**北 京 科 技 大 学**

**硕士学位研究生**

**选题报告及文献总结**

论文题目：基于机器学习的古代汉语

自动分词标注算法及语料库研究



指导教师： 皇 甫 伟

单 位：计算机与通信工程学院通信工程系

学 号： S20170665

作 者： 于 学 金

专业名称： 信息与通信工程

入学时间： 2017年09月

2018年7 月

**目 录**

[1 课题意义与选题依据 3](#_Toc5790)

[1. 1课题的意义 3](#_Toc4079)

[1. 2研究课题国内外现状 4](#_Toc7024)

[2研究目标与内容 8](#_Toc7814)

[2.1研究目标 8](#_Toc16552)

[2.2研究内容 8](#_Toc17285)

[2.3 关键问题与创新点 8](#_Toc31955)

[2.3.1 关键问题 8](#_Toc21033)

[2.3.2 创新点 9](#_Toc14740)

[3研究方案及技术路线 9](#_Toc19324)

[3.1研究方案 9](#_Toc505)

[3.2技术路线 10](#_Toc3031)

[4研究工作基础 11](#_Toc20590)

[4.1知识储备与研究条件 11](#_Toc23941)

[4.2可能遇到的问题及解决办法 13](#_Toc24927)

[4.2.1 可能遇到的问题 13](#_Toc24933)

[4.2.2 解决办法 13](#_Toc22467)

[5研究工作计划及进度安排 14](#_Toc1475)

[5.1研究工作计划及进度安排 14](#_Toc11049)

[5.2预期成果 14](#_Toc17591)

[参考文献 15](#_Toc21243)

# 1 课题意义与选题依据

## 1. 1课题的意义

随着科技不断发展，人工智能与生活的结合已逐渐成为潮流，人工智能领域也再度成为各大行业关注的焦点。“AI+”逐渐和“互联网+”一起推动整个社会科技的发展。人工智能企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器，该领域的研究包括机器人、语言识别、图像识别和自然语言处理等。人工智能从诞生以来，理论和技术日益成熟，应用领域也不断扩大，其中自然语言处理（NLP）是计算机科学，人工智能，语言学和人类（自然）语言之间的相互作用的领域，它融语言学、计算机科学、数学于一体。这一领域的研究将涉及自然语言即对人们经常使用的语言进行各种分析处理，所以它对于各种语言学的研究有着重要的意义。

中文自然语言处理是对汉语的各级语言单位(字、词、语句、文章等)进行自动加工处理的技术[1]，是自然语言处理技术在中文方向应用的体现，使计算机可以分析和处理中文自然文本。而计算机在进行中文自然语言处理时一般是以词为最小单位的，更深层次的语言语义分析，比如POS tagging, chunking, parsing都是以中文分词技术为基础的。我们知道，在英文文本中，单词之间是以空格作为自然分界符的。中文和英文比起来，有其自身的特点，就是中文以字为基本书写单位，句子和段落通过分界符来划界，但是词语之间没有一个形式上分界符。也就是说，从形式上看，中文没有“词"这个单位[3]。所以中文分词是汉语自然文本处理的基础问题之一。因此，中文分词技术是做中文自然语言处理必不可少的一项关键技术。

中文分词相关研究开始于20世纪80年代初，自2003年国际中文分词评测活动Bakeoff正式开展至今，中文自动分词技术有了长足的进步。中文分词技术发展到目前为止已经提出很多各具特色的方法，包括基于字典匹配、基于规则的方法、基于统计的方法和基于神经网络的方法等。经过多年的探索发展，中文分词技术已经进入了实用化阶段，广泛应用于机器翻译、信息检索、语义识别等领域[4]。

