通过查询生成增强段落表示以提升跨语言密集检索

Shengyao Zhuang 昆士兰大学 澳大利亚昆士兰州布里斯班 s.zhuang@uq.edu.au

Linjun Shou Microsoft STCA 中国北京 lisho@microsoft.com

Guido Zuccon 昆士兰大学 澳大利亚昆士兰州布里斯班 g.zuccon@uq.edu.au

arXiv:2305.03950v1 [cs.IR] 6 May 2023

摘要

有效的跨语言稠密检索方法依赖多语言预训练语言模型（PLM），需要经过训练，以同时涵盖相关性匹配任务和跨语言对齐任务。然而，可用于训练的跨语言数据往往十分有限。本文提出，不再使用更多跨语言数据进行训练，而是通过跨语言查询生成，利用原始段落语言之外的其他语言生成的查询来增强段落表示。这些增强后的表示用于推理时，使表示能够编码不同目标语言中的更多信息。跨语言查询生成器的训练无需额外的训练数据，只需使用稠密检索器所用的数据。查询生成器的训练也很有效，因为生成器的预训练任务（T5 文本到文本训练）与微调任务（生成查询）非常相似。在推理时使用生成器不会增加查询延迟，并且可以与任何跨语言稠密检索方法结合使用。在基准跨语言信息检索数据集上的实验结果表明，我们的方法能够提升现有跨语言稠密检索方法的效果。我们的方法实现以及所有生成的查询文件均已公开，地址为 https://github.com/ielab/xQG4xDR。

CCS 概念

• **信息系统** → **查询表示**；**语言模型**。

关键词

跨语言查询生成；跨语言检索；稠密检索器

ACM 引用格式：

Shengyao Zhuang、Linjun Shou 和 Guido Zuccon。2023。通过查询生成增强段落表示以提升跨语言密集检索。载于第*46*届 *ACM SIGIR* 信息检索研究与发展国际会议论文集（*SIGIR '23*），*2023*年*7*月*23*日至*27*日，台湾台北。ACM，纽约州纽约市，6页。https://doi.org/10.1145/3539618.3591952

允许为个人或课堂使用而免费制作本作品的全部或部分数字或纸质副本，但前提是不得为营利或商业利益而制作或分发副本，并且副本首页须附有本声明及完整引文。对于本作品中由作者以外的其他方拥有版权的组成部分，必须尊重其版权。注明出处后可进行摘要。若要以其他方式复制、再版、发布到服务器或重新分发到列表，须事先获得特定许可和/或支付费用。请通过 permissions@acm.org 申请许可。

SIGIR '23，*2023*年*7*月*23*日至*27*日，台北，台湾 © 2023 版权所有归所有者/作者所有。出版权授权予 ACM。ACM ISBN 978-1-4503-9408-6/23/07...$15.00 https://doi.org/10.1145/3539618.3591952

1 引言

基于预训练语言模型（PLM）的稠密检索器（DR）在纯英语段落检索任务中取得了显著成功 [12–14, 20, 21, 26, 28, 29, 40, 43, 44]。这些模型采用双编码器架构，利用 PLM 编码器将查询和段落编码为稠密嵌入。随后，它们在嵌入空间中执行近似最近邻（ANN）搜索。与传统的词袋方法相比，DR 能够进行语义软匹配，有助于克服段落检索中的词语不匹配问题 [33, 45]。

为了发挥密集检索器的语义建模能力，近期研究已将仅支持英语的密集检索器扩展至支持跨语言设置 [1, 2, 19, 22, 27, 31]，即查询和段落使用不同语言的情况。这是通过使用多语言预训练语言模型（PLM）实现的，例如用多语言 BERT [6], 取代仅支持英语的预训练语言模型。这种方法在该设置下尤为重要，因为由于不同语言之间匹配词项数量有限，传统的词袋方法难以奏效。相比之下，跨语言密集检索器（xDR）能够将不同语言的查询和段落编码到共享的嵌入空间中，从而支持跨语言的高效近似最近邻（ANN）搜索。然而，与仅支持英语的设置下的密集检索器相比，基于多语言预训练语言模型的 xDR 通常在跨语言段落检索任务上的效果较差 [1]。对此的一种解释是，在仅支持英语的设置下，密集检索器只需学习查询与段落之间的相关性匹配。相比之下，跨语言密集检索不仅需要学习相关性匹配任务，还需要学习如何对齐语义相近但语言不同的文本嵌入 [24, 41]。正是这种语言差异，使跨语言检索对 xDR 而言成为一项相对更具挑战性的任务。

基于这一假设，本文提出使用跨语言查询生成（xQG）来弥合xDR之间的语言鸿沟。我们的方法如图1所示。具体而言，我们使用特定语言的提示对多语言T5（mT5）模型进行微调，为每种目标语言的每个段落生成多个查询。在索引阶段，我们使用给定的xDR模型，将不同语言的段落及其相关生成查询编码为嵌入。最后，在将其添加到ANN索引之前，我们使用生成查询的嵌入增强原始段落嵌入。通过这种方式，我们将段落嵌入移动到更接近目标用户查询语言的空间中。我们的方法不会增加额外的查询延迟，并且可以应用于任何现有的xDR。

