电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

学士学位论文

BACHELOR THESIS

论文题目 基于语法树的文档级关系预测算法设计与验证

学科专业	计算机科学与技术
学　　号	2020080902004
作者姓名	朱旭东
指导老师	康昭　副教授
学　　院	计算机科学与工程学院（网络空间安全学院）

摘　要

文档级关系抽取旨在识别单个文档中实体对之间的关系。它需要处理多个句子并对这些句子进行推理。最先进的文档级关系抽取使用图形结构来连接文档中的实体，来捕获文档中的实体对的交互。但是这些方法没有充分利用在句子级关系抽取中被充分研究的语法信息。在本文中我们以将语法树融合到文档级关系抽取中为主要研究内容，重点研究了使用依赖语法树，依存语法树进行文档级关系抽取算法的实现，以及怎么调整依赖语法树和依存语法树的在文档级关系抽取中的权重问题。我们利用依存语法树来聚合整个句子信息，并为目标实体对选择有指导意义的句子。同时我们利用依赖语法树对整个文档进行细粒度的分析，并选择其中重要的单词增强目标实体对的信息。文档级关系抽取将同时利用依赖语法树和依存语法树进行预测。通过在不同领域的数据集上的实验结果证明了该方法的有效性。

关键词：文档级关系抽取，依赖语法树，依存语法树，信息抽取

ABSTRACT

Document-level Relation Extraction (DocRE) aims to identify relation labels between entity pairs within a single document. It requires handling several sentences and reasoning over them. State-of-the-art DocRE methods use a graph structure to connect entities across the document to capture interaction between entity pairs in the document. However, this is insufficient to fully exploit the rich syntax information in the document, which is widely used in sentence-level Relation Extraction(RE). In this thesis, we focus on integrating syntax trees into DocRE as the main research topic, and investigate the effective and effient implementation of DocRE algorithms using dependency syntax tree and constituency syntax tree, as well as how to adjust the weight of dependency syntax tree and constituency syntax tree in the extraction. It uses constituency syntax to aggregate the whole sentence information and select the instructive sentences for the pairs of targets. Meanwhile, it exploits the dependency syntax in a graph structure with constituency syntax enhancement and selects the most important words between entity pairs based on the dependency graph to enhance the information of target entity pairs. Finally, DocRE will integrate the dependency syntax and constituency syntax to predict. The experimental results on datasets from various domains demonstrate the effectiveness of the proposed method.

Keywords: Document-level Relation Extraction, Constituency Syntax, Dependency Syntax, Information Extraction

目 录

第一章 绪 论.....1
1.1 研究的背景.....1
1.2 研究的主要贡献与创新.....2
1.3 研究的结构安排.....3
第二章 关系抽取研究.....4
2.1 关系抽取的定义.....4
2.2 文档级关系抽取.....4
2.2.1 基于序列的文档级关系抽取.....4
2.2.2 基于注意力机制的文档级关系抽取.....6
2.2.3 基于图的文档级关系抽取.....9
2.3 本章小结.....11
第三章 基于语法树的文档级关系抽取研究.....13
3.1 基于依存语法树的文档级关系抽取.....13
3.2 基于依赖语法树的文档级关系抽取.....13
3.3 融合两种不同语法树结果的文档级关系抽取.....13
3.4 本章小结.....13
第四章 基于语法树的文档级关系抽取实验结果.....15
4.1 文档级关系抽取抽取数据集说明.....15
4.2 DocRED 数据集实验结果说明与分析.....16
4.3 医学数据集实验结果说明与分析.....16
4.4 两种不同的语法树权重分析.....16
4.5 依赖语法树路径距离分析.....16
4.6 消融实验结果说明与分析.....16
4.7 本章小结.....16
第五章 结论.....18
致 谢.....19
参考文献.....20

第一章 绪 论

1.1 研究的背景

关系提取是信息提取中的一项关键任务，旨在对非结构化文本中实体对之间的关系模式进行建模。它有两种特定的场景：句子级关系提取和文档级关系提取。与句子级关系提取 [1,2] 不同，文档级关系提取涉及识别和提取句子边界以外的实体之间的关系，为分析提供了更广泛的上下文，并且更具挑战性。这项任务可以从非结构化的大型文档（如科学论文、法律合同或新闻文章）[3] 中自动构建和填充知识库，以更好地理解实体之间的关系。因此，文档级关系提取更好地满足了实际需求，最近受到了越来越多的关注。

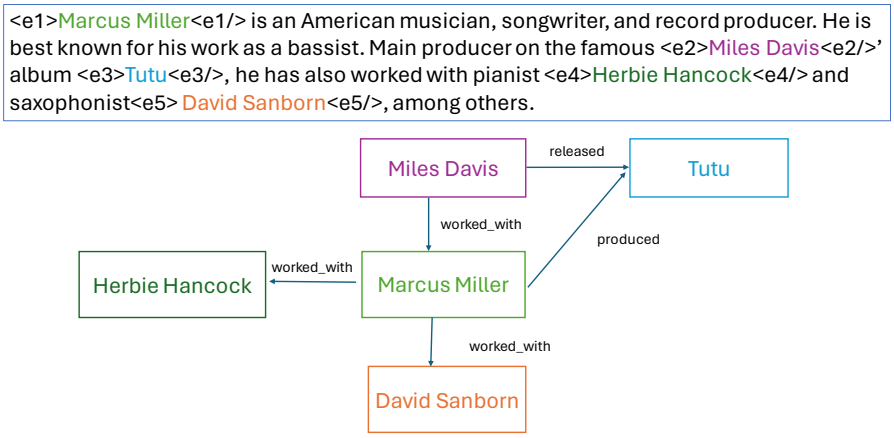
文档级关系抽取的目标是识别在同一文档内的实体对之间的关系。这是一项复杂的任务，因为它需要理解句子的内容并处理多个句子之间的交互。由于文档级关系抽取涉及多个句子，现在的研究有三个主要障碍 [4]：

