开题报告

一、 选题依据与价值

1.1 知识图谱文本生成任务的研究背景

文本生成，也称为自然语言生成（Natural Language Generation，NLG），是自然语言处理（Natural Language Processing，NLP）中最重要的子领域之一。它旨在从各种形式的输入数据（包括文本、图像、表格和知识图谱[1]）中生成合理且易于理解的自然语言文本。文本生成的基本目标是学习从输入数据到输出文本的映射函数。根据输入数据属性和类型的不同，又可细分为多种类别，数据到文本的生成即输入结构化的表格或知识图谱三元组生成有关结构化数据的描述性文本。“知识图谱（Knowledge Graph）”的概念由 Google 公司在 2012 年正式提出，指用于提升搜索引擎性能的知识库，通过对知识进行这种更加有序、有规律的组织，从而使用户更加智能、快速地获取和管理信息。知识图谱支撑起很多行业发展，如信息检索、问答系统、推荐系统、金融风控等等具体应用，在科研学术中也引发很多学者的研究，针对结构化的知识图谱数据实现文本生成（KG-To-Text Generation，KG2Text Generation）也是一个重要研究方向。

从概念上讲，知识图谱就是通过“关系”或“属性值”将“实体”节点联结起来，直观的展现事物概念之间的关系和属性，构成庞大的语义网络图结构，实现对现实世界的自然语言描述。RDF 三元组是知识图谱的基本组成部分，它是由实体关系或概念属性构成的，其基本展示形式可以形式化的描述为{头实体，关系，尾实体}或者{概念，属性，属性值}形式，即形成三元组的表达方式。RDF 三元组作为结构化知识的基本表示单位，能够更好的展现结构化知识的细粒度组成结构和内容，并且能够嵌入图信息展现为有向图的形式（如图 1-1），更加直观的展示了细粒度节点间的关系。因此结构化的数据到文本的生成任务就是选取以三元组$<S,R,O>$为输入，其中 S 表示头实体、R 表示实体间的关系、O 表示尾实体，生成一句以流畅性、丰富性、准确性为特性的自然语言文本序列 $y=<y_1,y_2,...y_n>$。如图 1-1 示例中，给定三元组<Barack Obama, president, United States>，生成符合三元组事实并且流畅的自然语言文本“Barack Obama is the 44th President of the United States.”。

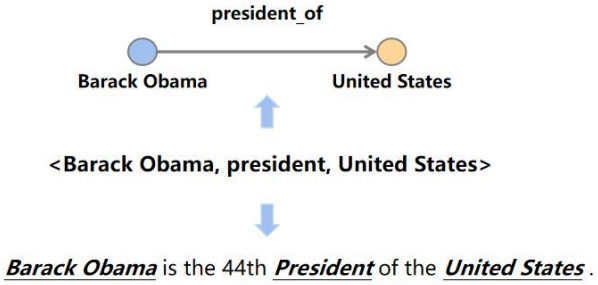

图 1-1 知识图谱文本生成的样例

基于结构化知识的文本生成技术是知识图谱在人工智能领域应用的一个重要方面，从数据到文本的生成技术富有研究意义，同时极具应用前景，我们期待有一天机器能够像人类一样具有写作能力，自动撰写出高质量自然语言文本，从而创新现有的文本创作方式。例如实现从数据到新闻的生成，从而极大改变编辑、记者的工作方式，实现媒体、报社行业的变革创新，该项技术甚至可以用来帮助学者进行学术论文撰写，进而改变科研创作模式。

1.2 知识图谱文本生成任务面临的挑战

如今随着计算机的计算能力和建模能力的大大提升，从知识图谱生成上下文流畅、语法正确的自然语句成为可能，但是能够生成与主题内容相关并且表现出整体的连贯性的自然语言文本又是一个当今学者面临的挑战。早期方法通常采用统计语言模型来建模单词的条件概率[2]，这种统计方法可能会受到数据稀疏问题的影响。随着深度学习技术的出现和发展，神经网络模型已经成为了文本生成的主流技术，并在生成自然语言文本方面取得了巨大的进步。深度神经网络生成模型通常采用序列到序列框架[3]，基于编码器-解码器方案生成目标文本，编码器首先将输入序列映射为固定大小的低维向量（称为输入嵌入），然后解码器根据输入嵌入进行解码生成目标文本。虽然文本生成的深度神经网络模型取得了成功，但仍然存在瓶颈，目前主要存在以下问题和挑战：

（1）知识图谱在文本生成中的结构化语义缺失问题

知识图谱中的图结构语义信息通常具有复杂的拓扑结构和多层次的关系。传统的文本生成模型难以直接处理这种复杂的图结构信息，因为它们主要关注文本序列的生成而忽略了图结构中实体之间的关联。这导致文本生成模型无法准确地捕捉到图结构中的语义关系和上下文信息。知识图谱中存在着大量的实体和关系，其中部分信息可能是缺失或不完整的。这些缺失的信息会导致文本生成模型在编码图结构时遇到困难，因为模型无法准确地推断实体之间的语义关系和连接，这会影响生成结果的准确性和一致性。此外，由于文本生成任务的早期神经模型仍然相对较浅。因此，这些模型难以对上下文和词义之间的关系进行建模，从而难以从上下文内容中更好理解和生成文本。因此针对知识图谱特有的的结构化信息特点来进行预训练语言模型的知识增强技术[4]成为了保证输出文本准确性的关键因素。

（2）任务级别的知识图谱文本生成的质量问题

常见神经网络文本生成模型所输出的文本虽然看似流畅，但往往存在多种逻辑错误，比如混淆了不同信息的输出位置、无中生有、缺乏语法正确性和逻辑性等等[5]，这些问题会影响文本的可读性和可理解性，进而影响知识图谱文本生成的实际应用效果，这些问题在进行大规模知识图谱的长文本生成时显得更为严重。除此之外，预训练语言模型通常在通用语料库中进行预训练，难以在特定的任务级别控制生成文本的多粒度属性，如知识图谱的任务特点、主题、领域信息等，具体来说主要表现为生成的文本与用户需求或目标不匹