2 相关工作

跨语言稠密检索器。 跨语言预训练语言模型的发展为此做出了重大贡献。

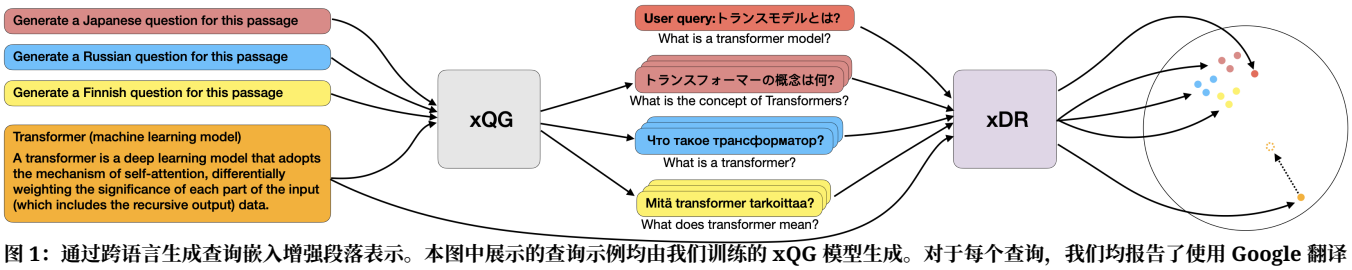

图 1：通过跨语言生成查询嵌入增强段落表示。本图中展示的查询示例均由我们训练的 xQG 模型生成。对于每个查询，我们均报告了使用 Google 翻译服务获得的相应译文。

跨语言密集检索（xDR） [4, 5, 8, 39]的发展。近期 xDR 方法的著名例子包括 CORA [2],，它采用生成器来促进检索训练数据挖掘；Sentri [31],，它提出了单编码器和自训练；以及 DR.DECR [19],，它利用平行查询和句子进行跨语言知识蒸馏。其中，我们的工作与 QuiCK [27], 最为相关，后者同样利用跨语言查询生成器进行 xDR。然而，与 QuiCK 在训练阶段将 xQG 用作 xDR 知识蒸馏中的教师模型不同，我们的方法直接利用 xQG 查询增强段落嵌入，无需进行任何 xDR 训练。

用于信息检索的查询生成。查询生成是一种成熟的技术，已广泛应用于各种检索模型，以提升检索性能 [11, 25]。此外，研究表明，该技术在稠密段落检索任务的领域自适应中具有良好效果 [23, 32, 34],，也能提升其他基于预训练语言模型的排序器的有效性 [37, 48]。在我们的方法中，我们同样依赖查询生成模型生成高质量查询，以用于下游段落嵌入增强任务。

增强稠密检索器的嵌入。 我们的方法依赖于有效增强稠密检索编码器编码的表示，这一方向近期也受到其他研究的关注。例如，Li等人 [16, 17] 将排名靠前的检索段落用作伪相关反馈，并使用Rocchio聚合函数增强查询嵌入。类似地，Zhuang等人 [47] 通过使用点击段落的嵌入来增强查询嵌入，拓展了这一思路。其他伪相关反馈方法也已得到广泛研究，以提升仅英语稠密检索 [18, 35, 36, 42, 46] 和跨语言密集检索 [3]。另一方面，HyDE方法利用大型预训练语言模型为给定查询生成假设段落，并直接使用该假设段落的嵌入执行搜索 [9]。尽管上述所有工作都聚焦于在查询时增强稠密检索器的查询嵌入，我们的工作则聚焦于在索引时增强段落嵌入，从而避免增加查询延迟开销。

3 方法

我们的方法由两个部分组成：（1）一个跨语言查询生成模型，可针对每个段落生成不同语言的高质量查询；（2）一个嵌入聚合函数，用生成查询的嵌入增强原始段落嵌入。

为获得 xQG 模型，我们使用带有标注相关性（q_t、p）的样本对对 mT5 模型进行微调，其中 t 是目标查询语言。我们采用类似于 docTquery-T5 目标函数的 seq2seq 目标函数 [25]；在我们的设置中，输入是带有特定语言提示的段落文本：

$$prompt\,(t,p) = Generate\ a\ [t]\ question\ for\ this\ passage:\ [p], \quad (1)$$

其中，$[t]$ 和 $[p]$ 分别是目标语言和段落文本的提示占位符。训练好 xQG 模型后，我们可以通过相应地替换占位符，为每种目标语言和每个段落生成一组查询：

$$Q_p = \bigcup_{t\in T} Q_p^t;\ Q_p^t = xQG\,(prompt\,(t,p)) \times n, \quad (2)$$

