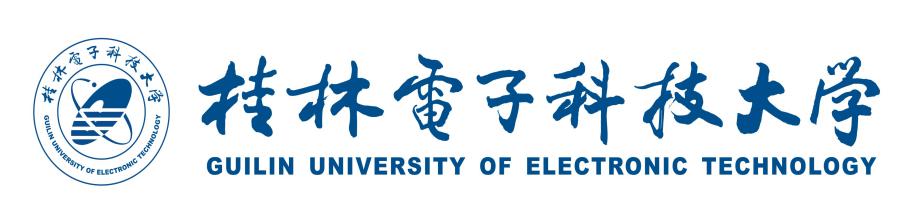
****

**毕业设计说明书**

题 目： 英语文本语篇关系

提取算法设计

院 （系）： 计算机与信息安全学院

专 业： 软件工程

学生姓名： 薛天杭

学 号： 1400330129

指导教师单位： 计算机与信息安全学院

姓 名： 黄桂敏

职 称： 教授

题目类型：🞎理论研究 🞎实验研究 🞎工程设计 🞎工程技术研究 🗹软件开发 2020年 6月 20日

**本科生毕业设计说明书（论文）独创性声明**

本人声明所呈交的毕业设计说明书（论文）是我个人在指导老师指导下进行的工作及取得的成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，文中不包含其他人或其它机构已经发表或撰写过的成果；也不包含为获得桂林电子科技大学或其他教育机构的学位或证书而使用过的材料。其他同志对本工作的启发和所做的任何贡献均已在毕业设计说明书（论文）中做了明确的说明并表示了谢意。

毕业设计说明书（论文）若有不实之处，本人承担一切相关责任。

本人签名： 日期：

**本科生毕业设计说明书（论文）使用授权说明**

桂林电子科技大学有权保留并向国家有关部门或机构送交毕业设计说明书（论文）的复印件，允许毕业设计说明书（论文）被查阅和借阅。本人授权桂林电子科技大学可以公布毕业设计说明书（论文）的全部或部分内容，可以采用影印、缩印或其它复制手段保存、汇编毕业设计说明书（论文）。

本人签名： 日期：

导师签名： 日期：

摘 要

随着科学的日渐发展，在人类的生产生活中机器所完成工作日益增多，自蒸汽时代以来，机器就已经成为了人类日常生活的一大助力，它可以进行能量交换、产生有用功、进行信息处理甚至代替人类劳动，是人类社会加速发展不可或缺的工具之一。到了当今这个时代，机器又因为人工智能被赋予了更多的用途。不管是在重工业、运输等传统行业，或者远程通讯、电话服务、科学发现等新兴行业，尤其是在计算机科学、医药、诊断、等方面，人工智能的出现从另一个角度重新定义了人们的生活，机器与人类的交互从未向现在这样频繁。

就目前而言，人工智能的研究方向主要包括图像识别、语言识别、智能机器人、[专家系统](https://baike.baidu.com/item/%E4%B8%93%E5%AE%B6%E7%B3%BB%E7%BB%9F/267819)和自然语言处理等。其中，自然语言处理(NLP)是人工智能研究方向一个相当重要的领域，该研究方向可以使用计算机作为工具，对书面形式或者口头形式的文本进行各式各样的处理和加工，使文本中的含义可以正确地被机器理解。该研究方向是解决人与计算机进行交互遇到的语言问题，同时以解决人与计算机的语言问题为基础，也能用来解决人与人之间因不同语种、不同语义等等遇到的各类语言问题。研究的方法主要是通过研究语言应用的模型，来建立对应的算法框架并对其进行完善、测评，最终设计出能够在我们日常的生活中可以使用的产品。

本论文设计了一个将英语文本中的语篇关系根据修辞结构理论(RST)提取出来的算法，并将提取出的语篇关系生成对应的语篇关系树用以展示。提取采用机器学习的方式通过训练Sequential模型来生成一个能够识别出英语文本中语篇关系的模型，并将其放入设计好的服务端，最终实现英语文本语篇关系的提取。

修辞结构理论是曼（W. Mann）和桑普森（S. Thompson）1987年在《修辞结构理论：一种文本组织的理论》一文中提出的，该理论基于文本局部之间关系对整个文本的组织进行一定程度上的描述，因其在各个语种中的适用性，该理论成为了文本组织最广泛使用的理论之一。在修辞结构理论中，文本可以先分为非重叠文本块，称为基本话语单元，随后在各基本话语单元之间建立各类联系，表示出整个文本的语篇关系。

关键词：人工智能；自然语言处理；修辞结构理论；机器学习

Abstract

With the development of science, machines have done more and more work in human production and life. Since the steam age, machines have become a great help to human daily life. They can exchange energy, generate useful work, process information or even replace human labor. They are one of the indispensable tools for the accelerated development of human society. In this era, machines are given more uses because of artificial intelligence. Whether it is in traditional industries such as heavy industry, transportation, or emerging industries such as remote communication, telephone service, scientific discovery, especially in computer science, medicine, diagnosis, etc., the emergence of artificial intelligence redefines people's lives from another perspective, and the interaction between machines and humans has never been so frequent.

At present, the research direction of artificial intelligence mainly includes image recognition, language recognition, intelligent robot, expert system and natural language processing. Among them, natural language processing (NLP) is a very important field of artificial intelligence research, which can use computer as a tool to process and process written or oral texts in various ways, so that the meaning of texts can be accurately understood by machines. The research direction is to solve the language problems encountered in the interaction between people and computers. At the same time, it is based on solving the language problems between people and computers. It can also be used to solve various kinds of language problems encountered between people due to different languages, different semantics and so on. The main research method is to establish the corresponding algorithm framework by studying the model of language application, improve and evaluate it, and finally design the products that can be used in our daily life.

This paper designs an algorithm to extract the text relations from English texts according to rhetorical structure theory (RST), and generates the corresponding text relation tree to display. Extraction uses machine learning to train sequential model to generate a model that can recognize the text relationship in English text, and put it into the designed server, and finally achieve the extraction of text relationship in English text.

The theory of rhetorical structure was put forward by W. Mann and S. Thompson in 1987 in the theory of Rhetorical Structure: a theory of text organization. The theory describes the organization of the whole text to a certain extent based on the relationship between parts of the text. Because of its applicability in various languages, the theory has become one of the most widely used theories of text organization. In the theory of rhetorical structure, text can be divided into non overlapping text blocks, which are called basic discourse units. Then, all kinds of connections are established between the basic discourse units to express the discourse relationship of the whole text.

**Keywords:** Artificial intelligence; natural language processing; rhetorical structure theory; machine learning

目 录

[引言 1](#_Toc42544043)

[1 绪论 1](#_Toc42544044)

[1.1 项目研究的背景 1](#_Toc42544045)

[1.2 项目研究的意义 1](#_Toc42544046)

[1.3 国内外研究现状 2](#_Toc42544047)

[1.4 本文主要的研究内容以及论文结构 2](#_Toc42544048)

[1.4.1 主要研究内容 2](#_Toc42544049)

[1.4.2 论文结构 2](#_Toc42544050)

[2 相关工作 4](#_Toc42544051)

[2.1 相关工作 4](#_Toc42544052)

[2.2 理论基础 5](#_Toc42544053)

[2.2.1 语篇关系提取 5](#_Toc42544054)

[2.2.2 修辞结构理论 6](#_Toc42544055)

[2.2.3 机器学习 7](#_Toc42544056)

[2.3 本章小结 8](#_Toc42544057)

[3 开发环境介绍 9](#_Toc42544058)

[3.1 PyCharm概述 9](#_Toc42544059)

[3.2 Python技术 9](#_Toc42544060)

[3.3 Python系统架构 10](#_Toc42544061)

[3.4 Keras简介 10](#_Toc42544062)

[3.5 TensorFlow简介 11](#_Toc42544063)

[3.6 本章小结 12](#_Toc42544064)

[4 系统分析 13](#_Toc42544065)

[4.1 可行性分析 13](#_Toc42544066)

[4.2 需求分析 13](#_Toc42544067)

[4.2 设计目标 13](#_Toc42544068)

[4.3 本章小结 14](#_Toc42544069)

[5 语篇关系提取算法设计 15](#_Toc42544070)

[5.1 问题描述 15](#_Toc42544071)

[5.2 算法综述 15](#_Toc42544072)

[5.3 设计思路 16](#_Toc42544073)

[5.4 模型设计与训练 17](#_Toc42544074)

[5.4.1 connection classifier 17](#_Toc42544075)

[5.4.2 relation classifier 18](#_Toc42544076)

[5.4.3 nuclearity classifier 19](#_Toc42544077)

[5.5 模型介绍 19](#_Toc42544078)

[5.5.1 全连接层 19](#_Toc42544079)

[5.5.2 Dropout方法 21](#_Toc42544080)

[5.5.3 总体介绍 23](#_Toc42544081)

[5.6 关系树的建立 23](#_Toc42544082)

[5.7 本章小结 24](#_Toc42544083)

[6 界面设计 25](#_Toc42544084)

[6.1 所用技术 25](#_Toc42544085)

[6.2 界面整体设计 25](#_Toc42544086)

[6.2.1 上传页面 25](#_Toc42544087)

[6.2.2 结果页面 26](#_Toc42544088)

[6.3 本章小结 26](#_Toc42544089)

