预训练模型在科学事实验证任务中的应用研究

1 任务背景

在过去几年中，随着互联网的发展和社交媒体的普及，包含虚假科学信息（如假科学普及、恶意欺骗和网络谣言）的网络内容一直在快速增长，并广泛传播，不仅会扰乱人们的思想、心理和行为，而且会引发社会震荡，危害公共安全，损害公众利益。因此，核实虚假的科学信息变得至关重要。然而，行业专家和审查工作人员的时间和精力是有限的，而谣言传播速度之快、范围之广、影响之深，又迫切要求人们尽早发现并阻止它们。

事实验证任务是指判断给定文本中的声明语句为支持、反对或信息不足的任务。事实验证任务可以帮助人们快速、准确地判断声明语句的真实性，从而避免因误传信息而产生的负面影响，因此事实验证任务在现代社会中具有重要的意义和应用价值。

科学事实验证是指，给定输入的科学声明，依据语料库中支持或反对声明的文章验证的真实性。例如，对于图 1.1 中输入的科学声明，科学事实验证模型进行处理和推断，最终输出声明的标签，以及支持该标签的相关证据。声明的标签分为支持、反对或信息不足三类，而相关证据则是模型从语料库中取得的。语料库由若干篇摘要构成，每篇摘要包括标题和若干个语句。在图 1.1 中，对于给定的声明，模型输出的标签为支持，表明输入的声明是正确的。模型需要从中选择证据语句，一篇摘要中可能有多个证据语句。图 1.1 用不同颜色标识了证据语句中直接证明的部分，且模型有 98%的信心做出正确判断。

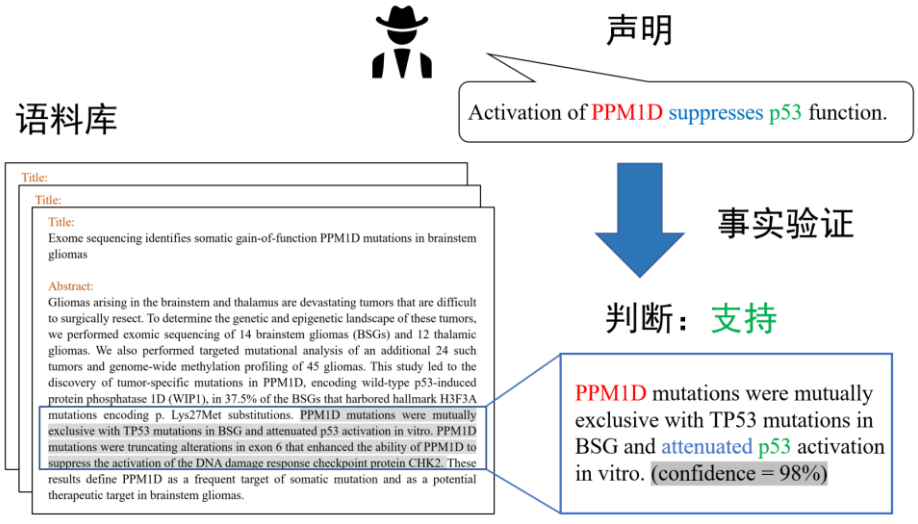

图 1.1 科学事实验证示例

在新闻报道、政治宣传和科学研究等领域中，事实验证任务已有了广泛的应用。在新闻报道中，事实验证任务可以帮助人们快速准确地判断某个新闻事件中的事实是否真实，从而保证报道的准确性和客观性；在政治宣传中，事实验证任务可以帮助人们识别并防止虚假信息和谣言的传播；在科学研究中，事实验证任务可以帮助人们鉴别并排除不可信的科学知识，从而提高科学研究的可靠性。

