嗷中图分类号：　　单位代号：10280

密　　级：公开　　学　号：

硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题目	面向不均衡样本的知识图谱构建与推理方法

作　者　陶曦

学科专业　计算机应用技术

导　师　王昊

完成日期　2023 年 5 月

姓　名：陶曦　　学号：20721546

论文题目：面向不均衡样本的知识图谱构建与推理方法

上海大学

本论文经答辩委员会全体委员审查,确认符合上海大学硕士学位论文质量要求。

答辩委员会签名：

主任：

委员：

导　师：

答辩日期：

姓　名：陶曦　　学号：20721546

论文题目：面向不均衡样本的知识图谱构建与推理方法

原 创 性 声 明

本人声明：所呈交的论文是本人在导师指导下进行的研究工作。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签 名：_____日 期：

本论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

（**保密的论文在解密后应遵守此规定**）

签 名：_____导师签名：_____日期：

上海大学工学硕士学位论文

面向不均衡样本的知识图谱构建与推理方法

姓　名：陶曦

导　师：王昊

学科专业：计算机应用技术

上海大学计算机工程与科学学院

2023 年 5 月

A Dissertation Submitted to Shanghai University for the Degree of Master in Engineering

Knowledge Graph Construction and Reasoning Methods for Unbalanced Data

MA Candidate：TAO XI

Supervisor：WANG HAO

Major：Computer Application

School of Computer Engineering and Science,

Shanghai University

March, 2023

摘要

近年来，各个领域逐渐开始重视文本中隐含着的丰富知识，而知识图谱作为一种信息载体，能够有效的存储与表达复杂文本中蕴含的丰富信息。在海量数据的背景下，为了能够快速定位与用户需求相匹配的数据，军事、新闻、搜索等领域均已应用知识图谱技术向外界提供可靠的知识服务。现有的方法在推理知识图谱中缺失的实体关系时，往往仅关注到高频样本，这大大限制了知识图谱的应用。因此，如何在样本分布不均衡的情况下进行知识图谱的构建与推理，是本文需要解决的核心问题。

面向不均衡样本的知识图谱构建与推理方法主要面临以下三个难点问题：一是文本数据具有大规模性，而其内蕴含的知识则具有稀疏性，导致难以从非结构化文本中准确、全面的抽取出隐含的实体与关系并构建知识图谱；二是知识图谱存在信息缺失，并且长尾分布效应的存在导致大部分类别为少样本，这大大提升了推理缺失信息的难度；三是随着信息的不断更新，在实际应用时模型可能需要面对零样本实体，即训练过程中从未出现过的实体，而现有的模型往往难以对该类实体进行有效应对。

针对上述研究难点，本文将针对性的研究以下三个研究内容：

1）基于层次特征表示的知识图谱构建方法。文本数据存在表达方式自由、知识分布稀疏等问题，这导致文本数据难以被有效利用。针对这一问题，本文设计了一种基于分区过滤网络的知识图谱构建模型。该模型根据联合抽取的两个子任务，将文本特征划分为三种类型的特征：实体特征、关系特征和共享特征，并通过特征隔离方法加强不同子任务间信息交互并避免它们的互相干扰，最终从复杂文本数据中构建出知识图谱。

2）面向长尾分布样本的变分异构静态推理方法。知识图谱构建方法能够将非结构化的文本转化为高度结构化的知识图谱。然而，该知识图谱中仍存在信息缺失与长尾分布的问题。因此，本文提出了基于变分异构注意力网络的静态推理算法，该算法通过动态计算邻域信息权重的方式实现信息的有效聚合，从而对中心实体进行有效的语义增强；同时，为了提高模型在长尾分布样本上

的表现，引入了变分信息瓶颈以提高模型在少样本上的性能。通过上述方法，有效的知识图谱中缺失的信息进行了推理，从而扩充与补全了已有的知识图谱。

3）面向零样本基于多层传播的动态推理方法。伴随着领域文本不断更新，实体的种类也随之不断增加。在这样的情况下，模型可能会遇到训练过程未曾观测到的零样本实体。为了能够推理零样本实体的关系，本文提出基于多层传播机制的知识图谱动态推理模型，通过实体信息传播模块和关系信息传播模块充分挖掘零样本实体与已有知识图谱之间存在的关联关系，从而生成零样本实体的向量表示，进而辅助模型进行推理与补全。

本文面向领域中的文本数据，探索知识图谱构建过程中涉及的联合抽取、静态推理和动态推理关键技术，最终形成富含丰富知识的高质量知识图谱，从而更好地为下游应用提供知识服务。

关键词：联合抽取、静态推理、动态推理、知识图谱

ABSTRACT

In recent years, various fields have gradually begun to attach importance to the rich knowledge hidden in texts, and knowledge graphs, as an information carrier, can effectively store and express the rich information contained in complex texts. In the context of massive data, in order to quickly locate data that matches user needs, knowledge graph technology has been applied in fields such as military, news, and search to provide reliable knowledge services to the outside world. Existing methods often only focus on high-frequency samples when inferring missing entity relations in knowledge graphs, which greatly limits the application of knowledge graphs. Therefore, how to construct and infer a knowledge graph in the case of imbalanced sample distribution is the core issue that needs to be addressed in this article.

The construction and inference methods of knowledge graphs for imbalanced data mainly face the following three difficult problems: firstly, text data has a large-scale nature, while the knowledge contained within it is sparse, making it difficult to accurately and comprehensively extract hidden entities and relations from unstructured text and construct knowledge graphs; Secondly, there is a lack of information in the knowledge graph, and the existence of the long tail distribution effect leads to a small sample size for most categories, which greatly increases the difficulty of inferring missing information; Thirdly, with the continuous updating of information, in practical applications, models may need to face zero sample entities, which have never appeared during the training process, and existing models often find it difficult to effectively respond to such entities.

In view of the above research difficulties, this paper will focus on the following three research contents:

1) Knowledge graph construction method based on hierarchical feature representation. Text data has problems such as free expression and sparse knowledge

distribution, which makes it difficult to effectively utilize text data. In response to this issue, this article designs a knowledge graph construction model based on partition filtering networks. This model divides text features into three types based on the joint extraction of two subtasks: entity features, relationship features, and shared features. Through feature isolation methods, it enhances information exchange between different subtasks and avoids mutual interference, ultimately constructing a knowledge graph from complex textual data.

2) Variational heterogeneous static reasoning method for long tailed distribution samples. The knowledge graph construction method can transform unstructured text into highly structured knowledge graphs. However, there are still issues with missing information and long tailed distribution in this knowledge graph. Therefore, this article proposes a static inference algorithm based on variational heterogeneous attention networks. This algorithm achieves effective aggregation of information by dynamically calculating neighborhood information weights, thereby effectively enhancing the semantics of the central entity; At the same time, in order to improve the performance of the model on long tailed distribution samples, a variational information bottleneck was introduced to improve the performance of the model on small samples. Through the above methods, the missing information in the effective knowledge graph is inferred, thereby expanding and completing the existing knowledge graph.

3) Dynamic inference method based on multi-layer propagation for zero samples. With the continuous updating of domain texts, the types of entities also continue to increase. In such cases, the model may encounter zero sample entities that were not observed during the training process. In order to infer the relationship between zero sample entities, this paper proposes a knowledge graph dynamic inference model based on multi-layer propagation mechanism. Through the entity information propagation module and relationship information propagation module, the correlation relationship between zero sample entities and existing knowledge

graphs is fully explored, and a vector representation of zero sample entities is generated to assist the model in inference and completion.

This article focuses on text data in the field and explores the key technologies involved in joint extraction, static inference, and dynamic inference in the process of constructing a knowledge graph, ultimately forming a high-quality knowledge graph rich in rich knowledge, thereby better providing knowledge services for downstream applications.

Keywords: Joint Extraction, Static reasoning, Dynamic reasoning, Knowledge Graph

目录

第一章 绪论

1.1 研究背景及意义

知识图谱内部蕴含着丰富的关联信息，本质上相当于一种语义网络[1]，因此往往会被当成知识库使用。而新闻、军事等领域中往往涉及大量实体，不同实体之间又存在着错综复杂的关系，知识图谱凭借其自身的特征，在描述与表达这些领域实体与关系时能够起到卓越的效果。例如，在军事推演领域中，可以通过由两个实体与一个关系构成的三元组（*J-10，打击，F22*）对作战形势中的“J-10”与“F22”之间的关联关系进行结构化的存储与描述，以此辅助指挥员快速洞悉复杂作战形势。

得益于知识图谱的高度结构化，许多领域均已开始着手构建知识图谱并为下游应用提供服务。如图 1.1所示，在军事推演领域中知识图谱可以以多种方式为指挥员提供数据支撑与知识服务[2]。例如，我们可以以知识图谱为基础构建智能问答机器人，通过人机对话的方式有效降低指挥员获取知识的时间成本，从而达到快速了解敌我实力对比的目的；在个性化推荐系统中，通过建模指挥员的信息需求，快速且准确地向指挥员提供相关的军事情报；在辅助决策信息系统中，则利用静态推理与动态推理技术分析作战单元之间可能存在或即将发生的军事行动，达到辅助指挥员提前制定应对措施的目的。

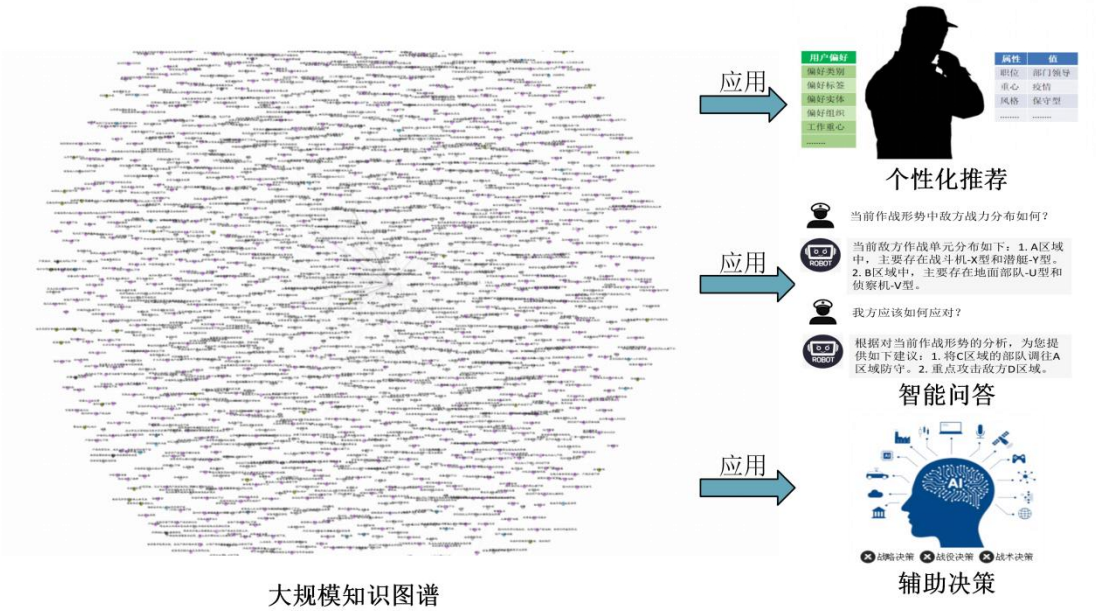

图 1.1　知识图谱在军事推演领域的应用，包括个性化推荐、智能问答与辅助决策。

然而，领域文本表达形式自由，知识分布稀疏，这为知识图谱的构建与利用带来了诸多困难。如何对海量领域文本数据进行挖掘，从中提炼出高质量的知识并以结构化的方式进行存储，成为目前不得不解决的问题，这也是知识图谱构建技术致力于解决的问题。同时，由于领域文本中知识具有稀疏性，这使得通过知识图谱构建技术获取的知识图谱往往备受信息缺失现象的困扰，并且知识图谱中存在严重的样本分布不均衡现象，如何在不均衡样本条件下对知识图谱中的缺失信息进行推理与补全，从而为用户与下游应用提供更加全面的知识服务，是知识推理技术必须解决的问题。

因此，本文以领域文本为切入点，研究面向不均衡样本的知识图谱构建与推理方法。首先，通过爬虫或其他技术获取各个领域的文本数据，并通过基于层次特征表示的知识图谱构建方法将非结构化文本数据转化为由结构化三元组形成的知识图谱；之后，针对样本分布不均衡与知识图谱信息缺失的问题，研究面向长尾分布样本的静态推理模型与面向零样本实体的动态推理模型对知识图谱的不完整性进行弥补；最后，搭建面向新闻领域的辅助决策信息系统将上述研究内容应用于实践当中，验证本文方法有效性的同时为决策者提供有效的数据支撑与知识服务。

1.2 研究问题

上文中介绍到，本研究的目的在于从领域文本中构建出高度结构化的知识图谱，并利用该知识图谱对实体关系进行推理与分析，从而为用户与下游应用提供可靠的知识服务。在以上过程中，本文面向领域文本数据，以知识图谱构建与推理为目标，深入探究这一过程中存在的三个主要研究问题：

1）面向结构复杂且知识稀疏文本的知识图谱构建方法

领域文本存在表达方式自由，有价值的信息分布稀疏的问题，因此需要通过知识图谱构建技术将有价值的信息抽取为实体关系三元组，并以知识图谱的形式将这部分数据进行高度结构化的存储。例如，通过三元组（*俄罗斯，进军，乌克兰*）对“*俄罗斯正在进军乌克兰*”这一事实进行结构化的表示与存储。知识图谱便是由大量这样的三元组汇集而成。然而，面对结构复杂、知识稀疏的领域文本，如何发现隐藏在非结构化文本中的实体之间的关联信息，进而获取其中的知识并进行存储与利用，是一个受到广泛关注的现实问题。知识图谱构建技术作为数据挖掘技术的重要组成部分，能够从非结构化文本中提取特定的实体以及实体间存在的关系。如何有效地对领域文本进行分析，提取其中有价值的信息并以三元组的形式将其保存，是构建知识图谱并向外界提供高质量的知识服务过程中亟待解决的问题。

2）面向长尾分布样本的知识图谱静态推理方法

通过知识图谱构建技术，我们能够从非结构化领域文本中抽取得到各类实体和实体间存在的复杂关系，并初步构建出知识图谱。然而，通过这种方式获取的知识图谱难免会存在信息缺失的问题，同时长尾分布效应的存在意味着大量的实体与关系仅出现较少次数，少量的实体与关系则在知识图谱中频频出现，这为模型的学习带来了困难。因此，如何在长尾分布数据上训练模型推理知识图谱中缺失的信息，以此弥补知识图谱中存在的不完整性，为下游应用提供更高质量的知识服务，是必须要解决的问题。为了解决这一难点问题，必须通过挖掘知识图谱中的长尾分布样本之间存在的关联信息，通过现有知识库对各个实体的缺失关系进行合理推断，从而弥补知识图谱中存在的信息缺失问题，并

向外界提供可靠的数据支撑。静态推理技术旨在整合并分析知识图谱中的已知信息，进而推理各个实体可能存在的关系。如何有效挖掘知识图谱中长尾分布样本的特性，对已知实体可能存在的关系进行合理推断，最终对知识图谱的内容进行扩充，是本研究必须解决的问题。

3）面向零样本实体的知识图谱动态推理方法

由于领域信息的不断更新，模型也必须不断的推理各类实体间可能存在的关系，以此应对外界的不断变化。然而，在这一过程中，模型可能需要推理已有知识图谱中不存在的实体之间的关系，以满足实际应用的需求。以新闻领域为例，随着新闻的不断发布，我们通过知识图谱构建技术从新闻中获取到三元组（*迈克·德瓦恩，攻击，拜登*），其中头实体“迈克·德瓦恩”在先前的知识图谱中从未出现过，属于零样本实体。而为了更好地为决策者提供服务，模型必须推理“迈克·德瓦恩”与其他实体间可能存在的关系。面向零样本实体的动态推理研究能够通过各类算法对零样本实体的特征进行捕获，从而实时推理该类实体可能存在的其他关系。如何全面挖掘零样本实体与已有知识图谱之间的关联信息，对零样本实体进行精确表述，是实现实时推理，提高知识图谱泛用性必须解决的难点问题。

1.3 国内外研究现状

1.3.1 知识图谱构建方法研究现状

知识图谱构建方法旨在从非结构化文本中识别出实体对以及实体对之间存在的关系，并以结构化的三元组对其进行保存，该方法一般包括实体抽取子任务和关系抽取子任务。知识图谱构建方法作为数据挖掘的重要手段，目前已应用于信息索引、问答系统等多个领域。目前，针对该领域的研究可以分为基于特征工程的知识图谱构建方法和基于神经网络的知识图谱构建方法两大类，其中基于神经网络的知识图谱构建方法按照编码方式又可以进一步划分为顺序编码方法与联合编码方法[3]。

1）基于特征工程的方法

基于特征工程的知识图谱构建方法本质是利用领域知识从文本数据中提取与任务相关的特征，并利用这部分特征来提高模型的性能。由于特征工程需要借助领域知识的缘故，这类方法需要结合数据特性设计相应的特征函数[4]。Yang 等人[5]将整数线性规划[6]和条件随机场[7]应用于知识图谱构建之中，并结合文本特性预先定义了大量的特征；Kate 等人[8]则提出一种“卡片金字塔”的模型将文本中潜在的实体和关系紧凑的编码在一起，以此实现知识图谱构建任务，此外该模型还使用动态规划和集束搜索的方式提高模型的速度；Yu 等人[9]和 Sing 等人[10]利用概率图模型来进行知识图谱构建，前者提出了一种具有任意图结构的联合概率判别模型以优化所有相关子任务，后者则设计了一种联合图模型表示任务之间的各种依赖关系；Miwa 等人提出了一种基于历史的结构化学习方法，同时引入了一个实体关系表，通过将抽取任务简化为填表任务以解决任务难以表达的问题。

然而，基于特征工程的方法存在着诸多缺陷：由于该方法需要借助领域知识设计特征函数，因此对研究者的语言学基础和领域知识有较高的要求，同时需要耗费大量的人工成本。此外，这类方法严重依赖自然语言处理工具，且模型的泛化性较差，在跨领域场景之中往往表现较差。

2）基于神经网络的方法

近年来，由于特征工程存在诸多弊端，加上机器算力的飞速提升，基于神经网络的方法在各个领域均取得了显著的成果，因此研究者们也将这种方法应用于知识图谱构建任务之中。不同于基于特征工程的方法，这一方法能够自动提取文本特征。降低人工成本的同时也大大提高了模型的泛化性。在此，我们根据编码方式的不同，将基于神经网络的方法进一步划分为顺序编码方法和联合编码方法两大类，具体如下：

（1）顺序编码方法

顺序编码方法是指按照子任务执行顺序依次捕获特定于子任务的特征，并根据这些特征依次执行各个子任务。例如 Bekoulis 等人[11]等人将词转化为词向量后，利用 BiLSTM[12]和条件随机场对实体抽取任务进行建模，并识别出

文本中潜在的实体，之后结合实体抽取结果与文本信息识别实体之间的关系，从而完成知识图谱构建任务；Ulges 等人[13]提出了一种基于注意力机制的 SqERT 模型，该模型先对实体对进行识别，之后将实体对以及实体对之间的局部语义信息作为输入进一步识别关系；由于上述方法均不能很好地重叠关系的问题，因此 Wei 等人[14]设计了一种新的级联二级制框架（cascade binary tagging framework），通过 Bert[15]获取预训练词向量后，先通过序列标注[16]的方式抽取出文本中的头实体，再结合头实体和其可能存在的关系识别尾实体，从而有效地解决了关系重叠的问题。这类方法虽然取得了不错的成果，但不同子任务间的信息传递是单向的，先执行的任务无法获取后执行任务的信息，这限制了模型的性能。其中，BiLSTM 的结构图如图 1.2 所示：

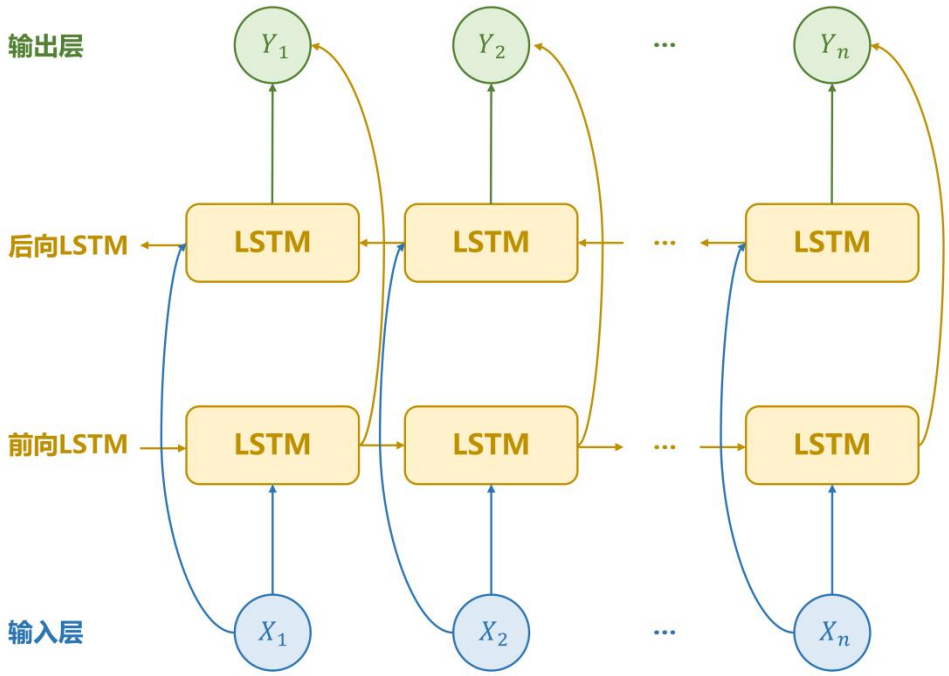

图 1.2　BiLSTM 结构图

（2）联合编码方法

与顺序编码不同，联合编码方法无需按照特定的顺序执行各个任务。一般而言，该类方法是从输入文本数据中提取包含丰富语义信息的特征向量，之后利用一个或多个解码器完成实体抽取与关系抽取任务。例如，Zheng 等人[17]首先将序列标注的方式应用于知识图谱构建任务之中，该方法通过 BiLSTM

充分挖掘各个单词的语义信息，并使用 LSTM 作为统一的解码器，通过序列标注的方式同时完成实体抽取任务与关系抽取任务；由于 Zheng 等人提出的方法无法解决关系重叠问题，Dai 等人[18]对这一标注方法进行了改进，通过基于位置的注意力方法与 CRF 层对每个单词进行多次标注，从而解决了关系重叠问题；除了序列标注方法外，Fu 等人[19]提出了 GraphRel 模型，该模型通过 BiLSTM 从输入文本数据中提取序列特征，之后依赖解析器创建依赖树，并利用 BiGCN 提取文本中的区域依赖特征，最后利用这部分特征分别抽取词与词之间可能存在关系以及词中包含的实体。上述的这类方法虽然均能够实现两个子任务间的信息交互，但这类方法直接融合两个子任务的特征而未能将其区分，这可能会导致子任务之间相互冲突并影响模型性能[20]。

1.3.2 静态推理方法研究现状

静态推理技术旨在通过对图谱中的已知信息进行深入分析，进而对图谱中缺失的事实进行补全。由于这一过程中不涉及图谱中未出现过的实体，因此这一任务又被称为静态推理任务[21]。国内外学者对这一领域进行了大量的研究工作，而基于知识表示学习的方法更是受到相关人员的广泛关注。基于知识表示学习的方法是通过将实体和关系映射到向量空间之中，从而能够通过向量之间的计算挖掘图谱中隐藏信息并完成推理任务[23]。在此，我们介绍四种基于表示学习的静态推理研究：基于翻译的模型、基于语义匹配的模型、基于神经网络的模型和基于图神经网络的模型。

1）翻译模型

基于翻译的模型将关系视作头实体到尾实体的翻译操作，其中最为经典的模型为 Borders 等人[24]提出的 TransE 模型，他们认为关系向量与尾实体向量之和应当近似等于头实体向量的值，并设计实验证明了这一思想的有效性。虽然这一方式能够简单高效地对图谱中的事实进行描述，但仍存在着无法描述实体之间一对多、多对多和多对一情况的问题。因此，Zhang 等人[25]设计 TransH 模型，通过为每一个关系定义一个超平面，实现同一实体在不同关系中拥有不同的意义，解决了 TransE 在复杂情况下表达能力不足的问题。同时，

TransR[26]则通过将实体与关系映射到不同的向量空间来提高 TransE 的表达能力，TransD[27]则在 TransR 的基础上引入了动态变化的投影矩阵以减少 TransR 中不同向量空间相互映射带来的计算成本与模型复杂度。之后，TranSparse[28]设计了动态稀疏矩阵以解决图谱中实体与关系分布的不平衡性问题，TransG[29]则使用一种新的生成模型解决了一个关系可能具有多重语义的问题。这些方法对实体和关系进行了一系列微妙的操作，有效地提高了知识推理的精度。

2）语义匹配模型

语义匹配模型注重挖掘实体与关系之间的潜在关联关系，并设计评分函数对三元组的置信度进行评估。2011 年 Nickel 等人[30]最先提出了 RESCAL 模型，该模型将整张图谱编码为一个三维张量，通过张量分解方法将关系建模为矩阵并进行表示学习，进而度量三元组的可信性。虽然该方法拥有不错的性能，但因模型拥有过多的参数导致容易出现过拟合现象。为此，Yang 等人[31]设计 DistMult 模型对 RESCAL 进行优化，使用对角矩阵代替 RESCAL 中的关系矩阵，这种方式一定程度上消除了 RESCAL 模型中存在的过拟合问题，但却不能很好地模拟非对称关系。在此基础上，Trouillon 等人[32]设计 ComplEX 模型从而将 DistMult 扩展到复数域之中，通过将节点向量映射到复数空间以实现非对称关系的建模。随着技术的不断发展，研究者们设计了众多语义匹配模型对图谱中缺失的事实进行推理[33-36]。然而，这类模型虽然取得了不错的成就，但仍存在着训练参数繁多、容易过拟合等问题[37]。

3）神经网络模型

鉴于神经网络模型在各个领域中取得的优异性能，研究人员们逐渐将神经网络模型融入到静态推理任务之中。2013 年 Chen 等人[38]提出神经张量网络模型 NTN，采用双线性张量层刻画实体之间的复杂联系并取得了显著的效果。之后，Dettmers 等人[39]设计包含卷积层、全连接层和点积层的 ConvE 模型，该模型将头实体向量与关系向量拼接为矩阵，之后对该矩阵进行二维卷积操作，将结果与候选实体集进行内积，经全连接层得到推理结果。Nguyen 等人[40]提出的 ConvKB 模型与 ConvE 不同，该模型构建正负样本，并将正负样本的

