基于机器学习的古代汉语切分标注算法及语料库研究

于学金

北京科技大学

密　级：公开

加密论文编号：

论文题目：**基于机器学习的**
古代汉语切分标注算法及语料库研究

学　　号：s20170665

作　　者：于学金

专 业 名 称：信息与通信工程

2019 年 12 月 20 日

基于机器学习的古代汉语切分标注算法及语料库研究

Study on segmentation tagging algorithm and corpus of ancient Chinese based on machine learning

研究生姓名：于学金

指导教师姓名：皇甫伟

北京科技大学计算机与通信工程学院

北京 100083，中国

Master Degree Candidate： Xuejin Yu

Supervisor： Wei Huangfu

School of Computer and Communication Engineering

University of Science and Technology Beijing

30 Xueyuan Road，Haidian District

Beijing 100083，P.R.CHINA

分类号：TP391　　密　　级：公开

ＵＤＣ：　　单位代码：１０００８

北京科技大学硕士学位论文

论文题目：基于机器学习的古代汉语切分标注算法及语料库研究

作　　者：于学金

指 导　教 师：皇甫伟　**单位：**北京科技大学

指导小组成员：　**单位：**

　　单位：

论文提交日期：2016 年 12 月 20 日

学位授予单位：北 京 科 技 大 学

致　谢

两年半的研究生生涯即将结束，在研究课题期间有过怀疑、有过否定也收获了独属于自己的成就感，庆幸自己能为古汉语自然语言处理的相关课题贡献一点小小的力量，也庆幸自己得到了许多人的帮助和关心。

首先感谢我的导师皇甫伟老师，从论文的开题到终稿，您给予了我无数的指导和鼓励。谢谢您曾抽出宝贵的休息时间为我反复修改论文；谢谢您在我对课题感到迷茫时的鼓励和安慰；还谢谢您在我偷懒时地不断鞭策。很庆幸也很骄傲能成为您的学生。

感谢我的三位舍友熊健、徐海祥、王梓楠，缘分让我们聚在一个宿舍，研究生生涯有你们的陪伴是我的幸运。

感谢乌尼日其其格同学和秦运慧师姐，我会一直记得你们对我的帮助。感谢师姐帮我修改文章思路，感谢乌尼日同学帮我调整论文细节等很多我没有注意到的工作。感谢你们对我的付出。

感谢我的其他实验室小伙伴们：刘娅汐、王欢、雷铠僖、安玮、沈一佳；还有已经毕业了的师兄师姐：王浩彬、黄鹤林、李佳轩，有你们的陪伴，使我们的实验室充满了色彩和欢乐，你们在我的研究生生涯中不仅在学业上为我提供帮助和指导，在生活上也无时无刻不在关心我，希望毕业后大家还能有机会重聚。

还特别感谢我的家人，感谢我的爸爸、妈妈，谢谢您们在我浮躁的时候的鞭策和告诫，在我低潮时期的劝勉和鼓励，如今在你们的鼓舞下我以快要毕业踏入社会，希望我能不辜负你们的期望，成为你们的骄傲。

摘 要

近年来，深度学习的浪潮渗透在科研和生活领域的方方面面，本文主要研究深度学习在自然语言处理，尤其是古汉语自然语言处理方面的应用。本文旨在利用计算机帮助古文研究者对古汉语完成断代、断句、分词及词性标注等特殊而繁琐的任务，其中的断句、分词是不同于英文自然语言处理的，中文自然语言处理所特有的任务，尤其是断句任务更是古汉语自然语言处理所特有的任务。利用计算机处理古代汉语的各种任务有助于提高语言工作者的工作效率，避免人为主观因素误差，这将他们从繁重的古汉语基础任务中解脱出来，从而将更多的精力投入到后续的授受、义理等内容方面上的研究。

本文使用长短期记忆神经网络作为主体，并针对不同的古汉语自然语言处理任务，设计不同的输入输出结构来搭建具体模型，训练集使用的是网络上公开下载的古汉语语料，并且我们对其中的部分上古汉语语料文本进行了自己手工标记。本文中设计的模型可对古汉语文本完成断代、断句、分词及词性标注的操作。本文涉及的的主要工作和创新点如下：

（1）使用长短期记忆神经网络作为主体构建古代文本断代模型。在断代模型当中，文本中的每一个字被转换成一串高维向量，然后将文本包含的所有向量送入模型分析它们之间的非线性关系。最终，模型会输出一个该段文本的年代类别标签。实验结果表明利用 Bi-LSTM(Bi-directional Long Short-Term Memory, Bi-LSTM)神经网络构造的模型能够很好的完成断代任务，断代的正确率能达到 80%以上。本文的断代模型提供了一种高效且准确的古文断代方法，这将节省古文研究工作者在文本断代过程中的时间消耗。

（2）针对某些古代汉语书籍原著中缺少标点符号的问题，本文提出一个断句模型。本部分我们通过深度神经网络对大量已经断句的古汉语文本进行学习，使断句模型自动学习到某一时期、某种题材的断句规则，从而在后面的古代汉语文献信息化过程中，可以将断句工作交给计算机来完成，减少部分古汉语工作者的任务量。

（3）上面两个任务的训练集相对较好获取，他们只需获取不同年代的古籍文本然后对文本添加上年代标记或者句读标记即可，想要训练分词模型则需要已经分好词的文本来做模型的训练集，而目前尚没有公开的具有分词和词性标注的古汉语语料库。因此本文通过手工标记部分语料的方法得到了少量的数据集对我们所设计的分词标注模型进行少量的实验，用以验证本文提出的分词标注模型可以较好的完成古汉语分词标注任务。

论文以 Bi-LSTM 网络为主要结构，建立了一系列针对古代汉语文本不同任务的模型。实验证明，在现有有限的古汉语语料库中本文提出的模型已具备较好的效果，并可以应用到后续更大语料库的构建当中，作为辅助工具帮助古汉语工作者对文本的标记工作。新产生的语料库又可继续用来训练模型提高模型的精度，以此构成语料库和模型互相促进提高的局面，促进古汉语信息化及大型古汉语语料库的构建。

关键词： 古汉语，自然语言处理，断代，断句，分词，词性标注

Study on segmentation tagging algorithm and corpus of ancient Chinese based on machine learning

Abstract

A new generation of information technology represented by big data has penetrated into health care, health management and many other fields. It effectively changes the statistical classification method and thinking pattern of traditional medicine, and provides the outstanding capability of data mining and disease risk assessment capabilities for human beings. The medical survey data of the population cohort are complex and high dimensional, which contain a huge number of attributes and individual differences. It is of great significance for data mining and disease risk classification, and facing technical challenges at the same time.

The cohort study data of breast cancer are chosen as the research data. Breast cancer has the highest incidence of all the malignant tumors in women worldwide. The breast cancer risk classification model can help reduce the incidence rate of breast cancer. It is necessary to build an efficient classification model to perform accurate and economical diagnoses. Only the respondents classified into the high risk group need further checks to determine the breast cancer patients. The classification model must have a low false-negative rate, must be low-cost and also can easy to be extended.

The main work and innovations in the field of data mining and disease risk classification of this paper include：

（1）Aiming at the characteristics of high dimension and imbalance of medical data, this paper proposes a one-class F-score feature selection method for feature selection, and establishes a Naive Bayesian classification model based on one-class F-score feature selection method. The experiment results show that, with the presented method, the false-negative rate is decreased to 0.09 and the area under the receiver operating characteristic curve (AUC) is 0.776 with 8 features selected only. Compared with related methods, our method leads to the lowest false-negative rate and the lowest number of features selected and has a certain clinical value. It shows that the one-class F-score feature selection is capable of dealing with high dimensional balance data classification.

（2）This paper proposes an improved one-class F-score induction feature selection based on genetic algorithm. The experimental results of the improved model showed that the AUC reached 0.823 and obtain a better classification effect.

（3）In order to support the promotion of the disease risk classification model

in China's huge population base, this paper focus on the closed-form solution of the aforementioned classification algorithm and the explicit assessment of the classification of disease risk in the further studies. On the one hand，this paper proposed a closed-form formulation to describe the classification process. On the other hand, we proposed a method to express the risk probability of illness by tree structure based on probability.

This paper sets up a series of data mining and classification algorithms on medical data base on the one-class F-score feature selection. It has the clinical guidance value on breast cancer and can be extended to data mining and risk classification of similar diseases.

Key Words： **Ancient Chinese, Natural language processing, Judging the age, Punctuation, Word segmentation, Part of speech tagging**

目　录

1 引言

1.1 课题研究背景及意义

数随着科技不断发展，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。

自然语言处理是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。这一领域的研究涉及自然语言，即人类使用的语言，所以它与语言学的研究有着密切的联系，是将语言、文字进行信息化的基础。中文自然语言处理是自然语言处理的一个重要部分，中文相比于英文有历史更悠久、词边界较难鉴定、句法更灵活等众多特殊性，近年来，国内针对中文的自然语言处理的相关研究也逐渐受到重视。伴随着深度学习热潮兴起，中文分词、词性标注、命名实体识别和句子结构化表示等中文自然语言处理的研究也在深度学习技术的推动下获得了长足的发展。随着中文信息化的程度越来越深，我们越来越发现蕴含着中华民族千百年智慧的古汉语书籍更加需要我们利用现代化的技术进行妥善的保存、处理。对于古汉语文章书籍的组织、采录、收集、整理、纂修、审定也逐渐转移到了计算机上。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。对于实体古籍来说我们需要将其数字化、信息化，分门别类存档入库，这有利于我们传承传统文化和保护先人的思想精华。对于已经入库的电子书来说，我们需要利用现代中文自然语言处理技术对其进行更深层次的结构化处理和数据挖掘。这对于现有古汉语书籍的保存、知识提取和历史研究将重要意义，将促进我国的数字人文建设。然而现实是当前古汉语数字化进展大多停留在入库阶段，其后期的分词、词性标注、命名实体识别、文本结构化处理、文本分类等研究较少，目前所实现的一些方法其精度也不是很高。限制目前各类方法精确度的一部分原因是，中文有据可查的文字源自公元前 14 世纪的殷商后期，这时形成了初步的甲骨文，距今已延续了三千多年，而这三千多年的演变过程中，中文的字义、词义和句法等也在不断的动态变化中。以古文翻译来说，“汤”，原指一切热水，现在仅指食物煮后所得的汁水或烹调后汁特别多的副食，又如“治”的本义是平治水患，所以字从“水”旁，后来扩大为泛指一切治理。由此可见不同时代的中文，会有不同时代的特色，并不是一成不变的。面对我们中文历史源远流长的情况，试图构造出一种普适于各种时代的模型是很难实现的。所以只有判定了古籍所在的大致时间，才可以更加有针对性的对古籍进行后续研究，提高研究的精度和效率。

清代以来中国文人也在古文断代领域进行不断的探索，包括刘师培、章太炎等在义理方面进行研究，之后还有高本汉、戴闻达和尤锐等利用现代汉语方法进行研究。以上方法无论是从内容上还是从词法上，都充分说明了古文的文本内容中含有足够用来断代和辨伪的信息量，这是利用机器学习解决该类问题的必要条件。本文利用深度学习的方法，通过对某些年代的古籍进行学习，使深度学习模型自动的学习到不同时期的古籍中的文法、词汇规律，得到可自动断代的深度学习模型。模型从古

籍中自动提取出的高维信息中不仅包含虚词的词法信息，也包含了其他所有词类的词法信息，以此作为参考再进行后续的研究将大大的解放学者们的时间和精力。本部分所研究的模型在文本断代方向将有重要的实践意义。

本文的第一部分主要内容试图从古籍时间判定的角度在中国古文自然语言处理领域进行一定的探索，本文的研究成果将对古文分词、词性标注、命名实体识别、文本结构化处理、文本分类等其他方面的研究有所帮助。从技术角度来说，古文的时间判定就是指模型接收一段文本，模型自动计算并输出一个年代标签。因此，从输入输出的关系来看，古文时间判定任务即为一个文本分类任务。目前的文本分类模型，大致可分为两类，一类是基于规则或基于概率统计的传统机器学习方法，另一类是基于 CNN(Convolutional Neural Networks, CNN)、RNN(Recurrent Neural Networks, RNN)、Self-Attention 的深度学习方法。其中，基于规则或概率的方法相对简单，易于实现，在特定领域能取得较好的效果。其优点是时间复杂度低、运算速度快。但是在基于规则和概率的方法中，需要考虑很多规则或特定条件来表述类别，因此需要通过领域专家定义和人工提取特征。结合深度学习方法来解决特定领域问题是近年来的一个趋势。18 年 Google 提出基于 Self-Attention 机制的 Bert(Bidirectional Encoder Representations from Transformers, Bert)模型，大有一统 NLP(Natural Language Processing, NLP)领域之势。但是 Bert 模型主要面向现代语言，其成功主要依赖于当下互联网时代的海量信息化的文本，例如 wiki 百科、各类新闻媒体以及网络评论留言等，通过大量的训练集才得以训练出 Bert 模型中的参数，然而这一切在语料资源相对缺乏的古汉语领域并不适用。因此，本文提出使用 Bi-LSTM 深度学习网络模型解决自动化古籍时间断定即古汉语文本分类任务。在判断好年代的基础上，从而可以对古代汉语进行更有针对性的进行跟深层次的处理和研究，例如后面要提到的中国古汉语特有的断句任务以及中文自然语言处理特有的分词任务等。

断句在古代叫“句读”。古人称文辞语意已尽处为“句”，语意未

尽的停顿处为“读”。古书是不加句读的，只是一个字一个字地排列，读者边读边断，直至终篇。句读涉及到整个古代汉语和古代文化知识领域。如果初学者缺乏句读训练，就谈不上读位古书了。正因为句读是一种综合性很强的知识技能，在研究领域中，句读的添加与校正全部是由古汉语工作者使用手工完成的，我们看到的已标点的古书是经反复点校后再印的，古汉语工作者利用他们夜以继日的调研，笔耕不辍，完成了大量的古籍句读工作，这给我们阅读古书带来了方便。但古代书播浩如烟海，因此研究一种针对古汉语的断句模型十分必要的。

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，使计算机可以分析和处理中文自然文本的技术[1]。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 等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，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。本文的第三部分主要内容主要针对中文分词和词性标注进行论述。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]。

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 研究内容

本课题的研究目的是利用现有成熟的基于深度学习的自然语言处理技术对中国古汉语建立一系列模型，旨在完成古代汉语的自动断代、断句及分词标注任务，解放部分古汉语工作者的繁琐劳动，将这部分繁琐工作机器去完成，从而加速古汉语信息化过程。研究内容从课题研究目的入手，可分为以下几个方面：

（1）为解决古代书籍断代的问题，本文提出使用双向长短期记忆神经网络作为主体构建古代文本断代模型。整理互联网上现有的已知年代的文本作为训练集对模型进行训练。利用 word2vec 模型将文本中的每一个字被转换成一串高维向量，然后将文本包含的所有文字的字向量送入模型分析它们之间的非线性关系。最终，模型会输出一个该段文本的年代类别标签。实验结果表明利用 Bi-LSTM 神经网络构造的模型能够很好的完成断代任务，断代的正确率能达到 80%以上。本文的断代模型提供了一种高效且准确的古文断代方法，这将节省古文研究工作者在文本断代过程中的时间消耗。

（2）针对某些古代汉语书籍原著中缺少标点符号的问题，本文提出一个断句模型。本部分我们通过深度神经网络对大量经过断句的古汉语文本进行学习，使断句模型自动学习到某一时期、某种题材的断句规则，从而实现输入一段无断句的文字序列，机器自动为其添加断句的效果。

（3）针对古汉语分词及词性标注任务，我们需要解决训练集获取的问题，分词标注任务需要已经分好词、标注好词性的文本来做模型的训练集，但目前目前尚没有公开的具有分词和词性标注的古汉语语料库。因此我们通过手工标记部分语料的方法得到了少量的数据集对我们所设计的分词标注模型进行少量的实验，用以验证本文提出的分词标注模型可以较好的完成古汉语分词标注任务。

1.3 论文组织结构

论文的整体安排如下：

第一章作为绪论部分，首先对论文的研究意义及研究背景进行了简要的阐述。之后将研究中面临的主要问题和所做的工作内容进行简单的梳理，通过系统地归纳总结帮助读者了解论文中面临问题的本质和相对应的解决方法。最后对论文的大体结构进行简略介绍，方便读者了解整篇文章的体系架构。

第二章对课题相关的研究内容进行了详细介绍和总结。除了介绍中国古代汉语研究领域里古代文人关于著作年代的判断方法外，还介绍了自然语言

方法在古汉语断代方面的应用；此外还介绍了自然语言处理领域分词及词性标注任务的研究现状，并对分词及词性标注的常用方法、算法进行了总结和优劣分析。

第三章首先根据古汉语语料较少的特点，选择了双向长短期记忆神经网络结构作为模型的主体，并介绍了断代模型的总体结构框架。之后针对断代模型的多层结构，分层依次讲解了每一层的构成及作用。本章首次将双向长短期记忆神经网络应用到古代汉语的断代问题上去，通过两组实验分析了模型的性能，并简要分析出了同一时期内同一书籍中及不同书籍之间具有互相独立而统一的关系。