1. 数量更多的潜在关系*:* 相比于单个句子，一个文档中包含的实体数量更多。由于我们需要预测每两个实体之间的关系，因此潜在的待预测关系随着实体的增加呈指数级别增加。
2. 处理实体共指: 一个实体通常在一个句子中只出现一次，而在一个文档中，它可以以多种形式出现多次。例如，在图 3 的句子 2 和句子 3 中，”He” 是”Marcus Miller” 的共同指称。
3. 处理长距离关系*:* 在文档级关系抽取中，我们通常需要预测跨句子之间的关系，这可能会导致实体对之间距离过长。然而，长句中通常包含不相关甚至有噪声的信息。*Huang* 等人 *[5]* 认为，在大多数情况下，最多只需要三个句子作为识别关系的支持证据。此外，这项工作指出，句子之间的重要性差异很大。不包含任何有价值的信息或关系的句子实际上阻碍了文档的理解。这也是我们这篇论文要解决的主要工作。

图 1.1是一个例子，它包括来自同一文档中的句子级关系和文档级关系。为了推断”Marcus Miller” 和”Herbie Hancock” 之间的关系，模型应该能够排除不相关实体的影响，并找出句子 3 中的”he” 一词指的是”Marcus Miller”。然而，由于现有的文档级关系提取模型 [6] 经常会被大量不相关的信息所淹没，从而无法捕捉到类似的关键信息。

由于以上这些原因，本文提出了新的一种文档级关系提取方法，通过融合语法信息作为辅助，来改进文档级关系提取。

图 1-1 文档级关系抽取的一个例子。图片上方是文档内容，文档中实体用不同的颜色标注，下方是文档级关系抽取的结果。



1.2 研究的主要贡献与创新

本文主要处理文档级关系提取的长距离关系。由于长句中通常包含无关甚至嘈杂的信息 [7]，对于文档级关系提取来说，隐含地学习指导性上下文是不够的 [6]。最近的一些文章通过预训练模型和注意力机制来学习文档中复杂的交互 [5, 8, 9]，他们通过创建更复杂、更强大的变体 transformer 来隐含的捕获实体之间复杂的互动。但是，更多的方法通过显示捕获实体互动信息来改善文档级关系抽取。他们首先构建由不同节点组成的文档图（例如实体的提及节点，实体节点，句子节点或者文档节点）[10, 11]，将指导性上下文转化为图。由于语法信息可以通过提供显式语法细化和子内容建模来帮助文档级关系抽取 [12]，最近的研究 [13, 14] 采用依赖图来合并语法信息和结构上下文。他们发现一个精心设计的图形可以帮助模型更好的捕获实体互动信息，缩短实体之间的距离。然而，正如 [6, 15] 所指出的，尽管预训练模型是用大量真实世界的文本数据进行训练的，但隐式学习的语法与黄金语法之间仍然存在很大差距。事实上，语法信息在句子级关系抽取中得到了广泛的应用 [16, 17]，但在文档级关系抽取场景下尚未得到充分的探索。

为了充分利用文档中的语法信息，本文融合了依存语法和依赖语法。我们主要采用依赖图和依存树来合并额外的语法信息，并使用依存树中的信息来进一步增强依赖图的表示。依赖语法描述了一个句子中单词之间的依赖关系，这些依赖关系对原始纯文本有很强的补充作用。而依存语法可以分层合理地组织单个句子的不同单词，消除了枚举不同单词组合的过程，同时保持了分层语法信息。在树形深度学习模块 [12] 的帮助下，我们可以收集具有适当权重的所有子内容，以解决句间依赖关系并融合依赖语法产生最终关系预测。

通过对三个公共文档级关系抽取数据集（DocRED [18]、CDR [19] 和 GDA [20]）的广泛实验下，我们证明我们的算法在很大程度上优于现有方法。我们在这篇文章中的主要贡献可以总结如下：

1. 我们建议利用依存语法树来聚合句子级信息，以弥补依赖语法树中的差距，并通过添加文档节点来改进依赖图，以减少实体对的距离，简化长句交互。
2. 我们分别用 Tree-LSTM 和 GCN 处理转化后的依赖语法和依存语法，并设置一个可学习的参数来调整它们的权重。
3. 实验结果表明，我们的模型在三个公共文档级关系抽取数据集上的测试优于现有方法，特别是在 DocRED 上，我们的模型将最先进方法的 IgnF1 提高了至少 1.56%

1.3 研究的结构安排

本研究的后续方式组织：第二章主要介绍文档级关系抽取的研究现状，在这一章中，我们对过往的研究结果进行总结。第三章介绍我们的研究工作，包括我们依赖语法和依存语法的使用方式，以及如何调节这两种语法的权重。第四章报告了我们研究的实验结果，包括在三个公共文档级关系抽取数据集上的不同指标的分数，与过往研究的分数对比以及分析；同时包括了对我们方法的每个模块的分析。