[7 实验分析 27](#_Toc42544090)

[7.1 实验环境 27](#_Toc42544091)

[7.2 实验数据 27](#_Toc42544092)

[7.2.1 训练数据获取 27](#_Toc42544093)

[7.2.2 实验数据与结果 27](#_Toc42544094)

[7.3 本章小结 31](#_Toc42544095)

[8 结束语 32](#_Toc42544096)

[8.1 总结 32](#_Toc42544097)

[8.2 不足与展望 32](#_Toc42544098)

[谢 辞 34](#_Toc42544099)

[参考文献 35](#_Toc42544100)

# 引言

自然语言处理这一项技术已经深入到了我们的日常生活中，以至于很多时候人们并没有发现它的存在。早年五笔输入法火热一时，当时并非没有拼音输入法，但是由于拼音输入法的重音字过多，无法准确选取，并且拼音输入法联想功能十分孱弱，导致人们打字输入远不如五笔输入法准确高效。直到后来随着自然语言处理的发展，使得人们的拼音输入法使用体验大幅度上升，同时拼音输入法又有着上手门槛低等特点，最终取代了五笔输入法的主导地位。从上述案例我们可以看到自然语言处理在人们生活中的应用。在本论文中，主要研究了自然语言处理这一个大的方向中的一个小分支：关系提取。具体工作是通过机器学习训练模型，将一篇英语文本的语篇关系以基于修辞结构理论的实体关系提取出来，并最终形成语篇关系树用以展示，可以直观地相认展示出整篇英语文本的整体结构。通过此操作可以为后续的翻译，理解等等提供一定程度的便利。

# 1 绪论

## 1.1 项目研究的背景

从有计算机开始，人们与机器的交互没有停止过，但是如何让机器明白人类的语言一直是一个难以解决的问题，初步的解决方案是规定好各种规则，人类输入的语言符合这种规则，人类再让机器按照既定规则处理人输入的数据，最典型就是从底层语言到高级语言，这个方法让机器能够明白人类所想要的数据处理并执行，但是这也带来了一定的上手难度，没有接触过编程的人很难去与计算机进行交互。后来产生了操作系统，人们能够通过简单的学习掌握计算机的使用方法。操作系统的出现极大地促进了人们生产力的发展，给人们的生产生活带来了极大的便利。

但是，目前人们使用电脑说到底还是在既定的规则下通过自己符合计算机操作章程的方式去使用计算机，计算机还是没有能够理解人类的想法。即使语言学家与计算机领域的专家设计出一定的文本规则提供给计算机让计算机去理解，在实际的使用中也显得较为迟钝愚蠢笨。人工智能的出现给这种现象带来了巨大的转机，不管是通过传统的机器学习方式还是后来的深度学习，人们通过设计模型并提供一定数据给机器学习，机器终于能够一定程度上“看懂”、“听懂”人类的语言和文本，给人们的生产生活带来了新的转变。

## 1.2 项目研究的意义

正如微软公司的创始人比尔·盖茨所说：“语言理解是人工智能皇冠上的明珠”，人工智能领域的研究离不开自然语言处理[1]。如果一台机器不能理解人类的文字、语言，那这样一台机器又何谈人工智能呢。通过对自然语言处理这门学科的研究，我们可以将可换搬运到现实。不管是当下火热的小爱音箱、天猫精灵，还是早就存在并广泛运用的机器翻译、聊天机器人，这些产品的设计与研发都离不开自然语言处理这一关键性技术，这些产品也随着自然语言处理这一课题的研究不断深入，不断提升产品服务的广度与精度。所以只有科技前沿的技术不断地进行迭代更新，人工智能等一种技术应用的领域越多、越广，我们才能在这个信息化时代享受到更多的便利。

## 1.3 国内外研究现状

作为一门热门新兴学科，自然语言处理在国外的研究相对于国内的研究领先较多，不管是Facebook通过训练Skip-Gram模型摆脱了语料库的限制，还是Google通过超参数调节在BiLSTM-CRF上面取得了98%的几乎接近人工准确率天花板的研究以及Duplex技术都向我们展示了当前世界上最先进的自然语言处理的相关研究。

在国内自然语言处理的研究也没有落后太多，华为主攻自然语言处理中的语音技术、机器翻译以及对话都已在华为的各种产品服务中的得到了广泛的应用，其次还有小米、淘宝等企业，均在自然语言处理这一领域有不少建树。

## 1.4 本文主要的研究内容以及论文结构

### 1.4.1 主要研究内容

本文主要研究了一个基于修辞结构理论的语篇关系提取算法设计，尝试从修辞结构理论的角度来对一篇英语文本进行系统性的关系划分，将一个个基本话语单元从整篇文章中提取出来，添加到一个语篇关系树中去，最终直观地展示出各基本话语单元之间的关系。

### 1.4.2 论文结构

本论文总体来说分为十个部分。

第一个部分是引言和绪论，主要通过拼音输入法的普及引入自然语言处理的重要性，随后介绍了一下自然语言处理方面的项目研究背景，意义以及国内外的研究现状。

接下来对自己所做的工作进行了细分，表明语篇关系提取这一研究课题在自然语言处理中的位置。展示了当前在语篇关系提取这一方面的研究以及各位前辈设计的各类语篇关系提取工具。其后针对语篇关系提取、修辞结构理论和机器学习这三个学科的基础理论这三个本论文所涉及到的核心学科的基础理论进行了介绍。

为了实现本论文所设计算法，既需要使用编译软件进行开发，也需要各类学习框架提供支持，所以之后就对算法及软件的开发环境进行了介绍。

其后对本论文所设计的算法从可行性分析，需求分析及设计目标进行了系统的分析，明确要设计出一个什么样的算法，实现什么样的效果以及该算法实现后的具体实际用处。

在介绍与分析完该算法的理论基础与实际价值后，在第五章进入了语篇关系提取算法的算法设计，首先通过简单的问题描述提出在设计时可能会遇到的问题以及需要克服的困难，然后算法设计中讲解问题的解决以及讲述算法设计的各项基本流程。

算法展示过于简陋，所以在算法之外还设计了两个简单的页面用以展示算法的提取结果，为多次调试提供方便。

之后通过实验分析来对实验结果做一个展示，先简单阐述实验所使用的环境配置，随后在实验数据中讲解训练用数据以及实验数据，在该章节最后根据实验数据评判出实验的结果。

最后对整个论文进行总结，客观地评价自己的实验成果，指出自己的不足以及待改进的部分，为后续的学习改进做好准备。

在本论文的最后对在毕业设计过程中给我提供帮助的各位老师同学表达感谢并列出了本论文所涉及的参考文献。

# 2 相关工作

## 2.1 相关工作

目前与自然语言处理相关的研究不胜枚举，但是有关英语文本间的语篇关系的研究目前却没有太多，对于机器翻译而言，只有明白了一篇文本之间的各个句子之间的含义才能更好地理解这个句子整个英语文本中所代表的含义，所以对于这方面的研究的不断深入就是不断提高机器对于长语句，难语句的翻译能力。从另一个方向来说，当机器能够轻松地理解人们的自然语言时，人工智能才能够真正地大放异彩，为我们的生活带来更多的便利。

随着计算机技术和机译软件的发展，机器翻译的能力相对于以前有了十分明显的进步。但是机译在英汉两种语言的翻译转换，译文的准确率和可读性以及词性、词义的辨识等方面还存在着较大的问题[2]。当机器翻译遇到长的、难的句子时，解决方案也只是在数据库中寻找类似的句子来进行对比，输出一个一个类似的结果，采取该方案的好处是显而易见的，通过不断丰富数据库，使其能够包含人们日常的各类对话，翻译的准确度可以不断提升直至逼近百分之百。

但是采取该方法的缺点也十分明显，由于人类的文明一直是向前发展的，所以我们无法保证经历若干年的发展，人们的日常沟通所使用的词汇还是跟现在类似，倘若差距较大，则数据库需要不断进行填充与更新，致使其成为一个极其庞大的数据。

在2016物联网识别、信息和知识国际会议中曾提到基于机器学习的信息提取取代了基于规则的方法，成为关系提取的主要方法。然而，机器学习方法仍存在许多缺点和局限性，如大部分关系提取系统侧重于二进制关系提取。在大规模数据中提取n-元关系是一种趋势。此外，不需要大规模训练数据的关系提取系统将来会随着一些新的算法更受欢迎。关系提取是构建QA系统和信息分析等新型智能应用的重要基础技术。

由此可见使用机器学习进行文本各语句单元间关系的提取是大势所趋，在这个方向上，许多前辈已经进行了大量的研究。

SPADE是一个句子级话语解析器。使用了两个概率模型，它们使用句法和词汇信息来分割和分析文本。通过动态编程方法克服了算法复杂性问题。虽然SPADE不支持全文话语解析，但它证明了句法和话语信息之间的相关性及其从经验上识别结构和关系的能力，特别是在不存在信号提示词的纯文本上。

此外，Reitter提供了一种完全基于功能丰富的监督学习的方法，以生成文本级话语解析树。他的算法依赖于训练一组支持向量机分类器来评分和标记跨距之间的关系。

在Sagae中，提出了一个基于减少移位解析的话语解析器。解析器在 RST-DT 上训练，对依赖项和组成树使用过渡算法，并使用其自己的语法和依赖项标记。与Marcu的原始图表分析方法一起，它在准确性和速度方面有了显著的提高。解析器还包括基于词汇句法功能训练的二进制分类器的语义分段。话语分段器的表现优于Reitter中介绍的分段，速度明显更快。