头实体、关系和尾实体均输入至模型中，并在后续卷积操作中使用一维卷积。通过这一方式，ConvKB 不仅能够关注到三元组的局部信息，还保留了三元组的整体结构信息，从而增强了知识之间的信息交互。

4）图神经网络模型

由于知识图谱天然具有图结构，因此基于图神经网络的模型逐渐引起了研究人员的关注。2018 年 Schlichtkrull 等人[41]提出 R-GCN 模型，该方法通过卷积操作将邻居信息聚合至中心实体以丰富实体的表示，是首先将图神经网络应用于静态推理任务的方法。受到这一思想的启发，Nathani 等人[42]在 g-GAT 模型中引入图注意力机制以提高表示学习的性能，该方法认为同一中心节点的不同邻居的重要性程度是不同的，因此在信息聚合的过程中应当对它们进行区分。之后，Zhang 等人[43]设计了分层注意力机制，通过实体层注意力和关系层注意力计算邻居实体的重要性，并使得相同关系下的相邻三元组的权重可以集体训练。此外，Zhao 等人[44]添加了额外的全局注意力机制，Fang 等人[45]则将注意力机制从对实体表示的学习引申到对关系表示的学习上。这些方法通过图神经网络有效地利用了图结构信息，并使得模型能够学习到更具表现力的实体表示与关系表示，从而有效地提升了模型的性能。

1.3.3 动态推理方法研究现状

静态推理任务能够很好地表示训练阶段观测到的实体，但是由于领域文本的不断更新，模型需要推理的实体也不断变化，在这一过程中模型可能推理训练过程中未被观测到的实体的关系。针对这一问题由 Hamaguchi 等人[46]首先最先提出，该问题一经提出便引起了学术界的广泛关注，现有的解决方案主要分为两大类：一是通过引入外部知识对零样本实体进行描述，即基于零样本实体文本描述的方法；二是挖掘零样本实体与已有知识图谱之间的关联关系以对零样本实体进行描述，基于辅助三元组关联信息的方法。具体如下：

1）基于零样本实体文本描述的方法

基于零样本实体文本描述的方法是利用新实体的文本描述形成实体的特征向量，从而将零样本实体映射到向量空间中并执行动态推理任务。例如，Xie

等人于 2016 年将 Freebase 知识库中的实体描述文本数据作为输入，通过 DKRL 模型提取并融合文本中各个词对应的词向量，从而增强实体的表示。Shi 和 Weninger 等人[48]则提出了一种以文本为中心的 ConMask 模型，该模型通过 MCRW 对文本描述信息进行预处理，仅保留与目标任务有关的文本片段，从而使得模型能够更好地关注有效信息，之后通过全卷机神经网络从文本片段中提取实体的特征，最终通过对候选实体进行评分与排序以得出推理结果。Shah 等人[49]等人则提出了 OWE 模型，该模型独立的训练模型分别从图谱和文本嵌入中获取实体的特征向量，并训练模型将基于文本的特征向量空间转化为基于图的特征向量空间，以此充分地利用文本描述信息与图结构信息。上述这些方法虽然都取得了不错的效果，但是由于多数情况下图谱中缺乏实体的文本描述信息，所以这类方法往往会受到文本描述的约束。

2）基于辅助三元组关联信息的方法

基于辅助三元组关联信息的方法是利用零样本实体与已知实体之间的有限联系，挖掘零样本实体与已有知识图谱之间存在的关联关系并对其进行表示。Hamaguchi[46]利用图神经网络挖掘零样本实体与已知实体之间的关联关系，并通过转移函数与池化函数生成零样本实体的向量表示，最后借助 TransE 模型完成动态推理任务；之后，Bi 等人[50]利用图神经网络挖掘节点之间的关联关系之后，使用卷积神经网络作为转移函数，从而以更少的参数学习到零样本实体更富有表现力的特征向量；为了让模型能够关注到邻域信息之间的差异性，Wang 等人[51]通过查询关系以判断邻居节点的重要性，而 Zhao 等人[52]则提出了一种基于图注意力的聚合函数来解决这一问题，均取得了很好的效果；而 Xie 等人[53]则提出了 Neighbor-T 模型解决这一问题，并在这一模型中通过统计学的方法对邻居实体进行有效区分。上述方法为解决零样本实体表征问题提供了很好的思路，在挖掘关联关系的过程中充分地考虑了邻居关系的特征，但却缺乏对邻域实体自身特性的考虑，因此还存在改进空间。

1.4 研究内容

针对非结构化领域文本数据，本文根据目前国内外研究现状与实际在知识图谱构建与推理过程中遇到的难点问题，提出了自己的解决方法。我们的研究内容可以概括为以下三点：

1）基于层次特征表示的知识图谱构建方法。为了从非结构化领域文本中提取结构化知识信息，形成知识图谱，提出了基于分区过滤网络的知识图谱构建模型，从而从海量结构复杂的领域文本中准确的抽取实体关系三元组，并向外界提供可靠的知识服务。

知识图谱构建模型一般通过对输入文本数据进行建模并从中提取特征，进而识别文本数据中的实体及其关系。这类方法一般包括两个子任务，即实体抽取与关系抽取，它们分别用于从非结构化文本中发现实体和实体间存在的关系。然而，基于顺序编码的方法两个子任务间的信息交互是单向的，这使得后执行任务的性能往往会受到先执行任务性能的限制。基于联合编码的方法虽然一定程度上解决了信息交互的问题，但仍存在不同任务间特征干扰与特征冲突的问题。因此，本文在第二章中提出了一种基于分区过滤网络的知识图谱构建模型。首先，文本嵌入层中针对非结构化领域文本，使用预训练模型 BERT 提取文本中蕴含的丰富语义信息，并借助隐藏层获取丰富的上下文知识以输入至下游模型；之后，为了加强不同子任务间的特征交互，使用实体嵌入层和信息门分别获取实体特征与关系特征，并利用共享特征提取器从二者中提取出共享特征以加强任务间的信息交互，并借助特征隔离方法消除交互过程中可能存在的信息冲突与信息冗余；最后，采用“填表”的方式抽取隐藏于文本中的实体，并结合抽取结果对实体间的关系进行预测，有效解决结构复杂、知识稀疏的领域文本的实体与关系抽取问题。

2）**面向长尾分布样本的变分异构静态推理方法。**针对知识图谱中存在的信息缺失与样本存在长尾分布效应的问题，提出基于变分异构注意力网络的静态推理模型对知识图谱中的缺失信息进行推理，最终形成高质量的知识图谱。

由于知识图谱天然具有图结构，因此对图谱中各个节点进行深入分析的时候往往会引入图卷积网络以更好地获取实体与关系的表征。然而，由于长尾分布效应的存在，已有的方法往往难以获取少样本实体与少样本关系的高质量表示；同时，在聚合邻域信息的过程中，如何对邻域实体与邻域关系的重要性进行有效评估也会影响到推理的精度。因此，本文在第三章中提出基于变分异构注意力网络的静态推理模型。首先，我们设计了一种新型的异构注意力网络对邻域实体和邻域关系的重要性进行计算，对邻域信息进行合理采样，以此提高模型在高频样本上的表现；之后，利用变分信息瓶颈模块对信息流进行控制，以此强化模型的泛化能力，提高模型在少样本上的表现；最后，使用 ConvKB 模型[40]进行解码，通过挖掘实体与关系表示之间的全局特性以完成静态推理任务。

3）面向零样本基于多层传播的动态推理方法。为了让模型能够对零样本实体的关系作出合理预测，提出基于多层传播机制的动态推理方法，通过实体信息传播层与关系信息传播层充分挖掘零样本实体与已有知识图谱之间的关联，从而实现知识图谱的动态推理。

为了能够获取训练过程中未曾出现过的实体的高质量表示，现有的方法主要通过如下两种方式：基于零样本实体文本描述的方法和基于关联信息的方法。基于零样本实体文本描述的方法容易因实体文本描述缺失而受到限制，并且文本描述限制了模型充分挖掘已有知识图谱中蕴含的丰富信息，且文本质量将对模型的性能产生直接影响；基于关联信息的方法则致力于通过挖掘零样本实体与已有实体之间存在的关联，对零样本实体的特征进行分析与表达，但在挖掘关联信息的过程中未能充分结合已有实体的固有特性。为此，本文第四章提出一种基于多层传播机制的动态推理模型。首先，在模型输入层，我们捕获所有实体的邻居信息，并对关系的方向性加以考虑；之后，通过实体信息传播层和关系信息传播层分别从邻域实体和邻域关系中挖掘关联信息，从而对零样本实体进行充分表达；最终，根据上一步生成的零样本实体的表示进行动态推理任务，以此保证模型对零样本实体的关系也具备推理能力。

本文总体研究路线如下：我们将海量领域文本作为切入点，通过知识图谱构建技术从非结构化领域文本中抽取蕴含丰富知识的三元组数据，以此初步构建出知识图谱，从而以高度结构化的形式向用户和下游应用提供信息服务；然后，针对知识图谱中存在的信息缺失、关联性弱的问题，我们通过静态推理技术对实体间缺失的关系进行合理分析与推测，从而更加全面的对实体进行描述；之后，利用动态推理技术对零样本实体的潜在关系进行实时推理，以此满足应用过程中对实时性的需求。

1.5 研究创新点

本文针对构建知识图谱过程中的实体关系联取、静态关系推理以及动态关系推理这三个方向进行了深入探索，在此将本文的创新点总结如下：

1）基于特征工程的方法严重依赖于领域知识与特征函数导致抽取成本过高；基于顺序编码的方法无法实现不同子任务之间的双向交互，基于联合编码的方法则存在特征交互过程中有冲突和冗余的问题，这限制了模型的性能。针对现有研究存在的问题，本文研究基于层次特征表示的知识图谱构建方法。具体而言，我们提出了基于分区过滤网络的知识图谱构建模型，该模型通过 Bert 模型对输入领域文本进行编码，之后利用实体嵌入层和关系嵌入层分别从文本语义向量中提取与实体抽取任务相关的特征和与关系抽取任务相关的特征。之后，利用共享特征分离层将这两部分特征切分为三部分，分别是仅与实体抽取任务相关的实体特征、仅与关系抽取任务相关的关系特征以及和两个任务均相关的共享特征，并通过共享特征实现不同子任务之间的信息交互，并引入特征隔离方法消除不同种类特征间的冲突与冗余。最后，将这些特征输入至相关的任务分区，并利用“填表”的方式完成知识图谱构建任务，从而将非结构化领域文本数据转化为易于保存和利用的结构化数据。

2）针对上一步构建的知识图谱存在的知识缺失问题，以及现有的模型难以很好的处理长尾分布样本的问题，本文研究面向长尾分布样本的变分异构静态推理方法。具体而言，我们提出一种基于变分异构注意力网络的静态推理模型，该模型通过设置查询向量从而在聚合邻域信息的过程中动态的查询来自邻

域实体和邻域关系的权重，进而有效的聚合重要的邻域信息，从而有效的学习到高频样本的表示；针对少样本，则引入了变分信息瓶颈模块辅助模型更好的捕获接结点的深层次特征，通过对结点的表示进行信息压缩并融合高斯噪音的方式增强了模型对少样本实体与少样本关系的表现能力，从而显著地提升了模型在少样本实体与少样本关系上的推理精度。通过基于变分异构注意力网络的静态推理方法，有效的缓解了样本存在长尾分布效应背景下的知识图谱信息缺失问题，从而更好为用户更好的了解领域知识奠定数据基础。

3）基于零样本实体文本描述的方法中目标文本难以获取，并且模型性能容易受到文本描述质量的制约；基于辅助三元组关联信息的方法在挖掘零样本实体与已有知识图谱之间关联关系的过程中，往往会忽略已知实体对关联信息挖掘的作用，这可能会限制模型对关联信息的充分挖掘。针对上述问题，本文研究面向零样本基于多层传播的动态推理方法。具体而言，我们提出一种基于多层传播机制的动态推理模型，该方法包含实体信息传播模块和关系信息传播模块，通过实体信息传播模块充分挖掘零样本实体与已知实体之间的关联关系，通过关系信息传播模块实现关系信息沿关系路径的有效传播，最终有效地通过辅助三元组关联信息实现零样本实体的充分表达，并且避免了对文本描述数据的依赖，最终实现动态推理任务，有效推广了我们模型在实际应用过程中的实用性与泛用性。

1.6 论文组织架构

本文分为七个章节，具体组织架构如图 1.3 所示：

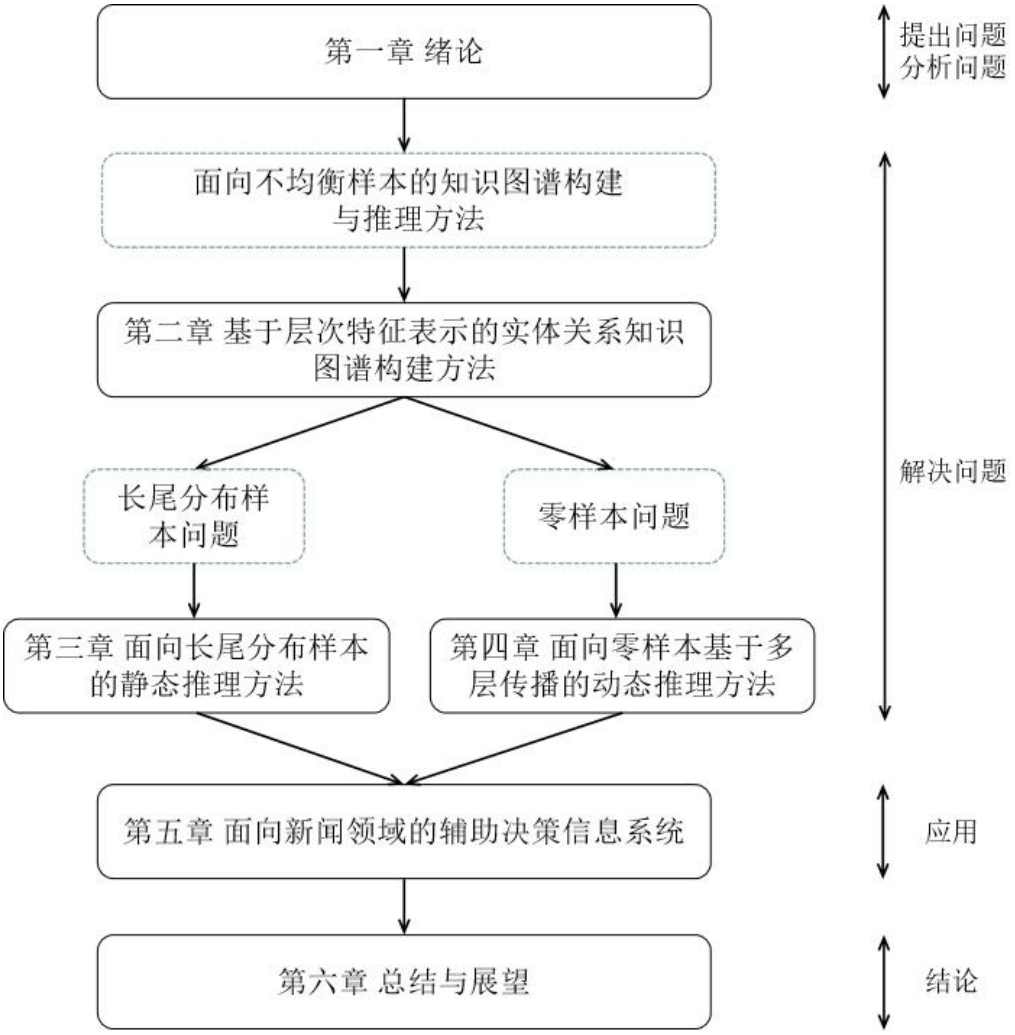

图 1.3　论文组织架构，其中实线表示论文实际章节，虚线表示拟解决的问题

在第一章中，首先分析了知识图谱构建与推理研究的研究背景与意义，从而引出本文需要研究的问题；之后，针对本文需要解决的研究问题，对相关领域的国内外研究现状进行了梳理，总结出先前研究值得借鉴的地方和不足之处；之后根据已有研究的不足之处提出我们的解决方案，并介绍我们的研究内容与研究创新点。

在第二章中，我们主要探讨了如何从领域文本中构建知识图谱，针对这一问题，我们提出了一种基于分区过滤网络的知识图谱构建模型。首先，针对知识图谱构建任务进行分析，介绍该研究方向目前存在的不足以及研究难点，并引出我们的模型；之后，用数学语言对联合抽取问题进行了定义；接着，详细阐述了提出的基于分区过滤网络的知识图谱构建模型；最后，在多个数据集上对该模型进行了相关的实验，以此验证了我们方法的有效性。

在第三章中，由于知识图谱中样本呈现长尾分布，并且知识图谱中存在信息缺失，因此，提出了基于变分异构注意力网络的静态推理模型。首先，我们指出了现有静态推理方法中存在的问题与难点，并提出基于变分异构注意力网络的静态推理模型；然后，对该任务涉及的相关概念进行了描述；之后，介绍了该模型的总体架构与细节实现；最后，通过在多个数据集上的对比实验、消融实验与超参数实验对该模型的有效性进行了验证。

在第四章中，针对零样本实体问题，提出了基于多层传播机制的动态推理模型。首先，结合实际应用场景分析了为什么必须解决零样本实体问题，以及在这一过程中存在的难点问题，并提出了基于多层传播机制的动态推理模型。然后，对该任务中相关的概念进行了公式化描述；之后，结合设计思路介绍了提出方法的整体框架与细节实现；最后，通过实验验证了该方法的有效性。

在第五章中，结合上述研究内容设计了面向新闻领域的辅助决策信息系统，从而将研究的内容与实际应用结合起来，以此证明上述方法的有效性。

在第六章中则对上述所有内容进行了总结，并对未来的工作进行展望。

第二章 基于层次特征表示的知识图谱构建方法

2.1 实体关系联合抽取问题的提出与定义

2.1.1 问题的提出

知识图谱构建方法往往是通过实体关系联合抽取技术实现的，该方法旨在从领域文本中发现并抽取实体对以及实体对之间存在的关系，并通过结构化的知识图谱对这些信息进行存储与表达，从而将难以直接利用的非结构化文本数据转化为结构化的三元组数据，并为下游应用提供可靠的知识服务。这一过程一般包含两个子任务，实体抽取任务与关系抽取任务，前者用于从领域文本中提取出所有的实体，后者则是用于判断实体之间是否存在某种关系。

表 2.1 知识图谱构建示例，通过实体抽取技术与关系抽取技术将非结构化新闻文本转化为结构化三元组。

新闻文本	三元组
德国经济部长彼得·阿尔特迈尔希望美国总统拜登上任后能够消除美欧间的贸易壁垒。	头实体（吴京） 关系（主演） 尾实体（《战狼》） 头实体（吴京） 关系 （导演） 尾实体（《战狼》）
《我和我的祖国》是由陈凯歌担任总导演，张一白、薛晓路、徐峥、宁浩等人联合执导的剧情片，于 2019 年 9 月 30 日在中国大陆上映。	头实体（陈凯歌） 关系 （导演） 尾实体（《我和我的祖国》） 头实体（薛晓路） 关系 （导演） 尾实体（《我和我的祖国》） 头实体（徐峥） 关系 （导演） 尾实体（《我和我的祖国》） 头实体（宁浩） 关系 （导演） 尾实体（《我和我的祖国》） 头实体（张一白） 关系 （导演） 尾实体（《我和我的祖国》）

知识图谱构建技术在各个领域中均有着广泛的应用，并且取得了良好的效果。以新闻领域为例，对于新闻文本“*德国经济部长彼得·阿尔特迈尔希望美国总统拜登上任后能够消除美欧间的贸易壁垒。*”，通过实体抽取子任务与关系抽取子任务后，最终能够将这段非结构化的新闻文本转化为如下结构化信息：“*彼得·阿尔特迈尔 任职 德国经济部长*”和“*拜登 任职 美国总统*”。通过上述方式，能够从海量新闻文本中构建知识图谱，为有效利用新闻文本中蕴含的丰富知识奠定基础。表 2.1 详细阐述了利用新闻文本进行知识图谱构建的过程与结果。

目前，知识图谱构建技术可以分为基于特征工程的知识图谱构建技术与基于神经网络的知识图谱构建技术，其中后者又可被进一步划分为基于顺序编码的知识图谱构建技术[54-58]和基于联合编码的知识图谱构建技术[17-20]。基于特征工程的知识图谱构建技术一般利用领域知识从文本数据中提取与任务相关的特征，并利用这部分特征完成实体抽取任务与关系抽取任务，最终构建出知识图谱；对于基于神经网络的知识图谱构建技术中，基于顺序编码的知识图谱构建技术通过子任务的执行次序依次对特征进行编码，一般而言，这类模型先从文本中提取实体特征并进行实体抽取任务，之后根据实体抽取结果编码关系特征，进而进行关系抽取任务；而基于并行编码的知识图谱构建技术则是从文本进行统一编码并提取特征，之后独立的进行实体抽取任务与关系抽取任务，或是直接对三元组进行建模，从文本中同时抽取实体与关系。

然而，通过分析，本文认为现有的知识图谱构建方法仍然存在着如下问题：

（1）基于特征工程的知识图谱构建模型虽然具有较强的可解释性，但是需要大量人工对文本内容进行分析，并设计相应的特征工程。这使得基于特征工程的知识图谱构建模型虽然具有较高的可解释性，但由于人力的有限，其召回率与泛化性通常较低，并且构具有较高的构建成本。

（2）基于顺序编码的知识图谱构建模型通过顺序编码实体特征与关系特征进行知识图谱构建任务，在这一过程中不同的编码顺序往往存在不同的含义，且对模型的性能影响较大，这提高了模型设计的难度。此外，顺序编码方式会使得对最先编码的特征无法参考之后编码的信息，即两个子任务间的信息仅能进行单向传递，这极大限制了模型提取特征的能力，进而为模型的性能带来了影响。

（3）基于联合编码的知识图谱构建模型无需按照特定的顺序对输入数据进行编码。然而，现有的方法在两个子任务进行信息交互的过程中，往往忽略了这部分信息间可能存在的冗余或冲突，这使得模型不能很好的利用这部分信息，从而限制了模型的性能。

针对这些问题，本文研究基于层次特征表示的知识图谱构建方法。具体而言，本文结合基于联合编码的知识图谱构建方法，设计了一种新的基于分区过

滤网络的知识图谱构建模型。对于输入的领域文本，该模型首先采用 BERT 模型提取文本特征，之后通过实体嵌入层和关系嵌入层初步提取实体特征与关系特征，之后再通过共享特征分离层从这两类特征中分离出共享特征，以此实现不同任务间信息的双向交互。之后，我们设计并应用各种特征隔离方法将不同类型特征映射到不同的向量空间之中，以此避免不同特征之间的冲突与冗余。最后，根据先前获取的各类特征进行实体抽取任务与关系抽取任务。我们的主要工作如下：

（1）设计一个新型的分区过滤网络，该网络能够从文本信息中获取实体特征与关系特征，并将二者重叠的部分分离出来，从而得到共享特征，以此实现不同类型特征之间的相对独立，并保证了不同子任务之间的信息交互。

（2）通过引入特征隔离方法将不同类型的特征映射到不同的向量空间，达到消除它们之间的冗余与冲突的目的，从而使得模型能够更好的进行实体抽取任务与关系抽取任务。

（3）我们在多个数据集上进行了一系列的对比实验与消融实验，实验结果验证了我们方法的有效性。

2.1.2 问题的定义

知识图谱构建任务旨在从输入的一段领域文本中抽取出所有的实体对以及实体对之间存在的关系，最终形成知识图谱。该任务可以定义为如下形式：对于输入的长度为 L 的领域文本 $S=\{s_1,s_2,...,s_L\}$，我们需要从中抽取若干对三元组，每个三元组包含两个实体和一个关系，即 $\tau_i=(e_{i1},r_i,e_{i2})$，其中 $e_{i1}=\{s_j,...,s_k\},(0<j<k<=L)$ 和 $e_{i2}=\{s_{j'},...,s_{k'}\},(0<j'<k'<=L)$ 为领域文本中蕴含的实体，而 $r_i\in R$ 则为我们预先定义的关系集合。实体关系联合抽取旨在发现领域文本中包含的所有实体对，并确定它们之间的关系。

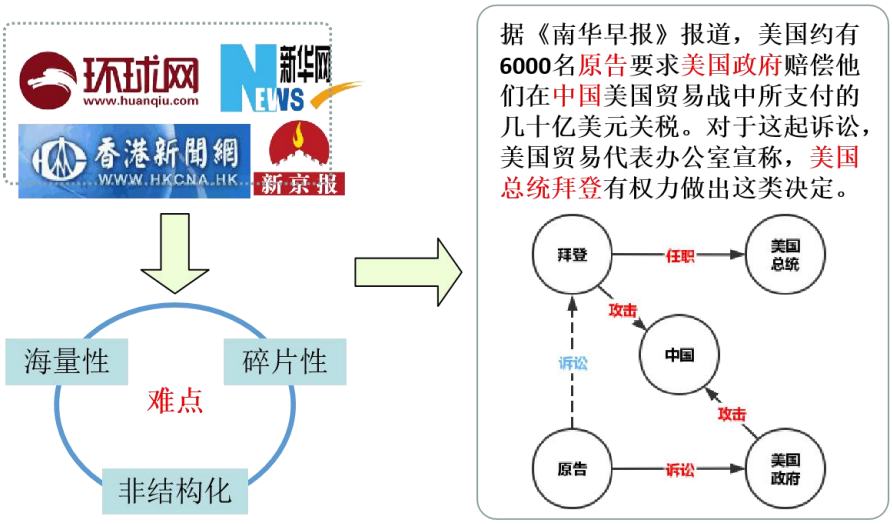

图 2.1　面向新闻领域的知识图谱构建示例，通过爬虫技术获取新闻语料，经数据清洗与自动标注方法获取原始数据集，进而训练模型并完成知识图谱构建。

以新闻领域为例，该领域的知识图谱构建过程如图 2.1 所示，首先我们从各大新闻网站中通过爬虫技术获取相关新闻语料；之后通过分句、去重、人工审核等方式完成对数据的初步清洗，并根据数据特性定义关系集合 R，包括*任职，发表，合作，地理，外交，支持，攻击，来自，检查，研究，竞争，管理，采访，隶属，雇佣*共计 15 种关系；继而，通过规则库生成原始标注数据，经人工校验后得到最终标注数据集。例如图中的新闻文本“*据《南华早报》报道，美国约有 6000 名原告要求美国政府赔偿他们在中国美国贸易战中所支付的几十亿美元关税。对于这起诉讼，美国贸易代表办公室宣称，美国总统拜登有权力做出这类决定*”中，通过规则库生成的原始标注数据集中包含实体对 $e_1 = $ ***拜登*** 与 $e_2 = $ ***中国***，它们之间的关系为“*任职*”，其中 $e_2 = \{s_{76}, s_{77}\}$，$e_2 = \{s_{33}, s_{34}\}$。最后，利用本文提出的基于分区过滤网络的知识图谱构建模型，学习数据集的特征，从而利用模型自动从新闻文本中提取实体及关系。