配，例如用户想要生成关于某个人物实体或特定领域知识图谱的子图描述，但是生成的文本中忽略了这个任务的领域相关主题，仅仅保证了生成文本的流畅性，并没有准确地涉及这些任务级别的语义结构的细粒度内容，这样的主题偏移现象显然与用户的要求相违背。这些问题都是任务级别的文本生成的质量问题导致的。

二、 国内外研究现状与发展状态

面向 KG2Text 的文本生成任务研究一直是研究者们关注的焦点，传统系统主要建立在基于规则和模板的算法上[6]。随着深度学习的快速进展，研究人员已经将注意力转移到神经网络结构的生成模型上，尤其是编码-解码（Encoder-Decoder）框架的流行，使得 Seq2Seq（Sequence to Sequence）模式成为了端到端的 KG2Text 生成技术中的主流方法。尽管神经模型在 KG2Text 生成方面取得了成功，但大规模数据集的可用性仍然是主要的性能瓶颈。因此近年来，预训练语言模型的范式正在蓬勃发展。近两年研究人员发现，扩展通过对模型大小或者数据集大小进行扩展能够有效地提高下游任务的模型性能。许多研究已经通过训练更大的语言模型[7]来探索性能极限，这些大规模的预训练语言模型（Large Language Model，LLM）在解决一系列复杂的 NLP 任务时展现出惊人的能力。

2.1 基于规则和模板的知识图谱文本生成方法

传统上基于规则和模板的数据到文本的生成方法需要解决两个主要问题[8]：内容规划和句法实现，即先规划出符合语义的文本结构，然后根据该结构进行最终文本的实现。Reiter 等人[9]又提出了更加细化的数据到文本生成的范式结构，可以归纳为一个四阶段的管道式（pipeline）结构：数据分析、数据解释、句子规划以及句法实现。

对于文本规划阶段，从生成决策器的出发，Paiva 等人[10]通过将大量文本的自由生成过程中的内部决策与输出结果的文本风格特征相关联，利用文本相关性来控制生成器，这与以前大多数基于经验的生成方法形成对比，提供了一种更有效的语言转换模型。但是上述方法都需要一个现有的手工句子规划器，因此 Dušek 等人[11]提出了基于 A*搜索的感知排序器，不需要在训练数据中进行细粒度对齐，而是使用深度依赖语法进行句子计划，生成器将对齐学习纳入句子规划器训练，并使用深度语法树和基于规则的表面实现步骤，以确保输出的语法正确性，与之前的方法不同的是该生成器不需要手工制作的基本句子规划器。

对于句法实现阶段，Dethlefs 等人[12]将句法实现作为序列标记任务，并将条件随机场（CRF）的使用与语义树相结合。提出了扩展的语境概念，其他句法实现方式容易受到局部特征的限制，CRF 能够考虑到整体的话语语境，适用于在丰富语义的上下文中进行建模，并结合语义树来获取丰富的语言信息，实现了更自然和低重复的句法实现效果。这种组合能够在多个话语中跟踪句法、语义和词汇特征之间的依赖关系。

还有其他工作将文本规划和句法实现结合到一个步骤中，不需要手工制作基础模块。Angeli 等人[13]提出了一个简单健壮的生成系统，它在一个统一的、独立于领域的框架中执

行内容选择和文本实现，使用一个端到端的过程，并将其中的子任务分解为一系列局部决策和分层排列，对每个决策都进行了判别性训练，从数据库记录中生成文本，逐步选择数据库记录、字段和相应的文本实现来描述它们，实现了一个具有领域通用性的生成系统。Dušek 等人[14]提出了一个基于序列到序列方法的自然语言生成器，结合集束搜索和 n-best 列表重新排序器来控制输出文本中的相关信息，并且通过实验比较了通过深度语法树的两步生成方法与直接生成字符串的比较，实验结果表明，直接生成文本方法更好，并且基于 n-gram 的分数更高。

基于规则和模板的方法能够更好地遵循人类指定的文本特性进行文本生成，但是也意味着需要更多的人工参与，从而失去了文本生成的流畅度和多样性，并且这些早期的工作通常集中在较小的特定领域的数据集上，使得文本生成模型的实际应用存在局限。

2.2 基于神经网络模型的知识图谱文本生成方法

2.2.1 端到端的神经网络模型

神经网络的重新兴起的部分原因是硬件的进步，可以支持资源密集型问题的学习，并且神经网络利用反向传播来学习特征表示，具有密集的、低维的和分布式的特点，使得它们非常适合捕捉语法和语义的特征[15]。编码器-解码器框架的流行，使得序列到序列（Seq2Seq）模式成为了端到端的文本生成技术中的主流方法，即将源语言中的可变长度输入序列映射到可变长度序列的目标语言中。基于注意力机制的模型结构的提出加速了编码器-解码器结构的发展，即在解码期间预测输出的某些部分时，对输入编码的部分有选择的进行关注。Liu 等人[16]提出了一种结构感知的 seq2seq 结构，该结构由领域级别的门控编码器和双注意力描述生成器组成。在编码阶段，我们通过字段门及其相应的字段值更新 LSTM 单元的单元内存，以便将字段信息合并到表示中。在解码阶段，提出了包含词级注意力和字段级注意力的双注意力机制，对生成的描述与表之间的语义关联进行建模。除了注意力机制，复制机制也是端到端的 KG2Text 生成任务的重要技术之一，复制机制首先被引入到循环神经网络中，以产生完全由输入序列元素组成的输出序列，Gu 等人[17]将复制引入到基于神经网络的 Seq2Seq 学习中，并提出了一个具有编码器-解码器结构的新模型 COPYNET，利用复制机制实现了输入序列中的某些片段被选择性地复制到输出序列中，不仅能够正常生成单词，而且能够复制输入序列的适当片段。See 等人[18]基于复制机制提出了一种指针生成模型，以两种正交的方式增强了标准的序列到序列的注意力模型。