Subba和Di Eugenio提出了一个基于归纳逻辑程序的词解析器，这是一种学习话语关系的第一阶逻辑方法。在其系统中，除了传统形式句法和话语提示信息外，规则还基于WordNet（Fellbaum，1998年）的语义信息、相似性度量、文本段之间的结构属性以及组成语义来学习。然后，采用移位减少分析模型来生成偏离过程的结构。作者采用一套定制的教学文本，手动分割成话语单元，并带有一套26种话语关系的自定义内容。对

在2007年Subba和Di Eugenio还提出了一种基于人工神经网络的统计话语分段。该系统采用多层感知器模型，并使用句法和词汇信息，特别是话语线索，对RST-DT进行培训用于减少过度拟合。

Le Thanh 等人提出了一种多步法算法，用于将文本分段，并将其根据范围组织成一个个连续级别的文本组织（句子级别、段落级别和文本级别）。其算法的多级特性减轻了组合爆炸效应。但是，在文本级别，算法必须对大量树进行评分，以提取最佳候选项。因此，这种方法对于大输入文本来说很不友好。此外，RST 语料中不支持在每个组织级别进行邻接假设的假设，因为大多数文档包含不能映射到有效子树的组织块。

Tofiloski等人认为，使用规则比自动学习方法有一定的优势，并呈现出一种基于规则的话语分段。他们提出的模型不依赖于特定的训练体，并通过插入更少但"质量"的边界支持高精度分割。但是，分段的完成方式不同于RST-DT 语料库中使用的分段指南。

## 2.2 理论基础

### 2.2.1 语篇关系提取

对于语篇关系提取的发展分为三个阶段，分别为基于规则的信息提取、传统的机器学习关系提取和开放信息提取。

(1)基于规则的信息提取

在机器学习还没有成为自然语言处理的人们方法之前，该提取方法较为普遍。通过语言以及计算机领域的专家设计专门的语篇关系提取规则，随后让计算机根据设计好的规则进行提取，但是由于基于规则所设计的提取系统高度上依赖于规则，因此该提取方法在通用化方面性能相对较差，且基于规则设计的提取算法要么过于简单，要么过于复杂，无法适用于各类场合，逐步被取代。

(2)基于传统机器学习的关系提取

由于机器学习的迅速发展，包括关系提取在内的众多自然语言处理研究领域都开始采用了机器学习算法，虽然早期基于机器学习的系统数量较少，但是由于机器学习系统相较于传统的规则系统在通用性能以及效率方面有明显的优势，所以机器学习方法逐渐成为了关系提取的主要方法，取代了基于规则的关系提取方法。

(3)开放信息提取

在之后随着互联网的发展，互联网上的文本信息已经逐步成为关系提取的主要目标。互联网上包含了大量的复杂的数据，如何利用互联网上未进行标记的非结构化的文本来提高关系提取系统的广义化成为了关系提取任务的新的挑战。随后就产生了许多无监督和半监督的机器学习方法，通过该方案完成公开信息提取任务。

传统的机器学习方法的基本思想是将关系提取问题转化为二进制分类问题，然后用标记文本对受监督系统进行训练。此外，受监督的关系提取方法包括基于功能的方法和内核方法。需要基于功能的方法手动设置文本的功能矩阵，而内核方法不手动标记。基于功能的方法在特定区域中可能有更好的结果。

### 2.2.2 修辞结构理论

修辞结构理论是由曼（W. Mann）和桑普森（S. Thompson）在1987年提出，该理论从文本的各个部分出发，认为文本是由各个部分根据一定的关系组合起来的，这个关系从一定程度上保证了文本的连贯性。此外，该理论还认为任何一篇文本中的每一个部分相对于其他部分来说都肯定担任着一个或者多个特定任务的角色，或者就是用来完成某一个整篇文章所需的特定功能。

修辞结构理论的关系可以分成两大类，分别为主从型的“核心/辅助关系”以及并列型的“多级核心关系”。多级核心各系共有五个，分别为对比、结合、列举、多级核心重述和序列。主从型关系进一步分为“表述”和“主题”关系[3]。

表述关系具体来讲就是通过该段文字加强读者对于摸个事物的某种倾向。类似于对核心文本所描述的内容的赞同或者接受程度。表述关系继续进行细分有背景、使能、证实、准备、重述、对照、让步、证据、动机和总结十类[4]。

主题关系则是让读者识别出文本所谈论的话题之间存在某种关系。该关系进行细分可以分成环境、详述、解释、非意愿性原因、另外、目的、非意愿性结果、除非、意愿性原因、无条件、意愿性结果、条件、评价、方式和解答十五个[5]。

修辞结构理论定义一组关系，以确定保存文本的两个相邻基本话语单元 （Elementary Discourse Units）的特定关系。根据交际内涵的重要性，这些关系分为两类：单核和多核关系。修辞关系也传达了关于哪个跨度更重要的信息来表达作者的目的。在单核关系中，其中一个涉及的基本话语单元被标记为核，这在表达交际意图方面相对重要;而另一种，提供支持或背景信息是标记为卫星。如果删除卫星的基本话语单元，实际想要展示的根本信息不回有太大的变化，没有实质性的影响。将文本用核以及卫星来展示出来成为了修辞结构理论的一个特点，这一个特点表明了核的转移可能影响话语焦点的转移，从而进一步影响综合交际意图的表达。

修辞结构理论的一个显著特征是，该理论可以应用于任何语言，包括计算机语言，并且它可以用来描述几乎所有类型的文本流派。这使得其成为了最广泛使用的理论之一，在应用面上有十分大的潜力，值得人们不断对这个理论进行补充和完善。

### 2.2.3 机器学习

机器学习就是对计算机一部分数据进行学习，然后对另外一些数据进行预测与判断。

该学科的核心是“使用算法解析数据，从中学习，然后对新数据做出决定或预测”。也就是说计算机利用以获取的数据得出某一模型，然后利用此模型进行预测的一种方法，这个过程跟人的学习过程有些类似，比如人获取一定的经验，可以对新问题进行预测。当前世界有关人工智能的研究大多集中在弱人工智能方面，即让机器拥有一定的观察和感知世界的能力，能够进行一定程度的理解以及推理。目前该领域的研究有望迎来突破性的发展，其缘由就是机器学习的不断发展。

总的来说机器学习是一门多领域交叉学科，涉及概率论、统计学、计算机科学等多门学科[6]。机器学习的概念就是通过输入海量训练数据对模型进行训练，使模型掌握数据所蕴含的潜在规律，进而对新输入的数据进行准确的分类或预测。

机器学习可以分为四类：

(1)监督学习，训练机器学习的模型的训练样本数据有对应的目标值，监督学习就是通过对数据样本因子和已知的结果建立联系，提取特征值和映射关系，通过已知的结果，已知数据样本不断的学习和训练，对新的数据进行结果的预测，通常应用在分类与回归中，主要的难点在于获取具有目标值的样本数据成本较高，训练集需要人工进行标注才能运用到机器学习中去。

(2)无监督学习，无监督学习跟监督学习的区别就是选取的样本数据无需有目标值，我们无需分析这些数据对某些结果的影响，只是分析这些数据内在的规律，通常运用在聚类分析上。相对于监督学习，无监督学习的优点在于数据不需要人工打标记，获取成本相对较低。

(3)半监督学习，半监督学习从原理上来讲是结合了监督学习和无监督学习产生的一种新的学习方法，通过半监督学习的方法在目前可以实现分类、回归、聚类的结合使用，是当下比较流行的方法。

(4)强化学习，强化学习是一种比较复杂的机器学习方法，强调系统与外界不断的交互反馈，它主要是针对流程中不断需要推理的场景，比如无人汽车驾驶，它更多关注性能。它是机器学习中的热点学习方法。

深度学习是目前关注度很高的一类算法，深度学习从原理上来讲就是解决机器学习中遇到的问题所采取的方法，跟上述的四中学习一样属于机器学习的子类。但是深度学习的分类方式与监督学习、半监督学习、无监督学习、强化学习这种分类方法不同，它采取了另一种分类方法即基于算法神经网络的深度来进行分类，通过该方法对机器学习进行划分可以将机器学习分成浅层学习算法和深度学习算法。浅层学习算法主要是针对一些结构化数据、半结构化数据的一些场景的预测，深度学习主要解决复杂的场景，比如图像、文本、语音识别与分析等。深度学习的过程就是利用深度神经网络来解决特征表达的一种学习过程，通过学习与模拟大脑的工作模式来解决生活中遇到的问题。

目前机器学习这一课题已经在我们的生活中得到了极为广泛的应用，有数据的地方机器学习就有用武之地。现在数据充斥在各个行业的各个领域，绝大部分行业都需要借助机器学习技术来指导下一步的决策方向。例如打车数据分析就可以根据四级在平台的历史数据，将其截单习惯、路线熟悉程度、路线拥堵情况、距离陈科远近、天气变化、乘客乘坐距离等信息输入到分类模型中，预测出司机简单的概率大小。