然而，由于科学领域缺乏大量的标记的训练数据，事实验证模型的性能受到了很大的限制。面向科学文本的事实验证对数据集和模型构建都提出了新的挑战。自然语言处理技术的最新进展是由深度神经网络模型推动的，但训练此类模型通常需要大量标记数据。在通用领域，可以通过众包的方式获得大规模的训练数据，比如政治声明在事实核查网站上很容易获得，并且可以由群众工作人员验证；但在科学领域，需要具有广泛领域知识的标注者来生成和验证科学事实。由于标记数据所需的广泛专业知识，科学领域的标记数据很难收集，且成本高昂。

最近的预训练语言模型如 BERT[1]主要在包含通用领域文本的数据集（如新闻文章和维基百科等通用领域的语料库）上进行训练和测试，而通用语料库和科学文献的数据分布有很大不同。若将预训练语言模型直接应用到基于科学文献的数据集上，则事实验证模型的性能受到了很大的限制，很难取得好的效果。综上所述，科学事实验证任务无法避免小样本场景的考验。

2 预训练技术

预训练语言模型是指在大规模语料库上进行预训练的模型，目的是使模型能够更好地理解自然语言。这些模型通常在大量文本数据上进行自监督学习，学习到语言的基础知识和统计规律，在此基础上，通过微调模型能够快速适应新的下游任务，如事实验证等。在过去的几年中，预训练已经成为自然语言处理领域的主流方法。

2.1 基于 Transformer 的双向编码表示

BERT (Bidirectional Encoder Representations from Transformers) 是一种基于 Transformer[2]架构的预训练模型。BERT 在大量未标注的文本数据上预训练，学习双向的上下文信息。具体来说，BERT 采用 Transformer 的多头自注意力机制，将文本序列作为输入，捕获序列中词层面和语句层面的关系。

在预训练阶段，BERT 使用两个预测任务进行训练：MLM (Masked Language Model) 和

NSP (Next Sentence Prediction)。MLM 是指随机掩盖输入文本中的 15%的词，让模型预测掩盖的词。在这些用于预测的词中，80%采用特殊符号（[MASK]）替换，10%采用一个任意词替换，剩余 10%情况下保持原词汇不变。

NSP 是指给定两个语句，模型需要预测这两个语句是否是连续的。具体而言，在用于 NSP 训练任务的语句 A 和语句 B 中，50%的语句 B 是语句 A 的下一句，即语句对是连续的，50%的语句 B 是从语料库随机选取的，即语句对是不连续的。

在微调阶段，BERT 在特定的下游任务上进行微调。预训练得到的参数用于初始化模型的参数，在此基础上，BERT 使用下游任务的标记数据微调参数，因此每个下游任务微调后的模型参数都不相同。由于保留了预训练参数的优势，BERT 可以很快地适应各种下游任务。

BERT 采用 WordPiece[3]方法分词，构造了大小为 30,000 的词表。输入序列的第一个词总是一个特殊的分类标记（[CLS]），它的对应输出汇总了整个序列的信息，可以用于分类任务。如果输入序列包含两个语句，BERT 使用一个特殊的标记（[SEP]）来分隔它们。此外，BERT 通过学习每个词的 Segment Embedding，也可以分辨该词属于第一个语句还是第二个语句。

例如，对于语句对分类任务，输入是语句 A 和语句 B，任务是判断这两个语句的类别标签。BERT 在输入的首部添加[CLS]符号，并将对应的输出作为文本的类别标签，BERT 用[SEP]符号来分割输入的语句对，如图 2.1 所示。