第二章 关系抽取研究

在这一部分，我们首先对关系抽取进行定义。然后我们介绍对文档级关系方法的回顾。

2.1 关系抽取的定义

在这篇文章中，我们遵循 Zhang 等人提出的定义 [21]。对于每个预测任务，我们有文档 $D = \{S_i\}_{i=1}^{n_s}$ 和实体集合 $E = \{e_i\}_{i=1}^{n_e}$。其中 n_s 是文档中句子的数量，n_e 是实体的数量，在句子级关系抽取中 n_s 被固定为 1，但是在文档级关系抽取中没有限制。文档中的句子被表示为 $S_i = \{w_j\}_{j=1}^{n_w^j}$，其中 w_i 是第 i 个句子中的第 j 个单词。每一个实体被表示为 $E_i = \{m_j\}_{j=1}^{n_w^i}$，其中 m_i 是第 i 个实体中的第 j 个提及。在句子级别设置中，一个实体只由一个提及表示，即 $\forall i \in [1, n_e], e_i = m_i$，但是在在文档级关系抽取中情况并非如此，其中实体可以通过多次提及来表示。为了获得实体的固定表示，以前的方法首先对所有提及的嵌入进行平均，即 $e_i = \frac{\sum_{j=1}^{n_m^j} m_j}{n_m^i}$，或者对所有提及的嵌入进行最大池化。然而，现在的方法都使用更加柔和的 logsumexp 表示，即 $e_i = \log \sum_{j=1}^{n_m^j} \exp(m_j)$，这样可以促进来自弱提及信息的积累 [22]。我们也采用这种方式表示每一个实体。

文档级关系抽取的目标是正确推断每个实体对 $(e_s, e_o)_{s,o=1,2,\ldots,n_e}; s \neq o$ 之间的关系。其中 e_s 是头实体，e_o 是尾实体。预测的关系是预定义的关系集合 R 或者 NA（没有关系）的子集。

2.2 文档级关系抽取

本节介绍了文档级关系抽取的最新进展，重点介绍基于序列的，基于注意力机制的和基于图的文档级关系抽取方法，并分析他们的优缺点。

2.2.1 基于序列的文档级关系抽取

执行文档级关系提取的一种方法是将文档视为一个长的增强的序列，并应用从句子级关系抽取导出的序列模型来识别特定实体之间的关系。

2.2.1.1 基于卷积神经网络的文档级关系抽取

Gu 等人 [23] 通过两个分离的模型来提取关系。他们首先将最大熵系统用于句子间的预测，卷积神经网络模型用于句子内的预测。然后，这两个结果被合并以

获得文档级别的关系。此外，他们还定义了一些启发式规则来从摘要中提取关系，例如将论文标题中提到的所有化学物质（或如有必要，将摘要中最频繁的化学物质）与摘要中提及的所有疾病相关联。

李等人 [24] 使用递归分段卷积神经网络进行化关系提取，并结合了领域知识、分段策略、注意力机制和多实例学习。该方法首先利用分段卷积神经网络来捕获单个实例的表示。然后，使用递归神经网络来聚合实例的表示，从而能够为实体对形成全面的文档级表示。

2.2.1.2 结合卷积神经网络和递归神经网络的文档级关系抽取

Mandya 等人 [25] 提出了一种长短期记忆和卷积神经网络的组合模型，该模型利用单词嵌入和位置嵌入来提取跨句子的 *n* 元关系。长短期记忆模型将结合单词嵌入和位置特征的变换矢量表示作为输入，将输出被馈送到卷积神经网络中。卷积神经网络被用来提取最重要的特征。这种方法的优点是独立于复杂的句法特征，如依赖树、共指消解或话语特征。因此被广泛应用到后续的改进模型中。

为了提取基因-疾病相关性，Wu 等人 [26] 还在其方法中结合了门控神经单元。在这里，卷积神经网络首先使用具有不同宽度的不同滤波器从单词表示中计算句子表示，以捕捉不同的特征。然后，门控神经单元将它们转换为文档表示。通过消融实验，他们发现，即使单独的卷积神经网络也可能捕捉到文档的一些重要特征，而额外的门控神经单元层可以提高其性能。这项研究后来被 Su 等人 [27] 进行了扩展。在 Wu 等人的系统之前引入了部分过滤步骤，来删除不包含任何实体对信息和模糊关系的段落。

2.2.1.3 以实体为中心的文档级关系抽取

在传统的句子级关系抽取中，对提及元组进行分类预测就已经足够了，但是在文档级关系抽取中却不是这样。因为任何实体都可能有多次提及，应用简单的提及级别分类方法会使模型在大量的提及中淹没，导致预测失败。为了解决这个问题，Jia 等人 [22] 提出通过对每个实体元组进行单一关系预测，使用以实体为中心的多尺度预测。它们通过对实体的所有上述表示应用 logsumexp 来获得实体表示。

2.2.1.4 基于序列的文档级关系抽取总结

这些方法通过将文档视为一个长的增强的序列，将句子级关系抽取的方法自然的映射到文档级关系抽取上。这些方法的优点是能够处理长文档，并且能够利用跨句子的关系信息。然而，当待预测实体对涉及到处理长距离并需要多跳推理

