Multi-CPR：用于段落检索的多领域中文数据集

原作者：Dingkun Long et al. 复现人：徐可

摘要

段落检索是信息检索 (IR) 研究中的一项基本任务。在英语领域，大规模注释数据集（例如 MS MARCO）的可用性和深度预训练语言模型（例如 BERT）的出现给段落检索系统带来巨大提升。然而，在中文领域，尤其是一些特定领域，由于缺乏大规模、高质量的标注数据，中文段落检索系统一直不够成熟。因此，本文提出了一种用于段落检索的新型多领域中文数据集。它涵盖了三个方面：电子商务、娱乐视频和医疗。每个领域的数据集包含数百万个段落和一定数量的人工标记的问题-段落对。同时进行了各种具有代表性方法的实验。实验结果证明，在一般领域的数据集上训练的检索模型的性能将不可避免地在特定领域下降，且在特定领域的数据集上训练的检索模型取得了较好的结果。这证明了构造特定领域的数据集的必要性。

关键词：段落检索；预训练模型

1 引言

大规模文档检索是信息检索中非常重要的一个问题。它是诸如开放域问答[1]、机器阅读理解[2]等下游任务的前提。近来，深度学习的高速发展使得其远远超越了传统模型的水平。然而，深度神经网络模型往往包含了大量的参数，优化参数需要大量的训练数据。因此，我们需要一个公开且高质量的文档检索数据集。

英文领域已经存在一大批规模大、质量高的数据集，如 SQuAD[3]、MS MARCO[4]等。在中文领域，尽管已经有了一些文档检索相关的数据集，如 Sogou-QCL[5]、Dureader[6]等。但是它们主要是通用领域的数据，能够用于某一单独领域的数据依然很少。

为了推进中文文档检索的发展，本文提出了 Multi-CPR 数据集。其包含三个领域的数据：电子商务、娱乐视频和医疗。每个领域都包含了百万个文档和足够数量的问题-文档对。

最后，本文实现了一系列具有代表性的方法。对于稀疏模型，有 BM25[7]和 doc2query[8]；对于稠密模型，有三组不同的和 DPR[1]相关的实验。实验结果表明，在特定领域数据集上训练的稠密模型的提升非常大，其表现大幅度超越了传统的稀疏模型。

2 相关工作

2.1 段落检索

段落检索的目的是在给定信息搜索查询的情况下从大型语料库中找出所有可能相关的段落[9]。传统的稀疏模型通常依赖于基于单词的检索方法，如 BM25[7]。近年来，随着文本表示学习研究[10]和深度预训练语言模型[11]的快速发展，稠密检索与预训练语言模型相结合已成为提高系统检索表现的流行方法。一般情况下，稠密模型在检索效果和对下游任务性能提升等方面明显优于传统的基于单词的检索模型[12]。

2.2 相关数据集

大规模高质量标注数据的出现，极大地促进了段落检索模型的发展。在这些数据集中，MS MARCO[4]是英语领域最具代表性的数据集。MS MARCO 是微软推出的文章排名数据集。文章排名任务的重点是从 880 万篇文章中对文章进行排名，这些文章是从必应搜索引擎的查询中收集的。约有 80.8 万个查询与相关段落对，以供训练使用。每个查询都与一个 (或几个) 标记为相关的段落相关联，没有明确表示为不相关的段落。

在中文领域，已有一些基于搜索引擎构建的数据集，如 Sougou-QCL[5]。Sougou-QCL 数据集包括 537366 个查询，900 多万个中文网页，以及点击模型评估的五种相关性标签。然而，该数据集集中在通用领域，并且标签是根据点击行为获得的，而不是人工的。Dureader[6]是最近发布的大型 MRC 中文数据集。数据分布主要集中在通用领域。它可以转换成一个信息检索数据集。虽然有一些通用的领域数据集，但特定领域的中文文章检索数据集还是很少。