## 2.3 本章小结

本章节首先介绍了有关语篇关系提取的相关工作，简单讲述了语篇关系提取这一课题的研究进程，将该领域各前辈所做出的作品进行了介绍与比较。同时在本章节的理论基础中对于语篇关系提取、修辞结构理论以及机器学习的相关基础理论进行了讲解，方便对下文所涉及内容的理解。

# 3 开发环境介绍

## 3.1 PyCharm概述

PyCharm是一款优秀的由Jetbrains公司研究开发的python IDE。它的优点包括且不限于Jetbrains公司开发的其他集成开发环境的优点，PyCharm拥有优秀的代码编写功能，能够极大限度地发挥Python这个语言的特性，而且由于其强大的可自定义与拓展性，人们在使用它时可以绑定Textmate、NetBeans等工具，为人们提高生产力有十分积极的作用。

在功能方面，除了一款优秀的IDE应有的编码协助、自动格式化、自动导入等功能，他还支持代码分析、Python重构以及Django，除了方便用户一键补全代码、提取方法或超类、导入所需变量/常量外还能方便用户使用Django开发Web。

此外随着更新，PyCharm在2.5之后也支持Google App引擎，能够在Google App Engine 服务器上运行或部署对应用户的应用与服务，并确保代码满足所需的沙盒环境。

## 3.2 Python技术

Python这一门语言是在上世纪九十年代初由Guido van Rossum所设计出来的。从编码方式等方面来说，其本身也是由许多语言发展而来，例如ABC、Modula-3、C、C++、Unix shell以及其他脚本语言等[7]。Python的源代码跟Perl一样遵循GPL协议。现如今，Python的发展模式是由Guido van Rossum指导推进，另外有一个强大的核心团队进行维护更新。

Python是一门面向对象的语言，在实际的使用中，它表现出了非常强大的功能。上到爬虫、游戏，下到简单的文字处理均可以通过Python实现。同时Python又有着易上手，易阅读，易维护的特点， 这使得它对于新手来说十分友好。此外由于其强大的兼容性、完整的数据库接口、可移植、可拓展、可嵌入的特点，使它在商业上的用途也十分广泛，尤其是随着近几年人工智能以及深度学习的发展，Python的使用规模也越来越大。

## 3.3 Python系统架构

Python的系统架构如图3-1所示。



图3-1 Python系统架构图

左边的文件组展示了了Python这门语言提供的模块、库以及用户自定义的模块。

中间的核心部分展示了Python的解释器，一般含有词法分析器(Scanner)，语法分析器(Parser)，编译器(Compiler)，虚拟机(Code Evaluator) [8]。 分别负责将Python代码切分，建立抽象语法书，生成指令集合，最终执行字节码。

右边展示了Python运行的时候的环境，分别体现了对象/类型系统(Object/Type Structures)，内存分配器(Memory Allocator)以及运行时候的状态信息(Current state of Python)与核心部分的关系[9]。在这当中，对象/类型系统包含了人们在开发设计时的创建的各类用户自定义的对象和类型以及各种内建对象，内存分配器可以在进行创建对象申请时申请内存，运行时候的状态信息就是一个状态机制管理，该部分用来维护解释器在不同状态之间的切换动作。

## 3.4 Keras简介

Keras是一个用Python编写的高级神经网络API，它可以在CNTK，Theano或者TensorFlow等平台上进行运行，使用这些平台提供的后端接口。支持快速的实验是Kears一个十分显著的开发重点，这个特性能够使其以最快的速度将用户脑海中的想法转换为对应的实验结果，这样用户就可以花费更多的研究在编程之外的方面，帮助研究者进行更好的研究。

Kears相对于其他深度学习框架，他的特点主要有四点：

1.用户友好，Kears对于用户而言十分友好，它把用户体验放在了第一位，为用户提供了一套直观且简单的API，将常见用例所需用户进行的操作数量降至最低，而且可以在用户错误时提供清晰和可操作的反馈。

2.模块化，在Kears中，模型可以被理解为由独立的、完全可配置的模块构成的序列或图。不同的模块可以根据简单的显著组装在一起。特别是损失函数、神经网络层、初始化方法、优化器、正则化方法、激活函数，它们都是构建一个新模型十分重要的模块。

3.易拓展，在Kears中，添加新的模块是轻松且容易的，且由于其能够轻松地创建可以提高表现力的新模块，Kears更加适合高级研究。

4.基于Python实现，Kears没有特定格式的单独配置文件，模型定义在Python代码中，这些代码紧凑，已与调试且易于拓展。

由于以上所提到的特点，Kears在Netflix，Uber，Yelp，Instacart等众多网站上使用，尤其受以深度学习作为产品核心的创业公司的欢迎。此外，它也深受研究人员的喜爱，被许多大型的科学组织研究人员使用，其中特别有名的就是欧洲核子研究组织和美国宇航局等等。

## 3.5 TensorFlow简介

TensorFlow是谷歌进行开发并提供给人们免费试用的一个开放源代码软件库，人们一般使用它进行高性能数值计算[10]。在使用的过程中，用户可以借助其灵活的架构将计算工作轻松地部署到各种各样的平台如CPU、GPU、TPU等等。TensorFlow最初是由google brain这个团队中的工程师和研究人员一起开发的，在当下最主要的应用就是为机器学习和深度学习提供强力支持。其所支持内容包括但不只限于基本的向量矩阵计算、各种优化算法、各种卷积神经网络和循环神经网络基本单元的实现以及各类可视化的辅助工具等等[11]。

TensorFlow的特点远不止其高度的灵活性以及可移植性，它还支持众多语言来构建和执行计算图，除此之外它丰富的算法库以及完善的文档也是其十分突出的优点。

由于上述特点，使得TensorFlow在图形分类、音频处理、推荐系统和自然语言处理等场景下都有丰富的应用。它已经在不知不觉中渗入到我们生活的方方面面。

## 3.6 本章小结

本章节对开发软件的环境进行了详细的描述，对编译器与python的技术以及系统架构进行了较为仔细的讲解，方便了解Python这一语言在机器学习方面的优势以及Pycharm这一编译器对于python语言的编写所能提供的帮助。同时对算法设计所涉及的框架Kears与后端TensorFlow也进行了简单的介绍，展现出该框架的优势以及选择此框架进行机器学习任务的设计的原因。

# 4 系统分析

## 4.1 可行性分析

本次设计的算法设计使用Python作为编程所使用的语言，其作为一个相对比较热门的用于编写各类深度学习模型的语言有许多前辈珠玉在前，使用Python来进行算法的设计与分析可以类比前辈的成果，让自己更快进步。其次由于Python的语法相对简单，在设计的过程中也可以更加投入地设计算法，不需要在编写的过程中花费过多精力。所以选择Python设计该算法从编码层次是可行的。

本次算法设计所需使用的数据量不是特别大，所以计算机的配置需求不需要过分高，本次使用的个人计算机中央处理器为i5-8400，为一颗六核心六线程的中央处理器，主频为2.8G赫兹，动态加速频率为4G赫兹，除此外TensorFlow也可调用图形处理器进行计算，所以图形处理器也显得尤为重要。本次使用的个人计算机图形处理器为英伟达GeForce GTX1070, 是英伟达于2016年5月7日发布的新一代图形处理器，该图形处理器使用的是GP104核心，采用了帕斯卡架构，显存频率为8000MHz，工艺水平上使用的是16nm制程工艺，单精度浮点数达到了6.5TFlops，接近GTX Titan X的7.0 TFlops。中央处理器i5-8400加上图形处理器英伟达GeForce GTX1070配合16GB的内存，基本能完成本次机器学习所需要的硬件需求。

## 4.2 需求分析

自然语言处理在我们的日常生活中的应用越来越广泛，作为其中的一个小分支，语篇关系分析在其中有很重要的作用，机器通过对英语文本进行语篇关系分析可以方便机器理解其中文本中的含义，也可以方便一个信息在各语种之间进行转换不丢失其原本的含义，确保消息的准确性。

该论文最终所设计出的算法以及用于提取的系统可以为后来研究这一方面的人提供一定的帮助，也可以帮助刚刚接触到修辞结构理论以及语篇关系分析的人更快地了解这个学科的内容。

## 4.2 设计目标

本次论文计划设计出一个能够提取出英语文本当中各个句子之间语篇关系的算法。以此为最终目标，则需要一个能够形容文本中各段文本间的关系来进行语篇关系的描述。经过了解与学习，我选择使用修辞结构理论来描述英语文本里面各段之间的语篇关系。

修辞结构理论作为美国功能语言学流派的主要代表之一，有着为语言的分析与生成提供形式的框架的宗旨，并且其规划和定义的二十多种结构关系已经能近乎全面地展示我们当下所遇到的英语文本中的各种语篇关系。所以提取规则采用修辞结构理论进行实行。

由于基于规则的信息提取需要大量的语言学知识且能够用计算机语言成功将其演化为一个个适用于各语种文本的规则，个人知识面有限，没有办法采取这种方法去实现最终的设计目标。且由于自己能力实在有限，实在无法实现基于互联网的开放信息提取，所以我最终考虑采用基于传统机器学习的关系提取来实现这一设计。

综上所述，我计划使用基于传统机器学习的方法训练模型，设计出一个能提取出一篇经过修辞结构理论划分基本话语单元的英语文本的算法，并且通过一个简单的软件将其直接地展示出来。

## 4.3 本章小结

在本章节中首先对于整个算法与系统从软件和硬件的层面进行了简单的分析，最终得出当前所采用的设备以及软件满足基本任务需求的结论。在这之后对最终设计出的算法及软件进行了简单的需求分析，明确该算法及软件的实际作用。最后指出了该算法以及软件最终需要完成的内容与效果，确定了一个准确的设计目标。

# 5 语篇关系提取算法设计

## 5.1 问题描述

在考虑如何将文本中的语篇关系提取出来时有过若干想法，最初的设想是设计一款基于规则提取的算法，使其能够根据英语文本中的各种单词将其的语篇关系进行划分，但是在实际查找一定的资料查找后发现该方法在没有雄厚的计算机以及语言学的基础很难实现，考虑到自己能力有限，最终将问题得解决方法定在了使用传统机器学习来完成算法的设计。

在计划好使用机器学习后需要解决的问题就变成了如何调整训练模型使其具备语篇关系提取的能力，采取什么样子的语篇关系提取结构，如何获取该语篇关系的对应训练数据。

## 5.2 算法综述

想要将文本中的语篇关系提取出来首先需要将一篇文章分为若干个连续的基本话语单元。

举个简单的例子：

Farm lending was enacted to correct this problem by providing a reliable flow of lendable funds.