由于知识图谱的三元组之间并不存在时序关系，而更多的是图的结构和节点之间的边，因而基于图结构的模型被提出，Marcheggiani 等人[19]提出了一种基于图卷积神经网络（Graph Convolutional Network，GCN）的编码器更好地利用了输入的三元组之间的结构信息，允许将图形数据显式编码到神经网络中。同时考虑到图结构中相邻节点的影响大小，解决由于信息过长导致信息丢失的问题，放大节点中重要部分的影响，Koncel-Kedziorski 等人[20]使用了图注意力网络（Graph Attention Network，GAT）来实现图到文本的生成，

由于 GCN、GAT 这类模型仅考虑图谱中已出现相邻节点的信息而忽略全局信息，所以作者添加了全局结点使得模型能够利用更为全面的全局信息，充分利用这些知识图的关系结构而不强加线性化或分层约束。

2.2.2 基于神经网络和管道模型相结合的方法

许多研究人员试图在 KG2Text 生成模型中对内容规划进行明确的建模，将管道式的生成方法与神经网络模型的优点相结合，进一步提升 KG2Text 的生成质量。由于神经网络的方法通常不会对文本生成过程中的内容顺序进行建模，但是参考人类的书写风格，在书写前需要对内容的顺序进行规划，因此 Sha 等人[21]提出了一个顺序规划的文本生成模型，其中文本的顺序信息通过基于链接的注意力显式捕获，然后，自适应门将基于链接的注意与传统的基于内容的注意结合起来，实现文本内容的规划和输出，并且使用了复制机制来解决生僻字的问题。Puduppully 等人[22]提出了两阶段的管道式生成方法，该管道使用基于注意力机制的指针网络从结构化数据中选择和规划重要信息，然后使用另一种编码器-解码器模型生成最终文本。首先生成一个内容计划，突出显示应该提到哪些信息以及以何种顺序进行书写，然后在考虑内容计划的同时生成文档。Chen 等人[23]提出了一个具有动态内容规划的神经数据到文本生成模型，在计划应该提到哪些数据时考虑先前生成的文本，从给定的结构化数据中利用先前生成的文本来动态地选择适当的内容。模型的动态内容规划机制易于与编码器-解码器框架集成，并具有自己的目标函数。

2.3 基于预训练语言模型的知识图谱文本生成方法

2.3.1 生成式语言模型

预训练语言模型使用大量未标记的文本数据进行预训练，可以在下游生成任务中进行微调。主要思路是首先在大规模语料库中预训练模型，然后在各种下游任务中对这些模型进行微调，以获得最先进的结果。由于 Transformer 模型[24]取得的巨大成就，几乎所有的预训练语言模型都采用了 Transformer 的骨干。

标准 Transformer 使用编码器-解码器架构，该架构由两个 Transformer 块堆栈组成。编码器以输入序列为源输入，而解码器则基于编码器-解码器的自注意机制生成输出序列。基于上述架构，T5[25]和 BART[26]等模型的提出显著提高了文本生成的质量。Raffel 等人[25]在迁移学习的启发下，引入一个将所有基于文本的语言问题转换为文本到文本格式的统一框架，基本思想是将每个文本处理问题视为“文本到文本”问题，即将文本作为输入并生成新文本作为输出，同时通过扩展模型和数据集来探索 NLP 迁移学习的前景和局限性。Lewis 等人[26]提出了用于序列到序列预训练模型的去噪自编码器，训练过程中首先用任意的噪声函数破坏文本，使用了新颖的填充方案来学习一个模型对原始文本进行重建，BART 在文本生成任务和文本理解任务上都能获得不错的效果。

基于 Transformer 的解码器架构，GPT 系列模型被提出，它们使用单向的自注意力掩码模型，每个 token 只能关注前一个 token 信息。GPT-1[27]探索了一种结合无监督预训练和监

督微调的半监督方法来完成语言理解任务，学习了一种通用的表征，这种表征可以在很少的为微调情况下转移到广泛的下游任务中。GPT-2[28]探索了语言模型对零样本生成任务的迁移能力，突出了数据体量的重要性，演示了语言模型可以在零样本设置中执行下游任务的潜力，且不进行任何参数或架构修改。GPT-3[29]表明海量模型参数可以显着改善下游生成任务，扩展语言模型的参数量可以极大地提高任务不可知的、少样本情况下的模型性能，有时甚至可以与之前最先进的微调方法相竞争。

2.3.2 优化方法

由于预训练语言模型已经在多个 NLP 领域中取得了非常显著的成功，因此学者们开始利用预训练语言模型和迁移学习来解决 KG2Text 的生成问题。目前有两种主流的迁移方法，对整个预训练语言模型进行微调[30]或使用前缀学习的轻量级微调方式[31]。