2.2　基于分区过滤网络的知识图谱构建模型

本文提出了基于分区过滤网络的知识图谱构建模型，该模型从文本信息中分离出实体特征、关系特征与共享特征，通过共享特征实现不同子任务间的有

效信息交互，通过将不同类型特征映射到不同向量空间实现特征间的相互隔离，从而避免了它们之间的相互干扰与冲突，最终完成知识图谱构建任务，模型总体架构如图 2.2 所示：

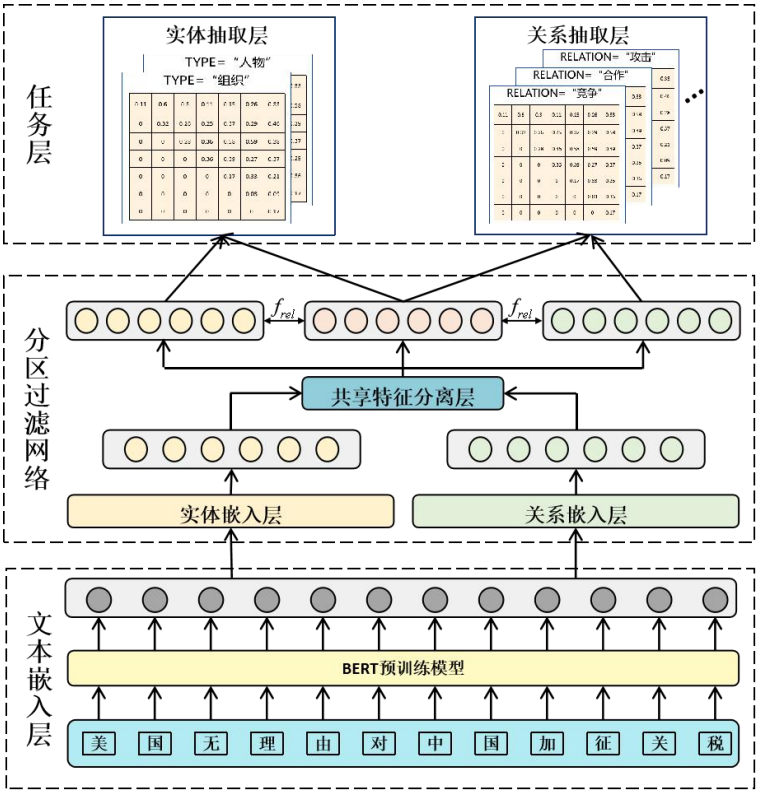

图 2.2　基于分区过滤网络的知识图谱构建模型架构图，通过分区过滤网络获得实体特征、关系特征与共享特征，最终完成实体抽取与关系抽取任务。

我们将在下文中介绍基于分区过滤网络的知识图谱构建模型的设计思路与实现方式。主要分为如下四个部分：在 2.3 节中介绍模型如何通过文本嵌入层从领域文本中提取高质量特征；2.4 小节中介绍我们如何设计分区过滤网络并从文本特征中提取不同类型的特征，以及如何通过设计特征隔离方法避免不同类型特征之间可能出现的冲突与冗余；2.5 节和 2.6 节则分别介绍了模型如何进行知识图谱构建任务以及训练过程中损失函数的计算方式。

2.3 文本嵌入层

以 BERT 模型为代表，基于 Transformer 架构的语言表示模型在 NLP 领域上取得了巨大的成功，为了获取更加丰富的初始化语义向量，我们采用了 BERT 模型从文本中提取原始特征，该模型的结构图如图 2.3 所示：

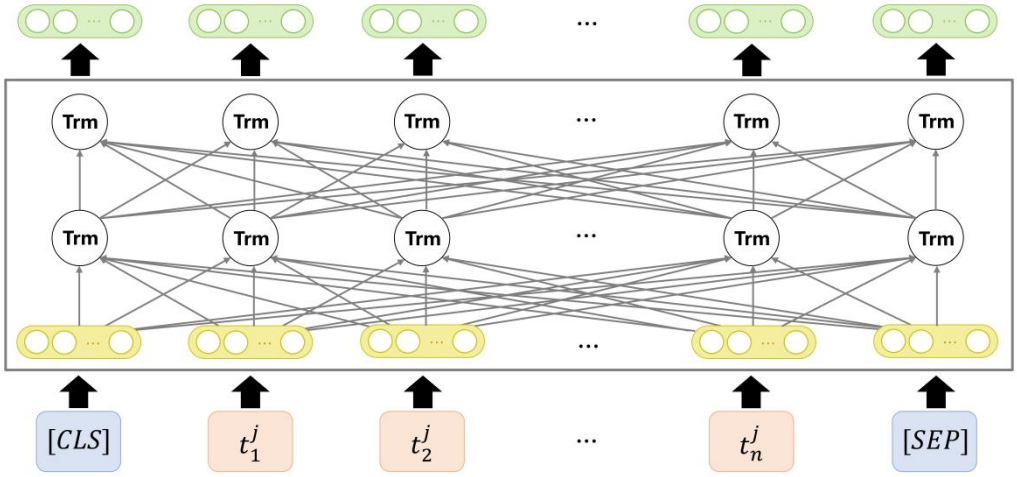

图 2.3 BERT 架构图，通过该模型能够从输入文本中提取到高质量的文本特征。

BERT 模型是 2019 年由谷歌提出的预训练模型，其训练集包括维基百科中所有的无标签文本数据和图书语料库，采用 MLM（Masked Lauguage Model，掩码语言模型）对模型进行预训练，通过多层 Transformer 结构提取融合丰富上下文信息的语义表征，并仅需对参数进行微调便可应用于各类自然语言处理任务之中。

对于输入的领域文本内容 $S = \{w_1, w_2, ..., w_n\}$，我们以字为单位将其输入至 Bert 模型中。以领域文本“英国是美国的主要盟友”为例，我们将其处理成 $input = \{CLS, 英, 国, 是, 美, 国, 的, 主, 要, 盟, 友, SEP\}$。在本文中，模型采用 bert-base-cased，输出词向量维度为 768。

获取包含丰富上下文信息的预训练词向量后，为了能够更好的利用上下文信息，我们参考 LSTM 模型的思想，将当前时刻的文本信息与上一时刻的隐藏层输出进行融合，从而在获取当前时刻信息的过程中进一步的考虑上下文信息，具体如下：

$$\hat{f}_t = tanh[L(x_t||h_{t-1})] \quad (2.1)$$

其中，L 表示线性层，x_t 为 BERT 模型在 t 时刻输出的预训练词向量，h_{t-1} 为上一时刻隐藏层输出的信息，具体结构如图 2.4 所示。在 $t=0$ 时 h_{t-1} 的取值为 0，其具体计算方法会在下文中给出。

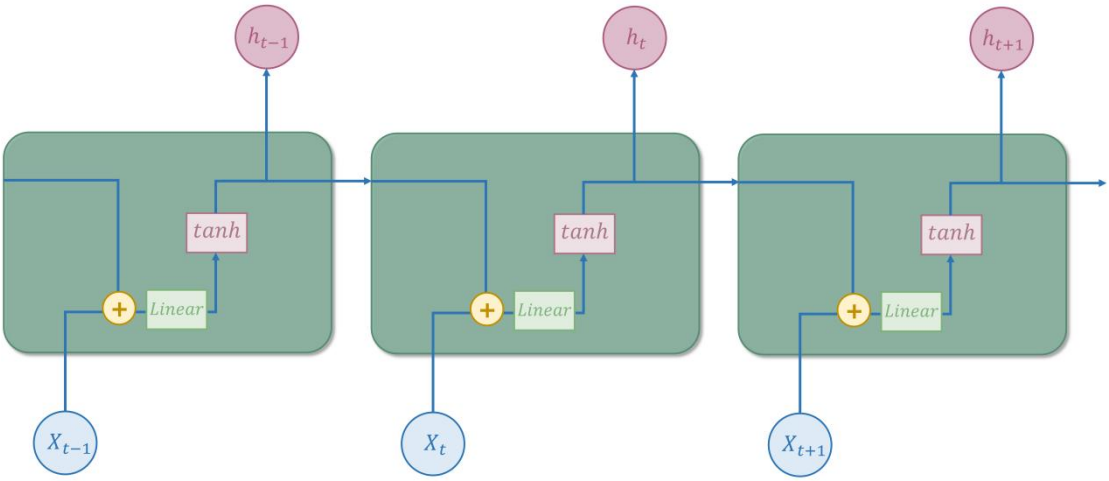

图 2.4　基于分区过滤网络的知识图谱构建模型文本嵌入层架构图。

2.4　分区过滤网络

为了能够更好的对信息进行区分，我们采用了实体嵌入层与关系嵌入层控制输入的信息流。实体嵌入层和关系嵌入层会保留与相应子任务有关的信息，并过滤与对应子任务无关的信息。以实体嵌入层为例，实体嵌入层的作用是将信息划分为对实体抽取任务有用的信息与对实体抽取任务无用的信息，并保留有用信息。在本文中，我们采用如下方式定义实体嵌入层与关系嵌入层：

$$\hat{e}_t = cumsum\{softmax[L(x_t \,||\, h_{t-1})]\} \quad (2.2)$$

$$\hat{r}_t = 1 - \hat{e} = 1 - cumsum\{softmax[L(x_t||h_{t-1})]\} \quad (2.3)$$

其中，cumsum 函数用于计算 softmax 函数输出数组的各项累加值，如对于输入(0.3, 0.3, 0.3, 0.1)，经 cumsum 函数后会输出(0.3, 0.6, 0.9, 1.0)。通过这种方式，我们能够将信息划分为两个部分，即与实体抽取任务相关的部分和与关系抽取任务相关的部分。之后，我们通过实体嵌入层与关系嵌入层计算得出共享门。遵循的思路是：与实体抽取任务相关的信息才能够通过实体嵌入层，

与关系抽取任务相关的信息才能通过关系嵌入层，若信息既能够通过实体嵌入层，又能够通过关系嵌入层，则说明该信息既与实体抽取任务相关，也与关系抽取任务相关。出于这样的考虑，我们通过如下方式计算得出共享门：

$$\hat{s}_t = \hat{e}_t \cdot \hat{r}_t \tag{2.4}$$

之后，我们通过实体嵌入层、关系嵌入层与共享门将输入的文本信息划分为三个部分，即实体信息、关系信息与共享信息，具体方式如下：

$$o_{s,t} = o_{s,t-1} \cdot f_{t-1} + o_{s,t} \cdot \hat{f}_t \tag{2.5}$$

$$o_{e,t} = o_{e,t-1} \cdot f_{t-1} + o_{e,t} \cdot \hat{f}_t \tag{2.6}$$

$$o_{r,t} = o_{r,t-1} \cdot f_{t-1} + o_{r,t} \cdot \hat{f}_t \tag{2.7}$$

如上式所示，我们通过实体嵌入层与关系嵌入层得到了三个分区，即实体分区、关系分区和共享分区。此处我们使用 η_e 表示与实体抽取任务有关的信息，η_r 表示与关系抽取任务有关的信息，η_s 表示与两个任务均有关的信息。由实体嵌入层与关系嵌入层的含义可知，η_e 为实体分区与共享分区内存储的信息，η_r 为关系分区与共享分区内存储的信息，即：

$$\eta_{e,t} = o_{\mathrm{e,t}} + o_{\mathrm{s,t}} \tag{2.8}$$

$$\eta_{\mathrm{r,t}} = o_{\mathrm{r,t}} + o_{\mathrm{s,t}} \tag{2.9}$$

$$\eta_{e,t} = o_{\mathrm{s,t}} \tag{2.10}$$

通过上述步骤，对特定子任务有关的信息被得以保留，而对特定子任务无关的信息则被过滤，而 η_s 则作为两个子任务之间的桥梁，保证二者直接的特征交互。 之后，我们通过激活函数从对应分区中提取所需特征：

$$\mu_{\mathrm{s,t}} = tanh(\eta_{s,\mathrm{t}}) \tag{2.11}$$

$$\mu_{e,t} = tanh(\eta_{e,t}) \text{-} \mu_{s,t} = tanh(\eta_{e,t}) \text{-} tanh(\eta_{s,t}) \tag{2.12}$$

$$\mu_{r,t} = tanh(\eta_{r,t}) \text{-} \mu_{s,t} = tanh(\eta_{r,t}) \text{-} tanh(\eta_{s,t}) \tag{2.13}$$

通过上述步骤，我们能够从领域文本中提取到三种类型的特征，即实体特征、关系特征和共享特征。但实验中我们发现这三部分特征具有较高的相关性，这意味着不同类型的特征之间可能存在着信息冗余或信息冲突。为了能够实现不同类型特征之间的相互隔离，我们设计了三种特征隔离方法，以此将不同类型的特征映射到不同的向量空间中，从而消除训练过程中可能存在的特征冗余或特征冲突：

$$\rho_1 = (\frac{\mu_e \cdot \mu_s}{|\mu_e| \cdot |\mu_s|})^2 + (\frac{\mu_r \cdot \mu_s}{|\mu_r| \cdot |\mu_s|})^2 \tag{2.14}$$

$$\rho_2 = \frac{1}{exp[dis(\mu_e, \mu_s)]} + \frac{1}{exp[dis(\mu_r, \mu_s)]} \tag{2.15}$$

$$\rho_3 = \{\frac{E[(\mu_e - \overline{\mu}_e)(\mu_s - \overline{\mu}_s)]}{\sigma(\mu_e) \cdot \sigma(\mu_s)} + \frac{E[(\mu_r - \overline{\mu}_r)(\mu_s - \overline{\mu}_s)]}{\sigma(\mu_r) \cdot \sigma(\mu_s)}\}^2 \tag{2.16}$$

其中，$|\mu|$为向量的模长，dis 为距离函数，本文采用的为欧式距离，σ 为方差，E 表示数学期望。通过上述的计算方法，我们能够保证每种方法最后的取值范围均在 0 到 2 之间，且取值越低，意味着不同类型的特征之间的冗余与冲突越少。

最后，我们使用这三部分特征来生成当前时刻的应当被保留的信息 f_t，进而得到当前时刻的隐藏状态 h_t，它们均将被输入到下一个时刻：

$$f_t = L(\eta_{e,t} || \eta_{r,t} || \eta_{s,t}) \tag{2.17}$$

$$h_t = tanh(f_t) \tag{2.18}$$

2.5 知识图谱构建任务层

在进行知识图谱构建任务之前，为了能够更好的捕获全局特征，使得模型在进行知识图谱构建任务的过程中能够有效的利用上下文信息，我们将每个时刻的特征进行融合，从而得到全局特征供模型使用：

$$\mu_e = maxpool\{tanh[L(\mu_{e,1} \| \mu_{s,1})], ..., tanh[L(\mu_{e,L} \| \mu_{s,L})]\} \tag{2.19}$$

$$\mu_{\mathrm{r}} = maxpool\{tanh[L(\mu_{\mathrm{r,1}} \parallel \mu_{\mathrm{s,1}})],...,tanh[L(\mu_{\mathrm{r,L}} \parallel \mu_{\mathrm{s,L}})]\} \quad (2.20)$$

我们的知识图谱构建模型以联合抽取方法为基础，具体可分为两个子任务，即实体抽取任务与关系抽取任务。对于实体抽取任务，我们的目标是识别领域文本中存在的所有实体及其类型。以新闻数据集为例，我们预先定义了一个实体类型集合 $\mathrm{T} = (\mathrm{people}, \mathrm{organization})$，对于每个类型 $\tau \in \mathrm{T}$，我们计算单词序列 $(\mathrm{s_i},...,\mathrm{s_j})$ 为 τ 类型实体的概率并记录为 $\mathrm{p_{ij}^{\tau}}$，具体计算方法如下：

$$\mu_{ij}^{e} = Elu[L(\mu_{e,i}||\mu_{e,j}||\mu_{e})] \quad (2.21)$$

$$p_{ij}^{\tau} = sigmoid\,[L(\mu_{ij}^{e})] \quad (2.22)$$

其中，μ_{ij}^{τ} 表示单词序列 $(\mathrm{s_i},...,\mathrm{s_j})$ 的特征，我们通过它来计算该序列为实体的概率。

对于关系抽取任务，我们的目标是找出实体及其关系。由于上一步已经成功从领域文本中抽取出实体，因此这部分仅需要抽取实体之间存在的关系。与实体抽取任务类似，一个预先定义的关系类型集合 $\mathrm{R} = (\mathrm{compete},...,\mathrm{collaborate})$，对于每个类型 $\gamma \in \mathrm{R}$，我们计算当以 $\mathrm{s_i}$ 为起始单词的实体作为主体，以 $\mathrm{s_j}$ 为起始的实体作为客体，它们之间关系为 r 的概率并记录为 $\mathrm{p_{ij}^{\gamma}}$，具体计算方法如下：

$$\mu_{ij}^{r} = Elu\,[L(\mu_{r,i}||\mu_{r,j}||\mu_{r})] \quad (2.23)$$

$$p_{ij}^{\gamma} = sigmoid\,[L(\mu_{ij}^{r})] \quad (2.24)$$

通过上述方式，模型能够计算出输入领域文本序列中任意子序列为实体的概率值，并能够进一步计算得出任意两个实体之间存在关系的可能性，为知识图谱构建任务奠定了基础。

2.6 损失函数

对于给定的训练集，我们模型的损失函数Loss由三部分组成，代表实体抽取任务损失函数的$Loss_e$，代表关系抽取任务损失函数的$Loss_r$和代表不同类型特征间冗余与冲突的$Loss_\rho$：

$$Loss_e = \sum_{p_{ij}^e \in E} BCELoss(p_{ij}^e, \widetilde{p}_{ij}^e) \tag{2.25}$$

$$Loss_r = \sum_{p_{ij}^\gamma \in R} BCELoss(p_{ij}^\gamma, \widetilde{p}_{ij}^\gamma) \tag{2.26}$$

其中。p_{ij}^e和p_{ij}^γ代表模型预测的实体与关系，而$\widetilde{p}_{ij}^e$和$\widetilde{p}_{ij}^\gamma$则代表数据集中真实存在的实体及关系，$Loss_\rho$的值则等于上文中相关性计算方法计算出的值，训练的目的是让损失值最低，因此此处我们将三个部分的损失值相加。

其中，我们通过为概率设置阈值的方式对模型的输出概率p进行分析，我们认为模型输出的概率值高于阈值ξ时表示模型预测该实体或关系存在，模型输出的概率值低于阈值ξ时则表示模型预测该实体或关系不存在。通过这一方式，我们的模型能够有效的完成知识图谱构建任务，并且可以能够解决领域文本中可能存在的三元组重叠的问题。

2.7 实验结果与分析

在本节我们通过实验来对本章提出的基于分区过滤网络的知识图谱构建模型进行评估。在2.7.1小节中我们对实验中涉及的数据集进行简单描述；在2.7.2小节则对知识图谱构建任务中常用的评价指标进行介绍，并给出实验的相关超参数；在2.7.3小节则对我们模型的性能进行对比与分析。

2.7.1 数据集描述

为了更好的评估我们模型的性能，我们在 NYT、WebNLG、ADE 三个公开数据集，新闻领域和百科领域的两个私有数据集上进行了实验验证。这五个数据集的具体情况如下：

WebNLG 数据集构建的初衷是为自然语言生成（Natural Language Generation，NLG）任务服务的，数据集中的三元组来源于 DBPedia，包括六个类别：宇航员、建筑、纪念碑、大学、运动队以及著作。该数据集包括 5019 条训练数据，500 条验证数据和 703 条测试数据，关系类型共 171 种。

NYT 数据集最开始被用于关系分类任务，其文本来源于《纽约时报》，数据集中的实体是通过斯坦福命名实体识别工具从文本中识别而来。其中，2005 年与 2006 年的新闻语料被用于构建训练集，2007 年的新闻语料则被用于构建训练集。该数据集包括 56195 条训练数据、4999 条验证数据和 5000 条测试数据，关系类型共 24 种。

ADE 数据集又名 Adverse Drug Effect，即药物不良反应语料库，该数据集通过文本的形式记录了药物及其不良反应，包含 3845 条训练数据和 427 条测试数据，关系类型共一种。

新闻领域的数据集是我们从互联网上的各大新闻网站中获取数据并构建的，这些新闻网站包括新华网、香港新闻网、人民日报、环球网和澎湃网。具体而言，我们通过爬虫技术从各个新闻网站中获得新闻语料，之后通过分句、去重和预先定义的规则库获得初始数据集；之后，我们通过人工复核的方式对数据集进行校正，最终形成新闻领域的数据集。该数据集包括 1849 条训练数据、200 条验证数据和 200 条测试数据，关系类型共 15 种。

百科领域的数据集则主要来源于百度百科，主要涉及的内容为各个领域的重要人物，如文学领域和艺术领域，其次还涉及这些人物的相关作品。我们通过对百度百科上的相关文本进行爬取、分句生成原始语料，并利用自适应模板与人工标注的方式构建该数据集。该数据集包括 55958 条训练数据、13417 条测试数据、11191 条验证数据，关系类型共 18 种。值得一提的是，该数据集

的训练集与验证集中大约有 35.7%的句子包含多个三元组，而在测试集这一比例则高达 54.6%，这意味着该数据集对模型的性能要求更高。

综上，本章所涉及的五个数据集中包含三个公共数据集（WebNLG 数据集、NYT 数据集和 ADE 数据集）以及两个实验室数据集（新闻数据集和百科数据集），其中 NYT 数据集与百科数据集为大规模数据集，WebNLG 数据集、ADE 数据集与新闻数据集则为小型数据集。这些数据集的数据分布情况如图 2.5 所示：

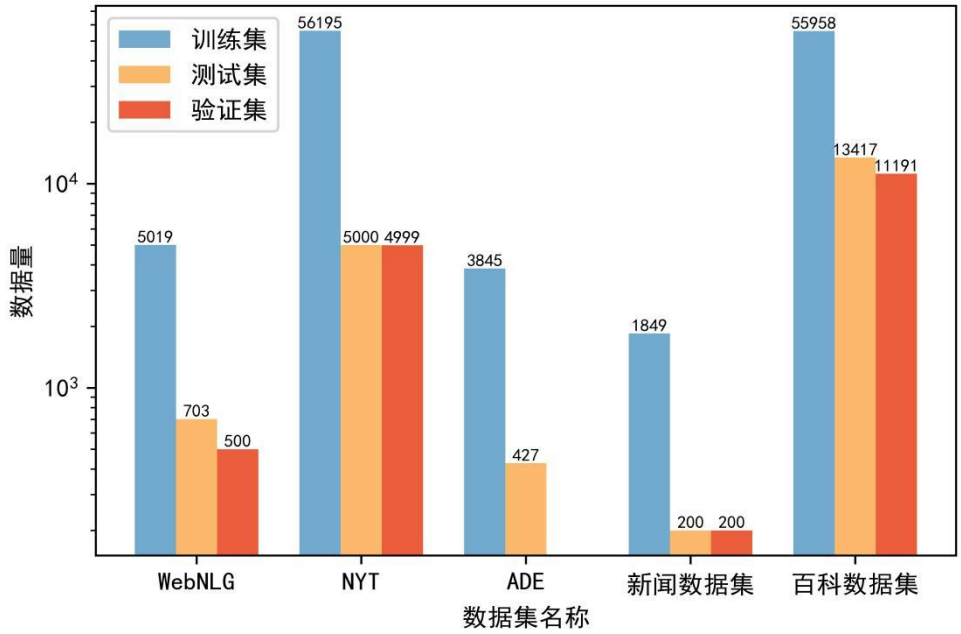

图 2.5　数据集数据分布一览，y 轴为 log 尺度。其中，NYT 数据集与百科数据集为大型数据集，WebNLG 数据集、ADE 数据集与新闻数据集为小型数据集

2.7.2　评价指标与参数设置

（1）评价指标

知识图谱构建任务是为了从文本中抽取出实体及关系，即三元组。在对模型进行评估的过程中，我们将测试集中真实存在的三元组个数记为 gold_num，而模型预测文本中存在的三元组个数记为 predict_num。对于预测结果，我们将模型预测的三元组与真实存在的三元组进行匹配，若实体和关系均正确，则认为预测正确，否则预测错误。通过这种方式，我们统计模型预测正确的三元

组个数，记为support_num，并对精确率（Precision，P）、召回率（Recall，R）和F1值这三个评价指标进行计算，具体计算方式如下：

精确率（Precision），该指标的含义为模型抽取的三元组中，抽取正确的三元组所占的比例，通过如下方式进行计算：

$$P = \frac{support\_num}{predict\_num} \tag{2.27}$$

召回率（Recall），该指标的含义为模型正确抽取的三元组在真实存在的三元组中所占的比例，通过如下方式进行计算：

$$R = \frac{support\_num}{gold\_num} \tag{2.28}$$

F1值（F-mean值），该指标为精确率与召回率的调和平均数，在多分类问题中被广泛应用，取值范围在0到1之间，通过如下方式进行计算：

$$F1 = \frac{2*P*R}{P+R} \tag{2.29}$$

（2）参数设置

在实验中我们使用到了一些超参数，主要包括学习率、训练epochs、相关性计算方法、dropout率、最大文本长度等。由于不同数据集的差异较大，因此在不同数据集上设置的超参数会略有差异，具体如表2.2所示：

表 2.2 超参数设置

超参数	**含义**	取值
lr	学习率	0.00002
epochs	训练轮次	100/200
embed_mode	编码方式	bert-base-cased
ξ	阈值	0.3/0.5
isolation	特征隔离方法	dis/pea/cos

2.7.3 结果分析

本小节主要包括对比实验结果分析和消融实验结果分析两个部分。其中，我们在 NYT 数据集与 WebNLG 数据集上进行了对比实验，以此证明本文提出的基于分区过滤网络的知识图谱构建模型的有效性。此外，还在 ADE 数据集、新闻数据集与百科数据集上进行了消融实验，用于分析我们的各个模块对模型性能的影响。

（1）对比实验

为了验证我们提出方法的有效性，我们在应用最为广泛的两个公共数据集 NYT 和 WebNLG 上进行了对比实验。同时，我们选取了 10 个模型作为对比实验的基线模型，以此充分证明我们方法的优越性。这 10 个基线模型分别是：GraphRel[19]、MHSA[59]、CasRel[14]、TPLinker[60]、SPN[61]、CGT[62]、CasDE[63]、RIFRE[64]、PRGC[65]和 TRN[66]。这些模型在知识图谱构建任务上均取得了较好的效果，并且具有一定的代表性。对比实验的结果如表 2.3 和表 2.4 所示：

