES	nDCG@1	53.09	56.45	**59.46**
	MRR	78.62	85.02	**86.70**		MRR	59.03	62.45	**64.85**
EN-JA	nDCG@1	75.92	85.10	**88.38**	ZH-FR	nDCG@1	54.01	55.82	**59.90**
	MRR	80.27	86.55	**89.97**		MRR	59.38	62.17	**64.57**
EN-ZH	nDCG@1	70.68	77.18	**81.31**	ZH-JA	nDCG@1	68.00	72.27	**75.63**
	MRR	77.08	80.02	**83.95**		MRR	72.00	75.69	**78.24**
ES-EN	nDCG@1	76.89	80.54	**84.18**	AR-EN	nDCG@1	53.13	61.61	**61.90**
	MRR	77.05	79.84	**84.00**		MRR	55.56	63.05	**63.77**
ES-FR	nDCG@1	67.83	68.80	**72.12**	AR-JA	nDCG@1	56.33	61.36	**62.38**
	MRR	67.46	67.72	**70.60**		MRR	60.78	65.30	**65.46**
ES-ZH	nDCG@1	72.63	76.87	**78.61**	AR-ZH	nDCG@1	57.84	56.43	**69.83**
	MRR	75.22	78.30	**79.24**		MRR	61.19	59.90	**75.93**
FR-EN	nDCG@1	72.45	79.86	**82.63**	JA-AR	nDCG@1	50.32	60.31	**64.07**
	MRR	72.70	78.39	**80.84**		MRR	58.27	66.22	**70.14**
FR-ES	nDCG@1	63.48	69.79	**71.83**	JA-EN	nDCG@1	61.25	64.06	**71.18**
	MRR	67.07	71.34	**73.49**		MRR	67.62	68.65	**75.64**
FR-ZH	nDCG@1	61.91	71.81	**75.24**	JA-ZH	nDCG@1	72.38	73.81	**77.47**
	MRR	66.53	74.77	**76.94**		MRR	77.02	77.16	**80.74**

从表 3 中可以观察到：

1）每个组件的移除都会导致一定程度的性能下降，这表明所提出的自指导机制和自适应系数都对模型在 CLIR 中的表现产生了正向影响；

2）比较“wo/ level”与“wo/ KL”，可以观察到，完全移除自指导机制时，性能下降更明显。具体而言，对于 JA-AR，“wo/ level”和“wo/ KL”分别导致 nDCG@1 下降了 3.76 个百分点和 13.75 个百分点；对于 FR-ZH，“wo/ level”和“wo/ KL”分别导致 MRR 下降了 2.17 个百分点和 10.41 个百分点。这证明了自指导机制的有效性，也表明自适应系数更好地发挥了自指导机制的指导学习作用，从不同语言角度全面实现了指导功能。

5.3.2 自适应层级系数的评估

在自适应自指导编码器中，本文提出了自适应系数学习，以实现基于不同语言视角的自指导功能。为了深入研究自适应系数学习组件的影响，本小节在两种不同的系数设置下进行了比较实验。实验结果如表 4 所列。“Same”表示在每个 Transformer 层上的 KL 散度损失的权重系数均为 1；“Linear”表示在第 j 个 Transformer 层上的 KL 散度损失的权重系数为 $j/10$。

表 4 自适应层级系数的消融实验

Table 4 Ablation study for adaptive hierarchical coefficients

(%)

Language Pair	Metric	Same	Linear	Ours	Language Pair	Metric	Same	Linear	Ours
EN-AR	nDCG@1	70.28	79.65	**82.16**	ZH-AR	nDCG@1	48.72	57.34	**60.78**
	MRR	75.02	82.33	**85.30**		MRR	54.51	63.75	**66.21**
EN-ES	nDCG@1	81.73	85.70	**85.8**	ZH-EN	nDCG@1	60.80	64.28	**66.08**
	MRR	84.57	86.31	**87.46**		MRR	64.79	68.79	70.81
EN-FR	nDCG@1	75.06	83.42	**84.60**	ZH-ES	nDCG@1	58.18	58.40	**59.46**
	MRR	77.75	85.02	**86.70**		MRR	64.57	63.92	**64.85**
EN-JA	nDCG@1	76.58	87.83	**88.38**	ZH-FR	nDCG@1	50.35	57.09	**59.90**
	MRR	79.95	85.43	**89.97**		MRR	54.84	62.87	**64.57**
EN-ZH	nDCG@1	75.34	72.55	**81.31**	ZH-JA	nDCG@1	74.87	72.27	**75.63**
	MRR	79.78	77.84	**83.95**		MRR	78.16	75.87	**78.24**