上下文内容时，这些模型不可避免地会遇到挑战。

2.2.2 基于注意力机制的文档级关系抽取

2.2.2.1 早期基于注意力机制的文档级关系抽取

Verga 等人 [28] 引入了双仿射关系注意力网络来计算实体之间的关系，它同时预测实体之间的句子内关系和句子间关系。为了实现这一点，它首先使用 transformer 层对令牌嵌入进行编码，然后它使用双仿射运算产生实体提及对之间的预测。最后，它使用 logsumexp 来汇集每个实体提及级别预测，以获得实体级别预测。该系统可以针对命名实体识别和关系预测进行联合训练，这可以提高对噪声的鲁棒性；此外，该系统在编码长序列方面相比于以往的方法更有效。

为了缓解正负关系之间的数据不平衡问题，Wang 等人 [29] 采用了两步微调过程和预训练模型。前者识别两次提及之间是否存在关系，后者仅使用先前识别的关系事实来训练模型来识别这些实体。

为了增强语言表示模型中的共指推理，Ye 等人 [8] 在各种下游任务上测试了他们的系统，这其中也包括文档级关系抽取。他们引入了一种称为提及参考预测的新的预训练任务。该任务利用了文章中的重复提及，并采用提及参考掩蔽来预测相应的参考，而不是依赖于监督的共同参考数据进行微调。此外，他们引入了基于新的训练目标，以鼓励基于上下文的单词选择和上下文敏感的表示，从而促进共指推理。

2.2.2.2 分层注意力模型

为了将推理信息从实体级聚合到句子级，再聚合到文档级，唐等人 [30] 使用层次推理网络进行预测。受 Bordes 等人 [31] 中给出的平移约束的启发，该约束对关系三元组 (r, e_s, e_o) 进行了建模。他们通过将实体提及和关系联合建模，获得了实体级别的表示信息。然后，他们应用语义匹配方法将实体级推理信息与每个句子向量进行比较，并使用多层双向长短期记忆模型获得句子级信息。最后，他们重新利用注意力机制来识别关键的句子级推理信息，这有助于以更全面的方式全面表征文档级推理。虽然这种方法十分有效，但是它的正确性受到广泛质疑，因为在文档级抽取中，两个实体对可以具有多个关系，一个实体可以与多个其他实体具有相同的关系。

Kuang 等人 [32] 认为，实体对之间的关系通常可以通过几个关键词来推断。为了从中受益，他们使用自注意记忆模块与预训练模型结合，通过寻找具有高交叉注意力实体对的单词嵌入来捕捉关键词特征。他们的模型非常简单，在专用和非专用数据集上都取得了很好的改进，并可应用于其他数据集。

2.2.2.3 基于实体表示的注意力模型

Han 和 Wang [4] 利用具有类型信息的文档级实体掩码方法来提供更多关于文档级关系预测中实体的信息。它在预训练模型的嵌入中引入了实体特征，用上下文信息和实体特定信息丰富了实体的表示。Han 和 Wang 首先使用预训练模型来获得实体提及的上下文化表示。然后，它使用一次性关系预测方法一次性处理所有实体对，因此文本只编码一次。

采用与 Jia 等人 [22] 相同的实体中心主义观点，Yu 等人 [33] 还解释了每一次提及之间的重要性，以代表关于特定关系的实体。他们使用关系特定提及注意力网络 (RSMAN) 将每个关系的基本语义合并到关系表示中，然后计算候选关系和给定实体对的提及表示之间的注意力。通过考虑这些事项，该模型通过对所有提及的表示进行加权聚合来获得实体的合成表示。这使得模型能够在不同的表示空间中捕获来自多个提及的信息，从而为不同的候选关系产生灵活的、特定于关系的实体表示。作为一种插入式方法，RSMAN 可以用作任何关系预测模型的方法中。

如前所述，在文档级关系预测中，一个文档包含多个实体对，这些实体对可以在文档中多次出现，并具有多种可能的关系。为了解决这些问题，Zhou 等人 [34] 在预训练模型的基础上使用了两种技术：自适应阈值和本地化文本池（ATLOP）。自适应阈值通过将多标签分类的全局阈值与可学习的依赖于实体的阈值进行切换来为每个实体对确定最佳阈值。而本地化文本池直接将注意力从预先训练的语言模型转移到有助于确定关系的相关上下文。这解决了对所有实体对使用相同实体嵌入的问题。ATLOP 通过与当前实体对相关的额外上下文丰富了实体嵌入，实现了更全面的表示，并为 DocRE 带来了巨大的飞跃。在我们的方法中，我们也采用了它的这两种技术。

2.2.2.4 基于远程监督标签的注意力模型

Xiao 等人 [35] 试图通过三个预训练任务来减少远程监督标签中引起的固有噪声：提及实体匹配，关系检测和关系事实对齐。提及实体匹配涉及文档内和文档间的匹配，第一种侧重于在文档内建立共引用连接，而另一种则捕获文档之间的实体关联；关系检测，主要检测被错误标记为负面的正面关系；关系事实对齐确保模型即使以不同的方式表示，仍然为同一实体对生成一致的表示，

接着，Tan 等人 [36] 提出了一种利用远程监督数据的多步骤方法。首先，他们使用注意力模块作为特征提取器，允许关注逻辑路径中的不同元素，并捕捉关系三元组（关系，头实体和尾实体）之间的相互依赖性。其次，为了处理类的不平衡，他们引入了一种关注长尾类的自适应焦点损失，使它们对整体损失的贡献