表 2.3 WebNLG 数据集对比实验结果

Model	P	R	F1
GraphRel[19]（2019）	44.7	41.1	42.9
MHSA[59]（2021）	89.5	86.0	87.7
CasRel[14]（2019）	93.4	90.1	91.8
TPLinker[60]（2020）	91.8	92.0	91.9
SPN[61]（2020）	93.1	93.6	93.4
CGT[62]（2021）	92.9	75.6	83.4
CasDE[63]（2021）	90.3	91.5	90.9
RIFRE[64]（2021）	93.3	92.0	92.6
PRGC[65]（2021）	94.0	92.1	93.0
TRN[66]（2022）	93.5	92.7	93.1
Proposed	**94.2**	**94.1**	**94.2**

表 2.3 显示了我们的模型与基准模型在 WebNLG 数据集上的对比实验结果，实验结果表明我们的模型相较于基线模型在精确率、召回率和 F1 值上均取得了显著的提升，相较于 2022 年提出的 TRN 模型，我们的模型在这三个指标上分别提升了 0.7%、1.4%和 1.1%，这充分的证明了我们所提出的基于分区过滤网络的知识图谱构建模型的有效性。

表 2.4 NYT 数据集对比实验结果

Model	P	R	F1
GraphRel[19]（2019）	63.9	60.0	61.9
MHSA[59]（2021）	88.1	78.5	83.0
CasRel[14]（2019）	89.7	89.5	89.6
TPLinker[60]（2020）	91.3	**92.5**	91.9
SPN[61]（2020）	93.3	91.7	92.5
CGT[62]（2021）	**94.7**	84.2	89.1
CasDE[63]（2021）	90.2	90.9	90.5
RIFRE[64]（2021）	93.6	90.5	92.0
PRGC[65]（2021）	93.3	91.9	**92.6**
TRN[66]（2022）	93.0	92.3	**92.6**
Proposed	92.3	**92.5**	92.4

与此同时，从表 2.4 中我们可以发现，我们模型在 NYT 数据集上取得了不错的效果，但仍与部分模型存在轻微的差距。为了探究这一现象出现的原因，我们对 NYT 数据集与 WebNLG 数据集的特性进行了深入分析。最终，我们认为出现这一现象的原因是 WebNLG 数据集与 NYT 数据集在数据质量上存在显著的差异。上文中提到，NYT 数据集来源于《纽约时报》，并且是通过远程监督的方式产生的。该方式虽然能够省去人工标注的成本，但是却使得数据的质量难以得到保证。而在我们的模型中，不同子任务间信息的相互传递会使得这种误差被进一步放大，因此导致了这一现象的出现。而 WebNLG 数据集的数据质量较高，不存在上述问题，因此在该数据集上我们的模型表现优异，明显高于其他基线模型。

（2）共享特征消融实验

为了进一步探究我们设计的分区过滤网络是否能够对模型的性能产生提升，我们进一步设计了三组实验对该方法进行了验证，它们分别为“Proposed”、“Proposed *w/o share*”和“Proposed *w overlap*”，为了能够更好的阐明我们的实验方法，我们将结合图 2.6 对我们的方法进行详细描述：

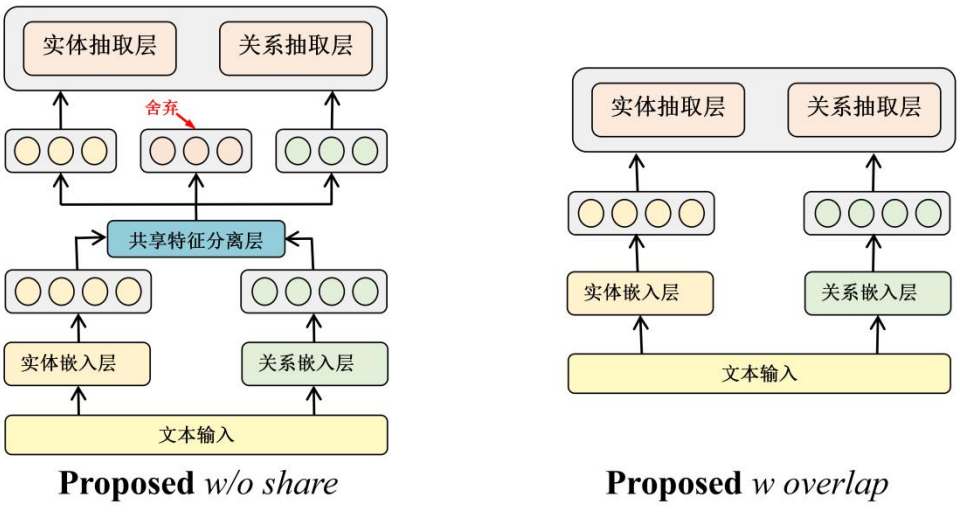

图 2.6　共享特征消融实验模型结构图，通过移除模型中的部分模块验证方法的有效性

第一组实验为“Proposed *w/o share*”，在该实验中，我们通过实体嵌入层与关系嵌入层从文本信息中提取得到实体特征与关系特征，之后，我们使用共享特征分离层识别并舍弃实体特征与关系特征中重合的部分，即共享特征。通过这一方式，能够得到特定于实体抽取任务的实体特征和特定于关系抽取任务的关系特征，而对实体抽取任务与关系抽取任务均有效的共享特征则被舍弃。之后，将这些特征用于进行知识图谱构建任务，该任务包含实体抽取任务与关系抽取任务。

第二组实验为“Proposed *w overlap*”，在这组实验中，我们以实体嵌入层与关系嵌入层为基础捕获实体特征与关系特征，但不对二者之间共同拥有的特征进行分离与删除，共享特征被包含在实体特征与关系特征之中。相较于第一组实验所使用的方式，共享特征被包含在了实体特征与关系特征之中，并未被舍弃，但由于其未被分离出来，因此可能会导致模型在进行知识图谱构建任务时难以有效利用这部分信息。

第三组实验为我们最终提出的方法，在此记作“Proposed”，其具体方法与第一组实验类似，区别在于我们将分离出的共享特征一同作为知识图谱构建任务的输入，以此使得模型能够更好的进行信息交互，并在此基础上引入了特征隔离方法避免不同类型特征之间的相互冲突。设计完消融实验的具体方法后，我们在公共数据集 ADE 与实验室的新闻数据集上进行了实验，实验结果如表 2.5 所示：

表 2.5 消融实验结果，实验结果表明了共享特征分离层的有效性

Dataset	Model	P	R	F1
ADE 数据集	Proposed *w/o share*	78.30	**85.96**	81.95
	Proposed *w overlap*	82.02	83.44	82.72
	Proposed	**82.15**	85.65	**83.86**
新闻数据集	Proposed *w/o share*	**97.99**	73.00	83.67
	Proposed *w overlap*	90.56	81.50	85.79
	Proposed	87.24	**85.50**	**86.36**

实验结果显示，在新闻数据集上，即使不将实体特征与关系特征中共同包含的共享特征进行分离，而是将它们一同输入至任务层中，模型的性能较舍弃共享特征相比也具有明显的提高，F1 得分上涨了 2.12%，而进一步将共享特征输入至模型中后，模型的 F1 得分再次提高了 0.57%。在 ADE 数据集上的实验结果也与之类似，模型的 F1 得分分别提高了 0.77%与 1.14%。这一实验结果表明共享特征对知识图谱构建任务具有显著的作用，能够有效的提升模型在实体抽取与关系抽取上的性能，同时将共享特征分离出来并单独输入至任务层中能够让模型更有效的利用这部分特征。

与此同时，我们注意到，在新闻数据集上随着我们逐渐突出模型中共享特征的作用，虽然模型的召回率在逐渐上升，但是其精确率却逐步下降，最终使得模型的 F1 得分稳步上升。我们认为这是因为随着两个子任务间有效信息交互的不断提升，模型在抽取三元组时考虑的愈发全面，因此导致这一现象的出现。

对于上述实验结果，我们的结论如下：对于完全不包含共享特征的模型而言，由于该方法子任务间的信息交互被人为的去除，这使得在训练过程中不同

子任务是完全独立的，因此使得模型的性能较差；而对于第二组实验，虽然不同子任务间通过实体嵌入层与关系嵌入层存在部分交互，但特定于任务的特征与共享特征未能被有效区分开，使得模型在难以充分这部分信息，从而一定程度上限制了模型的性能；而第三组实验则通过共享特征分离层从文本信息中获取到了三个类型的特征，并通过共享特征实现不同子任务间的有效信息交互，使得模型能够充分的利用各种类型的信息，从而性能最好。总的而言，实验结果进一步证明了我们方法的有效性。

（3）特征隔离方法消融实验

上一步我们通过实验证明了引入共享特征能够有效的提升模型在各个数据集上的表现，这一部分我们则尝试探索引入特征隔离方法能否有效的降低不同类型的特征之间的冗余与冲突并提升模型的性能，以及不同的特征隔离方法对模型的性能影响是否存在区别，如果存在区别，何种特征隔离方法对模型的性能提升最为明显。

表 2.6　消融实验结果，结果显示无论通过何种特征隔离方法对不同分区的特征进行区分，均能有效提升模型的性能。

Dataset	Model	P	R	F1
ADE 数据集	Proposed-NONE	82.12	83.28	82.69
	Proposed-PEA	81.08	85.17	83.08
	Proposed-DIS	81.87	85.49	83.64
	Proposed-COS	**82.15**	**85.65**	**83.86**
新闻数据集	Proposed-NONE	**96.82**	76.00	85.15
	Proposed-PEA	87.24	**85.50**	**86.36**
	Proposed-DIS	94.01	78.50	85.56
	Proposed-COS	95.09	77.50	85.40
百科数据集	Proposed-NONE	74.53	47.85	58.28
	Proposed-PEA	74.60	48.37	58.69
	Proposed-DIS	67.55	**54.35**	60.24
	Proposed-COS	**76.92**	49.52	**60.25**

为此，我们在 ADE 公共数据集、新闻数据集和百科数据集上分别进行了四组实验，从而对我们的猜想进行验证。其中，Proposed-None 仅通过共享特

征分离层将从实体特征与关系特征中分离出二者共有的共享特征，并直接输入至模型中用于知识图谱构建任务；Proposed-PEA、Proposed-DIS、Proposed-COS 则是在分离出共享特征的基础上分别采用上文中 ρ_1、ρ_2、ρ_3 的特征隔离函数将不同类型的特征映射到不同的向量空间之中，进而再输入至模型中。具体的实验结果如表 2.6 所示。

在三个数据集上的实验结果显示，无论引入何种特征隔离方法，模型的性能较不引入任何特征隔离方法时均有显著提升，这意味着将不同类型的特征映射到不同的向量空间中是有效的。通过对不同分区特征进行相互隔离，在 ADE 数据集上我们的模型的 F1 得分上升了 1.17%，其他指标也均有提升；在新闻数据集上，我们的模型的 F1 得分上升了 1.21%；在百科数据集上则上升了 1.97%。这在三个数据集上进行的实验说明通过引入特征隔离方法能够有效的避免不同类型特征之间的冲突与冗余，从而提高模型的性能。并且，在所有特征隔离方法中，以余弦函数为基础设计的特征隔离方法对模型性能的提升最大。

2.8 本章小结

本章主要研究如何从非结构化领域文本中抽取得出实体对与实体对之间的关系，最终形成该领域的知识图谱。为了能够从海量非结构化领域文本中挖掘领域知识，从而向用户与下游应用提供高度结构化的知识，必须通过知识图谱构建技术对领域文本进行挖掘。现有的基于神经网络的方法可以分为基于顺序编码的联合抽取技术和基于联合编码的联合抽取技术。其中，基于顺序编码的联合抽取技术按一定次序提取各个子任务所需的特征，但因缺乏子任务间特征的双向交互而限制了模型的性能；基于联合编码的联合抽取方法同时编码实体抽取子任务和关系抽取子任务所需的特征，但在这一过程中则存在不同子任务间特征交互弱或冗余信息过多的问题，为模型的性能带来影响。

面对上述难点问题，本文研究基于层次特征表示的知识图谱构建方法。具体而言，本文提出了基于分区过滤网络的知识图谱构建模型，首先通过 BERT 模型获得高质量的文本嵌入；之后，利用实体嵌入层与关系嵌入层将文本嵌入

信息划分为实体信息与关系信息；最后，从共享特征分离层从中提取出实体特征、关系特征与共享特征，并利用特征隔离方法将它们映射到不同的向量空间中以消除冲突与冗余，从而实现不同任务间有用特征的交互与无用特征的隔离。本章实验部分，我们在五个数据集上进行了广泛的对比实验与消融实验，从而对我们方法的有效性进行了评估与分析。

综上所述，我们的模型能够有效的从领域文本中抽取出实体及关系三元组，并通过知识图谱对这部分有用信息进行结构化的表达与存储，为后续工作奠定数据基础。

第三章 面向长尾分布样本的变分异构静态推理方法

3.1 静态推理问题的提出与定义

3.1.1 问题的提出

在上一章节中，本文详细介绍了基于层次特征表示的知识图谱构建方法，并通过该技术成功构建了包含丰富知识的知识图谱，为下游应用奠定了坚实的数据基础。然而，由于现实世界的复杂多样，从领域文本中构建的知识图谱往往伴随着连通性不足的缺陷，这使得实体间的关联信息常常存在缺失，这限制了知识图谱质量的提升，并导致在实际应用的过程中，各个应用难以对实体间存在的丰富语义信息进行全面的分析与利用。同时，长尾分布现象的存在使得知识图谱中的大部分实体与关系出现频次较少，而少数实体与关系则占据了绝大部分数据，这进一步加大了对知识图谱分析与利用的难度。为了解决上述问题，研究者们设计了众多的模型对知识图谱中的关联信息进行深入的挖掘与分析，进而推理其中缺失的信息，以此解决图谱中存在的不完整性问题，这些模型被统称为静态推理模型。

静态推理模型的任务是通过对已有三元组集合组成的知识图谱深入分析，对知识图谱中存在的缺失信息进行推理，从而生成新的三元组以弥补知识图谱的不完整性。目前，现有的知识推理模型可分为如下三组：基于距离的模型、基于翻译的模型和基于图神经网络的模型。其中，基于距离的模型使用两个实体之间的语义距离来衡量三元组的合理性，这类方法虽然设计思路简单明了，为后续研究提供了良好的参考；基于翻译的模型则将每个三元组 $\tau_{ijk}=(e_i,r_k,e_j)$ 中的关系作为从头实体到尾实体的翻译操作，但是由于知识图谱中存在一对多关系与多对多关系，因此相关研究人员不得不设计更为复杂的翻译方法对知识图谱中的复杂关系进行建模；基于图神经网络的方法的提出则是因为知识图谱天然具有图结构，为了有效利用知识图谱中的图结构信息，研

究人员设计了一系列的图注意力机制[43][44][69]以更好地挖掘图谱中蕴含的丰富信息，并取得了不错的效果。

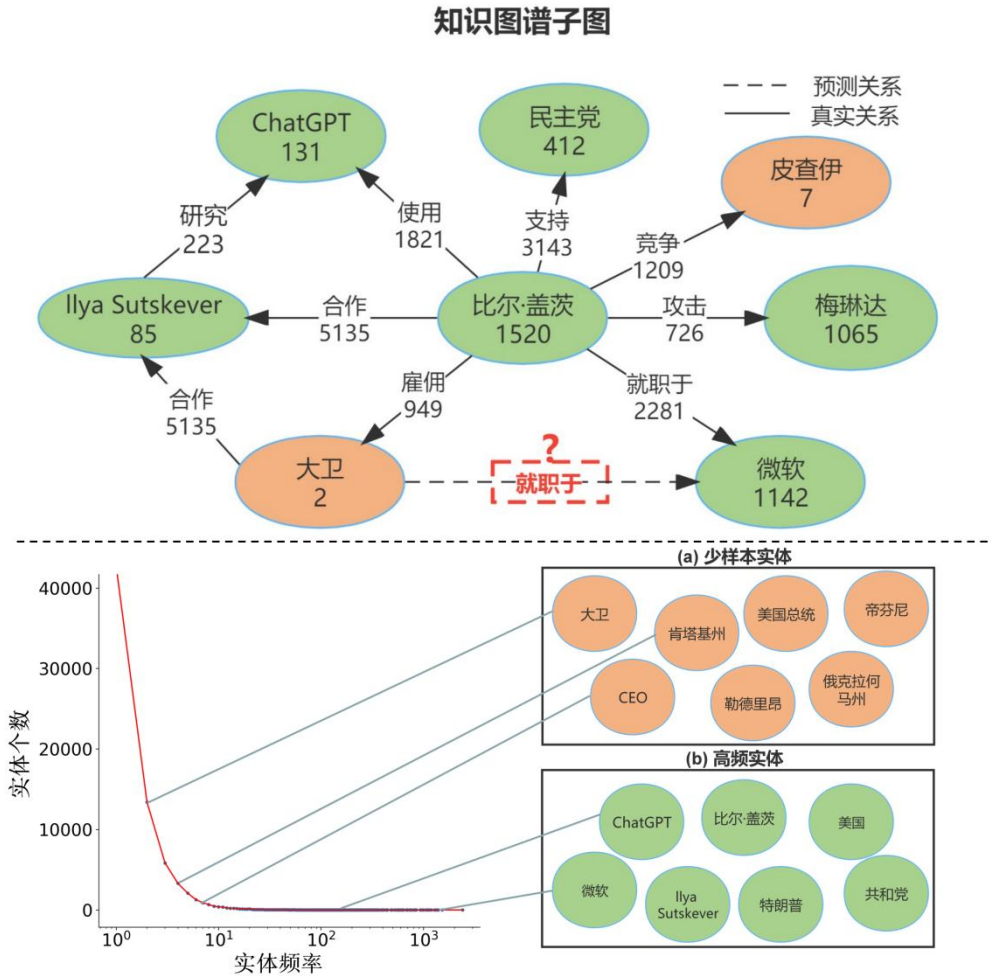

图 3.1　知识图谱子图与实体分布图。知识图谱子图中实体与关系下方的数字表示它们在整个知识图谱中出现的次数。图中表明大部分实体出现频次极低，而少量实体则出现频次极高。

然而，如图 3.1 所示，当我们利用图神经网络来学习实体和关系的表示时，我们观察到以下现象：第一，“合作”关系在该子图中与不同的头实体相连，而当其连接到“比尔盖茨”实体时显然比连接到“大卫”时更加重要，这是因为前者是一位举世闻名的人物。这一现象说明即使是相同的关系，当其被连接到不同的头实体时，它对中心实体的贡献也可能存在显著差异。第二，对于尾实体“ChatGPT”，当关系为“研究”时“ChatGPT”更能够体现中心实体的特性，因为 ChatGPT 的研究者比 ChatGPT 的使用者更能代表中心实体的特性，

这说明同一尾实体对中心实体的贡献可能因关系不同而发生变化。（iii）知识图谱中的实体和关系呈现长尾分布，由于缺乏足够的数据，模型很容易提取冗余特征。基于这些现象，我们总结了先前方法的三个缺点：

（1）基于距离的模型与基于翻译的模型未能充分利用图谱中存在的图结构信息，从而在特征提取的过程中忽略了很多有价值的内容，导致模型的性能受到限制。

（2）现有的模型在提取实体和关系的特征时，能够很好的提取出高频实体与高频关系的特征，但是在少样本实体与少样本关系上则表现较差。但是长尾分布效应导致了少样本实体与少样本关系普遍存在，因此现有模型的性能也受到了严重限制。

（3）基于图神经网络的模型在学习实体表示时，现有的方法往往会聚合与实体紧密相连的其他实体的信息，并设计各类算法计算这些邻域实体的权重，但忽略了相邻关系和相邻实体可以相互作用并共同确定当前中心节点的语义的事实。

因此，为了克服这些挑战，本文研究面向长尾分布样本的静态推理方法。具体而言，本文提出了一种新的变分异构注意力网络，该方法在聚合邻域信息的过程中动态计算邻居实体与关系的权重。此外，为了提高模型在长尾分布样本上的表现，本文在编码器和解码器之间引入了变分信息瓶颈模块，以此提取编码器输出信息中的深层次特征，加强模型的泛化性和与对长尾分布样本的表现能力。最终，将得到的高质量实体与关系向量表示用于静态推理任务中。本章节的主要贡献如下：

（1）本文通过信息瓶颈来控制编码器到解码器的信息流，它有助于模型更好地捕获潜在的有用的特征，提升了模型的泛化性和在长尾分布样本上的表现。

（2）我们设计了一种新的图注意力机制，在聚合相邻节点信息的过程中，根据三元组的特征动态调整关系和尾实体的权值，并让它们在更新节点嵌入时相互优化。

（3）在 FB15k-237、FB15k 和 WN18RR 数据集上的实验结果证明了我们方法的有效性。

3.1.2 问题的定义

在进行静态推理任务时，首先我们定义：通过知识图谱构建模型得到的三元组集合 T 形成了一张知识图谱 $G = \{E, R\}$，其中 $E = \{e_1, e_2, \cdots, e_n\}$表示知识图谱中出现过的所有实体构成的集合，$R = \{r_1, r_2, \cdots, r_n\}$表示实体之间存在的所有关系组成的集合。每个三元组 τ 由一个头部实体e_i、一个尾部实体e_j和它们之间的关系r_k组成。它可以正式表示为$\tau_{ijk} = (e_i, r_k, e_j)$。静态推理任务是指通过对已有图谱 G 的分析，推理出图谱中缺失的信息并组成新的三元组，新三元组中的实体与关系均来自于已有的实体集 E 和关系集 R。

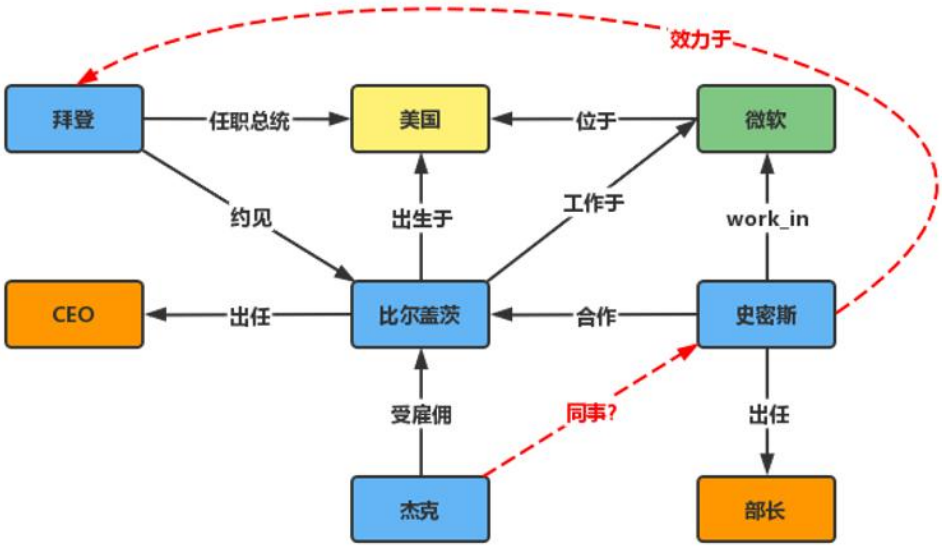

图 3.2 新闻领域中静态推理示例，图中实线表示已有知识图谱中存在的关系，虚线表示通过静态推理方法得出的关系。

图 3.2 展示了新闻领域中静态推理任务的示例，由于新闻文本中的知识有限，已构建的知识图谱中不可避免的存在缺失信息$\tau = (杰克，同事，？)$，而静态任务是通过分析已有知识图谱中的信息，对候选集中的实体进行排序，并选择最可能的实体作为三元组缺失信息的答案。

而在军事推演领域中，我们以实体表示作战单元，实体间的关系表示作战单元之间的行为，在不完全观测条件下，我方往往难以洞悉敌方所有作战单元的行为，因此静态推理任务旨在通过对当前作战形势的合理分析，挖掘并发现敌方作战单元潜在的行为并为决策者提供决策依据。

3.2 基于变分异构注意力网络的静态推理模型总体框架

基于变分异构注意力网络的静态推理模型（Variational Information Bottleneck for Heterogeneous Attention Network，ViHAN）采用的是编码-解码器模型架构（Encoder-Decoder）[70]。这是一种较为常见的模型架构，其特性是通过编码器获取实体与关系的高质量词嵌入，再通过解码器完成指定任务，是一种典型的端到端学习算法。

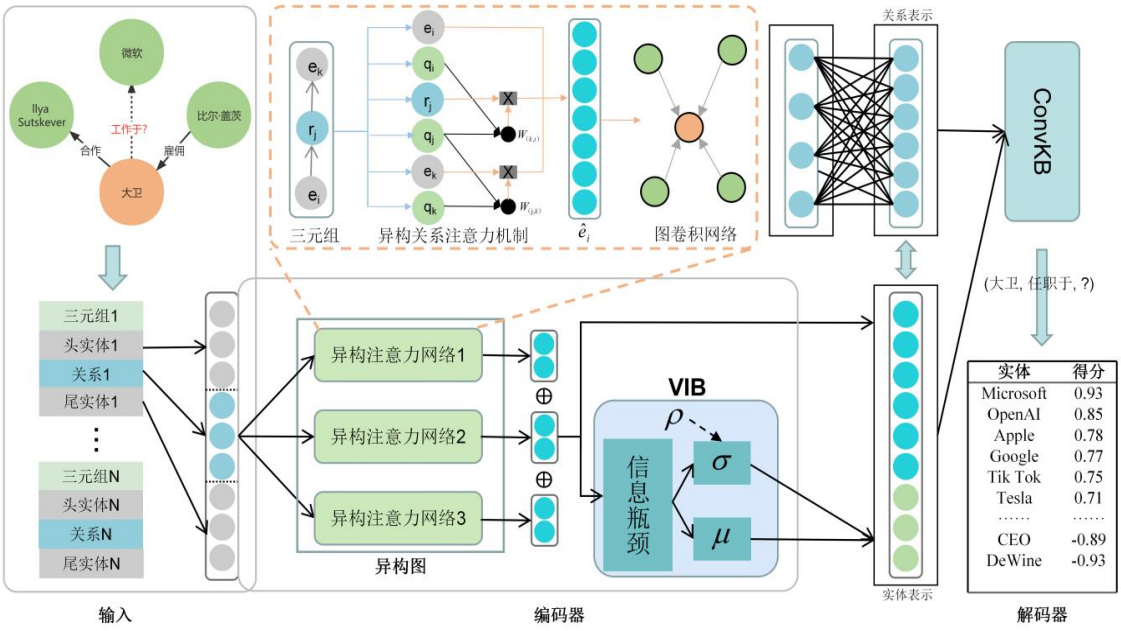

图 3.3 基于变分异构注意力网络的静态推理模型，通过异构注意力网络有效聚合来自邻域的信息，并利用信息瓶颈提高模型在少样本实体与少样本关系上的性能。

具体来说，对于知识图谱中的每个三元组，我们首先通过异构注意力机制分别计算尾实体和关系的重要性程度，并根据计算结果对中心实体的邻域信息进行聚合，同时结合实体本身的语义信息，以获得最终的语义表示。为了让模型能够更好的提取少样本的特征，本文引入变分信息瓶颈来对特征进行过滤并