第四章对于古代汉语书籍没有标点符号的特点，利用字符标签的形式对输入的一句或多句古汉语文本进行标记，标记出应该含有标点符号的位置。本部分首先介绍了模型的数据来源和整体结构，然后介绍模型的代码实现，最后通过部分真实数据进行了一定的实验分析，分析证明模型的正确了较高，可以当做断句辅助工具供古汉语工作人员参考。

由于第五章模型任务的特殊性，第五章首先阐述了魔性训练集的数据来源及预处理，介绍了几种模型分类效果的评估标准。然后提出了本章的主要内容：基于双向长短期记忆神经网络的古代汉语分词及词性标注一体化系统。针对一体化系统，本部分创新性的提出了将两种标签进行一体化输出的编码方式，使得模型的输出可以同时带有分词及词性标注标签。关于模型分词及词性标注效果的评估，本文利用少量的手工标记的数据集对模型分别进行了分词实验和词性标注实验两部分实验，实验证明本部分提出的一体化系统在古代汉语分词及词性标注任务上有不错的效果，后期若有更加充足、准确的数据集后，该模型的准确率将可以达到更高。

第六章对研究工作进行了总结分析，并对下一步的研究方向和计划进行阐述。

2 研究综述

2.1 断代综述

在古籍断代领域，自晚清以来，刘师培、章太炎等学者对古籍的断代停留在古文经学的视角,主要是以书面文献为资料的授受、义理研究，但在这些方向上对于断代的研究结果没有太多的信服力。二十世纪初，瑞典学者高本汉(Bernhard Karlgren)著《左传真伪考》(On the Authenticity and the Nature of Tso Chuan)，第一次将西方现代语言学的方法应用到中文古籍的断代和辨伪真伪问题。高本汉利用西方现代语言学方法，考察了《左传》的 7 组虚词，分析、归纳《左传》中的语法体系，并将其与先秦其它古籍的语法体系相对照，以此作为《左传》断代的依据。胡适的《左传真伪考》中评价其研究方法“完全用文法学的研究来考订《左传》”，是一种“开山的工作”。依据高本汉的语言学方法，美国汉学家李克考察了《管子》的虚词，除了某些篇章太短无法分析，其它篇章的分析与高本汉总结的“公元前 3 世纪的标准语”颇为契合[]。荷兰汉学家戴闻达(Duyvendak)也依据高本汉归纳的(先秦经典中)虚词的语法特征考察《商君书》的真伪，发现辨伪的“标准”仅依赖高本汉归纳的那少数几组虚词还不能奏效，要完成《商君书》的辨伪，还需要考察更多的虚词[]。20 世纪六七十年代，杜百胜承继高本汉从虚词入手为古籍断代的方法，对虚词做了穷尽式的研究，从历史语言学的角度，描写、分析虚词的语法特征，以确定古籍的年代。在 2l 世纪初，尤锐继续从事古籍年代的考订。尤锐将先秦时期便存在的尚未完全定型的文本称为“著作底本”，这是先秦典籍的核心部分，汉以后有窜入，但是居于少数，而这个著作底本在文献成型过程中的里程碑作用不能被轻易忽视。他的研究从高本汉、杜百胜所从事的虚词语法研究转移到词汇的研究。他的论文《战国时期的词汇变化》(Lexical Changes in Zhanguo Texts)考察了公元前 5 世纪到 3 世纪的史书和子书中的词汇差异，共选择了 7 个有使用时间差异的词汇，用以确定各“著作底本”的年代[]。

以上现代语言学方法虽然不局限于义理和授受等思想内容上的研究，

但仍受限于人力。大多数研究仅选择某几个关键的虚词文法或者有时间差异的词汇进行研究，其研究的可靠性受词汇选择的影响较大。杜百胜意识到不能仅取个别词汇进行研究，他利用穿孔卡片将文本所有的文句都抄录一遍，再逐句分析。即便如此也仅仅覆盖了古籍中的虚词部分。为追求可靠性的这种穷尽式地考察文本的语法、词汇，以现代语言学方法从事中国古籍的辨伪与断代，在现今仍然大有可为。

从技术角度来说，古文的时间判定就是指模型接收一段文本，模型自动计算并输出一个年代标签。因此，从输入输出的关系来看，古文时间判定任务即为一个文本分类任务。目前的文本分类模型，大致可分为两类，一类是基于规则或基于概率统计的传统机器学习方法，另一类是基于 CNN、RNN、Self-Attention 的深度学习方法。其中，基于规则或概率的方法相对简单，易于实现，在特定领域能取得较好的效果。其优点是时间复杂度低、运算速度快。但是需要考虑很多规则或特定条件来表述类别，因此需要通过领域专家定义和人工提取特征。

结合深度学习方法来解决特定领域问题是近年来的一个趋势，Collobert 和 Tang 分别将 CNN 和 RNN 应用到文本分类中，18 年提出基于 Self-Attention 机制的 Bert 模型，在不同的 NLP 任务中均有很好的表现。但是 Bert 模型主要面向现代语言，其成功主要依赖于当下互联网时代的海量信息化的文本，例如 Wiki 百科、各类新闻媒体以及网络评论留言等，通过数以 T 计的训练集才得以训练出 Bert 模型中 200M 的模型参数，然而这一切在语料资源相对缺乏的古汉语领域并不适用。因此，本文使用 LSTM 深度学习网络模型解决自动化古籍时间断定即古汉语文本分类任务，该模型主要有两个优点，一是不借助人工提取规则特征，二所需数据量比基于 Self-Attention 机制的模型相对较少。

2.2 断句综述

古代书籍没有标点符号，诵读时根据文章所作的停顿，或在古书上按停顿加的圈点，就叫“断句”。断句在古代叫“句读”。古人称文辞语意已尽处为“句”，语意未尽的停顿处为“读”。古书是不加句读的，只是一个

字一个字地排列，读者边读边断，直至终篇。古代教育童蒙读书非常重视句读能力的训练。《礼记学记》说：“一年视离经辨志”。“ 离经”就是指断句，也就是现在所说的句读能力的训练。谚语说：“学问如何看点书”。意思是学问的大小可以从句读能力上看出，因为断句的正确与否，就看对书中文字懂不懂，文字懂了句也就晰对了。如果不位而硬断，自然会乱点鸳鸯谱，同一段文字就会文义迎异。比如古人描绘龙：“其形有九似：头似驼，角似侄，冈似兔，耳似牛，项似蛇，匠似蜃，鳞似鲤，爪似鹰，掌似虎。”如果断成：“其形有九：似头，似角，似鹿限，似兔耳，似牛头…”这样错误的断句，意思变成龙有九种形状，并且面目全非，显然是很可笑的。所以正确地句读是提高古汉语阅读能力、正确理解古书的重要基本。

研究一篇文言文，应该先通读这篇文章，大体上弄清这篇文章写了什么内容，思考它表达了什么意思，具有什么结构，它属于什么文体等，然后再来给文章断句。断句有下面一些基本方法：

弄通文意断句。给文言文断句，首先要阅读全文，了解文意，这是断句的先决条件，如果想当然地断下去，就容易发生错断。通读全文，搞清属于什么文体，写了什么内容，想表达什么意思。要注意文言文单音词占多数的特点，抓住几个关键的字词翻译以理解文段大意。

利用对话标志断句。常以“曰”、“云”、“言”为标志，两人对话，一般在第一次问答写出人名，以后就只用“曰”而把主语省略。遇到对话时，应根据上下文判断出问者、答者，明辨句读。

借助文言虚词断句。古人的文章没有标点符号，为了明辨句读，虚词就成了重要的标志。尤其是一些语气词和连词的前后往往是该断句的地方。文言文，多用虚词来表达语气或感情。句首发语词：夫、盖、至若、若夫、初、唯、斯、今、凡、且、窃、请、敬等常用于一句话的开头，在它们的前面一般要断开。句尾词：也、矣、焉、耳等经常用于陈述句尾；耶、与（欤）、邪（耶）等经常用于疑问句末尾；哉、夫等经常用于感叹句尾。其后面一般要断开。疑问语气词：何、胡、安、曷、奚、盍、焉、孰、孰与、何如、奈何、如之何、若之何等词或固定结构之后，一般可构成疑问句，只要贯通上下文意，就可断句。复句中的关联词：虽、虽然、纵、纵使、向使、假使、苟、故、是故、则、然则、或、况、而况、且、若夫、至于、至若、已而、于是、岂、岂非，在它们的前面一般要断开。其它的如：以、于、为、则、而，往往用于句中，在他们的前后一般就不断句；（“而”表转折而且后面为一个比较长和完整的句子时，“而”前面要断开）。

找出动词，明确句意。古汉语中，句子多以动词或形容词谓语为中心。

找出了动词或形容词谓语，也就区分出独立的句子，明确了语句的意思，从而正确断句。比如：马无故亡而入胡/人皆吊之，句中动词有“亡”“入”“吊”，因此可区分出两个句子。②其马将胡骏马而归/人皆贺之，句中动词有“将”“归”“贺”，可区分出两个句子。

借助名词（代词）断句。一般完整的句子都有主谓宾，而主语一般由名词或代词充当。名词一般为文章陈述、描写、说明或议论的对象，在它们的前后往往要进 行断句。名词（代词）一般也常常用作句子的主语和宾语，因此，找出文中反复出现的名词或代词，就基本上可以断出句读了。常见代词有：吾、余（表示“我”），予、尔、汝（女）、公、卿、君、若（表示“你”）、彼、此、其、之（表示“他”）。

借助语法结构断句。文言语法中有一些固定结构，如：“……者，……也”、 “不亦……乎”、 “何……之有”， “孰与……乎”、 “为……所……”、“受……于……”等，根据这些结构也可断句。

利用总分关系断句。文言文中常用总说分承或分说总承的写法，掌握了这个写法对断句很有帮助。如《谋攻》的最后一段：“故知胜有五/知可以战与不战者胜 /识众寡之用者胜/上下同欲者胜/以虞待不虞者胜/将能而君不御者胜。”这显然是总说分承的写法了。再如“老而无妻曰鳏/老而无夫曰寡/老而无子曰独/幼 而无父曰孤/此四者天下之穷民而无告者”，这显然是分说总承的写法了。

借助对比、对偶、排比、顶真等修辞断句。文言中常有对偶句、排比句，抓住这个特点断句，常能收到断开一处、接着断开几处的效果。例：秦孝公据崤函之固拥雍州之地君臣固守以窥周室有席卷天下包举宇内囊括四海之意并吞八荒之心当是时也商君佐之内立法度务耕织修守战之具外连横而斗诸侯于是秦人拱手而取西河之外……这一段文字之中，“据崤函之固/拥雍州之地”是对偶；“席卷天下/包举宇内/囊括四海/并吞八荒”是排比；“内” “外”是对照。根据这样的语言特点，断句也就容易多了。顶真是文言文中常见的形式。句子前后相承，前一句做宾语的词，在后一句又作了主语。例如：“畏惧则存想，存想则目觑。”（王充《订鬼》）根据这一特点，我们也可以确定句读。“名不正则言不顺/言不顺则事不成/事不成则礼乐不兴/礼乐不兴则刑罚不中/刑罚不中则民无所措手足”。

利用对称句式。注意古文讲究整齐对称、行文中上下句常用相同的字数和相同的结构的特点。如“故福之为祸/祸之为福/化不可极/深不可测也”，句式工整，都为四字一句，据此可正确断句。

实词断句法。即在读懂全文，了解所点断文章的大致内容的基础上，通过找名词与动词来组句，先断开能断的句子。如果是叙述性的文章，就要弄懂故事的基本情节；若有人物对话，就要弄清谁与谁对话，讲的什么话。如是说理性文章，则要弄明白谈了哪些问题，表明了怎样的观点。同现代汉语语法一样，古文中的主语、宾 语一般是名词（代词），谓语多是动词，主语、谓语与宾语是句子的主干，而谓语是句子的核心。因此，抓住谓语动词，分析动词与它前后词语之间的关系，就能正确断句。如：乡 人 管 彦 少 有 才 而 末 知 名 裒 独 以 为 必 当 自 达 拔 而 友 之 男 女 各 始 生 便 共 许 为 婚。找出句中几个名词，句子基本就断开了：乡人管彦少有才而未知名/裒独以为必当自达/拔而友之/男女各始生/便共许为婚。动词断句法的难点在于介于两个动词之间的名词或名词性短语，它们属上作前一动词的宾语还是属下作后一动词的主语，这要结合具体语境，反复推敲。如上句“马无故亡而入胡人皆吊之”、“其马将胡骏马而归人皆贺之”， 是断在“胡”后还是断在“人”后，是断在“归”后还是断在“人”后，颇费思量。根据语境，“胡”应为胡地，“归”意为自己家里，与后文“人皆贺之”意义关联。如果“归人”的话，就谈不上“贺”了。再从句式上看，“人皆吊之”与“人皆贺之”，句式整齐对称。根据以上分析，即可正确断句。所要补充指出的是句子中一些专有名词（如地名、人名、官职名、器物名等）和连得很紧的词语（如双音节的词语）间不能点断。

虚词断句法。古人的文章没有标点符号，为了明辨句读，虚词就成了重要的标志。尤其是一些语气词和连词的前后，往往是应该断句的地方。我们只要抓住了这些虚词，了解它们在句中的位置，断句也就准确迅速了。

发语词和句首助词常出现于句首，有领起全句的作用，其前自可断句。复句中的关联词一般也用在句首，在这些关联词前可点断。句末语气词等常用在句末，其 后往往能断句。连词如“以、于、为、而、则”等经常出现在句中，后面不能断开。

“曰”字断句法。文言文中对话、引文常常用“曰”、“ 云”、“ 言”为标志，一般情况下碰到它们都要停顿，且大多用冒号顿开，后面“曰”的内容一般要加双引号。如果两人对话，一般在第一次问答出现人名，以后就只用“曰”， 而把主语省略。例如：太 宗 谓 太 子 少 师 萧 禹 曰 朕 少 好 弓 矢得 良 弓 十 数 自 谓 无 以 加 近 以 示 弓 工 乃 曰 皆 非 良 材 朕 问 其 故 工 曰 木 心 不 直 则 脉 理 皆 邪 弓 虽 劲 而 发 矢 不 直。正确答案为：太宗谓太子少师萧禹曰/朕少好弓矢/得良弓十数/自谓无以加/近以示弓工/乃曰/皆非良材/朕问其故/工曰/木心不直/则脉理皆邪/

弓虽劲/而发矢不直。这个文言语段划线部分是转述一段对话，三次“曰”的出现、两问两答的过程都可成为断句的参考。

除此之外，断句的方法还有很多，如根据押韵规律断句、根据间隔反复断句、特殊句式等断句，要想断句准确度更高，就要综合运用这些方法。

综上所述，句读涉及到整个古代汉语和古代文化知识领域。如果初学者缺乏句读训练，就谈不上读位古书了。正因为句读是一种综合性很强的知识技能，在研究领域中，句读的添加与校正全部是由古汉语工作者使用手工完成的，我们看到的已标点的古书是经反复点校后再印的，古汉语工作者利用他们夜以继日的调研，笔耕不辍，完成了大量的古籍句读工作，这给我们阅读古书带来了方便。但古代书播浩如烟海，因此需要研究一种针对古汉语的断句模型。

2.3 分词综述

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。

2.3.1 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，如图 2-1 所示。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。

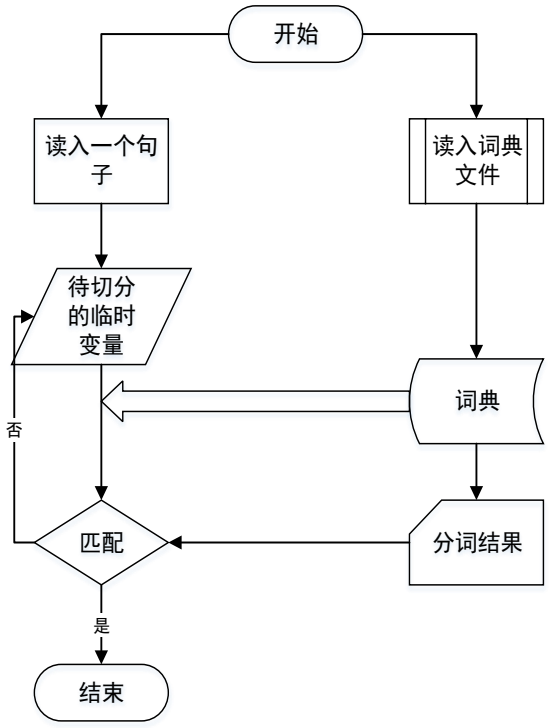

图 2-1 字典法分词流程图

2.3.2 基于统计的方法

1、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间的增大而有所减慢。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。

2.3.3 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。

2.3.4 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。

2.3.5 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。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。

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，古汉语的自动分词及词性标注技术仍是中文自然语言处理的一项重要难题。

2.4 词性标注综述

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。

2.4.1 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。

2.4.2 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。

2.4.3 统计与规则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。

2.5 本章小结

本章针对古代汉语尤其是上古汉语自然语言处理的一些特殊问题，选取三个具有代表性的方向利用现代自然语言处理技术进行了深入研究，它们分别是古汉语书籍断代问题、非平衡数据分类的研究现状和特征选择的研究现状。