随着计算机技术和自然语言处理技术的发展，对于古汉语文章书籍的组织、采录、收集、整理、纂修、审定也大多转移到了计算机上。古代汉语，是与现代汉语相对而言的，古代汉族群众的语言。广义的古代汉语的书面语有两个系统：一个是先秦口语为基础形成的上古汉语书面语及其后人用这种书面语写成的作品，也就是我们所说的文言；另一个是六朝以后在北方方言的基础上形成的古代白话[2]，狭义的古代汉语书面语就是指文言文。由于古汉语的专业性，古汉语自然文本的采录、处理和分析过程大多由专业的古汉语研究者来操作，整个过程十分消耗人力物力。自然语言处理技术与古汉语处理的结合使人们可以用处理一般文本的方式处理晦涩的古汉语，而无需再花大量的时间和精力去学习和检索不符合现代人习惯的古汉语语法。

古汉语分词研究可以服务于古汉语学术研究，是古汉语自然语言处理领域后续的机器翻译、情感分析和语义识别等工作的基础；与此同时，词语的标注是NLP任务预处理中的重要步骤，再进行句法分析就容易多了，对于字、词用法灵活的古代汉语来说词语标注也至关重要。古汉语分词、标注研究对于古文字学、出土文献以及古史等古汉语人文研究具有重要意义。不仅如此，针对古汉语的研究对于现代汉语处理也具有一定的帮助作用，因为在现代汉语中仍然存在不少古汉语语句词汇的存留，现代汉语文本中也会存在古代汉语的诗句、文章的引用，针对于现代汉语语言特点设计的自然语言处理系统在面对古汉语的诗句时处理有效性必将受到影响，因此古汉语分词及标注系统的研究也将是现代汉语研究的重要补充。

然而目前对于中文分词及标注系统的大量研究成果主要是针对现代汉语，在古汉语分词及标注领域的研究成果相对较少。因为古代汉语在文字、词法和句法等诸多方面与现代汉语有很大的不同，例如古汉语在文字上是使用繁体字，而现代汉语大多用简体字；词法上古汉语词类活用更为丰富，词类分工并不明确，现代汉语词汇意思大多固定，词类活用的例子并不多见；句法上，古汉语在判断句中大多以名词或名词短语作谓语，现代汉语的判断句中用‘是’做谓语；从词汇构成方面来看，现代汉语以Bakeoff-2003和Bakeoff-2005训练语料库词为例，表一说明现代汉语中单音词和双音词占语料库的绝大部分，其中单音词占54%，双音词占39.3%[4]。古汉语这边以上古、中古汉语训练集为例，其中单音节词占所有词的比例仅为25%，但是其使用频率为80%，远远高于双音节及其他多音节词。

表**1** Bakeoff-2003和Bakeoff-2005训练语料库词长频率分布

|  |  |  |  |
| --- | --- | --- | --- |
| 词长 | AS2003 | AS2005 | CityU2003 |
| 1 | 0.5447 | 0.5712 | 0.4940 |
| 2 | 0.3938 | 0.3787 | 0.4271 |
| ≥3 | 0.0615 | 0.0501 | 0.0789 |

基于古汉语与现代汉语的以上不同，尤其是词频、 词类活用的问题，对古汉语分词标注系统增加了很大的难度，许多对现代汉语的大量研究成果不能直接应用于古代汉语处理领域中去。在古代汉语有关的领域中，如古汉语学术研究、古汉语文章检索与校对、自动翻译等，均以古汉语分词及标注为基础，若古汉语分词及标注正确率不能达到实际应用水平，则上述领域均寸步难行。所以本文决定针对古代汉语分词及标注系统做专题性的研究。

## 1. 2研究课题国内外现状

从1983开始，国内外的研究者在中文分词及标注领域进行了广泛的研究，提出了一些有效的算法。

### 1.2.1分词算法

分词算法方面，早先的算法大致可分为以下几类：基于字符串匹配的分词方法、基于统计的分词方法、基于规则的分词方法。从2008年往后，随着机器学习的重新兴起，机器学习的各种模型也被逐渐引入到中文分词应用当中。应该看到，不管是基于词表的切分方法，还是基于统计、基于规则或基于机器学习的切分方法，每一种方法都有自己的优点和一定的局限性。

