C-Pack：推进中文普通话嵌入的打包资源

Shitao Xiao ♠ **Zheng Liu** ♠† **Peitian Zhang** ♠ **Niklas Muennighoff** ⋆

♠北京人工智能学会 ⋆拥抱的脸

stxiao@baai.ac.cn {zhengliu1026,namespace.pt,n.muennighoff}@gmail.com

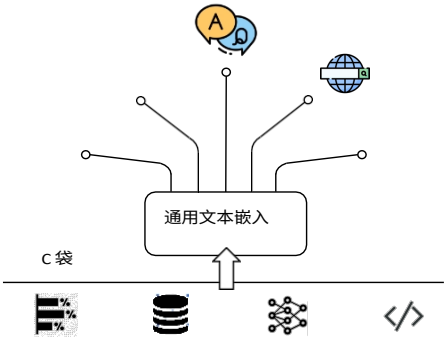
摘要

我们介绍 **C-Pack**，它是一个资源包，极大地推动了通用中文嵌入领域的发展。**C-Pack** 包括三个关键资源。1) **C-MTEB** 是一个全面的中文文本嵌入基准，涵盖 6 个任务和 35 个数据集。2) **C-MTP** 是一个海量文本嵌入数据集，由已标注和未标注的中文语料库组成，用于训练嵌入模型。3) **C-TEM** 是一个涵盖多种规模的嵌入模型系列。我们的模型在 **C-MTEB** 上的表现优于之前所有的中文文本嵌入模型，发布时最高可达 +10%。我们还整合并优化了 **C-TEM** 的整套训练方法。除了普通中文嵌入资源，我们还发布了英文文本嵌入的数据和模型。英文模型在 MTEB 基准测试中达到了最先进的性能；同时，我们发布的英文数据是 中文数据的 2 倍。所有这些资源都在 https://github.com/FlagOpen/FlagEmbedding 上公开发布。

1 导言

文本嵌入是自然语言处理和信息检索领域的一个长期课题。通过用潜在语义向量表示文本，文本嵌入可以支持各种应用，如网络搜索、问题解答和检索增强语言建模（Karpukhin 等人，2020；Lewis 等人，2020；Guu 等人，2020）。最近大语言模型（LLM）的流行使文本嵌入变得更加重要。由于 LLMs 的固有局限性，如世界知识和行动空间，通过知识库或工具使用的外部支持是必要的。文本嵌入对于连接 LLM 与这些外部模块至关重要（Borgeaud 等人，2022 年；Qin 等人，2023 年）。

†. 通讯作者



C-MTEB C-MTP C-TEM 食谱

图 1：**支持一般中文嵌入的 C-Pack 资源。**

由于应用场景多种多样，因此需要一个统一的嵌入模型，以处理任何应用场景（如问题解答、语言建模、转换）中的各种用途（如检索、排序、分类）。然而，学习通用文本嵌入比学习特定任务的文本嵌入更具挑战性。以下因素至关重要：

- **数据**开发通用文本嵌入技术提出了更高的 de-嵌入对训练数据的*规模*、*密度*和*质量*都有影响。要实现嵌入的高分辨力，可能需要超过数亿的训练实例（Izac- ard 等人，2021；Ni 等人，2021b；Wang 等人，2022b），这比典型的特定任务数据集，如 MS MARCO（Nguyen 等人，2016）和 NLI （Bowman 等人，2015；Williams 等人，2017）要大得多。除了规模，训练数据还需要从广泛的来源收集，以提高不同任务间的通用性（Izacard 等人，2021 年；Wang 等人，2022b）。最后，规模和多样性的增加可能会带来噪音。因此，在利用收集到的数据进行嵌入训练之前，必须对其进行适当的清理（Wang 等人，2022b）。

• **训练**通用文本嵌入的训练有赖于两个关键要素：a
这需要合适的骨干编码器和适当的训练配方。虽然可以求助于 BERT（Devlin 等人，2018 年）和 T5（Raffel 等人，2020 年）等通用预训练模型，但通过使用大规模无标记数据进行预训练，可以大大提高文本嵌入的质量（Izacard 等人，2021 年；Wang 等人，2022b）。此外，训练通用文本嵌入不是依靠单一算法，而是需要复合配方。特别是，它需要以嵌入为导向的预训练来预处理文本编码器（Gao 和 Callan，2021 年），通过复杂的负抽样来提高嵌入的可辨别性（Qu 等人，2020 年），以及基于指令的微调（Su 等人，2022 年；Asai 等人，2022 年），以整合文本嵌入的不同表示能力。

• **基准**。另一个前提条件是建立适当的基准、
在这里可以全面评估文本嵌入所需的所有能力。BEIR （Thakur 等人，2021 年）提供了 18 个集合，用于评估嵌入在不同检索任务（如问题解答和事实检查）中的一般性能。随后，MTEB（Muennighoff 等人，2022a）对嵌入式进行了更全面的评估，并对 BEIR 进行了扩展。它整合了 56 个数据集，可对文本嵌入的所有重要功能（如检索、排序、聚类等）进行联合评估。

总之，通用文本嵌入的发展需要从数据、编码器模式、训练方法和基准测试等多方面综合推动。近年来，这一领域取得了持续进展，例如 Contriever（Izacard 等人，2021 年）、E5（Wang 等人，2022 年 b）和 Ope- nAI 文本嵌入（Neelakantan 等人，2022 年）的工作。然而，这些著作大多面向英语世界。相比之下，由于一系列限制因素，通用中文嵌入模型的竞争力严重不足：既没有准备充分的训练资源，也没有合适的基准来评估通用性。

为应对上述挑战，我们提出了一个名为 **C-Pack 的**资源包，从以下几个方面促进通用中文电子书的发展。

MTEB 的中文扩展。[1]**C-MTEB** 收集了属于 6 类任务的 35 个公开数据集。得益于 **C- MTEB** 的规模和多样性，中文嵌入的所有主要能力都能得到可靠的测量，使其成为评估中文文本嵌入通用性的最合适基准。

- **C-MTP**（中文海量文本对）。我们创建了一个包含 1 亿对文本的海量训练数据集、

C-MTP 整合了从 "悟道"（Yuan et al.，2021）中收集的标注数据和非标注数据，" 悟道 " 是用于预训练中文语言模型的最大语料库之一。**C-MTP 不仅**数据量大，而且经过清洗以确保数据质量。

- **C-TEM**（中文文本嵌入模型）。我们提供了一系列训练有素的模型，用于

中文普通文本嵌入。有三种可选的模型大小：小型（24M ）、基本型（102M）和大型（326M），用户可以灵活地权衡效率和效果。我们的模型在通用性方面实现了巨大飞跃

- **C-MTEB**（中文大文本嵌入基准）。该基准是作为

：在 **C-MTEB** 的所有方面，**C-TEM 都**大大优于之前所有的中文文本嵌入模型。**C-TEM** 除了可以直接应用外，还可以利用额外的数据进行微调，以获得更好的特定任务性能。

- **培训配方**。在重新获取资源的同时，我们整合并优化了培训方法。

它包括面向嵌入的文本编码器的预训练、通用对比学习和特定任务的微调。训练秘诀的发布将有助于社区重现最先进的方法，并在此基础上不断进步。

总之，C-Pack 为人们**应用**通用中文文本嵌入提供了一个可选项。它大大推进了**培训**和**评估**工作，为这一领域的未来发展奠定了坚实的基础。

2 C 袋

在本节中，我们首先介绍 C-Pack 中的资源：基准 **C-MTEB** 、训练数据 **C-MTP** 和模型类 **C-TEM** 。然后，我们将讨论训练配方，它能让我们根据所提供的资源训练最先进的通用中文嵌入模型。