对于整个预训练语言模型进行微调方面，Chen 等人[30]提出了一种基于知识的预训练方法 KGPT，这是一种远程监督学习算法，利用大规模未标记的网络文本来预训练数据到文本模型。预先训练的模型可以在各种数据到文本生成任务上进行微调，以生成特定于任务的文本。Deng 等人[32]研究了在少样本场景下的具有逻辑形式的结构化数据到文本的生成，提出了一个统一的逻辑知识条件文本生成框架 LOGEN，利用容易获得的领域内语料库来进行自我训练，使用少量种子数据进行训练以生成伪逻辑形式，并逐步扩大生成逻辑形式的训练语料库。关于前缀调优[31]，它为利用预训练的语言模型提供了一种有效的训练范式。具体来说，它冻结了预训练语言模型的参数，只调整了一组预先准备的可训练连续向量到语言模型，该模型由更少的参数组成。在此基础上，Chen 等[33]提出在提示学习和适配学习之间架起桥梁，提出了 inducer-tuning 来扩展前缀调优的优势，将提示符（包括硬提示和软提示）视为内核方法中的“诱导变量”，将预训练的语言模型中的注意力模块与核估计器作类比，解释并激发了前缀调优的潜在机制。

2.4 研究现状总结

分析当前研究成果，如表 2-1 所示：

表 2-1 研究现状总结

研究内容		方法	方法内容	年份
基于规则和模板的方法	文本规划阶段	Paiva 等人[10]	捕获生成文本的表面语言特征，利用文本相关性来控制生成器	2005
		Dušek 等人[11]	使用深度依赖语法树并提出基于 A*搜索的感知排序器	2015
	句法实现阶段	Dethlefs 等人[12]	将条件随机场与语义树相结合实现语境的扩展	2013
	文本规划和句法实现联合实现	Angeli 等人[13]	独立于领域的框架中执行内容选择和文本实现	2010
		Dušek 等人[14]	提出了一个基于序列到序列方法的自然语言生成器	2016

基于神经网络模型的方法	端到端的模型	Liu 等人[16]	结构感知的端到端模型由领域门控编码器和双注意力描述生成器组成	2018
		COPYNET[17]	将复制机制应用于具有编码器-解码器结构的新模型	2016
		pointer-generator[18]	利用混合的指针生成器网络和覆盖追踪的方式生成文本	2017
		Marcheggiani 等人[19]	基于图卷积神经网络的编码器显式编码结构信息	2018
		GraphWriter[20]	使用了图注意力网络并添加全局结点来实现图到文本的生成	2019
	神经网络和管道模型相结合	Sha 等人[21]	提出了一种表到文本生成的顺序规划方法	2018
		Puduppully 等人[22]	提出了两阶段的管道式生成方法并保留端到端的训练	2019
		NDP[23]	具有动态内容规划的神经数据到文本生成模型	2021
基于预训练语言模型的文本生成方法	基于 Transformer 的编码器-解码器架构的 PLM	T5[25]	将所有基于文本的语言问题转换为文本到文本格式的统一框架中	2020
		BART[26]	实现了序列到序列的去噪自编码器预训练语言模型	2020
	基于 Transformer 的解码器架构的 PLM	GPT-1[27]	结合无监督预训练和监督微调的半监督方法来完成语言理解任务	2018
		GPT-2[28]	探索了语言模型对零样本生成任务的迁移能力	2019
		GPT-3[29]	实现了海量模型参数可以显着改善下游生成任务	2020
	基于预训练语言模型的微调	KGPT[30]	利用大规模未标记的网络文本来预训练数据到文本模型	2020
		LOGEN[32]	利用自我训练和基于内容和结构一致性的伪逻辑形式样本	2023
	基于前缀调优	inducer-tuning[33]	将提示符（包括硬提示和软提示）视为内核方法中的“诱导变量”	2022

三、 研究目标与研究内容

3.1 研究目标

综合分析 KG2Text 生成任务的发展历程和研究现状，并针对现有基于预训练语言模型的知识图谱生成式方法存在的问题，本课题的主要研究目标是结合知识图谱输入的结构化信息，构建出能更好编码并输出知识图谱结构信息的语言模型，同时保证输出文本在特定任务级别的生成文本质量的准确率，对生成的自然语言文本进行评估。因此本课题的研究目标主要分为三部分，首先针对知识图谱在文本生成中的结构化语义缺失问题，利用去噪

自编码器训练模式，学习跨范围的实体之间的结构化语义信息，提升上下文理解能力。然后针对任务级别的知识图谱文本生成的质量问题，利用前缀调优技术将大规模的预训练语言模型更好地应用于知识图谱文本生成的任务中，更好地引导预训练语言模型扩展到更细粒度的任务和属性级别的知识图谱文本生成中，为特定领域的文本生成任务提供更加可靠和高效的解决方案。最后综合评估本课题提出的模型在知识图谱文本生成中的质量并，且在人物报告生成场景中测试本课题提出模型的有效性，通过案例分析来证明本课题模型的应用价值。

3.2 研究内容

基于上述研究目标并结合问题定义，根据知识图谱到文本生成的应用场景和现实需求，本课题的具体研究内容主要包括以下几个部分：

（1）针对知识图谱在文本生成中的结构化语义缺失问题，提出面向知识图谱文本生成任务的预训练语言模型，使其更好地理解和编码跨范围的实体之间的结构化语义信息；

（2）针对任务级别的知识图谱文本生成的质量问题，提出面向知识图谱文本生成任务的预训练语言模型前缀微调技术，提升模型在任务级别的语义感知能力，在任务级别控制文本的多样性、连贯性和准确性的同时，保证生成文本满足任务所需属性要求；

