**BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation**

BGE-M3:基于自知识蒸馏的多语言、多功能、多粒度文本嵌入

**摘要**

本文提出了一种新的嵌入模型，称为M3-嵌入，它以其在多语言、多功能和多粒度方面的表现而闻名。可以支持100多种工作语言，从而在多语言和跨语言检索任务上产生了新的最先进的性能。它可以同时执行嵌入模型的三种常见检索功能：密集检索、多向量检索和稀疏检索，为现实信息检索应用提供了统一的模型基础。它能够处理不同粒度的输入，从短句到多达8192个token的长文本。本文提出了一种新的自知识蒸馏方法，可以将来自不同检索功能的相关性分数作为教师信号进行集成，以提高训练质量。本文还优化了批处理策略，实现大批处理规模和高训练吞吐量，以确保嵌入的鉴别性。M3-嵌入是第一个实现如此强的通用性的嵌入模型。代码已开源<https://github.com/FlagOpen/FlagEmbedding>。

**引言**

嵌入模型是自然语言处理中DNN应用程序的一种关键形式。它们在潜在空间中对文本数据进行编码，其中数据的底层语义可以通过输出嵌入来表示。随着预训练语言模型的出现，文本嵌入的质量得到了极大的提高，使它们成为信息检索（IR）的必要组件。

基于嵌入的信息检索的一种常见形式是密集检索，可以基于嵌入相似度检索查询的相关答案。此外，嵌入模型还可以应用于其他信息检索任务，如多向量检索，基于多个嵌入的交互分数计算细粒度的关联。以及稀疏或词汇检索，通过输出嵌入估计每个项的重要性。

尽管文本嵌入的广泛流行，但现有的方法在多功能性方面仍然有局限性。首先，大多数嵌入模型都是为英语定制的。其次，现有的嵌入模型通常是针对一个单一的检索功能进行训练的。然而，典型的信息检索系统需要多种检索方法的复合工作流。第三，由于训练成本高昂，因此训练具有竞争力的长文档检索器只能支持短文本输入。

为了解决上述挑战，本文引入了M3嵌入。M3能够支持100多种世界语言，通过学习不同语言的公共语义空间，既可以实现每种语言内的多语言检索，也可以实现不同语言之间的跨语言检索。此外，它还能够生成多功能嵌入来支持不同的检索功能，不仅支持密集检索，还支持稀疏检索和多向量检索。最后，M3嵌入支持不同的输入粒度，从句子和段落等短文本输入，到多达8192个token的长文本输入。

这种多功能嵌入模型的有效训练是一个重大的挑战。

首先，提出了一种新的自知识蒸馏框架，其中多重检索功能可以联合学习和相互强化。在M3嵌入中，[CLS]嵌入用于密集检索，而来自其他标记的嵌入用于稀疏检索和多向量检索。基于集成学习原理，这种异质预测器可以作为一个更强的预测器组合。因此，将来自不同检索函数的相关性分数整合作为教师信号，通过知识蒸馏来增强学习过程。

其次，优化了批处理策略，以实现较大的批处理规模和较高的训练吞吐量，这有助于嵌入的鉴别性。

最后，进行了全面和高质量的数据管理。数据集由三个来源组成： 1)从大量的多语语言语料库中提取无监督数据，2)密切相关的监督数据的集成，3)稀缺的训练数据的合成。这三个数据源相互补充，并应用于训练过程的不同阶段。

本文工作贡献如下：

1. 提出了M3嵌入模型，这是第一个支持多语言、多功能和多粒度的模型。
2. 提出了一种新的自知识蒸馏和高效处理策略的训练框架。还对训练数据进行了高质量的管理。
3. 模型、代码和数据都是公开的，这为文本嵌入的直接使用和未来的开发提供了关键资源。