1. https://huggingface.co/spaces/mteb/leaderboard

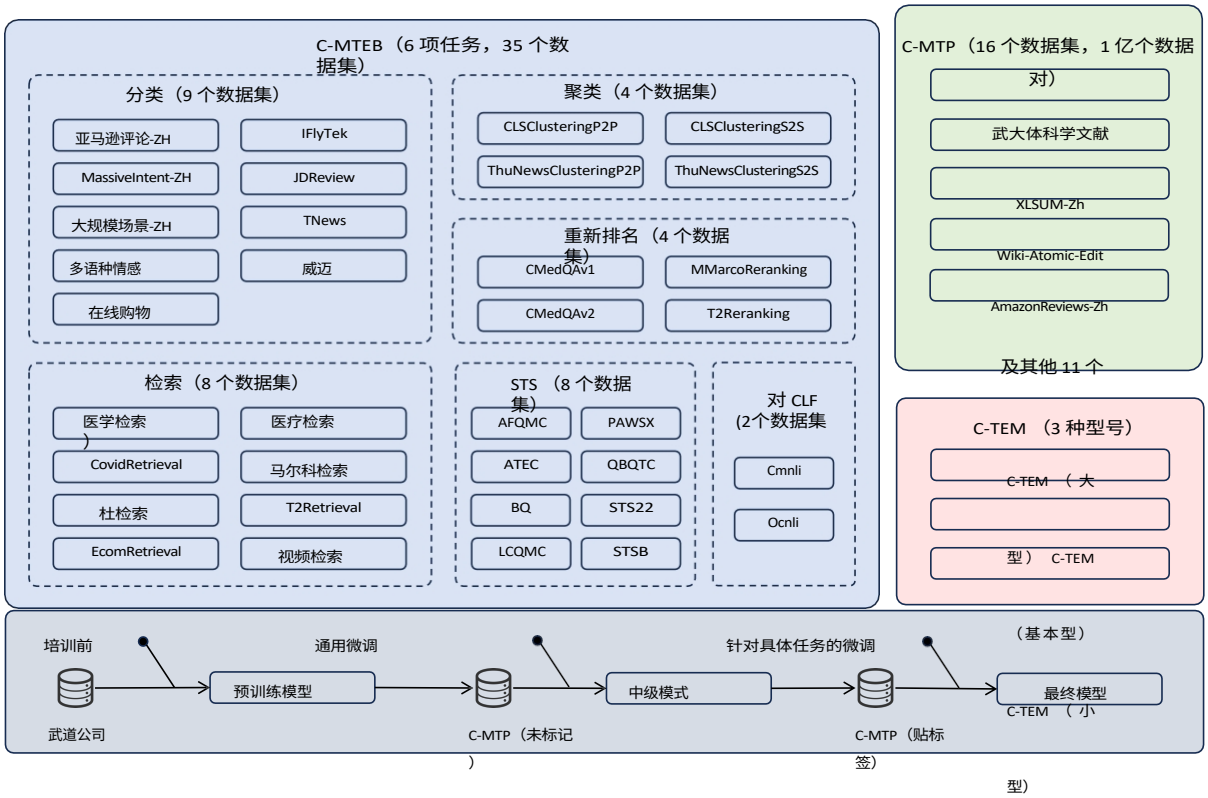


图 2：**C-Pack 概述。C-MTEB** 是中文文本嵌入的基准。**C-MTP** 是大规模中文嵌入训练数据集。**C-TEM** 是最先进的中文嵌入模型。底部显示的是训练配方。

2.1 基准：C-MTEB

C-MTEB 的建立是为了全面评估中文嵌入的通用性（图 2）。在过去几年中，业界提出了研究中文文本嵌入的基本数据集，如 CMNLI（Xu 等，2020a）、DuReader（He 等，2017）、T^2 Ranking（Xie 等，2023）。然而，这些数据集都是独立策划的，而且只关注文本嵌入的一种特定能力。因此，我们创建了 **C-MTEB，**目的是：**1）**全面收集相关数据集；**2）**对数据集进行分类；**3）**对评估管道进行标准化和整合。

其中，我们总共收集了 35 个公共数据集，所有这些数据集都可以用来评估 Chi-nese 文本嵌入。所收集的数据集根据其可能评估的嵌入能力进行了分类。共有 6 组评估任务：检索、重新排序、STS（语义文本相似性）、分类、配对分类和聚类，涵盖了中文文本嵌入的主要有趣方面。请注意，每个类别都有多个数据集。同一类别的数据集收集自不同的研究机构，相互之间具有互补性。

确保充分评估相应的能力。

每项任务的性质及其评估指标简要介绍如下。

- **检索**。检索任务包括测试查询和大型语料库。对于每个查询
它能在语料库中找到前 k 个相似文档。检索质量可以通过不同截点下的排名和召回率指标来衡量。在这项工作中，我们采用了 BEIR（Thakur 等人，2021 年）的设置，将 NDCG@10 作为主要指标。
- **重新排序**。重新排序任务是通过测试查询及其候选文档列表来完成的。
文档（一个正面文档加 N 个负面文档）。对于每个查询，它都会根据嵌入相似度对候选文档重新排序。MAP 分数被用作主要指标。
- **STS**（语义文本相似性）。STS（Agirre 等人，2012、2013、2014、2015、2016 年）
任务是根据两个句子的嵌入相似性来测量它们之间的相关性。按照 Sentence-BERT 的原始设置（Reimers 和 Gurevych，2019 年），斯皮尔曼相关性是根据给定的标签计算的，其结果被用作主要指标。

数据集	C-MTP (无标签)	C-MTP (有标签)
消息来源	吴道、CSL、XLSUM-Zh、Amazon-Review-Zh、CMRC 等。	²-Ranking, mMARCO-Zh、DuReader, NLI-Zh
尺寸	*100M*	*838K*

表 1：**C-MTP 的组成**

• **分类**。分类任务重新使用 MTEB 的逻辑回归分类器

(Muennighoff 等人，2022a），其中提供的标签是根据输入嵌入进行预测的。平均精度被用作主要指标。

• **成对分类**。该任务处理一对的输入句子，其关系呈现为二值化标签。这种关系是通过嵌入相似性来预测的，其中平均预分是主要的衡量标准。

• **聚类**。聚类的任务是将感光元件和光学元件进行分组。将有意义的信息分组。按照最初的在 MTEB（Muennighoff 等人，2022a）的最终设置中，它使用迷你批次 k-means 方法进行评估，批次大小等于 32，k 等于迷你批次中的标签数。V-measure 分数被用作主要指标。

最后，嵌入式在每个任务中的能力是通过该任务所有数据集的平均性能来衡量的。嵌入式的总体通用性是通过 **C-MTEB** 中所有数据集的平均性能来衡量的。

2.2 训练数据：C-MTP

我们策划了最大的数据集 **C-MTP**，用于一般中文嵌入的训练。配对文本是文本嵌入训练的数据基础，例如一个问题和它的答案、两个仿写句子或两个关于同一主题的文档。为确保文本嵌入的通用性，配对文本需要同时具备大规模和多样化的特点。因此，**C-MTP** 的收集来源有两个：一是海量无标注数据的整理，*即 C*-**MTP（无标注）**；二是标注数据的综合收集，*即 C*-**MTP（有标注）**。数据收集过程简要介绍如下。

• **C-MTP（未标记）**。我们寻找多种不同的在这些语料库中，我们可以提取语义丰富的从纯文本中提取成对的结构，如副词-短语、标题-正文。我们的主要数据来源于开放的网络语料库。最能体现