以乳腺癌风险分类模型为背景，从医学角度介绍了疾病与表浅数据的关系，并详细阐述了疾病风险分类模型尤其是 乳腺癌风险分类模型的研究现状。同时罗列了应用较广泛的几种已有乳腺癌模型，并对几种模型进行了总结分析。最后指出针对现有乳腺癌风险分类模型不适合中国女性筛查的情况下建立适合中国女性乳腺癌筛查模型的重要性。

针对建立模型面临的两个挑战之一——非平衡数据的分类问题，本章首先介绍了非平衡数据的概述以及这种现象对分类器的分类效果产生的影响。同时总结了现有研究中非平衡数据分类的发展现状，此外还对几种常用解决方法进行了详细的阐述并指出各自的优缺点。

针对研究面临的另一个挑战——特征选择问题，本章首先对特征选择进行了概述介绍，其次总结了生成特征子集的几种方式、各自优缺点以及目前的研究方向。最后阐述了特征选择中如何对特征进行评价及其对应的评价准则，为后续章节提出新的特征选择法提供了理论依据和实验可行性。

3 古代文本断代模型

本部分要解决古代书籍时间判定的问题，首先是获取待判定的一部书籍或者书籍中的一段文本，其字符序列表示为 S，将此文本送入模型 M，模型进行计算并输出一个年代标签 T。

$$S = \{x_1, x_2, x_3, ..., x_n\} \tag{3-1}$$

$$T = M[g(\{x_1, x_2, x_3, ..., x_n\})] \tag{3-2}$$

这个过程中有以下几点关键技术需要注意，文本序列的向量化表示、模型总体结构和以及长短期记忆神经网络记忆单元结构。古代文本断代模型的结构框图如下图所示：

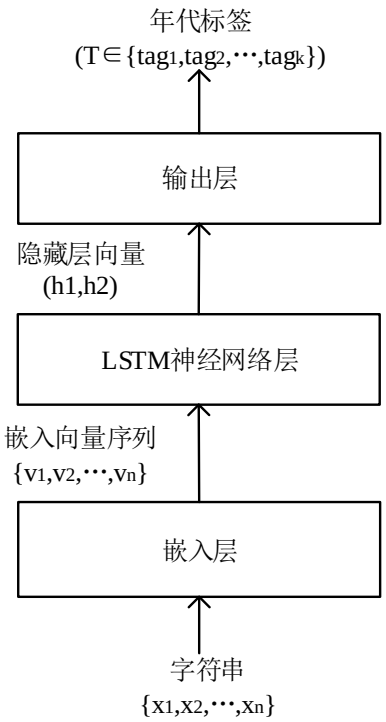

图 3-1 古代文本断代模型结构框图

3.1 模型结构

古籍断代模型接收一段古代文本作为输入送入模型，文本首先被送入嵌入层使模型获得该段文本的向量化表示。然后字符的嵌入向量被逐字送入双层 LSTM 神经网络层中分别计算正向和反向的隐藏向量。最终将两个隐藏向量串联送入输出层计算出最终的预测结果。模型的细节如图二所示：

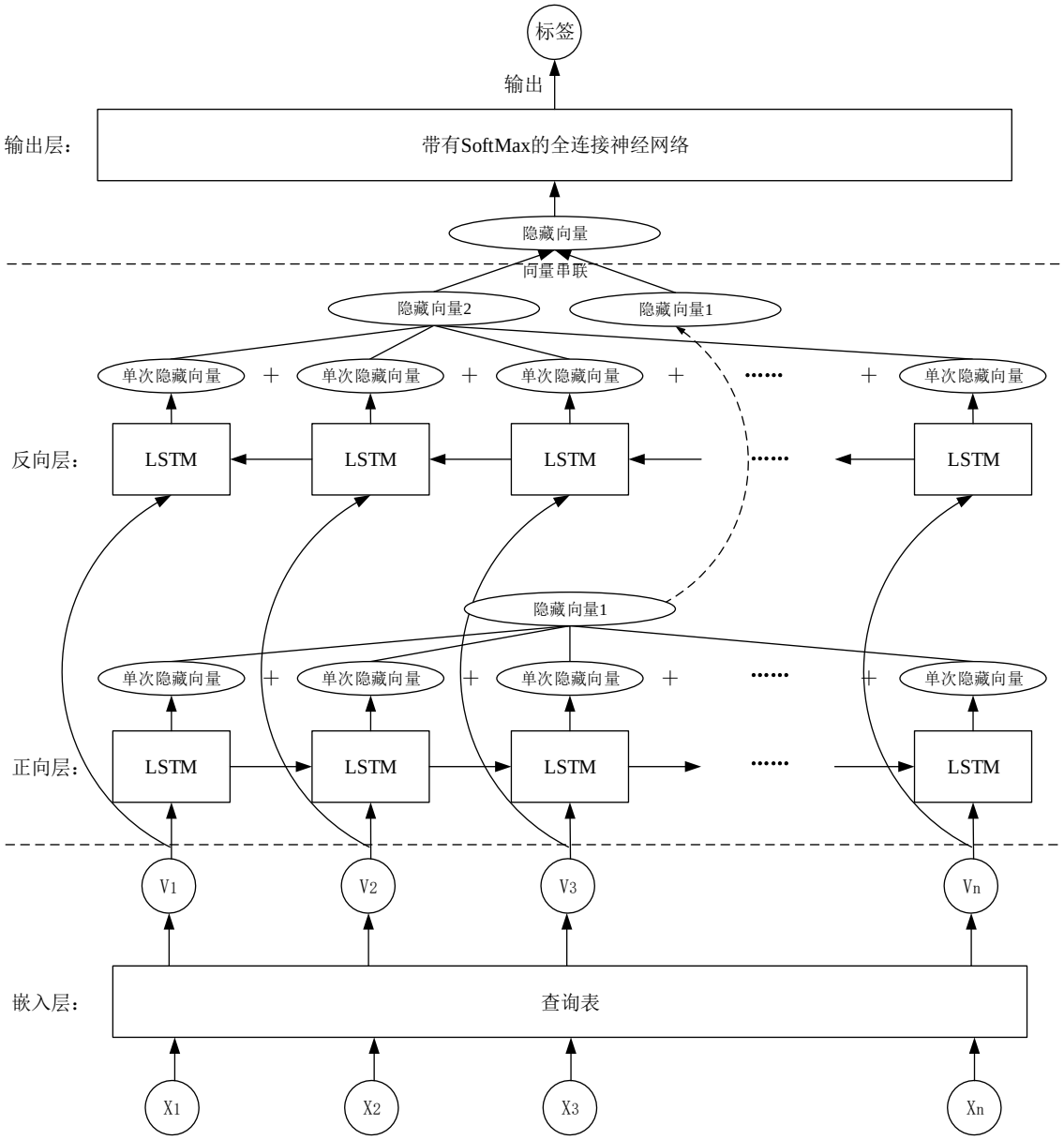

图 3-2 古籍断代模型结构图

模型主要分为三层，嵌入层、LSTM 神经网络层和输出层。其中嵌入层使用的是 word2vec 中的 CBOW 模型，即通过文本上下文预测中间字的方式实现中间字的向量表示；神经网络层使用双层反向的 LSTM 作为主体，第一层接受文本的正向输入，输出一个隐藏层向量V_1，第二层接受文本的逆向输入，输出一个隐藏层向量V_2，然后将两个向量串联，由于 LSTM 具有记忆上文信息的能力，因此两个隐藏向量相当于将全文信息编码成一个向量表示；输出层为一个全连接网络，用于将上层得到的向量进行解码、计算，输出相应的预测年代。下面本文将对模型的三层结构进行详细的介绍。

3.1.1 嵌入层

在式 1 中，文本字符序列 $S=\{x_1,x_2,x_3,...,x_n\}$ 仍是人类所能阅读的文字形式，$\{x_1,x_2,x_3,...,x_n\}$ 等为一系列汉字的有序组合，式 2 中 g 映射的作用是将人类所能理解的文字序列形式的 S 转化成计算机所能理解的向量化表示 V：

$$V=\{v_1,v_2,v_3,...,v_n\}=g(S)=g(\{x_1,x_2,x_3,...,x_n\}) \tag{3-3}$$

其中 $\{v_1,v_2,v_3,...,v_n\}$ 分别为 $\{x_1,x_2,x_3,...,x_n\}$ 对应的每个字的向量化表示，这个将文字映射为向量表示的过程就叫做字嵌入。

目前字嵌入方式有两种，一种是对所有字符进行 one-hot 编码，但 one-hot 编码有编码过长，不同字之间向量相互垂直，没有语义联系，不能表示位置信息等缺点，因此我们不使用 one-hot 编码；另一种是稠密的低维向量表示(Dristributed representation)，它的思路是通过训练，将每个词都映射到一个较短的词向量上来，一方面可以解决 One-hot 编码过长的问题，另一方面向量也可携带一定的语义信息。Google 在 2013 年开源的 word2vec 是一种利用神经网络进行字嵌入训练的一个语言模型。他假设字向量是服从分布式假设的，如果两个词的上下文时相似的，那么他们语义也是相似的。word2vec 模型的训练输入是某一个特征词的上下文相关的词对应的词向量，而输出就是这特定的一个词的词向量，如图所示。训练完毕后，输入层的每个单词与矩阵 W 相乘得到的向量的就是我们想要的词向量(word embedding)，这个矩阵也叫做查询表(look-up table)。

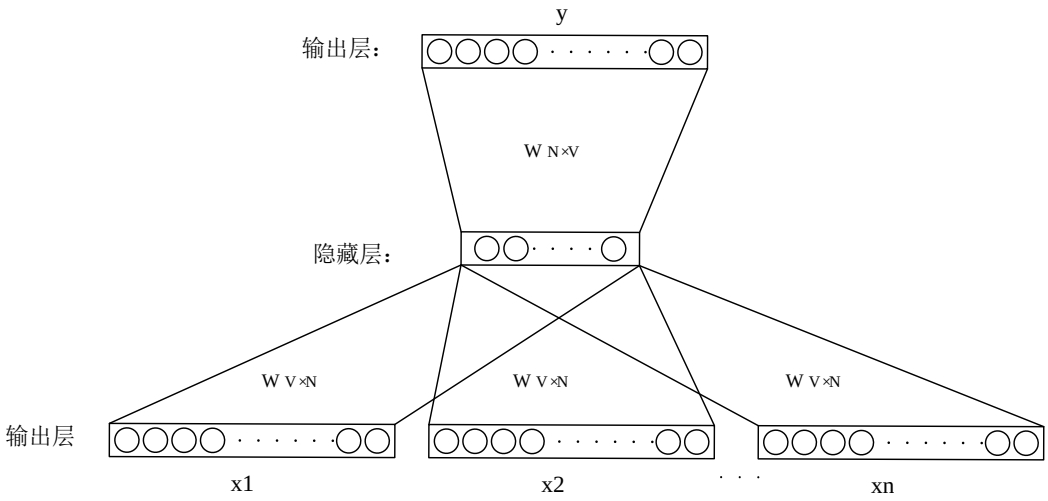

图 3-3 word2vec 网络结构示意图

Word2vec 模型是非监督的，资料获取不需要很大的成本，我们可以通过 word2vec 模型在大量的未标注的语料上学习，就可以学习到比较好的向量表示，可以学习到词语之间的一些关系。比如男性和女性的关系距离，时态的关系，学到这种关系之后我们就可以把它作为特征用于后续的任务，从而提高模型的泛化能力。

同文字的向量化道理相同，时间标签同样需要考虑人类标签和机器理解的问题，本文使用的时间标签为年代标签，如汉代、唐代中期，首先将此各种时间标签 $T_m \in T_M$ 按时间顺序排序 $\{T_1,T_2,T_3,...,T_n\}$，由于年代数相对较少，因此可以将朝代标签进一步表示为 one-hot 编码，方便模型通过 Softmax 评估出某一朝代。

3.1.2 长短期记忆神经网络层

LSTM 神经网络通过输入门、输出门和遗忘门这些门结构来控制长期记忆的遗忘与保存，LSTM 记忆单元在时间维度上展开的结构图如图所示。

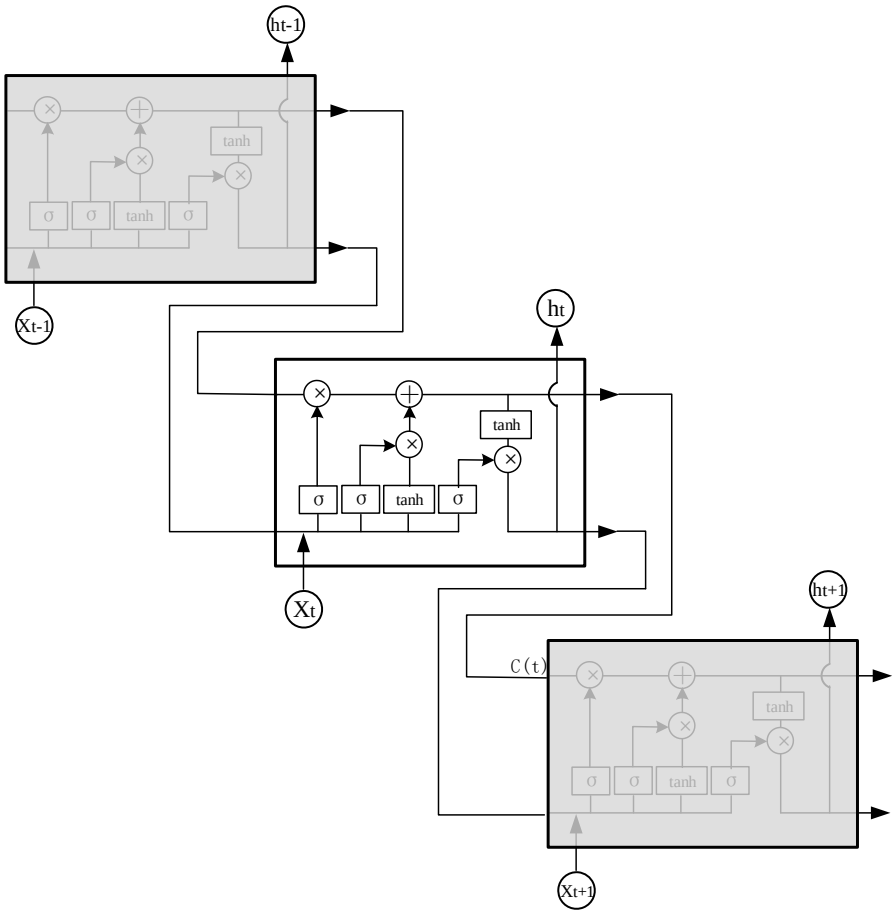

图 3-4 LSTM 记忆单元在时间维度展开结构图

LSTM 记忆神经元中控制信息的门结构公式如下：

$$i^{(t)} = \sigma(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + W_{ic}c^{(t-1)}) \tag{3-4}$$

$$f^{(t)} = \sigma(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + W_{fc}c^{(t-1)}) \tag{3-5}$$

$$c^{(t)} = f^{(t)} \bullet c^{(t-1)} + i^{(t)} \bullet \phi(W_{cx}x^{(t)} + W_{ch}h^{(t-1)}) \tag{3-6}$$

$$o^{(t)} = \sigma(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + W_{oc}c^{(t)}) \tag{3-7}$$

$$h^{(t)} = o^{(t)} \bullet \phi(c^{(t)}) \tag{3-8}$$

公式中，$i^{(t)}$ 为输入门，σ 是 sigmoid 函数，sigmoid 函数的值域为[0，1]，起到将输入门向量中的每个元素限制在 0 到 1 之间。因此，输入门向量和另一个向量的哈达玛乘积就是对向量的每个维度进行一定的缩放，即可理解为保存或者遗忘向量中的某些信息。ϕ 是 tanh 函数，用来将向量的每个元素映射到[-1，1]之间。$c^{(t)}$ 是长期信息，用来保留距离当前时刻较远处的有用信息，$f^{(t)}$ 是遗忘门，二者相乘可以用来控制长期信息中的某些信息是否继续保存到下一时刻，结合当前时刻的输入可以得到新的 $c^{(t)}$，此外还有 $o^{(t)}$ 输出门，用来控制输出信息，从而影响到当前时刻的输出 $h^{(t)}$ 以及下一时刻的输入。正是这些门结构使得 LSTM 可以有效的保留前文有用信息、遗忘前文中的无用信息，从而能做出更好预测。