这一句话通过修辞结构理论将其划分为基本话语单元就可替换成：

【Farm lending was enacted】

【to correct this problem】

【by providing a reliable flow of lendable funds.】这三个部分。

其后使用预先设定好的修辞或话语关系将连续的基本话语单元彼此放在关系中，最终生成一个树状结构，该树状结构就能比较直观地展示出文本的各文本之间的关系。

比如上述文字即可转换成一个十分简单的关系树如图5-1所示

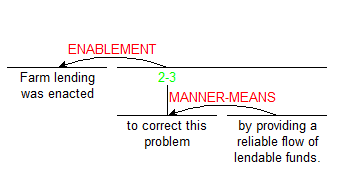


图5-1 简单的关系树

To correct this problem 对于前面的 Farm lending was enacted 起到一个使能的作用，其后的 by providing a reliable flow of lendable funds 又对前面所说起到了一个解释的作用。

当然这是单核的情况下，在RST中，多核关系可以采用任意数量的参数。例如图5-2，一个文本中可能包含若干个核心，从而导致表示文本建语篇关系的时候需要使用n元树来进行表示。通过将n元树成为二进制树可以简化提取时的复杂程度。这种转换可以实现连续嵌套的多核关系的展示，如图5-2所示。

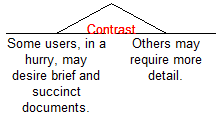


图5-2 多核心文本



图5-3 n元树转化为二进制树

## 5.3 设计思路

首先将文本划分为基本的话语单元，对每个基本话语单元进行分段，随后通过算法对两个基本话语单元的关系进行评估，判断这两个基本话语单元之间是否可能存在关系，将两个最有可能由修辞关系连接的基本话语单元合并为两个基本话语单元的语篇关系树，之后不断进行重复，直至最终完成所有基本话语单元的判断，结束，输出该语篇关系树。该方法可以保证最终输出的语篇关系树接近完美。

整个工作流程可由图5-4表示



图5-4 整体工作流程图

在实际的设计中，语篇关系标记的任务通过进行有监督的分类任务来进行模型训练。

## 5.4 模型设计与训练

模型用于修辞结构理论的标签识别，总共可以分为三个部分。

### 5.4.1 connection classifier

connection classifier部分的主要效果是在上一个的基础上将各个基本话语单元形成的简单的语篇关系树整合为最终一个整体的语篇关系树，具体模型如下图5-5所示，其中第一层全连接层也是输入层，数据由此进入模型进行运算，最后一层全连接层为输出层，输出预测的数据用来与正确结果进行核对。

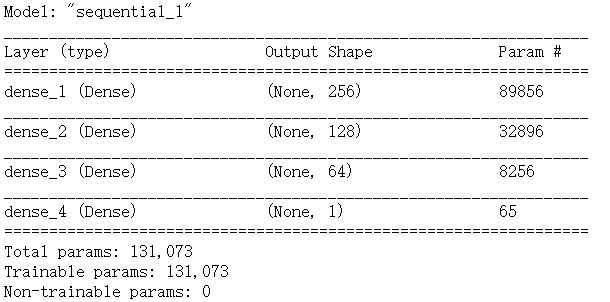


图5-5 connection classifier模型

### 5.4.2 relation classifier

relation classifier部分的主要效果是将文本中的各基本话语单元的关系识别出来，具体模型如下图5-6所示，其中第一层全连接层为输入层，数据由此进入模型进行运算，最后一层全连接层为输出层，输出预测的数据用来与正确结果进行核对，Dropout数值设置为0.3。

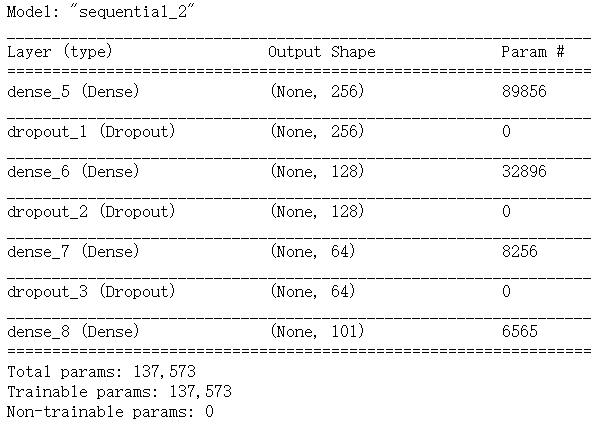


图5-6 relation classifier模型

### 5.4.3 nuclearity classifier

nuclearity classifier部分的主要效果是判断出上面所获得的各类语篇关系之间的层级关系，为后续生成一个语篇关系提供基础，具体模型如下图5-7所示，其中第一层全连接层为输入层，数据由此进入模型进行运算，最后一层全连接层为输出层，输出预测的数据用来与正确结果进行核对，Dropout数值设置为0.5。

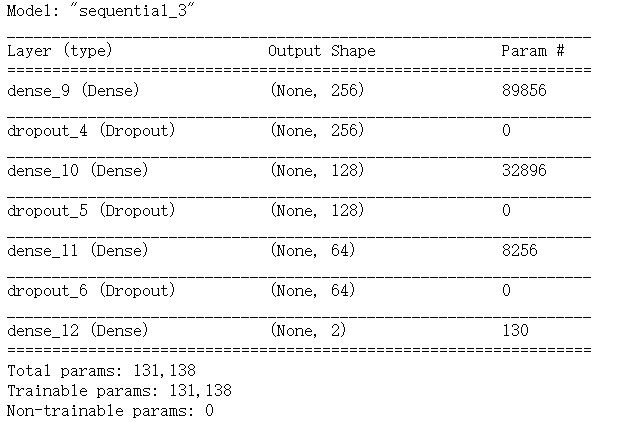


图5-7 nuclearity classifier模型

## 5.5 模型介绍

上述三个机器学习模型均由全连接层以及Dropout组成。

### 5.5.1 全连接层

全连接层指的是层中的每个节点都会连接它下一层的所有节点，它是模仿人脑神经结构来构建的，其实质就是矩阵相乘。由于它在数学上满足交换律和结合律，因此可以用并行化来加速计算，cuBLAS就是Nvidia为深度学习提供的数学加速运算库，它已经集成到Pytorch、Tensorflow这些深度学习框架中。

在上述模型中，全连接层体现的主要效果如图5-8所示。



图5-8 全连接层前向传播

全连接层对模型的影响参数总共有三个：1.全连接层的总层数（长度），2.全连接层的神经元数（宽度），3.激活函数。

图5-9通过一个简单的网络展示了全连接层的工作过程：



图5-9 全连接层工作过程简图

其中，X1、X 2、X 3为全连接层的输入，A1、A 2、A 3为输出，则可有如下公式如公式5.1，公式5.2，公式5.3：

1（公式5.1）

2（公式5.2）

3（公式5.3）

也可写为如下矩阵：

=\*+

其中W和B分别为对应的权重系数以及偏置系数。

### 5.5.2 Dropout方法

在机器学习的模型中，如果遇到模型的参数太多，但是实验所能使用的训练样本又太少的时候，训练出来的模型相比于其他模型十分容易产生过拟合的现象。过拟合这个现象是在训练神经网络时经常遇到的问题，该问题的具体表现在：模型在训练数据上的表现为损失函数较小，预测准确率较高；但是在测试数据上的表现为损失函数比较大，预测准确率较低。

由于过拟合在机器学习中十分常见，并且如果训练出模型出现过拟合的情况，那么最终训练出的模型几乎完全不能使用。人们在进行机器学习时为了解决过拟合的问题，一般都会采用模型集成的方法来进行规避，该方法通俗来讲就是训练多个模型进行组合。虽然该方法确实有一定的效果，但是如果采取该方法，进行训练模型所花费的大量时间就成为了另一个的问题，而且该操作不仅使模型训练的时间成本大大增加，测试模型所需花费的时间也会大大增加。