武大语料库（Yuan 等人，2021 年）是用于预训练中文模型的最大的格式化数据集。我们为其中的每篇文章提取（标题、段落），形成文本对。按照同样的方法，我们还从其他类似的网络内容（如知乎、百科、新闻网站等）中收集此类文本对。除了公开的网络内容，我们还探索了其他公开的中文数据集来提取文本对，如 CSL（科学文献）、Amazon-Review-Zh （评论）、Wiki Atomic Edits（转述）、CMRC（机器阅读理解）、XLSUM-Zh（摘要）等。在这些数据集中，配对结构非常明显，可直接提取用于增强 **C-MTP（无标记）**。

从网络和其他公共资源中整理出来的文本对并不能保证是密切相关的。因此，数据质量可能是一个主要问题。在我们的工作中，我们采用了一种简单的策略，在将数据添加到 **C-MTP（无标记）**之前对其进行过滤。特别是，我们使用了第三方模型：Text2Vec- Chinese[2]对每一对文本的关系强度进行评分。我们根据经验选择了一个 0.43 的阈值，并放弃了得分低于阈值的样本。通过这样的操作，我们从未加工的语料库中筛选出了 1 亿个文本对。尽管操作简单，但我们发现，在人工审核样本时，它能有效去除无关的文本对，并为在 **C-MTP（无标记）**上训练的模型带来强大的经验性能。

- **C-MTP（带标记）**。 下列标记数据集是为 **C-MTP（标注）**收集的，原因是其质量和多样性：T^2 -Ranking (Xie et al., 2023)、DuReader (He et al., 2017; Qiu et al., 2022)、mMARCO (Bonifacio et al., 2021)和 NLI-Zh[3](包括 ATEC[4], BQ[5], LCQMC[6]、PAWSX[7]、CNSD[8]). 共有 838,465 个配对文本。虽然其规模远小于 **C-MTP（未标注）**，但大部分数据都是由人工标注的，因此确保了发布的高可信度。此外，**C-MTP（已标注）**还全面覆盖了文本嵌入的不同功能，如重新三值、排序、相似性比较等，有助于提高嵌入模型的通用性

2. https://huggingface.co/GanymedeNil
3. https://huggingface.co/datasets/shibing624/nli_zh
4. https://github.com/IceFlameWorm/NLP_Datasets/tree/master/ATEC
5. http://icrc.hitsz.edu.cn/info/1037/1162.htm
6. http://icrc.hitsz.edu.cn/info/1037/1162.htm
7. https://arxiv.org/abs/1908.11828
8. https://github.com/pluto-junzeng/CNSD

在微调之后。

鉴于规模和质量上的差异，**C- MTP（无标签）**和 **C-MTP（有标签）**被应用于不同的训练阶段，这共同导致了嵌入模型的强大性能。详细分析将在我们的训练配方中进行。

2.3 型号级别：C-TEM

我们为社区提供了一整套训练有素的嵌入模型。我们的模型采用类似于 BERT 的架构，其中最后一层的特殊标记 [CLS] 的隐藏状态经过训练后用作嵌入。模型有三种不同规模：大型（3.26 亿个参数）、基本型（1.02 亿个参数）和小型（2400 万个参数）。大规模模型实现了最高的一般表示性能，在目前公开的模型中遥遥领先。小规模模型在经验上也具有竞争力。

中的公共可用模型和其他模型选项。

C-TEM；此外，它速度更快，重量更轻，适合处理海量知识库

和高通量应用。由于全面覆盖了不同尺寸的机型，人们可以根据自己的需求灵活权衡运行效率和表现质量。

如前所述，**C-TEM** 中的模型经过了良好的训练，在各种任务中具有很强的通用性。同时，如果 1) 嵌入应用于特定场景，2) 针对应用场景提供训练数据，这些模型还可以进一步微调。经验证明，与 **C-TEM** 中的原始模型以及 BERT 等其他通用预训练编码器的微调模型相比，微调后的模型在应用中可能会带来更好的性能。换句话说，**C- TEM 不仅为**人们提供了直接使用的嵌入式编码，还为人们开发更强大的嵌入式编码奠定了基础。

2.4 培训食谱

• **预训练**。我们的模型是在海量纯文本上进行预训练的。

以便更好地支持嵌入任务。特别是，我们利用了武大语料库（Yuan 等，2021 年），这是一个用于中文模型预训练的巨大而高质量的数据集。我们利用了 RetroMAE（Liu 和 Shao，2022 年；Xiao 等，2023 年）中提出的 MAE 式方法，该方法简单而高效。污染文本被编码到其嵌入中，然后在轻量级解码器的基础上从中恢复出干净的文本：

$$\min \sum_{x \in \mathcal{X}} - \log \mathrm{Dec}(\mathbf{x} | \mathbf{e}), \quad \mathbf{e} \leftarrow \mathrm{Enc}(\tilde{\mathrm{X}}).$$

(Enc、Dec 表示编码器和解码器，X、$\tilde{X}$表示干净文本和污染文本）。

• **通用微调**。预训练模型在 **C-MTP（无标记）**上进行微调，通过对比学习，即学习如何从反面样本中辨别配对文本：

$$\text{分钟} \sum_{(p,q)} - \log \frac{e^{(\mathbf{e}p, \mathbf{e}q)/\tau}}{e^{(\mathbf{e}p, \mathbf{e}q)/\tau} + \sum_{Q'} e^{(\mathbf{e}p, \mathbf{e}_{q'})/\tau}}$$

C-TEM 的训练配方与 C-Pack 一起完全向公众出租（图 2）。我们的训练方法由三个主要部分组成：**1)** 使用纯文本进行预训练；**2)** 使用 C-**MTP（无标签）**进行对比学习；**3)** 使用 **C-MTP（有标签）**进行多任务学习。

(p 和 q 是配对文本，$q' \in Q'$ 是负数样品，τ 为温度）。一个关键因素是对比学习的关键在于负样本。我们并不刻意挖掘硬负样本，而是纯粹依靠批内负样本（Karpukhin 等人，2020 年），并采用大批量（多达 19200 个）来提高嵌入的判别能力。

- **针对具体任务的微调**嵌入

用 **C-MTP（标注）**对模型进行进一步微调。标注数据集的规模较小，但质量较高。然而，所包含的任务类型不同，其影响可能相互矛盾。在这里，我们采用了两种策略来缓解这一问题。一方面，我们利用基于指令的微调（Su 等人，2022 年；Asai 等人，2022 年），对输入进行区分，帮助模型适应不同的任务。对于每个文本对（p, q），任务特定指令 I_t 会被附加到文本对上。

查询方：$q' \leftarrow q + I_t$。该指令是一个口头的提示语，指定任务的性质，例如 "*搜索查询的相关段落*"。另一方面，负面采样也进行了更新：除了批内负面样本外，还为每个文本对（p, q）挖掘了一个硬负面样本 q'。硬负面样本是从任务的原始语料库中按照（Xiong et al.）

模型	尺寸	检索	STS	对 CLF	CLF	重新排名	群组	平均
Text2Vec (base)	768	38.79	43.41	67.41	62.19	49.45	37.66	48.59
Text2Vec（大）	1024	41.94	44.97	70.86	60.66	49.16	30.02	48.56
罗托（大）	1024	44.40	42.79	66.62	61.0	49.25	44.39	50.12
M3E（基础）	768	56.91	50.47	63.99	67.52	59.34	47.68	57.79
M3E（大）	1024	54.75	50.42	64.30	68.20	59.66	**48.88**	57.66
多种。E5（基础）	768	61.63	46.49	67.07	65.35	54.35	40.68	56.21
多种。E5（大）	1024	63.66	48.44	69.89	67.34	56.00	48.23	58.84
OpenAI-Ada-002	1536	52.00	43.35	69.56	64.31	54.28	45.68	53.02
BGE (小)	512	63.07	49.45	70.35	63.64	61.48	45.09	58.28
BGE（基础）	768	69.53	54.12	77.50	67.07	64.91	47.63	62.80
BGE （大）	1024	**71.53**	**54.98**	**78.94**	**68.32**	**65.11**	48.39	**63.96**