**相关工作**

本节从一般文本嵌入、神经检索嵌入模型、多语言嵌入三个方面进行了综述。

近年来，在文本嵌入领域取得了实质性的进展。一个主要的驱动力是预训练语言模型的流行，其中数据的潜在语义可以被强大的文本编码器有效地编码。此外，对比学习的进展是另一个关键因素，特别是负抽样的改进和知识蒸馏的开发。

嵌入模型的一个主要应用是神经检索。通过度量与文本嵌入的语义关系，可以根据嵌入的相似度检索到输入查询的相关答案。基于嵌入的检索方法最常见的形式是密集检索，其中，文本编码器的输出被聚合（例如，通过[CLS]或平均池化）来计算嵌入的相似度。另一种常见的替代方法是多向量检索，它对文本编码器的输出应用细粒度的交互来计算嵌入的相似度。最后，文本嵌入还可以转换为术语权重，这有利于稀疏或词汇检索。通常，上述检索方法都是通过不同的嵌入模型来实现的，没有一种现有的方法能够统一所有这些功能。

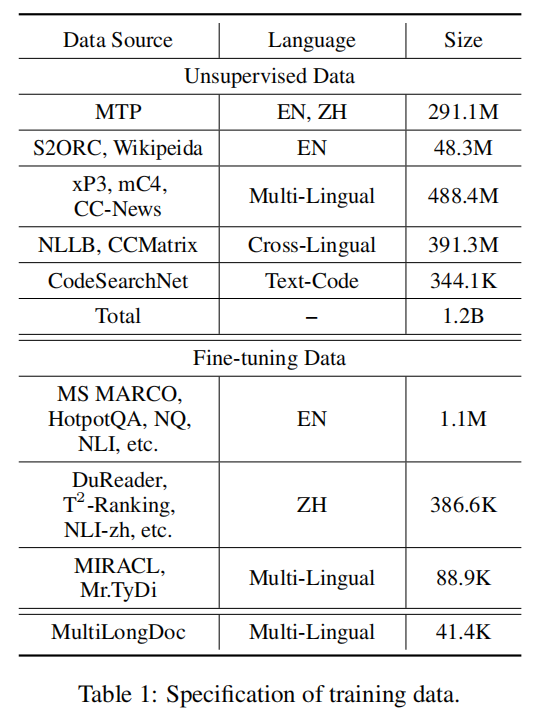
尽管有了巨大的技术进步，但大多数现有的文本嵌入只为英语开发。为了缓解这个问题，研究人员从多个方向做出了不断的努力。一个是预训练的多语言文本编码器的开发，如mBERT、mT5、XLM-R。另一个方法是管理多语言文本嵌入的训练和评估数据，例如，MIRACL、mMARCO、Mr.TyDi、MKQA。与此同时，多语言文本嵌入不断发展，如mDPR、mContriever、mE5等。然而，鉴于与英语模式之间的显著差距以及不同语言之间的巨大不平衡，目前的进展还远远不够。

**M3嵌入**

M3嵌入实现了三倍的多功能性。它支持多种语言，并处理不同粒度的输入数据。此外，它还统一了文本嵌入的通用检索功能**。**形式上，给定一个任意语言x中的查询q，它能够从语料库中检索语言为y的文档d。在这里，fn∗(·)属于任何一个函数：密集、稀疏/词汇或多向量检索；y可以是另一种语言，也可以是与x相同的语言。

**数据收集**

从三个来源进行了全面的数据收集：来自未标记语料库的无监督数据，来自已标记语料库的微调数据，以及通过合成进行的微调数据（如表1所示）。这三个数据源相互补充，它们应用于培训过程的不同阶段。无监督数据是通过提取丰富的语义结构，如标题体、标题抽象、指令输出等。为了学习跨语言语义匹配的统一嵌入空间，从两个翻译数据集NLLB和CCMatrix中引入了并行句子。对原始数据进行过滤，以去除潜在的不良内容和低相关性样本。总共提供了12亿对文本对，包括194种语言和2655对跨语言样本。