1. 基于字符串匹配的方法

基于字符串匹配[5-7]的方法又叫做机械分词法，这种方法要事先准备一个“充分大的"词典，然后将待切分的句子按照一定的扫描规则与词典中的词条进行匹配，如果匹配成功，则将这个词切分出来，否则进行其他相关处理。按照扫描方向的不同分为正向匹配和逆向匹配；按照不同长度优先分配的情况，分为最大匹配和最小匹配；按照与词性标注过程是否相结合，又可以分为单纯分词方法和分词与标注相结合的一体化方法。常用的基于字符串匹配的方法有正向最大匹配分词法、逆向最大匹配分词法、最少切分分词法、双向匹配法四种。在基于字符串匹配法中，算法的运行速度快，程序复杂度低，成本较低。但算法的性能很大程度上取决于字典，然而因为总会有新词出现，所以字典不能包含所有的词语，对于存在大量未登录词的情况，字典匹配法效果并不好。



图1.2.1 字典匹配法示意图

1. 基于统计的方法

基于统计[8-11]的分词方法是从概率的角度出发，单个字出现在词组中的联合概率是比较大，因此当相邻的字越经常出现，则越有可能是一个词组。因此字与字相邻共现的频率或概率能够较好的反映成词的可信度。因此可对语料中相邻共现的各个字的组合的频度进行统计，计算它们的相关度。这种方法首先切分与词典能匹配成功的所有可能的词，即找出所有候选词条，然后运用统计语言模型和决策算法得出最优的切分结果。由于纯粹从统计的角度出发，因此在统计意义上某些经常出现在一起的字并不能构成完整的词语，例如“上的"、“下的”、“这一”等在文本中会大量的互邻同现，但他们却分属于不同的词；并且统计语言模型和决策算法在很大程度上决定了解决歧义的方法，需要大量的标注语料，并且分词速度也因搜索空问的增大而有所减慢。基于统计的分词方法所应用的主要的统计量或统计模型有：互信息、N元文法统计模型、隐马尔科夫模型和最大熵模型等。这些统计模型主要是利用词与词之间的联合出现概率作为分词判断的信息。基于统计的方法利用可能性或者评分机制来判断是否将词进行分割，而不是仅依赖于字典匹配。这种基于统计进行分割的做法主要有三个缺点：一、这些方法仅识别未登录词而不判断这些词的种类；二、这些基于统计的方法多不结合语法信息和语言知识，因此在统计意义上某些经常出现在一起的字并不能构成完整的词语，导致分词出错，这就需要在识别之后再耗费人力去验证。三、未登录词识别在很多系统中是和分词系统分开的，例如文献[15]假设未登录词大多被识别为单字，所以他采用在基本分词后，添加级联层来检测未登录词。

1. 基于规则的分词方法

该方法主要基于句法、语法分析，并结合语义分析[12-14]。通过对上下文内容所提供信息的分析对词进行定界，它通常包括三个部分：分词子系统、句法语义子系统、总控部分。在总控部分的协调下，分词子系统可以获得有关词、句子等的句法和语义信息，用来对分词歧义进行判断。例如Wu(2003a)尝试了将分词与句法分析技术融为一体的方法，用整个句子的句法结构来消除不正确的切分。这种方法对消解组合型歧义比较有效，但组合型歧义在切分歧义中毕竟占少数，而在频繁出现的交集型歧义的消解方面，使用句子分析器并没有明显优势。而且这类方法试图让机器具有人类的理解能力，需要使用大量的语言知识和信息。由于汉语的复杂性，难以将各种语言信息和规则组织成机器可直接读取的形式，需要消耗大量的人力整理规则。

1. 组合方法

文献[16][17]提出了更加完善的基于字典和统计的系统。他们将未登录词的识别和分词系统结合到一个统一的系统中。其中[16]系统是利用基于加权有限状态传感器，文献[17]是基于线性混合模型，线性模型是源于广泛用于模式分类的线性判别函数[18]并由Collins等人[19]引入到自然语言处理中，线性混合模型可以更灵活的利用字库中的统计信息，这也使[17]在未登录词识别性能上优于[16]。除此之外，最大熵模型、条件随机场模型、最大熵模型[20-24]也都将分词和未登录词识别结合在一起进行。然而这些所有基于统计和规则的方法，在生成模型的时候还是依赖于手工提取的特征，耗费大量的时间和人力资源，且算法的正确率极大的取决于规则的准确性与完整性。