其中，T是目标语言集合，Q_p是段落 p 对应的所有生成查询的集合，其中每种目标语言各包含 n 个生成查询。在实验中，我们采用 top-k 采样方案 [7] 来生成查询，并设置 $k = 10$。我们发现，这些简单的特定语言提示能够有效引导 T5 模型以相应语言生成所需输出。

对于嵌入聚合函数，我们采用一种类似 Rocchio 的聚合函数，这与先前聚合稠密表示的工作 [16, 47]类似。我们使用此聚合函数来获取语料库中所有段落的嵌入：

$$emb(p, Q_p, \theta, \alpha) = (1-\alpha)\cdot\theta(p) + \alpha \sum_{\hat{q}_t \in Q_p} \theta(\hat{q}_t), \quad (3)$$

其中，$\hat{q}_t$是为目标语言 t生成的查询，θ是跨语言密集检索编码器模型，θ（。）是跨语言密集检索编码器生成的文本嵌入。超参数 $\in [0,1]$ 是增强比例，。该比例用于控制分配给原始段落嵌入和生成的查询嵌入的权重。

4 实验设置

We 设计实验以回答以下研究问题：

研究问题1： 生成的查询数量如何影响跨语言密集检索的效果?
研究问题2： 增强比例 α 如何影响跨语言密集检索的效果?
研究问题3： 为每种语言生成的查询如何影响跨语言密集检索的效果?

数据集与评估。 我们在 XOR-TyDi [1], 上训练并评估我们的方法。这是一个跨语言开放式检索问答基准数据集。该数据集的训练集中约有 15k 个经标注的相关段落-查询对，开发集中有 2k 个段落-答案对。训练集和开发集中的查询均使用七种类型各异的语言（阿拉伯语、孟加拉语、芬兰语、日语、韩语、俄语和泰卢固语），而段落则使用英语。语料库中约有 18M 个段落。我们使用训练集来训练 xQG 模型和

表 1：使用不同骨干预训练语言模型训练的跨语言密集检索模型。零样本表示仅使用 NQ 训练数据的英语子集进行训练（因此，在跨语言任务中为零样本）。与基础模型（不含 xQG）的统计差异标记为 ⋆ ($p < 0.05$)。(a) R@2kt

模型	Ar	Bn	Fi	Ja	Ko	Ru	Te	平均值
XLM-R	28.5	26.3	29.3	22.4	34.0	17.3	34.9	27.5
XLM-R + xQG	**30.1**	**29.6**	**31.2**	**23.7**	**36.1**	**19.4**	**38.2**	**29.8**⋆
mBERT	41.1	49.0	52.2	**37.3**⋆	48.1	33.3	47.9	44.1
mBERT + xQG	**42.4**	**54.9**⋆	**54.1**	33.6	**52.3**⋆	**33.8**	**52.5**⋆	**46.2**⋆
mBERT（零样本）	32.6	25.0	38.2	29.5	38.9	27.0	39.9	33.0
mBERT（零样本）+XQG	**33.9**	**28.9**⋆	**43.0**⋆	**32.8**⋆	**41.4**	**29.5**	**42.0**	**36.0**⋆

(b) R@5kt

模型	Ar	Bn	Fi	Ja	Ko	Ru	Te	平均值
XLM-R	38.5	33.6	37.9	**32.8**	42.8	28.3	47.5	37.3
XLM-R + xQG	**38.8**	**41.1**⋆	**39.8**	**32.8**	**43.2**	**30.8**	**49.6**	**39.4**⋆
mBERT	49.2	57.6	**58.6**	42.7	57.5	**41.4**	55.9	51.8
mBERT + xQG	**51.5**	**60.9**⋆	58.3	**43.6**	**58.6**	**41.4**	**60.1**⋆	**53.5**⋆
mBERT（零样本）	38.5	36.5	47.5	38.2	48.1	35.0	48.7	41.8
mBERT（零样本）+xQG	**43.7**⋆	**40.8**⋆	**50.0**⋆	**40.2**	**49.8**	**39.7**⋆	**51.7**	**45.1**⋆

xDR 模型。我们使用每 m 千个词元的召回率（R@$m$kt）评估我们的方法及基线方法的效果，这是该数据集的官方评估指标。该指标计算最小答案出现在检索段落前 m = 2k 5k 2 5 个词元中的查询所占比例。按照该数据集的常见做法，我们考察，（R@ kt 和 R@ kt）。统计显著性差异通过采用邦弗朗尼校正的双尾 t 检验进行报告。

基线。遵循 XOR-TyDi [1], 上的常见做法，我们通过使用不同的多语言预训练语言模型（PLM）初始化 DPR [14]，将其调整到跨语言设置。具体而言，在我们的实验中，我们使用 BERT [6] 的多语言变体（mBERT）和 XLM-RoBERTa [5]（XLM-R）来实现这一目的。除了标准的监督基线外，我们还探索了 xQG 嵌入增强方法如何改进零样本 xDR 模型：该 xDR 使用 mBERT 初始化，但仅使用英语段落-查询对进行训练，并直接应用于 XOR-TyDi 跨语言检索任务。