3.1.3 输出层

输出层是一个全连接网络，结构如图所示：

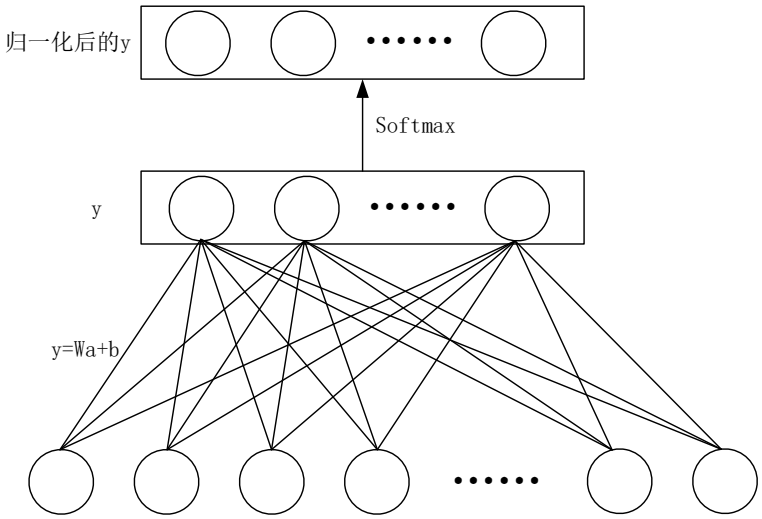

图 3-5 输出层网络结构图

他接受一个上层传入的 $1\times 2n$ 维度的向量 $\{a_1, a_2, a_3, ..., a_{2n}\}$，其中 n 为

LSTM 神经网络输出的隐藏层向量维度。输出层包括一个 $2n\times k$ 维度的 W 权重矩阵以及一个 $1\times k$ 维度的偏置矩阵，其中 k 为所有分类的类别数，输出层的输出为 $y=Wa+b$ ， y 是一个 $1\times k$ 维的向量，最后通过 Softmax 函数对其进行输出归一化，归一化后的某一维度的数值即可视作该维度所表示年代的预测概率。Softmax 公式如下。

$$\sigma(z)_j=\frac{e^{z_j}}{\sum_{k=1}^{K}e^{z_k}} \tag{3-9}$$

3.2 实验

实验所用的数据集为从网络的开放数据库下载的不同年代的古籍。根据古籍所处具体时期的不同，将其分为了春秋战国时期、后汉时期、南北朝时期、宋元时期及明清时期五个时间段并分别使用 T1、T2、T3、T4 以及 T5 表示，对应关系如下表所示。

表 2 时间标签与年代对应表

时间标签	年份	时期
T1	公元前 770-公元前 221	春秋战国
T2	公元 25 年-公元 220 年	后汉
T3	公元 420 年-公元 589 年	南北朝
T4	公元 960-公元 1368 年	宋元
T5	公元 1368 年-公元 1912 年	明清

在上述的每个时期中，我们分别选择著成于当前时代的古籍作为训练集，各个时期具体书目如下表所示，总计 644 万字。

表 3 各时期书目表

时间标签	书名	字数（万字）

T1	《春秋公羊传》《论语》《春秋左传》《国语》《道德经》《尚书》《吕氏春秋》	90
T2	《道行般若经》《法镜经》《佛说兜沙经》	20
T3	《洛阳伽蓝记》《颜氏家训》《齐民要术》《世说新语》	43
T4	《大唐三藏取经诗话》《窦娥冤》《全相平话》《关大王独赴单刀会》《新编五代史平话》《赵盼儿风月救风尘》	76
T5	《红楼梦》《金瓶梅》《水浒传》《西游记》《儒林外史》	415

模型有两个评估指标：一个是单句分类正确率 Ps；另一个是书籍分类正确率 Pb：

Ps=N 时间判定正确的句子条数/总句子条数 (3-10)

Pb=时间判定正确的书籍数/总书籍数 (3-11)

其中一本书籍时间判定的决策取决于书籍内部所有句子决策的投票结果，统计所有句子判定结果后，投票数最多的时间分类即判定为该书的时间段。在实现方面，模型通过 Python 的 TensorFlow-GPU 框架实现，所用的硬件配置为 8G-cpu 1080Ti 显卡。

训练流程图如下图所示：

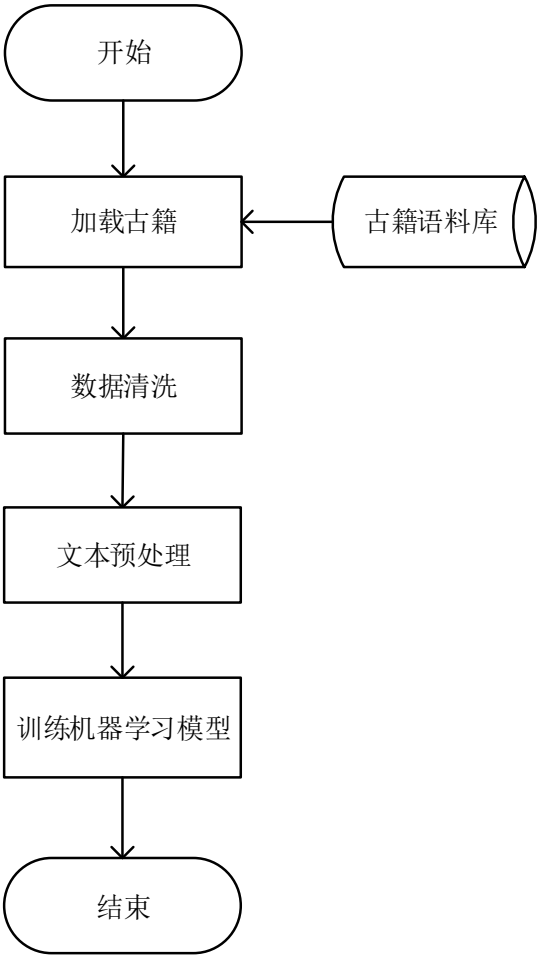

图 3-6 断代模型训练流程图

3.2.1 参数选择

模型涉及到一些超参数的选择，我们进行了大量的实验来比较不同超参数情况下模型的训练过程。本文中我们选择对不同维度的隐藏层(Hidden Layer Dimension，HLD）和单词嵌入向量(Embedding Layer Dimension，ELD)进行比较。下图为在不同超参情况下训练过程的训练精度曲线。可见，当 HLD 为 64，ELD 为 64 时，收敛速度最快，在其他参数下虽然收敛速度较慢但也能达到较高的精度。

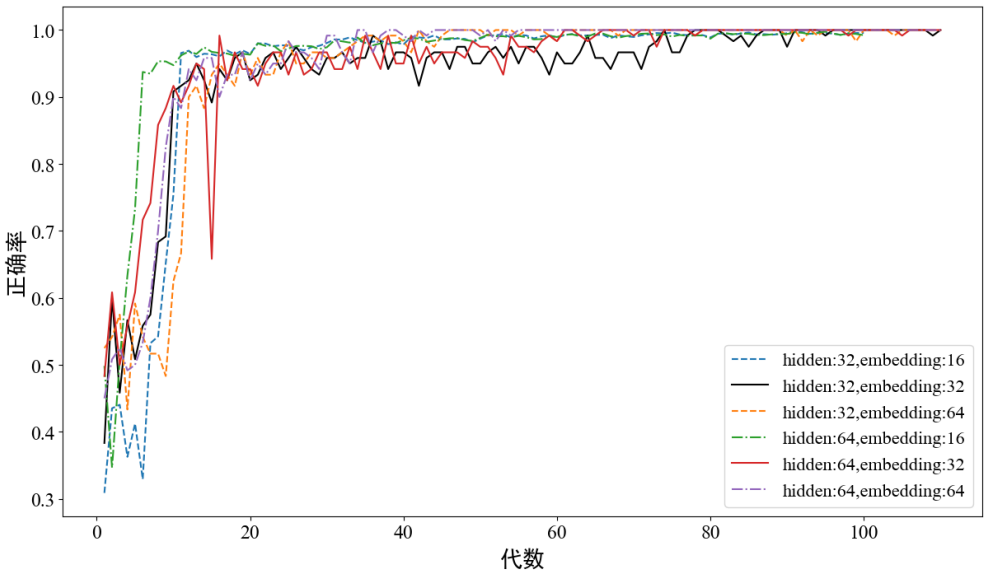

图 3-7 不同超参下训练过程的训练精度曲线

同时，我们比较不同参数时模型在测试集上的表现，实验结果如表 4 所示。表中可见，虽然 HLD 为 64，ELD 为 64 的参数在训练集中收敛速度、精确度表现较好，但是在测试集中其表现却不够优秀，而 HLD 为 32，ELD 为 32 时在训练集上虽然在收敛速度上略逊于 HLD=64，ELD=64 的参数设置的模型，但其在测试集上的正确率高于 HLD=64，ELD=64 的模型。我们初步猜想这种情况产生的原因是模型在 HLD=64，ELD=64 的时候训练过程中过拟合了，模型对训练集拟合程度过高，而对于其他数据集其表现将有所欠缺。因此，我们后面的实验将统一使用 HDL=32、ELD=32 的参数设置下训练得到的模型进行。

表 4 不同超参情况下模型在部分测试集上的表现情况

隐藏层向量维度 (HLD)	字嵌入向量维度 (ELD)	正确率
32	16	0.921
32	32	0.937
32	64	0.910
64	16	0.925
64	32	0.912
64	64	0.925

3.2.2 断代实验

在第一个实验中，各个时期的所有书中都选出部分章节作为训练集，保

留每本书的一部分章节不参与训练作为最后的测试集，以此来保证某个时期内的所有书籍均有部分语句用于训练。如图所示：

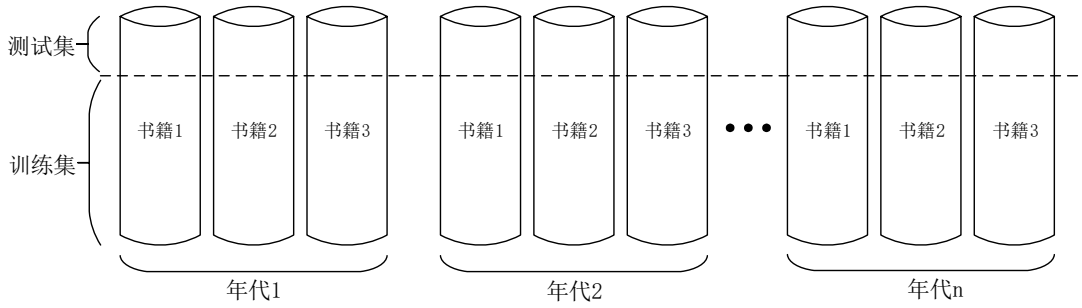

图 3-8 实验一中训练集和测试集的获取方法示意图

然后我们将测试集中的文本送入模型，来预测他们的年代。下表为实验结果。表中的每一行表示输入为某时代的古籍文本时，模型的预测为不同时代的结果的句子条数。实验表明，当所有书籍均有部分语句用于训练时，判断一个句子为正确时期的概率很大。这说明了某一时代内同一本书中的句法语法结构基本一致，模型学习了部分章节的结构信息后，可以较好的适用于同本书的其他章节中。

表 5 实验一结果

输入 \ 输出	T1	T2	T3	T4	T5
T1	15672	536	524	359	186
	(88.3%)	(7.5%)	(5.7%)	(2.5%)	(0.9%)
T2	821	5563	675	416	669
	(4.6%)	(78.1%)	(7.3%)	(2.9%)	(3.2%)
T3	897	485	7531	1069	561
	(5.0%)	(6.8%)	(82.2%)	(7.4%)	(2.7%)
T4	111	239	231	11569	1239
	(0.6%)	(3.3%)	(2.5%)	(80.6%)	(6.0%)
T5	239	298	196	924	17880
	(1.3%)	(4.1%)	(2.1%)	(6.4%)	(87.1%)
共计	17740	7121	9157	14337	20535
	(100%)	(100%)	(100%)	(100%)	(100%)

我们也做了另外一个实验，在该实验中我们在每个时期中挑选其中几本

书作为测试集不参与训练，用该时期剩下的书籍文本训练模型，如图 3-9，去观察模型对训练集之外的书籍的断代效果。

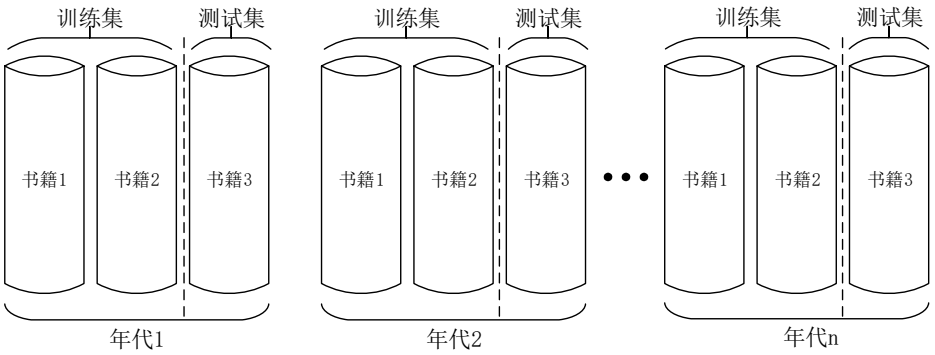

图 3-9 实验二中训练集和测试集的获取方法示意图

我们以《左传》为例进行了实验。在训练集没有输入《左传》任何文本的情况下，我们输入《左传》文本，让模型预测其年代。模型对《左传》的预测结果如图 3-10。在所有《左传》的句子中，模型将其中 1749 句预测为春秋战国时期，将 891 句判断为后汉时期，712 句判断为南北朝时期，104 句判断为宋元时期，121 句判断为明清时期。在训练集中不包括某本书的情况下，模型对该书的单句断代正确率(48.9%左右)相比实验一降低了不少，可见一本书籍在词法、句法上有一定的独立性。而从整体上看，我们利用一本书中的所有句子断代结果为该本书进行头片，投票占比最多的时代即判断为该书的时代。在这种情况下，图中可见，我们将《左传》判断为 T1 标签，即模型最终判断《左传》是春秋战国末期的书籍，此结果与中国古籍研究领域目前的共识是一致的，也印证了我们模式的正确性。该实验证明了同一时期内的某一本书虽然同其他书籍在词法句法上有一定的独立性，但是总体而言同一时期内的书籍之间也是有潜在联系的，也有一定的统一性，因此模型可通过学习同一时期内一定量的书籍来判断其他书籍的信息。

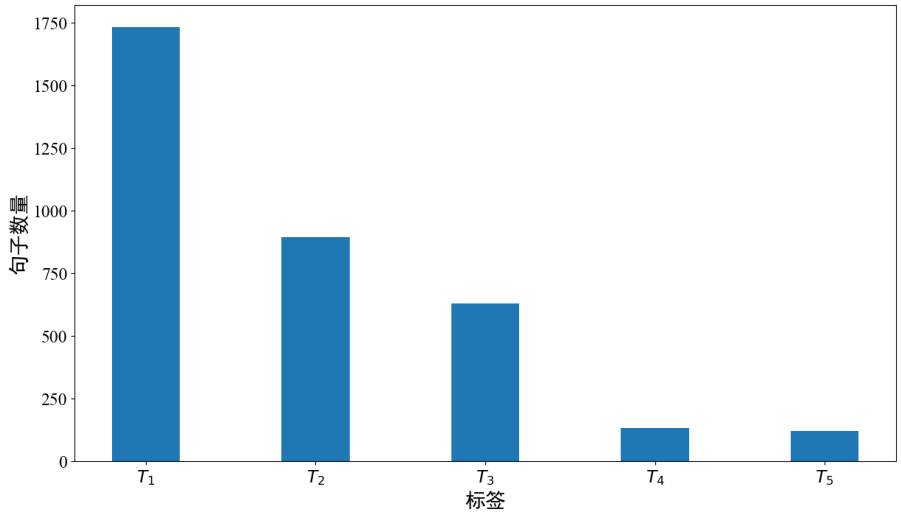

图 3-10 模型对《左传》输出结果

实验结果表明，如果选取某一时期所有古籍的部分文本作为训练集，其余文本作为测试集（实验一），该模型的实验正确率可以达到 90%。但是，若测试集的文本所属的书籍未参与模型的训练（实验二），古籍句子的正确率会降低，但综合某本古籍的所有句子来看，仍然可以通过投票原则正确判断古籍的年代。以上两个实验可以看出，同一部古籍的词汇和语法规则相对统一，同一历史时期不同古籍之间既有互相独立的一面又一定程度上统一、互相联系。

3.2.3 结论

本部分使用了 Bi-LSTM 网络模型实现了古籍断代的任务。本部分通过实验展示了不同情况下模型的预测结果对断代任务有一定的参考价值，证明了 Bi-LSTM 在文献量较少的古汉语领域也能训练出正确率不错的模型从而避免了使用 Bert 等大规模模型。因此，我们在本部分提出了一个针对古籍断代的有效模型并可以投入实用，本模型可用于辅助古籍工作人员的断代工作或者将断代标签作为输入辅助其他 NLP 任务的解决。

4 古代汉语断句模型

前面提到几种主要用到的断句方法：根据标志性的词断句，根据文言虚词断句，根据语法结构、句式结构断句，根据互文、对偶、排比等修辞手法断句，根据反复结构断句，根据古代文化常识断句，根据前后向承的句子断句，根据总分关系或分总关系断句等。由此可见句读是一种综合性很强的知识技能，在研究领域中，古汉语工作者利用他们夜以继日的调研，笔耕不辍，完成了古代汉语中大量的古籍句读工作，这给我们阅读古书带来了方便。但古代书播浩如烟海，因此研究一种针对古汉语的断句模型十分必要的。

4.1 数据来源及预处理

本部分可使用的数据来源较为开放灵活，为针对后面部分所使用的上古语料，本部分也从互联网的开放数据集上下载带有标点的上古汉语文本。文本包括《尚書》、《詩經》、《周易》、《儀禮》、《周禮》、《禮記》、《春秋公羊傳》、《左傳》、《國語》、《論語》、《孟子》、《莊子》、《呂氏春秋》、《老子》、《孝經》、《史記》、《春秋繁露》等在内的 30 余本著作，共计 300 万字。

本部分的重点是对无标点的古代汉语文本进行断句，而不过分纠结于具体标记什么标点。因此在所有的数据集的基础上，我们编写 Python 脚本将文本中的所有标点均使用“/”符号代替。转换过程如下：