1. 基于机器学习的方法

2010年之后，随着神经网络的重新兴起，自然语言处理也开始利用神经网络模型进一步发展，大量基于神经网络[35-39]的中文分词系统开始被提出，示意图如如图1.2.2所示。滑动窗口输入字符嵌入[25]、标签嵌入的方法[26]；利用门控制递归神经网络并对n-gram特征进行建模(GRNN)的方法[27]；长短期记忆神经网络方法(LSTM)[28-29]；结合GRNN和LSTM提取深层特征信息的方法[30]；基于过渡(transition-based)模型的方法[31]。系统的不断完善，分词性能也不断提高，但是其结构基本不变。



图1.2.2 基于神经网络的分词系统示意图

近几年来新的机器学习方法和大规模计算技术在汉语分词中的应用，分词系统的性能一直在不断提升，基于MSRA、AS、PKU等主流测试集的分词准确率均已突破95%[40-43]。在一些通用的书面文本上，如新闻语料，领域内测试（训练语料和测试语料来自同一个领域）的性能已经达到相当高的水平。但是，跨领域测试的性能仍然很不理想，例如用计算机领域或者医学领域的测试集测试用新闻领域的数据训练出来的模型。由于目前具有较好性能的分词系统都是基于有监督的学习方法训练出来的，需要大量有标注数据的支撑，而标注各个不同领域的语料需要耗费大量的人力和时间，因此，如何提升汉语自动分词系统的跨领域性能仍然是目前面临的一个难题。

### 1.2.2标注算法

汉语词性标注同样面临许多棘手的问题， 其主要难点可以归纳为三个方面：1、汉语是一种缺乏词形态变化的语言，词的类别不能像印欧语那样， 直接从词的形态变化上来判别；2、常用词兼类现象严重；3、研究者主观原因造成的困难。汉语词性标注与分词一样，是中文信息处理面临的重要的基础性问题，而且两者有着密切的关系。

1. 基于统计模型的词性标注方法

I.Marshall建立的LOB语料库词性标注系统CLAWS是基于HMM模型的词性标注方法的典型代表[44]，该系统通过对n元语法概率的统计优化，实现133个词类标记的合理标注。实现基于HMM的词性标注方法时，模型的参数估计是其中的关键问题。算法随机地初始化HMM的所有参数，这将使词性标注问题过于缺乏限制。还有另外一个问题需要注意，就是模型参数对训练语料的适应性。由于不同领域语料的概率有所差异，HMM的参数也应随着语料的变化而变化。因此当对原有的训练语料增加新的语料以后，模型的参数需要重新调整；而且在经典HMM理论框架下，利用标注过的语料对模型初始化以后，已标注的语料就难以再发挥作用。

1. 基于规则的词性标注方法

基于规则的词性标注方法是人们提出较早的一种词性标注方法， 其基本思想是按兼类词搭配关系和上下文语境建造词类消歧规则。 早期的词类标注规则一般由人工构造， 如美国布朗大学开发的TAGGIT词类标注系统。 刘开瑛(2000)曾按兼类词搭配关系构造了词类识别规则库，针对动名词兼类现象，归纳出了9条词性鉴别规则，包括：并列鉴别、同境鉴别、区别词鉴别和唯名形容词鉴别规则等，并结合词类同现概率实现了汉语词性标注系统。随着标注语料库规模的逐步增大，可利用资源越来越多，以人工提取规则的方式越来越难以实现，规则的提取越来越难。

1. 统计方法与规则方法相结合的词性标注方法

理性主义方法与经验主义方法相结合的处理策略一直是自然语言处理领域的专家们不断研究和探索的问题，对于词性标注问题也不例外。周强(1995)给出了一种规则方法与统计方法相结合的词性标注算法，其基本思想是，对汉语句子的初始标注结果（每个词带有所有可能的词类标记）， 首先经过规则排歧，排除那些最常见的、语言现象比较明显的歧义现象，然后通过统计排歧，处理那些剩余的多类词并进行未登录词的词性推断，最后再进行人工校对，得到正确的标注结果。这样做有两个好处：一方面利用标注语料对统计模型进行参数训练，可以得到统计排歧所需要的不同参数；另一方面，通过将机器自动标注的结果（规则排歧的或统计排歧的）与人工校对结果进行比较，可以发现自动处理的错误所在，从中总结出大量有用的信息以补充和调整规则库的内容。但是，该方法中容易产生规则与统计的作用域不明确的问题。