表 2：**各种模型在 C-MTEB 上的性能。**

3 实验

我们进行实验研究的目的如下。**P1**.在 **C-MTEB** 上对不同的中文文本嵌入进行广泛评估。**P2**.通过 **C-TEM** 对文本嵌入进行经验验证。**P3**.探索 **C-MTP** 的实用价值。**P4**.探索训练配方带来的影响。

我们将以下流行的中文文本嵌入模型作为实验基准：Text2Vec-Chinese[9]基础模型和大型模型；Luotuo[10]；M3E[11]多语言 E5 （Wang 等人，2022b）和 OpenAI 文本嵌入 Ada 002。[12].第 2.1 节中介绍的主要指标针对 **C-MTEB** 中的每个任务进行了报告。

3.1 总体评价

如表 2 所示，我们在 **C-MTEB** 上对 **C-TEM** 与流行的中文文本嵌入进行了广泛评估。[13]我们得出以下结论。

首先，我们的模型远远优于现有的中文文本嵌入模型。我们的模型不仅在平均性能方面具有压倒性优势，而且在 **C-MTEB 的**大多数任务中都有显著提高。改进最大的是检索任务，其次是 STS、词对分类和重新排序。这些方面是文本嵌入最常见的功能，在搜索引擎、开放域问题解答和检索增强等应用中得到了广泛应用。

大型语言模型。虽然这些优势

9. https://huggingface.co/shibing624
10. https://huggingface.co/silk-road/luotuo-bert-medium
11. https://huggingface.co/moka-ai
12. https://platform.openai.com/docs/guides/embeddings
13. 我们的 **C-TEM** 模型在表格中被命名为 **BGE**。

虽然在分类和聚类任务中，C-TEM 的表现并不突出，但我们的表现仍与其他最具竞争力的模型相当或略胜一筹。上述观察结果验证了 **C-TEM** 的强大通用性。*我们的模型可以直接用于支持不同类型的应用场景。*

其次，我们观察到模型大小和嵌入维度的扩大带来的性能增长。特别是，当嵌入模型从小到大扩展时，平均性能从 58.28 提高到 63.96。除了平均性能的提高，所有评估任务的性能也都有所提高。与其他两个基线模型（Text2Vec 和 M3E）相比，我们的模型的扩展影响更为一致和显著。值得注意的是，尽管我们的模型规模缩小了很多，但我们的小型模型在经验上仍然具有竞争力，其平均性能甚至高于许多现有模型的大规模选项。因此，*它为人们提供了在嵌入质量和运行效率之间权衡的灵活性*：人们可以使用我们的大规模嵌入模型来处理高精度应用，也可以在高吞吐量场景下改用小规模模型。

3.2 详细分析

我们研究了 **C-MTP** 和**训练配方**的具体影响。相应的实验结果见表 3 和表 4。

首先，我们分析了训练数据 **C-MTP 的**影响。如前所述，**C-MTP** 由两部分组成。1) **C-MTP（无标记）**，用于通用微调；这一阶段产生的模型被称为中间模型（intermedi-tuning）。

模型	尺寸	检索	STS	对 CLF	CLF	重新排名	群组	平均
M3E（大）	1024	54.75	50.42	64.30	68.20	59.66	48.88	57.66
OpenAI-Ada-002	1536	52.00	43.35	69.56	64.31	54.28	45.68	53.02
w.o. 指导	1024	70.55	53.00	76.77	68.58	64.91	**50.01**	63.40
BGE-i	1024	63.90	47.71	61.67	**68.59**	60.12	47.73	59.00
BGE-i w.o. 预培训	1024	62.56	48.06	61.66	67.89	61.25	46.82	58.62
BGE-f	1024	**71.53**	**54.98**	**78.94**	68.32	**65.11**	48.39	**63.96**

表 3：**消融训练数据、C-MTP 和训练配方。**

2) **C-MTP（带标记），**在 BGE-i 的基础上进一步进行特定任务的微调；这一阶段产生的模型称为最终检查点，记为 BGE-*f*。根据我们的实验结果，**C-MTP（无标记）**和 **C-MTP（有标记）**都对嵌入的质量有很大的帮助。

关于 **C-MTP（无标注）**，尽管该数据集主要是从无标注语料库中整理出来的，但仅凭该数据集就为在其上训练的嵌入模型带来了强大的经验性能。与其他基线（如 Text2Vec、M3E 和 OpenAI 文本嵌入）相比，BGE-i 已经取得了更高的平均性能。进一步观察这些性能可以发现更多细节。一方面，**C-MTP（无标记）**对嵌入的检索质量有很大影响，BGE-i 在这一属性上明显优于基线。另一方面，由于 *BGE-i 在*其他方面（如 STS 和聚类）**的**表现与基准线接近，因此嵌入的一般能力主要是通过 **C-MTP（未标记）**建立起来的。*这使我们的嵌入模型处于非常有利的地位，可以进一步改进*。

至于 **C-MTP（带标记）**，数据集要小得多，但质量更好。在对 **C-MTP（带标记）**进行另一轮微调后，经验优势在最终检查点 *BGE-f* 上得到显著扩大，平均性能从 59.0（BGE-i）跃升到了 59.0（*BGE*-f）。63.96（BGE-*f*）。由于 **C- MTP（标注）**中的文本对主要来自检索和 NLI 任务，因此在检索、重新排序、STS 和文本对分类等密切相关的任务中，**C- MTP 的**改进最为显著。而在其他任务上，则保持或略微提高了性能。*这表明，高质量和多样化标记数据的混合使用能够为一项任务带来实质性的全面改进。*

预训练的嵌入模型。

我们进一步探讨了**训练配方的**影响，特别是对比学习、特定任务微调和预训练。

我们的训练方法有一个显著特点，那就是我们采用了较大的批量进行对比学习。根据之前的研究，嵌入模型的学习可能会受益于负样本的增加（Izacard 等人，2021 年；Qu 等人，2020 年；Muennighoff，2022 年）。鉴于我们对批内负样本的依赖性，需要尽可能扩大批次规模。在我们的实施过程中，我们采用了梯度检查点和跨设备嵌入共享的复合策略（Gao 等，2021b），这使得最大批次规模达到 19200 个。通过对 bz：256、2028、19,200 进行并行比较，我们*观察到随着批量大小*（以 bz 表示）*的扩大，嵌入质量得到了持续改善*。最显著的改进体现在检索性能上。这可能是由于检索通常是在大型数据库中进行的，因此嵌入需要具有很高的区分度。

另一个特点是在特定任务微调期间利用指令。特定任务指令是一种硬性提示。它区分了嵌入模型的激活，使模型能更好地适应各种不同的任务。我们在进行消融研究时取消了这一操作，记为 "*w.o. Instruct*"。与这种变化相比，原始方法 BGE-f 的平均性能更好。此外，BGE-f 在重新三重分类、STS、配对分类和重新排序方面也有更明显的经验优势。所有这些方面都与最后阶段的训练数据密切相关，即 **C-MTP（标注**）**，在这一阶段，模型将在一小部分任务上进行微调。这表明，*使用指令可能会大大有助于特定任务的微调。*

还有一个特点是，我们使用了一种特殊的

批量大小	256	2,048	19,200
检索	57.25	60.96	**63.90**
STS	46.16	46.60	**47.71**
对 CLF	**62.02**	61.91	61.67
CLF	65.71	67.42	**68.59**
重新排名	58.59	59.98	**60.12**
群组	**49.52**	49.04	47.73
平均	56.43	57.92	**59.00**