添加符合高斯分布的干扰信息，从而提升模型对少样本实体与少样本关系的表示能力。最后，借助 ConvKB[40]分析三元组的全局嵌入属性，并对缺失的信息进行预测，图 3.3 详细阐述了该模型的工作流程。

我们将在下文中介绍基于变分异构注意力网络的静态推理模型的设计思路与实现方式。主要分为如下两个部分：在 3.3 节中介绍模型的编码器部分，包括如何有效利用知识图谱中的图结构信息以生成高质量的实体与关系表示，以及如何利用变分信息瓶颈提高模型在长尾分布样本上的性能；3.4 小节则介绍了模型的解码器部分，即模型如何通过编码器输出的高质量特征进行静态推理任务。

3.3 特征编码器

编码器的任务是从知识图谱中学习到节点的隐藏特征，为解码器提供高质量的向量表示。该编码器由以下两个部分组成：1）异构注意力网络，该模块通过对三元组的分析，在学习过程中实时查询当前三元组中各个元素的重要性程度，之后依据查询到的重要性程度对中心实体的邻域信息进行语义增强，通过该部分模型能够有效提取高频样本的特征；2）变分信息瓶颈模块，该模块负责控制编码器与解码器之间的信息流，该部分能够有效提升模型捕获少样本特征的能力。其详细原理描述如下：

3.3.1 异构注意力网络

该模块的出发点来源于这样的一个常识：并非所有的邻居对于中心实体都同等重要，同一主体做出的不同关系应当被分配不同的注意力分数，同一主体的相同关系作用于不同的客体上时也应当被分配不同的注意力分数。基于如上考虑，我们为每个节点添加一个额外的查询向量q，该向量被查询各个节点在当前三元组 $\mathrm{t}=(\mathrm{e_i},\mathrm{r_k},\mathrm{e_j}\}$中的重要性。实体与关系的查询方法分别如下：

$$W_{(\mathrm{k},i)}=q_k^T q_i \tag{3.1}$$

$$W_{(\mathrm{j},k)}=q_j^T q_k \tag{3.2}$$

上述公式中，$W_{(k,i)}$ 表示当前三元组中关系 r_k 向头实体 e_i 查询到的重要性分数，该分数与实体与关系的种类均有关。类似地，$W_{(j,k)}$ 表示尾实体 e_j 向关系 r_k 查询到的重要性分数。通过上述方式，我们能够对结合三元组特性为邻居节点动态分配权重。在这一过程之后，本文将当前三元组中的邻域信息聚合到中心实体中以丰富其语义表示，如下所示：

$$\alpha_i = W^H \left[e_i \,\|\, relu(W_{(k,i)} r_k) \,\|\, relu(W_{(j,k)} e_j) \right] \tag{3.3}$$

其中，α_i 是融合了邻域信息的当前三元组的集成表示，W^H 为用于收集邻域信息的矩阵，激活函数最终选取了 relu 函数，“||”符号表示向量间的拼接操作。

通过上述方式，我们能够有效聚合单个三元组中蕴含的丰富信息，之后，我们需要将不同三元组的信息聚合到相应的实体表示之中。由于不同三元组的重要性也存在着差异，因此它们应当在计算过程中被给予不同程度的注意力分数。为此，我们首先计算了每个三元组的相对注意力得分 η_i，如下所示：

$$\eta_i = softmax[\sigma(W^R \alpha_i)] = \frac{exp[GELU(W^R \alpha_i)]}{exp[\sum_{\chi_i \in \chi} \sigma(W^R \alpha_i)]} \tag{3.4}$$

其中，σ 为激活函数，χ 表示图谱中的所有三元组所构成的集合，χ_i 表示头实体为 e_i 所有三元组。通过 softmax 函数、GELU 激活函数和权重矩阵 W^R 计算得出每个三元组的相对注意力分数 η_i 后，我们通过该分数对来自不同三元组的信息进行加权融合，通过这种方式，我们将所有的邻域信息融合到实体的表示之中，如下所示：

$$\hat{e}_i = BatchNorm(\sum_{\chi_i \in \chi} \eta_i \alpha_i) \tag{3.5}$$

其中，BatchNorm 为标准化函数，用于避免少样本实体因邻域信息过少而导致的不公平现象。通过上述方式计算得出实体表示 $\hat{e}$ 后，为了能够获得更加

丰富的语义表示信息，我们设置了多个异构注意力网络以提取更加多元的特征，并将来自于不同注意力头的输出进行拼接，计算过程如下：

$$\hat{e}'_i = \frac{1}{M} \|_{m=1}^{M} BatchNorm(\sum_{\chi_i \in \chi} \hat{\eta}_i^m \alpha_i^m) \tag{3.6}$$

其中，M 是我们使用注意力头的个数。由于上述拼接操作导致关系的维度与实体的维度不一致，为了能够让关系的维度与实体的维度保持一致，我们通过一个矩阵对关系的表示进行了线性变换操作，如下：

$$\hat{r}^f = rW^r \tag{3.7}$$

通过这样的方式，实体的语义表示中已经蕴含了丰富的邻域信息，但是它依然缺乏来自自身的语义信息。为此，我们通过权重矩阵 W^e 将来自邻域的信息与实体自身的信息进行加权融合，具体过程如下：

$$\hat{e}^f = eM^e + \hat{e}' \tag{3.8}$$

3.3.2 变分信息瓶颈模块

由于通用特征提取器不可避免地会提取与目标任务无关的特征，这一现象在训练集中出现次数较少的少样本实体与少样本关系上尤为明显。为了解决这一问题，我们引入该变分信息瓶颈模块对来自编码器的信息流进行控制，其目的在于消除特征提取过程中的不相关特征和冗余信息，以此提高模型在少样本实体和少样本关系上的表现。该模块首先对实体和关系的表示进行压缩，并通过两个线性层和两个激活函数提取它们的潜在特征。之后，我们使用神经网络拟合输入向量的均值和方差，并通过均值与方差对其进行重构，在重构的过程中剔除冗余信息并添加高斯噪声提升模型的鲁棒性，如下所示：

$$\mathrm{E} = [\mu(\hat{e}^f) + \rho \Sigma(\hat{e}^f)] \| \hat{e}^f \tag{3.9}$$

$$\mathrm{R} = [\mu(\hat{r}^f) + \rho \Sigma(\hat{r}^f)] \| \hat{r}^f \tag{3.10}$$

其中，ρ 表示服从高斯分布的噪声信息，μ 和 Σ 分别表示均值和方差，它们输出的维度是输入的四分之一。E 和 R 是重构后的实体和关系的向量表示。

通过上述方式，我们一定程度上剔除了实体与关系中的冗余特征，从而学习到了更加富有表现力的实体与关系表示。

3.3.3 训练目标

对于编码器的训练过程，我们主要参考了 TransE[24]的思想，希望三元组 $(e_i, r_k, e_j) \in \chi$ 能够满足 $\vec{e}_i + \vec{r}_k \approx \vec{e}_j$，因此我们设计了如下的打分函数：

$$s_{\tau_{ijk}} = \| E_i + R_j - E_j \| \tag{3.11}$$

其中，$\tau_{ijk} = (e_i, r_k, e_j)$ 表示三元组集合中的任意一条三元组记录，包含事实三元组样本与非事实三元组样本。对于事实三元组样本，我们希望 $s_{\tau_{ijk}}$ 的取值能够尽可能的低，而对于非事实三元组样本，则希望 $s_{\tau_{ijk}}$ 的取值尽可能的高。出于这样的目的，我们使用了 hinge 损失函数来训练我们的编码器，具体表现为：

$$L_G = \sum_{\tau'_{ijk} \in \chi'} \sum_{\tau_{ijk} \in \chi} \max(s_{\tau'_{ijk}} - s_{\tau_{ijk}} + \gamma, 0) \tag{3.12}$$

其中，χ' 表示非事实三元组样本集，χ 表示事实三元组样本集，γ 为超参数，用来保证损失函数不会出现负值。在训练的过程中，事实三元组的得分将逐渐趋向于 0，而非事实三元组的得分则将趋向于阈值 γ。

与此同时，在训练的过程中变分信息瓶颈模块也应当被加以限制，我们希望该模块能够尽可能多的保留与静态推理任务相关的信息，同时对输入信息进行压缩。因此，整体的损失函数被设计如下：

$$L = \beta KL[p_\theta(E \mid \hat{e}^f), r(E)] + \beta KL[p_\theta(R \mid \hat{r}^f), r(R)] + L_G \tag{3.13}$$

其中，β 为超参数，它的作用是保证两个部分的损失值接近相等，根据实验情况我们最终将该值设置为 1e-4。$p_\theta(R \mid \hat{r}^f)$ 表示在 R 已知的情况下 $\hat{r}^f$ 的后验概率的数学期望。r() 则表示 $p_\theta()$ 的数学期望。

3.4 长尾分布样本解码器

完成对编码器的训练之后，我们能够得到实体与关系的高质量词嵌入，之后便是训练解码器对实体的关系进行推理。在综合考量下，我们选取了ConvKB作为模型的解码器，该解码器通过应用卷积神经网络提取三元组上的不同特征，具体方式是通过多个滤波器对三元组进行特征提取，并以此为依据判断三元组的有效性。由于样本由事实三元组和非事实三元组组成，ConvKB的打分函数可以表示为：

$$F_D(\chi_i) = W \cdot \|_{m=1}^{M} ReLU(E_i, R_k, E_j) * \zeta_m \tag{3.14}$$

其中，ζ_m表示卷积层中滤波器的个数，该参数为超参数。我们将所有滤波器的输出结果进行拼接，并将其与矩阵$W \in R^{1\times MD}$进行点乘操作“$\cdot$”得到三元组的最终得分，最后使用Adam优化器训练解码器，具体损失函数为：

$$L_D = \sum_{\chi_i \in (\chi \cup \chi')} \log\{1 + \exp[l_\chi \cdot F_D(\chi_i)]\} + \frac{\lambda}{2} \| W \|_2^2 \tag{3.15}$$

上述表达式中，我们对权重矩阵W进行了L_2正则化运算，并且对于事实三元组样本，l_χ的取值为1，对于非事实三元组，其取值则为-1，其目的是使该模型能够有效地区分事实三元组与非事实三元组。

3.5 实验结果与分析

本节对上一节提出的基于变分异构注意力网络的静态推理模型进行实验评估与验证，并进一步探讨了该模型对少样本实体的有效性。其中，我们在3.5.1小节中对实验所使用的数据集进行了简单的介绍；在3.5.2小节介绍了静态推理任务，并介绍其评价指标与模型参数设置；在3.5.3小节对实验结果进行了展示与分析。

3.5.1 数据集描述

为了验证基于变分异构注意力网络的静态推理模型的有效性，我们在四个数据集上对模型的性能进行了验证。

我们选取了三个被广泛使用的大型基准数据集 WN18RR、FB15k-237 和 FB15k 以验证我们模型的优越性。其中，FB15k 是知识图谱 Freebase 的子集，包含 14951 个实体和 1345 种关系，FB15k-237 则是 FB15k 去除了反向关系的数据集，包含 14541 个实体和 237 种关系。WN18RR 则是 WordNet 的子集，具有 18 个关系和 40943 个实体。

新闻领域的数据集源自联合抽取章节中所形成的新闻数据集，该数据爬取于新华网、香港新闻网、人民日报、环球网和澎湃网等新闻网站，通过上文中设计的联合抽取模型从原始领域文本中得到结构化的知识图谱作为数据集，它描述了新闻领域中不同实体之间存在的复杂关系。由于从不同领域文本中可能抽取出相同的实体关系三元组，因此对这部分数据进行了去重处理，这导致该数据集的数据量略低于上一章节。通过上述方式最终形成的数据集包含 1420 个实体与 15 种关系。

表 3.1 数据集数据统计

数据集	实体集	关系集	训练集	验证集	测试集
WN18RR	40943	11	86835	3034	3134
FB15k-237	14541	237	272115	17536	20446
FB15k	14951	1345	483142	50000	59071
新闻数据集	1420	15	1224	145	150

从上述关于数据集的介绍中不难得出，平均而言，FB15k-237 数据集与 FB15k 数据集中实体之间的关系更多，而 WN18RR 与新闻数据集中实体之间的关系则相对稀疏。这意味着前者中的实体较后者中的实体有更多的邻域信息可以挖掘。各个数据集的详细描述如表 3.1 所示。

3.5.2 评价指标与参数设置

根据前文介绍，静态推理任务是为了预测实体之间可能存在的关系，具体方式为给定头实体与关系方式，通过打分函数计算该关系可能作用于哪个尾实体上。因此，在进行模型性能评估的过程中，研究者们往往会对实体集中的所有实体进行打分，并将得分最高的实体作为预测的结论。在训练过程中，我们将训练样本分为事实三元组样本与非事实三元组样本，其中事实三元组样本指训练集中存在的客观事实，而非事实三元组样本则是指通过替换事实三元组样本中的头实体或尾实体得到的负样本。在对模型进行评估时，则是将待预测的缺失实体依次填充为实体集中的所有实体进行打分，并根据不同实体的得分高低作为评判模型好坏的标准。本次实验中，我们采用了平均排名 Mean Rank(MR)、平均倒数排名 Mean Reciprocal Rank(MRR)、HIT@1、HIT@3 和 HIT@10 五个评价指标作为模型性能的评估标准，具体如下：

（1）MR

平均排名 MR 是指训练后的模型对事实三元组与非事实三元组进行打分时，事实三元组排名的平均值。通过上述定义不难得出，MR 值越小，说明模型预测的实体越接近正确答案，即模型的性能越好，通过如下方式计算：

$$MR = \frac{1}{|S|}\sum_{i=1}^{|S|} rank_i \tag{3.16}$$

其中，$|S|$是测试集中三元组的个数，$rank_i$ 表示第 i 个三元组的推理得分排名。

（2）MRR

平均倒数排名 MRR 与平均排名 MR 类似，是指事实三元组排名的倒数的平均值。因此 MRR 的取值范围在 0-1 之间，且该值越接近 1，说明模型的性能越好，其计算公式如下：

$$MRR = \frac{1}{|S|}(\sum_{i=1}^{|S|} \frac{1}{rank_i}) \tag{3.17}$$

（3）HIT@N

HIT@N 指的是所有样本中，正确答案的排名在前 N 的概率。对于测试集中的每一条数据，若正确答案的排名在前 N，则给对应的 HIT 计数加 1，最终将 HIT 的计数值除以测试集样本个数作为 HIT@N 的值。同 MRR，该值取值范围在 0-1 之间，越接近 1，说明模型的效果越好，可通过如下方式得到：

$$HIT@N = \frac{1}{|S|}\sum_{i=1}^{|S|} I(rank_i \le N) \tag{3.18}$$

其中，本次实验选取 N=1、3、10 作为实验结果的评价指标，I 表示 indicator 函数，当传入条件为真时取值为 1，否则取值为 0。

在实验中我们使用到了一些超参数，主要包括学习率、编码器训练 epochs、解码器训练 epochs、训练过程中正负样本比值等，具体如表 3.2 所示：

表 3.2 超参数设置

超参数	含义	取值
lr	学习率	0.001
Encoder-epochs	编码器训练轮次	3600
Decoder-epochs	解码器训练轮次	200
n-head	注意力头个数	2、3

3.5.3 结果分析

为了验证我们模型的有效性，我们在公共数据集 WN18RR 和 FB15k-237 上进行了对比实验与消融实验，并与其他八个基线模型（DistMult[31]、ComplEx[32]、ConvE[39]、RotatE[71]、g-GAT[42]、ParamE-Gate[72]、DualE[73]、D-AEN[45]）进行对比。

同时，为了验证各个模块的有效性，我们进行了消融实验探究了各个模块对模型性能的影响。同时，为了进一步证明变分信息瓶颈模块能够有效的提升模型的鲁棒性与泛化性，我们在涉及少样本实体与少样本关系的数据上进行了进一步的实验。最后，我们通过超参数实验探究了注意力头个数的不同对模型性能的影响，

（1）对比实验结果分析

表 3.3 FB15k-237 数据集对比实验结果

Methods	MR	MRR	Hit@1	Hit@3	Hit@10
DistMul[31]（2014）	512	0.199	28.1	30.1	44.6
ComplEx[32]（2016）	546	0.194	2.78	29.7	45.0
ConvE[39]（2018）	245	0.225	31.2	34.1	49.7
RotatE[71]（2019）	177	0.338	24.1	37.5	53.3
g-GAT[42]（2019）	210	0.518	46.0	54.0	62.6
ParamE-Gate[72]（2020）	-	0.399	31.0	43.8	57.3
DualE[73]（2021）	**91**	0.365	26.8	40.0	55.9
D-AEN[45]（2023）	164	0.429	33.7	47.1	61.1
Our work	169	**0.567**	**50.8**	**59.4**	**67.8**

在公共数据集 FB15k-237 和 FB15k 上的实验结果如表 3.3 和表 3.4 所示。实验结果显示我们的模型在 MRR、Hit@1、Hit@3、Hit@10 四个指标上均优于基线模型，虽然在 MR 指标上次于 DualE 模型，但也取得了不错的效果。这表明我们设计的基于变分异构注意力网络的静态推理模型是有效的。

表 3.4 FB15k 数据集对比实验结果

Methods	MR	MRR	Hit@1	Hit@3	Hit@10
DistMul[31]（2014）	42	0.798	-	-	89.3
ComplEx[32]（2016）	-	0.692	59.9	75.9	84.0
ConvE[39]（2018）	51	0.657	55.8	72.3	83.1
RotatE[71]（2019）	32	0.699	58.5	78.8	87.2
g-GAT[42]（2019）	33	0.868	82.4	90.2	93.5
DualE[73]（2021）	**21**	0.813	76.6	85.0	89.6
Our work	29	**0.896**	**86.7**	**91.7**	**94.1**

此外，表 3.5 展示了在公共数据集 WN18RR 上的对比实验结果。从表中不难看出，我们的模型在 MR 和 Hit@10 指标上显著优于其他基准模型，在其他指标上则次于部分模型。同时，我们注意到在 FB15K-237 上，DualE 模型的

Hit@1 得分比我们的模型低约 20%，但在 WN18RR 上则比我们的模型高 7%左右。通过对这两个数据集的深入分析，我们认为出现这种差异的原因在于 FB15k-237 较 WN18RR 拥有更加丰富的邻域信息，而 WN18RR 则较 FB15k-237 则拥有更加明显的层次关系。因此，强调学习实体层次关系的 DualE 等模型在 WN18RR 数据集上表现更好，而我们使用邻域信息的模型则在 FB15K-237 上表现良好。

表 3.5 WN18RR 数据集对比实验结果

Methods	MR	MRR	Hit@1	Hit@3	Hit@10
DistMul[31]（2014）	7000	0.444	41.2	47.0	50.4
ComplEx[32]（2016）	7882	0.449	40.9	46.9	53
ConvE[39]（2018）	4464	0.456	41.9	47.0	53.1
RotatE[71]（2019）	3340	0.476	42.8	49.2	57.1
g-GAT[42]（2019）	2037	0.429	36.1	47.7	57.2
ParamE-Gate[72]（2020）	-	0.489	46.2	50.6	53.8
DualE[73]（2021）	2270	**0.492**	**44.4**	51.3	58.4
D-AEN[45]（2023）	2248	0.484	44.3	50.0	56.1
Our work	**1447**	0.474	39.8	**51.7**	**60.9**

为了证明这一观点，我们统计了三个公共数据集中各个实体平均拥有的一阶近邻实体个数与二阶近邻实体个数，其统计结果如表 3.6 所示。数据显示，FB15k-237 平均每个实体拥有 18.71 个一阶近邻实体，1684.75 个二阶近邻实体，FB15k 与之类似，各个实体也包含了丰富的邻域信息。相对而言，WN18RR 平均每个实体仅有 2.12 个一阶近邻实体，6.64 个二阶近邻实体。这种差异正是造成模型在不同数据集上实验结果存在差异的原因。

表 3.6 邻域信息统计情况，WN18RR 数据集的邻域信息明显低于其他数据集

数据集	实体数	一阶邻域（平均）	二阶邻域（平均）
FB15k-237	14541	18.71	1331.60
FB15k	14951	32.32	5800.87
WN18RR	40943	2.12	6.42

为了进一步证明我们的方法的有效性，我们在 FB15k-237 数据集中选择了一些在训练集中出现相对较少的少样本关系，并将测试集中与这些少样本关系相连的实体进行可视化。因为 WN18RR 数据集仅有 11 种关系，所以并未进行此项对比实验。我们模型的可视化效果与 TransE 的可视化效果如图 3.4 所示。通过二者对比不难发现，我们的模型可以在少样本关系背景下更好的对实体进行区分，如对于相似实体（与“sports team location”和“location of ceremony”相关的实体），我们模型相较于 TransE，明显能够更好的对这两类实体进行区分，我们认为导致 TranE 无法很好区分这类实体的原因是“俱乐部”等实体在某些情况下也可以表示地理概念，因此其无法很好地将这类实体与“庆典位置”相区分。总的来说，实验结果证明了我们的模型能够学习到更高质量的表示，因此能够很好的对 few-shot 关系下的相关实体进行区分。

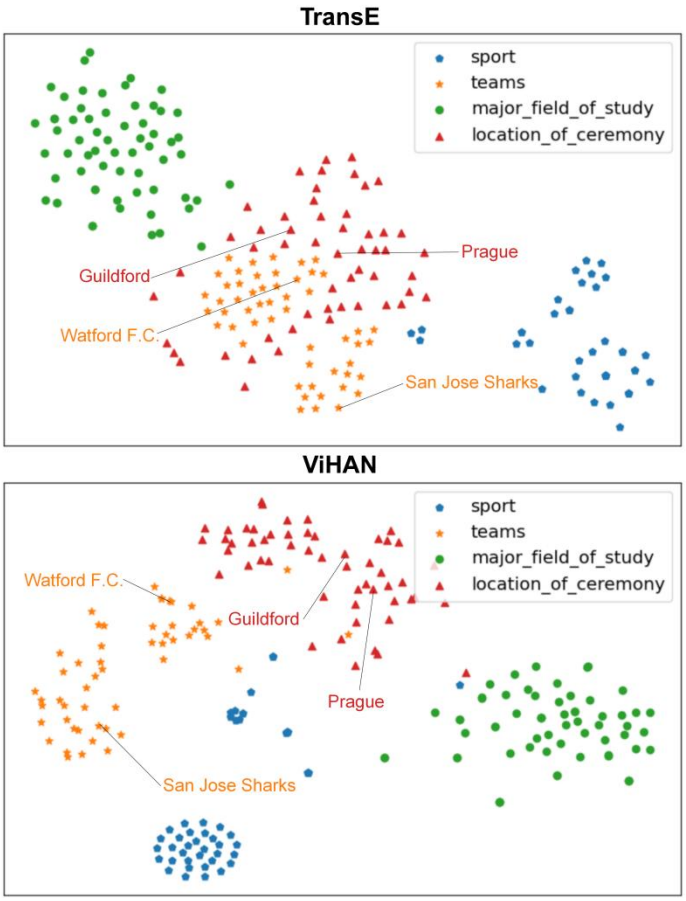

图 3.4 TransE 与 ViHAN 实体表示可视化，图中显示 TransE 模型难以有效区分少样本关系下的各个实体，而我们的模型则能够有效区分这部分实体。

（2）消融实验结果分析

表 3.7 FB15k-237 数据集消融实验结果，结果显示变分信息瓶颈模块能够有效提升模型在少样本实体上的性能

Methods	Tail	Head	MR	MRR	Hit@1	Hit@3	Hit@10
HAN	10	-	296	0.125	5.76	12.75	25.93
ViHAN	10	-	**280**	**0.190**	**13.53**	**17.28**	**33.74**
HAN	20	-	306	0.159	8.49	15.22	32.53
ViHAN	20	-	**242**	**0.215**	**14.10**	**21.96**	**38.94**
HAN	-	10	545	0.489	43.85	49.73	60.43
ViHAN	-	10	**285**	**0.537**	**47.86**	**56.15**	**64.93**
HAN	-	20	421	0.526	46.80	54.35	64.37
ViHAN	-	20	**235**	**0.577**	**51.77**	**60.17**	**69.14**

我们进行了消融实验，以探索变分信息瓶颈是否可以帮助模型捕获深层次的隐藏特征，并提高模型在少样本实体上的表现。为此，我们统计了训练集上各个实体出现的频次并对其进行排序，之后，将出现次数最少的 10%实体记为“tail-10”实体，出现次数最少的 20%实体记为“tail-20”实体。之后，通过这两类实体对测试集进行处理，从而分别探究这两类少样本实体出现在测试数据的头实体和尾实体上时模型的表现性能。

表 3.8 WN18RR 数据集消融实验结果，结果显示变分信息瓶颈模块能够有效提升模型在少样本实体上的性能

Methods	Tail	Head	MR	MRR	Hit@1	Hit@3	Hit@10
HAN	10	-	3447	0.107	6.18	10.55	20.73
ViHAN	10	-	**2907**	**0.147**	**9.81**	**15.27**	**24.36**
HAN	20	-	2263	0.310	25.57	32.51	42.73
ViHAN	20	-	**1958**	**0.345**	**28.21**	**37.46**	**45.21**
HAN	-	10	2242	0.282	16.86	**36.77**	44.96
ViHAN	-	10	**2130**	**0.340**	**26.23**	36.70	**47.78**
HAN	-	20	2023	0.354	27.07	39.82	50.73
ViHAN	-	20	**1886**	**0.417**	**35.35**	**44.11**	**53.94**

根据前文中的分析，变分信息瓶颈能够剔除无关特征与冗余特征，提高模型在少样本实体上的性能。表 3.7 和表 3.8 的实验结果证明了我们的推测。实验结果显示无论 few-shot 实体出现在头实体位置还是尾实体位置，引入变分信息瓶颈之后均能够有效的提升模型的性能，这一结果与我们的猜想一致，因此证明了我们方法的有效性。