此外，从标记的语料库中收集了相对较小但多样化和高质量的微调数据。

最后，生成合成数据，以减轻长文本检索任务的缺点，并引入额外的多语言微调数据（记为MultiLongDoc）。具体来说，从数据集中抽取冗长的文本，并从中随机选择段落。然后，使用GPT-3.5来基于这些段落生成问题。生成的问题和抽样的文章对微调数据构成了一个新的文本对。

**混合检索**

M3嵌入统一了嵌入模型的三种常见的检索功能，即密集检索、词汇（稀疏）检索和多向量检索。

**密集检索**

输入查询q基于文本编码器转换为隐藏状态Hq。使用特殊令牌“[CLS]”的规范化隐藏状态来表示查询：同样，也可以将文本p嵌入为。因此，查询和文本之间的相关性得分是通过两个嵌入之间的内部积来衡量的。

**词汇检索**

输出嵌入还用于估计每个术语的重要性，以促进词汇检索。对于查询中的每个术语t（一个术语对应于一个标记），术语的权重计算为

其中，是将隐藏状态映射到一个浮点数的矩阵。如果一个项t在查询中出现多次，则只保留最大权重。用同样的方法来计算文章中每一项的权重。基于估计项的权重，通过在查询和文章中共存的项（记为q∩p）的联合重要性来计算查询和文章之间的相关性得分

**多向量检索**

作为密集检索的一种扩展，多向量方法利用整个输出嵌入来表示查询和文章：

其中，是可学习的投影矩阵。根据ColBert，使用后期交互来计算细粒度的相关性得分：

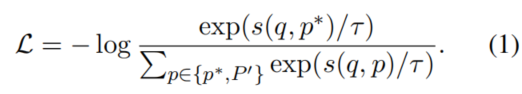
N和M是查询和文章的长度。

由于嵌入模型的多功能性，检索过程可以在一个混合的过程中进行。首先，每种方法都可以单独检索到候选结果（多向量方法由于成本较高，可以免除这一步骤）。然后，根据综合相关性评分，对最终的检索结果进行重新排序：



**自我知识蒸馏**

通过训练嵌入模型来区分正样本和负样本。对于每种检索方法，希望查询的正样本分配比负样本更高的分数。因此，训练过程是为了最小化InfoNCE的损失，其一般形式由以下损失函数表示：



这里，p∗和P’表示查询q的阳性和阴性样本；s（·）是中的任何函数。

不同的检索方法的训练目标可能会相互冲突。因此，原生的多目标训练可能不利于嵌入的质量。为了便于优化多个检索函数，提出在自知识蒸馏的基础上统一训练过程。具体来说，基于集成学习原理，不同检索方法对其异质性的预测可以整合为一个更准确的相关性得分。在最简单的形式中，积分可以只是不同预测分数的总和：