表 6 原句与利用 Python 对其去标点后的实例

原句	子曰：“参乎！吾道一以贯之。”曾子曰：“唯。”子出，门人问曰：“何谓也？”曾子曰：“夫子之道，忠恕而已矣。”
处理后	子曰/参乎/吾道一以贯之/曾子曰/唯/子出/门人问曰/何谓也/曾子曰/夫子之道/忠恕而已矣/

4.2 模型构建

本部分使用的主要结构是 Bi-LSTM 神经网络。模型对输入的每一个字进行预测，预测标签有“s”、“n”两种，分别表示此处有句读(/)和此处无句读。如图 4-1 所示：

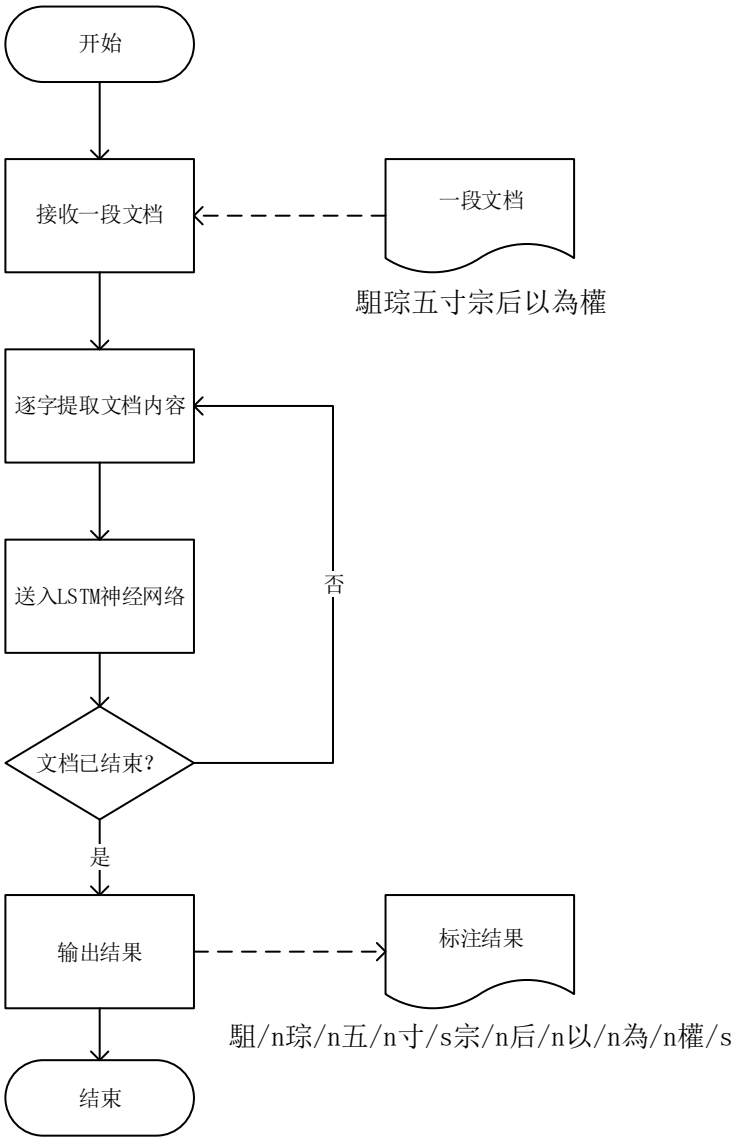

图 4-1 模型处理断句任务流程图

模型对每个字的输出和正确答案比较求出当前输出的代价 C，并利用反向传播算法调整模型当中的各种参数。

4.3 实验及效果展示

4.3.1 实验配置

模型是 Python TensorFlow-GPU 下搭建，所使用的机器的配置为：

表 7 机器的配置表

名称	配置

处理器	Intel Xeon(R) CPU ES-2620 v4@2.10GHz×16
内存	62.8GiB
显卡	GeForce GTX 1080 Ti/PCIe/SSE2×2
显卡内存	11264MB×2
系统	Ubuntu 16.04 64-bit

4.3.2 网络搭建伪代码

模型的搭建全部使用 Python 中 TensorFlow 框架实现，下面为使用 TensorFlow 框架搭建 Bi-LSTM 神经网络的伪代码。

Bi-LSTM 网络搭建

输入： X_inputs,Y_inputs,layer_num,batch_size,timestep_size,hidden_size,class_num

1：cell←调用 TensorFlow 的 rnn 类中的 LSTMCell 方法初始化一个 LSTM cell

2：inputs←对 X_inputs 调用查询表查询，查询各字的对应嵌入向量，得到嵌入向量列表

3：while layer<layer_num do

4： cell_fw，cell_bw←调用 TensorFlow 包中 rnn 类的 MultiRNNCell 方法,初始化正向 LSTM 层和反向 LSTM 层网络对象

5：end while.

6：initial_state_fw，←根据 batch_size 初始化正向和反向 LSTM 网络的初始参数

7：output_fw←[]

8：state_fw←initial_state_fw

9：for timestep←1,timestep_size do

10： output_fw,state_fw←将 inputs 输入到 cell_fw 中，返回前向网络输出和当前参数状态

11：end for

12：inputs←reverse(inputs) 将原 input 前后取反作为后向网络的输入

13：for timestep←1,timestep_size do

14： output_bw,state_bw←将 inputs 输入到 cell_bw 中，返回后向网络输出和当前参数状态

15：end for

16：Hidden_vector←concatenate(ouput_fw,output_bw)将前向输出向量和后向输出向量串联

17：全连接神经网络:

18：bilstm_output←Hidden_vector

19：W←根据 hidden_size 和 class_num 初始化全连接层的权重矩阵

20：b←根据 class_num 初始化全连接层的偏置矩阵

21：y_pred←W×bilstm_output+b 将 Bi-LSTM 层的输出送入全连接层，返回 y_pred

22：cost←调用 TensorFlow 的 reduce_mean 方法利用 corss entropy 计算模型输出的误差代价

23：optimizer←初始化一个 Adam 优化器

24：调用 Adam 优化器优化所有网络参数

4.3.3 结果展示

我们使用上古古籍的文本对模型进行训练，模型的输入是最大长度为 100 的无标点文本，模型输出每个字的断句标记。模型的正确率 R 定义：

$$正确率(R)=\frac{系统正确断句的字数量}{输入所有字总数}$$

训练过程的正确率曲线图如下图所示：

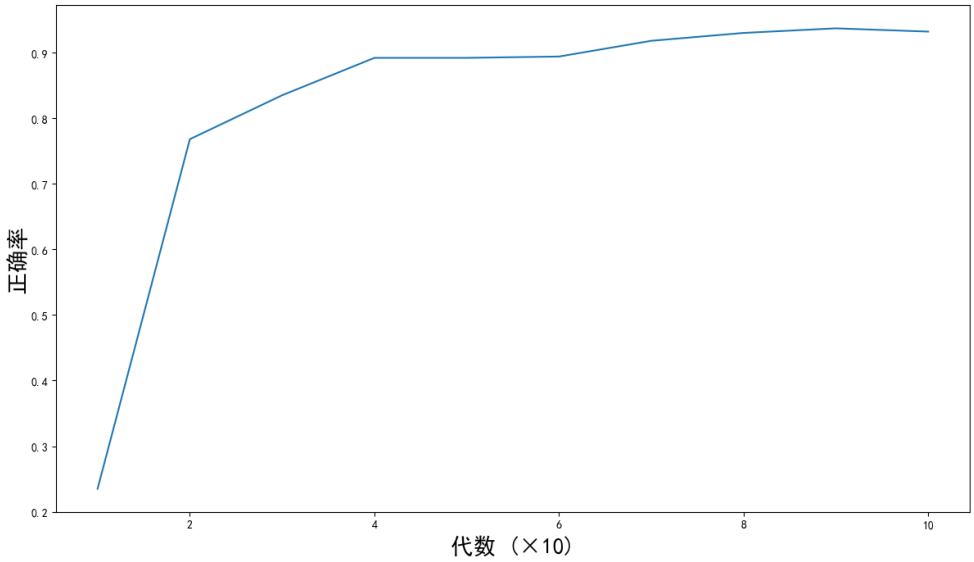

图 4-2 训练过程的断句正确率曲线图

图中可见在训练集上，模型的标注正确率可以达到 90%。训练完成之后，我们也在测试集上进行了一定的测试，模型也有不错的断句效果。部分实验结果如下所示，其中(-)表示此处应有断句，而模型模型输出有所遗漏。(+)此处不应有断句，而模型输出误将此处进行了断句。

实例一：

正确断句：

從者曰 長子近 且城厚完 襄子曰 民罷力以完之 又斃死以守之 其誰與我 從者曰 邯鄲之倉庫實 襄子曰 浚民之膏澤以實之 又因而殺之 其誰與我

模型输出：

從者曰 長子近(-)且城厚完 襄子曰 民罷力以完之 又斃死以守之 其誰與我(-)從者 (+)曰 邯鄲之倉庫實(-)襄子曰 浚民之膏澤以實之 又因而殺之

其誰與我

误分(+)：1 处　遗漏(-)：3 处

实例二：

正确断句：

故君子操權一政以立術　立官貴爵以稱之　論勞舉功以任之　則是上下之稱平　上下之稱平　則臣得盡其力　而主得專其柄　天地設　而民生之　當此之時也　民知其母而不知其父　其道親親而愛私

模型输出：

故君子操權一政以立術　立官貴爵　(+)以稱之　論勞舉功以任之　則是上下之稱平(-)上下之稱平　則臣得盡其力(-)而主得專其柄　天地設(-)而民生之當此之時也　民知其母而不知其父　其道親親而愛私

误分(+)：1 处　遗漏(-)：3 处

实例三：

正确断句：

駔琮五寸　宗后以為權　大琮十有二寸　射四寸　厚寸　是謂內鎮　宗后守之　駔琮七寸　鼻寸有半寸　天子以為權　兩圭五寸　有邸　以祀地　以旅四望

模型输出：

駔琮五寸　宗后以為權　大琮十有二寸　射四寸　厚寸　是謂內鎮(-)宗后守之　駔琮七寸　鼻寸有半寸　天子以為權　兩圭五寸　有邸(-)以祀地　以旅四望

误分(+)：0 处　遗漏(-)：2 处

实例四：

正确断句：

君仁莫不仁　君義莫不義　君正莫不正　一正君而國定矣　孟子曰　有不虞之譽　有求全之毀　孟子曰　人之易其言也　無責耳矣

模型输出：

君仁莫不仁　君義莫不義　君正莫不正　一正君而國定矣　孟子曰　有不虞之譽(-)有求全之毀(-)孟子曰　人之易其言也　無責耳矣

误分(+)：0 处　遗漏(-)：2 处

实例五：

正确断句：

陽也者稹理而堅 陰也者疏理而柔 是故以火養其陰而齊諸其陽 則轂雖敝不蔌 轂小而長則柞 大而短則摯 是故六分其輪崇 以其一為之牙圍 參分其牙圍而漆其二

模型输出：

陽也者稹理而堅(-)陰也者 (+)疏理而柔 是故以火養其陰而齊諸其陽 則轂雖敝不蔌 轂小而長則柞 大而短則摯 是故六分其輪崇 以其一為之牙圍 參分其牙圍而漆其二

误分(+)：1 处　遗漏(-)：1 处

由上面的实例可见，模型做出的断句判断整体正确，但也有遗漏和误分等情况的发生，错误主要发生的情况是遗漏问题，即某处本应有句读，而模型未在此处断句。因此从断句结果上来说，本模型的误分情况较少，当本模型在某处添加句读时的可信度较高，可当做参考，古汉语工作者在可接受的范围内，工作人员只需在模型输出的基础上查漏补缺即可。总体来说，本部分提出的一个可信度较高的古代汉语断句模型，完全可以承担辅助古汉语工作者断句的任务，为工作者提供辅助参考。

5 古代汉语分词、标注系统及数据库建设

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，中文没有显著的‘词’这个单位。所以中文分词是汉语自然文本处理的基础问题之一，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。

5.1 数据来源及预处理

由于当前尚无开放的已完成分词标注的古汉语语料库，因此，本部分使用的数据来是源于网上下载的少量开放数据集，并在后期在实验室的同学们帮助下，参考中央研究院的分词标准对部分语料进行手工标记所得。数据内容包括《尚书》、《礼记》、《论语》等多本上古古籍的部分章节。调查数据包含超过 10,000 条语句，平均每条语句包含 9 个字。由于本部分所使用的语料库较小，无法对模型进行充分的训练，因此本部分的重点为提出一种可同时用于分词及词性标注的模型并证明该模型具有一定的实用性，后序继续可以使用该模型对预料进行大致的分词及词性标注，所产生的带有分词标签及词性标签的语料在经过人工确认后又可继续用来训练模型以提高模型的对上古汉语分词及词性标注任务的拟合度。

5.1.1 标签编码

首先，我们对语料标签进行标签转换，将语料的分词标记和词性标记转成统一的二元标签结构，如 3.1（a-c）式所示，T_d 为词性标签列表，其中元素 d_1、d_2 等分别表示不同的词性标记。T_c 为分词标签列表，其中的 c_1、c_2 等表示不同的分词标记，携带不同的分词信息。而 T 则表示二元标签组，其中的每一个标签是由 T_d 的转置点乘 T_c 得到的矩阵，则 T 中的每一个标签都分别携带了词性标记信息和分词标记信息式。

$$S = \{x_1, x_2, x_3, ..., x_n\} \tag{5-1}$$

$$T_d = [d_1, d_2, d_3, ..., d_m] \tag{5-2}$$

$$T_c = [c_1, c_2, c_3, ..., c_n] \tag{5-3}$$

$$T \in (T_d^{\mathrm{T}} \cdot T_c)=\begin{Bmatrix} d_1c_1 & d_1c_2 & \ldots & d_1c_n \\ d_2c_1 & d2c_2 & \ldots & d_2c_n \\ \ldots & \ldots & \ldots & \ldots \\ d_mc_1 & d_mc_2 & \ldots & d_mc_n \end{Bmatrix} \tag{5-4}$$

预处理过程如图 5-1 所示，首先将词W_n拆分成单字Z_n，并且单字Z_n也携带该字隶属词的词性标记信息d_m，例如将“瀑布/N”转换为‘瀑/N’和‘布/N’，然后使用分词标签为每个字打上分词标记，即将中文分词当成分类任务[34]。最后将词性标签作为第一维标签，分词信息标签作为第二维标签组合成二元标签组的结构。

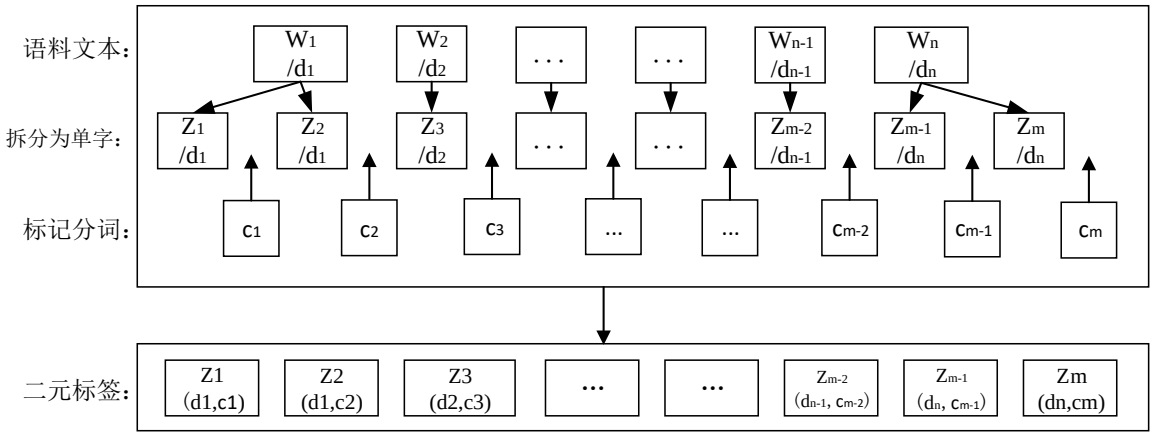

图 5-1 文本预处理过程示意图

本模型的输出不同于分词或词性标记单任务模型的单标签输出，本模型输出的是经过编码的二元标签组，标签的两个维度分别表示词性标记和分词信息，可以使网络更加充分考虑字、词性和分词之间的内在关联，相较先分词后进行词性标注这样的“两步走”方法增加了词性和分词之间的内在关联，使一体化模型的精确度更高。

针对一体化模型中的输入、输出层，我们提出两种二元标签结构的编码方式，如图 5-2 和 5-3 所示，若二元标签组包括 m 种词性标记，用d_m表示，n 种分词标记，用c_n表示，则编码方式一为：对(m×n)种不同词性标记和分词标记的自由组合结果进行编号，即每一个不同的二元标签有自己固定的从 0 至(m×n)的某一编号，然后对其进行 One-hot 编码，即一串(m×n)位的 0/1 序列，其中除某一位为 1 以外，其他均为 0，为 1 的位对应的编号即为对应的二元标签，如图 5-2 所示。编码方式二：对(m×n)种二元标签进行二维编码，第一维有 m 列，表示词性标注信息，第二维有 n 列，表示分词信息，则二元标签组的编码为一串(m×n)位的 0/1 序列，其中前 m 位中仅有一位为 1，