（3）结合组织机构知识图谱，以人物报告生成场景为例进行案例分析，综合评估生成文本的质量和模型的有效性。

3.2.1 基于去噪自编码的知识图谱文本生成预训练语言模型

针对知识图谱在文本生成中的结构化语义缺失问题，由于预训练语言模型在各种自然语言处理(NLP)任务中都表现出出色的语言理解能力，本课题借助预训练语言模型在 NLP 多领域中的应用成效，研究去噪自编码器的预训练模式，注于序列到序列(Seq2Seq)的训练目标，从而可以在给定噪声损坏的情况下重建原始文本，通过在大规模文本数据上进行预训练，使模型学习跨范围的实体之间丰富的结构化语义信息，提升语义结构感知能力和上下文理解能力。过去的研究主要在模型编码器的输入序列中使用噪声损坏来重建文本，本课题根据生成式语言模型的特性在解码器中研究给定噪声损坏实现文本还原的能力，利用外部语料实现知识增强的序列到序列预训练，使得语言模型能够完全理解生成语句的含义并进一步进行推理输出。因此针对带有掩码标记的输入结构，本任务旨在结合跨范围的实体之间的结构化语义信息和文本事实对输入文本的完整信息进行自编码重构，捕获输入和已输出内容的上下文语义，保证了输出语句的实体对齐和知识三元组检索，弥补了知识图谱在文本生成中的结构化语义缺失的问题，为下游任务奠定基础。

3.2.2 面向任务级知识图谱文本生成的前缀微调技术

在上一任务的基础之上，将预训练语言模型适应于下游应用场景，进一步解决任务级别的知识图谱生成文本质量问题。“prefix-tuning”是一种用于预训练语言模型的轻量级微调技术，它将大规模的语言模型参数冻结，任务转为优化一个小的、任务导向的连续向

量，通过采用模块化控制前缀进行参数的训练。利用前缀调优技术将大规模的预训练语言模型更好地应用于知识图谱文本生成的任务中，通过静态前缀和动态前缀相结合的方式提供关于任务级别的语义结构信息和属性级别信息分别对输出文本的生成效果和内容进行控制，任务级别的语义结构信息主要包括领域任务的特性，例如生成一段人物报告描述，而属性级别的信息可以包括知识图谱到文本生成的三元组集合的领域信息，或者它可以指定所需输出文本的简化目标长度或者惩罚机制等来控制文本的多样性、连贯性和准确性。任务级别的输入信息能够在模型的每一层集成特定于任务的静态提示，同时还允许特定于输出文本的属性控制信息作为动态的指导信号。使用前缀提示可以更好地引导预训练语言模型扩展到更细粒度的目标质量的文本生成中。本课题研究在固定整个预训练语言模型的参数的情况下，只对少量的前缀信息进行调优的模型效果，为模型在多种类型的下游任务中的轻量级调优技术奠定基础。

3.2.3 面向知识图谱文本生成任务的案例分析

综合上述的研究，生成文本质量的评估也是非常关键的研究内容，文本的生成质量直观地展现出语言模型的性能，文本生成领域对于生成文本质量的评估并没有统一的标准，研究对于生成文本质量更加公平有效的评价方式，可以避免耗费高昂的人工评估的成本和大量的时间成本。同时为了进一步验证提出模型的有效性和可迁移性，本课题在人物报告生成的场景中进行实验，输入人物信息相关的知识子图，即多组与某个任务节点相关联的三元组输入信息，输出该人物相关的知识子图的文本信息介绍，以了解该人物的画像信息细节内容，从而在真实案例中分析语言模型生成文本的质量。

四、 实施方案及可行性分析

本课题总体方案设计如图 4-1 所示，下面将对主要模块的实施方案进行详细介绍。

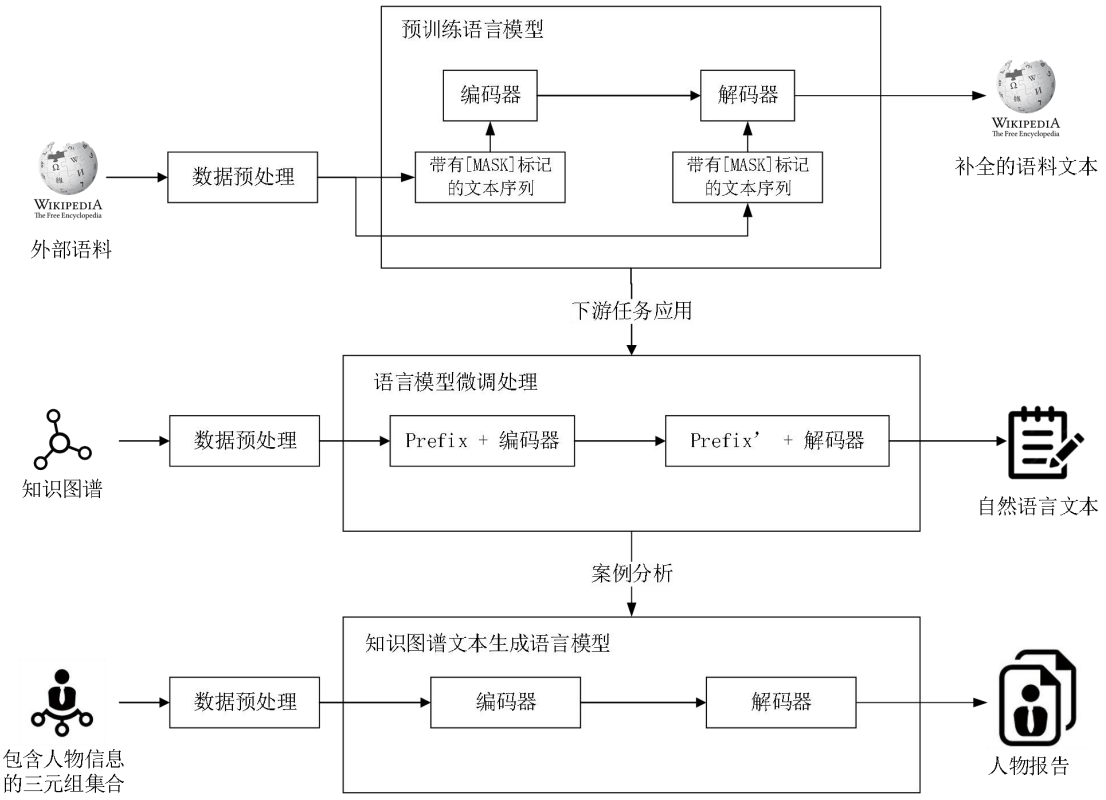

图 4-1 知识图谱文本生成方案总体结构图

4.1 总体任务定义

一般来说，一个长度为 n 的文本序列可以表示为序列 $y =< y_1, y_2, \ldots y_n >$，其中$y_1$是从词汇库 V 中抽取出来的。文本生成的任务旨在生成流畅且正确的自然语言文本。在大多数情况下，文本生成任务可以根据输入的数据，例如文本、图像、表格和知识库等等，生成满足某些语言特性的语言文本，例如流畅性、自然性和连贯性。将输出文本所需的特征定义为特征集 K。因此，文本生成的任务可以定义为：