虽然中文自然语言处理发展迅速，但如前面所说的，古代汉语在文字、词汇和语法等诸多方面与现代汉语有所不同，现代汉语的先进研究成果不能直接应用于古代汉语处理领域中去。在数据集、分词标准等问题并没有形成共识，没有一个类似于现代汉语中Bakeoff数据库一样的作为统一测试集的数据库。且现代计算机方面研究者大多并不熟悉古汉语，缺乏相应的古汉语常识和知识，对于古汉语的语言习惯、语言规则并不如现代汉语这样熟悉，也造成研究者无法深入到古汉语处理研究当中去。

# 2研究目标与内容

## 2.1研究目标

本论文拟完成下列目标：

1. 完成一种面向古汉语的正确率可满足社会科学人员工作需求的分词及标注系统。
2. 制作一个古汉语的语料库网站，语料库在中央研究院上古语料库基础上进行完善和校订，包含上古、中古及近代词性信息和初步的分词结果。

## 2.2研究内容

根据本文的研究目标，拟研究内容如下：

1. 结合古汉语单字利用率高，一字多义的特点，研究一种一字多嵌的方法，将一个字根据词性的不同映射为多个向量，以方便计算机处理。
2. 研究未登录词识别的方法，可分别对古汉语中的形态衍生词、事实描述词和命名实体的识别。
3. 结合(1)(2)，研究符合古汉语预料特点的实用分词标注系统结构。
4. 研究数据收集，数据整理和数据清洗的方法。
5. 研究数据库的字段设置，数据库模式和交互方式等数据库建设方面问题。

## 2.3 关键问题与创新点

### 关键问题

本论文的关键问题包括：

1. 古汉语规则的提取。古汉语语言结构与语言习惯不同于现代汉语，现阶段算法大多是面向现代汉语的，自然语言处理研究者大多提取的是现代汉语语法规则，然而这些规则难以应用于古汉语中。
2. 模型的组合搭建。古汉语分词标注研究中，统计模型和机器学习模型各有利弊，模型如何组合搭建才能达到最好的性能需要深入调研研究。
3. 歧义消除和未登录词的识别一直是分词系统的关键问题，在分词过程中，分词与歧义消除和未登录词识别过程若分为两个阶段，则粗分词阶段的准确率会直接影响歧义消除和未登录识别阶段。所以这两阶段如何处理是本论文的关键问题之一。
4. 数据库以及分词标准的确定。古汉语的数据库和各类标准并不完善，因此在系统的研究过程中的数据如何收集，以何种形式存储，以何种格式呈现也是需要面对的关键问题。

### 2.3.2 创新点

根据现在的研究情况，拟创新点如下：

(1) 对于古汉语语法规则提取困难的问题，利用机器自动学习语法规则，利用神经网络模型完成分词和标注，减少人工规则提取。

(2) 在分词与歧义消除、未登录词识别两阶段的关系上，采用基于字的分词方法，即给每个字添加一个标签来标定字在词中的位置，这样就可以将分词问题表示为序列的标记问题，以字为基础则可以平等的看待登录词和未登录词，即可同时处理分词与歧义消除、未登录词识别两个阶段。