（3）超参数实验结果分析

为了探究不同注意力头个数是否会影响模型捕获实体特征的能力，本文设计了注意力头个数的超参数实验。具体而言，基于变分异构注意力网络的静态推理模型利用多个异构注意力网络来充分捕获实体的不同类型特征，并将它们的输出结果进行融合，从而更好的学习实体的表示。在此，为了探究异构注意力网络的个数是否会对模型的性能产生显著影响，我们设计了该超参数实验并观察模型在各个指标上性能的变化，新闻数据集上的实验结果如图 3.5 所示：

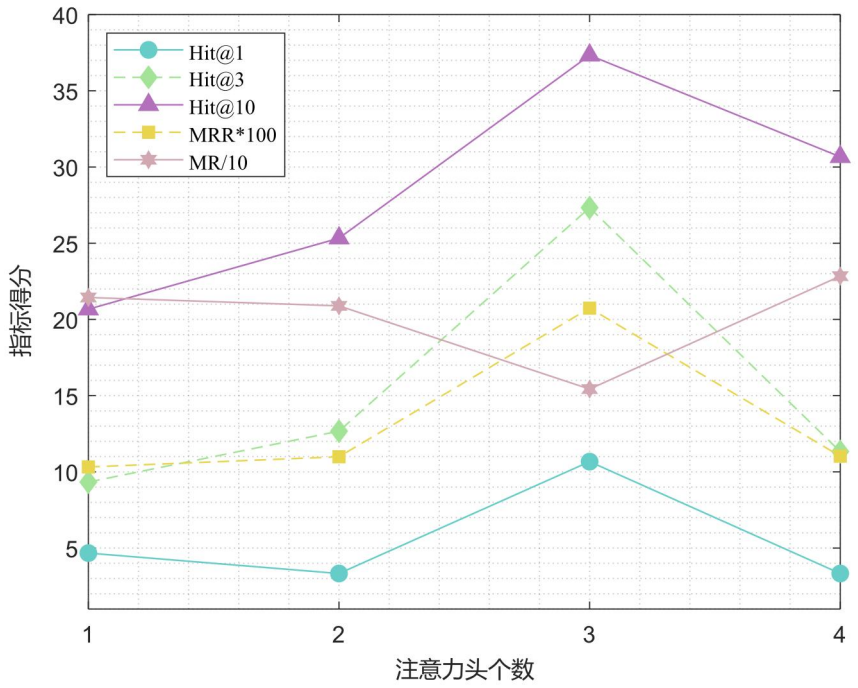

图 3.5　注意力头个数对于模型捕获实体特征能力的影响，其中 MRR 指标越低说明性能越好，其余指标越高说明性能能越好。

此外，我们还在公共数据集 FB15k-237、FB15k 和 WN18RR 三个公共数据集上进行了相同的实验，其实验结果如表 3.9 所示：

表 3.9 注意力头个数对于模型捕获实体特征能力的影响，其中 MRR 指标越低说明性能越好，其余指标越高说明性能能越好。

Dataset	Model	MRR	Hit@1	Hit@3	Hit@10
FB15k-237	HAN *w 1 head*	0.495	41.7	53.2	63.7
	HAN *w 2 head*	0.508	42.9	54.7	65.8
	HAN *w 3 head*	0.514	45.2	54.1	63.5
	ViHAN *w 1 head*	0.528	45.2	56.6	66.5
	ViHAN *w 2 head*	**0.567**	**50.8**	**59.4**	**67.8**
	ViHAN *w 3 head*	0.548	48.4	57.6	66.8
WN18RR	HAN *w 1 head*	0.439	36.5	48.4	57.2
	HAN *w 2 head*	0.434	35.2	47.7	58.7
	HAN *w 3 head*	0.449	36.5	50.0	60.0
	ViHAN *w 1 head*	0.459	37.9	50.4	59.9
	ViHAN *w 2 head*	**0.474**	**39.8**	**51.7**	**60.9**
	ViHAN *w 3 head*	0.462	38.4	50.8	59.9
FB15k	HAN *w 1 head*	0.826	76.2	87.9	92.4
	HAN *w 2 head*	0.886	85.9	90.6	92.7
	HAN *w 3 head*	0.890	86.0	91.4	93.9
	ViHAN *w 1 head*	0.853	80.3	85.2	92.8
	ViHAN *w 2 head*	0.889	85.9	91.1	93.7
	ViHAN *w 3 head*	**0.896**	**86.7**	**91.7**	**94.1**

在多个数据集上的实验结果充分的证明了注意力头的个数会显著影响模型的捕获实体特征的效果。随着注意力头个数的不断增加，模型的性能总体上呈现先上升后下降的趋势。我们认为导致这一现象的原因是随着注意力头个数的增加，模型能够更加捕获更多的实体特征，但当注意力头个数过多时则会出现过拟合现象，上述实验结果也验证了这一结论。

3.6 本章小结

本章针对知识图谱构建过程中遇到的信息缺失的问题，进行了针对性的静态推理研究。对于通过知识图谱构建模型获得的知识图谱而言，虽然其内部蕴含着丰富的知识，但仍然存在着关键实体信息缺失的问题，这对决策者准确把控行业发展动向并规避潜在风险带来困难。现有的基于图注意力机制的模型虽然能够很好的对高频样本缺失的信息进行推理，但是却难以推理少样本间的缺失信息，同时在聚合邻域信息的过程中考虑欠佳，这些因素均对模型的性能产生了影响。

基于上述问题，本文进行面向长尾分布样本的静态推理方法。具体而言，本文结合现有静态推理方法中存在的问题针对性的设计了基于变分异构注意力网络的静态推理模型。首先，在学习实体表示的过程中通过异构关系注意网络动态计算邻域信息的重要性，从而让编码器能够有效的聚合邻域信息，提升了模型对高频实体的表示能力；之后，通过变分信息瓶颈对编码器输出的特征进行控制，剔除冗余特征的同时引入高斯噪音，增强模型的泛化性与鲁棒性，有效提高了模型在少样本实体与少样本关系上的表现；最后，通过 ConvKB 模型对编码器的最终输出进行解码并执行静态推理任务。对比实验与消融实验的结果显示我们的方法不仅公共数据集上取得了优异的性能，更是在少样本实体与少样本关系上表现良好，这足以证明我们方法的有效性。

综上所述，我们提出的基于变分异构注意力网络的静态推理模型能够充分挖掘不同实体间潜在的关联关系，进而有效地推理关键实体间的潜在关系，在少样本实体与少样本关系上也取得了不错的效果，能够有效的对知识图谱的不完整性进行弥补。

第四章 面向零样本基于多层传播的动态推理方法

4.1 动态推理问题的提出与定义

4.1.1 问题的提出

上一章本文详细介绍了面向长尾分布样本的静态推理方法，通过对已有知识图谱的深入分析推理对知识图谱中的缺失信息进行补全，不仅在高频实体和高频关系上取得了卓越的效果，在少样本实体与少样本关系上也表现良好。之后，本文针对实际应用过程中可能出现会零样本实体的问题，研究面向零样本实体的知识图谱动态推理方法。

由于领域文本不断更新，快速扩增，在这样的背景下，模型在推理的过程中往往遇到的不全是训练时曾观测到的实体，这意味着在实际应用的过程中往往需要对原有知识图谱中不存在的实体的关系进行推理。这类在训练时未曾观测到的实体被称为零样本实体（zero-shot entity）。由于在实际应用过程中已有知识图谱难以涵盖所有的实体，因此这一问题在应用过程中十分普遍。为了解决零样本实体问题，研究人员设计相关算法利用有限的辅助集信息计算零样本实体的向量表示并进行推理，这些方法被称为动态推理模型。

动态推理方法旨在解决模型无法对训练阶段未观测到的实体进行精确预测的缺陷。例如，在新闻领域中，“大卫·莫伦斯”作为马里兰州国家国民和传染病研究所的科研人员，在新闻文本中较少出现，这可能已有知识图谱中缺失该实体的信息，使得推理任务难以进行。针对这一问题，现有的方法可以分为基于零样本实体文本描述的方法和基于辅助三元组关联信息的方法。基于零样本实体文本描述的方法主要利用外部文本提供的丰富信息对零样本实体进行表示，通过从文本中获取实体特征以解决已有知识图谱中缺失该实体特征的问题，从而完成动态推理任务；而基于辅助三元组关联信息的方法则通过有限辅助三元组挖掘零样本实体与已有知识图谱之间存在的关联信息，进而对零样本实体进行表示并完成动态推理任务。

上述方法虽然都可以用于面向零样本实体的动态推理任务，但通过分析，我们认为这些方法存在以下局限性：

（1）基于零样本实体文本描述的方法虽然能够利用模型从外部文本中提取实体与关系的表示，但是外部文本的获取难度与成本较高，同时从外部文本中提取实体与关系特征需要额外进行大量的计算，进一步加大了该方法的成本。除此之外，外部文本的质量将直接影响到模型的性能，这对模型的性能产生了制约。

（2）基于辅助三元组关联信息的方法通过挖掘已有知识图谱与零样本实体之间的关联，从而得到零样本实体的高质量表示。然而，现有的方法在挖掘关联信息的过程中未能充分的考虑邻域实体与邻域关系的特征，使得模型无法充分的挖掘实体间的关联关系，对模型的性能产生影响。

为了解决上述难点问题，本文研究面向零样本基于多层传播的动态推理方法。具体而言，本文提出一种基于多层传播机制的动态推理模型。该模型通过实体传播层和关系传播层分别从知识库中获取实体关联信息与关系关联信息，通过融合这两类信息生成零样本实体的高质量表示，从而进行动态推理任务。通过上述方式，我们无需花费资源获取外部知识并引入至模型中，并有效的从有限辅助知识中生成零样本实体的向量表示，实现知识图谱的动态补全，我们的主要工作如下：

（1）设计了一种新的关联信息挖掘方法，在对零样本实体进行表示的过程中无需借助外部文本的信息。

（2）通过多层传播机制捕获零样本实体的邻居实体与邻居关系中蕴含的丰富上下文信息，在挖掘关联信息的过程中充分考虑了已有知识图谱的内在特性。

（3）在多个数据集上的实验结果证明了我们方法的有效性。

4.1.2 问题的定义

动态推理模型需要解决的问题是：随着领域文本的不断更新，已有知识图谱也会不断的扩展，期间知识图谱构建模型可能会从领域文本中抽取出包含零

样本实体的三元组。为此，我们将新抽取出的包含零样本实体的三元组作为辅助集，模型则是要通过有限的辅助集信息对零样本实体进行表达，进而完成动态推理任务。在此，我们定义训练时可被观测到的知识图谱为 $G=\{E,R\}$，其中 E 和 R 分别表示训练过程中出现过的实体集合与关系集合。而测试集中可能会出现训练集中不存在的实体，这种零样本实体集被记为$E^o=\{e_1^o,e_2^o,\cdots,e_n^o\}$。

与传统的推理任务相比，动态推理任务中零样本实体的语义嵌入是缺失的，因此模型无法直接使用该类实体的嵌入信息，而是必须利用有限的辅助知识对该类实体进行表示。在这种情况下，推理任务是正确识别包含零样本实体的三元组 $\tau_{ijk}=(e_i,r_k,e_j^o)$ 或 $\tau_{ijk}=(e_i^o,r_k,e_j)$，该三元组包含由已知实体 $e\in E$ 与零样本实体 $e^o\in E^o$ 组成。

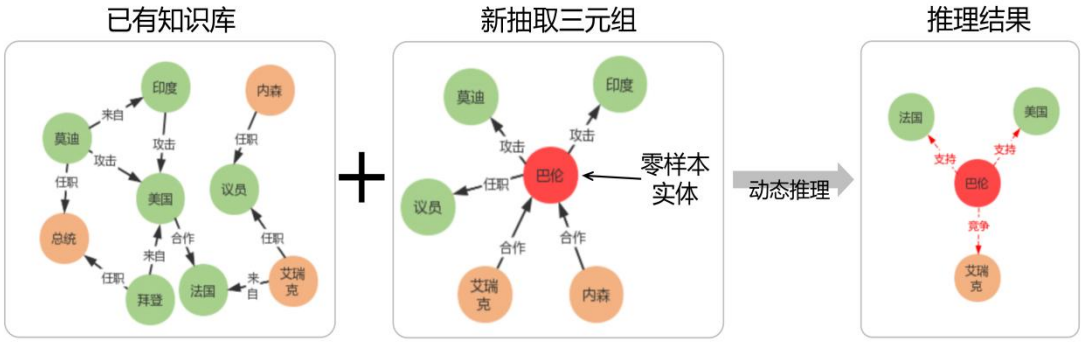

图 4.1　新闻领域动态推理示例，图中红色实体表示训练阶段未出现过的实体，黑色连线表示通过知识图谱构建方法新抽取出的关系，红色连线表示需要推理的关系。

图 4.1 展示了新闻领域动态推理的示例，由于领域文本的不断更新，通过知识图谱构建技术能够从邻域文本中获得新的三元组（巴伦，合作，艾瑞克）和（巴伦，任职，议员），而“巴伦”这一实体在先前的训练集中从未出现过。动态推理的任务就是判断“巴伦”与其他已知实体之间是否可能存在某种关系。

在军事推演领域中，这一问题则更为重要。上文提到，军事推演领域中我们以实体表示作战单元，关系表示作战单元间的行为。由于作战单元种类繁多，而推演数据往往规模有限，因此部分作战单元在以往推演数据中可能从未出现过。而在对当前作战形势进行实时分析的过程中，可能会出现训练集中未曾观测到的作战单元对我方实施打击或侦察等行为，动态推理任务则是通过对这类

实体已知行为的分析，判断其可能会执行的潜在作战行动，从而为指挥员提供可靠的作战参考。

4.2 基于多层传播机制的动态推理模型总体框架

为了解决零样本实体问题，本文提出了基于多层传播机制的动态推理模型。在模型的输入层中，我们将关系的方向性纳入考虑，从而更加充分的利用零样本实体的有限辅助信息；之后，通过实体信息传播层与关系信息传播层充分挖掘零样本实体与已有知识图谱之间的关联关系，经池化层后生成零样本实体的高质量表示，并进行动态推理任务。模型总体架构如图 4.2 所示：

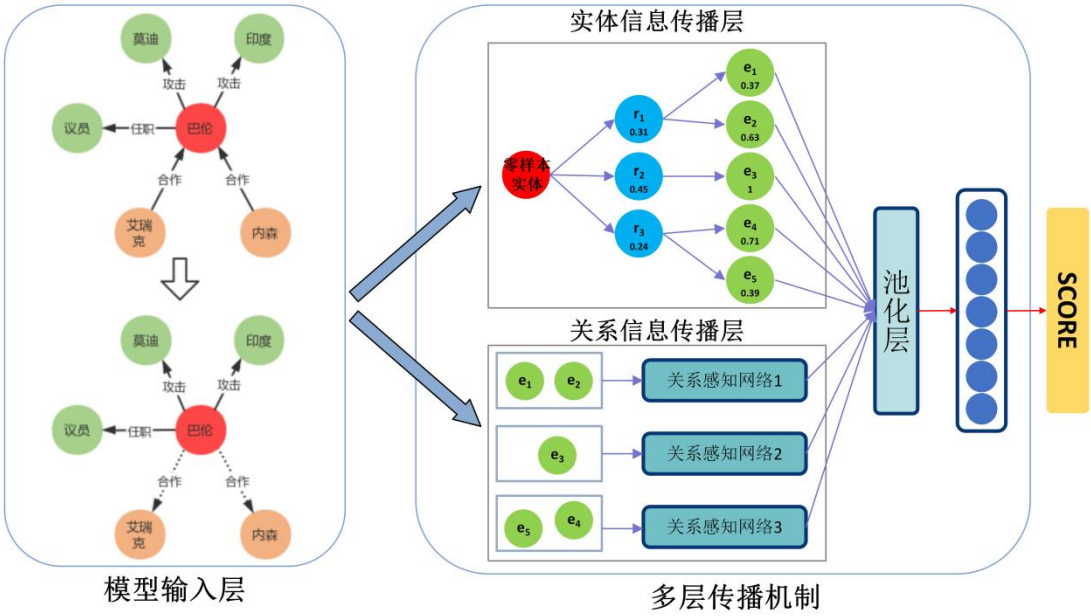

图 4.2 基于多层传播机制的动态推理模型架构图，在模型输入层中考虑关系的方向性，通过多层传播机制获取零样本实体的高质量表示，最终完成动态推理任务。

下文将详细描述基于多层传播机制的动态推理模型的设计思路与实现方式，主要分为如下三个部分：在 4.3 节中介绍模型输入数据的形式以及如何对输入数据进行预处理，为后续任务奠定基础；4.4 节中介绍多层传播机制的具体实现方式，包括实体信息传播模块和关系信息传播模块的设计动机和设计思路；4.5 节则详细描述了模型的输出形式以及损失函数的计算方法。

4.3 模型输入层

根据上文所述，模型的输入主要包括两部分，即不包含零样本实体的训练集和包含零样本实体的辅助集。为了在模型中引入关系的方向性，我们对训练集和辅助集中的任意的一条从e_i指向e_k的数据(e_i, r_j, e_k)，均生成了一条反向的从e_k指向e_i的数据(e_k, r_j', e_i)，以此让模型能够识别关系的方向性，我们将r_j'称为r_j的反向关系。之后，对于已知实体集$E = \{e_1, e_2, ..., e_n\}$中的每个实体$e$，我们将训练集中出现的所有实体关系对组合在一起以便于后续计算；对于零样本实体集$E^o = \{e_1^o, e_2^o, ..., e_3^o\}$，我们则将辅助集中出现的所有实体关系对组合在一起，以实体e_i为例，具体方式如下：

$$\mathcal{N}_{e_i} = \{(r_1, e_1), (r_2, e_2), \cdots, (r_n, e_n)\} \tag{4.1}$$

其中，(r_j, e_j)表示存在形如(e_i, r_j, e_j)的三元组，对于多跳邻域，我们采用(r_j, N_{e_j})的方式递归获取多跳邻域信息，使得模型能够更加充分对邻域信息进行利用。

此外，在对模型进行训练与测试的过程中，需要对数据集进行负采样以完成后续任务。由于知识图谱中实体间不仅仅存在一对一的关系，还可能存在一对多、多对一或多对多的关系，因此使用随机替换头尾实体的方式生成负样本可能会使得负样本中实体出现频率与数据集中的频率存在较大差异，使得样本不能很好的反映数据集真实情况。为此，我们参考伯努利采样方法，在采样的过程中统计实体在数据集中的分布情况以决定如何选取负样本。

具体而言，对于正样本三元组$\tau_{ijk} = (e_i, r_k, e_j)$，我们统计存在关系$r_k$的三元组中头实体$e_i$对应的尾实体个数$n_h$和尾实体$e_j$对应的头实体个数$n_t$，并利用这两个统计数据计算概率$p$，具体计算方式如下：

$$p = \frac{n_h}{n_h + n_t} \tag{4.2}$$

上式概率p能够有效反应数据集中实体的分布情况，因此对于事实三元组集合，我们以概率p对其进行伯努利采样生成负样本，即以p的概率将事实三元组中的头实体替换为实体库中的随机非真实体，以1-p的概率将事实三元组中的尾实体替换为实体库中的随机非真实体，最终生成与原数据集拥有相同特性的负样本集合并输入至模型中。

4.4 多层传播机制

多层传播机制旨在通过利用已有的知识图谱和有限辅助集，学习邻域信息与实体之间的映射关系，从而生成零样本实体的向量表示。它由实体传播模块和关系传播模块两部分构成。

4.4.1 实体信息传播模块

实体信息传播模块旨在从邻域实体中还原中心实体的表示，从而让模型能够在缺乏中心实体信息的情况下，依然能够通过邻域信息推测出中心实体的表示。然而，在传统的推理模型中[41][76]，大部分模型在聚合邻域实体信息的过程中需要对邻域信息的权重进行随机初始化，或根据中心节点的特性计算邻域信息的权重，这些方法在动态推理任务中往往表现得无法尽如人意，为此，我们设计了一种针对零样本实体的分层注意力机制，包括实体注意力层与关系注意力层，以此有效的聚合零样本实体的邻域信息，从而更好的生成零样本实体的向量表示。

关系注意力层的设计动机如下：当同一实体被不同关系所连接时，其对中心实体的影响力往往是不一样的。例如，在舆情分析领域中，通过“合作”相连的邻域实体往往比通过“采访”相连的邻域实体更加重要，因为通过“合作”相连的邻域实体与中心实体更可能具有相似的立场与目标，而“采访”则不能表现出这一点。

从这一动机出发，我们首先计算三元组 $\tau_{ijk}=(e_i,r_k,e_j)$ 的关系注意力层分数，由于零样本实体的存在，三元组中可能存在一个实体缺失语义向量，我们在此通过正向关系与反向关系的转化将零样本实体的位置固定为头实体，并在计算过程中不使用头实体的语义向量，具体如下：

$$a_r = W_1\gamma \tag{4.3}$$

$$\alpha_r = softmax[\sigma(a_r \cdot v)] = \frac{exp[\sigma(a_r \cdot v)]}{\sum_{r' \in \chi_{e_i}} \sigma(a_{r'} \cdot v)} \tag{4.4}$$

其中，$W_1 \in \mathcal{R}^{d\times d}$ 和 $v \in \mathcal{R}^d$ 为训练参数，$\gamma \in \mathcal{R}^d$ 是与关系种类一一对应的系数，σ 为激活函数，在此处我们将 LeakyRelu 函数作为激活函数，该函数对大于零的输入不做任何处理，对于小于 0 的输入则乘以 0.2 的系数。通过上述计算，我们能够获得关系层的注意力系数 α_r，它表示邻域实体在关系层注意力分数。值得注意的是，α_r 能够被关系 r 下的其他三元组显式共享，这促进了不同三元组之间的信息交流。

之后，我们计算邻域实体的实体注意力层分数。其出发点与关系注意力层的出发点类似：即使是同一个关系下的不同实体，其对中心节点的影响力往往也存在较大差异。以新闻数据集中的“竞争”关系为例，当“竞争”对象为“美国”时，该实体应当被给予更高程度的关注，而当“竞争”对象为“斯里兰卡”时，该实体则应当被给予较低程度的关注，这是因为两个国家无论是在国际影响力还是在综合国力上均存在显著差异。出于这样的考虑，我们通过如下方式计算邻域实体的实体注意力层得分：

$$b_{r,e_j} = W_2[a_r \| e_j] \tag{4.5}$$

$$\beta_{r,e_j} = softmax[\sigma(b_{r,e_j} \cdot u)] = \frac{exp[\sigma(b_{r,e_j} \cdot u)]}{\sum_{e_j' \in \chi_{e_i,r}} \sigma(b_{r,e_j'} \cdot u)} \tag{4.6}$$

其中，$W_2 \in \mathcal{R}^{d\times 2d}$和 $u \in \mathcal{R}^{d}$为训练参数，$e_j \in \mathcal{R}^{d}$则是尾实体的语义向量，通过上述计算，我们能够获得邻域实体e_j在关系 r 下的注意力分数，并将其记作β_{r,e_j}。

在获取关系层注意力分数和实体层注意力分数后，我们通过将二者相乘得到邻域实体的注意力分数，通过这种方式，我们便能够动态地计算邻域实体的重要性权重，并根据它们的重要性控制其传递到中心实体的信息流，具体过程如下：

$$\xi_{r,e} = \alpha_r \cdot \beta_{r,e} \tag{4.7}$$

$$e_\alpha = \xi_{r,e} e \tag{4.8}$$

其中，e 表示邻居实体的向量表示，e_α表示当前三元组邻居实体向中心实体传递的信息流。通过上述方式，我们成功通过实体信息传播模块获取来自邻域实体的丰富上下文信息。

4.4.2 关系信息传播模块

实体信息传播模块主要介绍了中心实体如何从邻域实体处获得丰富语义信息的，虽然在这一过程中考虑了关系的影响，但依然缺乏直接来自于关系的丰富信息。因此，我们借用 Hamaguchi 等人提出的传播模型[46]的思想，依据关系特征对邻域实体信息进行转换，从而在聚合邻域信息的过程中更好地融合关系中蕴含的丰富上下文信息。

总的而言，对于任意的实体 e，我们对其包含正向关系与反向关系的实体关系对集合$\mathcal{N}_e = \{(r_1, e_1), (r_2, e_2), \cdots, (r_n, e_n)\}$进行计算，从而在信息传播的过程中有效融合关系信息，计算整体过程如下：

$$S_e^h = \{T_{r_i}^h(e_i)\} \tag{4.9}$$

$$S_e^t = \{T_{r_i}^t(e_j)\} \tag{4.10}$$

$$e_\beta = S_e^h \cup S_e^t \tag{4.11}$$

其中，S_e^h 表示实体 e 在$\mathcal{N}_e$中的正向实体关系对经过正向关系转移函数 T^h 得到的向量表示，S_e^t 表示实体 e 在$\mathcal{N}_e$中的逆向实体关系对经过逆向关系转移函数 T^t 得到的向量表示。我们对这两部分信息取并集，从而得到包含实体 e 的每个三元组从关系传播模块中获取的信息 e_β。其中，转移函数 T 的计算方式如下所示：

$$T_r^h(e_i) = ReLU\{BatchNorm[Linear_r(e_i)]\} \tag{4.12}$$

$$T_{r'}^t(e_j) = ReLU\{BatchNorm[Linear_{r'}(e_j)]\} \tag{4.13}$$

上述公式中 $Linear_r$ 和 $Linear_{r'}$ 分别表示正向关系和逆向关系的转换层，由一个线性层组成，能够根据输入关系的不同对实体向量进行转换，从而在传播过程中融入关系特征。BatchNorm 表示标准化层，目的是控制线性层的输出结果拥有相同的取值范围，这意味着经不同类型关系的转换层输出的特征会被映射到同一特征空间。此外，它还可以使得模型能够更加均衡地关注到各类特征并加快其收敛速度。最后通过 ReLU 激活函数对输出神经元进行过滤，通过消除部分神经元的输出结果，生成稀疏性网络，从而降低参数间的相互依存关系，提高训练速度并缓解模型过拟合的问题。

4.4.3 池化层

通过上述步骤我们分别获取了融合邻域实体信息的 e_α 和蕴含关系信息的 e_β，之后，我们通过池化函数 Q 对上一步输出的信息进行聚合，从而多个三元组中得到实体的最终表示。模型分别探究了两种池化方式对结果的影响，分别为：对包含实体 e 的所有三元组传播过来的信息流求平均以及求最大值，具体方式如下：

$$x = e_\alpha + e_\beta \tag{4.14}$$

$$e_{avg} = Q(S)_{avg} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.15}$$

$$e_{max} = Q(S)_{max} = max(x_1, x_2, ..., x_n) \tag{4.16}$$

上述公式中不同的池化方式分别代表着不同的含义。x_i 代表某个三元组经实体传播层和关系传播层传递到中心节点的关联信息，因此对所有三元组传递传递的信息求最大值代表从邻域信息中选取最具代表性的信息来表示中心实体，而取平均值则是代表均衡考虑邻域信息并对中心实体进行表示。