$$y = f_K(x) \tag{4-1}$$

其中f_K表示根据输入 x 生成特定文本属性 K 的文本序列 y 的生成函数。本课题研究的输入 x 由单个或多个形如$< S, R, O >$的三元组构成，其中 S 表示头实体、R 表示实体间的关系、O 表示尾实体，输出序列 y 则是符合输入三元组事实的一段自然语言文本。

4.2 数据集准备

本课题实验中使用的数据集包括 E2E 数据集[34]和 WebNLG 数据集[35]。

4.2.1 E2E 数据集

E2E 数据集是一个广泛用于自然语言生成任务的数据集，其目标是评估端到端

（End-to-End）生成模型的能力。该数据集主要用于评估基于神经网络的模型在生成自然语言描述的任务上的性能。E2E 数据集的主要任务是根据餐馆的属性生成一些列餐馆描述文本。输入可以是一些特定的餐厅信息，例如餐厅的位置、菜单、评分等。模型需要根据这些输入生成相应的自然语言描述，如餐厅的介绍、推荐菜品等。它由 42,061,547 和 630 个用于训练、验证和测试集的实例组成。平均输入长度为 28.5，平均输出长度为 27.8。此数据集的数据以键值格式存储，结构化数据可以被看作是<属性，值>对，因此在处理过程中将餐馆的名称作为头实体，与属性、值构成三元组结构，即<餐馆名称，属性，值>，然后再输入到语言模型中进行处理。具体数据集信息如表 4-1 所示。

4.2.2 WebNLG 数据集

WebNLG 挑战的意图是在于将 RDF 三元组生成流畅的自然语言文本，由描述事实实体关系的 RDF 三元组和一个或多个人工生成的参考文本组成。WebNLG 数据集包含 9,674 组 RDF 三元组和 25,298 个参考文本，输入三元组个数不一，最多包含七个三元组输入集合，这些三元组是从 DBPedia 抽取出来的，每个参考文本还与它实现的三元组顺序配对。输入 x 是一个<主题，属性，对象>三元组的序列。平均输出长度为 22.5。测试数据跨越 15 个领域，其中 10 个是在训练中看到的。测试集由两部分组成，前半部分包含训练数据中看到的 DB 类别，后半部分包含 5 个未见过的类别。因此 WebNLG 同时兼具大量数据对和多样化、多领域的实体关系，能够更好的训练出泛化性能较好的生成模型，符合本研究方向和目标。由于数据集是较为标准、干净的，导入数据集之后，只需要修复数据集中的一些拼写错误，并且对一些缺失数据进行处理，标记句子中的实体和关系，并处理好大小写的规范，最后再输入到语言模型性中进行文本生成。具体数据集信息如表 4-1 所示。

表 4-1 数据集统计信息

数据集	领域	训练集	验证集	测试集
E2E	酒店信息	42,061	4,672	4,693
WebNLG	15 种 DBPedia 类别信息	34,338	4,313	4,222

4.3 基于去噪自编码的知识图谱文本生成预训练语言模型

4.3.1 任务定义

本任务为输入带有掩码的描述序列，经过优化可以自动回归地预测被屏蔽的标记的 Seq2seq 预训练的语言模型基本框架。给定一个序列长度为 n 的句子，$X = \{x_1, x_2, x_3, \cdots, x_n\}$，$X_{predict} = \{x_i, x_{i+1}, x_{i+2}, \cdots, x_j\}$是一个从$x_i$开始的任意跨度的文本，其中 $1 \le x_i < x_j \le n$。X_{mask}是$X_{predict}$序列当中的 token，并且用[Mask]标志将源输入序列的内容进行替换。因此本掩码端到端的预训练模型任务就是最大化条件概率 $P(X_{predict}|X_{mask})$。

4.3.2 带有噪声输入的编码器和解码器联合预训练模型

由任务定义可以看出，给定一个被噪声损坏的句子，例如输入的带有掩码标记的文本

“Obama is the President of [MASK] [MASK] [MASK]”，想要还原文本时，标准的 Seq2Seq 预训练可以逐个预测掩码的文本片段为“the United States”，即当预测“States”时，之前的基础真值序列“the United”被送入解码器，这种生成模型是基于“the United”这一事实预测生成“States”，但并不代表这样的模型已经充分理解了输入文本中的“Obama”和“the United States”实体之间结构化的事实关系。因此为了不局限于时序序列所表达的信息，让模型更好的去理解跨范围的实体之间的结构化语义信息，本课题提出的处理方式与之前的方法截然不同，如图 4-2 示例所示，将“the United”也替换为“[MASK] [MASK]”标记输入解码器中。这样实体片段中的每个令牌都是在没有真实的实体指示的情况下生成的，预训练将通过学习有关实体的上下文事实和跨范围的结构化信息来推理并生成正确的文本。这样不仅能够促进编码器-解码器结构的联合训练给解码器提供更有用的信息，还能让解码器预测连续的序列片段，以提升解码器的语言建模能力。