(3)采用分词与标注相结合，利用多种统计模型和机器学习模型来分别判断各种词的标注并进行评分，最后利用Viterbi算法选择最佳路径，在分词的同时完成识别和标注。

(4) 完成一个有分词标准的古汉语语料库。

# 3研究方案及技术路线

## 3.1研究方案

通过深入学习和调研，拟定研究方案如示意图3-1所示。



图3-1 研究方案示意图

如上图所示，

(1) 输入长篇带有标点的文本，首先使用标点’。’’？’和’！’来检测句子边界，将长篇文本划分为多句话。

(2) 输入一句话首先分割出所有可能的词语并暂存这些词语，分别对这些词利用不同的词性模型进行评分，例如“百姓聞王鐘鼓之聲，管籥之音，舉欣欣然有喜色。”中“欣欣然”这个词需要利用命名实体识别的模型、形态衍生词模型和事实描述词模型进行评分。我们首先将所有词分为两大类：登录词和未登录词。其中的未登录词又分为形态衍生词、事实描述词、命名实体和新词。我们将对所有类分别建模，计划对登录词利用搜索树搜索，形态衍生词等未登录词先对全文进行预识别，识别出可能的新词组合，然后利用机器学习的算法进行建模评分。在利用机器学习等数学方法进行建模之前，不可避免的需要进行汉字的向量化处理，而对于古汉语的一字多义的特点，可能需要对统一汉字进行多种向量化的处理，在后期分别对多种向量化的结果进行评分。

1. 利用Viterbi算法计算最大可能性的一条路径，即确定各个词的分词结果和词性。
2. 在确定了词的类别后，根据词语类别进行定制化输出。例如“欣欣然”输出“欣然+[形态衍生次(重复类)]”带有标注信息的分词结果，也可直接输出“欣欣然”。

## 3.2技术路线

拟将按图3-2所示技术路线对本课题进行研究。



图3-2 技术路线图

研究过程中用到的方法：数据库、数据结构、线性混合模型、感知机算法、支持向量机和神经网络方法等。

研究过程中用到的实现手段： Python、Matlab、实验室服务器。

# 4研究工作基础

## 4.1知识储备与研究条件

**1.知识储备**

本课题是做了充分调研和学习后，并有一定的工作基础上提出的。主要包括阅读了关于自然语言处理相关文献，研究了国内外在该领域的研究现状，总结了在这方面现存的可发展空间和不足。同时也仿真了许多相关实验、储备了数学基础知识。掌握了数据结构、数据库、线性混合模型、梯度下降算法、支持向量机、神经网络等基本算法和概念的应用和实现。在此基础上，还有以下两点重要的储备：

1. 中央研究院古汉语语料库

中央研究院古汉语语料库是应汉语史研究需求而建构的语料库。这个语料库又可依是否经过断词及加标词类而分成两类，即未加标的素语料库以及有标注的标记语料库。目前素语料库所收集的语料已含盖上古汉语(先秦至西汉) 中古汉语(东汉魏晋南北朝) 近代汉语(唐五代以后)大部分的重要语料，并开放使用；在标记语料库方面，上古汉语至近代汉语大部分语料均已完成标注的工作。

1. 部分古汉语分词代码已经实现

经过前期部分准备工作，已利用Python编写完成初步的古汉语分词系统，包括基于词典的古汉语分词系统，基于神经网络的古汉语分词系统，并且分词准确率可以达到85%，后期工作可在此基础上继续优化和完善。

**2.研究条件**

本课题的研究和论文的撰写工作是在北京科技大学“网络楼”中进行的，该研究所已在数据处理和数据结构相关研究等方面积累了一定的经验和技术成果，同时具有部分中央研究院上古汉语语料数据，为我的研究工作奠定了强大的理论基础和必备条件。“网络楼”研究所作为学校重点建设的研究基地和科研团队之一，实验设施齐全，仿真系统完备，同时提供了快速的接入互联网搜索和大容量的服务器设施，为本课题的展开和实施提供了良好的仿真分析平台和实验条件。