（续表）

Language Pair	Metric	Same	Linear	Ours	Language Pair	Metric	Same	Linear	Ours
ES-EN	nDCG@1	80.54	83.15	**84.18**	AR-EN	nDCG@1	49.11	53.02	**61.90**
	MRR	79.84	80.53	**84.00**		MRR	50.58	55.09	**63.77**
ES-FR	nDCG@1	66.05	68.80	**72.12**	AR-JA	nDCG@1	54.80	61.81	**62.38**
	MRR	64.16	67.22	**70.60**		MRR	58.11	**65.59**	65.46
ES-ZH	nDCG@1	77.07	77.35	**78.61**	AR-ZH	nDCG@1	60.58	58.47	**69.83**
	MRR	78.26	78.17	**79.24**		MRR	63.36	62.47	**75.93**
FR-EN	nDCG@1	82.44	80.73	**82.63**	JA-AR	nDCG@1	62.83	63.05	**64.07**
	MRR	80.61	78.89	**0.84**		MRR	68.94	69.54	**70.14**
FR-ES	nDCG@1	69.79	71.75	**71.83**	JA-EN	nDCG@1	63.06	64.06	**71.18**
	MRR	71.34	72.28	**73.49**		MRR	67.78	70.14	**75.64**
FR-ZH	nDCG@1	63.62	71.49	**75.24**	JA-ZH	nDCG@1	75.42	75.90	**77.47**
	MRR	67.37	75.11	**76.94**		MRR	79.13	79.68	**80.74**

从表 4 的结果可以得出以下结论：

1）自适应系数在大多数情况下表现出显著优势，尽管在某些语言组合的部分指标上略低于“Linear”。具体来说，对于 ZH-AR，自适应系数学习的 nDCG@1 得分比“Same”和“Linear”分别高出 12.06 个百分点和 3.44 个百分点。

2）在大多数情况下，“Linear”设置的表现优于“Same”，这表明每一层的结果分布对目标函数学习的有效性并不相同。“Linear”设置下每一层的 KL 损失权重不相同，是因为不同层的作用不同，模型低层往往学习语法知识，而高层往往学习语义知识[31]。相比之下“Same”假设每一层的有效性相同，而实验结果表明这并不正确。

5.3.3 效率评估

为了展示本文方法的效率，将本文方法与基于 LLMs 的方法在训练时间和推理时间上进行了比较。为了公平对比，两种方法的机器配置相同。具体来说，基于 LLMs 的方法使用 LLaMA2 作为骨干。训练和推理时间的比较结果如图 3 所示。可以观察到：

1）在训练阶段，除 EN-ZH 语言组合，两种方法的大部分语言组合的时间消耗差异并不显著。对于 EN-ZH 语言组合，基于 LLMs 的方法的训练时间几乎是本文方法的两倍。

2）在推理阶段，两种方法的时间消耗差异十分明显，基于 LLMs 的方法通常比本文方法花费更长时间。具体来说，对于 EN-ZH，基于 LLMs 的方法的推理时间是本文方法的两倍多。本文方法在训练和推理时间上均优于基于 LLMs 的方法，主要原因在于本文方法采用的骨干网络的参数量（mBERT，参数量为 1.1 亿）远小于基于 LLMs 的模型（Llama2-7B，参数量为 70 亿），因此模型在前向计算和反向传播过程中的计算量更少，从而减少了整体的计算时间。值得注意的是，本小节所比较的基于 LLMs 的推理时间是对每个查询设置 10 个候选文档。如果两种方法保持相同数量的候选文档，基于 LLMs 的方法的推理时间会更长。