基本的模型框架采用基于 Transformer 架构的编码器-解码器模型，并且基于 BART 的大模型架构，BART 模型吸纳借鉴了 BERT 和 GPT 的优点，编码器以输入序列为源输入，而解码器则基于编码器-解码器的自注意机制生成输出序列。在给定句子剩余部分的情况下，重构一个句子片段，编码器以一个随机屏蔽片段（几个连续的 token）的句子作为输入，而解码器则在含有屏蔽片段的某一段序列的基础之上，尝试预测并输出这个被屏蔽的片段内容，以此更有利于训练模型的鲁棒性和对于知识的学习能力。预训练阶段使用英文维基百科作为预训练数据的来源，并且对于文本中的实体个数进行筛选，以提高模型训练的效果，实体 token 可以通过文本对齐维基百科数据集来提取。为了避免预训练过程中的数据泄漏，还需要丢弃与下游数据集中样本重叠的预训练数据。同时也以一定的处理规则对掩码序列的长度进行设置，针对源输入序列的 30%作为掩码序列，对于掩码序列的 80%使用[Mask]标记，10%使用随机 token 进行替换，10%不做处理。

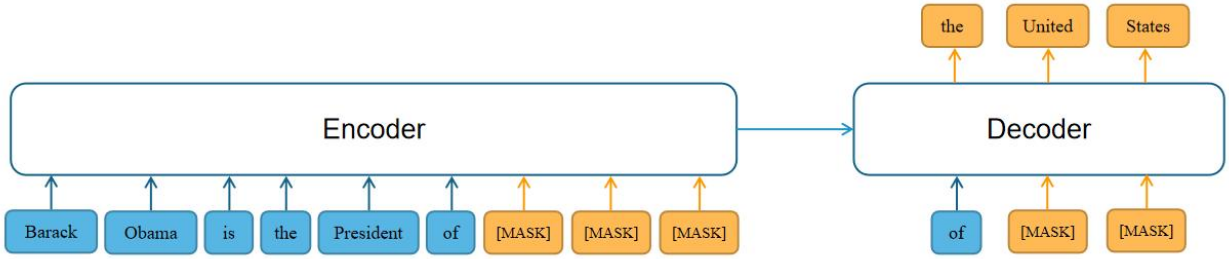

图 4-2 带有噪声输入的编码器和解码器模型文本示例

4.4 面向任务级知识图谱文本生成的前缀微调技术

4.4.1 任务定义

为了实现生成文本的可控性并提高生成文本的准确性，本任务定义一个的任务级别的前缀为P_θ和一组属性级别的控制前缀C_θ，任务级别的前缀微调的目的是最大化特定于任务的语义结构的性能，例如 KG2Text 任务；属性级别的前缀微调目的是生成具有特定目标属

性的文本，与任务性能无关，所选择的属性可以提供关于输入的附加信息，例如知识图谱三元组的集合领域信息，或者它可以指定所需输出的某些方面，例如文本简化的目标长度。这种属性级别的指导信息称为 G，属性级信息或指导指示在处理给定输入 x 时使用哪些控制前缀。针对平行行语料库$\mathcal{Z} = \{(X^j, Y^j, G^j)\}_{j=1,..,N}$，其中$G^j$表示样本 j 的所有属性级条件信息。

因此本任务的目标是通过梯度下降优化最终的推理参数 θ，并且保证预训练语言模型的底层参数 Φ 保持不变，如公式 4-1 所示：

$$\theta^* = \arg\max_\theta \sum_{j=1}^{N} \log p(Y^j|X^j, G^j; P_\theta, C_\theta, \phi) \quad (4\text{-}1)$$

4.4.2 静态前缀和动态前缀相结合的微调技术

研究依赖输入内容的动态提示策略，使用动态提示和静态提示相结合的方式，既维护一个大型的静态提示组件来指定任务本身，又维护具有特定目标属性并与任务性能无关的动态提示。为预训练语言模型提供了一个参数有效的起点，固定语言模型的大部分参数，将其与可训练任务表示相结合，允许模型学习与特定任务相关的信息。此外，引入属性级参数将生成文本的内容引导到所需的方向，并为模型提供数据点级的信息。一般特定的任务参数可以适应模块化控制前缀，这些前缀根据每个输入 x 的特点而变化。这种参数划分可以扩展为细粒度控制，以帮助下游任务提升模型性能。因此前缀提示可以利用输入级信息，并且固定大部分语言模型参数，实现参数有效的轻量级微调方法。对于属性前缀，研究可能的标签属性，例如一篇文章的新闻域，并且对该特定属性的提示前缀实现生成文本的控制。具体的模型示意如图 4-4 所示。

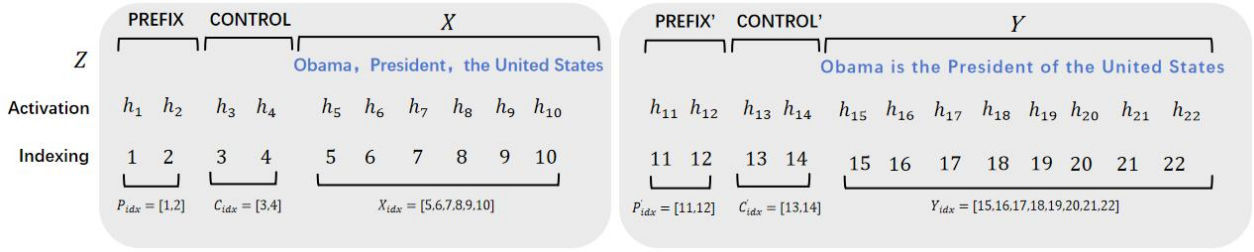

图 4-4 静态前缀和动态前缀相结合的微调模型示意图

4.5 人物报告生成的应用实现

针对文本生成的一个可应用场景，实现人物报告的生成应用。首先通过收集、整理和结构化处理的方式，构建一个包含人物相关信息的知识图谱。该知识图谱包括人物的属性、关系和上下文信息，用于提供生成报告所需的基础知识。在生成报告之前，需要对知识图谱中的数据进行预处理，包括语言处理任务，如分词、词性标注和实体识别，以及数据清洗和规范化，以确保输入数据的一致性和准确性。根据报告的要求和生成目标，选择适当的输入建模方法，涉及将知识图谱中的人物属性、关系和上下文信息编码为适当的向量表