对应词性标签的 One-hot 编码，后 n 位有一位为 1，对应分词标签的 One-hot 编码。如图 5-3 所示。这样二元标签组就表示成了一串计算机能处理的 0/1 序列。

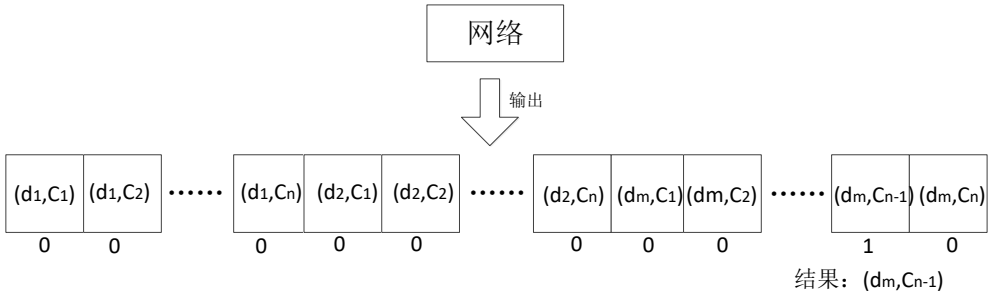

图 5-2 编码方式一示意图

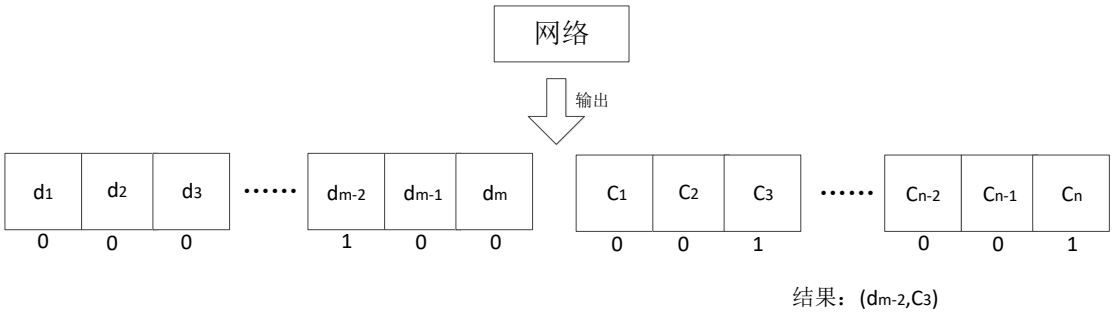

图 5-3 编码方式二示意图

5.2 分类模型的评估标准

准确率和召回率是广泛用于信息检索和统计学分类领域的两个度量值，用来评价结果的质量。其中精度是检索出相关文档数与检索出的文档总数的比率，衡量的是检索系统的查准率；召回率是指检索出的相关文档数和文档库中所有的相关文档数的比率，衡量的是检索系统的查全率。一般来说，Precision 就是检索出来的条目（比如：文档、网页等）有多少是准确的，Recall 就是所有准确的条目有多少被检索出来了。

$$召回率\left(Recall\right)=\frac{系统判断正确的数量}{系统所有正确总数} \tag{5-5}$$

$$准确率\left(Precision\right)=\frac{系统判断正确的数量}{系统判断正确的数量+系统判断错误的数量} \tag{5-6}$$

准确率和召回率是互相影响的，理想情况下肯定是做到两者都高，但是一般情况下准确率高、召回率就低，召回率低、准确率高，当然如果两者都

低，那是什么地方出问题了。通常，我们希望准确率和召回率均越高越好，但事实上这两者在某些情况下是矛盾的。

比如我们只搜出了一个结果，此结果是正确的，求得 precisin 等于 1。但是由于只搜出一个结果，recall 值反而很低，接近于 0。所以需要综合考量，因此便引入了 F-measure。F-measure 又称 F-score，其公式为：

$$F_{\beta}=(1+\beta^{2})\cdot\frac{precision\cdot recall}{(\beta^{2}\cdot precision)+recall} \tag{5-7}$$

其中，当 β=2 时，更加注重召回率；β=0.5 时，更加重视准确率；当 β=1 时，就是 F1-score：

$$F_{1}=2\cdot\frac{precision\cdot recall}{precision+recall} \tag{5-8}$$

F-measure 综合了 precision 和 recall，其值越高，通常表示算法性能越好，下表展示了召回率、准确率和 F-score 之间的关系。

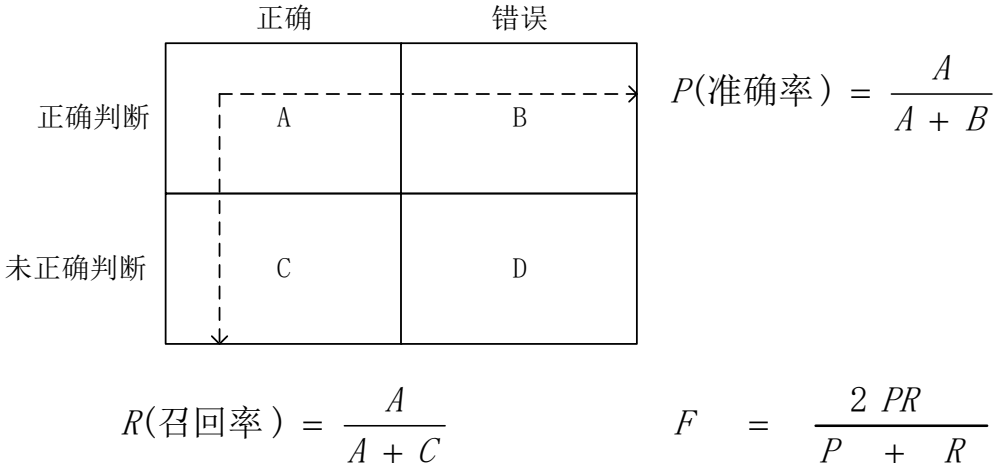

图 5-4 两类分类的混淆矩阵示意图

5.3 模型架构

本部分将介绍基于 Bi-LSTM 的古汉语自动分词及词性标注一体化模型的相关细节，图 5-5 为模型的训练流程图。其中，数据预处理是指将原预料库中 W/d 即词/词性标记形式的内容处理为模型所需要的 $Z/(d_m,c_n)$ 即单字/二元标签组形式的内容。编码指将计算机不能理解的二元标签组转变成计算机可以处理的数字序列，以便送入神经网络进行计算。本节将包括语料库数据的预处理，网络结构和标签编码三个部分。

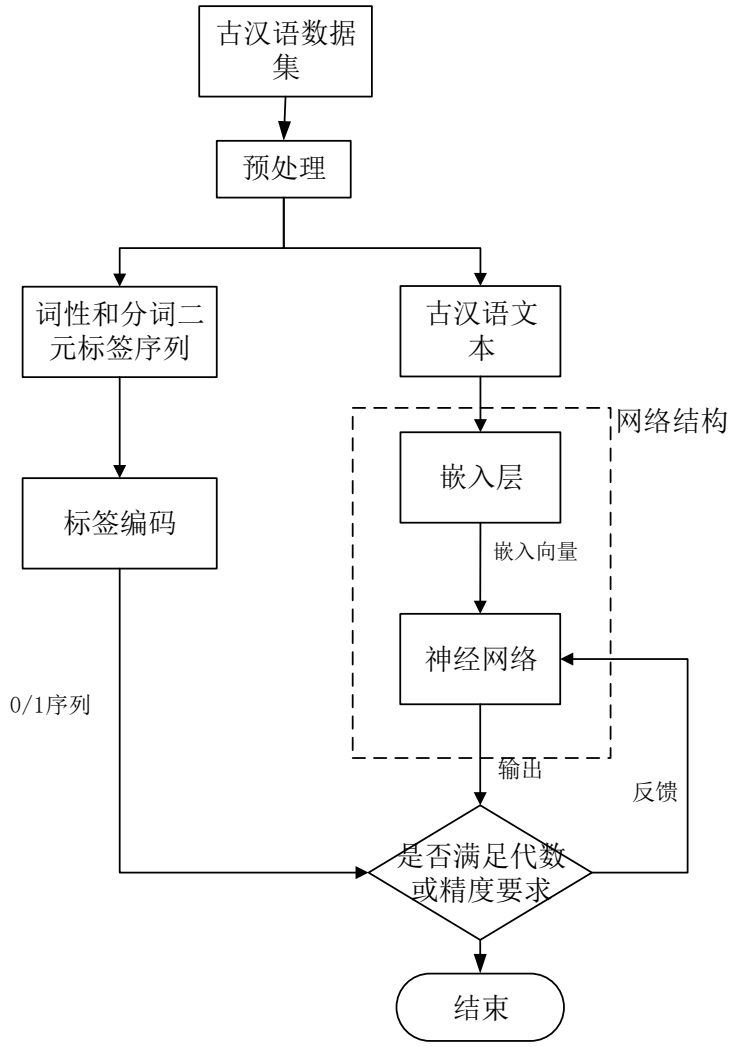

图 5-5 模型的训练流程图

5.3.1 网络结构

本模型利用 Bi-LSTM 网络进行训练。模型在获取一段文本作为输入后，首先将该句中所有字转化成字向量[35][36]，然后送入 Bi-LSTM 网络进行训练。图 5-6 是基于 Bi-LSTM 的自动分词及词性标注一体化模型总体示意图。

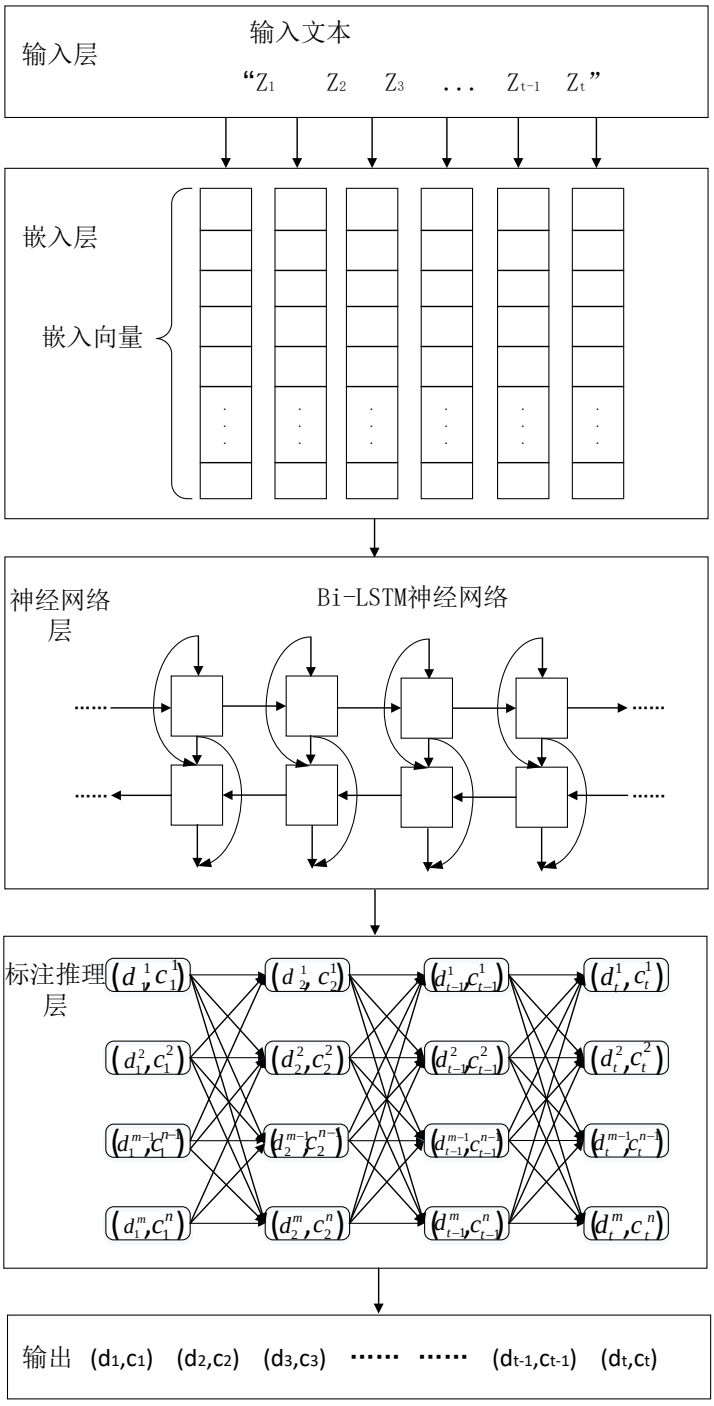

图 5-6 基于 Bi-LSTM 的自动分词及词性标注模型结构示意图

使用神经网络处理文字数据的第一步是将它们表示为分布式向量，也称为嵌入向量[37]。前面部分已经介绍过，在中文自然语言处理过程中有一个从训练集中提取出的大小为$|C|$的字符字典C，每个字符$c \in C$表示为实值向量（字符嵌入）$v_c \in R^d$其中 d 是向量空间的维数。然后将所有字符嵌入堆叠成嵌入矩阵$M \in R^{d \times |C|}$。对于字符$c \in C$，查找表 C 检索相应的字符嵌入$v_c \in R^d$。

查找表和嵌入层之间可以被视为简单的投影层，每个字符嵌入通过其查找表索引到相应的列操作来实现，然后将字符向量送入到相应的神经网络中。

在 NLP（Natural Language Processing）任务中常用循环神经网络（Recurrent Neural Network，RNN）和长短期记忆神经网络（Long Short-Term Memory，LSTM），LSTM 是递归神经网络 RNN 的扩展。RNN 每次的输出取决于前一次的输出。对于给定序列 $x(1:\ n)=(x(1),x(2),...,x(t),...,x(n))$，RNN 通过 $h(t)=g(Uh^{(t-1)}+Wx^{(t)}+b)$ 更新其隐藏状态，其中 g 是一个非线性函数。虽然已经证明 RNN 在语音识别，语言建模和文本生成等许多任务上都很成功，但是，它学习过程中的梯度消失和爆炸问题会影响训练过程中长期依存信息的学习。

LSTM 通过其特殊的具有门结构的神经元解决了传统 RNN 网络中梯度消失和弥散的问题。LSTM 的基本胞元行为都是由三个“门”控制，如图 5-7，即输入门 i，遗忘门 f 和输出门 o。门上的操作被定义为逐位相乘，因此如果门是非零矢量，则门可以缩放相应的输入值作为输入；如果门是零矢量，则门可以将输入清零。T 时刻输出门的输出将被馈送到下一时刻即 t+1 时刻的神经网络作为输入。

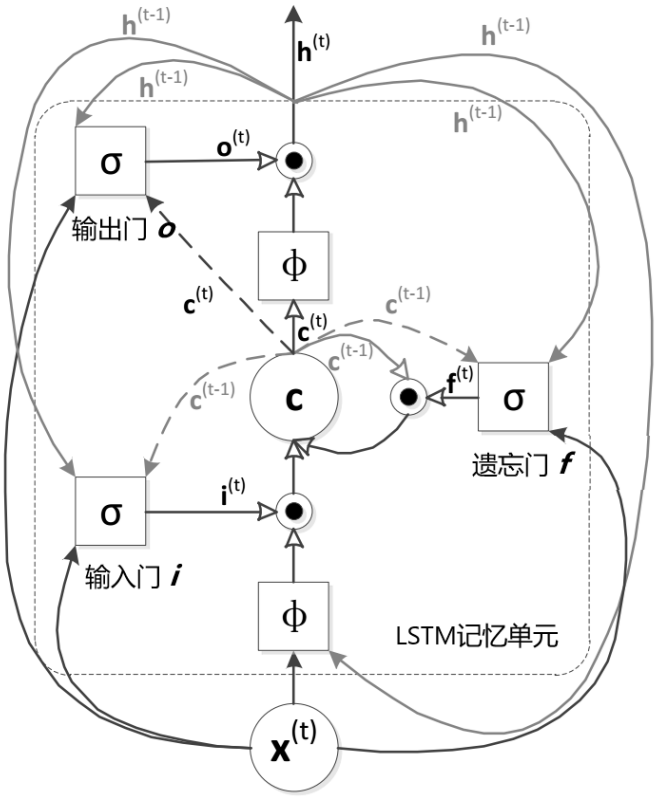

图 5-7 长短期记忆神经网络记忆单元

（1）嵌入层解决的是古汉语自然语言的数字化表示并嵌入到模型的问题。古汉语是一种自然语言无法被计算机直接处理，因此古汉语语句在通过神经网络模型进行处理之前必须首先将其数字化。

（2）模型神经网络层是古汉语分词与词性标注一体化模型的核心，是采用深度学习方法进行古汉语分词特征学习与古汉语词性标注特征学习的重要的中间层，本部分模型网络层采用 Bi-LSTM 神经网络，其最主要的作用是针对输入的字向量数据，在训练过程通过不断调整神经网络参数来进行高度抽象的古汉语特征的学习，使之自动提取出字、分词、词性标注和上下文之间的内在关联。

（3）标注推理层本质上是一个隐马尔科夫过程，它可以在句子层级上利用上下文标注信息推断当前字对应的词位标记或当前词对应的词性标记。古汉语所具有的语法关系会影响标记，标注推理层通过标记之间前后的互相影响计算一个句子的最优标注分数。

