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。Natural language processing is an important area in the field of computer science and artificial intelligence.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。It studies various theories and methods that enable effective communication between humans and computers in natural language.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。Natural language processing is a science that integrates linguistics, computer science, and mathematics.这一领域的研究将涉及自然语言，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。Research in this field is the basis for informatizing language and words.中文自然语言处理是自然语言处理的一个重要部分，Chinese natural language processing is an important part of natural language processing.中文相比于英文有词边界较难鉴定、句法更灵活等特殊性，Chinese has many characteristics such as fuzzy word boundaries and flexible syntax.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。In recent years, relevant research on Chinese natural language processing has also received increasing attention.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。Along with the deep learning craze, the study of Chinese natural language processing such as Chinese word segmentation, part-of-speech tagging, named entity recognition and sentence structure representation has also made great progress under the impetus of deep learning technology.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。With the deepening of Chinese informationization, with the deepening of Chinese informationization, we are increasingly discovering that ancient Chinese books containing the wisdom of the Chinese nation for thousands of years need us to use modern technology for proper preservation and processing.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，For the physical ancient books, we need to digitize, informatize them, and archive them into different categories. 这有利于我们传承传统文化和保护先人的思想精华。This is conducive to the inheritance of traditional culture and the protection of the essence of ancient thoughts.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。For e-books that have already been stored, we need to use modern Chinese natural language processing to perform deeper structured processing and data mining on ancient texts.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。This will have important significance for historical research, preservation of existing ancient Chinese books, and knowledge extraction, which will greatly promote the construction of digital humanities in China.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。However, the reality is that the current progress of digitization in ancient Chinese mostly stays in the storage stage, and there are few studies on subsequent word segmentation, part-of-speech tagging, named entity recognition, text structure processing, text classification, etc. Moreover, the accuracy of some methods currently implemented is not Very high.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。Part of the reason for limiting the accuracy of current methods is that the Chinese text originated from the late Shang Dynasty in the 14th century BC, when the initial Oracle was formed, which has lasted for more than 3,000 years. In the course of the evolution of three thousand years, the meaning, grammar and syntax of Chinese are constantly changing.以古文翻译来说，In the case of ancient Chinese translation, “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，‘*汤*’(*soup*) originally referred to as all hot water, now only refers to the soup obtained after cooking the food.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。Another example is ‘*治*’（）, which originally refers to the 'governance flood', so the word is from the side of *“氵”*(which means water) and later expanded to refer to all governance.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。This shows that Chinese in different periods will have different language characteristics.，semantics and grammar are not static. 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。In the face of the long history of Chinese and the change of Chinese structure over time, it is difficult to construct a model suitable for various periods.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。Therefore, only by judging the approximate time of the ancient books, can we conduct more follow-up research on ancient books and improve the accuracy and efficiency of the research.因此，本文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。in this paper, we attempt to explore the field of Ancient Chinese natural language processing from the perspective of Dating Ancient Chinese Texts. The research results of this paper will help to study the ancient Chinese word segmentation, part-of-speech tagging, named entity recognition, text structure processing, text classification and other aspects.

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。From a technical point of view, the time judgment of ancient text means that the model receives a piece of text, and the model automatically calculates and outputs an age label.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。Therefore, from the perspective of the relationship between input and output, the task of determining the age of ancient texts is a text classification task.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。The current text classification models can be roughly divided into two categories, one is traditional machine learning method which is based on rules or probability, and the other is deep learning method which is based on CNN, RNN or self-Attention.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。The rule-based and the probability-based approach are relatively simple, easy to implement, and work well in specific areas. Their advantages are low time complexity and fast computing speed. However, in the rule-based and the probability-based methods, many rules or specific conditions need to be considered, so it is necessary to define and manually extract features by domain experts.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，\*\*\*和\*\*\*分别将cnn和RNN应用到自然语言处理中，18年Google提出基于self-attention机制的bert模型，大有一统NLP领域之势。Combining deep learning methods to solve specific domain problems is a trend in recent years. \*\*\* and \*\*\* apply cnn and RNN to natural language processing respectively. In 2018, Google proposed a bert model based on self-attention mechanism. It reaches the top level in all kinds of tasks in nlp.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，However, the bert model is mainly oriented to modern languages, and its success mainly depends on the massive informational texts of the current Internet era, such as wiki encyclopedia entries, various news media, and commentary messages.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。200M model parameters in the bert model can be trained through a large number of training sets. However, this is not applicable in the ancient Chinese field where the corpus resources are relatively lacking.因此，本文提出使用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，Therefore, this paper proposes to solve the problem of dating age of ancient books by using LSTM deep learning network model.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。The model has two main advantages. One is that there is no need to manually extract rule features. Second, the training set data required by lstm is less than that required by the bert model.