表 4：**批量大小的影响。**

我们使用预先训练的文本编码器来训练 **C-TEM，**而不是使用 BERT 和 RoBERTa 等常见选择（Liu 等人，2019 年）。为了探究其影响，我们用广泛使用的中文-RoBERTa 替换了预训练文本编码器。[14]，记为 "BGE- *i* w.o. pre-train"。根据与 *BGE-i* 的比较，*预训练文本编码器显著提高了检索能力，同时在其他方面也保持了类似的性能。*

4 相关工作

通用文本嵌入的重要性已得到广泛认可，这不仅是因为它在网络搜索和问题解答等典型应用中的广泛应用（Karpukhin 等人，2020 年），还因为它在增强大型语言模型中的基础作用（Lewis 等人，2020 年；Guu 等人，2020 年；Borgeaud 等人，2022 年；Izacard 等人，2022 年；Shi 等人，2023 年）。与传统的特定任务方法相比，通用文本嵌入需要广泛适用于不同的场景。近年来，人们在这一领域不断努力，提出了一系列著名的作品，如 Contriever（Izacard 等，2021 年）、GTR（Ni 等，2021b）、sentence-T5（Ni 等，2021a）、Sentence-Transformer（Reimers 和 Gurevych，2019 年）、E5（Wang 等，2022a）、OpenAI 文本嵌入（Neelakantan 等，2022 年）等。尽管这仍是一个未决问题，但最近的研究强调了以下重要因素。首先，希望训练数据是大规模和多样化的，这样嵌入模型才能从中学习识别不同类型的语义关系（Izacard 等人，2021；Wang 等人，2022b；Neelakantan 等人，2022）。其次，必须扩大嵌入模型的规模，因为大型文本编码器在不同的应用场景中更具通用性（Muennighoff，2022；Ni et al、

14. huggingface.co/hfl/chinese-roberta-wwm-ext-large

2021b,a），这与关于缩放 LLM 重要性的观察结果一致（Hoffmann 等人，2022 年；Rae 等人，2021 年；Brown 等人，2020 年；Chowdhery 等人，2022 年；Srivastava 等人，2022 年；Gao 等人，2021a；Li 等人，2023a；Allal 等人，2023 年；Muennighoff 等人，2023b）。第三，必须通过预训练（Liu 和 Shao，2022；Wang 等人，2022a）、负采样（Izacard 等人，2021；Wang 等人，2022a）和多任务微调（Su 等人，2022；Asai 等人，2022；Sanh 等人，2021；Wei 等人，2021；Muennighoff 等人，2022b，2023a；Chung 等人，2022）对训练配方进行优化。除此之外，建立适当的基准来评估文本嵌入的通用性也至关重要。与以往针对特定任务的评估（如 MS-MARCO（Nguyen 等，2016 年）、SentEval（Con- neau 和 Kiela，2018 年））不同的是，需要大量增加基准，以便评估 EM- bedding 在各种任务中的性能。BEIR（Thakur等人，2021 年；Kamalloo等人，2023年）是其中一项具有代表性的工作，它可以在不同的检索任务中对em- beddings进行评估。后来，MTEB（Muennighoff 等人，2022a）对其进行了扩展，可以对文本嵌入的所有主要方面进行全面评估。

鉴于上述分析，我们可以得出结论：一般文本嵌入对资源的依赖性很强，需要大量的要素，如数据集、模型和基准。因此，创建和公开发布相应的资源至关重要。

5 结论

我们提出了 C-Pack，以推动通用中文嵌入的进展。C-Pack 由三个核心资源组成 **1)** 基准 **C-MTEB**，涵盖 6 个主要嵌入任务和 35 个数据集，是评估中文嵌入通用性最全面的基准。**2)** 训练数据 **C-MTP**，由海量未标注语料库和高质量标注数据集组成。其前所未有的规模、多样性和质量有助于提高我们嵌入模型的通用性。**3)** 具有经验竞争力的模型 **C-TEM**。它们的不同规模为人们提供了在效率和嵌入质量之间进行权衡的灵活性。在提供这些资源的同时，还提供了整个培训配方。C-Pack 的公开发布促进了普通中文嵌入的使用，也为其未来的发展铺平了道路。

参考资料

Eneko Agirre、Carmen Banea、Claire Cardie、Daniel Cer 、 Mona Diab 、 Aitor Gonzalez-Agirre 、 Weiwei Guo 、 Inigo Lopez-Gazpio 、 Montse Maritxalar、Rada Mi- halcea 等人，2015 年。Semeval-2015任务2：语义文本相似性、英语、西班牙语和可预测性试点。*第 9 届语义评估国际研讨会（SemEval 2015）论文集*，第 252-263 页。

Eneko Agirre 、 Carmen Banea 、 Claire Cardie 、 Daniel M Cer、 Mona T Diab、 Aitor Gonzalez-Agirre、Weiwei Guo、Rada Mihalcea、German Rigau 和 Janyce Wiebe。2014.Semeval-2014任务10：多语言语义文本相似性。In *SemEval@ COLING*, pages 81-91.

Eneko Agirre、Carmen Banea、Daniel Cer、Mona Diab、Aitor Gonzalez Agirre、Rada Mihalcea、German Rigau Claramunt 和 Janyce Wiebe。2016.Semeval- 2016 任务 1：语义文本相似性、单语和跨语言评估。In *SemEval-2016.10th International Workshop on Semantic Evalua- tion; 2016 Jun 16-17; San Diego, CA.Stroudsburg (PA)：p. 497-511.*ACL（计算语言学协会）。

Eneko Agirre、Daniel Cer、Mona Diab 和 Aitor Gonzalez-Agirre。2012.Semeval-2012任务6：语义文本相似性试验。In * *SEM 2012：第一卷：主会议和共享任务论文集，以及第二卷：第六届语义评估国际研讨会（SemEval 2012）论文集*，第385-393页。

Eneko Agirre 、 Daniel Cer 、 Mona Diab 、 Aitor Gonzalez- Agirre 和 Weiwei Guo。2013.* sem 2013 共享任务：语义文本相似性。*第二届词法和计算语义学联合会议（* SEM），第1卷：主会议和共同任务会议记录：语义文本相似性，*第32-43页。

Loubna Ben Allal, Raymond Li, Denis Kocetkov, Chenghao Mou, Christopher Akiki, Carlos Munoz Ferrandis, Niklas Muennighoff, Mayank Mishra, Alex Gu, Manan Dey, et al.Santacoder: Don't reach for the stars! *arXiv preprint arXiv:2301.03988*.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih.2022.带指令的任务感知检索。*arXiv 预印本 arXiv:2211.09260*.

Luiz Bonifacio、 Vitor Jeronymo、 Hugo Queiroz Abonizio、Israel Campiotti、Marzieh Fadaee、Roberto Lotufo 和 Rodrigo Nogueira。2021 年：*ArXiv preprint arXiv:2108.13897*.

Sebastian Borgeaud 、 Arthur Mensch 、 Jordan Hoff-mann 、 Trevor Cai、Eliza Rutherford、Katie Milli-can、George Bm Van Den Driessche、Jean-Baptiste

Lespiau, Bogdan Damoc, Aidan Clark, et al.通过检索三狮标记改进语言模型。*国际语言学习会议*，第 2206-2240 页。PMLR.

Samuel R Bowman、Gabor Angeli、Christopher Potts 和 Christopher D Manning。2015.用于学习自然语言推理的大型注释语料库》，*arXiv preprint arXiv:1508.05326*.

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等，2020 年。语言模型是少数学习者。*神经信息处理系统进展*》，33：1877-1901。

Daniel Cer、Mona Diab、Eneko Agirre、Inigo Lopez- Gazpio 和 Lucia Specia。2017.Semeval-2017任务1：语义文本相似性--多语言和跨语言重点评估。

Jing Chen, Qingcai Chen, Xin Liu, Haijun Yang, Daohe Lu, and Buzhou Tang.2018.bq语料库：用于句子语义等价性识别的大规模特定领域中文语料库。*2018年自然语言处理实证方法会议论文集*》，第4946-4951页。

Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann 等，2022 年。Palm：*ArXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Bar- ret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al.Scaling instruction-finetuned language models. *ArXiv preprint arXiv:2210.11416*.