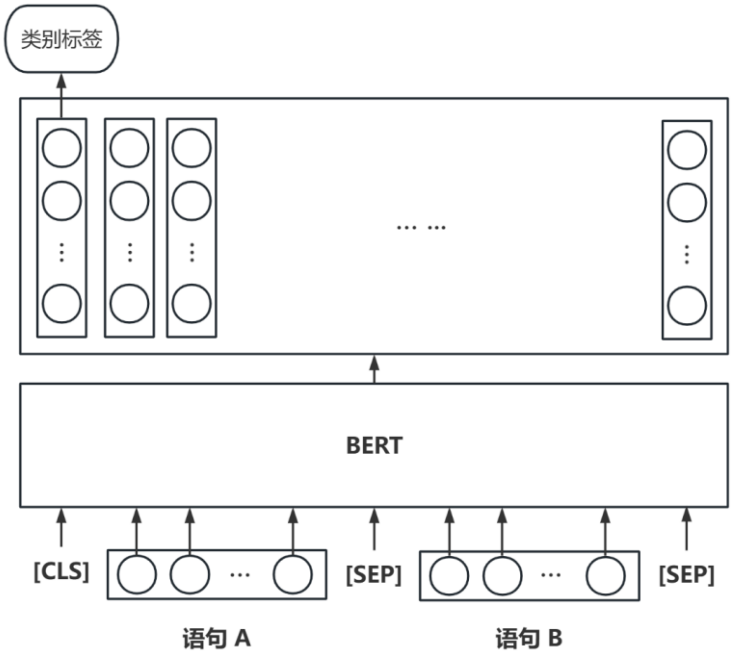

图 2.1 BERT 语句对分类任务示意图

给定一个词，它输入到 BERT 的词嵌入表示为 Token Embedding，Segment Embedding 和 Position Embedding 之和，如图 2.2 所示，输入语句 A 和语句 B 分别有两个词。

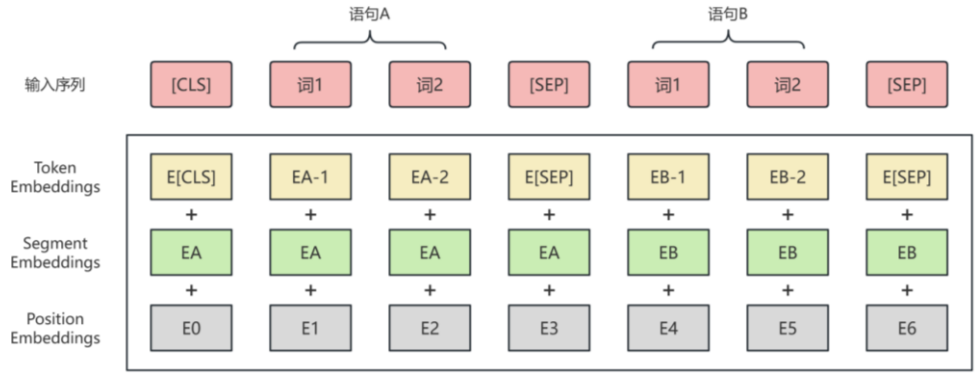

图 2.2 BERT 词嵌入示意图

BERT 将文本序列中的每个词的词嵌入向量作为模型输入，模型输出则是融合整个序列语义信息后的向量表示。BERT 内部由多层 Transformer Encoder Layer 堆叠而成，每一层 Transformer Encoder Layer 的输入和输出在形式上是完全相同的，如图 2.3 所示。