更大。最后，他们利用知识提炼来弥合手动注释数据和远程监督数据之间的差距。它由一个用少量人工注释数据训练的教师模型组成，该模型对大量远程监督数据生成预测，从而为预训练学生模型提供软标签。

2.2.2.5 基于结构推理的注意力模型

Xu 等人 [37] 基于当时文档级关系抽取模型的一个问题，即上下文推理和结构推理的阶段是分开的、因此缺乏对上下文表示的结构指导，从而引入了结构化自注意网络（SSAN），将结构依赖性包含在编码网络内和整个系统中。他们利用自注意机制的一个新扩展 [38]，通过考虑两个实体结构，一个用于同一句子中的实体对的共现，另一个用于单个实体的共指，有效地对其构建块内和所有网络层自下而上的提及依赖性进行建模。除了实体提及之间的依赖性，Xu 等人 [37] 还使用注意力模块来捕捉实体提及与句内非实体之间的关系。它们包含了一个额外的与注意力模块并行的模块，该模块以上下文化的查询键表示为条件，对结构依赖性进行建模。这种设计使模型能够受益于结构相关性的指导，确保上下文信息和结构信息有效集成。

2.2.2.6 基于联合学习的注意力模型

Ma 等人 [39] 解决了学习文档级关系抽取和证据提取的两个问题：高资源消耗和缺乏人工注释数据。他们使用证据引导注意机制进行的文档级关系提取，通过在证据提取上指导 ATLOP 模型 [9] 来完成证据提取和节省资源消耗。Ma 等人使用手动标注的实体对特定的局部上下文嵌入进行计算，并将注意力权重作为证据得分结合在标记上。这种方法允许他们引入来自证据提取的知识，而不引入额外的可训练参数或昂贵的矩阵计算。为了缓解数据的缺乏，他们使用了与 Tan 等人 [36] 类似的知识提取技术，利用远程监督数据来训练学生模型。

2.2.2.7 基于注意力的文档级关系抽取模型小结

综上所述，基于注意力的文档级关系抽取模型的研究已经取得了长足的进步。目前，除了早期的一些朴素方法，现有的方法大致可以分为以上五类，这些方法可以将实体嵌入、关系嵌入和上下文信息结合起来，以获得比基于序列的模型更好的文档级关系抽取结果。这些模型的共同特点是使用注意力机制来捕捉不同实体和关系之间的依赖关系，并将它们融合到统一的表示中。它们的优点在于能够捕捉到实体和关系之间的复杂关系，并能够在不同的表示空间中捕获来自多个提及的信息。这些模型的缺点是需要它们要么需要额外的远程监督数据，要么需要设计复杂的分层结构。而且由于实体之间的距离可能较长，它们中大多数方法

没有办法有效的处理长距离交互。然而，它们中的一些思想和假设十分有效（如第2.2.2.3 节的 ALTOP 的自适应阈值和本地化文本池等），因此在最近的研究中被广泛采用。

2.2.3 基于图的文档级关系抽取

在这一节中，我们首先在第2.2.3.1 节介绍早期的一些图的文档级关系抽取方法。然后我们将现有的基于图的文档级关系抽取模型分为五类：分别在第2.2.3.2节中介绍了联合任务的图文档级关系抽取方法，在第2.2.3.3节中介绍了以实体为中心的图文档级关系抽取方法，在第2.2.3.4节中介绍了基于异构图的图文档级关系抽取方法，在第2.2.3.5节中介绍了基于分层网络的图文档级关系抽取方法，在第2.2.3.6节中介绍了基于最优图结构的图文档级关系抽取方法。

2.2.3.1 早期的图文档级关系抽取方法

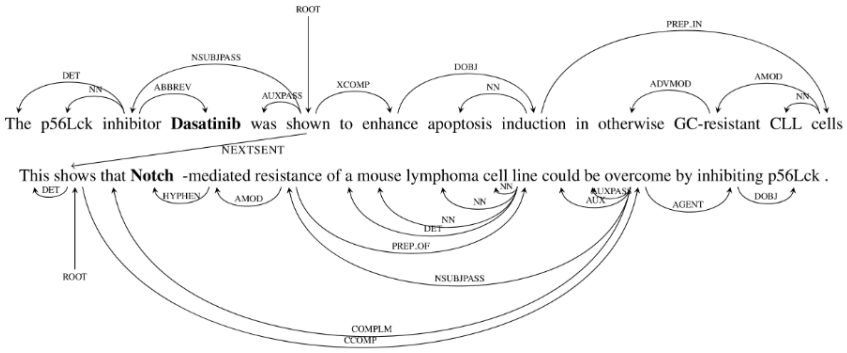


图 2-1 Quirk 和 Poon [40] 构建的文档图

Quirk 和 Poon [40] 引入了文档图的概念。它提供了一种统一的方法，通过将单词表示为节点，将一些启发式的关系（例如，依赖关系，邻接关系等）表示为边，来导出对待预测的关系进行分类的特征，图2.2.3.1是这种文档图的一个示例。这为文档级关系抽取的所有基于图的方法铺平了道路。Peng 等人 [41] 决定将这种图表示作为其框架的主干，其中包含一种新的图形长短期记忆模型。在经典的长短期记忆模型体系结构中，每个单元只有一个前置点（因此只有一个遗忘门）。图形长短期记忆模型的情况并非如此，通过从树形长短期记忆模型 [42] 中获得灵感，它可以具有多个前置点，包括通过不同边与同一单词连接的点，并为每个单元添加与先例数量一样多的遗忘门。从输入文本获得的单词嵌入被作为图形长短期记忆模型的输入，该图捕捉每个单词的上下文表示。然后，实体的上下文表示通过这种图形的连接进行组合，并由关系分类器使用。在多单词实体的情况下，Peng 等