示形式，以供后续的文本生成模型使用。接下来使用训练好的文本生成模型，将构建的知识子图作为输入，并生成符合要求的人物报告。文本生成模型根据输入的知识图谱信息和生成要求，生成连贯、准确且具有可读性的人物报告，包括描述人物的背景、特征、经历等相关信息。最后对生成的人物报告进行评估，以衡量生成文本的质量和合理性，评估指标可以包括常用的机器评估指标和语义一致性、流畅性、信息完整性等。

4.6 实验设计

4.6.1 模型性能评估指标

在本课题中，模型性能评估的标准采用常用的精确率（Precision）、召回率（Recall）、F-1 值（F-1 score）等评估指标。通过在文本生成领域常用的 E2E 数据集与其他基线模型的实验对比，在 BLEU 和 METEOR 评价指标上验证本课题方法的有效性，并且在开放域数据集 WebNLG 上应用本课题所建模型，从而证明本课题方法的泛化性。BLEU 计算公式如式 4-1 所示，首先计算修正的 N -gram 精度 p_n，然后通过 w_n 进行几何平均，根据机器生成文本长度 c 和参考译文长度 r 来计算长度惩罚因子 BP，最终得到 BLEU 得分。METEOR 的计算方法如公式 4-2 所示，F_{means} 是调和平均值，它赋予召回率更大的权重，P 是单图精度，R 是单图召回率，其中∝为可调控的参数，m 为候选翻译中能够被匹配的一元组的数量，c 为候选翻译的长度，r 为参考摘要的长度。

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{4-1}$$

其中，$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}$

$$w_n = 1/N$$

$$\mathrm{MENTOR} = (1 - \mathrm{Penalty}) \times \mathrm{F_{means}} \tag{4-2}$$

其中，$\mathrm{F_{means}} = \frac{\mathrm{PR}}{\propto \mathrm{P} + (1-\propto)\mathrm{R}}$

$$\mathrm{P} = \frac{\mathrm{m}}{\mathrm{c}}$$

$$\mathrm{R} = \frac{\mathrm{m}}{\mathrm{r}}$$

4.6.2 对比实验

针对本课题提出的串行模型设计，面向知识图谱文本生成任务的预训练语言模型采用

GCN[19]、KGPT[30]、BART[26]模型作为对比的基线模型，GCN 是一种基于图卷积神经网络的编码器，允许将图形数据显式编码到神经网络中，从而更好地利用了输入的三元组之间的结构信息。KGPT 模型是一种远程监督学习算法，利用大规模未标记的网络文本来预训练数据到文本模型。并且可以在各种数据到文本生成任务上进行微调，以生成特定于任务的文本。BART 模型提出了用于序列到序列预训练模型的去噪自编码器，并且在文本生成任务和文本理解任务上都能获得不错的效果。

面向知识图谱文本生成任务的预训练语言模型前缀微调技术采用 GPT-2[28]、T5[25]、Li 等人[31]模型作为对比的基线模型，GPT-2 探索了语言模型对零样本生成任务的迁移能力，演示了语言模型可以在零样本设置中执行下游任务的潜力。T5 模型是一个将所有基于文本的语言问题转换为文本到文本格式的统一框架，通过扩展模型和数据集来探索 NLP 迁移学习的前景和局限性。Li 等人最早提出前缀调优，它为利用预训练的语言模型提供了一种有效的训练范式。

4.6.3 消融实验

本课题研究了预训练语言模型和前缀微调的串行模型生成效果，因此需要通过消融实验验证每个模块设计的性能和作用，只使用预训练语言模型进行文本生成任务没有进行前缀微调、使用基本的 baseline 模型进行前缀微调，以及使用本课题提出的预训练语言模型进行前缀微调，并在文本生成任务中应用微调后的模型。对每个消融实验模型，拟采用 BLEU 和 METEOR 评估指标在 WebNLG 数据集上进行消融实验评估。

4.7 可行性分析

本课题的研究目的是基于预训练语言模型实现知识图谱的自然语言文本生成。该课题可以用于生成知识图谱包含的结构信息和文本事实的语言描述，在例如人物或机构的报告生成中实现应用，有助于更好地将结构化的知识图谱知识传递到下游的对话或问答任务中。

在技术方案中，本课题构建了一个数据到文本生成的预训练语言模型，可以使用现有的预训练语言模型架构，针对数据到文本生成任务进行模型设计和训练，并结合知识图谱三元组结构进行特定任务的建模。同时使用了前缀微调技术，可以使用现有的前缀微调方法和技术，根据任务需求和特定的前缀设置，对预训练语言模型进行微调，以提高模型在特定任务上的性能。预训练语言模型以及下游任务中的微调技术最近几年被越来越广泛的使用，近几年高质量的论文模型也大都采用了预训练语言模型及高效微调的技术[19]，因此在技术方案中本课题也沿用了这一主流方法思路，充分利用这些方法在 NLP 领域的优势。

在本课题提出前，本人已经对知识图谱自然语言生成方法进行了系统的学习和梳理，对相关算法和研究有了大致的了解，并学习了本课题实施方案中的相关技术基础。

综上所述，本课题的研究能够在计划时间内达到预期目标。