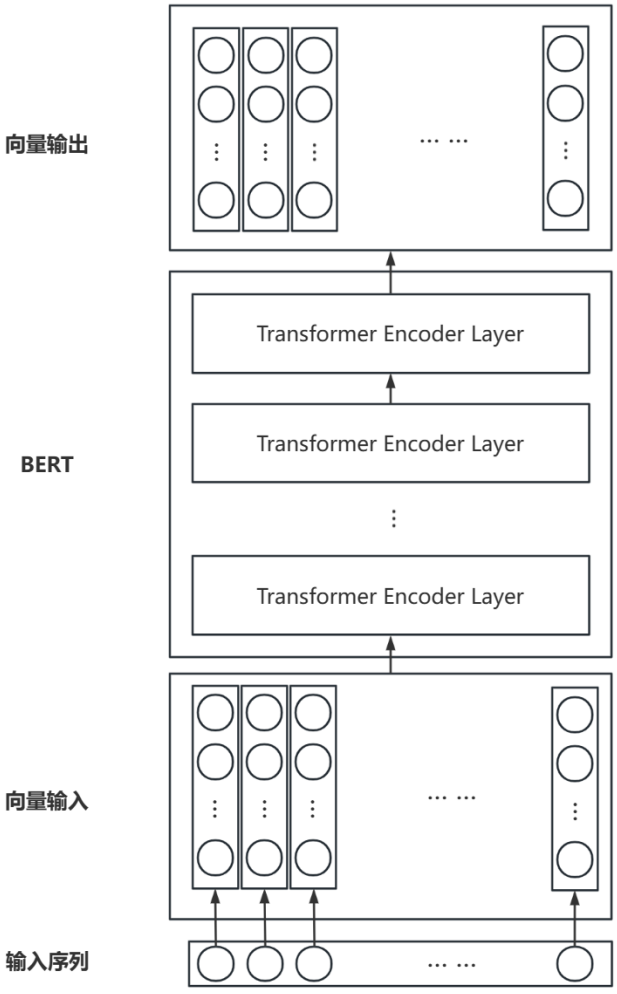

图 2.3 BERT 模型结构示意图

Transformer 是基于多头自注意力机制的编码器-解码器架构。其中，编码器由多层相同的 Transformer Encoder Layer 堆叠而成。每一层分为两个子层，第一层是多头自注意力机制，第二层是全连接的前馈神经网络，每个子层的输出都要经过残差连接和归一化（Layer Normalization），最终每一层的输出跟输入在形式上有相同的维度，如图 2.4 所示。

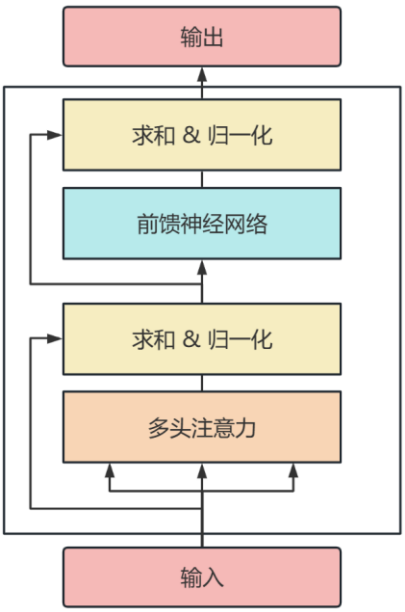

图 2.4 Transformer Encoder Layer 结构示意图

2.2 基于 BERT 的预训练模型

SCIBERT[4]专门针对科学文献的语言特点对 BERT 进行了优化，旨在提高科学文献领域的任务表现。由于科学文献领域专业词汇的复杂性和多样性，与 BERT 一样采用 WordPiece 方法，SCIBERT 在其科学语料库上分词并构造了新的词表 SCIVOCAB，其中包含了从科学文献中抽取的专业术语和常用名称。SCIVOCAB 的大小为 30K，与 BERT 的词表大小相同，但分词结果与 BERT 的词表只有 42%的重合的，这说明了科学文献与通用领域文本数据分布的差异。由于科学文献具有其独特的语言特点和专业性，在用于预训练的语料库方面，SCIBERT 采用来自 Semantic Scholar 的 1.14M 篇科学文献。SCIBERT 的预训练数据集覆盖多个学科领域的科学文献，包括 18%来自计算机科学领域的文章和 82%来自生物医学领域的文章。SCIBERT 不只使用文章的摘要，而是使用整篇文章训练，文章的平均长度为 154 个语句，整个语料库的词数为 3.17B。

RoBERTa[5] (Robustly optimized BERT approach) 在 BERT 的基础上作出了改进，详细分析了超参数调整和训练集大小等方面对模型性能的影响。RoBERTa 采用动态掩码机制，在每次将序列输入模型时，对序列随机掩码。当数据量增大时，动态掩码机制的作用更加明显。