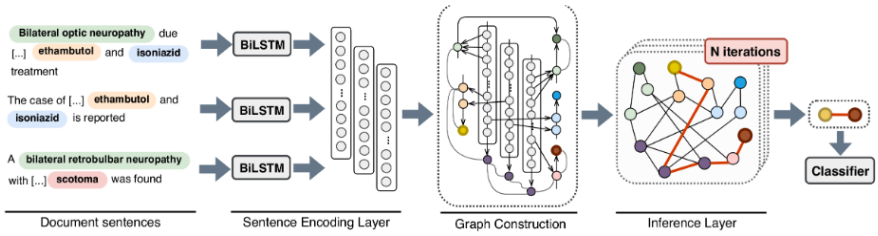


图 2-2 Christopoulou [43] 等人用于句子级关系抽取的基于图的随机行走的神经网络。该模型对句子进行编码，并将其输入迭代算法，以生成目标实体节点之间的边缘表示。为简洁起见，未显示某些节点连接。

人计算其单词表示的平均值来获得实体的嵌入向量，尽管现在大多采用 logsumexp 操作来获得更好的结果，但这种方法在当时是首创。至于图形长短期记忆模型的反向传播，他们将文档图拆分为两个有向无环图。一个有向无环图包含从左到右的线性连接链接相邻的单词以及其他前向指向依赖项（前向传递），而另一个有向无环图则涵盖从右到左的线性连接链接相邻的单词和后向指向依赖性（后向传递）

Christopoulou 等人 [43] 提出了一种仅用于句子级关系抽取的基于图的随机行走的神经网络，同时处理多对关系，并将句子表示为向无环图，这种方法的详细过程参见图2.2.3.1。他们后来在将他们的方法扩展到文档级别，引入部分连接的文档图而不是以前的完全连接的文档，使得文档级关系抽取向前迈出了一大步。与仅包含实体节点和单个边缘类型的句子级图相比，这种新的文档图由异构类型的节点和边组成，当实体提及可用时，这种方法可以使用多实例学习来进一步提高模型的性能。节点之间的连接基于预定义的文档级交互，目标是使用图中的其他现有边生成实体到实体的边表示，从而推断实体到实体关系。在这项任务中，Christopoulou 等人 [44] 还发现，文档级信息有助于找到句内关系。同年年初，同一团队 [45] 中提出应用图卷积网络实现消息传递。在该方法中，利用依赖语法边、相邻句子和单词边、自节点边以及共引用边来构造文档图，以连接句子的依赖树。Guo 等人 [46] 还提出通过注意力引导来扩展关系抽取的图卷积网络结构，以有效地关注任务的相关子图，而不修剪一些重要信息。

根据 Quirk 和 Poon [40] 中引入的依赖图，Gupta 等人 [47] 捕捉了两个实体之间的最短依赖路径，同时考虑了句内和句间关系。它们使用双向循环神经网络来表示通过依赖子树的最短依赖路径上的每个单词。他们提出的方法，基于句子间依赖的神经网络，通过将句子解析树的根节点连接到后续树的根，结合最短依赖路径来处理句子间关系。此外，他们采用了一种增强依赖路径，其中双向循环神经网络对句子间最短依赖路径上每个单词的依赖子树进行建模，然后提取显著的

语义特征。在大多数情况下，增强依赖路径模型比最短依赖路径模型更好，然而，他们没有在 DocRE 的基准数据集上进行测试。当将他们的方法与 [41] 进行比较时，他们指出，由于其非循环结构架构，他们的系统更简单、更高效。

此外，Minh Tran 等人 [48] 发现基于图的过往模型只关注图的边缘表示，而忽略了节点的表示。因此，这些模型错过了实体及其提及中包含的潜在重要线索，并阻止模型利用节点和边缘表示之间的关系，以及利用实体的表示向量之间的相似性找到相同实体的节点。因此，他们扩展了 Christopoulou 等人 [44] 的 EoG 模型。他们通过合并三个损失来提高模型表现：一个用于节点表示，一个用于结点边缘表示一致性，以及一个用于实体提及表示一致性。

2.2.3.2 基于联合任务的图文档级关系抽取

2.2.3.3 以实体为中心的图文档级关系抽取

2.2.3.4 基于异构图的图文档级关系抽取

2.2.3.5 基于分层网络的图文档级关系抽取

2.2.3.6 基于最优图结构的图文档级关系抽取

2.3 本章小结

1

第三章 基于语法树的文档级关系抽取研究

3.1 基于依存语法树的文档级关系抽取
3.2 基于依赖语法树的文档级关系抽取
3.3 融合两种不同语法树结果的文档级关系抽取
3.4 本章小结

1

第四章 基于语法树的文档级关系抽取实验结果

为了全面的评估我们的模型，我们在来自不同领域的三个文档级数据集上评估了我们的算法模型。我们在第 4.1节介绍了这些模型，在第 4.2和 4.3节说明了我们的结果。然后，我们在第 4.4节分析了两种不同的语法树权重，在第 4.5节分析了依赖语法树路径距离。最后，我们在第 4.6节进行了消融实验，并总结了我们的研究结果。