3 本文方法

3.1 本文方法概述

DPR[1]是文档检索领域常用的模型。其模型示意图如图 1 所示：

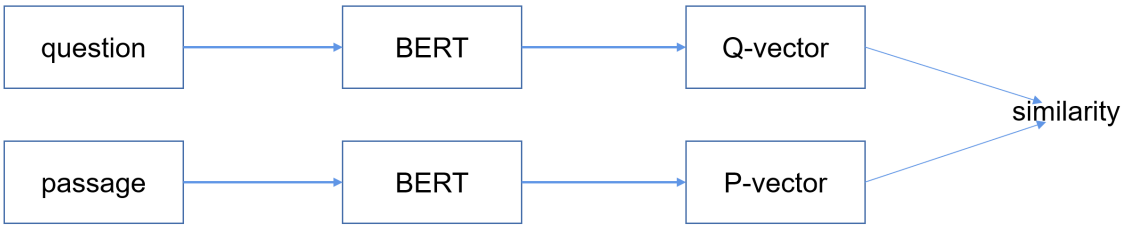

图 1: DPR 的双塔模型

DPR 整体来看分为两个部分：一个 question encoder，一个 passage encoder。两个 encoder 是独立不相关的，但是其预训练模型是一样的。question encoder 对用户输入的 question（也可称作 query）进行编码，passage encoder 对语料库的文档进行编码。分别得到 q-vector 和 p-vector 后，利用相似度函数，就可得出 question 和 passage 的相似性得分。

DPR 采用的相似度函数非常简单：

$$sim(q,p) = E_Q(q)^T E_P(p)\cdot$$

3.2 损失函数定义

$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -log\frac{e^{sim(q_i,p_i^+)}}{e^{sim(q_i,p_i^+)} + \sum_{j=1}^{n} e^{sim(q_i,p_{i,j}^-)}}$$

其中，p^+、p^- 分别指与 question 相关的 passage 和不相关的 passage

4 复现细节

4.1 与已有开源代码对比

我复现了原文的两组实验结果，分别是 DPR-1 和 DPR-2，工作主要集中在数据预处理部分。

对于 DPR-1，原文的描述是：在不同领域的数据上面进行训练的 DPR；DPR-2 与其不同的地方是：不停地在同一领域的数据上面进行训练的。DPR-2 的数据比较容易处理，因为原始数据集已经切分好

了不同领域的数据，只需要利用分词器对其进行编码即可。过程伪代码如下：

Procedure 1 分词器编码 query 和 passage

Input: q_p_relation *qrels,* corpus *c,* queries *q*
Output: tokenized query passange *span*
for *line* ***in*** *qrels* **do**
    $query$ = q(line(0))
    $passage$ = c(line(2))
    $encoded\_query$ = Tokenizer(query)
    $encoded\_passage$ = Tokenizer(passage)
    $span.add([encoded\_query, encoded\_passage])$
end
return span

DPR-1 的数据，作者没有直接给出，需要自己基于原始数据集构造。原始数据集由三个领域的数据构成。由于我们需要根据 query_id 和 passage_id 取得对应的 query 和 passage 内容，而每个领域的 id 取值是有重叠的，所以这里需要给它们加上前缀码。过程伪代码如下：

Procedure 2 整合三个领域的 corpus

Input: corpuses *c*
Output: merged_corpus *mc*
for *corpus* ***in*** *c* **do**
    merged_corpus[corpus.domain][pid] = prefix + corpus[pid]
end
return merged_corpus

Procedure 3 整合三个领域的 query

Input: queries *q*
Output: merged_queries *mq*
for *query* ***in*** *q* **do**
    merged_query[query.domain][qid] = prefix + query[qid]
end
return merged_query

Procedure 4 整合三个领域的 qprels

Input: qprels *qpr*
Output: merged_qprels *mqpr*
for *qprels* ***in*** *qpr* **do**
    merged_qprels[qprels.domain][qid] = prefix + qprels[qid]
    merged_qprels[qprels.domain][pid] = prefix + qprels[pid]
end
return merged_qprels

4.2 创新点

1. 在具体实验过程中，为了验证 in-batch negative 对 DPR 训练结果的影响，我把 batch size 调到了 16，原文的 batch size 是 32，结果差异较明显。

2.DPR-1 的训练数据作者并没有具体说明怎么处理，我通过整合数据集实现了 DPR-1 的效果

5 评估矩阵

文档检索的效果用 1.MRR@10 2.Recall@1k。它们的计算公式如下：

$$MRR@k = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{rank_i}$$

其中，$rank_i$ 是推荐的物品的排位

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

6 实验结果分析

将训练得到的模型，拿去做各个单一领域的文档检索任务。总共 2 个实验组，每组进行三次文档检索实验，因此总共 6 个结果。

我复现的 DPR-1 的结果如图 2

video		ecom		medical	
MRR@10	Recall@1000	MRR@10	Recall@1000	MRR@10	Recall@1000
0.2612	0.91	0.2733	0.905	0.3046	0.71

图 2: DPR-1 结果

我复现的 DPR-2 的结果如图 3

video		ecom		medical	
MRR@10	Recall@1000	MRR@10	Recall@1000	MRR@10	Recall@1000
0.2639	0.905	0.2572	0.906	0.3029	0.714