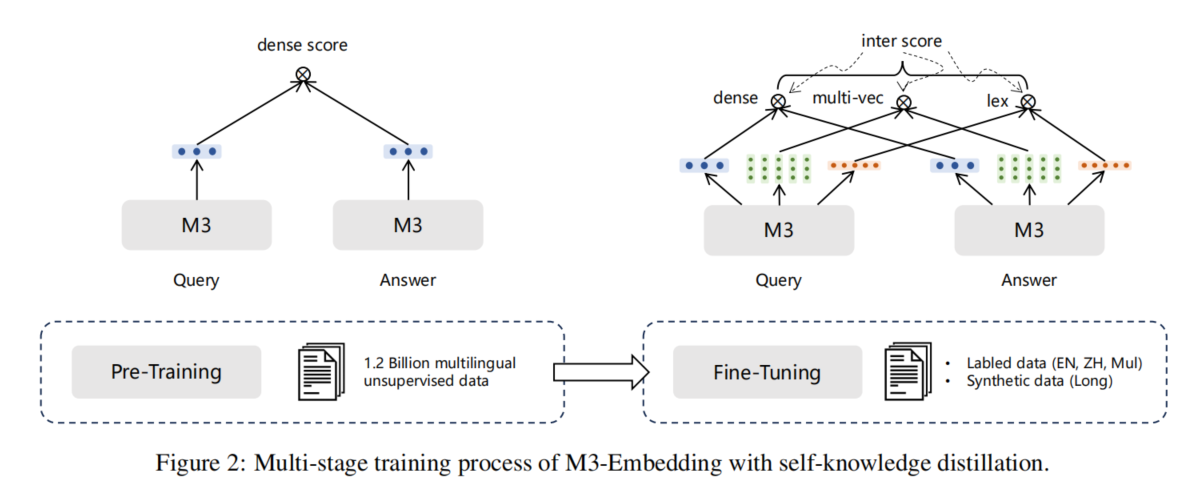


在以往的研究中，嵌入模型的训练质量可以受益于知识蒸馏，知识蒸馏利用了来自另一个排分模型的细粒度软标签。在这里，简单地使用s\_inter作为教师，其中每种检索方法的损失函数修改为：



最后，推导了在L和L’的线性组合下，自认识精馏的最终损失函数：

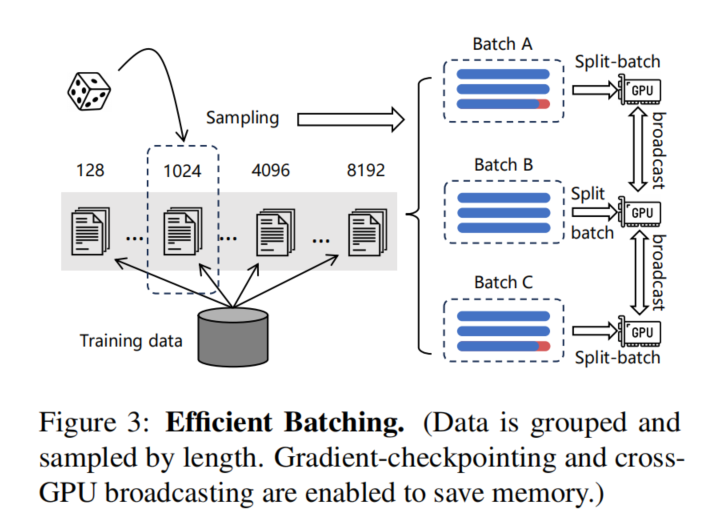
整个培训过程是一个多阶段的工作流程（如图2）。使用XLMRoBERTa预训练模型进一步通过RetroMAE方法作为基本文本编码器。首先，利用大量的无监督数据对文本编码器进行预训练，其中只有密集检索以对比学习的基本形式进行训练。将自知识蒸馏法应用于第二阶段，其中对嵌入模型进行了微调，建立了三种检索功能。在这一阶段同时使用标记数据和合成数据，其中按照ANCE方法为每个查询引入硬负样本。



**高效分组**

嵌入模型需要从不同的多语言数据中学习，以完全捕获不同语言的一般语义。它还需要保持batch的大小尽可能大（其中可以利用大量的批内负样本），以确保文本嵌入的区别性。由于GPU的内存和计算能力的限制，通常会将输入数据截断成短序列，以进行高吞吐量的训练和大批量的批处理。然而，对于M3嵌入并不是一个可行的选择，因为它需要从短序列和长序列数据中学习，才能有效地处理不同粒度的输入。在本文的工作中，通过优化批处理策略来提高训练效率，这使高训练吞吐量和大批处理规模。