实现细节。我们使用 Google[1] 提供且可在 Huggingface 库 [38]中获取的多语言 T5-base 检查点初始化 xQG 模型。我们使用 XOR-TyDi 训练集中的段落 - 查询对对该检查点进行微调。我们使用批量大小为 128、学习率为 1e-4，对 xQG 模型训练 1600 个步骤。训练完成后，我们针对每种目标语言为每个段落生成 5 个查询，总共产生约 7 * 5 * 18M = 630M 个查询。我们使用四个 A100 GPU 生成所有查询；这一过程耗时约 70 小时。我们在 Huggingface Hub [2] 上发布生成的查询，以便社区重复利用这一资源 [30]。在训练 xDR 时，我们使用 Tevatron 密集检索器训练工具包 [10],，该工具包使用 BM25 困难负例段落。我们使用批量大小为 128 训练 xDR，使用 mBERT base[3] 或 XLM-R base[4] 检查点进行初始化，并在 XOR-TyDi 训练集上训练 40 个训练轮次。对于每个训练样本，我们将对比损失中的困难负例段落数量设为 7，并采用批内负例训练。我们使用 1e-5 的学习率

[1]https://huggingface.co/google/mt5-base [2] https://huggingface.co/datasets/ielab/xor-tydi-xqg-augmented [3] https://huggingface.co/bert-base-multilingual-cased [4] https://huggingface.co/xlm-roberta-base

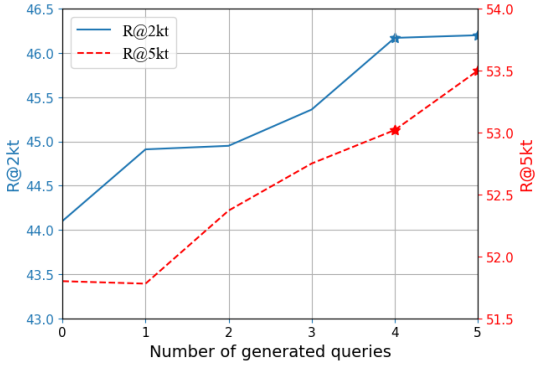

图 2：为目标语言生成的查询数量的影响。分数为所有语言的平均值。与不进行增强（查询数量为 $n = 0$）相比，具有统计显著性的提升以星号（$p < 0.05$）标注。

对于基于 mBERT 的跨语言密集检索器，以及对于基于 XLM-R 的跨语言密集检索器，取值为 2e-5。对于零样本跨语言密集检索器，我们采用与在 XOR-TyDi 上训练的基于 mBERT 的跨语言密集检索器相同的训练配置，但使用自然问题（NQ）训练数据 [15]，其中包含仅使用英语查询和段落的训练样本。

5 结果

5.1 主要结果

表 1 展示了以 XLM-R 和 mBERT 骨干预训练语言模型初始化、并在 XOR-TyDi 数据集上训练的 xDR 模型的效果。零样本表示仅在 NQ 数据集的英语子集上训练的模型。在这些实验中，我们使用 xQG 生成的所有查询，并将增强比例设为 $\alpha = 0.01$，以增强 xDR 的段落嵌入。

对于 R@2kt 指标，使用 mBERT 初始化的跨语言密集检索优于使用 XLM-R 初始化的跨语言密集检索，平均 R@2kt 得分达到 44.1，而 XLM-R 的平均得分为 27.5。我们的 xQG 段落嵌入增强方法提升了基于 XLM-R 的跨语言密集检索，平均得分达到 29.8，相较于其基线效果具有统计显著性改进 ($p < 0.05$)。同样，mBERT 的效果通过 xQG 得到提升，平均得分达到 46.2，这同样是统计上显著的

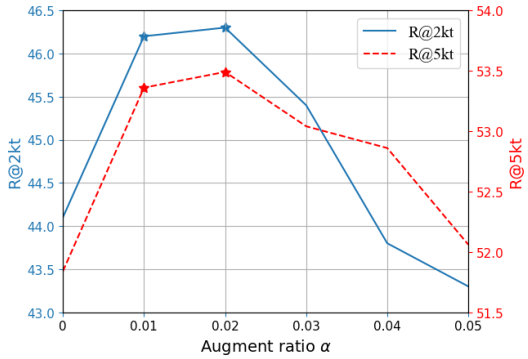

图 3：增强比例的影响 α。分数为所有语言的平均值。与不进行增强的情况（$\alpha = 0$）相比，具有统计显著性的提升以星号（$p < 0.05$）标示。

与其对应的基线相比有显著提升（$p < 0.05$）。零样本 mBERT 模型的平均值 R@2kt 为 33.0；与 xQG 结合后，该指标也有所提升，平均得分达到 36.0。这一提升具有统计显著性（$p < 0.05$）。R@5kt 也呈现出类似趋势。总体而言，我们发现 xQG 能够显著提升所有经研究的跨语言密集检索模型。就各语言的效果而言，除 mBERT 在日语（Ja）上的 R@2kt 以及 mBERT 在芬兰语（Fi）上的 R@5kt 外，xQG 几乎提升了所有语言上的所有模型。