Alexis Conneau 和 Douwe Kiela.2018.Senteval：*ArXiv preprint arXiv:1803.05449*.

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018.Bert: Pre-training of deep bidirectional transformers for language understand- ing. *arXiv preprint arXiv:1810.04805*.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, et al.Massive：包含 51 种不同类型语言的 100 万例多语言自然语言理解数据集。*arXiv 预印本 arXiv:2204.08582*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou.2021a.少量语言模型评估框架。

Luyu Gao 和 Jamie Callan。2021.Condenser: a pre-training architecture for dense retrieval. *ArXiv preprint arXiv:2104.08253*.

Luyu Gao、Yunyi Zhang、Jiawei Han 和 Jamie Callan。2021b.在内存受限设置下扩展深度对比学习批量大小。*arXiv 预印本 arXiv:2101.06983*.

高天宇、姚兴成、陈丹琪。2021c.Simcse：*ArXiv preprint arXiv:2104.08821*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang.2020.检索增强语言模型预训练。*机器学习国际会议*，第 3929-3938 页。PMLR.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al.Dureader: a chinese ma- chine reading comprehension dataset from real-world applications. *ArXiv preprint arXiv:1711.05073*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al.训练计算最优的大型语言模型。*arXiv preprint arXiv:2203.15556*.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence S Moss.2020.Ocnli：*ArXiv preprint arXiv:2010.05444*.

Gautier Izacard、Mathilde Caron、Lucas Hosseini、Se- bastian Riedel、Piotr Bojanowski、Armand Joulin 和 Edouard Grave。2021.无监督密集形成检索与对比学习。*arXiv preprint arXiv:2112.09118*.

Gautier Izacard、Patrick Lewis、Maria Lomeli、Lucas Hosseini、Fabio Petroni、Timo Schick、Jane Dwivedi-Yu、Armand Joulin、Sebastian Riedel 和 Edouard Grave。2022.用三元增强语言模型进行少量学习。*ArXiv 预印本 arXiv:2208.03299*.

Ehsan Kamalloo、Nandan Thakur、Carlos Lassance、Xueguang Ma、Jheng-Hong Yang 和 Jimmy Lin。2023. 酿造 beir 的资源：*ArXiv preprint arXiv:2306.07471*.

弗拉基米尔-卡尔普欣、巴拉斯-奥格˘uz、苏元民、帕特里克-刘易斯、吴莱德尔、谢尔盖-埃杜诺夫、陈丹琪、易文涛。2020.开放域问题解答的密集通道检索。*arXiv预印本arXiv:2004.04906*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Red- field, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Ken- ton Lee, et al. 2019.自然问题：问题解答研究的基准。自然问题：*问题解答研究的基准。*

计算语言学协会，7：453- 466。

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Hein- rich Küttler, Mike Lewis, Wen-tau Yih, Tim Rock- täschel, et al.知识密集型神经网络任务的检索增强生成。*神经信息处理系统进展》*，33：9459-9474。

Jingyang Li and Maosong Sun.2007.文本分类的可扩展术语选择。*2007 年自然语言处理和计算自然语言学习经验方法联合会议论文集》（EMNLP-CoNLL）*，第 774-782 页。

Jingyang Li, Maosong Sun, and Xian Zhang.2006.作为中文文本分类特征的字和词的比较与半定量分析。*第 21 届国际计算语言学会议暨第 44 届计算语言学协会年会论文集》*，第 545-552 页。

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al.Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*.

Yudong Li, Yuqing Zhang, Zhe Zhao, Linlin Shen, Wei-jie Liu, Weiquan Mao, and Hui Zhang.2022.Csl: A large-scale chinese scientific literature dataset. *ArXiv preprint arXiv:2209.05034*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang.2023b.利用多阶段对比学习实现通用文本嵌入。*arXiv 预印本 arXiv:2308.03281*.

Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang.2018.Lcqmc：大规模中文问题匹配语料库。*第 27 届计算语言学国际会议论文集》*，第 1952-1962 页。

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man- dar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.2019. 罗伯塔：*ArXiv preprint arXiv:1907.11692*.

Zheng Liu 和 Yingxia Shao.2022.Retromae：*ArXiv preprint arXiv:2205.12035*.

龙定坤、高琼、邹宽、徐光伟、谢鹏军、郭瑞杰、徐健、蒋冠军、邢璐茜和杨平。2022.Multi-cpr：用于段落检索的多域中文数据集。*第 45 届国际信息检索研究与发展大会（ACM SIGIR Conference on Research and Development in Information Retrieval）论文集*，第 3046-3056 页。

Julian McAuley and Jure Leskovec.2013.隐藏的因素和隐藏的主题：用评论文本理解评分维度。RecSys '13, New York, NY, USA.美国计算机协会。

尼克拉斯-穆恩尼格霍夫2022.Sgpt：用于语义搜索的 Gpt 句子嵌入。*arXiv 预印本 arXiv:2202.08904*.

Niklas Muennighoff、Qian Liu、Armel Zebaze、Qinkai Zheng、Binyuan Hui、Terry Yue Zhuo、Swayam Singh、Xiangru Tang、Leandro von Werra 和 Shayne Longpre。2023a.Octopack：*ArXiv preprint arXiv:2308.07124*.

Niklas Muennighoff、Alexander M Rush、Boaz Barak、Teven Le Scao、Aleksandra Piktus、Nouamane Tazi、Sampo Pyysalo、Thomas Wolf 和 Colin Raffel。2023b.扩展数据约束语言模型。*arXiv 预印本 arXiv:2305.16264*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers.2022a.Mteb：*ArXiv preprint arXiv:2210.07316*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al.*ArXiv preprint arXiv:2211.01786*.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 通过对比预训练进行文本和代码嵌入。*arXiv 预印本 arXiv:2201.10005*.

Tri Nguyen、Mir Rosenberg、Xia Song、Jianfeng Gao、Saurabh Tiwary、Rangan Majumder 和 Li Deng。2016.Ms marco：人类生成的机器阅读理解数据集。

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang.2021a.Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *ArXiv preprint arXiv:2108.08877*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gus- tavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al.大型对偶编码器是可泛化的重三维编码器. *arXiv预印本 arXiv:2112.07899*.

秦宇佳、梁世豪、叶一宁、朱昆仑、严岚、卢雅茜、林彦凯、丛昕、唐相如、钱彪等 2023.Toolllm：促进大型语言模型掌握 16000 多个真实世界 apis。*arXiv 预印本 arXiv:2307.16789*.

Yifu Qiu, Hongyu Li, Yingqi Qu, Ying Chen, Qiaoqiao She, Jing Liu, Hua Wu, and Haifeng Wang.2022.Dureader_retrieval：*ArXiv preprint arXiv:2203.10232*.

曲颖琦、丁雨辰、刘静、刘凯、任瑞阳、赵伟昕、董大祥、吴华、王海峰。2020.Rocketqa：用于开放域问题解答的密集段落检索的优化训练方法。*arXiv preprint arXiv:2010.08191*.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susan- nah Young, et al.扩展语言模型：来自训练地鼠的方法、分析和见解。*arXiv 预印本 arXiv:2112.11446*.

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J Liu。2020.用统一的文本到文本转换器探索迁移学习的极限》。*机器学习研究期刊》*，21（1）：5485-5551。

Nils Reimers 和 Iryna Gurevych.2019.Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv preprint arXiv:1908.10084*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al.*ArXiv preprint arXiv:2110.08207*.