具体来说，训练数据按序列长度进行分组进行预处理（如图3）。当生成一个小批量时，训练实例从同一组中采样。由于序列长度相似，它显著减少了序列填充（用红色标记），促进了更有效的GPU的利用。此外，为不同GPU采样训练数据时，随机种子总是固定的，这确保了负载平衡，并最小化了每个训练步骤中的等待时间。此外，在处理长序列训练数据时，将小batch进一步划分为子batch。使用梯度检查点迭代地编码每个子批处理，并收集所有生成的嵌入。该方法可以显著增加批处理的大小。例如，当处理长度为8192的文本时，批处理大小可以增加20倍以上。最后，广播来自不同GPU的嵌入，允许每个设备在分布式环境中获得所有的嵌入，扩大了batch内负样本的规模。



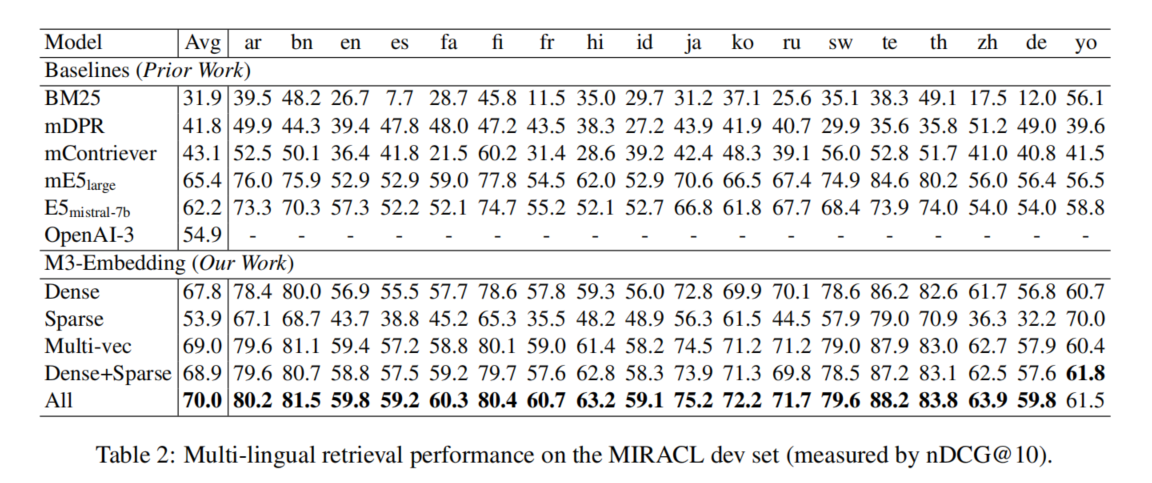
然而，用户可能缺乏足够的计算资源或数据来训练长文本模型。因此，本文还提出了一种MCLS策略来增强模型的长文本能力。该策略利用多个CLS令牌来捕获在推理过程中应用的文本语义。

**实验**

在三个任务来评估模型：多语言检索、跨语言检索和长文本检索。

**多语言检索**

使用Pyserini评估本文方法，并使用nDCG@10作为主要评估指标。在实验中，对BM25使用了与M3相同的标记器（即XLM-Roberta的标记器）。使用来自XLM-Roberta的相同词汇表还可以确保这两种方法具有相同的检索延迟。如表2所示，所有三种方法的协作（表示为All）带来最好的性能，在每种单独的检索上也取得了极好的性能。