总之，在 R@2kt 和 R@5kt 指标上，mBERT 的表现均优于 XLM-R。使用 xQG 增强段落嵌入，在统计上显著提升了这两种骨干网络的效果。

5.2 RQ1：生成查询数量的影响

图 2 展示了在使用 mBERT 跨语言密集检索时，采用不同数量的生成查询来增强段落嵌入所产生的影响。结果表明，增加生成查询的数量有利于提升 R@2tk 和 R@5tk。每种目标语言使用 4 个或更多生成查询时，提升在统计上显著。尽管曲线尚未趋于平稳，表明使用更多生成查询可能会进一步提升效果，但受计算资源限制，我们的实验对每种目标语言最多使用了 5 个生成查询。

5.3 研究问题2：增强比例的影响

我们在图 3 中报告了增强比例 α 对跨语言密集检索（xDR）效果的影响。较高的 α 值意味着在嵌入聚合过程中，为生成的查询嵌入分配更高的权重，而 $\alpha = 0$ 则表示不进行增强。如图所示，即使较小的 α 值（0.01）也能显著提升 R@2kt 和 R@5kt 两项指标。当 $\alpha = 0.02$ 时，检索效果达到最佳。然而，使用更高的 α 值并不会带来进一步的性能提升；相反，当 $\alpha > 0.05$ 时，甚至可能降低检索效果。基于这些结果，我们得出结论：嵌入聚合函数中的增强比例会对跨语言密集检索（xDR）的效果产生显著影响，而采用较小的 α 值有助于提升检索效果。

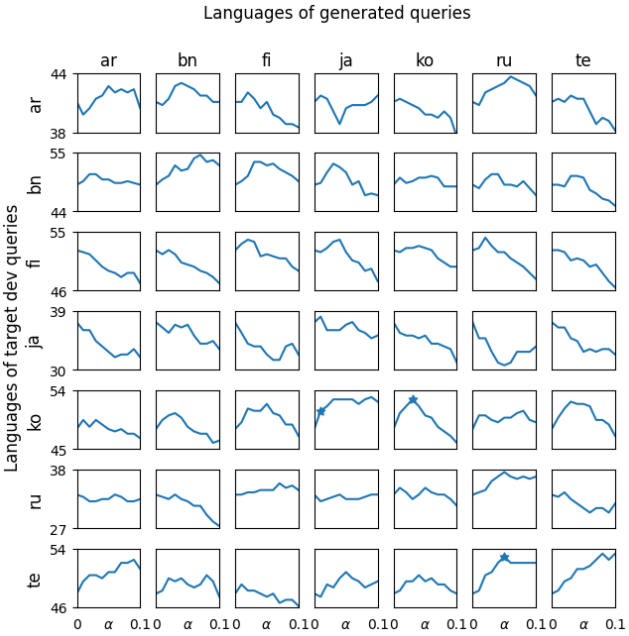

图4：由源语言生成的查询（列）增强的目标语言（行）的 R@2kt。我们绘制 α，范围为 0 到 0.1，步长为 0.01（x 轴）。相对于不进行增强的结果（$\alpha = 0$），具有统计显著性的更优结果以星号（$p < 0.05$）标记。

5.4 研究问题3：各语言的影响

在之前的实验中，我们使用为一个段落生成的所有查询来增强原始段落嵌入，而不考虑生成查询所用的语言。接下来，我们分别研究使用各语言生成的查询对效果的影响。我们在图4中对此进行分析，其中按不同的 α 值（横轴）绘制各目标语言（行）的 R@2kt。

对角线上的图表显示，当使用为同一语言生成的查询增强段落嵌入时，跨语言密集检索的效果会提升。值得注意的是，我们观察到，不同语言的 α最佳值各不相同，而且效果提升并不总是具有统计显著性。我们还观察到，对于某些语言，为其他语言生成的查询可以提升另一种语言中目标查询的效果。例如，使用为日语（Ja）生成的查询可以提升韩语（Ko）中目标查询的效果，而使用俄语（Ru）生成的查询则有助于提升泰卢固语（Te）中的目标查询效果。这些结果表明，为任一单一语言生成的查询嵌入也可能为其他语言的目标查询提供有用信息。

6 结论与未来工作

本文提出一种段落嵌入增强方法，以提升跨语言密集检索器的效果。该方法可应用于任何跨语言密集检索器，无需进一步训练检索器，也无需更改检索流程。我们通过实验证明，该方法对采用不同骨干网络的跨语言 DPR 方法均有效。不过，我们的实证研究有一项局限：未在其他密集检索器架构上评估该方法。我们将这项研究的扩展留待