4.1 文档级关系抽取抽取数据集说明

表 4-1 三个公共数据集的统计数据

统计数据	DocRED	CDR	GDA
# 训练集	3053	500	23353
# 验证集	1000	500	5839
# 测试集	1000	500	1000
# 关系数量	96	2	2
# 平均每篇文章的句子数量	8.0	9.7	10.2

尽管文档级关系抽取方法可能因为要提取的关系类型（基因-疾病关系，一般类型的关系等）而有所不同，但大多数关系抽取方法都是在通用数据集和医学数据集上进行评估的。因此我们选择了一个通用数据集和两个医学数据集来评估我们的模型，表 4.1列出了这些数据集的统计数据。

- **DocRED** [18] 是一个从维基百科中人为标注的大型数据集。它包含从维基百科中采样的 5053 个黄金注释文档，132275 个实体、56354 个关系事实和 96 个关系类，每篇平均长度为 196.7 个单词，超过 40.7% 的关系对是跨句关系事实。实体部分包括经典的实体标签，如 PERSON、ORGANIZATION 和 LOCATION；而实体关系部分使用了 96 种关系类型，涵盖了广泛的科目，包括科学（33.3%）、艺术（11.5%）、时间（8.3%）、个人生活（4.2%）等。
- **CDR** [19] 是一个生物医学的文档级关系抽取数据集，由 1500 篇 PubMed 的摘要组成。这些摘要被随机分为三个相等的部分进行训练、验证和测试。这个数据集的预测任务是预测化学品和疾病之间的二元关系。
- **GDA** [20] 也是一个生物医学的文档级关系抽取数据集，包含 30192 篇摘要。这个大型数据集是通过对公开数据库收集的基因-疾病关联中的 30192 篇摘要（测试集为 1000 篇，其余按 80:20 的比例在训练集和验证集之间划分）进

行远程监督构建的。这个数据集的预测任务是预测基因和疾病之间的二元关系。

4.2 DocRED 数据集实验结果说明与分析

4.3 医学数据集实验结果说明与分析

4.4 两种不同的语法树权重分析

4.5 依赖语法树路径距离分析

4.6 消融实验结果说明与分析

4.7 本章小结

1

第五章 结论

1

致 谢

在攻读计算机学士学位期间，首先热烈感谢我的导师康昭教授。经过风风雨雨的研究，我得到了他的无私关怀和支持。在此特别表达感谢之情。我还要感谢我的一直以来的帮助者们，包括学院的老师、同学、同事，以及所有支持和关心我的人。我也要感谢我的家人，他们给予我强大的内心支持和生活的安定和稳定。

参考文献

[1] Dixit K, Al-Onaizan Y. Span-level model for relation extraction[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, 5308-5314.

[2] Lyu S, Chen H. Relation classification with entity type restriction[C]. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 2021, 390-395.

[3] Delaunay J, Tran T H H, González-Gallardo C-E, et al. A comprehensive survey of document-level relation extraction (2016-2022)[J]. arXiv preprint arXiv:2309.16396, 2023.

[4] Han X, Wang L. A novel document-level relation extraction method based on bert and entity information[J]. IEEE Access, 2020, 8(): 96912-96919.

[5] Huang Q, Zhu S, Feng Y, et al. Three sentences are all you need: Local path enhanced document relation extraction[C]. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Online, 2021, 998-1004.

[6] Bai J, Wang Y, Chen Y, et al. Syntax-BERT: Improving pre-trained transformers with syntax trees[C]. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 2021, 3011-3020.

[7] Gupta P, Rajaram S, Schütze H, et al. Neural relation extraction within and across sentence boundaries[C]. Proceedings of the AAAI conference on artificial intelligence, 2019, 6513-6520.

[8] Ye D, Lin Y, Du J, et al. Coreferential Reasoning Learning for Language Representation[C]. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 2020, 7170-7186.

[9] Zhou W, Huang K, Ma T, et al. Document-level relation extraction with adaptive thresholding and localized context pooling[C]. Proceedings of the AAAI conference on artificial intelligence, 2021, 14612-14620.

[10] Zeng S, Xu R, Chang B, et al. Double graph based reasoning for document-level relation extraction[C]. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 2020, 1630-1640.

[11] Liu H, Kang Z, Zhang L, et al. Document-level relation extraction with cross-sentence reasoning graph[C]. Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2023, 316-328.

[12] Duan Z, Li X, Li Z, et al. Not just plain text! fuel document-level relation extraction with explicit syntax refinement and subsentence modeling[C]. Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, 2022, 1941-1951.

[13] Sahu S K, Christopoulou F, Miwa M, et al. Inter-sentence relation extraction with document-level graph convolutional neural network[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, 4309-4316.

[14] Wei Y, Li Q. Sagdre: Sequence-aware graph-based document-level relation extraction with adaptive margin loss[C]. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2022, 2000 2008.

[15] Sundararaman D, Subramanian V, Wang G, et al. Syntax-infused transformer and bert models for machine translation and natural language understanding[J]. arXiv preprint arXiv:1911.06156, 2019.