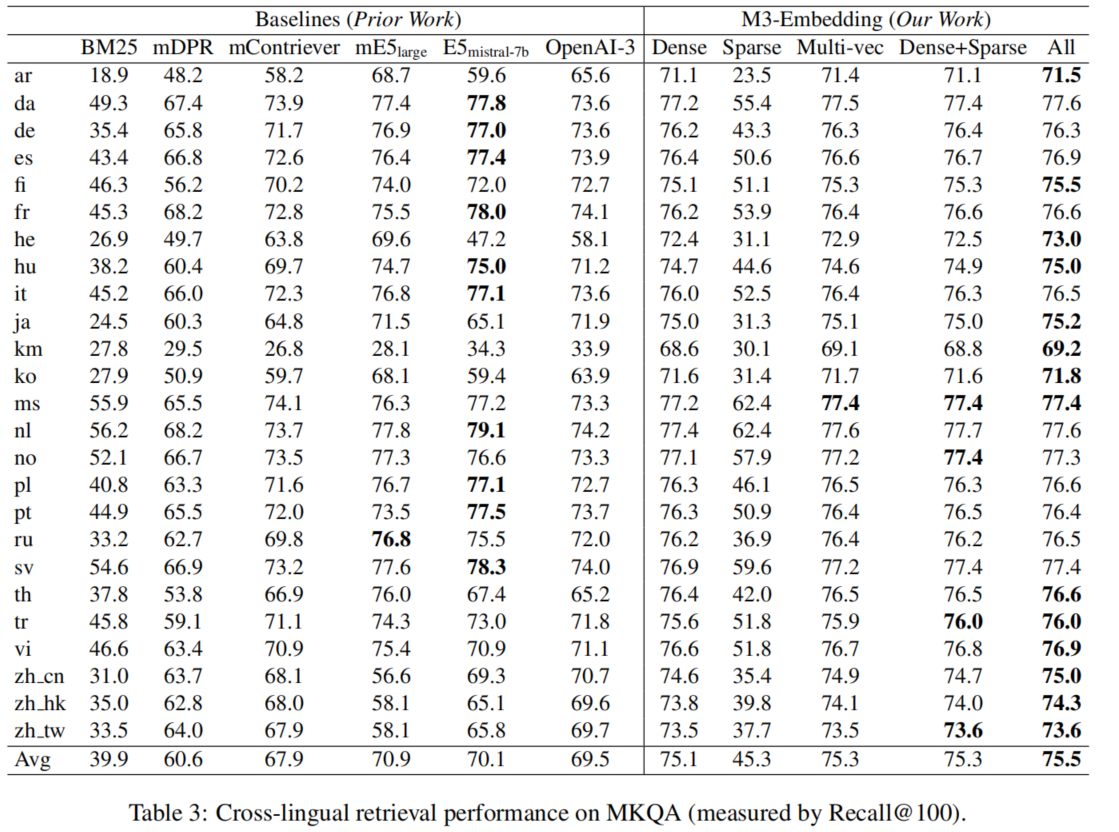


**跨语言检索**

使用MKQA基准测试对跨语言检索性能进行了评估。对于每个查询，它都需要从英语维基百科语料库中检索标注的段落。在实验中，使用BEIR9提供的语料库。根据之前的研究（Karpukhin et al.，2020），使用Recall@100作为主要指标。

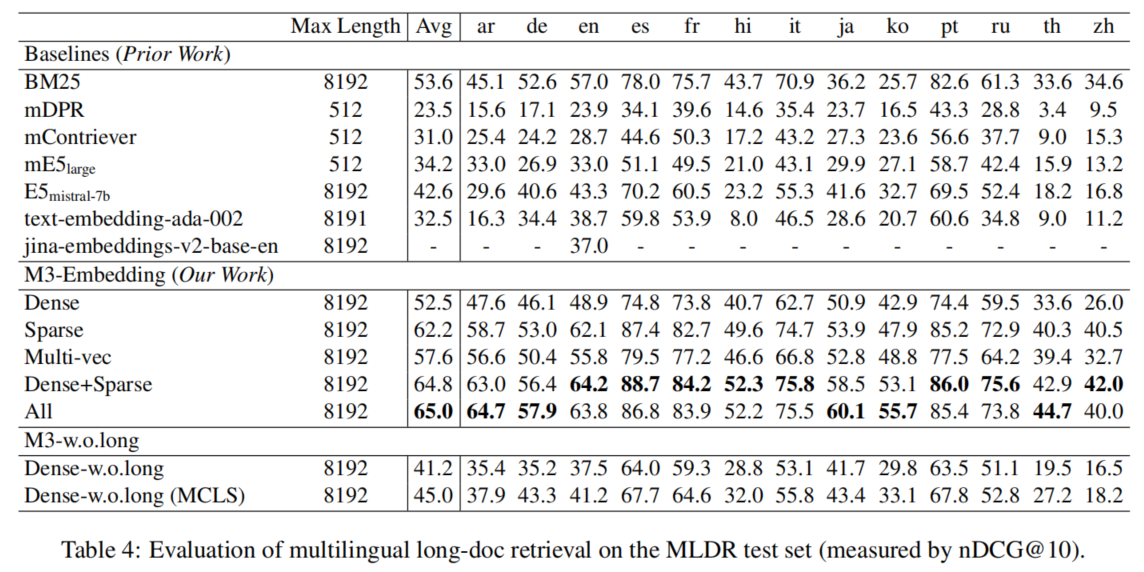
实验结果如表3所示。与在多语言检索中观察到的类似，M3嵌入产生优越的性能。它的密集检索功能明显优于其他基线方法。不同检索方法的协作带来了进一步的改进，导致了跨语言检索的最佳经验性能。