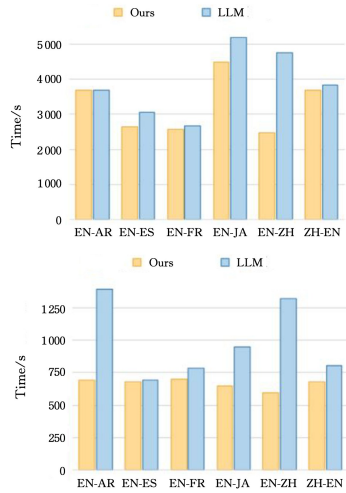

图 3 本文方法和基于 LLMs 的方法的耗时对比

Fig. 3 Duration comparison of the proposed method and LLM-based method

5.3.4 收敛性评估

为了进一步证明本文方法的效率，将其与基于 LLMs 的方法进行收敛性的比较。训练过程中的损失值变化情况如图 4 所示。

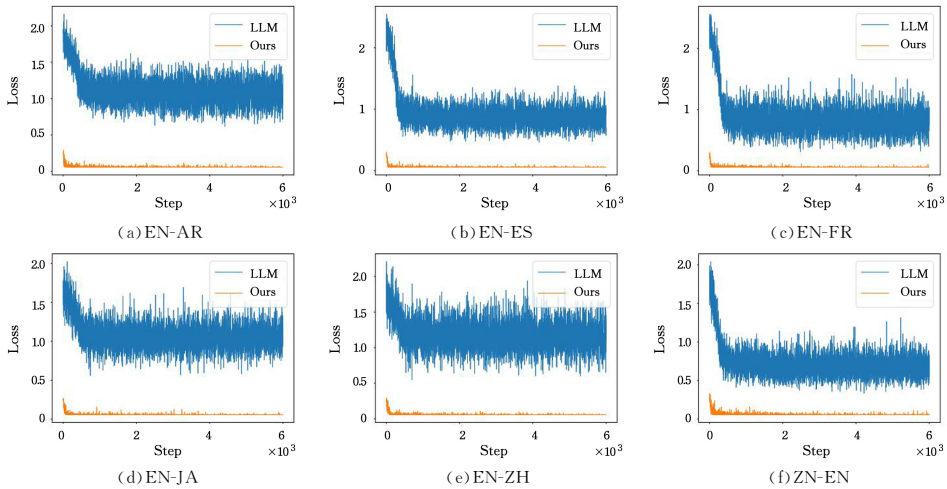

图 4 本文方法与基于 LLMs 的方法在部分语言组合上的 loss 曲线对比

Fig. 4 Loss curves comparison of the proposed method and LLM-based method on some language pairs

从图中可以清楚地观察到：1)本文方法的收敛速度显著快于基于 LLMs 的方法；2)基于 LLMs 的方法在训练过程中损失值的波动剧烈，而本文方法相对平滑且更稳定。

结束语 本文提出了一种基于对齐查询的全新跨语言信息检索方法。具体来说，对于跨语言数据，该方法利用对比学习来建模查询和文档之间的语义相似性。在此基础上，进一步提出了一种自适应的自指导编码器，通过单语言学习来引导跨语言学习。具体而言，该方法首先在与文档相同的语言中引入语义对齐查询，并将对比学习应用于单语言数据。然后，设计了一种自指导机制并采用自适应系数学习，利用单语言检索结果的概率分布来引导跨语言检索。本文在 22 对语言对上进行了广泛的实验。结果表明，所提出的方法的平均 MRR(平均倒数排名)指标达到了当前最优性能。具体来说，本文方法在高资源语言对中相比次优基线提高了 15.45 个百分点的平均 MRR，在低资源语言对中相比次优基线提高了 18.9 个百分点的平均 MRR。此外，与基于 LLMs 的方法相比，本文方法在训练时间和推理时间上更快，且在收敛性能上也表现出显著的提升。

未来，将进一步扩展该方法的通用性。首先，由于当前方法基于 mBERT 进行文本嵌入表示，而 mBERT 在知识储备和最大上下文长度方面存在一定局限，考虑到目前快速发展的 LLMs 在知识量与文本长度上有很大改善，探索以 LLMs 作为文本表征基座来进行 CLIR 的方法具有重要意义。其次，在跨语言检索任务中，何种语言的知识文档包含回答问题证据的先验不可知，通常需要事先指定检索目标语种。因此，研究自适应选择检索目标语种的方法，将极大地提高跨语言检索系统的灵活性。

参 考 文 献

[1] HUANG Z, YU P, ALLAN J. Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation[C]// Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. 2023: 1048-1056.

[2] ZHANG S, LIANG Y, GONG M, et al. Modeling sequential sentence relation to improve cross-lingual dense retrieval[J]. arXiv: 2302.01626, 2023.

[3] LI Z J, LI S H. Survey on Web-based Question Answering [J]. Computer Science, 2017, 44(6): 1-7, 42.

[4] YU Y Y, CHAO W H, HE Y Y, et al. Cross-language Knowledge Linkage Based on Bilingual Topic Model and Bilingual Word Vectors [J]. Computer Science, 2019, 46(1): 238-244.

[5] WANG Y, REN R, LI J, et al. REAR: A Relevance-Aware Retrieval-Augmented Framework for Open-Domain Question Answering[J]. arXiv: 2402.17497, 2024.

[6] ZHUANG H, QIN Z, HUI K, et al. Beyond yes and no: Improving zero-shot llm rankers via scoring fine-grained relevance labels[J]. arXiv: 2310.14122, 2023.

[7] SUN W, YAN L, MA X, et al. Is ChatGPT good at search? investigating large language models as re-ranking agents[J]. arXiv: 2304.09542, 2023.

[8] QIN Z, JAGERMAN R, HUI K, et al. Large language models are effective text rankers with pairwise ranking prompting[J]. arXiv: 2306.17563, 2023.

[9] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report[J]. arXiv: 2303.08774, 2023.

[10] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models [J]. arXiv: 2302.13971, 2023.

[11] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv: 2307.09288, 2023.