与 BERT 不同，RoBERTa 总是使用长度为最大长度 512 的序列进行训练。NSP 是 BERT 的预训练任务之一，即给定两个语句，让模型预测这两个语句是否是连续的。RoBERTa 的实验结果表明，移除 NSP 任务后，模型在下游任务的表现结果跟 BERT 相当或者更优。RoBERTa 使用大的 Batch Size 训练，不但可以提升模型的优化速度，而且可以增强模型的性能。RoBERTa 采用更大的 BPE 词表，其中包含 50K 子词单元。

2.3 基于 Kernel 机制的图注意力网络

事实验证模块使用最先进的模型 KGAT[6] (Kernel Graph Attention Network)。KGAT 是基于 Kernel 机制的图注意力网络，具备细粒度的证据选择和推理能力。给定证据语句集，KGAT 首先构造证据图，图中的结点为声明和证据，边为完全连接。KGAT 使用两组核，在边上的核（Edge Kernel）负责汇总信息并在结点之间传播信息，在结点上的核（Node Kernel）负责选择更加匹配声明的证据。KGAT 能够整合这些信息，进行更加精确的学习和推理。

3 科学事实验证模型

3.1 三级流水线结构

事实提取和验证任务通常采用三级流水线结构：摘要检索、证据选择和事实验证。三级流水线结构如图 3.1 所示，包括每一级模块的输入和输出。对于第一级摘要检索，给定声明和摘要集合，摘要检索模型进行相关度分析，选择出前 3 个跟声明最相关的摘要。对于第二级证据选择，给定声明和已经选出的 3 个摘要，证据选择模型进行相关度分析，在 3 个摘要中选择出跟声明相关的证据，构成证据集合。对于第三级事实验证，给定声明和证据集合，事实验证模型进行声明和证据的联合推理，最终得到声明的标签。

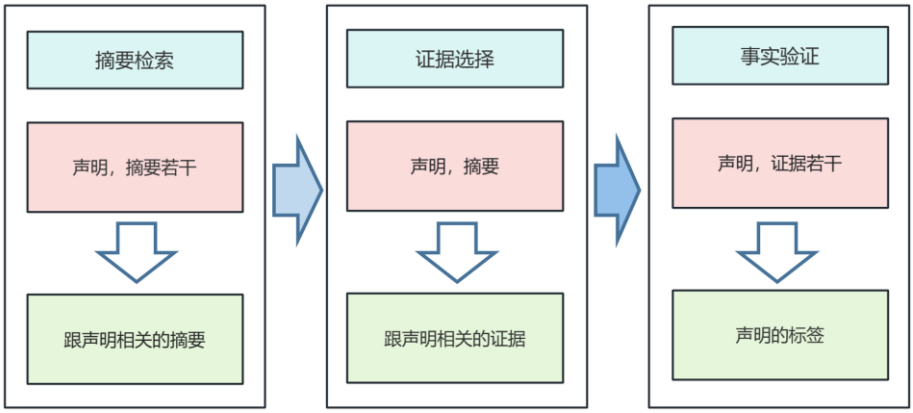

图 3.1 三级流水线结构图

3.2 继续训练方法

对于小样本场景下的事实验证任务，多阶段训练在提高性能方面具有很大的优势。为了应对事实验证任务在科学领域上的小样本场景，我们通过继续训练（Continuous Training），让预训练语言模型学习领域知识，以提高模型的事实验证能力。

继续训练是指在一个已经训练好的模型的基础上，继续用新的数据对模型进行训练，从而进一步提升模型的性能和适应性。在面对新的任务或数据时，继续训练可以使模型具有更好的迁移性和适应性。一方面，它可以通过利用先前的知识和经验，避免重新从头开始训练模型，从而节省时间和资源；另一方面，继续训练能够将特定领域内的知识转移到预训练语言模型中，以提高其理解能力。实验中采用两种继续训练方法，分别是基于证据预测的训练和基于掩码语言模型的训练。