图 3: DPR-2 结果

原文的 DPR-1 的结果如图 4

video		ecom		medical	
MRR@10	Recall@1000	MRR@10	Recall@1000	MRR@10	Recall@1000
0.2537	0.9340	0.2704	0.9210	0.3270	0.7470

图 4: 原文 DPR-1 结果

原文的 DPR-2 的结果如图 5

video		ecom		medical	
MRR@10	Recall@1000	MRR@10	Recall@1000	MRR@10	Recall@1000
0.2627	0.9350	0.2894	0.9260	0.3388	0.7690

图 5: 原文 DPR-2 结果

原文的 DPR-2 的效果均超过了 DPR-1 的效果。这证明在单一领域数据上面进行训练能够提升模型的表现，尽管其数据量小了很多。

我复现的 DPR-2 在某一些方面能够比 DPR-1 的效果好，但是在另一些方面缺表现不好。这可能和 batch size 有关。DPR 的优化方法是 in-batch negative，如果 batch size 过小，那么就会导致学习的负样本不足；而 DPR-1 本身数据集够大，因此负样本数量足够，所以受 batch size 的影响相对 DPR-2 小一点。

7 总结与展望

本文先介绍了中文文档检索领域发展的一个突出的问题：大规模高质量的数据集不足，从而导致人们对模型的理解不够深入，影响了中文文档检索领域的发展。之后分别介绍了 Multi-CPR 数据集和

DPR 模型。接着重点说明了复现的一些细节，主要集中在数据处理部分。最后对比了复现的结果和原文的结果，发现对于 DPR，batch size 是一个比较重要的参数。

本次实验不足的地方主要集中在 batch size 参数实验部分。原文的 batch size 是单卡 32，如果调成 64，会不会提升模型的表现？由于硬件条件的限制，batch size 调成 32 就已经爆显存了（batch size 调成 16 大概占用了 10GB 的显存）。未来可以多进行一些不同 batch size 的实验。

最后，中文也仅仅是世界上诸多语言的一种。更有前途也更具有挑战性的方向是多语言文档检索模型的研究[13]。也就是说，模型认识更多语言的文字，能够进行不同单语言的检索任务。如法语、西班牙语、阿拉伯语等，而不仅仅是英文和中文。

参考文献

[1] KARPUKHIN V, OĞUZ B, MIN S, et al. Dense passage retrieval for open-domain question answering [J]. arXiv preprint arXiv:2004.04906, 2020.

[2] NISHIDA K, SAITO I, OTSUKA A, et al. Retrieve-and-read: Multi-task learning of information retrieval and reading comprehension[C]//Proceedings of the 27th ACM international conference on information and knowledge management. 2018: 647-656.

[3] RAJPURKAR P, ZHANG J, LOPYREV K, et al. Squad: 100,000+ questions for machine comprehension of text[J]. arXiv preprint arXiv:1606.05250, 2016.

[4] NGUYEN T, ROSENBERG M, SONG X, et al. MS MARCO: A human generated machine reading comprehension dataset[C]//CoCo@ NIPs. 2016.

[5] ZHENG Y, FAN Z, LIU Y, et al. Sogou-qcl: A new dataset with click relevance label[C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 1117-1120.

[6] HE W, LIU K, LIU J, et al. Dureader: a chinese machine reading comprehension dataset from real-world applications[J]. arXiv preprint arXiv:1711.05073, 2017.

[7] ROBERTSON S, ZARAGOZA H, et al. The probabilistic relevance framework: BM25 and beyond[J]. Foundations and Trends® in Information Retrieval, 2009, 3(4): 333-389.

[8] NOGUEIRA R, LIN J, EPISTEMIC A. From doc2query to docTTTTTquery[J]. Online preprint, 2019, 6.

[9] CAI D, YU S, WEN J R, et al. Block-based web search[C]//Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. 2004: 456-463.

[10] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: A review and new perspectives [J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1798-1828.

[11] HE P, LIU X, GAO J, et al. Deberta: Decoding-enhanced bert with disentangled attention[J]. arXiv preprint arXiv:2006.03654, 2020.

[12] GAO L, CALLAN J. Unsupervised corpus aware language model pre-training for dense passage retrieval [J]. arXiv preprint arXiv:2108.05540, 2021.

[13] ZHANG X, THAKUR N, OGUNDEPO O, et al. Making a MIRACL: Multilingual Information Retrieval Across a Continuum of Languages[J]. arXiv preprint arXiv:2210.09984, 2022.