[16] Xu Y, Jia R, Mou L, et al. Improved relation classification by deep recurrent neural networks with data augmentation[C]. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 2016, 1461-1470.

[17] Qin H, Tian Y, Song Y. Relation extraction with word graphs from n-grams[C]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 2021, 2860-2868.

[18] Yao Y, Ye D, Li P, et al. DocRED: A large-scale document-level relation extraction dataset[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, 764-777.

[19] Li J, Sun Y, Johnson R J, et al. Biocreative v cdr task corpus: a resource for chemical disease relation extraction[J]. Database, 2016, 2016.

[20] Wu Y, Luo R, Leung H C, et al. Renet: A deep learning approach for extracting gene-disease associations from literature[C]. Research in Computational Molecular Biology: 23rd Annual International Conference, RECOMB 2019, Washington, DC, USA, May 5-8, 2019, Proceedings 23, 2019, 272-284.

[21] Zhang Z, Yu B, Shu X, et al. Document-level relation extraction with dual-tier heterogeneous graph[C]. Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), 2020, 1630-1641.

[22] Jia R, Wong C, Poon H. Document-level n-ary relation extraction with multiscale representation learning[C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, 2019, 3693-3704.

[23] Gu J, Sun F, Qian L, et al. Chemical-induced disease relation extraction via convolutional neural network[J]. Database, 2017, 2017: bax024.

[24] Li H, Yang M, Chen Q, et al. Chemical-induced disease extraction via recurrent piecewise convolutional neural networks[J]. BMC medical informatics and decision making, 2018, 18: 45-51.

[25] Mandya A, Bollegala D, Coenen F, et al. Combining long short term memory and convolutional neural network for cross-sentence n-ary relation extraction[J]. arXiv preprint arXiv:1811.00845, 2018.

[26] Wu Y, Luo R, Leung H C, et al. Renet: A deep learning approach for extracting gene-disease associations from literature[C]. Research in Computational Molecular Biology: 23rd Annual International Conference, RECOMB 2019, Washington, DC, USA, May 5-8, 2019, Proceedings 23, 2019, 272-284.

[27] Su J, Wu Y, Ting H-F, et al. RENET2: high-performance full-text gene disease relation extraction with iterative training data expansion[J]. NAR Genomics and Bioinformatics, 2021, 3(3): lqab062.

[28] Verga P, Strubell E, McCallum A. Simultaneously self-attending to all mentions for full-abstract biological relation extraction[C]. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, 2018, 872-884.

[29] Wang H, Focke C, Sylvester R, et al. Fine-tune bert for docred with two-step process[J]. CoRR, 2019, abs/1909.11898.

[30] Tang H, Cao Y, Zhang Z, et al. Hin: Hierarchical inference network for document-level relation extraction[J]. Advances in Knowledge Discovery and Data Mining, 2020, 12084: 197-209.

[31] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]. Advances in Neural Information Processing Systems, 2013, .

[32] Kuang H, Chen H, Ma X, et al. A keyword detection and context filtering method for document level relation extraction[J]. Applied Sciences, 2022, 12(3).

[33] Yu J, Yang D, Tian S. Relation-specific attentions over entity mentions for enhanced document-level relation extraction[C]. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, 2022, 1523-1529.

[34] Zhao C, Zeng D, Xu L, et al. Document-level relation extraction with context guided mention integration and inter-pair reasoning[J]. ArXiv, 2022, abs/2201.04826.

[35] Xiao C, Yao Y, Xie R, et al. Denoising relation extraction from document-level distant supervision[C]. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 2020, 3683-3688.

[36] Tan Q, He R, Bing L, et al. Document-level relation extraction with adaptive focal loss and knowledge distillation[C]. Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 2022, 1672-1681.

[37] Xu B, Wang Q, Lyu Y, et al. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(16): 14149-14157.

[38] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. CoRR, 2017, abs/1706.03762.

[39] Ma Y, Wang A, Okazaki N. Dreeam: Guiding attention with evidence for improving document-level relation extraction[C]. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Dubrovnik, Croatia, 2023, (toappear).

[40] Quirk C, Poon H. Distant supervision for relation extraction beyond the sentence boundary[C]. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Valencia, Spain, 2017, 1171-1182.

[41] Peng N, Poon H, Quirk C, et al. Cross-sentence n-ary relation extraction with graph LSTMs[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 101-115.

[42] Tai K S, Socher R, Manning C D. Improved semantic representations from tree-structured long short-term memory networks[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), Beijing, China, 2015, 1556-1566.

[43] 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), Melbourne, Australia, 2018, 81-88.

[44] Christopoulou F, Miwa M, Ananiadou S. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs[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), Hong Kong, China, 2019, 4925-4936.

[45] Sahu S K, Christopoulou F, Miwa M, et al. Inter-sentence relation extraction with document-level graph convolutional neural network[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, 4309-4316.

[46] Guo Z, Zhang Y, Lu W. Attention guided graph convolutional networks for relation extraction[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, 241-251.

[47] Gupta P, Rajaram S, Schütze H, et al. Neural relation extraction within and across sentence boundaries[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 6513-6520.

[48] Tran H M, Nguyen M T, Nguyen T H. The dots have their values: Exploiting the node-edge connections in graph-based neural models for document-level relation extraction[C]. Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 2020, 4561-4567.