未来工作。我们还注意到，我们的方法依赖于能够生成高质量查询的 xQG 模型。然而，近期一项研究表明，基于 T5 的查询生成容易产生幻觉；生成器除了生成效果很好的查询，也会生成质量较差的查询，从而降低检索效果 [11]。该研究随后提出使用交叉编码器排序器过滤掉部分无效生成查询；这一做法可进一步提升效果。我们将该方法适配到 xQG 场景留待未来工作。

参考文献

[1] Akari Asai、Jungo Kasai、Jonathan H. Clark、Kenton Lee、Eunsol Choi 和 Hannaneh Hajishirzi。2021。XOR QA：跨语言开放检索问答。载于 北美计算语言学协会分会人类语言技术会议。[2] Akari Asai、Xinyan Yu、Jungo Kasai 和 Hanna Hajishirzi。2021。一个适用于多种语言的问答模型：采用跨语言稠密段落检索。神经信息处理系统进展 34（2021），7547–7560。[3] Ramraj Chandradevan、Eugene Yang、Mahsa Yarmohammadi 和 Eugene Agichtein。2022。通过伪相关反馈学习丰富跨语言检索的查询表示。载于 第*45*届国际 *ACM SIGIR* 信息检索研究与发展会议论文集。1790–1795。[4] Zewen Chi、Li Dong、Furu Wei、Nan Yang、Saksham Singhal、Wenhui Wang、Xia Song、Xian-Ling Mao、He-Yan Huang 和 Ming Zhou。2021。InfoXLM：一种跨语言语言模型预训练的信息论框架。载于 *2021*年北美计算语言学协会分会人类语言技术会议论文集。3576–3588。[5] Alexis Conneau、Kartikay Khandelwal、Naman Goyal、Vishrav Chaudhary、Guillaume Wenzek、Francisco Guzmán、Édouard Grave、Myle Ott、Luke Zettlemoyer 和 Veselin Stoyanov。2020。大规模无监督跨语言表示学习。载于 第*58*届计算语言学协会年会论文集。8440–8451。[6] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019。BERT：用于语言理解的深度双向 Transformer 预训练。载于 *2019*年北美计算语言学协会分会人类语言技术会议论文集（*NAACL-HLT 2019*），美国明尼苏达州明尼阿波利斯，*2019*年*6*月*2*日至*7*日，第*1*卷（长篇与短篇论文），Jill Burstein、Christy Doran 和 Thamar Solorio（编）。计算语言学协会，4171 4186 10 18653 1 19 1423 –。[7] 2018 https://doi.org/。/v/n - Angela Fan、Mike Lewis 和 Yann Dauphin。。分层神经故事生成。载于 第*56*届计算语言学协会年会论文集（第*1*卷：长篇论文）。889–898。[8] Fangxiaoyu Feng、Yinfei Yang、Daniel Cer、Naveen Arivazhagan 和 Wei Wang。2022。与语言无关的 BERT 句子嵌入。载于 第*60*届计算语言学协会年会论文集（第*1*卷：长篇论文）。878–891。[9] Luyu Gao、Xueguang Ma、Jimmy Lin 和 Jamie Callan。2022。无需相关性标签的精准零样本稠密检索。*arXiv* 预印本 *arXiv:2212.10496*（2022）。[10] Luyu Gao、Xueguang Ma、Jimmy J. Lin 和 Jamie Callan。2022。Tevatron：一种高效灵活的稠密检索工具包。*ArXiv* abs/2203.05765（2022）。[11] Mitko Gospodinov、Sean MacAvaney 和 Craig Macdonald。2023。Doc2Query：少即是多。*arXiv* 预印本 *arXiv:2301.03266*（2023）。[12] Sebastian Hofstätter、Sophia Althammer、Michael Schröder、Mete Sertkan 和 Allan Hanbury。2020。通过跨架构知识蒸馏改进高效神经排序模型。*arXiv* 预印本 *arXiv:2010.02666*（2020）。[13] Sebastian Hofstätter、Sheng-Chieh Lin、Jheng-Hong Yang、Jimmy Lin 和 Allan Hanbury。2021。通过均衡的主题感知采样，高效地训练有效的稠密检索器。载于 第*44*届国际 *ACM SIGIR* 信息检索研究与发展会议论文集。113–122。[14] Vladimir Karpukhin、Barlas Oguz、Sewon Min、Patrick Lewis、Ledell Wu、Sergey Edunov、Danqi Chen 和 Wen-tau Yih。2020。用于开放域问答的稠密段落检索。载于 *2020*年自然语言处理经验方法会议论文集（*EMNLP*）。计算语言学协会，线上，6769–6781。https://doi.org/10.18653/v1/2020.emnlp-main.550[15] Tom Kwiatkowski、Jennimaria Palomaki、Olivia Redfield、Michael Collins、Ankur Parikh、Chris Alberti、Danielle Epstein、Illia Polosukhin、Matthew Kelcey、Jacob Devlin、Kenton Lee、Kristina N. Toutanova、Llion Jones、Ming-Wei Chang、Andrew Dai、Jakob Uszkoreit、Quoc Le 和 Slav Petrov。2019。自然问题：问答研究基准。计算语言学协会汇刊 （2019）。[16] Hang Li、Ahmed Mourad、Shengyao Zhuang、Bevan Koopman 和 Guido Zuccon。2023。利用深度语言模型和稠密检索器进行伪相关反馈：成效与陷阱。*ACM* 信息系统汇刊 （2023）。