## 4.2可能遇到的问题及解决办法

### 4.2.1 可能遇到的问题

(1) 已有的上古语料数据并不完善，由于大陆与台湾所使用的字符编码不同以及古汉语中存在部分生僻字，语料库中存在部分字符无法显示的问题。

(2) 目前将统计模型与机器学习模型结合并且同时进行分词与标记的实际应用较少，所以在具体算法模型的选择上会有一定的困难。

### 4.2.2 解决办法

(1) 不断搜索其他数据库，寻找相同的语句对已有数据库进行修复，或者对编码有问题的语句在程序运行时进行忽略，避免有编码问题的字符对整个系统所产生的噪声干扰。

(2) 查阅文献，阅读相关理论书籍，了解不同模型的优缺点，多进行不同模型之间的尝试，不断加深对各种模型之间组合的理解。

# 5研究工作计划及进度安排

## 5.1研究工作计划及进度安排

表5-1 研究工作进度表

|  |  |
| --- | --- |
| 2018年7月—2018年9月 | 文献调研和理论学习。阅读文献，研究数据，实时了解国内外中文分词现状和并寻找古汉语相关数据库，学习最新的研究方法，统计、机器学习等数学基础。 |
| 2018年10月—2018年12月 | 完成简单的分词系统的设计与仿真，并继续深入研究古汉语规则及数据的存储结构等问题。 |
| 2019年1月—2019年3月 | 评估多种分词方案的效果并进行测试和完善，设计实现统计与机器学习相结合的古汉语分词系统。 |
| 2019年4月—2019年6月 | 弥补系统的各种缺陷，不断提高分词正确率，反复进行效果的测试与完善。 |
| 2019年7月—2019年9月 | 完成课题研究的总结和论文撰写。 |

## 5.2预期成果

本论文拟完成：(1) 一种基于神经网络并利用全文统计信息的中文分词系统并且该系统可以在分词的同时进行词性标记。即将一段文字作为输入，系统输出分词结果的同时，标记词的某些属性信息。(2) 生成一个20M左右大小的语料库，该数据库可以网站的形式呈现在互联网上。语料库包含上古、中古及近代汉语文章中初步的分词结果和词性信息，可供后续实验训练和测试用。(3)在核心期刊或国际会议上发表1到2篇学术论文。