Teven Le Scao, Angela Fan, Christopher Akiki, El- lie Pavlick, Suzana Ilic´, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022a.布鲁姆：一个 176b 参数的开放式多语言语言模型。*arXiv 预印本 arXiv:2211.05100*.

Teven Le Scao, Thomas Wang, Daniel Hesslow, Lu- cile Saulnier, Stas Bekman, M Saiful Bari, Stella Bideman, Hady Elsahar, Niklas Muennighoff, Jason Phang, et al. 2022b.如果有一百万 gpu 小时，该训练什么语言模型？*arXiv preprint arXiv:2210.15424*.

Weijia Shi、Sewon Min、Michihiro Yasunaga、Min- joon Seo、Rich James、Mike Lewis、Luke Zettle- moyer 和 Wen-tau Yih。2023.Replug：检索增强黑箱语言模型。*ArXiv 预印本 arXiv:2301.12652*.

Aarohi Srivastava、Abhinav Rastogi、Abhishek Rao、Abu Awal Md Shoeb、Abubakar Abid、Adam Fisch、Adam R Brown、Adam Santoro、Aditya Gupta、Adrià Garriga-Alonso 等，2022 年。超越模仿游戏：*ArXiv preprint arXiv:2206.04615*.

Hongjin Su, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, Tao Yu, et al.一个嵌入器，任何任务：指令调整的文本嵌入。*arXiv 预印本 arXiv:2212.09741*.

Nandan Thakur、Nils Reimers、Andreas Rücklé、Ab- hishek Srivastava 和 Iryna Gurevych。2021.Beir：*ArXiv preprint arXiv:2104.08663*.

James Thorne、Andreas Vlachos、Christos Christodoulopoulos 和 Arpit Mittal。2018.Fever: a large-scale dataset for fact extraction and verification. *ArXiv preprint arXiv:1803.05355*.

王亮、杨楠、黄小龙、焦斌兴、杨林军、蒋大新、Rangan Majumder 和魏福如。2022a.Simlm：*ArXiv preprint arXiv:2207.02578*.

王亮、杨楠、黄小龙、焦斌兴、杨林军、蒋大新、Rangan Majumder 和魏福如。2022b.弱监督对比预训练的文本嵌入。*arXiv preprint arXiv:2212.03533*.

Jason Wei、Maarten Bosma、Vincent Y Zhao、Kelvin Guu、Adams Wei Yu、Brian Lester、Nan Du、An- drew M Dai 和 Quoc V Le。2021.Finetuned lan- guage models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Adina Williams、Nikita Nangia 和 Samuel R Bowman。2017.通过推理进行句子理解的广覆盖挑战语料库》，*arXiv preprint arXiv:1704.05426*.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao.2023.Retromae-2：用于预训练面向检索的语言模型的双工屏蔽自动编码器。

Xiaohui Xie, Qian Dong, Bingning Wang, Feiyang Lv, Ting Yao, Weinan Gan, Zhijing Wu, Xiangsheng Li, Haitao Li, Yiqun Liu, et al.T2ranking：*ArXiv preprint arXiv:2304.03679*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk.2020.用于密集文本检索的近似近邻负向对比学习. *arXiv preprint arXiv:2007.00808*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al.Clue: A Chinese language understanding evaluation benchmark. *ArXiv preprint arXiv:2004.05986*.

Liang Xu, Xuanwei Zhang, and Qianqian Dong.2020b.Cluecorpus2020：用于预训练语言模型的大规模中文语料库。*arXiv preprint arXiv:2003.01355*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge.2019.Paws-x：用于转述识别的跨语言对抗数据集。*arXiv preprint arXiv:1908.11828*.

杨志林、齐鹏、张赛正、Yoshua Bengio 、 William W Cohen 、 Ruslan Salakhutdinov 和 Christopher D Manning 。 2018.Hotpotqa ： *ArXiv preprint arXiv:1809.09600*.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang.2021.Wudaocorpora： 用于预训练语言模型的超大规模中文语料库。*AI Open*, 2:65-68.

Sheng Zhang, Xin Zhang, Hui Wang, Jiajun Cheng, Pei Li, and Zhaoyun Ding.2017.使用端到端字符级多尺度cnns的中医问答匹配。*应用科学*，7（8）：767.

Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu.2018.用于中文医疗问题答案选择的多尺度注意力交互网络。*IEEE Access*, 6:74061-74071.

A C-MTEB 数据集

名称	网址	说明	分类。	测试样品
		分类		
亚马逊评论分类（Muennighoff 等人，2022a；McAuley 和 Leskovec，2013 年）	https://hf.co/datasets/mteb/多条评论	阿玛的情绪 zon 评论	s2s	5,000
IFlyTek（Xu 等人，2020a）	https://hf.co/datasets/C-MTEB/IFlyTek-classification	应用程序脚本的长文本分类	s2s	2,600
JDReview (https://hf.co/datasets/kuroneko5943/jd21)	https://hf.co/datasets/C-MTEB/JDReview 分类	iPhone 评论	s2s	533
MassiveIntentClassification（Muennighoff 等人，2022a；FitzGerald 等人，2022）	https://hf.co/datasets/mteb/amazon_massive_intent	亚马逊 Alexa 虚拟助手的实用程序标注了相关意图	s2s	16,500
大规模情景分类（Muennighoff 等人，2022a；FitzGerald 等人，2022）	https://hf.co/datasets/mteb/amazon_massive_scenario	亚马逊 Alexa 虚拟助手用相关场景注释的实用程序	s2s	16,500
多语言情感》（McAuley 和 Leskovec）、2013)	https://hf.co/datasets/C-MTEB/多语种感官分类	Ama- zon 评论的情绪	s2s	3,000
在线购物 (https://github.com/SophonPlus/ChineseNlpCorpus/blob/master/datasets/online_shopping_10_cats/intro.ipynb)	https://hf.co/数据集/C-MTEB/网上购物分类	情感分析用户评论网站	s2s	1,000
TNews（Xu 等人，2020a）	https://hf.co/datasets/C-MTEB/新闻分类	短文分类新闻分类	s2s	10,000
Waimai (https://github.com/SophonPlus/ChineseNlpCorpus/blob/master/datasets/waimai_	https://hf.co/	情感分析	s2s	1,000

10k/intro.ipynb)	数据集/C-MTEB/ 外卖分类	的用户评论 外卖平台
聚类		

CLSClusteringP2P （Li 等人，2022 年）	https://hf.co/datasets/C-MTEB/CLSClusteringP2P	标题分组 + CLS 数据集摘要。根据主要类别对 13 组数据进行聚类。	p2p	10,000
CLSClusteringS2S （Li 等人，2022 年）	https://hf.co/datasets/C-MTEB/C-MTEB/CLSClusteringS2S	对 CLS 数据集中的标题进行聚类。根据主要类别对 13 套数据进行聚类。	s2s	10,000
ThuNewsClusteringP2P （Li 等人，2006 年；Li 和 Sun，2007 年）	https://hf.co/datasets/C-MTEB/ThuNewsClusteringP2P	标题分组 + THUCNews 数据集摘要	p2p	10,000
ThuNewsClusteringS2S （Li 等人，2006 年；Li 和 Sun，2007 年）	https://hf.co/datasets/C-MTEB/ThuNewsClusteringS2S	对 THUC 新闻数据集的标题进行聚类	s2s	10,000
配对分类				
Cmnli （Xu 等人，2020a,b；Conneau 和 Kiela，2018；Williams 等人，2017）	https://hf.co/datasets/C-MTEB/CMNLI	中文多流派 NLI	s2s	139,000
Ocnli （Hu 等人，2020 年）	https://hf.co/datasets/C-MTEB/OCNLI	原始中文自然语言推理数据集	s2s	3,000
重新排名				
T2Reranking （Xie et al、2023）	https://hf.co/datasets/C-MTEB/T2Reranking	T2Ranking A：大型中文通过率排名基准	s2p	24,382
马可检索（Bonifacio 等人，2021 年）	https://hf.co/datasets/C-MTEB/Mmarco-ranking	mMARCO 是 MS MARCO 段落排序数据集的多语言版本	s2p	7,437