Dropout可以比较有效的缓解过拟合的发生，在一定程度上达到正则化的效果。

假设我们要训练这样一个神经网络，如图5-10所示。



图5-10 标准的神经网络

假设上图的输入是x，最终的输出结果为y。进行正常的流程，第一步就是将x通过网络前向传播，在这之后将误差反向传播，通过反向传播进行反馈，根据反馈来决定下一次更新的参数，随后继续让网络进行学习。假如上述神经网络使用Dropout之后，过程就会变成如下另一种流程：

(1) 首先随机挑选一部分网络中的隐藏神经元进行临时的删除，输入输出的神经元与之前普通的神经网络保持一致。具体表现如下图5-11中所示，临时被删除的隐藏神经元通过虚线表示。



图5-11部分临时被删除的神经元

(2) 之后过程与普通的神经网络类似，将x通过修改后的网络前向传播，之后将得到的损失结果通过修改后的神经网络进行反向传播。在部分训练样本执行完这个过程后，在没有被删除的神经元上按照随机梯度下降法更新对应的权重系数以及偏置系数。

(3) 随后就是恢复被删掉的神经元然后不断的重复上述过程，每一次都对权重系数以及偏置系数进行部分更新（保留在神经网络中的部分得到更新，被删除的神经元参数不进行变动）。

Dropout在神经网络中的主要就是用来解决过拟合，其工作与原理主要可以总结为三方面。

1. 取平均，假设我们使用相同的数据对五个不同的神经网络进行训练，一般来说得到的五个结果不会完全相同，此时对获得的五个结果取平均或者取相同数据较多的结果作为最终结果，那么该结果的正确率会得到显著的提升。采取类似于这种的综合起来取平均的方法可以十分有效地防止过于拟合的问题。该理论的原理是不同的神经网络可能会产生不同的过拟合，取平均可能会让部分拟合之间相互抵消，最终达到整体上减少过拟合的效果。

2. 减少神经元之间复杂的共适应关系，该原理主要是因为引用Dropout可以使两个神经元不一定每次都在一个经过临时删除神经元后的网络中出现。此举会保证权值的更新不再被存在固定关系的隐含节点影响，有效避免了某些特征只有在其他特定特征下才有效果的情况，通过此举迫使网络去学习鲁棒性更高的特征，简单来讲就是保证最终设计出的神经网络不会对部分特定的线索过于敏感，即使没有特定线索的存在，它也可以从众多的其他线索中学习到更加共性的特征。

3. Dropout类似于性别在生物进化中的角色，各类物种为了在环境中生存下去一般都会倾向于努力适应当前的环境，假设环境突变，物种就难以做出及时的反应。性别的出现让物种可以更快进行繁衍与变异，提高诞生可以适应新环境变种的概率与效率。有效的阻止过拟合的方法，就是避免物种因环境改变导致灭绝。

### 5.5.3 总体介绍

在上述全连接层中曾提到的激活函数在模型中的主要作用就是增加了模型的非线性表达能力。除开激活函数，剩余的变量就只有全连接层的总层数以及全连接层的神经元数这两个因素。

在算法的设计过程中曾经尝试使用两层全连接层以及三层全连接层来完成语篇关系的提取任务，但是结果都不是很理想。

使用两层全连接层进行提取最主要的问题在于不管是训练时或者是测试时的准确率都偏低，使用三层全连接层时准确率有了一定的提升，但是由于数据量较小，最终的训练结果依然不是很理想。

最后决定采用四层全连接层进行训练，在训练初期确实准确率提高十分明显，但是在测试时遇到了训练数据损失较小、准确率较高，但是测试时准确率十分低的情况，通过查询相关资料了解到这种情况属于过拟合，因此引入了Dropout作为解决过拟合的解决方法，同时通过尝试，Dropout设置为0.5时nuclearity classifier模型以及connection classifier模型表现最好。训练时的数据与测试时的数据准确率均在百分之九十以上。

然而relation classifier在Dropout设置为0.5时表现相当糟糕，在尝试后将其Dropout设置为0.3则可获得相对较高的准确率并且测试集所表现出的情况与测试集的数据没有太大出入。

事实证明最终的模型在训练时准确率较高且在测试时的表现与训练时没有太大差异。

## 5.6 关系树的建立

首先确定好所有的基本单元之间可能会出现的关系，本次设计总共使用了18种修辞关系。用R来表示其中的一个个节点，R=｛R1，R2，R3，R4，······Rn，｝，对于前面的所有数组均有Ri =< RRi , Lefti , Righti >当RRi ∈{Attribution; Background; Cause; Compassion; Condition; Contrast ; Elaboration ; Enablement ; Evaluation ; Explanation ; Joint ; Manner-Means ; Summary ; Temporal ; Topic-Change ; Topic-Comment ; Same-Unit ; Textual-Organization }且Lefti , Righti∈｛Nucleus ; Satellite｝时。

要想最终能够生成一个完整的语篇关系树，对于每个Ri 三个数值均不能为空。在应用的时候根据RRi 的修辞关系来决定Lefti , Righti 这两个数值是Nucleus还是Satellite。

除以上条件外，还需要保证生成树的所有叶节点都是基本话语单元，所有非叶节点都携带有修辞关系。

确保上述条件成立后开始设计语篇关系树，首先创建一个列表，按从左到右的读取顺序对输入文本的所有基本话语单元进行压缩。随后调用之前训练好的relation classifier模型对各基本话语单元之间的修辞关系进行预测。选择可能性最大的两个基本话语单元，通过训练好的nuclearity classifier模型对两个语篇关系进行判断，最后通过训练好的connection classifier模型判断将其放在关系树的哪一个部分。

确定好上述基本话语单元在语篇关系树中的位置后即可将这两个基本话语单元从列表中删除并加入到新建的语篇关系树中（第一组基本话语单元分别位于根的左右两侧的子级），之后继续对下面的基本话语单元按照上述步骤进行判断与删除，直至列表为空，语篇关系树建立完毕。

## 5.7 本章小结

本章节主要的工作就是介绍本次毕业设计所设计的算法，首先描述了设计算法所需要解决的问题，随后在算法综述中将语篇关系提取的基本原理进行了讲述并在随后所计划设计的算法建立基本的设计思路，描述清楚整个设计出的系统应该如何工作，流程大致是什么样子。

本章节的重点就是三个使用多重全连接层进行设计的算法模型，这三个模型通过调参，添加Dropout处理过拟合，最终基本实现语篇关系提取的功能，在讲解完三个算法模型后，也对最终的语篇关系树如何建立进行了简单的介绍，包括语篇关系树中存在的关系以及每一个节点的各类内容。

# 6 界面设计

## 6.1 所用技术

得益于Pycharm极强的拓展性，使得用户可以在Pycharm上进行简单的网页的编写，由于本次设计算法为主要核心，所以页面使用HTML进行了简单的设计，基本满足使用要求。

HTML即超文本标记语言。该语言是一种标志性的语言。它主要内容由一系列标签组成．通过特定的标签可以将网络上的各式各样的文档格式进行统一，将分散的Internet资源连接为一个缜密的逻辑整体供用户使用[12]。

超文本是一种组织信息的方式，它通过超级链接方法将文本中的文字、图表与其他信息媒体相关联。这些相互关联的信息媒体可能在同一文本中，也可能是其他文件，或是地理位置相距遥远的某台计算机上的文件。这种组织信息方式将分布在不同位置的信息资源用随机方式进行连接，为人们查找，检索信息提供方便[13]。

## 6.2 界面整体设计

### 6.2.1 上传页面

输入界面主要完成文本的上传工作，通过简单的提示，提醒用户进行文本上传。上传文本支持TXT格式，通过选择页面选择自己打算进行语篇关系分析的文本，点击提交完成上传。文本文件传输到服务端后再由服务端进行分析。

由于计划功能比较单一，所以最终设计出的页面也比较简单，具体页面如下图6-1所示。



图6-1 上传页面

该页面可通过选择文件按钮进行文件浏览，最终通过提交按钮完成文件的提交，实现了上传页面所需的基础功能。

### 6.2.2 结果页面

结果页面的主要工作是向用户展示提取出的语篇关系，让用户能直观地看到结果。

最终所设计的页面主要分为三个部分：

(1)分析结果的下载链接。通过该链接可以将文档的提取结果进行保存，文档会下载至浏览器默认的下载路径。下载的文档为TXT格式，打开即可直观地看到文档的提取结果。

(2)原文内容。页面中不展示原文内容，用来与下面的提取结果进行比对，核对提取出的结果是否正确，根据原文内容阅读提取结果可以方便用户对提取的结果进行判断。

(3)提取结果。该部分的作用就是让用户能够直接在页面对提取结果进行阅读与判断，不需要再去各类文档中寻找，也方便用户对结果进行比对，考虑需不需要进行保存。

页面显示效果如下图6-2所示。



图6-2 结果页面

在该页面用户可以通过提示文字点击即可进行文档的下载，原文内容显示在中部，在原文下方展示通过关系提取获得的语篇关系，实现了结果页面所需的基础功能。

## 6.3 本章小结

在本章节我展示了我所设计的用来展示算法的两个页面，用户可以通过该页面根据提示自主进行相关操作，完成英语文本的语篇关系提取以及提取结果的下载。虽然页面较为简陋，但是两个页面基本实现了所需的功能，整体显示效果较为清爽整洁，在算法设计的过程中为我一次次进行测试提供了极大地便利。

