一种无结构金融公告多元关系抽取方法

周映彤1　孟剑1　郭岩1　刘悦1　贺广福1　董琳2　程学旗1
（1. 中科院计算所 网络数据科学与技术实验室，北京 100190；2. 国家计算机网络应急技术处理协调中心，北京 100029）

论文摘要

金融公告信息披露了企业运营的关键数据，具有很高的价值。无结构金融公告中涉及到复杂的财务关系，即多元关系。本文设计了基于依存分析树和频繁子图挖掘的垂直域多元关系抽取方法TextMining，可以大大降低对数据集的依赖。进一步，受图卷积神经网络启发，设计了垂直域优化的FTA-GCN算法，在构建的适用金融公告数据集上，算法能够特别关注金融公告中常见的名词实体为核心的多元关系，实验结果表明算法具有良好的抽取效果。

系统模型

- 基于频繁子树挖掘和拓展的TextMining算法
- 基于TextMining和改进的注意力图卷积模型在Attention层编码融合的FTA-GCN抽取算法

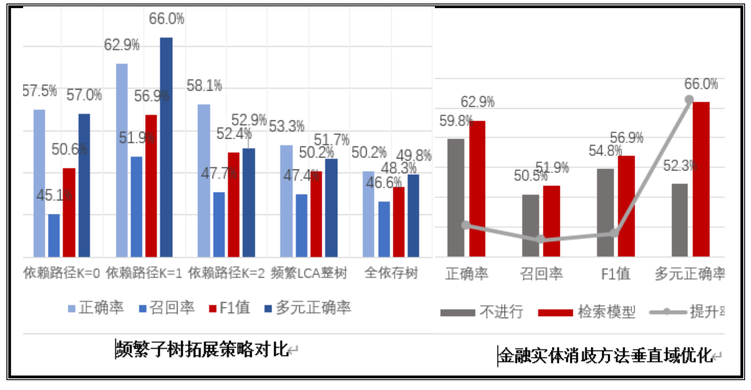
论文简介

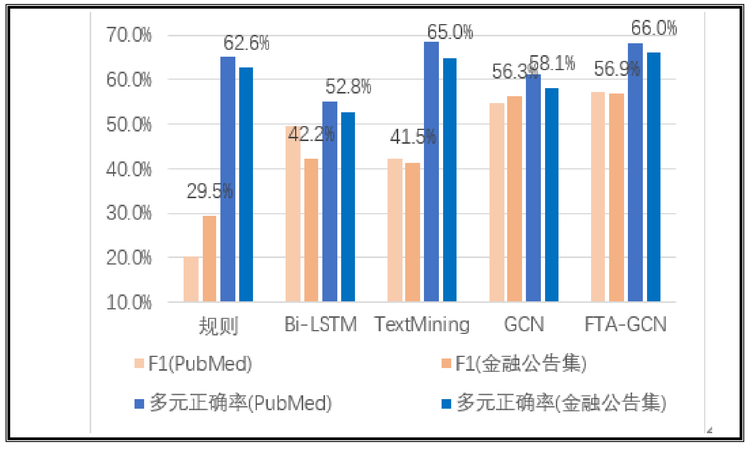
本文希望通过TextMining频繁子树挖掘和拓展算法作为移除句子中无关信息的一种新的方式。过于激进地“剪枝” 可能导致关键信息的丢失。以频繁子树为中心点的依赖路径拓展技术，保留依存分析树中到频繁子树距离为K以内的节点，有适应于依存树多元关系抽取的优势。为了更有效地利用这部分结构信息，本文进一步提出TextMining和改进的注意力图卷积模型在Attention层编码融合的FTA-GCN算法。

算法原理

从有效利用依存句法分析树结构出发，首先提出了基于频繁子树挖掘和拓展的TextMining算法，进行了无监督多元关系抽取实验。然后在金融公告多元关系抽取任务中实现了多头自注意力机制引导的图卷积模型。最后基于TextMining和改进的注意力图卷积模型在Attention层编码融合提出了FTA-GCN抽取算法。频繁子树结构信息被有效利用，提高召回率和非动词实体的关注度。在实体词典优化（垂直域优化）、实体消歧优化、词向量嵌入优化、剪枝策略优化、对非动词关注度提升等方面进行了细粒度实验，评价多元关系抽取效果。

实验仿真





论文结论

本文从有效利用依存句法分析树结构出发，首先提出了基于依存关系树频繁子树挖掘的TextMining算法，进行了TextMining无监督多元关系抽取实验。然后基于依存句法分析树，在金融公告多元关系抽取任务中实现了多头自注意力机制引导的图卷积网络抽取模型。最后基于TextMining和改进的注意力图卷积模型融合提出了FTA-GCN抽取算法。本章介绍了TextMining和FTA-GCN的算法和流程，分别在PubMed、自建金融公告数据集上，与基于规则、传统LSTM和现有GCN[16]的Baseline模型进行了对比实验。同时，在实体词典优化（垂直域优化）、实体消歧优化、词向量嵌入优化、剪枝策略优化、对非动词关注度提升等方面进行了细粒度实验，评价多元关系抽取效果，结果表明，研究算法在无结构金融公告的信息抽取任务上有效、鲁棒，具有实用性。

西安电子科技大学 XIDIAN UNIVERSITY