5.4 实验及性能分析

本模型通过 Python 下的 TensorFlow 框架搭建，使用的数据包括《尚书》、《礼记》、《诗经》、《论语》等含有分词和词性标记的上古书籍文本。该语料库中使用普及化标记标记上古汉语文本，且所有语料文本均使用 Unicode 编码。实验选取上古语料中 75%文本作为训练集对模型进行训练，25%作为测试集进行后期实验性能分析和比较。训练过程中利用 Adam(Adaptive Moment Estimation)方法对参数进行调节，该方法能分别计算每个参数的自适应学习率，可使模型参数达到较好的效果。图 5-8 是训练过程中二元标签组的两种编码结构在不同步长情况下的精确度随代数变化的比较图。图中可见，第一种编码方式的情况下，学习率为 1e-2 的时候上升速度最快，但精确度最高稳定在 92.7%；1e-3 的学习率曲线上升速度虽然不如 1e-2 快，但能达到更高的精确度 97.7%；1e-4 学习率情况下由于学习率过小，正确率曲线在 300 批次之后上升十分缓慢，最终只可达到 91.8%。与此同时，编码方式二的情况下，正确率曲线在不同步长参数下的变化趋势基本相同，但在三种步长下结构二所能达到的最优正确率仅为 1e-3 情况下的 96.8%，达不到方式一中 1e-3 时的 97.7%正确率。造成这种情况的原因，一方面可能因为编码方式二是分前后两段的，则分词和词性标注的过程独立性相较方式一更高，不能充分的利用到分词标志和词性标志之间的内在联系，另一方面可能是因为参数未设置到方式二的最佳状态。在下面的实验中，我们均选择 1e-3 步长下利用方式一编码训练得到的模型参数进行实验，以获得最好的实验效果。

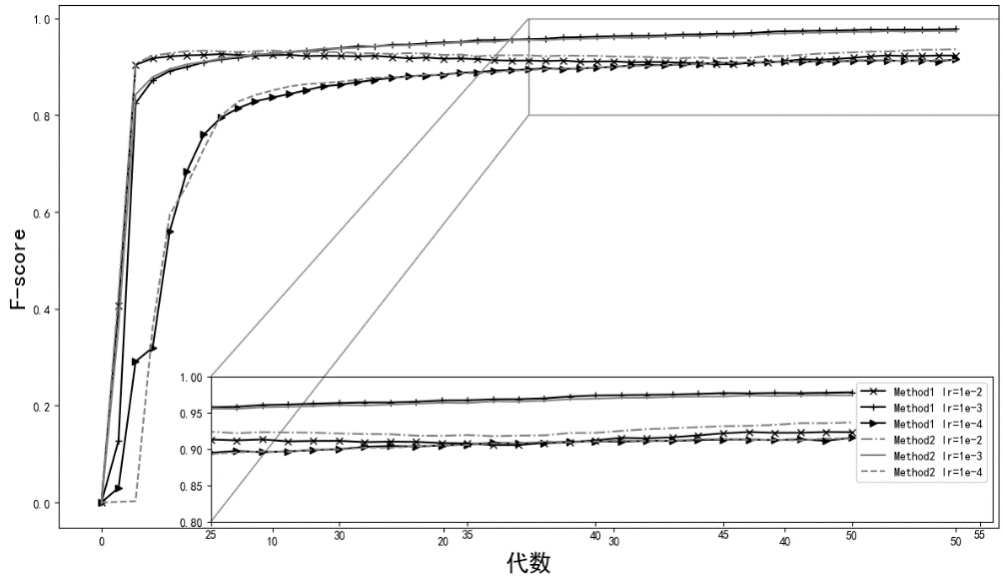

图 5-8 两种输出结构下不同学习率下正确率随批次增加的变化示意图

本部分将本文提出的方法同基于字典的分词方法以及基于马尔科夫链的标注方法分别进行性能上的比较。该部分使用 F 值（F-measure）作为主要指标评估标准对不同方法的分词标注结果进行统一衡量。

$$R = \frac{c}{N} \tag{5-9}$$

$$P = \frac{c}{c+e} \tag{5-10}$$

$$F = \frac{2 \times P \times R}{P+R} \tag{5-11}$$

其中 N 为标准分割的单词数，e 为错误标注的单词数，c 为分词器正确标注的单词数，R 为召回率，P 为精度。一个完美的分词器其 F 值为 1，F 值越大说明分词器分词标注效果越好。

5.4.1 分词

字典匹配法中使用的字典是包含上古语料库中提取出的包含所有词语的理想词典，即语料库中所有的词均包含在字典中，上古汉语中 70%以上都是单字成词，对此本文针对性的使用最大长度为 4 的基于字典的分词方法。图 5-9 为在不同字典大小情况下字典法 F 值的比较图，可见字典的大小对分词精确度有直接影响，这里利用无未登录词的理想字典同本文模型进行比较，图中可见理想字典匹配法 F 值可达 0.94。

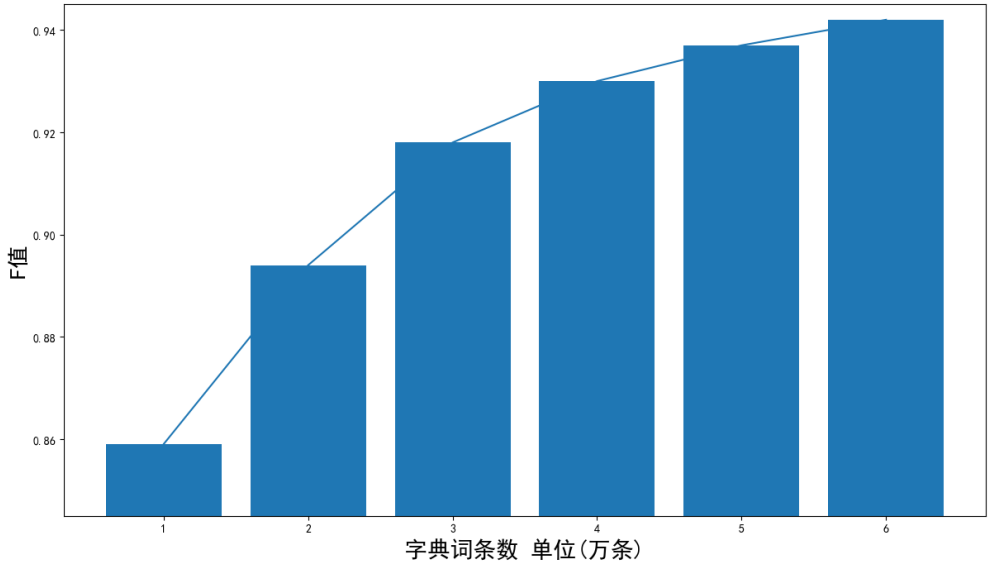

图 5-9 不同大小字典分词 F 值比较图

图 5-10 是不同句长的句子个数分布图，图 5-11 是利用上古预料库 80%语料训练的 Bi-LSTM 一体化模型并利用 25%作为测试集进行的测试结果条形图。图 5-11 中可以看出，句子长度大多分布在 4 字到 15 字之间。实验中一体化模型的测试集不包含于训练集当中，但利用训练集训练得到的模型对测试集的分词及标签 F 值仍然可以达到 0.97，效果优于理想字典匹配法的 0.94 的单任务自动分词 F 值。图 5-10 显示在句长小于 5 的情况下，标签的正确率稍低，原因是句长过短，LSTM 网络从上下文中提取的信息有限，影响了标签正确率。虽然句长小于 10 的情况下分词 F 值略低，但图 5-11 所展示的句长分布可见包含五个字以下的句子在上古文本中所占比例并不大，所以对于短句对于正确率的影响也可控制在可接受的范围之内。

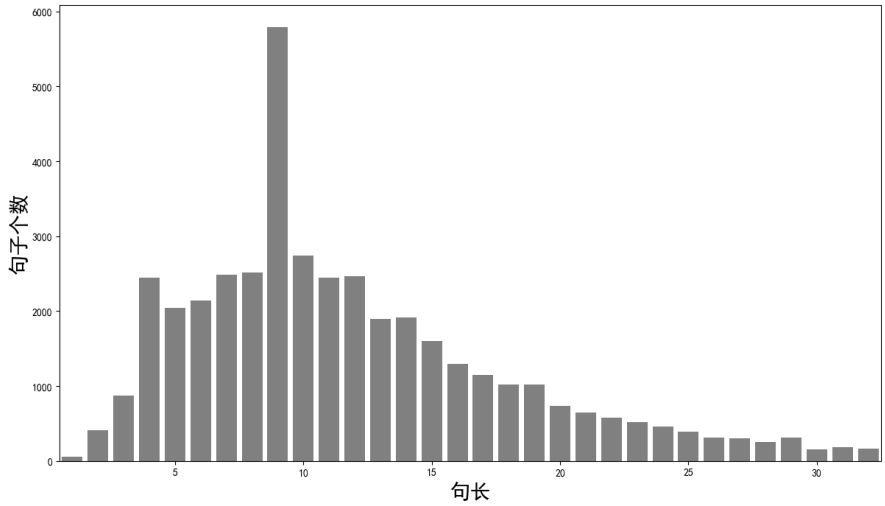

图 5-10 句长分布图

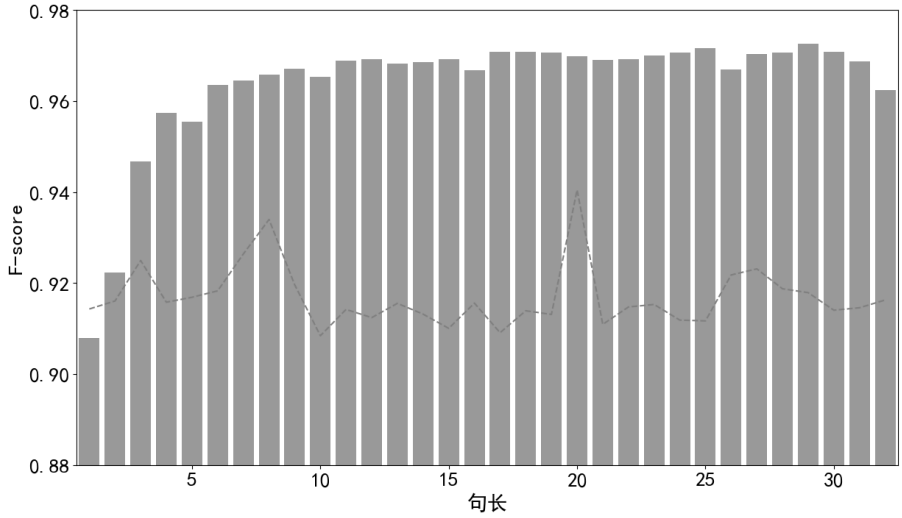

图 5-11 Bi-LSTM 一体化模型在不同句长下 F 值

5.5 词性标注

本部分我们将一体化模型同隐马尔科夫链模型（Hidden Markov Model）进行词性标注上的比较，这里使用的 HMM 模型使用的参数是通过统计所有上古语料库文本得到的。HMM 模型在已经分好词的现代汉语上进行词性标记可达到 95%的准确率，而应用到上古汉语中则效果较差，如图 5-12 中绿色条形图所示，在不同句子长度的情况下，其 F 值仅 0.89 左右，图中蓝色条形

图为本一体化模型在不同句长下的词性标注 F 值情况。可以看出现代汉语与古代汉语在句长分布，一词多义等方面的不同对 HMM 方法具有一定的消极影响。而本文的一体化模型是基于上下文信息以及字词内部深层次的关联来输出标签，对上古汉语具有较好的适应性。

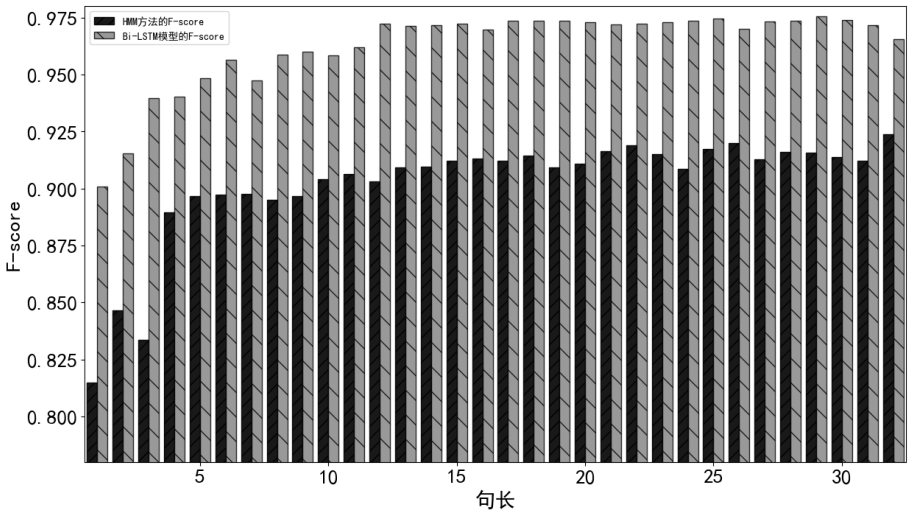

图 5-12 HMM 方法与一体化模型在不同句长下的词性标注 F 值比较图

通过上述实验可得，本文提出的一体化模型在分词效果上比基于理想字典的匹配法正确率高，在词性标注上效果优于利用所有语料进行参数设置的 HMM 法。查阅文献可知，当前对于古汉语来说，分词所能达到的最佳性能为文献[38]中达到的 97.7%正确率，词性标注目前能达到的最佳性能为文献[39]中达到的 97%正确率。如果将以上两种分词和词性标注单任务领域的最佳方法串联在一起，则其正确率将会变为二者正确率的乘积，又将有一定程度上的损失而无法达到 97%。而本文的一体化模型从最开始就将分词和词性标注任务一同考虑，不存在错误扩散的问题，结果更加精确。

5.5.1 结果分析

正如古汉语计算语言学家尉迟治平的呼吁:“我们期望能有可以用于汉语史电子文献自动分词、自动断句、自动标注的软件早日问世，专家只需对结果刊谬补缺，这将大大减轻属性式标注的劳动强度，加快工作进度[40]。”

本模型不仅可输出可信度最大的二元标签，而且可输出第二大、第三大等标签的可信度作为辅助供使用者对结果刊谬补缺。表 8 是通过一体化模型对句子进行自动分词及词性标注的一个实例，其中 word 是输入句子的单字，

y_right 是输入汉字的正确二元标签，y_pred 为本模型的输出二元标签，Same 表示 y_right 与 y_pred 是否一致，‘1’表示在该位置 y_right 与 y_pred 标签相同，‘0’则表示二者不同。

表 8 《睡虎地秦墓竹簡・为吏之道》实例

输入	民	心	將	移	乃	難	親
正确标签	(N,n)	(N,s)	(ADV,s)	(Vt,s)	(ADV,s)	(Vt,s)	(VI,s)
预测标签	(N,n)	(N,s)	(ADV,s)	(Vt,s)	(N,n)	(N,s)	(VI,s)
是否相同	1	1	1	1	0	0	1

如图 5-13 展示了每个字前三名标签的分数，其中绿色条形对应的分数为正确标注，可以看出前 4 个字最大与次大标签评分差距较大，且评分最大标签均为正确标签。而第五、第六个字最大与次大标签评分差距较小，评分差距较小是程序对该字判断模糊的表现，该两字的正确答案均为程序判断的第二可能标签。所以增加前三名标签作为辅助可以覆盖更多的正确标签，这一特性将为古汉语研究者接下来的编辑和校对提供便利。

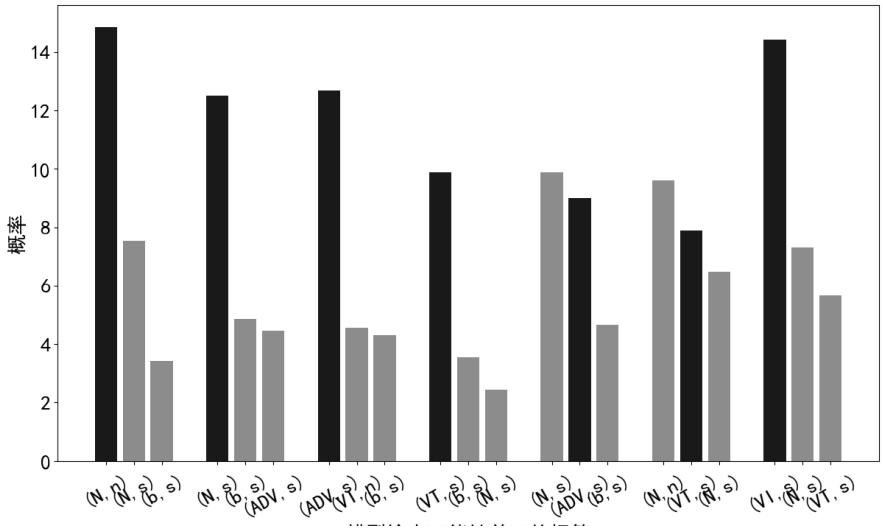

图 5-13 实例中每个字的前三名评分条形图

5.5.2 本章小结

古汉语自动分词及词性标注是古汉语研究电子化的基础，在中文研究领域逐渐受到重视。由于古汉语与现代汉语的一些不同，直接使用现代汉语的相关技术有一定的困难。本文针对古汉语自动分词及词性标注技术：一、提出一种二元组标签；二、提出一种自动分词及词性标注的一体化设计方案，