# 7 实验分析

## 7.1 实验环境

具体操作环境如下表7-1所示：

表7.1 所有实验环境

|  |  |
| --- | --- |
| **操作系统** | Windows 10专业版 64位操作系统 |
| **中央处理器** | Intel(R) Core(TM) i5-8400 2.80GHz |
| **图形处理器** | NVIDIA GTX1070 |
| **内存** | 16GB |
| **Python版本** | Python 3.6 |
| **Django版本** | Django 2.1.5 |
| **Keras版本** | Kears 2.3.1 |
| **Numpy版本** | Numpy 1.18.4 |
| **Spacy版本** | Spacy 2.1.4 |
| **TensorFlow版本** | TensorFlow 1.13.1 |

## 7.2 实验数据

### 7.2.1 训练数据获取

训练数据由RST-DT获得，RST-DT是保存有大量文档的语料库，里面的文本均使用基本话语单元进行了分段并且每一篇文章都有其对应的使用修辞结构理论进行分割的语篇关系。该语料库于2002年发行，收录了385篇从《华尔街日报》转录的文章。这些文章的来源是宾夕法尼亚树银行英语语料库的子集，该语料库由马库斯等人在1993年构建。在使用了语料库后才能进行模型的训练以及对最终模型做出评估。

虽然最初的RST集由24个话语关系组成，随着这类研究的发展，话语关系已经发展到通过78中修辞关系来完成关系注释，这其中包括53个单核关系以及25个多和关系。这些多样的语篇关系可以为研究者提供不小的帮助，基本上能适用于觉大多数研究者所需要的应用类型和表达力。

### 7.2.2 实验数据与结果

如果想要评价模型的好坏需要引入一个“混淆矩阵”的概念

True Positive(真正，TP)：将正类预测为正类数

True Negative(真负，TN)：将负类预测为负类数

False Positive(假正，FP)：将负类预测为正类数误报 (Type I error)

False Negative(假负，FN)：将正类预测为负类数→漏报 (Type II error)

混淆矩阵如下表7-2所示。

表7.2 混淆矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **实际类别** | **预测类别** | | | |
|  | **Yes** | **No** | **总计** |
| **Yes** | TP | FN | P（实际为Yes） |
| **No** | FP | TN | N（实际为No） |
| **总计** | P’（被分为Yes） | N’（被分为No） | P+N |

衡量一个分类模型的好与坏最重要的影响因素就是其准确率高低，准确来讲准确率就是用分类的最终结果表示为正确的样本数除以所有进行实验的样本数。一般的情况下，正确率越高，代表分类模型的分类效果越好。

准确率的计算公式7.1：（公式7.1）

由于在本次设计中使用sigmoid函数作为神经元的激活函数，所以最终采取二元交叉熵代价函数作为损失函数，使用该函数可以避免训练过程太慢，加快训练的周期[14]。

在本章节的图片中黄色曲线代表准确率（ACC），蓝色曲线代表损失（LOSS）。

#### (1)connection classifier实验数据

在设计的过程中，首先设计的是包含两个全连接层的简单模型，在实际训练时的表现如图7-1所示，测试时的表现如图7-2所示。

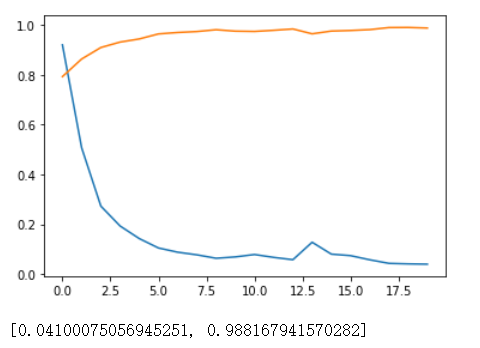
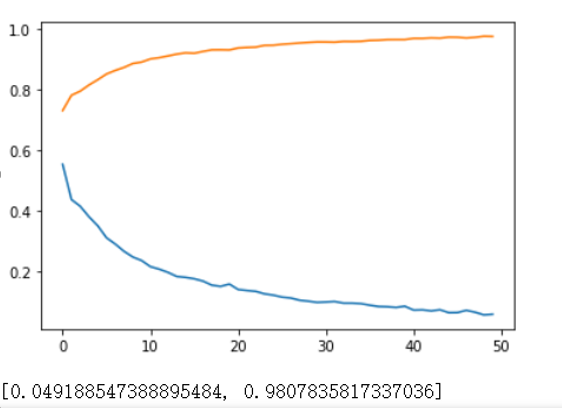


图7-1 connection classifier训练 图7-2 connection classifier测试

通过上面两幅图可以看出在整个模型训练的过程中该模型表现较为稳定，准确率是随着训练量的增加不断增加并趋于稳定，最终的准确率在百分之九十五以上，达到了设想中的准确率。但是在尝试过使用四重全连接层后，我发现使用该方法可以使模型的准确率提高更快。如图7-3以及图7-4，分别为两层全连接层以及四层全连接层时的训练结果。

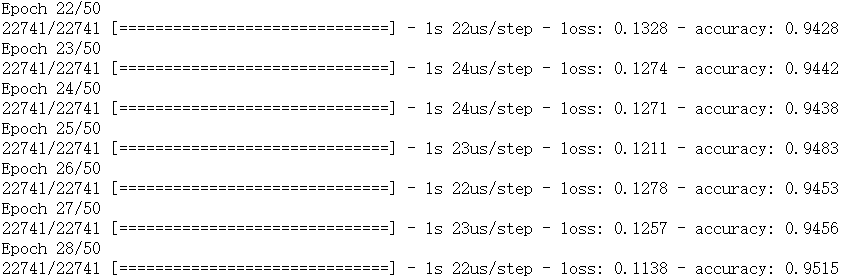


图7-3

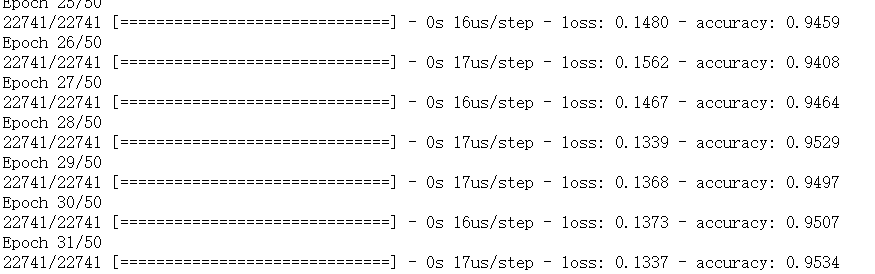


图7-4

由上述两幅图不难发现使用四层全连接层的机器学习模型准确率提高得更快，损失更小，所以最终决定采用由四层全连接层构成的connection classifier模型。

#### (2) relation classifier实验数据

在这个模型上我直接使用了由四层全连接层构建的机器学习模型来进行训练，但是在实际训练时遇到了训练数据的结果表现良好，但是测试数据表现与训练数据在损失与准确率上相差极大的情况，经过相关资料的查询，知道了这种现象是由于过拟合引起，于是加入了Dropout用来解决过拟合的问题。尚未引入Dropout的数据如图7-5以及图7-6所示。

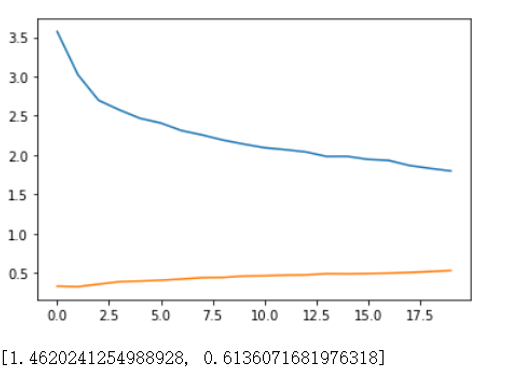
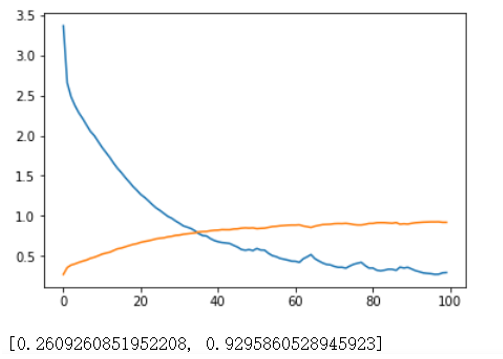


图7-5 relation classifier训练 图7-6 relation classifier测试

引入Dropout后的训练结果如图7-7以及图7-8所示。

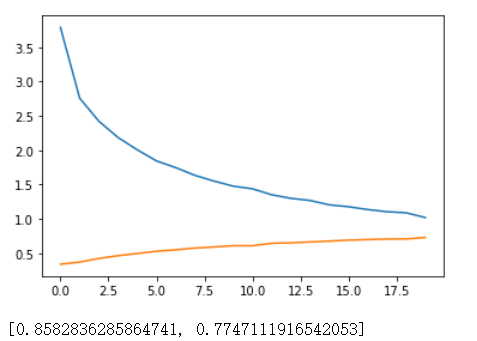
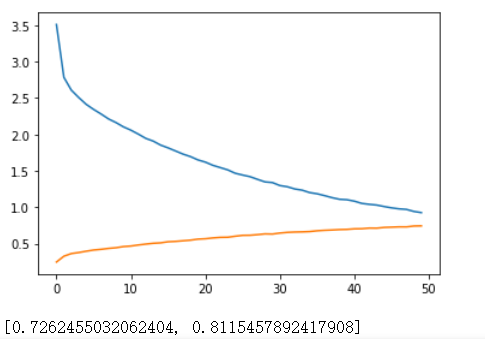