# 参考文献

1. 曹卫峰. 中文分词关键技术研究[J]. 南京理工大学. 硕士学位论文, 2009.桂智明, 陈彩. 基于语义的移动对象轨迹知识发现研究[J]. 计算机工程, 2009, 35(16):14-16.
2. 汉生, 古汉语, 云之, 等. 中国古代的语言和逻辑[M]. 社会科学文献出版社, 1998.
3. 周强. 基于语料库和面向统计学的自然语言处理技术[J]. 计算机科学, 1995 (4): 36-40.
4. 邱冰, 皇甫娟. 基于中文信息处理的古代汉语分词研究[J]. 微计算机信息, 2008, 24(24): 100-102.
5. Wu Z, Tseng G. Chinese text segmentation for text retrieval: Achievements and problems[J]. Journal of the American Society for Information Science, 1993, 44(9): 532-542.
6. Sproat R, Shih C. Corpus-based methods in Chinese morphology[J]. Tutorial at the 19th COLING, 2002.
7. Sun M, Tsou B K. A review and evaluation on automatic segmentation of Chinese[J]. Contemporary Linguistics, 2001, 3(1): 22-32.
8. Lin M Y, Chiang T H, Su K Y. A preliminary study on unknown word problem in Chinese word segmentation[C]//Proceedings of Rocling VI Computational Linguistics Conference VI. 1993: 119-141.
9. Chang J S, Su K Y. An unsupervised iterative method for Chinese new lexicon extraction[J]. International Journal of Computational Linguistics & Chinese Language Processing, Volume 2, Number 2, August 1997, 1997, 2(2): 97-148.
10. Teahan W J, Wen Y, McNab R, et al. A compression-based algorithm for Chinese word segmentation[J]. Computational Linguistics, 2000, 26(3): 375-393.
11. Dai Y, Loh T E, Khoo C S G. A new statistical formula for Chinese text segmentation incorporating contextual information[C]//Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999: 82-89.
12. Ma W Y, Chen K J. Introduction to CKIP Chinese word segmentation system for the first international Chinese Word Segmentation Bakeoff[C]//Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17. Association for Computational Linguistics, 2003: 168-171.
13. Wu A, Jiang Z. Statistically-enhanced new word identification in a rule-based Chinese system[C]//Proceedings of the second workshop on Chinese language processing: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 12. Association for Computational Linguistics, 2000: 46-51.
14. Chen K J, Bai M H. Unknown word detection for Chinese by a corpus-based learning method[C]//International Journal of Computational Linguistics & Chinese Language Processing, Volume 3, Number 1, February 1998: Special Issue on the 10th Research on Computational Linguistics International Conference. 1998, 3(1): 27-44.
15. Ma W Y, Chen K J. Introduction to CKIP Chinese word segmentation system for the first international Chinese Word Segmentation Bakeoff[C]//Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17. Association for Computational Linguistics, 2003: 168-171.
16. Gao J, Li M, Huang C N, et al. Chinese word segmentation and named entity recognition: A pragmatic approach[J]. Computational Linguistics, 2005, 31(4): 531-574.
17. Sproat R, Gale W, Shih C, et al. A stochastic finite-state word-segmentation algorithm for Chinese[J]. Computational linguistics, 1996, 22(3): 377-404.
18. Duda R O, Hart P E, Stork D G. Pattern classification[M]. John Wiley & Sons, 2012.
19. Collins M, Duffy N, Park F. Parsing with a single neuron: Convolution kernels for natural language problems[J]. 2001.
20. Xue N. Chinese word segmentation as character tagging[J]. International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing, 2003, 8(1): 29-48.
21. Och F J, Ney H. Discriminative training and maximum entropy models for statistical machine translation[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002: 295-302.
22. Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[C]//Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, 2004: 22.
23. Peng F, Feng F, McCallum A. Chinese segmentation and new word detection using conditional random fields[C]//Proceedin
24. Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.
25. Zheng X, Chen H, Xu T. Deep learning for Chinese word segmentation and POS tagging[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 647-657.
26. Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(Aug): 2493-2537.
27. Chen X, Qiu X, Zhu C, et al. Gated recursive neural network for Chinese word segmentation[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015, 1: 1744-1753.
28. Chen X, Qiu X, Zhu C, et al. Long short-term memory neural networks for chinese word segmentation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1197-1206.
29. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
30. Xu J, Sun X. Dependency-based gated recursive neural network for chinese word segmentation[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2016, 2: 567-572.
31. Zhang M, Zhang Y, Fu G. Transition-based neural word segmentation[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016, 1: 421-431.
32. Cai D, Zhao H. Neural word segmentation learning for Chinese[J]. arXiv preprint arXiv:1606.04300, 2016.
33. Cai D, Zhao H, Zhang Z, et al. Fast and accurate neural word segmentation for Chinese[J]. arXiv preprint arXiv:1704.07047, 2017.
34. Yang J, Zhang Y, Dong F. Neural word segmentation with rich pretraining[J]. arXiv preprint arXiv:1704.08960, 2017.
35. Schalkoff R J. Artificial neural networks[M]. New York: McGraw-Hill, 1997.
36. Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]//Eleventh Annual Conference of the International Speech Communication Association. 2010.
37. Mikolov T, Kombrink S, Burget L, et al. Extensions of recurrent neural network language model[C]//Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011: 5528-5531.
38. Sak H, Senior A, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling[C]//Fifteenth annual conference of the international speech communication association. 2014.
39. Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112.
40. Krogh A, Larsson B È, Von Heijne G, et al. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes[J]. Journal of molecular biology, 2001, 305(3): 567-580.
41. Bagnall A, Lines J, Bostrom A, et al. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances[J]. Data Mining and Knowledge Discovery, 2017, 31(3): 606-660.
42. Han H, Lei W, Hua Y, et al. Effective Neural Solution for Multi-Criteria Word Segmentation[J]. arXiv preprint arXiv:1712.02856, 2017.
43. Liu X. Comparisons of Features for Chinese Word Segmentation[C]//International Conference on Geo-Spatial Knowledge and Intelligence. Springer, Singapore, 2017: 492-499.
44. Marshall I N. Consciousness and Bose-Einstein condensates[J]. New ideas in Psychology, 1989, 7(1): 73-83.
45. 周强. 规则和统计相结合的汉语词类标注方法[J]. 中文信息学报, 1995, 9(3): 1-10.