[12] LIU J P, SU J S, HUANG D G. Incorporating Language-specific Adapter into Multilingual Neural Machine Translation[J]. Computer Science, 2022, 49(1): 17-23.

[13] ELAYEB B, ROMDHANE W B, SAOUD N B B. Towards a new possibilistic query translation tool for cross-language information retrieval[J]. Multimedia Tools and Applications, 2018, 77: 2423-2465.

[14] AZARBONYAD H, SHAKERY A, FAILI H. A learning to rank approach for cross-language information retrieval exploiting multiple translation resources[J]. Natural Language Engineering, 2019, 25(3): 363-384.

[15] KISHIDA K, KANDO N. Two-stage refinement of query translation in a pivot language approach to cross-lingual information retrieval: An experiment at CLEF 2003[C]// Workshop of the Cross-Language Evaluation Forum for European Languages. Berlin: Springer, 2003: 253-262.

[16] TASHU T M, KONTOS E R, SABATELLI M, et al. Mapping Transformer Leveraged Embeddings for Cross-Lingual Document Representation[J]. arXiv: 2401.06583, 2024.

[17] LIN J A, BAO C Z, DONG J F, et al. Multilingual Text-Video Cross-Modal Retrieval Model via Multilingual-Visual Common Space Learning[J]. Journal of Computer Science, 2024, 47(9): 2195-2210.

[18] ZOU A, HAO W N, JIN D W, et al. Study on Text Retrieval Based on Pre-training and Deep Hash [J]. Computer Science, 2021, 48(11): 300-306.

[19] QIU X, WANG Y, SHI J, et al. Cross-Lingual Transfer for Natural Language Inference via Multilingual Prompt Translator [J]. arXiv: 2403.12407, 2024.

[20] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised cross-lingual representation learning at scale[J]. arXiv: 1911.02116, 2019.

[21] LUO F, WANG W, LIU J, et al. VECO: Variable and flexible cross-lingual pre-training for language understanding and generation[J]. arXiv: 2010.16046, 2020.

[22] LITSCHKO R, VULIĆ I, PONZETTO S P, et al. On cross-lingual retrieval with multilingual text encoders[J]. Information Retrieval Journal, 2022, 25(2): 149-183.

[23] YANG E, NAIR S, CHANDRADEVAN R, et al. C3: Continued pretraining with contrastive weak supervision for cross language ad-hoc retrieval [C] // Proceedings of the 45th International

ACM SIGIR Conference on Research and Development in Information Retrieval. 2022:2507-2512.

[24] ZHENG H,ZHANG X,CHI Z,et al. Cross-lingual phrase retrieval[J]. arXiv:2204.08887,2022.

[25] MA X,ZHANG X,PRADEEP R,et al. Zero-shot listwise document reranking with a large language model[J]. arXiv:2305.02156,2023.

[26] CHEN X T,YE J J,ZU C,et al. Robustness of GPT Large Language Models on Natural Language Processing Tasks [J]. Journal of Computer Research and Development,2024,61(5):1128-1142.

[27] IZACARD G,CARON M,HOSSEINI L,et al. Unsupervised dense information retrieval with contrastive learning[J]. arXiv:2112.09118,2021.

[28] QU Y,DING Y,LIU J,et al. RocketQA:An optimized training approach to dense passage retrieval for open-domain question answering[J]. arXiv:2010.08191,2020.

[29] XIONG L,XIONG C,LI Y,et al. Approximate nearest neighbor negative contrastive learning for dense text retrieval[J]. arXiv:2007.00808,2020.

[30] HUANG Z H,YANG S Z,LIN W,et al. Knowledge Distillation:A Survey [J]. Journal of Computer Science,2022,45(3):624-653.

[31] JAWAHAR G,SAGOT B,SEDDAH D. What does BERT learn about the structure of language? [C]// 57th Annual Meeting of the Association for Computational Linguistics. 2019.

[32] SUN S,DUH K. CLIRMatrix:A massively large collection of bilingual and multilingual datasets for Cross-Lingual Information Retrieval[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020:4160-4170.

[33] PENEDO G,MALARTIC Q,HESSLOW D,et al. The RefinedWeb dataset for Falcon LLM:outperforming curated corpora with web data,and web data only[J]. arXiv:2306.01116,2023.

[34] ZHENG L,CHIANG W L,SHENG Y,et al. Judging LLM-as-a-judge with MT-bench and chatbot arena[J]. arXiv:2306.05685,2024.

[35] LIU J,SONG Y,XUE K,et al. Fl-tuning:Layer tuning for feed-forward network in transformer[J]. arXiv:2206.15312,2022.

LI Junwen, born in 2001, postgraduate. His main research interests include natural language processing and information retrieval.

ZHU Yan, born in 1984, Ph.D, associate professor. Her main research interest is graph theory and its applications.

(责任编辑:何杨)