（1）基于证据预测的训练。为了提升预训练语言模型在科学领域方面的推理能力。使用来源于科学论文的数据集 SCIFACT 进行监督学习，给定声明和证据，让 BERT 预测它们的相关标签为 1 的概率，使用此概率与真实标签的交叉熵作为损失函数L_r来优化 BERT，见公式（3.1），得到继续训练后的模型 BERT-RP（Rationale Prediction）。相比于 BERT，经过基于证据预测的训练后的 BERT-RP 具备更强的推理能力。

$$L_r(c,e) = -[y^* \cdot logp(y_r|c,e) + (1-y^*) \cdot log(1-p(y_r|c,e))] \qquad (3.1)$$

其中，$p(y_r|c,e)$为给定声明和证据语句，其相关标签y_r的预测概率。

（2）基于掩码语言模型的训练。为了使模型更好地理解特定科学领域内相关词汇的语义，用掩码对词汇进行替换，并要求模型对目标词汇进行预测。基于掩码语言模型的训练，通过在特定科学领域的语料库上自监督学习，模型能够接触到来自特定科学领域语料库的语言，从而能够学习到新的术语，并且能够更好地捕捉到这些新术语的上下文或语义。

与 BERT 一致，基于掩码语言模型的训练随机掩盖输入文本中的 15%的词，让模型预测掩盖的词是什么。在这些用于预测的词中，80%采用特殊符号（[MASK]）替换，10%采用一个任意词替换，剩余 10%情况下保持原词汇不变。基于掩码语言模型的继续训练使用与医学主题相关的 CORD-19 数据集[9]，训练预训练语言模型学习医学领域知识。

3.3 模型微调设置

在三级流水线中，不同的模块采用不同的模型微调。在摘要检索和证据选择模块，选择 SCIBERT 作为预训练模型，使用 SCIFACT 作为微调数据集。摘要检索模块的训练目标是对

于给定的声明和摘要，预测两者是否相关，即声明和摘要的相关标签为 1 的概率；证据选择模块的训练目标是对于给定的声明和证据，预测两者是否相关，即声明和证据的相关标签为 1 的概率是否大于相关标签为 0 的概率。摘要检索和证据选择模块都使用交叉熵作为二分类任务的损失函数。

而在事实验证模块，选择 RoBERTa (Large)作为预训练模型，KGAT 作为事实验证模型，使用 SCIFACT 和 FEVER 作为训练数据集，训练目标是对于给定声明和证据集合，预测声明的标签，使用交叉熵作为多分类任务的损失函数 L，见公式（3.2）：

$$L = -[y^* \cdot logP(y|G) + (1 - y^*) \cdot log(1 - P(y|G))] \quad (3.2)$$

其中，y^*表示真实的声明标签，$P(y|G)$为 KGAT 对标签的预测概率。

超参数设置方面，在证据选择模块中，总共训练 20 个 epoch。设置一批训练数据的大小为 8，输入和输出语句的最大序列长度设为 512，梯度累加步数设为 4，即每训练 4 步更新一次模型参数。微调开始后，初始学习率为 1e-5，之后随着当前的训练步数不断增大，以余弦方式衰减学习率，直到训练结束，参数更新总步数为 520 次。

摘要检索和事实验证模块的超参数设置相同，总共训练 5 个 epoch。设置一批训练数据的大小为 8，输入和输出语句的最大序列长度设为 256，梯度累加步数设为 4，即每训练 4 步更新一次模型参数。微调开始后，初始学习率为 2e-5，之后随着当前的训练步数不断增大，以线性方式衰减学习率，直到训练结束，参数更新总步数为 44815 次。证据选择和事实验证的模型微调超参数信息如表 3.1 所示。