此外，还可以观察到以下有趣的结果。首先，性能差距不像MIRACL那么显著，像E5mistral-7b这样的竞争基线能够在一些测试语言上产生类似甚至更好的结果。然而，基线在许多其他语言中也容易表现不佳，特别是那些资源匮乏的语言，如ar、km、he等。相比之下，M3嵌入在所有语言中保持相对稳定的性能，这很大程度上归因于它在全面的无监督数据上的预训练。其次，虽然M3嵌入（Sparse）仍然优于BM25，但与其他方法相比性能较差。这是因为由于查询和文本以不同的语言呈现，跨语言检索只有非常有限的共存术语。



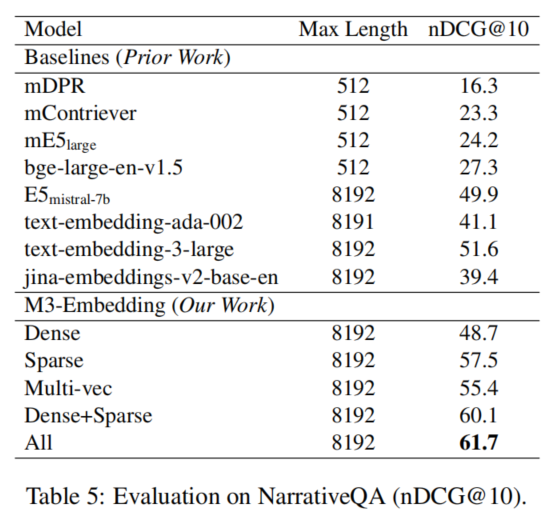
**多语言长文本检索**

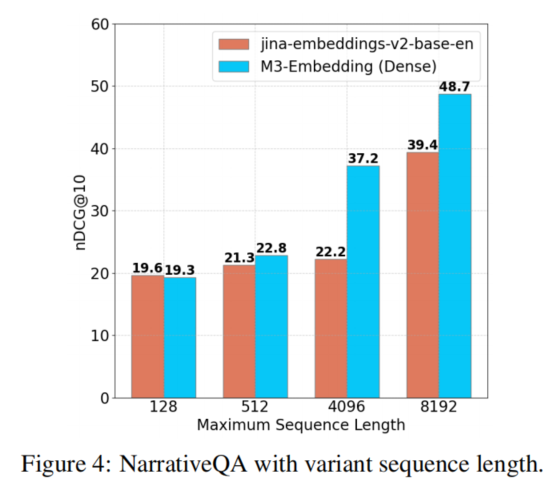
通过两个基准来评估较长序列的检索性能：MLDR和NarrativeQA。对MLDR的评价结果见表4。M3（Sparse）是一种更有效的长时间文档检索方法，它比密集的方法又提高了约10分。此外，多维检索也令人印象深刻，比M3（密集）提高了5.1个+点。最后，结合不同的检索方法，平均性能为65.0。