[17] Hang Li、Shengyao Zhuang、Xueguang Ma、Jimmy Lin 和 Guido Zuccon。2023。使用 Pyserini 中的稠密检索器进行伪相关反馈。载于 第 *26* 届澳大拉西亚文档计算研讨会（*ADCS '22*）论文集。[18] Hang Li、Shengyao Zhuang、Ahmed Mourad、Xueguang Ma、Jimmy Lin 和 Guido Zuccon。2022。使用伪相关反馈改进稠密检索的查询表示：一项可复现性研究。载于 信息检索进展：第 *44* 届欧洲信息检索研究会议（*ECIR 2022*），挪威斯塔万格，*2022* 年 *4* 月 *10—14* 日，论文集，第一部分。Springer，599–612。[19] Yulong Li、Martin Franz、Md Arafat Sultan、Bhavani Iyer、Young-Suk Lee 和 Avirup Sil。2022。从英语检索器学习跨语言信息检索。载于*2022* 年北美计算语言学协会分会会议：人类语言技术论文集。4428–4436。[20] Sheng-Chieh Lin、Jheng-Hong Yang 和 Jimmy Lin。2020。使用紧耦合教师蒸馏用于排序的稠密表示。*arXiv* 预印本 *arXiv:2010.11386*（2020）。[21] Sheng-Chieh Lin、Jheng-Hong Yang 和 Jimmy Lin。2021。使用紧耦合教师进行稠密检索知识蒸馏的批内负样本。载于 第 *6* 届 *NLP* 表示学习研讨会（*RepL4NLP- 2021*）论文集。163–173。[22] Shayne Longpre、Yi Lu 和 Joachim Daiber。2021。MKQA：用于多语言开放域问答的语言多样化基准。计算语言学协会会刊 9（2021），1389–1406。[23] Ji Ma、Ivan Korotkov、Yinfei Yang、Keith Hall 和 Ryan McDonald。2021。通过面向领域的合成问题生成实现零样本神经段落检索。载于 第 *16* 届欧洲计算语言学协会分会会议：主会场论文集。1075–1088。[24] Suraj Nair、Eugene Yang、Dawn Lawrie、Kevin Duh、Paul McNamee、Kenton Murray、James Mayfield 和 Douglas W Oard。2022。构建跨语言稠密检索模型的迁移学习方法。载于 信息检索进展：第 *44* 届欧洲信息检索研究会议（*ECIR 2022*），挪威斯塔万格，*2022* 年 *4* 月 *10—14* 日，论文集，第一部分。Springer，382–396。[25] Rodrigo Nogueira 和 Jimmy Lin。2019。从 doc2query 到 docTTTTTquery。[26] Yingqi Qu、Yuchen Ding、Jing Liu、Kai Liu、Ruiyang Ren、Wayne Xin Zhao、Daxiang Dong、Hua Wu 和 Haifeng Wang。2021。RocketQA：面向开放域问答的稠密段落检索优化训练方法。载于 *2021* 年北美计算语言学协会分会会议：人类语言技术论文集。5835– 5847。[27] Houxing Ren、Linjun Shou、Ning Wu、Ming Gong 和 Daxin Jiang。2022。使用查询生成器增强双编码器的跨语言稠密检索。载于 *2022* 年自然语言处理经验方法会议论文集。3107–3121。[28] Ruiyang Ren、Shangwen Lv、Yingqi Qu、Jing Liu、Wayne Xin Zhao、Qiaoqiao She、Hua Wu、Haifeng Wang 和 Ji-Rong Wen。2021。PAIR：利用以段落为中心的相似关系改进稠密段落检索。载于 计算语言学协会研究成果：*ACL-IJCNLP 2021*。2173–2183。[29] Ruiyang Ren、Yingqi Qu、Jing Liu、Wayne Xin Zhao、Qiaoqiao She、Hua Wu、Haifeng Wang 和 Ji-Rong Wen。2021。RocketQAv2：稠密段落检索与段落重排序的联合训练方法。载于 *2021* 年自然语言处理经验方法会议论文集。2825–2835。[30] Harrisen Scells、Shengyao Zhuang 和 Guido Zuccon。2022。减少、再利用、回收：绿色信息检索研究。载于 第 *45* 届国际 *ACM SIGIR* 信息检索研究与发展会议论文集。2825–2837。[31] Nikita Sorokin、Dmitry Abulkhanov、Irina Piontkovskaya 和 Valentin Malykh。2022。用你的母语问我任何问题。载于 *2022* 年北美计算语言学协会分会会议：人类语言技术论文集。395–406。[32] Nandan Thakur、Nils Reimers、Andreas Rücklé、Abhishek Srivastava 和 Iryna Gurevych。2021。BEIR：用于信息检索模型零样本评估的异构基准。载于 第三十五届神经信息处理系统会议数据集与基准赛道（第二轮）。https://openreview. net/forum?id=wCu6T5xFjeJ [33] Nicola Tonellotto。2022。神经信息检索讲义。*arXiv* 预印本 *arXiv:2207.13443*（2022）。[34] Kexin Wang、Nandan Thakur、Nils Reimers 和 Iryna Gurevych。2022。GPL：用于稠密检索无监督领域自适应的生成式伪标签。载于 *2022* 年北美计算语言学协会分会会议：人类语言技术论文集。2345–2360。[35] Xiao Wang、Craig Macdonald、Nicola Tonellotto 和 Iadh Ounis。2021。多表示稠密检索的伪相关反馈。载于 *2021* 年 *ACM SIGIR* 信息检索理论国际会议论文集。297–306。[36] Xiao Wang、Craig Macdonald、Nicola Tonellotto 和 Iadh Ounis。2023。ColBERT- PRF：用于稠密段落和文档检索的语义伪相关反馈。*ACM Web Transactions* 17，1（2023），1–39。