Natural language processing is an important area in the field of computer science and artificial intelligence. It studies various theories and methods that enable effective communication between humans and computers in natural language. Natural language processing is a science that integrates linguistics, computer science, and mathematics. Research in this field is the basis for informatizing language and words. Chinese natural language processing is an important part of natural language processing. Chinese has many characteristics such as fuzzy word boundaries and flexible syntax. In recent years, relevant research on Chinese natural language processing has also received increasing attention. Along with the deep learning craze, the study of Chinese natural language processing such as Chinese word segmentation, part-of-speech tagging, named entity recognition and sentence structure representation has also made great progress under the impetus of deep learning technology. With the deepening of Chinese informationization, with the deepening of Chinese informationization, we are increasingly discovering that ancient Chinese books containing the wisdom of the Chinese nation for thousands of years need us to use modern technology for proper preservation and processing. For the physical ancient books, we need to digitize, informatize them, and archive them into different categories. This is conducive to the inheritance of traditional culture and the protection of the essence of ancient thoughts. For e-books that have already been stored, we need to use modern Chinese natural language processing to perform deeper structured processing and data mining on ancient texts. This will have important significance for historical research, preservation of existing ancient Chinese books, and knowledge extraction, which will greatly promote the construction of digital humanities in China. However, the reality is that the current progress of digitization in ancient Chinese mostly stays in the storage stage, and there are few studies on subsequent word segmentation, part-of-speech tagging, named entity recognition, text structure processing, text classification, etc. Moreover, the accuracy of some methods currently implemented is not Very high. Part of the reason for limiting the accuracy of current methods is that the Chinese text originated from the late Shang Dynasty in the 14th century BC, when the initial Oracle was formed, which has lasted for more than 3,000 years. In the course of the evolution of three thousand years, the meaning, grammar and syntax of Chinese are constantly changing. In the case of ancient Chinese translation, ‘汤’(soup) originally referred to as all hot water, now only refers to the soup obtained after cooking the food. and another example is ‘治’（governance）, which originally refers to the 'governance flood', so the word is from the side of “氵”(which means water) and later expanded to refer to all governance. This shows that Chinese in different periods will have different language characteristics.，semantics and grammar are not static. In the face of the long history of Chinese and the change of Chinese structure over time, it is difficult to construct a model suitable for various periods. Therefore, only by judging the approximate time of the ancient books, can we conduct more follow-up research on ancient books and improve the accuracy and efficiency of the research. in this paper, we attempt to explore the field of Ancient Chinese natural language processing from the perspective of Dating Ancient Chinese Texts. The research results of this paper will help to study the ancient Chinese word segmentation, part-of-speech tagging, named entity recognition, text structure processing, text classification and other aspects.

From a technical point of view, the time judgment of ancient text means that the model receives a piece of text, and the model automatically calculates and outputs an age label. Therefore, from the perspective of the relationship between input and output, the task of determining the age of ancient texts is a text classification task. The current text classification models can be roughly divided into two categories, one is traditional machine learning method which is based on rules or probability, and the other is deep learning method which is based on CNN, RNN or self-Attention. The rule-based and the probability-based approach are relatively simple, easy to implement, and work well in specific areas. Their advantages are low time complexity and fast computing speed. However, in the rule-based and the probability-based methods, many rules or specific conditions need to be considered, so it is necessary to define and manually extract features by domain experts. Combining deep learning methods to solve specific domain problems is a trend in recent years. Moschitti [Moschitti, A., Pang, B., & Daelemans, W. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).]and Tang[Tang, Duyu, Bing Qin, and Ting Liu. "Document modeling with gated recurrent neural network for sentiment classification." Proceedings of the 2015 conference on empirical methods in natural language processing. 2015.] apply Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to natural language processing respectively. In 2018, Google proposed a BERT model based on self-attention mechanism. It reaches the top level in all kinds of tasks in NLP. However, the BERT model is mainly oriented to modern languages, and its success mainly depends on the massive informational texts of the current Internet era, such as wiki encyclopedia entries, various news media, and commentary messages. 200M model parameters in the BERT model can be trained through a large number of training sets. However, this is not applicable in the ancient Chinese field where the corpus resources are relatively lacking. Therefore, this paper proposes to solve the problem of dating age of ancient books by using LSTM deep learning network model. The model has two main advantages. One is that there is no need to manually extract rule features. Second, the training set data required by LSTM is less than that required by the BERT model.