4.5 动态推理任务层

我们在信息传播模块中分别利用实体信息传播模型与关系信息传播模型从已有知识图谱和有限辅助集中获取零样本实体的信息，并将两个模型的输出结果的平均值作为零样本实体的向量表示，从而进行最终的动态推理任务。在此，我们训练一个打分函数对三元组的可靠性进行评分，具体而言，我们遵从经典的 TransE[24]模型的思想对打分函数进行设计，对于一个三元组 $\tau_{ijk} = (e_i, r_k, e_j)$，其计算过程如下：

$$f(e_i, r_k, e_j) = \left\| e_i + r_k - e_j \right\| \tag{4.17}$$

根据 TransE 的思想，我们希望头实体向量 e_i 与关系向量 r_k 之和应当尽可能接近尾实体向量 e_j。因此，当输入三元组为事实三元组时，我们希望打分函数输出的结果尽可能接近于 0，而当输入三元组不是事实三元组时，我们希望打分函数的结果尽可能的高。

为了防止训练过程中模型盲目追求非事实三元组获得更高的得分，我们采用绝对边界目标损失函数（Absolute-Margin）对模型进行训练，从而让模型能够更好的区分正负样本之间的差异性，具体如下：

$$L = \sum_{\tau \in \chi} \sum_{\tau' \in \chi'} f(\tau) + max[(\lambda - f(\tau'), 0)] \tag{4.18}$$

其中，$\tau \in \chi$ 表示三元组为事实三元组，$\tau' \in \chi'$ 表示三元组为非事实三元组，λ 为阈值超参数，用于限制模型通过给非事实三元组过高的评分以获得更低的

损失值。通过这种方式，模型在训练的过程中会倾向于给事实三元组接近于 0 的评分，而给非事实三元组接近于阈值超参数λ的评分。

4.6 实验结果与分析

在本部分我们将对上文中提出的基于多层传播机制的动态推理模型进行实验评估。我们根据实际应用过程中遇到的零样本实体问题，对已有的数据集进行了重新构建并开展实验。4.6.1 小节介绍了我们是构建包含零样本实体的数据集的过程，4.6.2 小节介绍了我们的评价指标与实验超参数设置，4.6.3 小节则对我们的实验结果进行了展示与分析。

4.6.1 数据集构建与描述

这一部分，我们主要通过已有数据集构建面向零样本实体的新数据集，以还原实际场景中可能出现的零样本实体问题。已有数据集包括百科数据集、金融数据集和公共数据集 WordNet11。其中，百科数据集通过前文中提出的知识图谱构建模型与静态推理模型构建，金融数据集则主要描述了企业与企业之间和企业家与企业之间存在的关联关系，百科数据集、金融数据集与公共数据集 WordNet11 的相关统计数据如下：

表 4.1 百科数据集数据统计

数据集	实体集	关系集	训练集	验证集	测试集
百科数据集	65156	18	96301	5288	5146
金融数据集	5912	8	14869	1316	911
WordNet11	38696	11	112581	5218	21088

之后，为了能够让原数据集更加贴合实际应用场景，使之包含零样本实体，我们对其进行了如下步骤：（1）随机选取若干实体构成零样本实体集；（2）通过零样本实体集构建新的训练集与测试集。具体步骤如下：

（1）随机选取若干实体构成零样本实体集：我们在百科数据集中随机选取 N 个三元组构成集合，并通过这一集合生成零样本实体集。之后我们将集

合中每个三元组的头实体取出作为零样本实体候选集，记为“head-N”；将尾实体取出作为零样本实体候选集，记为“tail-N”；将头尾实体均取出作为零样本实体候选集，记为“both-N”。通过上述方式，我们可以获得多种类型的零样本实体候选集合。之后，为了保证零样本实体与已有知识图谱存在关联关系，我们将候选实体集中与已知实体不存在任何联系的实体删除，得到最终的零样本实体集合。

（2）通过零样本实体集构建新的训练集与测试集：在这一步骤中，我们首先通过上一步骤获取的零样本实体集合对训练集进行划分。训练集中不包含零样本实体的三元组将被作为新数据集的训练集，以此模拟更新数据前的已有知识图谱；训练集包含一个零样本实体的三元组将被作为新数据集的辅助集，以此模拟在更新数据过程中被新抽取出的包含零样本实体的数据；训练集中包含两个零样本实体的三元组则会被舍弃；对于测试集，我们则简单地删除了不包含任何零样本实体的三元组，并将剩余的三元组作为新数据集的测试集，以模拟实际应用过程中涉及零样本实体的推理任务。

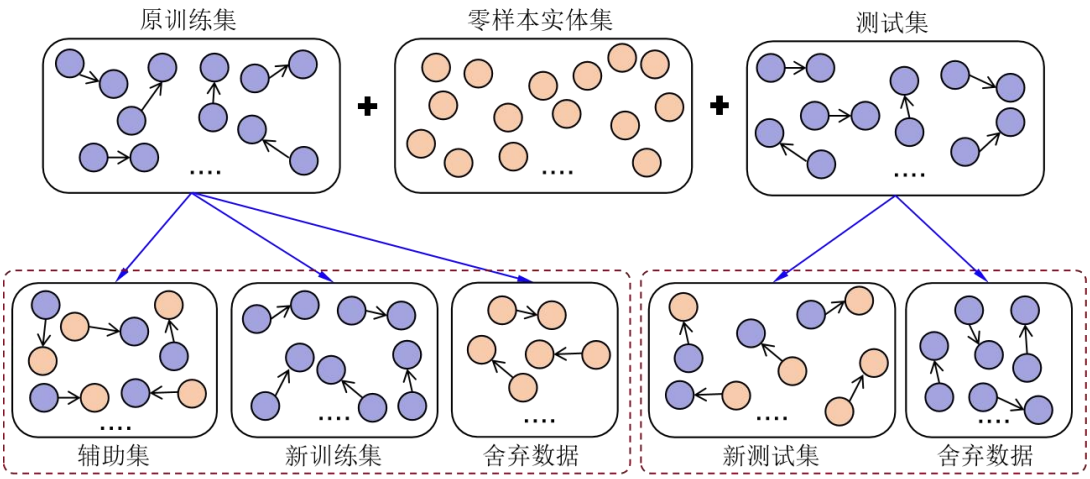

图 4.3　数据集构建流程，首先从原训练数据中随机选取部分三元组，从而获得零样本实体集，再利用零样本实体集对训练集和测试集进行重构，形成最终的数据集。

上述步骤可以简单为如图 4.3 所示的过程。对于百科数据集我们分别取 N=1000 与 N=3000 构建新数据集，通过上述方式，我们从原始数据集中构建了六个面向零样本实体的新数据集，它们分别是：百科-1000（Head）、百科-

1000（Tail）、百科-1000（Both）、百科-3000（Head）、百科-3000（Tail）和百科-3000（Both）。这些数据集的详细统计信息如下所示：

表 4.2 百科数据集数据统计

	Head		Tail		Both	
N	1000	3000	1000	3000	1000	3000
实体集	65156	65156	65156	65156	65156	65156
零样本实体集	786	2056	975	2827	1100	2819
训练集	67542	54030	91687	84384	65953	50867
测试集	2826	4058	1166	2412	2984	4512
辅助集	28609	41804	4606	11807	28693	41267

与百科数据集类似，公开数据集 WordNet11 也通过同样的方式构建为新的数据集，WordNet11 为 WordNet 数据集的子集，该数据集是由普林斯顿大学的心理学家、语言学家和计算机工程师联合设计的一种基于认知语言学的英语词典，它不是光把单词以字母顺序排列，而且按照单词的意义组成一个单词的网络。通过 WordNet11 构建的面向零样本实体的数据集的详细信息如表 4.3 所示：

表 4.3 WordNet11 数据集数据统计

	Head		Tail		Both	
N	1000	3000	1000	3000	1000	3000
实体集	38192	38180	38180	38089	38165	37978
零样本实体集	348	1034	942	2627	1238	3319
训练集	108197	99963	96968	78763	93364	71097
验证集	4613	4184	3999	3122	3799	2166
测试集	994	2969	986	2880	690	2708
辅助集	4352	12376	15277	31770	18638	38285

金融数据集则与上述两个数据集略有不同，由于金融数据集中包含的实体种类相对较少，仅包含 5912 种实体，为百科数据集的十分之一。在这种情况下，若仍然取 N=3000 则会导致零样本实体集过大，已有知识图谱中的信息量

过少，不利于模型的训练与推理，为此，在综合考虑下，在金融数据集上我们选取了 N=500 和 N=1000 两种方式构建新数据集，具体如下所示：

表 4.4 金融数据集数据统计

	Head		Tail		Both	
N	1000	3000	1000	3000	1000	3000
实体集	5912	5912	5912	5912	5912	5912
零样本实体集	475	877	472	877	828	1388
训练集	11961	9749	11986	9688	10263	7295
测试集	320	518	300	538	468	750
辅助集	2770	4615	2759	4729	3670	5755

4.6.2 评价指标与参数设置

（1）评价指标

对于本章的动态推理任务，我们将该任务转化为对三元组合理性的评判，即三元组分类任务。因此，我们将评判的准确率作为模型的评价指标。具体而言，我们通过评分函数对三元组进行合理性分析，对于模型评分小于阈值的三元组，我们认为其为事实三元组，而对于模型评分大于阈值的三元组，我们则认为其为非事实三元组。之后，我们将模型的预测情况与测试样本的真实情况进行对比并计算精确率，具体公式如下：

$$P = \frac{correct\_num}{test\_num} \tag{4.19}$$

（2）参数设置

在本次实验过程中我们涉及到了一些重要的超参数，主要包括向量维度、训练次数、阈值、邻域深度、最大邻居个数、学习率等。其中，阈值指的是模型判断输入三元组是否为事实三元组的阈值、邻域深度指的是模型最多能够从几阶邻居中聚合信息、最大邻居个数指的是当邻域信息过多时，模型每次迭代时最多使用的邻居节点个数。这些参数的具体设定如表 4.5 所示：

表 4.5 超参数设置

超参数	含义	取值
dim	向量维度	200
epochs	训练次数	10000
λ	阈值	300
depth	邻域深度	1
sample_size	最大邻居个数	64
lr	学习率	0.1

4.6.3 结果分析

为了验证我们方法的有效性，我们在三个数据集上展开实验对模型的性能进行了验证。对比实验中，我们选取了经典的 TransE 模型与基于图注意力机制的 g-GAT[42]、IGNN[50]模型和 D-AEN[45]模型作为基准模型，并与它们进行对比实验。在消融实验中，我们探究了对关联信息采样时采用不同聚合方式对模型性能的影响，具体方式在上文中已给出。

表 4.6 WordNet11 数据集实验结果

		Head		**Tail**		**Both**	
Method	**Pooling**	1000	3000	1000	3000	1000	3000
TransE[24](2013)	Max	58.1	56.3	55.2	54.2	56.8	56.8
	Avg	63.0	60.2	63.8	63.9	65.3	63.9
g-GAT[42](2019)	Max	62.2	62.8	63.0	61.9	67.2	65.1
	Avg	75.6	71.8	72.0	72.0	75.6	**76.0**
IGNN[50](2020)	Max	76.4	73.3	72.5	62.6	67.2	58.5
	Avg	83.5	80.5	78.7	72.8	75.4	67.9
D-AEN[45](2023)	Max	80.6	79.7	79.4	71.9	72.9	66.7
	Avg	81.3	81.2	79.6	73.6	76.1	72.1
Proposed	Max	82.2	77.3	76.0	67.7	70.1	62.8
	Avg	**86.0**	**83.2**	**82.9**	**74.5**	**80.1**	74.6

如表 4.6 所示，在公开数据集 WordNet11 上的实验结果表明，无论采用何种方式对邻域关联信息进行聚合，我们提出的基于多层传播机制的动态推理模型的性能均优于其他基准模型。这说明我们的模型较其他模型具有更好的特征提取能力，能够从已有知识图谱中更好的还原零样本实体的特征，从而在动态推理任务中取得更好的表现。此外，我们注意到聚合邻域关联信息时，采用 Avg 方式均衡的考虑中心节点的所有关联信息可以取得更好的效果，而不是使用 Max 方式仅考虑关联信息中最重要的特征。

与此同时，我们注意到对于百科数据集而言，当选取三元组中的尾实体作为零样本实体时，模型能够更加精确的推断零样本实体可能存在的关系。如上述实验结果所示，当选取尾实体为零样本实体时，模型的性能较选取头实体与选取头实体和尾实体提高了约 13%。为了探究这一现象出现的原因，我们对训练集的数据分布进行了分析。通过统计分析，我们发现在训练集中共存在 43234 种实体曾被作为尾实体，而仅有 23621 种实体曾被作为头实体，这意味着将选取头实体作为零样本实体时，知识库将缺失更多的关键信息。这一现象的存在导致了模型在不同数据集上表现的差异。

表 4.7　百科数据集与金融数据集实验结果

		Head		**Tail**		**Both**	
DataSet	**N**	Max	Avg	Max	Avg	Max	Avg
百科数据集	**1000**	66.7	**74.9**	69.8	**87.0**	66.2	**73.7**
	3000	65.2	**70.8**	67.9	**83.9**	64.6	**70.3**
金融数据集	**500**	70.6	**73.5**	68.3	**73.2**	67.9	**72.4**
	1000	63.3	**69.7**	68.0	**72.5**	65.2	**68.7**

4.7　本章小结

本章针对实际应用过程中可能遇到已有知识图谱中缺失相应实体信息的问题，进行了面向零样本基于多层传播的动态推理方法研究。由于领域文本的不断更新，模型需要不断地对各类实体的关系进行推理，在这一过程中往往会出现训练时未观测到的零样本实体，并为推理的准确性带来挑战。现有的方法主

要分为两类：基于零样本实体文本描述的方法和基于辅助三元组关联信息的方法。然而，前者可能因缺少文本描述而受到约束，后者在挖掘关联信息的过程则往往忽略了邻域实体本身的特征，导致模型的性能受到影响。

针对上述问题，我们设计了基于多层传播机制的动态推理模型，该模型属于基于辅助三元组关联信息的方法，模型在挖掘关联信息的过程中通过实体传播模块和关系传播模块充分的考虑邻域节点特征并挖掘关联信息。通过这一方式，我们的方法不仅能够避免文本描述对模型性能的约束，同时解决了挖掘关联信息方法中实体特征缺失的问题，使得模型在实验中取得了更好的效果。

综上所述，我们提出的基于多层传播机制的动态推理模型能够有效地挖掘辅助三元组中蕴含的关联信息，以此提供实体与关系高质量的特征向量，并在零样本实体问题上取得良好的结果。

第五章 面向新闻领域的辅助决策信息系统

5.1 应用背景

知识图谱是一种超级知识库，它描述了客观世界中实体之间的相互作用关系。它能够对非结构化新闻文本数据进行有效整合，并将其转化为成高质量的结构化数据。在大数据时代，知识图谱主要用于数据结构化处理、解析、关联以及后续分析等任务之中。它具有强大的数据语义分析处理能力并有效描述了实体之间的关联关系，改进了数据的存储与展现方式。因此，知识图谱技术已经从最初被搜索引擎应用的技术逐渐成为各行各业研究的热点。

正是因为知识图谱拥有强大的数据分析能力，所以它能够有效的帮助决策者处理海量信息并提供可靠的建议[77][78]。因此，本文结合决策者在制定政策过程中的信息需求，以新闻领域的文本数据为基础，将上文所提到的基于层次特征表示的知识图谱构建方法、面向长尾分布样本的静态推理方法以及面向零样本基于多层传播的动态推理方法进行工程化实现，集成为面向新闻领域的辅助决策信息系统。其中，基于层次特征表示的知识图谱构建方法能够从海量新闻文本中发现各类实体以及实体之间存在的复杂关系，并以结构化的方式对其进行存储；面向长尾分布样本的静态推理方法能够挖掘实体间存在的潜在关系，从而向决策者提供更加全面的知识服务；面向零样本基于多层传播的动态推理方法则能够对零样本实体的关系进行准确推理，从而避免重新训练模型，在应用时达到实时推理各个实体关系的目的。

综上所述，本系统能够将难以利用的海量新闻文本转化为高度结构化的知识图谱，并运用推理技术向决策者提供高质量的信息服务，从而在速度、维度、精度和广度上满足决策者对于信息的需求，有效应对瞬息万变的局势，为智能决策的发展提供技术支持。

5.2 系统整体架构设计

为了保证系统中各个子模块之间的低耦合性，便于后续对系统代码进行高效的维护与复用，因此采用分层架构设计构建面向新闻领域的辅助决策信息系统。系统总体能够分为四个部分：数据获取层、数据存储层、系统功能层与系统交互层，如图 5.1 所示：

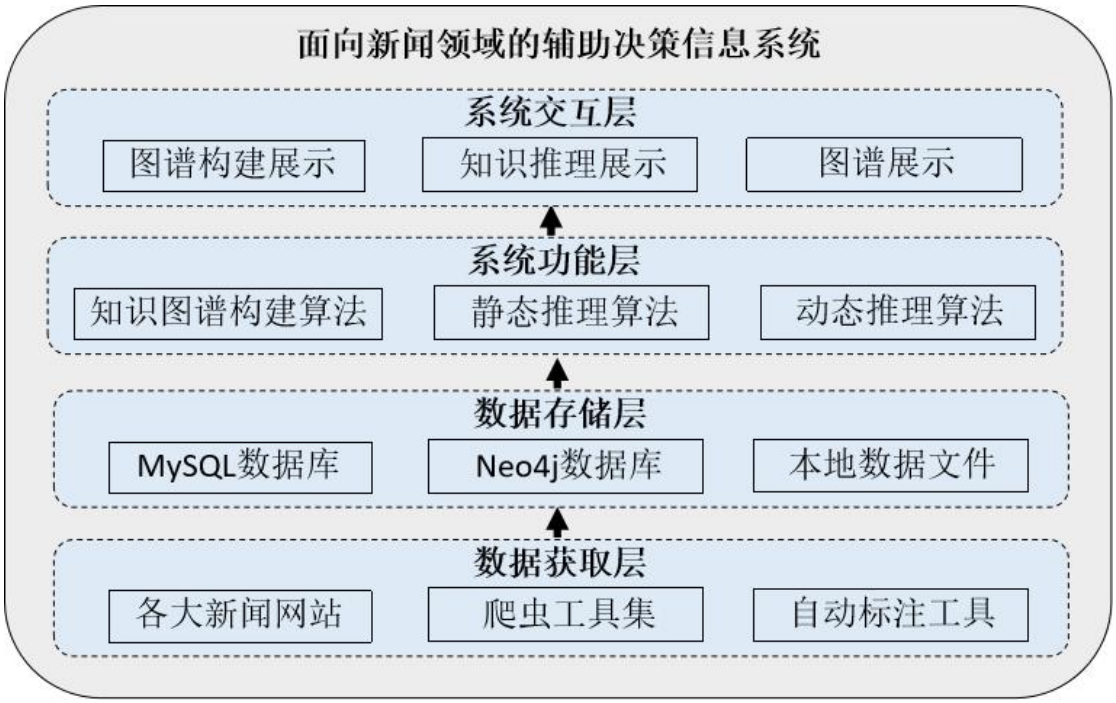

图 5.1 面向新闻领域的辅助决策信息系统结构图，包含系统交互层、系统功能层、数据存储层与数据获取层

5.2.1 数据获取层

数据获取层旨在从各大新闻网站中获取、处理新闻文本数据，将其转化为可用于深度学习的数据集。其中，新闻网站包括新华网、香港新闻网、人民日报、环球网和澎湃网等网站。具体而言，我们通过爬虫技术从相应网站上获取新闻文本后，利用自动标注工具与人工校验的方式将其转化为可靠的知识图谱构建数据集，为后续任务提供数据支撑。

5.2.2 数据存储层

数据存储层主要是根据需求的不同，将上一步生成的各类数据进行分类存储，从而为后续任务提供可靠的数据服务。例如，对于模型训练与测试所需的数据和训练好的模型文件，我们通过本地数据文件的方式进行存储；而对于前后端交互的数据以及爬取的原始新闻语料数据，通过 MySQL 数据库进行存储；对于构建的知识图谱，则利用 Neo4j 数据库对其进行存储。通过这种方式，能够对不同数据进行有效的整合，便于各类任务的执行。

5.2.3 系统功能层

系统功能层包括两个部分，模型的训练与模型的调用。其中，模型的训练是指通过数据存储层的接口获取训练所需数据对模型进行训练，并将训练得到的模型存储于本地数据文件之中；模型的调用则是指根据用户的需求调用相应的模型并将结果反馈给调用者，从而为用户提供可靠的信息服务。

5.2.4 系统交互层

系统交互层是展现给用户浏览的部分，用户可以通过该模块对所需功能进行调用，从而满足自身的需求。该模块包括实体关系抽取展示、知识推理展示和知识图谱展示，能够实现用户与人工智能模型之间的高效交互，为行业风险评估与政策制定提供信息服务。

5.3 系统功能模块设计

基于上述系统整体架构，我们开发出了一套易于操作的面向新闻领域的辅助决策信息系统。该系统能够调用模型实时抽取新闻文本中的实体与关系，并根据抽取结果对已有知识图谱进行更新，同时向决策者提供丰富的信息服务，辅助其决策。系统主要基于 Web 框架 Django 开发，数据库以 MySQL 为主，前端则以 element 框架为主，除首页外，还包括“知识图谱构建”、“知识推理”和“知识图谱展示”三个界面，能够通过侧边菜单栏访问相应的功能界面。

5.3.1 知识图谱构建模块

通过点击左侧菜单栏的“知识图谱构建”模块，系统跳转至知识图谱构建功能界面。该模块依靠本文第二章中研究的基于层次特征表示的知识图谱构建方法，用户输入新闻文本后，点击“抽取”按钮，系统会自动调用模型并对文本的内容进行识别，此外，若此时抽取出的实体为零样本实体，则在前端界面中会以红色字体标出，若抽取出的实体并非零样本实体，则以黑色字体在前端界面进行显示，如图 5.2 所示：

图 5.2 知识图谱构建模块展示，通过输入互联网上的新闻文本能够实现对已有知识图谱内容的实时更新

此外，点击“重置”按钮，能够快速清除前端显示的输入内容与抽取结果，将页面恢复至初始化状态；点击“写入图谱”按钮，则系统会将本次抽取结果更新至已有知识图谱中，从而实现知识库的实时更新功能；点击“舍弃数据”按钮，则系统会忽略本次抽取结果，不对已有知识图谱做任何操作。

5.3.2 知识推理模块

通过点击左侧菜单栏的“知识推理”模块，系统自动跳转至知识推理功能界面。该功能模块主要应用了本文研究的面向长尾分布样本的静态推理方法与面向零样本基于多层传播的动态推理方法，其功能的实现如下：

首先，用户通过输入实体名称并点击“搜索”按钮对实体间的关系进行查询。系统会根据用户输入的信息对已有知识图谱的内容进行检索，并将检索结

果反馈给用户。通过知识图谱构建模型获取的实体间关系会以黑色字体进行显示，而通过知识推理模型推理出的实体间关系则会以红色字体进行区分，在这一过程中调用何种推理模型取决于用户检索的实体是否存在于已有的知识关系知识图谱当中。具体如图 5.3 所示：

图 5.3　知识推理模块功能展示，通过输入实体名称并点击检索按钮，会在左侧“实体关系三元组”中显示实体间存在的关系，点击真实关系（黑色关系）则会在右侧“实体关系详情”中显示具体信息

此外，当用户点击实体间的关系时，右侧则会根据点击关系的不同显示不同的信息。当点击通过知识图谱构建方法得出的关系时，如“攻击关系”，右侧会显示抽取的相关信息，包括实体信息、实体间关系、源自哪段新闻文本、出自哪篇新闻、数据来源网址和数据发布日期。而当点击通过知识推理模型得出的关系时，如“竞争关系”，右侧则会显示推理模型推理时的尾实体得分排序，如图 5.4 所示：

图 5.4　知识推理模块功能展示，通过点击推理所得关系（红色关系），右侧“尾实体排序”中会显示出最可能与当前头实体存在该关系的五个尾实体，“实体关系详情”与“尾实体排序”通过 display 属性实现切换

5.3.3　图谱展示模块

通过上述知识图谱构建模块和知识推理模块，用户能够快速调用本文所提出的各个模型。最后，我们将上述结果进行汇总，以知识图谱为媒介，向决策者提供高质量的信息服务。

通过点击左侧菜单栏的“图谱展示”模块，进入该功能模块。在该功能模块中，输入需要查询的实体名称后，点击“搜索”按钮，后端会访问数据库并将通过知识图谱的形式展示该实体的一阶近邻信息。其中，点击“精确搜索”时，需要数据库中的实体名称与输入框中输入的实体名称完全匹配方可获取相应的信息；点击“模糊搜索”时，则后端会查找与输入实体名称最为相似的实体的相关信息并进行展示，如搜索“特朗普”时会自动匹配至“唐纳德·特朗普”，以此降低用户获取信息的难度。具体情况如图 5.5 所示：

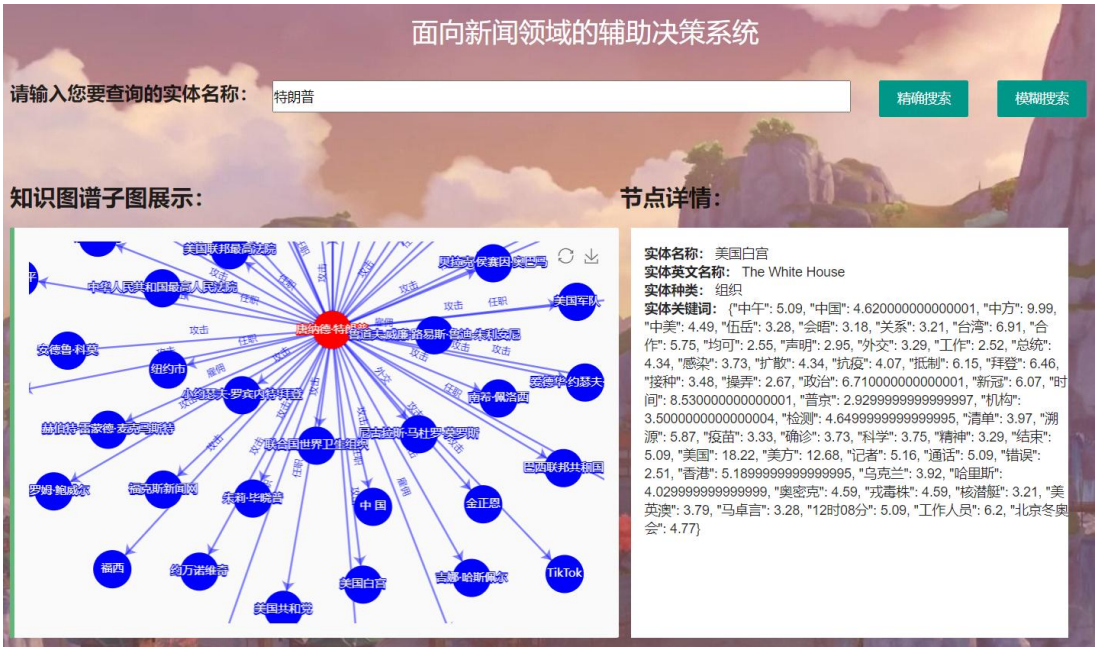

图 5.5　知识图谱展示，点击“美国白宫”节点时，右侧显示该节点的详细信息，包括实体名称、实体英文名称、实体种类与实体关键词

此外，通过点击子图中各个节点，即实体，右侧会显示该实体的相应信息，如中英文名称、实体种类与实体关键词等信息，此外还能够获取到实体涉及新闻、事件和意图等信息，但与本文研究内容关联性不强，因此在此不做展示；而点击知识图谱子图中的各个边，即关系时，右侧则会显示与知识推理模块小节中相同的实体关系详细信息，如图 5.6 所示