同时考虑分词和标注；三、将 Bi-LSTM 用到了古汉语分词标注领域，有一定的创新性。实验效果表明，本模型的自动分词及词性标注正确率可以达到 97%，较现有的 HMM 和字典法有一定的提升，且优于当前性能最佳的分词和词性标注方法串联后得到的处理结果。但是本文的工作还比较初步，预计接下来将继续完善该模型以下不足：对数据进行预处理，利用其他数据库将语料库中的误码、错码补全；使用更加细致的上古词类标记，而不使用简化版的普及化标记；后期可尝试加入预先训练的字嵌入向量，已有实验表明，此步骤可提高 NLP 任务的性能，我们将对该结论在古汉语领域继续加以验证；继续寻找更优的参数，例如字符嵌入时嵌入向量的维度，梯度下降时的步长等。

6 总结与展望

6.1 总结

基于我国现阶段古代汉语研究和现代机器学习方法结合相对欠缺的现象，本文针对古代汉语提出了的三项基本目标：实现自动古籍断代、实现古代文本自动切分、实现古代文本自动标注以及实现古籍的结构化入库。为了实现古代汉语信息化以及古代汉语的现代化研究，本文将深度学习与古汉语研究的基本任务相结合，利用 Bi-LSTM 深度神经网络完成古汉语任务，并利用得到的模型处理古籍文本，最终将古籍文本进行切分及标注，并入库管理。

本文的贡献可总结为以下几个方面：

（1）为解决古代书籍断代的问题，本文提出使用双向长短期记忆神经网络作为主体构建古代文本断代模型。整理互联网上现有的已知年代的文本作为训练集对模型进行训练。利用 word2vec 模型将文本中的每一个字被转换成一串高维向量，然后将文本包含的所有文字的字向量送入模型分析它们之间的非线性关系。最终，模型会输出一个该段文本的年代类别标签。实验结果表明利用 Bi-LSTM 神经网络构造的模型能够很好的完成断代任务，断代的正确率能达到 80%以上。本文的断代模型提供了一种高效且准确的古文断代方法，这将节省古文研究工作者在文本断代过程中的时间消耗。

（2）针对某些古代汉语书籍原著中缺少标点符号的问题，本文提出一个断句模型。本部分我们通过深度神经网络对大量经过断句的古汉语文本进行学习，使断句模型自动学习到某一时期、某种题材的断句规则，从而实现输入一段无断句的文字序列，机器自动为其添加断句的效果。

（3）针对古汉语分词及词性标注任务，我们需要解决训练集获取的问题，分词标注任务需要已经分好词、标注好词性的文本来做模型的训练集，但目前目前尚没有公开的具有分词和词性标注的古汉语语料库。因此我们通过手工标记部分语料的方法得到了少量的数据集对我们所设计的分词标注模型进行少量的实验，用以验证本文提出的分词标注模型可以较好的完成古汉语分词标注任务。

6.2 展望

从模型数据角度来说，现有的古代汉语预料主要是未进行分词及词性标注的，本文中所有使用到的分词及词性标注训练集均为人工标注的，因此数

据量较小，训练所得的模型正确率也有限。下一步研究工作将考虑结合机器学习和人工纠错，不断扩充古汉语语料库，不断循环迭代，模型和语料库互相促进，提高模型性能以及语料库容量。

从模型算法角度来说，由于数据量较小的关系，现有模型均基于Bi-LSTM神经网络，下一步研究工作可以采用其它大规模自然语言处理模型进行建模，可考虑使用现代汉语模型加在古汉语数据集上的 fine-tune 的方法训练模型。

参考文献

[1] Torre LA， Bray F， Siegel RL， et al. Global cancer statistics，2012. CA Cancer J Clin. 2015， 65(2)：87-108.J. Clerk Maxwell， A Treatise on Electricity and Magnetism，3rd ed.，vol. 2. Oxford：Clarendon，1892，pp.68–73.

[2] Desantis C E， Fedewa S A， Mpsh A G S， et al. Breast cancer statistics， 2015： Convergence of incidence rates between black and white women[J]. Ca A Cancer Journal for Clinicians， 2015， 66(1)：31–42.

[3] Chen W， Zheng R， Baade P D， et al. Cancer statistics in China， 2015[J]. Ca A Cancer Journal for Clinicians， 2016， 66(2)：115–132.

[4] Gail M H， Brinton L A， Byar D P， et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually.[J]. Jnci Journal of the National Cancer Institute， 1990， 81(24)：1879-86.

[5] Gustafsso K， Aronsson G， Marklund S， et al. The World Bank. Int J Behav Med. 2014. 21(1)：77-87.

[6] Gustafsso K， Aronsson G， Marklund S， et al. Population and social conditions. Int J Behav Med. 2014. 21(1)： 77-87.

[7] Liyuan L. A pilot study on risk factors and risk assessment score screening model for high-risk population of breast cancer[D]. Shandong University， 2010：16.

[8] He H， Garcia E A. Learning from Imbalanced Data[J]. IEEE Transactions on Knowledge & Data Engineering，2009，21(9)：1263-1284.

[9] Demuth H B， Beale M H， De Jess O， et al. Neural Network Design[C]// Wseas International Conference on Circuits. Martin Hagan，2014.

[10] Lei L， Yang D. A GA-based feature selection and parameters optimization for support vector regression.[C]// International Conference on Natural Computation， Icnc 2011， Shanghai， China， 26-28 July. 2011：335-339.

[11] Kerr N L， Maccoun R J， Kramer G P. Bias in judgment：Comparing individuals and groups[J]. Psychological Review，1996，103(4)：

687--719.

[12] Herman WH，Smith PJ，Thompson TL，et a1.A new and simple question naire to identify rmple at increased risk for undlagnosed diabetes[J].Diabetes Care，1995，18(3)：382-387．

[13] Ahmed M H， Abdu T A M. Logistic Regression Analysis[M]. Springer Netherlands， 2008.

[14] Ruige JB，Neeling JN，Kostense PJ，et a1.Pefronnanee of an NIDDM screning questionnaire based on symptoms and risk factors[J].Diabetes Care.1997.20(4)：491-496．

[15] Jaana LD，PiIjo IP，Markku P，et al.Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention：follow-up of the Finnish Diabetes Prevention Study[J].Lancet，2006，368(9548)：1673-1679．

[16] 郑莹，吴春晓，张敏璐. 乳腺癌在中国的流行状况和疾病特征[J]. 中国癌症杂志， 2013， 23(8)：561-569.

[17] 赵洁玉，徐卫云. 乳腺癌风险评估及预测模型的研究进展[J]. 临床外科杂志， 2013， 21(7)：566-568.

[18] Adamscampbell L L， Makambi K H， Palmer J R， et al. Diagnostic accuracy of the Gail model in the Black Women's Health Study.[J]. The Breast Journal， 2007， 13(4)：329–331.

[19] Wen Y C. Validation of the Gail model for predicting individual breast cancer risk in a prospective nationwide study of 28，104 Singapore women[J]. Breast Cancer Research， 2013， 14(1)：1-12.

[20] Claus E B， Risch N， Thompson W D. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction[J]. Cancer， 1994， 73(3)：643-51

[21] van Asperen C J， Jonker M A， Jacobi C E， et al. Risk estimation for healthy women from breast cancer families： new insights and new strategies.[J]. Cancer Epidemiology Biomarkers & Prevention， 2004，13(1)：87-93.

[22] Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors[J]. Statistics in Medicine， 2012， 10(23)：1111-30.

[23] Jacobi C E， Bock G H D， Siegerink B， et al. Differences and similarities in breast cancer risk assessment models in clinical practice：which model to choose?[J]. Breast Cancer Research and Treatment， 2009，

115(2)：381-90.

[24] Hall M， Reid J L， Pruss D， et al. BRCA1 and BRCA2 mutations in women of different ethnicities undergoing testing for hereditary breast-ovarian cancer.[J]. Cancer， 2009， 115(10)：2222–2233.

[25] Kwong A， Wong C H， Suen D T， et al. Accuracy of BRCA1/2 mutation prediction models for different ethnicities and genders： experience in a southern Chinese cohort.[J]. World Journal of Surgery， 2012， 36(4)：702-13.

[26] Liyuan L. A pilot study on risk factors and risk assessment score screening model for high-risk population of breast cancer[D]. Shandong University， 2010：16.

[27] D. Hand， H. Mannila and P. Smyth， 2001.“Principles of data mining”， MIT.

[28] Chawla N V， Japkowicz N， Kotcz A. Editorial： special issue on learning from imbalanced data sets[J]. Acm Sigkdd Explorations Newsletter， 2004， 6(1)：1-6

[29] Maloof M A. Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown[J]. ICML-2003 Workshop on Learning from Imbalanced Data Sets II， 2010.

[30] Kubat M. Addressing the Curse of Imbalanced Training Sets：One-Sided Sampling[C]// Proceedings of the International Conference on Machine Learning (ICML-97. 1997.

[31] Chawla N V， Bowyer K W， Hall L O， et al. SMOTE：synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research， 2011， 16(1)：321-357.

[32] 古平，欧阳源遊. 基于混合采样的非平衡数据集分类研究[J]. 计算机应用研究， 2015(2)：379-381.

[33] Tomar D， Agarwal S. Prediction of defective software modules using class imbalance learning[J]. Applied Computational Intelligence & Soft Computing， 2016， 2016(1)：1-12

[34] Erfani S M， Rajasegarar S， Karunasekera S， et al. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning[J]. Pattern Recognition， 2016， 58：121-134

[35] PATRICIA RIDDLE， RICHARD SEGAL， OREN ETZIONI. REPRESENTATION DESIGN AND BRUTE-FORCE INDUCTION IN A

BOEING MANUFACTURING DOMAIN[J]. Applied Artificial Intelligence，1994， 8(1)：125-147.

[36] Kubat M， Holte R， Matwin S. Learning When Negative Examples Abound[C]// European Conference on Machine Learning. Springer-Verlag， 2000：146-153.

[37] Weiss G M. Mining with rarity： a unifying framework[J]. Acm Sigkdd Explorations Newsletter， 2004， 6(1)：7-19.

[38] Wu G， Chang E Y. Class-boundary alignment for imbalanced dataset learning[J]. Icml Workshop on Learning from Imbalanced Data Sets，2003：49--56.

[39] Barandela R， Sánchez J S， GarcíA V， et al. Strategies for learning in class imbalance problems ☆[J]. Pattern Recognition， 2003，36(3)：849-851.

[40] Joshi M V， Kumar V， Agarwal R. Evaluating Boosting Algorithms to Classify Rare Classes： Comparison and Improvements[M]// Evaluating boosting algorithms to classify rare classes： comparison and improvements. 2001：257-264.

[41] Kiryu T. Introduction to Data Mining[J]. 2010， volume 16(472)：127-130(4).

[42] Tahani H， Keller J M. Information fusion in computer vision using the fuzzy integral[J]. IEEE Transactions on Systems Man & Cybernetics， 1990， 20(3)：733-741.

[43] Keller J M， Gader P， Tahani H， et al. Advances in fuzzy integration for pattern recognition[J]. Fuzzy Sets & Systems， 1994， 65(2-3)：273-283.

[44] K. J. Friston， C. D. Frith， P. F. Liddle，等. Functional Connectivity： The Principal-Component Analysis of Large (PET) Data Sets[J]. Journal of Cerebral Blood Flow & Metabolism Official Journal of the International Society of Cerebral Blood Flow & Metabolism，1993，13(1)：5-14.

[45] 西奥多里蒂斯[希腊]. 模式识别： 第 2 版[M]. 电子工业出版社， 2004.

[46] Liu H， Motoda H， Setiono R， et al. Feature Selection： An Ever Evolving Frontier in Data Mining.[J]. 2010， 10：4-13.

[47] 王娟， 慈林林， 姚康泽. 特征选择方法综述[J]. 计算机工程

与科学， 2005， 27(12)：68-71.

[48] Chandrashekar G， Sahin F. A survey on feature selection methods [J]. Computers & Electrical Engineering， 2014， 40(1)：16-28.

[49] Selima S Z， Alsultanb K. A simulated annealing algorithm for the clustering[J]. Pattern Recognition， 1991， 24(10)：1003-1008.

[50] Trelea I C. The particle swarm optimization algorithm：convergence analysis and parameter selection[J]. Information Processing Letters， 2003， 85(6)：317-325.

[51] Kononenko I， Kononenko I. Analysis and extension of RELIEF[C]// The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. 1994.

[52] Xu L， Yan P， Chang T. Best first strategy for feature selection[C]// International Conference on Pattern Recognition. IEEE， 1988：706-708 vol.2.

[53] 姚旭， 王晓丹， 张玉玺，等. 特征选择方法综述[J]. 控制与决策， 2012， 27(2)：161-166.

[54] 徐燕， 李锦涛， 王斌，等. 基于区分类别能力的高性能特征选择方法[J]. 软件学报， 2008， 19(1)：82-89.

[55] Jain A K，Duin R P W，Mao J. Statistical Pattern Recognition：A Review[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2000， 22(1)：4-37.

[56] Battiti R. Using mutual information for selecting features in supervised neural net learning[J]. IEEE Transactions on Neural Networks，1994， 5(4)：537-550.

[57] Sulaiman M A， Labadin J. Feature selection based on mutual information[C]// International Conference on It in Asia. IEEE， 2015.

[58] Yu L， Liu H. Efficient Feature Selection via Analysis of Relevance and Redundancy[J]. Journal of Machine Learning Research，2004， 5(12)：1205-1224.

[59] Fleuret， Fran&#. Fast Binary Feature Selection with Conditional Mutual Information[J]. Journal of Machine Learning Research， 2004， 5(3)：1531-1555.

[60] Ding C， Peng H. Minimum Redundancy Feature Selection from Microarray Gene Expression Data[C]// Bioinformatics Conference， 2003.

Csb 2003. Proceedings of the. IEEE， 2003：523.

[61] Hall M A. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning[C]// Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 2000：359-366.

[62] 陈思，郭躬德，陈黎飞. 基于聚类融合的不平衡数据分类方法[J]. 模式识别与人工智能， 2010， 23(06)：772-780.

[63] Fawcett T. An introduction to ROC analysis[J]. Pattern Recognition Letters， 2006， 27(8)：861-874.

[64] Ilango B S， Ramaraj N. A hybrid prediction model with F-score feature selection for type II Diabetes databases[C]//Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India. ACM， 2010：13.

[65] Zhang M L， Peña J M， Robles V. Feature selection for multi-label naive Bayes classification[J]. Information Sciences An International Journal， 2009， 179(19)：3218-3229.

[66] Akay M F. Support vector machines combined with feature selection for breast cancer diagnosis[J]. Expert Systems with Applications，2009， 36(2)：3240-3247.

[67] Liyuan L. A pilot study on risk factors and risk assessment score screening model for high-risk population of breast cancer[D]. Shandong University， 2010：16.

[68] 邢文训. 现代优化计算方法[M]. 清华大学出版社， 1999.

[69] 沈崇圣. 遗传算法中常用选择算子在 MATLAB 中的实现[J]. 上海应用技术学院学报(自然科学版)， 2003， 3(3)：199-202.

[70] 张林波. 并行计算导论[M]. 清华大学出版社， 2006

作者简历及在学研究成果

一、 作者入学前简历

起止年月	学习或工作单位	备注
2010 年 09 月至 2014 年 06 月	在北京科技大学通信工程专业攻读学士学位	

二、 在学期间从事的科研工作

三、 在学期间所获的科研奖励

四、 在学期间发表的论文

[1] **Xiaoli Lin**， Wei Huangfu， Fei Wang， Liyuan Liu， Keping Long. A breast cancer risk classification model based on the features selected by a novel F-score index for the imbalanced multi-feature dataset[c]. 8th International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (Cyberc 2016)，October 13-15， 2016， Chengdu.

独创性说明

本人郑重声明：所呈交的论文是我个人在导师指导下进行的研究工作及取得研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写的研究成果，也不包含为获得北京科技大学或其他教育机构的学位或证书所使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中做了明确的说明并表示了谢意。

签名：_____ 日期：_____

关于论文使用授权的说明

本人完全了解北京科技大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后应遵循此规定）

签 名：_____ 导 师 签 名：_____ 日 期：_____

学位论文数据集

关键词*	**密级***	**中图分类号***	**UDC**	**论文资助**
学位授予单位名称*		**学位授予单位代码***	**学位类别***	**学位级别***
北京科技大学		10008		
论文题名*		**并列题名**		**论文语种***
作者姓名*			**学号***	
培养单位名称*		**培养单位代码***	**培养单位地址**	**邮编**
		10008	北京市海淀区学院路 30 号	100083
学科专业*		**研究方向***	**学制***	**学位授予年***
论文提交日期*				
导师姓名*			**职称***	
评阅人	**答辩委员会主席***		**答辩委员会成员**	
电子版论文提交格式 文本（ ） 图像（ ） 视频（ ） 音频（ ） 多媒体（ ） 其他（ ） **推荐格式：**application/msword；application/pdf				
电子版论文出版（发布）者		**电子版论文出版（发布）地**		**权限声明**
论文总页数*				
共 33 项，其中带*为必填数据，为 22 项。				