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.....2
第二章 关系抽取研究.....3
2.1 句子级关系抽取.....3
2.2 文档级关系抽取.....3
2.2.1 文档级关系抽取的预测目标.....3
2.2.2 基于序列的文档级关系抽取.....3
2.2.3 基于注意力机制的文档级关系抽取.....3
2.2.4 基于语法树的文档级关系抽取.....3
2.3 本章小结.....3
第三章 基于语法树的文档级关系抽取研究.....5
3.1 基于依存语法树的文档级关系抽取.....5
3.2 基于依赖语法树的文档级关系抽取.....5
3.3 融合两种不同语法树结果的文档级关系抽取.....5
3.4 本章小结.....5
第四章 基于语法树的文档级关系抽取实验结果.....7
4.1 文档级关系抽取抽取数据集说明.....7
4.2 DocRED 数据集实验结果说明与分析.....7
4.3 医学数据集实验结果说明与分析.....7
4.4 两种不同的语法树权重分析.....7
4.5 依赖语法树路径距离分析.....7
4.6 消融实验结果说明与分析.....7
4.7 本章小结.....7
第五章 结论.....9
致 谢.....10
参考文献.....11

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 **??**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。

<e1>Marcus Miller<e1/> is an American musician, songwriter, and record producer. He is best known for his work as a bassist. Main producer on the famous <e2>Miles Davis<e2/>' album <e3>Tutu<e3/>, he has also worked with pianist <e4>Herbie Hancock<e4/> and saxophonist<e5> David Sanborn<e5/>, among others.

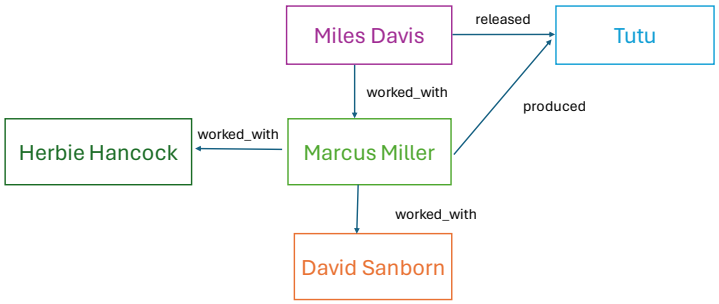

图 1-1 文档级关系抽取的一个例子。图片上方是文档内容，文档中实体用不同的颜色标注，下方是文档级关系抽取的结果。

1.2 研究的主要贡献与创新

1.3 研究的结构安排

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。

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.