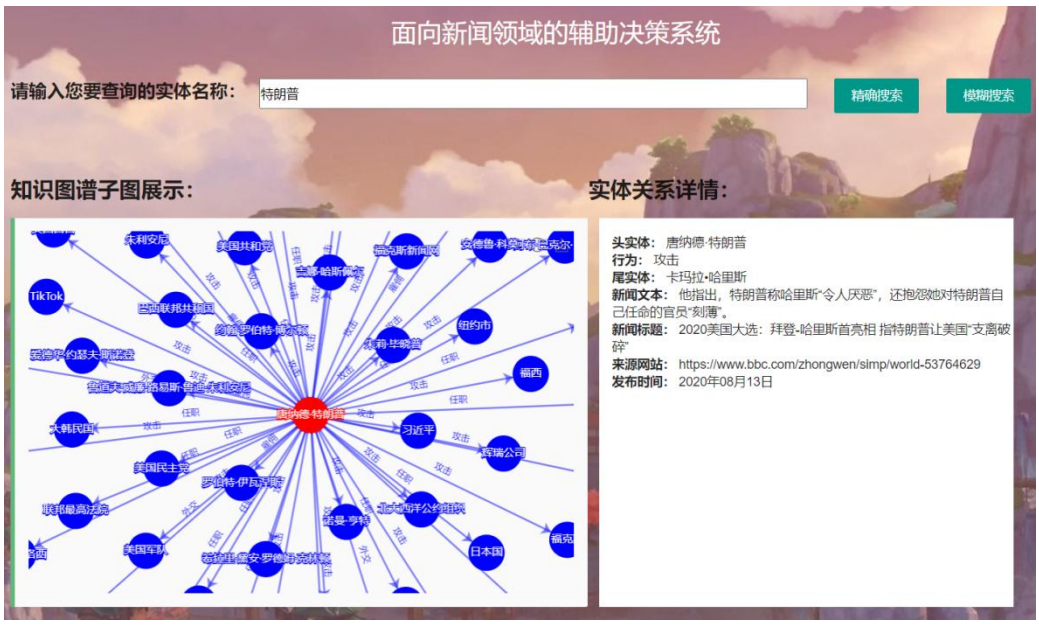

图 5.6　实体关系详情展示，点击子图中的边，右侧“实体关系详情”中所显示信息

5.4 小结

本章将上文中研究的基于层次特征表示的知识图谱构建方法、面向长尾分布样本的静态推理方法和面向零样本基于多层传播的动态推理方法应用于实际场景之中，形成面向新闻领域的辅助决策信息系统，并向外界提供信息服务。该系统包括数据获取层、数据存储层、系统功能层和系统交互层。其中，数据获取层主要介绍了新闻文本数据的来源以及处理方法；数据存储层则介绍了系统各类数据存储的形式；系统功能层和系统交互层则对应该系统的后端与前端，后端通过调用模型奠定数据基础，前端则利用数据为决策者提供可靠的信息。

总的来说，本章将高效利用新闻文本以辅助决策作为出发点，针对这一过程中所存在的问题为切入点，通过设计面向新闻领域的辅助决策信息系统将上文中所提出的模型应用于实践之中，从而在决策制订的过程中为决策者提供有力的信息支持。

第六章 总结与展望

6.1 本文总结

知识图谱是一种通过结构化三元组整合领域文本中丰富知识的语义知识库，它通过节点和边精确描述实体之间的复杂关联关系，进而向外界提供高度结构化的数据。然而，知识图谱中往往存在着样本分布不均衡的问题，在训练的过程中，部分样本出现频次较高，部分样本出现频次较低，部分样本甚至从未在训练阶段中出现过，这大大增加了知识图谱推理的难度，并严重影响了知识图谱在各个领域中的应用。

为此，本文研究面向不均衡样本的知识图谱构建与推理方法，旨在向用户与下游应用提供可靠的知识服务。然而，领域文本作为非结构化数据，难以被直接分析与利用，并且有价值的信息往往分布稀疏且离散，这使得知识图谱的构建困难重重，同时样本的不均衡性也为知识图谱推理带来了重大挑战。综上，我们将问题总结如下：

1）领域文本规模庞大且结构复杂，如何从这部分数据中挖掘出分布稀疏的知识，并以结构化的三元组进行存储与表达，实现知识图谱的构建并为后续任务奠定坚实的数据支撑；

2）为了提高知识图谱的连通性与完整性，必须对其内部缺失的信息进行推理。而由于知识图谱中的样本存在长尾分布效应，即大部分样本出现频次较低，这使得模型难以保证在少样本实体上的推理精度，进而对知识图谱的质量产生了影响。

3）随着领域文本的不断扩增，模型可能需要对训练过程中未观测到的实体进行推理。如何挖掘零样本实体与已有知识图谱之间的关联关系，从而推理零样本实体与其他实体之间存在的潜在关联关系，是研究中必须要解决的问题。

为了解决上述问题，本文以领域文本作为基础，提出了面向不均衡样本的知识图谱构建与推理方法。本文的研究内容如下：

（1）基于层次特征表示的知识图谱构建方法。为了能够从领域文本中构建出高度结构化的知识图谱，当前的主流方法是通过联合抽取模型获取文本中的实体及其关系，该模型一般包含实体抽取与关系抽取两个子任务。然而，现有的基于深度学习的联合抽取模型往往无法实现不同子任务间的有效信息交互。针对这一问题，本文提出了基于分区过滤网络的知识图谱构建模型。在模型中，通过 Bert 模型从领域文本中提取高质量的信息，通过实体嵌入层和关系嵌入层分别提取两个子任务所需的特征，之后利用共享特征分离层从中获取实体特征、共享特征和关系特征，并设计额外的损失函数将不同类型的特征映射到不同的向量空间中，从而实现它们之间的相互隔离，最终有效的提升了知识图谱构建任务的精确率与召回率。

（2）面向长尾分布样本的静态推理方法。为了弥补知识图谱中存在的连通性不足的问题，目前往往通过静态推理模型对已有知识图谱进行全面分析并得到实体与关系的高质量表示，进而有效的发现实体间存在的潜在关系。然而，由于长尾分布效应的存在，一些样本出现次数较少，这使得模型难以从中提取到有效的特征；同时，邻域信息中蕴含着丰富的知识，如何有效聚合邻域信息以增强中心实体的表示也是亟待解决的问题。因此，本文提出基于异构注意力机制的静态推理模型。该方法采用典型的编码器-解码器架构[70]，编码器负责生成高质量的语义向量，解码器则分析三元组的全局嵌入特征并完成静态推理任务。具体而言：首先，我们设计了异构注意力网络动态计算邻居实体与邻居关系的重要性，从而有效的聚合邻域信息，增强中心实体的表示；之后，引入变分信息瓶颈对编码器与解码器之间的信息流进行控制，过滤与任务无关的特征并提高模型在长尾分布样本上的性能；最后，将 ConvKB 作为解码器，通过应用卷积神经网络提取三元组上的不同特征，从而对三元组进行打分并完成静态推理任务。

（3）面向零样本基于多层传播的动态推理方法。动态推理主要针对的是实际应用过程中可能会出现的零样本实体（训练时未被模型观测到的实体）问题，旨在通过挖掘并利用零样本实体与已有知识图谱的关联关系实现知识图谱的动态推理。因此，如何有效的挖掘零样本实体与已有知识图谱之间的关联关

系，有效的对这部分实体进行表示，是实现动态推理任务必须解决的问题。针对这一问题，我们提出基于多层传播机制的动态推理模。首先，在模型的输入部分，我们捕获实体的所有邻居实体，并将关系的方向性也纳入考虑，同时构建正负样本作为模型的输入；之后，构建实体信息传播模块和关系信息传播模块分别获取来自邻域实体和邻域关系的关联信息，从而全面的挖掘零样本实体与已有知识图谱之间的关联关系并对其进行表达；最终，我们参考 TransE 的思想进行动态推理任务。通过这一方式，我们设计的动态推理模型能够有效挖掘零样本实体与已有知识图谱之间的关联关系，从而对零样本实体的可能关系进行实时推理，有效提高知识图谱的质量。

本文以领域文本数据为基础，提出了面向不均衡样本的知识图谱构建与推理方法，最终获得了包含丰富知识的知识图谱，从而为用户和下游应用提供可靠的知识服务。

6.2 研究展望

为了能够从海量非结构化领域文本中提取出有价值的知识，通过知识图谱对这部分知识进行存储，并为下游应用高效赋能，本文结合遇到的困难研究了如下内容：（1）基于层次特征表示的知识图谱构建方法实现了从领域文本中抽取出实体与关系的三元组，并以知识图谱的方式对其进行存储，为后续任务奠定了数据基础；（2）提出面向长尾分布样本的静态推理方法，有效提升了模型在少样本上的表现能力，同时弥补了知识图谱中存在的连通性不足的问题，实现知识图谱的静态推理；（3）针对领域文本不断扩增带来的零样本实体问题，提出了面向零样本基于多层传播的动态推理方法，通过利用有限的辅助信息挖掘零样本实体与已有知识图谱之间的关联关系，完成动态推理任务。我们的研究在知识图谱构建与知识图谱推理任务中均取得了较好的效果，能够有效的从领域文本中生成高质量的知识图谱。然而，通过进一步调研，上述研究内容依然存在不足之处，需要对其进行改善，具体如下：

（1） 基于层次特征表示的知识图谱构建方法中，我们通过将特征类型将特征进行划分，从而实现不同子任务间信息交互与信息隔离之间的平衡。然而，

当训练数据不足时模型难以取得很好的结果，因此后续研究可以引入少样本学习的思想，进一步提高模型的适用范围与抽取效果。

（2） 面向长尾分布样本的静态推理方法中，本文通过变分信息瓶颈模块有效的提升了模型在长尾分布样本上的性能，并设计了一种新的图注意力机制以有效的聚合邻域信息。然而，上述方法未能很好的对实体间的层次关系进行建模，因此在实体层次关系明显的数据集上仍存在一定的改进空间，下一步将研究如何对实体间的层次关系进行建模，从而进一步提升模型的性能。

（3） 面向零样本基于多层传播的动态推理方法中，本文针对零样本实体的问题，设计了实体信息传播模块和关系信息传播模块充分挖掘零样本实体与已有知识图谱之间的关联关系，为利用辅助三元组关联信息解决零样本实体问题提供新的解决思路。然而，由于模型属于基于辅助三元组关联信息的方法，因此在辅助信息较少的情况下难以对零样本实体进行有效描述，从而导致模型动态推理的精度下降。因此，后续将研究如何有效融合外部文本信息与内部关联信息，以此增强零样本实体的表征，从而更好的进行动态推理任务。

参考文献

[1] Wang, Q. , Mao, Z. , Wang, B. , & Guo, L. . (2017). Knowledge graph embedding: a survey of approaches and applications[J]. IEEE Transactions on Knowledge & Data Engineering, 29(12), 2724-2743.

[2] 夏毅,兰明敬,陈晓慧,罗军勇,周刚,何鹏.可解释的知识图谱推理方法综述[J].网络与信息安全学报,2022,8(05):1-25.

[3] Yan Z, Zhang C, Fu J, et al. A partition filter network for joint entity and relation extraction[J]. arXiv preprint arXiv:2108.12202, 2021.

[4] 张少伟,王鑫,陈子睿,王林,徐大为,贾勇哲.有监督实体关系联合抽取方法研究综述[J].计算机科学与探索,2022,16(04):713-733.

[5] Yang B, Cardie C. Joint inference for fine-grained opinion extraction[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013: 1640-1649.

[6] Dantzig G B. Reminiscences about the Origins of Linear Programming[R]. STANFORD UNIV CA SYSTEMS OPTIMIZATION LAB, 1981.

[7] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.

[8] Kate R, Mooney R. Joint entity and relation extraction using card-pyramid parsing[C]//Proceedings of the Fourteenth Conference on Computational Natural Language Learning. 2010: 203-212.

[9] Yu X, Lam W. Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach[C]//Coling 2010: Posters. 2010: 1399-1407.

[10] Singh S, Riedel S, Martin B, et al. Joint inference of entities, relations, and coreference[C]//Proceedings of the 2013 workshop on Automated knowledge base construction. 2013: 1-6.

[11] Bekoulis G, Deleu J, Demeester T, et al. Joint entity recognition and relation extraction as a multi-head selection problem[J]. Expert Systems with

Applications, 2018, 114: 34-45.

[12] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers). 2016: 207-212.

[13] Eberts M, Ulges A. Span-based joint entity and relation extraction with transformer pre-training[J]. arXiv preprint arXiv:1909.07755, 2019.

[14] Wei Z, Su J, Wang Y, et al. A novel cascade binary tagging framework for relational triple extraction[J]. arXiv preprint arXiv:1909.03227, 2019.

[15] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[16] KATIYAR A, CARDIE C. Going out on a limb: joint extraction of entity mentions and relations without dependency trees[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Jul 30-Aug 4, 2017. Stroudsburg: ACL, 2017: 917-928.

[17] Zheng S, Wang F, Bao H, et al. Joint extraction of entities and relations based on a novel tagging scheme[J]. arXiv preprint arXiv:1706.05075, 2017.

[18] Dai D, Xiao X, Lyu Y, et al. Joint extraction of entities and overlapping relations using position-attentive sequence labeling[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 6300-6308.

[19] Fu T J, Li P H, Ma W Y. Graphrel: Modeling text as relational graphs for joint entity and relation extraction[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. 2019: 1409-1418.

[20] Huang Z, Liang L, Zhu X, et al. An Improved Partition Filter Network for Entity-Relation Joint Extraction[C]//Neural Computing for Advanced Applications: Third International Conference, NCAA 2022, Jinan, China, July 8–10, 2022, Proceedings, Part I. Singapore: Springer Nature Singapore, 2022:

129-141.

[21]倪立旺.(2020).基于深度学习的知识图谱补全技术研究(硕士学位论文,电子科技大学). https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202002&filename=1020735515.nh

[22]Arora S. A survey on graph neural networks for knowledge graph completion[J]. arXiv preprint arXiv:2007.12374, 2020.

[23]丁建辉,贾维嘉.知识图谱补全算法综述[J].信息通信技术,2018,12(01):56-62.

[24]Bordes A, Usunier N, Garcia-Durán A, et al. Translating embeddings for modeling multi-relational data[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2. Red Hoook: Curran Associates Inc., 2013: 2787-2795.

[25]Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes[C]//Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2014: 1112-1119.

[26]Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion[C]//Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2015: 2181-2187.

[27]Ji G, He S, Xu L, et al. Knowledge graph embedding via dynamic mapping matrix[C]//Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers). Stroudsburg: Association for Computational Linguistics, 2015: 687-696.

[28]Ji G, Liu K, He S, et al. Knowledge graph completion with adaptive sparse transfer matrix[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. 2016: 985-991.

[29]Xiao H, Huang M, Hao Y, et al. TransG: A generative mixture model for knowledge graph embedding[J]. arXiv preprint arXiv:1509.05488, 2015.

[30]Nickel M, Tresp V, Kriegel H P. A three-way model for collective learning on

multi-relational data[C]//Proceedings of the 28th International Conference on International Conference on Machine Learning. Madison: Omnipress, 2011: 809-816.

[31] Yang B, Yih S W, He X, et al. Embedding entities and relations for learning and inference in knowledge bases[C]//Proceedings of the International Conference on Learning Representations (ICLR) 2015. ICLR, 2015: 1-13.

[32] Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction[C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48. Brookline: Microtome Publishing, 2016: 2071-2080.

[33] Trouillon T, Dance C R, Gaussier É, et al. Knowledge graph completion via complex tensor factorization[J]. Journal of Machine Learning Research, 2017, 18: 1-38.

[34] Liu H, Wu Y, Yang Y. Analogical inference for multi-relational embeddings[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. Brookline: Microtome Publishing, 2017: 2168-2178.

[35] Kazemi S M, Poole D. SimplE embedding for link prediction in knowledge graphs[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2018: 4289-4300.

[36] Balažević I, Allen C, Hospedales T. TuckER: Tensor factorization for knowledge graph completion[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg: Association for Computational Linguistics, 2019: 5185-5194.

[37] 宋兴豪. 基于知识表示学习的知识图谱补全算法研究[D]. 成都: 西南科技大学, 2020.

[38] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks

for knowledge base completion[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 1. Red Hook: Curran Associates Inc., 2013: 926-934.

[39] Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2D knowledge graph embeddings[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. Palo Alto: AAAI Press, 2018: 1811-1818.

[40] Nguyen D Q, Nguyen T D, Nguyen D Q, et al. A novel embedding model for knowledge base completion based on convolutional neural network[J]. arXiv preprint arXiv:1712.02121, 2017.

[41] Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks[C]//The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3 – 7, 2018, Proceedings 15. Springer International Publishing, 2018: 593-607.

[42] Nathani D, Chauhan J, Sharma C, et al. Learning attention-based embeddings for relation prediction in knowledge graphs[J]. arXiv preprint arXiv:1906.01195, 2019.

[43] Zhang Z, Zhuang F, Zhu H, et al. Relational graph neural network with hierarchical attention for knowledge graph completion[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(05): 9612-9619.

[44] Zhao Y, Zhou H, Xie R, et al. Incorporating global information in local attention for knowledge representation learning[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 1341-1351.

[45] Fang H, Wang Y, Tian Z, et al. Learning knowledge graph embedding with a dual-attention embedding network[J]. Expert Systems with Applications, 2023, 212: 118806.

[46] Hamaguchi, T. , Oiwa, H. , Shimbo, M. , & Matsumoto, Y. . (2017). Knowledge

Transfer for Out-of-Knowledge-Base Entities : A Graph Neural Network Approach. 1802-1808. 10.24963/ijcai.2017/250.

[47] Xie R, Liu Z, Jia J, et al. Representation learning of knowledge graphs with entity descriptions[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2016, 30(1).

[48] Shi, B. , & Weninger, T. . (2018). Open-world knowledge graph completion[C]. Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[49] Shah H, Villmow J, Ulges A, et al. An open-world extension to knowledge graph completion models[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 3044-3051.

[50] Bi Z, Zhang T, Zhou P, et al. Knowledge transfer for out-of-knowledge-base entities: Improving graph-neural-network-based embedding using convolutional layers[J]. IEEE Access, 2020, 8: 159039-159049.

[51] Wang P, Han J, Li C, et al. Logic attention based neighborhood aggregation for inductive knowledge graph embedding[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 7152-7159.

[52] Zhao M, Jia W, Huang Y. Attention-based aggregation graph networks for knowledge graph information transfer[C]//Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II 24. Springer International Publishing, 2020: 542-554.

[53] Xie J, Jiang J, Xiao J, et al. Neighbor-T: neighborhood transformer aggregation for enhancing representation of out-of-knowledge-base entities[C]//International Conference on Computer Application and Information Security (ICCAIS 2021). Bellingham, SPIE, 2022: 1-11.

[54] Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg:

Association for Computational Linguistics, 2016: 2124-2133.

[55] Huang Y Y, Wang W Y. Deep Residual Learning for Weakly-Supervised Relation Extraction[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2017: 1803-1807.

[56] Qin P, Xu W, Wang W Y. Robust Distant supervision relation extraction via deep reinforcement learning[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2018: 2137-2147.

[57] Christopoulou F, Miwa M, Ananiadou S. A walk-based model on entity graphs for relation extraction[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg: Association for Computational Linguistics, 2018: 81-88.

[58] Zeng X, Zeng D, He S, et al. Extracting relational facts by an end-to-end neural model with copy mechanism[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 506-514.

[59] Liu J, Chen S, Wang B, et al. Attention as relation: learning supervised multi-head self-attention for relation extraction[C]//Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence. 2021: 3787-3793.

[60] Wang Y, Yu B, Zhang Y, et al. TPLinker: Single-stage joint extraction of entities and relations through token pair linking[J]. arXiv preprint arXiv:2010.13415, 2020.

[61] Sui D, Chen Y, Liu K, et al. Joint entity and relation extraction with set prediction networks[J]. arXiv preprint arXiv:2011.01675, 2020.

[62] Ye H, Zhang N, Deng S, et al. Contrastive triple extraction with generative transformer[C]//Proceedings of the AAAI conference on artificial intelligence.

2021, 35(16): 14257-14265.

[63] Ma L, Ren H, Zhang X. Effective cascade dual-decoder model for joint entity and relation extraction[J]. arXiv preprint arXiv:2106.14163, 2021.

[64] Zhao K, Xu H, Cheng Y, et al. Representation iterative fusion based on heterogeneous graph neural network for joint entity and relation extraction[J]. Knowledge-Based Systems, 2021, 219: 1-9.

[65] Zheng H, Wen R, Chen X, et al. PRGC: Potential relation and global correspondence based joint relational triple extraction[J]. arXiv preprint arXiv:2106.09895, 2021.

[66] Wang Z, Yang L, Yang J, et al. A Triple Relation Network for Joint Entity and Relation Extraction[J]. Electronics, 2022, 11(10): 1535.

[67] Bordes A, Weston J, Collobert R, et al. Learning structured embeddings of knowledge bases[C]//Proceedings of the AAAI conference on artificial intelligence. 2011, 25(1): 301-306.

[68] Bordes A, Glorot X, Weston J, et al. Joint learning of words and meaning representations for open-text semantic parsing[C]//Artificial intelligence and statistics. PMLR, 2012: 127-135.

[69] Han Y, Fang Q, Hu J, et al. Gaeat: Graph auto-encoder attention networks for knowledge graph completion[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020: 2053-2056.

[70] Vaswani, A. , Shazeer, N. , Parmar, N. , Uszkoreit, J. , Jones, L. , & Gomez, A. N. , et al. (2017). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[71] Sun Z, Deng Z H, Nie J Y, et al. Rotate: Knowledge graph embedding by relational rotation in complex space[J]. arXiv preprint arXiv:1902.10197, 2019.

[72] Che F, Zhang D, Tao J, et al. Parame: Regarding neural network parameters as relation embeddings for knowledge graph completion[C]//Proceedings of the

AAAI Conference on Artificial Intelligence. 2020, 34(03): 2774-2781.

[73] Cao Z, Xu Q, Yang Z, et al. Dual quaternion knowledge graph embeddings[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(8): 6894-6902.

[74] Xie R, Liu Z, Jia J, et al. Representation learning of knowledge graphs with entity descriptions[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016: 2659-2665.

[75] Fu C, Li Z, Yang Q, et al. Multiple interaction attention model for open-world knowledge graph completion[C]//Web Information Systems Engineering-WISE 2019: 20th International Conference, Hong Kong, China, January 19-22, 2020, Proceedings 20. Springer International Publishing, 2019: 630-644.

[76] Jia N, Cheng X, Su S, et al. CoGCN: Combining co-attention with graph convolutional network for entity linking with knowledge graphs[J]. Expert Systems, 2021, 38(1): 1-13.

[77] 田鹏,李定主,陈勇,李天宇,莫瑞峰,乔沛昊.基于作战指挥规则图谱的辅助决策方法[J].火力与指挥控制,2022,47(03):105-110.

[78] 钟昊,郭永贞,宗孝鹏. 基于知识图谱的智能决策辅助系统研究[C]//中国指挥与控制学会.第九届中国指挥控制大会论文集.兵器工业出版社,2021:133-138.DOI:10.26914/c.cnkihy.2021.011158.

作者在攻读硕士学位期间公开发表的论文

【1】Xi Tao, Hao Wang, Xiangfeng Luo and Pinpin Zhu. Few-Shot Link Prediction using Variational Heterogeneous Attention Networks. The 35th International Conference on Software Engineering and Knowledge Engineering (2023).

【2】Jin, W., Zhao, B., Yu, H., Tao, X. Improving embedded knowledge graph multi-hop question answering by introducing relational chain reasoning. Data Min Knowl Disc 37, 255–288 (2023). https://doi.org/10.1007/s10618-022-00891-8

作者在攻读硕士学位期间所参加的项目

【1】 上海市优秀学术带头人项目“千万级中小微企业动态知识图谱构建方法”。项目编号：20XD1401700；执行期限：2020.08-2023.09。

【2】 国家级军民融合项目，XXX 事件预测。执行期限：2021.02-2022.06。

致谢

本论文由王昊老师与骆祥峰老师指导完成，在此对两位老师的悉心指导与深切关怀表示真切的感谢。老师们虽然工作繁忙、时间紧张，但仍然在百忙之中抽出时间在学术与人生上给予我指导，并为我提供良好的科研与实践环境，使我得以将研究成果写在祖国的江山大地上。老师们在工作上勤勤恳恳，在育人上兢兢业业，在学术上勇于创新，这些优良的品质永远使我学习的目标。在此，谨向王昊老师与骆祥峰老师致以深深的敬意和由衷的感谢。

我要感谢实验室师兄、师姐们，无论是在学术上还是在实践中，你们都为我提供了许多宝贵的意见与支持，为我提供了至关重要的帮助。同时，我也要感谢我的同门、师弟和师妹们，是你们日日夜夜的陪伴与无微不至的关心才让我的步伐始终坚定，让我的信念不曾动摇。在此，诚挚的向实验室的各位同学们表达谢意，是你们的支持让我一直走到了这里。

我还要感谢我的父母，无论是经济上还是精神上，他们都是我最坚实的后盾，是他们给予我努力学习的信心与力量。我背负着他们的期望一路走来，也正是这份期望让我始终充满前进的力量。

最后，感谢所有帮助过我的家人、老师、朋友与同学，在这里，我仅用一句话来表明我无法言语的心情：感谢你们！