[37] Yujing Wang、Yingyan Hou、Haonan Wang、Ziming Miao、Shibin Wu、Qi Chen、Yuqing Xia、Chengmin Chi、Guoshuai Zhao、Zheng Liu 等。[无日期。]一种用于文档检索的神经语料索引器。载于神经信息处理系统进展。[38] Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Rémi Louf、Morgan Funtowicz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander M. Rush。2020。Transformer：最先进的自然语言处理技术。载于*2020 年自然语言处理经验方法会议论文集：系统演示*。计算语言学协会，38–45 页。[39] Ning Wu、Yaobo Liang、Houxing Ren、Linjun Shou、Nan Duan、Ming Gong 和 Daxin Jiang。2022。用于多语言稠密检索的无监督上下文感知句子表示预训练。*arXiv* 预印本 *arXiv:2206.03281*（2022 年）。[40] Lee Xiong、Chenyan Xiong、Ye Li、Kwok-Fung Tang、Jialin Liu、Paul N Bennett、Junaid Ahmed 和 Arnold Overwijk。2020。用于稠密文本检索的近似最近邻负对比学习。载于国际学习表征会议。[41] Eugene Yang、Suraj Nair、Ramraj Chandradevan、Rebecca Iglesias-Flores 和 Douglas W Oard。2022。C3：通过对比式弱监督进行持续预训练，以实现跨语言即席检索。载于第 *45* 届国际 *ACM SIGIR* 信息检索研究与发展会议论文集。2507–2512 页。

[42] HongChien Yu、Chenyan Xiong 和 Jamie Callan。2021。利用伪相关反馈改进稠密检索的查询表示。在 第*30*届 *ACM* 信息与知识管理国际会议论文集。3592–3596。[43] Jingtao Zhan、Jiaxin Mao、Yiqun Liu、Jiafeng Guo、Min Zhang 和 Shaoping Ma。2021。利用困难负例优化稠密检索模型训练。第*44*届国际 *ACM SIGIR* 信息检索研究与发展会议论文集（2021）。[44] Jingtao Zhan、Jiaxin Mao、Yiqun Liu、Min Zhang 和 Shaoping Ma。2020。RepBERT：用于第一阶段检索的上下文化文本嵌入。*arXiv* 预印本 *arXiv:2006.15498*（2020）。[45] Wayne Xin Zhao、Jing Liu、Ruiyang Ren 和 Ji-Rong Wen。2022。基于预训练语言模型的稠密文本检索：综述。*arXiv* 预印本 *arXiv:2211.14876*（2022）。[46] Yunchang Zhu、Liang Pang、Yanyan Lan、Huawei Shen 和 Xueqi Cheng。2022。LoL：针对伪相关反馈的查询改写损失设计的比较性正则化损失。在 第*45*届国际 *ACM SIGIR* 信息检索研究与发展会议论文集。825–836。[47] Shengyao Zhuang、Hang Li 和 Guido Zuccon。2022。稠密段落检索中的隐式反馈：一种反事实方法。在 第*45*届国际 *ACM SIGIR* 信息检索研究与发展会议论文集。18–28。[48] Shengyao Zhuang、Houxing Ren、Linjun Shou、Jian Pei、Ming Gong、Guido Zuccon 和 Daxin Jiang。2022。利用查询生成弥合可微分搜索索引中索引与检索之间的差距。*arXiv* 预印本 *arXiv:2206.10128*（2022）。