为了探究M3嵌入在长文档检索中的竞争力的原因，通过从微调阶段删除长文档数据(表示为w.o.long)经过这个修改，Dense-w.o.long仍然可以优于大多数基线，这表明它的经验优势在训练前阶段就已经建立。还提出了一个简单的策略MCLS来解决没有数据或没有GPU资源进行文档检索微调情况。实验结果表明，MCLS可以显著提高文档检索性能（41.2→45.0）。

用narrativeQA（表5）进行了进一步的分析，其中有与MLDR相似的观察结果。此外，随着序列长度的增加，本文方法逐渐扩大了其比基线的优势（图4），这反映了它在处理长输入方面的优越性能。

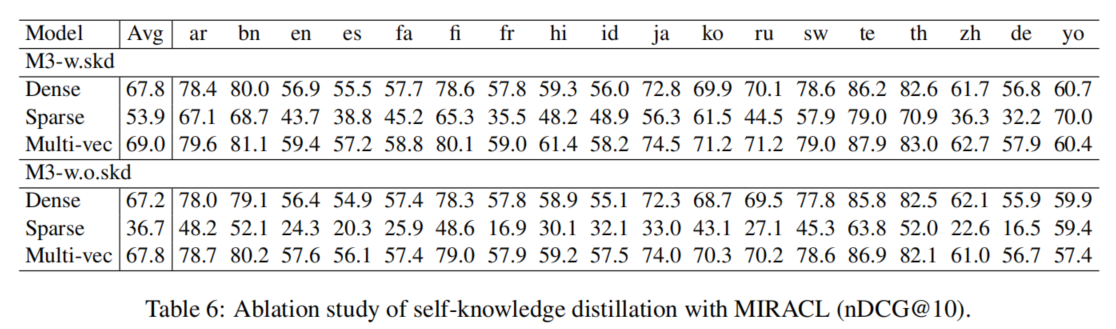




**消融实验**

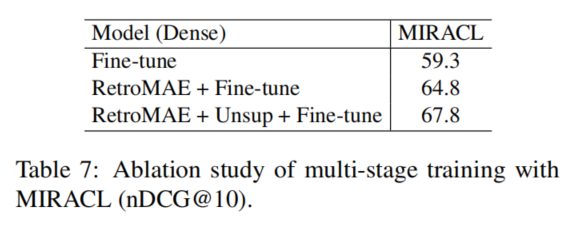
**自知识蒸馏**

本消融研究旨在分析自知识蒸馏的影响。具体来说，禁用了蒸馏处理，并独立训练每种检索方法（记为M3-w.o.skd）。根据对MIRACL的评价（表6），原始的方法，即M3 w.skd，在所有情况下都比消融方法，即致密、稀疏、多维数都具有更好的性能。值得注意的是，这对稀疏检索的影响更为明显，这表明了密集检索方法和稀疏检索方法之间的不兼容性。



**多阶段训练的影响**

还进行了实验来探讨不同阶段的影响。Fine-tuning表明直接从xlm-roberta模型进行微调，RetroMAE+Fine-tuning是指对使用RetroMAE训练的模型进行微调。同时，RetroMAE+Unsup+Fine-tuning涉及到对使用RetroMAE训练的模型进行微调，然后对无监督数据进行预训练。结果汇总见表7。可以看到，RetroMAE可以显著提高检索性能，而对无监督数据进行预训练可以进一步提高嵌入模型的检索能力。



**结论**

在本文中，提出了M3嵌入技术，它在支持多语言检索、处理不同粒度的输入和统一不同的检索功能方面实现了显著的多功能性。研究人员对训练数据进行全面、高质量的管理，通过自知识蒸馏优化学习过程，并通过有效的批处理改进训练量大小。实验结果验证了M3嵌入的有效性，在多语言检索、跨语言检索和多语言长文本检索任务方面具有优越的性能。