图7-7 relation classifier训练 图7-8 relation classifier测试

（引入Dropout） （引入Dropout）

由上面四幅图不难看出在没有引入Dropout之前，训练模型的损失以及准确率均已经到了一个比较低的程度，但是在测试中表现出较高的损失以及较低的准确率。这个现象在引入Dropout得以解决。后两幅图虽然准确率还没有达到较高的水平，维持在百分之七十到百分之八十之间，但是通过增加数据量以及迭代次数即可将其的准确率逐渐提高部分。

#### (3)nuclearity classifier实验数据

在尝试过上面两个模型后，该模型直接由四层全连接层以及三个Dropout构成，最终的实验数据如图7-9以及图7-10所示。

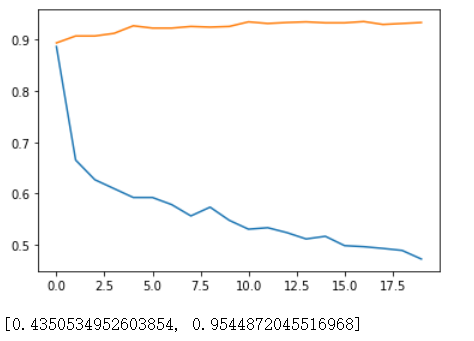
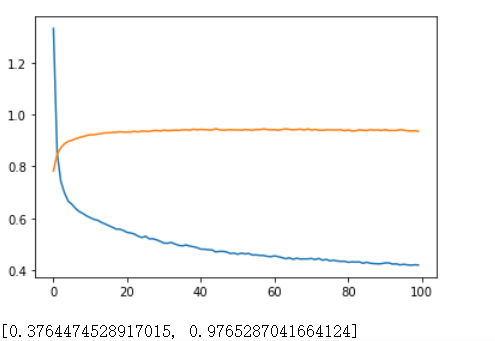


图7-9 nuclearity classifier训练 图nuclearity classifier测试

由上面两幅图可以看出通过一定量的训练样本的训练，损失在不断下降，准确率在不断提高，最终的准确率大致在百分之九十五以上，基本满足项目的设计要求，在测试集的测试中也与训练集所获得的结果没有太大出入，没有出现过拟合的现象。

## 7.3 本章小结

在本章节中我对实验的硬件以及软件环境进行了介绍，简单描述了熟练数据的来源。着重对三个机器学习模型的训练数据进行了展示与介绍，对于在训练的过程中遇到的主要问题如训练的结果不如人意以及过拟合等问题，也通过修改修炼模型进行了完善，最终的提取结果基本都在百分之九十五以上，但是relation classifier模型直至最终也没有能够达到一个比较高的准确率，需要以后对机器学习有根深入的学习来设计出更有效的算法模型，提高最终的准确率。

# 8 结束语

## 8.1 总结

在本次毕业设计中，我收获了许多，接触到了许多自己之前想去了解、学习却没有付出实际行动的方向，最大的收获就是通过本次毕业设计对人工智能这一个领域有了新的认知。起初的我对于人工智能领域一直是一知半解，然而在我通过一段时间的学习之后，懂得了其中的一些基本原理与运作方法，让我对这门学科燃起了极大地兴趣。在整个设计的过程中，我也充分运用到了自己在本科阶段所学习的基础知识，在设计，编码，测试等环节都力求规范化，体现了本科所学知识在实际中的用处，为自己的大学生生涯划上了一个完美的句号。

本论文主要的研究方向是自然语言处理中的语篇关系提取这一个课题，该研究方向对自然语言处理也有着十分大的影响，倘若能够实现较为完美的语篇关系提取，那么在句法分析与语义分析上就会长足的进展，例如机器翻译、信息提取、文本情感分析以及信息提取等自然语言理解方面的研究也会去的比较大的突破。

在本次毕业设计中，我设计完成了一个能够提取英语文本中的语篇关系的算法，该算法提取的前提条件是英语文本经过了基于修辞结构理论的基本话语单元划分。通过建立模型，训练模型，使最终设计出的软件能够将一篇英语文本的语篇关系用语篇关系树的形式展现出来，最终完成的作品基本完成了前期计划的目标。

## 8.2 不足与展望

由于自己在设计的过程中，许多知识都是随学随用，给项目增加了许多麻烦，而且由于掌握不牢固的缘故，对于很多知识还是一知半解，不能够做到熟练灵活的运用。虽然最终的算法基本实现了前期计划的目标，但是该项目还是有部分不足之处有待改进，主要存在的不足有三个方面：

(1)没有能够实现对纯英语文本的分析，由于个人能力有限，没有能够完成对于基本话语单元界限区分的算法模型。最终实现的算法模型只能够对已经进行基本话语单元划分的文本进行语篇关系提取。

(2)relation classifier算法模型的准确率偏低，虽然通过增加全连接层以及调整各参数进行试验，但是最终训练出的模型准确率依然在百分之六十五到百分之七十五，不能够进行准确有效的提取。

(3)关系树的建立相对混乱，由于第二条所提到的relation classifier算法模型的准确率偏低，导致在建立两个基本话语单元关系时可能会出现错误，而一个语篇关系树是由下而上逐步生成的，开始时的部分错误可能导致最终提取出的结果不是很让人满意。

上述三个方面均有改进的空间，可以通过之后的学习，促使自己掌握更多扎实的理论基础与实践技术，完成对这个系统的改善。

本次毕业设计虽然总体来说完成了基础的任务，但是最终的成品距离完美还有着相当大的差距，还存在许多可以改进的地方。提取语篇关系所设计的算法主要采用的是多重的全连接层配合Dropout解决过拟合的问题来进行设计，在设计的过程中克服了许多困难，希望本次设计的系统可以提供给后来学习这一方面的同学一定的帮助。

# 谢 辞

在本次毕业设计论文即将完成之际，我想向曾经给我帮助和支持的人们表示衷心的感谢。首先要感谢我的毕业设计导师，他在这次毕业设计的过程中给了我极大的支持与鼓励，提供给了我大量的资料和建议。老师从一开始论文方向的选定，到最后整篇论文的完成，都给予了我非常耐心的指导，细心地给我指出错误。老师丰富的研究经验在我进行毕业设计的整个设计的过程中帮助我解决了许多难以处理的难题。

此外我也十分感谢教育我的母校以及各科任课老师，学校里浓厚的学术氛围，舒适的学习环境以及老师们的谆谆教诲帮助我掌握了本科阶段所需的扎实的理论基础与一定的实践能力，为我本次进行毕业设计提供了坚实的基础。

最后感谢宿舍的七位舍友，他们与我一起度过了美好的大学时光，建立了深厚非凡的友谊，衷心祝愿他们前程似锦、事业有成。

# 参考文献

1. 黄恒,孙常龙(常龙),李琳琳(琳嫣),张琼(复歌),王睿(马司),骆卫华,吴晨(绥猷),刘晓钟,司罗.阿里巴巴的NLP应用与发展[J].人工智能,2018(01):120-131.
2. 马鸣皓.机器翻译质量评估刍议[J].宁波教育学院学报,2019,21(06):76-78+93.
3. Jose M. Chenlo, Alexander Hogenboom, David E. Losada. Rhetorical Structure Theory for polarity estimation: An experimental study[J]. Data & Knowledge Engineering, 2015, 94(PB):135-147.
4. Marir F, Haouam K. Rhetorical structure theory for content-based indexing and retrieval of Web documents[C]. international conference on information technology research and education, 2016: 160-164.
5. DeUzeda VR,Pardo TA,Nunes MD,et al.Evaluation of Automatic Text Summarization Methods Based on Rhetorical Structure Theory[C]. intelligent systems design and applications, 2018: 389-394.
6. 欧华杰.大数据背景下机器学习算法的综述[J].中国信息化,2019(04):50-51.
7. Zhang J J, Fung P. Automatic Parliamentary Meeting Minute Generation Using Rhetorical Structure Modeling[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2017, 20(9): 2492-2504.
8. 孙爱珍.语境共现词汇链的自动提取及与语篇衔接之关系(英文)[J].Chinese Journal of Applied Linguistics,2016,34(04):3-14+127.
9. Tu M, Zhou Y, Zong C, et al. A Novel Translation Framework Based on Rhetorical Structure Theory[C]. meeting of the association for computational linguistics, 2017: 370-374.
10. 练睿婷. 语篇标注语料库的建设研究. 中国中文信息学会.第四届全国学生计算语言学研讨会会议论文集[C]. 北京:中国中文信息学会. 2018:284-290.
11. Zhang S. Entity relation extraction to free text[C]. international conference natural language processing, 2017: 1-5.
12. Li K, Zhang J, Yao C, et al. Automatic Relation Extraction from Text: A Survey[C]. the internet of things, 2016: 83-86.
13. Chengcheng L. Automatic Text Summarization based on Rhetorical Structure Theory[C]. international conference on computer application and system modeling, 2018.
14. 张晓敏. 深度学习在文本分类中的研究与应用[D]. 成都: 电子科技大学,2019.