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 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 [**?**, **?**] 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）[**?**] 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。

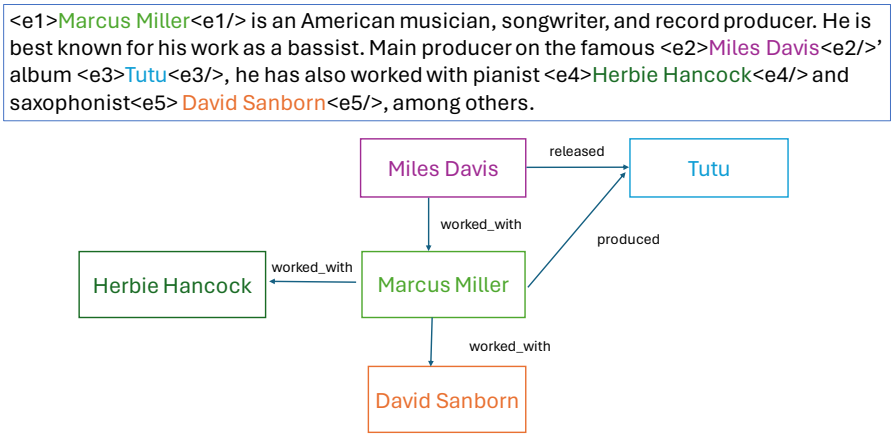
文档级关系抽取的目标是识别在同一文档内的实体对之间的关系。这是一项复杂的任务，因为它需要理解句子的内容并处理多个句子之间的交互。由于文档级关系抽取涉及多个句子，现在的研究有三个主要障碍 [**?**]：

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 *[?]* 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 **??**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 [**?**] 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 [?], 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 [?]。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 [?, ?]，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）[?, ?] ，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 [?]，最近的研究 [?, ?] 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 [?, ?] 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 [?, ?]，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 [?] 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 [**?**]、CDR [**?**] 和 GDA [**?**] ）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

2.1 句子级关系抽取

2.2 文档级关系抽取

2.2.1 文档级关系抽取的预测目标

2.2.2 基于序列的文档级关系抽取

2.2.3 基于注意力机制的文档级关系抽取

2.2.4 基于语法树的文档级关系抽取

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

4.1 文档级关系抽取抽取数据集说明

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。