表 3.1 模型微调设置的超参数信息

超参数名	证据选择模块	事实验证模块
Epoch	20	5
Batch 大小	8	8
最大序列长度	512	256
梯度累加步数	4	4
初始学习率	1e-5	2e-5
参数更新总步数	520	44815
学习率衰减策略	余弦	线性

实验所用服务器的配置信息如表 3.2 所示，CPU 型号为 Intel Xeon 3.00GHz，内存大小为 504.54GB，GPU 型号为 NVIDIA A800，显存大小为 80GB。

表 3.2 服务器配置信息

服务器组件	组件信息
CPU	Intel Xeon 3.00GHz
内存	504.54GB
GPU	NVIDIA A800
显存	80GB

3.4 数据集与评估方法

在数据集准备方面，选用三个数据集，分别是 SCIFACT[7]，FEVER[8]和 CORD-19[9]。进行数据预处理后，分别用于不同模块的训练。模型测试使用 SCIFACT 数据集。

SCIFACT 用于训练摘要检索、证据选择模块和事实验证模块。SCIFACT 是科学声明与证据组成的数据集，包括从基础科学到临床医学跨领域的高质量文章，由 1409 个标注过的科学声明和 5183 篇科学文章组成。所有声明都被分类为支持、反对或信息不足。其中，训练集、验证集和测试集分别包含 809 个、300 个和 300 个声明。

FEVER 用于训练事实验证模块。FEVER 是声明与证据组成的数据集，由 185455 个标注过的声明和 5416537 篇维基百科文档组成。SCIFACT 来自科学语料，FEVER 来自通用语料，SCIFACT 和 FEVER 声明数之比为 1:132，文章数之比为 1:1045，说明科学领域缺少大量标记过的训练数据。

CORD-19 用于针对医学主题的继续训练，训练预训练语言模型学习医学领域知识。CORD-19 是新型冠状病毒肺炎和相关冠状病毒文章组成的数据集，在其语料库中，大约有 40%的文章是关于新型冠状病毒肺炎的。

在模型评估方面，与 SCIFACT 一致，评估模型性能的指标采用精确率、召回率和F_1值。这些评估指标由 FEVER 评分启发而来，分别在摘要层面和语句层面，给模型所选证据的正确性打分。

3.5 实验结果与分析

基准线模型主要来自于 Wadden 等人的模型[7]。基准线模型在摘要检索模块采用 TF-IDF 算法，而在证据选择和事实验证模块采用 RoBERTa (Large)模型。其中，证据选择模块同样使用 SCIFACT 数据集进行微调，而事实验证模块使用 FEVER 和 SCIFACT 进行训练。

实验采用的验证集来自 SCIFACT 数据集，本研究采用的科学事实验证模型和基准线模

型在验证集上的性能对比如表 3.3 所示。实验结果表明，在验证集上，本研究采用的科学事实验证模型在证据层面和摘要层面都要优于基准线模型，其中本研究的模型在证据层面和摘要层面的精确率分别达到了 76.42%和 82.46%，精确率和F_1值明显优于基准线模型。

表 3.3 验证集实验结果

	证据层面			摘要层面		
模型	精确率	召回率	F_1值	精确率	召回率	F_1值
RoBERTa (Large)	46.51	38.25	41.98	46.6	46.4	46.5
本研究模型	76.42	44.26	56.06	82.46	44.98	58.20

实验采用的测试集来自 SCIFACT 数据集，本研究采用的科学事实验证模型和基准线模型在测试集上的性能对比如表 3.4 所示。实验结果表明，在测试集上，本研究采用的科学事实验证模型在证据层面和摘要层面都要优于基准线模型，其中本研究的模型在证据层面和摘要层面的精确率分别达到了 60.59%和 76.98%，精确率和F_1值明显优于基准线模型。

表 3.4 测试集实验结果

	证据层面			摘要层面		
模型	精确率	召回率	F_1值	精确率	召回率	F_1值
RoBERTa (Large)	38.6	40.5	39.5	46.6	46.4	46.5
本研究模型	60.59	44.05	51.02	76.98	48.20	59.28