CMedQAv1 (Zhang et al., 2017)	https://hf.co/datasets/C-MTEB/CMedQAv1-排名	中文社区医疗问题解答	s2p	2,000
CMedQAv2 (Zhang et al., 2018)	https://hf.co/datasets/C-MTEB/C-MTEB/CMedQAv2-排名	中文社区医疗问题解答	s2p	4,000

检索

T2Retrieval （Xie 等人，2023 年）	https://hf.co/datasets/C-MTEB/T2Retrieval	T2Ranking：大规模中文段落排名基准	s2p	24,832
马可检索（Bonifacio 等人，2021 年）	https://hf.co/datasets/C-MTEB/MMarcoRetrieval	mMARCO 是 MS MARCO 段落排序数据集的多语言版本	s2p	7,437
DuRetrieval （Qiu 等人，2022 年）	https://hf.co/datasets/C-MTEB/DuRetrieval	从网络搜索引擎检索段落的大规模芝麻基准测试	s2p	4,000
CovidRetrieval （Qiu 等人，2022 年）	https://hf.co/datasets/C-MTEB/CovidRetrieval	COVID-19 新闻报道	s2p	949
CmedqaRetrieval （Qiu 等人，2022 年）	https://hf.co/datasets/C-MTEB/CmedqaRetrieval	在线医疗咨询文本	s2p	3,999
EcomRetrieval （Long 等人，2022 年）	https://hf.co/datasets/C-MTEB/EcomRetrieval	从阿里巴巴收集的通道检索数据集 搜寻器 商业领域	s2p	1,000
医学检索（Long 等人，2022 年）	https://hf.co/datasets/ C-MTEB/MedicalRetrieval	系统的电子 从阿里巴巴收集的通道检索数据集	s2p	1,000
视频检索（Long 等人，2022 年）	https://hf.co/datasets/C-MTEB/VideoRetrieval	医疗领域的搜索引擎系统 从阿里巴巴收集的通道检索数据集 视频领域的搜索引擎系统	s2p	1,000
STS				

AFQMC （Xu 等人，2020a）	https://hf.co/datasets/C-MTEB/AFQMC	蚂蚁金融问题匹配库	s2s	3,861
ATEC (https://github.com/IceFlameWorm/NLP_Datasets/tree/master/ATEC)	https://hf.co/datasets/C-MTEB/ATEC	ATEC NLP sen- 句对相似性竞赛	s2s	20,000
BQ (Chen et al., 2018)	https://hf.co/datasets/C-MTEB/BQ	银行问题语义相似性	s2s	10,000

LCQMC (Liu et al., 2018)	https://hf.co/datasets/C-MTEB/LCQMC	A 大型中文 问题匹配语料库。	s2s	12,500
PAWSX（Yang 等人，2019 年）	https://hf.co/datasets/C-MTEB/PAWSX	经翻译的 PAWS 评估对	s2s	2,000
QBQTC （Xu 等人，2020a）	https://hf.co/datasets/C-MTEB/QBQTC	QQ 浏览器查询标题语料库	s2s	5,000
STSB（Cer 等人，2017 年）	https://hf.co/datasets/C-MTEB/STSB	翻译 STS-B 翻译成中文	s2s	1,360
STS-22（Muennighoff 等人，2022a）	https://hf.co/datasets/mteb/sts22-crosslingual-sts	中国新闻	p2p	656

表 5：**C-MTEB 数据集概览。**

B C-MTP 构成

我们挖掘了来自不同领域的大规模数据对。表 6 显示了各数据的详细信息。

数据源	文本对类型	# 对数	网址
cmrc2018	(查询，上下文)	9,669	https://huggingface.co/datasets/cmrc2018
杜蕾斯	(查询，上下文)	96,486	https://github.com/baidu/DuReader
simclue	$(句子_a, 句子)_b$	388,779	https://github.com/CLUEbenchmark/SimCLUE
csl	(标题、摘要)	394,846	https://arxiv.org/abs/2209.05034
多个亚马逊评论	(标题、正文)	157,762	https://huggingface.co/datasets/amazon_reviews_multi
维基原子编辑	(句子，已编辑)	1,213,688	https://huggingface.co/datasets/wiki_atomic_edits
mlqa	(问题，上下文)	70,594	https://huggingface.co/datasets/mlqa
xlsum	(标题、摘要) （标题、文本）	89,505	https://huggingface.co/datasets/csebuetnlp/xlsum
武都	(标题、段落)	37,318,330	https://data.baai.ac.cn/details/WuDaoCorporaText
杂项	-	60,260,341	-

表 6：**每个数据集的详细信息。**Misc 数据来自互联网，包括质量保证数据、纸质数据和新闻数据。

C 英语模型

利用我们的方法，我们还训练了一套英文文本嵌入模型，如表 7 所示。在撰写本文时，我们的英文 BGE 模型在英文 MTEB 基准（Muennighoff et al.我们的模型明显优于更大的模型，如拥有

71 亿个参数的 SGPT Bloom（Muennighoff，2022；Scao 等人，2022a,b）。我们将先前的先进水平绝对值提高了 1.1（Li 等人，2023b）。除了使用英文数据外，我们的训练方法与中文模型相同。我们首先在无监督数据集上进行微调，包括维基百科、CC-net、StackExchange、Reddit、S2orc 等数据集以及来自句子转换器的数据集。[15]然后，我们在有监督数据集上进一步微调，包括 NLI（Gao 等人，2021c）、FEVER（Thorne 等人，2018）、NQ（Kwiatkowski 等人，2019）、HotpotQA（Yang 等人，2018）、Quora、StackExchange Duplicates 和 MEDI（Su 等人，2022）。

15. https://huggingface.co/datasets/sentence-transformers/embedding-training-data

型号名称	暗淡。	平均	检索	群组	对 CLF	重新排名	STS	总结	CLF
BGE （大）	1024	**64.23**	**54.29**	46.08	**87.12**	**60.03**	**83.11**	31.61	**75.97**
BGE（基础）	768	63.55	53.25	45.77	86.55	58.86	82.4	31.07	75.53
BGE (小)	384	62.17	51.68	43.82	84.92	58.36	81.59	30.12	74.14
GTE （大型）	1024	63.13	52.22	**46.84**	85.00	59.13	83.35	31.66	73.33
GTE（基本型）	768	62.39	51.14	46.2	84.57	58.61	82.3	31.17	73.01
E5（大）	1024	62.25	50.56	44.49	86.03	56.61	82.05	30.19	75.24
教员-XL	768	61.79	49.26	44.74	86.62	57.29	83.06	32.32	61.79
E5（基础）	768	61.5	50.29	43.80	85.73	55.91	81.05	30.28	73.84
GTE (small)	384	61.36	49.46	44.89	83.54	57.7	82.07	30.42	72.31
OpenAI Ada 002	1536	60.99	49.25	45.9	84.89	56.32	80.97	30.8	70.93
E5（小）	384	59.93	49.04	39.92	84.67	54.32	80.39	31.16	72.94
ST5 (XXL)	768	59.51	42.24	43.72	85.06	56.42	82.63	30.08	73.42
MPNet（基础）	768	57.78	43.81	43.69	83.04	59.36	80.28	27.49	65.07
SGPT Bloom (7.1B)	4096	57.59	48.22	38.93	81.9	55.65	77.74	**33.60**	66.19

表 7：**英语模型在 MTEB 上的表现。**