基于Bi-LSTM-CRF的中文命名实体识别的研究

摘要

关键词

Abstract

Keywords:

[第一章 绪论 2](#_Toc536366931)

[1.1研究背景 2](#_Toc536366932)

[1.2国内外研究现状 2](#_Toc536366933)

[1.3研究内容及工作 2](#_Toc536366934)

[1.4论文的组织结构 3](#_Toc536366935)

[第二章 命名实体识别的相关研究 3](#_Toc536366936)

[2.1引言 3](#_Toc536366937)

[2.2 传统命名实体识别方法介绍 3](#_Toc536366938)

[2.3 word2vec技术和深度学习技术介绍 3](#_Toc536366939)

[2.4 基于深度学习方法的命名实体识别 3](#_Toc536366940)

[2.5 本章小结 3](#_Toc536366941)

[第三章 改进的字符级嵌入方法的BiLSTM-CRF模型 3](#_Toc536366942)

[3.1 引言 3](#_Toc536366943)

[3.2 相关工作 3](#_Toc536366944)

[3.3 3](#_Toc536366945)

[3.4 实验结果及分析 4](#_Toc536366946)

[3.5 本章小结 4](#_Toc536366947)

[第四章 基于集成方法的CRF命名实体识别优化方法 4](#_Toc536366948)

[4.1 引言 4](#_Toc536366949)

[4.2 相关工作 4](#_Toc536366950)

[4.3 4](#_Toc536366951)

[4.4 实验结果及分析 4](#_Toc536366952)

[4.5 本章小结 4](#_Toc536366953)

[第五章 命名实体识别在企业风险识别中的应用 4](#_Toc536366954)

[第六章 总结与展望 5](#_Toc536366955)

[6.1 工作总结 5](#_Toc536366956)

[6.2 不足与展望 5](#_Toc536366957)

# 第一章 绪论

## 1.1研究背景及意义

互联网自诞生于上个世纪以来，正一步一步的改变这地球上每个人的生活。特别是伴随移动通信技术的革新，结合互联网技术和移动通信技术的移动互联网正深刻地改变着生活的方方面面。饮食、购物、交通、居住、社交、娱乐等等方面，在当下的移动互联时代都有了新的运作生态。随着移动互联技术的成熟和在“摩尔定律”下硬件成本的降低，人们越来越容易地融入互联网时代，根据中国互联网络信息中心发布的数据，截至2018年6月，我国网民规模已经超过了8亿，渗透率近六成，而其中移动互联网用户比例高达98%。海量的用户在使用互联网时也正有意无意地创造着海量的数据，而海量的数据中蕴藏着巨大的价值。这些数据的类型包括数值型数据、文本型数据、图片型数据、视频型数据、音频型数据等，利用好这些不同类型的数据可以创造出大量的经济价值和社会价值。

自然语言处理技术（Natural language processing）是计算机信息工程的一个子领域，目标便是处理和分析海量的文本数据，使得计算机程序可以利用词法、语法、语义等信息对自然语言文本完成识别、理解与输出等任务，例如词语分割、命名实体识别、关系抽取、机器翻译、自然语言生成、问答系统、情感分析等等。自然语言技术通过规则学习、统计学习等方法的研究与探索日臻成熟。近十年，表示学习、深度神经网络类机器学习技术给自然语言处理技术带来了新的探索与发展，在部分自然语言处理问题上可以达到良好而稳定的结果。自然语言处理技术在各行各业有着多种应用：社交媒体上的评论文本数据可以用来辅助监测舆情舆论的走向；财经新闻中包含诸多经济数据、公司运营情况，利用这些文本数据可以辅助量化交易的执行；利用新闻媒体中的海量文本数据，我们可以对用户兴趣话题进行建模，高效地为读者进行内容过滤和兴趣推荐；机器翻译技术可以将不同语言为载体的文献自动翻译，促进不同文化间的沟通和交流；知识图谱技术可以链接不同的人和组织，构造知识库，服务与多种商业应用。

命名实体识别（Named-entity recognition），又称实体抽取技术、实体分块技术，是自然语言处理技术的一个子问题。目标在于将非结构化文本中提及的命名实体抽取出来，例如人名，组织名，地点名，医疗术语法规术语，时间，数量，货币价值等等。例如在财经文章中需要准确地抽取企业名称、重要人物名称、货币价值等命名实体；在政治新闻中需要准确地抽取政治人物名称、国家地理名称、组织机构名称、事件名称等命名实体；在判决文书文本中，需要抽取出当事人名称、处罚条款、量刑情况、关联组织等信息。可以说，命名实体识别问题是自然语言处理最基础的任务之一，命名实体识别的准确率、召回率的高低直接影响着后续自然语言处理问题，例如信息抽取、文本分类、文本摘要、问答系统等等研究方向。

因而研究中文命名实体识别问题，对于中文自然语言处理技术的研究有着关键性的地位。通用命名实体技术对于中文命名实体识别有着不错的效果，然而中文与其他许多种语言的构词、语法有着诸多不同。特别是词语的边界模糊、无大小写和时态词型的变化、一词多性，一字多义、简称方法独特等等独特之处，因而相适应地根据中文语言的特点对通用命名实体技术进行优化有着很强的必要性。因而针对中文语言的各种特点，相适应地研究如何提高命名实体识别的效果就有着重要的意义。

## 1.2国内外研究现状

## 1.3研究内容及工作

因而本文针对上述问题采取了以下创新方法进行改进（1）在特征表示方面利用环境信息调整字词的语义向量，加入单个汉字的主题特征，提供更丰富可靠的信息用于中文命名实体模型的训练（2）针对中文词语边界不明、词性词义多变的特点，在序列标注模型的基础上，利用约束条件和滑动窗口方法对序列分块分类，提高识别的准确率和召回率（3）最终将基于这些改进方法的中文命名实体识别模型应用到企业产品风险监测系统中。

## 1.4论文的组织结构

# 第二章 命名实体识别的相关研究

## 2.1引言

## 2.2 传统命名实体识别方法介绍

## 2.3 序列标注方法相关研究

## 2.4 基于word2vec技术和深度神经网络的命名实体识别

### 2.4.1word2vec技术

### 2.4.2深度神经网络

## 2.5 主题模型相关研究

## 2.6 本章小结

# 第三章 改进的字符级嵌入方法的BiLSTM-CRF模型

## 3.1 引言

## 3.2 相关工作

## 3.3

## 3.4 实验结果及分析

## 3.5 本章小结

# 第四章 基于集成方法的CRF命名实体识别优化方法

## 4.1 引言

## 4.2 相关工作

## 4.3

## 4.4 实验结果及分析

## 4.5 本章小结

# 第五章 命名实体识别在企业风险识别中的应用

# 第六章 总结与展望

## 6.1 工作总结

## 6.2 不足与展望