通过分析实验结果，无论是在语句层面还是在摘要层面，本研究采用的科学事实验证模型在精确率方面都比基准线有 30%左右的显著提升，说明本研究模型在面向小样本的事实验证方面，具备提供高质量且可信的结果的能力。模型性能的全面提升，证明了事实验证模块的 KGAT 模型的多证据细粒度的推理能力，也证明了继续训练方法使模型能够识别并理解领域内术语的有效表现。

4 结论与展望

事实验证任务旨在使用可信的语料库自动验证语句的正确性。近年来虚假科学信息传播迅速，危害公共安全，而审查人员有限，社会迫切需要自动化工具帮助核查，事实验证任务可以帮助人们快速鉴别不可信的科学信息。然而，科学事实验证任务面临小样本场景的考验。科学事实数据的生成和验证，需要数据标注者具备广泛的专业知识，导致科学领域缺乏大量标记的训练数据。在机器学习中，小样本学习任务容易面临过拟合，限制了事实验证模型的

性能。若直接使用少量样本进行训练，很难建立高质量的科学事实验证模型。预训练语言模型虽然引入了额外的数据集，但是主要在包含通用领域文本的数据集上进行训练，而通用语料库和科学文献的数据分布存在很大差异，因此预训练语言模型在科学事实验证任务上的性能受到了限制。

本研究采用事实提取和验证三级流水线，即摘要检索、证据选择和事实验证，并对预训练语言模型进行微调，使模型能够完成事实验证任务。本研究使用继续训练方法，训练预训练语言模型学习领域知识，使模型能够完成面向小样本的事实验证任务。本研究使用SCIFACT 数据集，一个由专家标注的科学事实验证数据集，评估测试模型性能。结果表明，模型在摘要层面和证据层面的精确率分别提升到 77%和 61%，跟基准线相比有 30%左右的显著提升，说明本研究采用的模型在面向小样本的事实验证方面具备提供高质量且可信的结果的能力。

在未来，需要继续对科学事实验证系统的适用范围和验证质量等性能方面继续做出改进。在模型算法方面，需要继续研究改进预训练语言模型和事实验证模型的结构，以提高模型在事实验证任务上的性能；研究继续训练方法，以更好地提高模型在小样本任务上的表现；研究提高大模型的推断速度，以便更快响应人们的需求。在数据集处理方面，尝试扩展语料库，将来自其他科学领域的数据用于训练模型，探索其他科学领域的数据对科学事实验证模型效果的影响。

希望通过预训练模型在科学事实验证任务中的应用研究探索，推动事实验证任务的发展，为解决实际应用中的问题提供更可靠的解决方案。

参考文献

[1] Devlin J, Chang M, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of NAACL. 2019: 4171–4186.

[2] Ashish V, Noam S, Niki P, et al. Attention is all you need[C]// Proceedings of NeurIPS. 2017: 6000–6010.

[3] Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv preprint arXiv:1609.08144, 2016.

[4] Beltagy I, Lo K, Cohan A. SCIBERT: A pretrained language model for scientific text[C]// Proceedings of EMNLP. 2019: 3615–3620.

[5] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[6] Liu Z, Xiong C, Sun M, et al. Fine-grained fact verification with kernel graph attention network[C]// Proceedings of ACL. 2020: 7342–7351.

[7] Wadden D, Lo K, Wang L, et al. Fact or fiction: Verifying scientific claims[C]// Proceedings of EMNLP. 2020: 7534–7550.

[8] Thorne J, Vlachos A, Christodoulopoulos C, et al. FEVER: a large-scale dataset for fact extraction and VERification[C]// Proceedings of NAACL. 2018: 809–819.

[9] Wang L L, Lo K, Chandrasekhar Y, et al. Cord-19: The covid-19 open research dataset[J]. arXiv preprint arXiv: 2004.10706, 2020.