Y3690040

分类号 G250
学 号 2015114004

南京農業大學

硕士学位论文

《诗经》与其注疏文献的句子对齐研究

王姗姗

指导教师 何琳 教授 王东波 副教授
学科门类 管理学
一级学科 图书情报与档案管理
二级学科 图书馆学
研究方向 智能信息处理
答辩日期 二〇一八年五月二十六日

万方数据

0156783

The Study of the Sentence Alignment between Shijing with its Commentations and Annotations



By

WANG SHANSHAN

A Thesis

Presented to Nanjing Agricultural University

in Partial Fulfillment of the Requirement for the Degree of

Master of Management Science

in

Library Science

Supervised by

Professor HE LIN

Nanjing Agricultural University

Nanjing, China

May 2018

万方数据

原 创 性 声 明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者（需亲笔）签名：王娜娜　　2018 年 6 月 1 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权南京农业大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编学位论文。

保密□，在_____年解密后适用本授权书。本学位论文属于不保密□。

（请在以上方框内打“√”）

学位论文作者（需亲笔）签名：王娜娜　　2018 年 6 月 1 日

导师（需亲笔）签名：　　2018 年 6 月 1 日

万方数据

目 录

摘要 ..... **I**

ABSTRACT ..... **III**

第一章 绪 论 **1**

1.1 研究背景 1
1.2 研究目的与意义 1
1.2.1 研究目的 1
1.2.2 研究意义 2
1.3 研究内容 2
1.4 创新点 3
1.5 研究思路和研究方法 3
1.5.1 研究思路 3
1.5.2 研究方法 4

第二章 研究综述 **5**

2.1 《诗经》注疏文献版本流变 5
2.2 《诗经》及其注疏文献的相关研究 6
2.3 先秦文献句子对齐的相关研究 7
2.4 相似度计算研究 9
2.4.1 相似度计算相关研究 9
2.4.2 词语相似度计算相关研究 10
2.4.3 组块相似度计算相关研究 11
2.4.4 句子相似度计算国外相关研究 12
2.4.5 句子相似度计算国内相关研究 13

第三章 《毛诗正义》文本分析 **15**

3.1 《毛诗正义》行文结构 15
3.2 训诂语言分析 16
3.2.1 训诂术语及训诂格式 17

万方数据

3.2.2 训诂内容 19
3.2.3 训诂方法 23
3.2.4 训诂体例 25
第四章 语料库构建 27
4.1 数据预处理 27
4.1.1 数据校对 27
4.1.2 语料标注 30
4.2 对齐规范设计 31
4.2.1 外观布局 31
4.2.2 标注规范 32
4.3 自动分词 34
4.3.1 自动分词的流程 34
4.3.2 机器学习模型的简介和特征模板的构建 35
4.3.3 机器学习分词实验与结果分析 37
第五章 对齐实验 41
5.1 对齐算法的总体设计 41
5.2 篇和章的对齐 42
5.3 训诂术语正则匹配实验 43
5.3.1 训诂术语与正则表达式 43
5.3.2 正则表达式匹配实验 44
5.3.3 正则表达式匹配结果及分析 46
5.4 相似度算法对齐 47
5.4.1 gensim 简介 47
5.4.2 相似度计算实验 48
5.4.3 相似度算法对齐实验及结果分析 50
第六章 对齐实验效果测评 53
6.1 测评方法 53
6.1.1 测试语料 53
6.1.2 测试方法 55
6.1.3 测评指标 56
6.2 训诂术语正则匹配实验测评及结果分析 57

万方数据

6.3 GENSIM 相似度计算实验测评及结果分析 60
第七章 总结与展望 63
7.1 本文总结 63
7.2 存在的不足 63
参考文献 65
附录一 繁体字校对样例 69
附录二 词性标记集合及特征模板 71
附录三 《毛诗正义》训诂术语模式库 73
附录四 自动分词-全特征模板样例 75
附录五 正则对齐实验结果 77
附录六 相似度对齐实验结果 79
致 谢 81
攻读学位期间已发表的学术论文 83

万方数据

图目录

图 2-1 《毛诗注疏》版本流变 6
图 4-1 数据校对和语料标注前的《毛诗正义》行文格式 30
图 4-2 数据校对和语料标注后的《毛诗正义》行文格式 30
图 4-3 《诗经》顺序标注 33
图 4-4 《行露》诗名经-注-疏初步匹配结果 33
图 4-5 《行露》诗内容经-注-疏初步匹配结果 33
图 4-6 《诗经》自动分词流程 34
图 4-7 《诗经》手工分词语料部分样例 35
图 4-8 23 组特征组合模板的性能综合比较 40
图 5-1 句子对齐实验流程 41
图 5-2 “篇和章的对齐”结果 42
图 5-3 “注”一次匹配结果 47
图 5-4 corpus 特征词及词频 49
图 5-5 “引文”特征词及词频 49
图 5-6 “引文”特征词二维矩阵转换 49
图 5-8 主题向量转换代码 50
图 5-9 LSI 相关度前 15 的 topics 51
图 5-10 相关度排第十的 topic 51
图 6-1 正则表达式匹配“注”一次匹配的前后两部分内容 54
图 6-2 gensim 相似度匹配“注”一次匹配的前后两部分内容 55
图 6-3 已标注的标准数据集示例 56
图 6-4 “引文”和“注”第一次匹配结果 58
图 6-5 “引文”和“注”第一次匹配结果配“引文”内容 58
图 6-6 匹配“注”的第一部分和第二部分内容 59
图 6-7 “引文”与“注”和“疏”匹配结果 61
图 6-8 “引文”与“注”和“疏”匹配得到的词频词表 61

万方数据

表目录

表 4-1 繁体字校对记录 28
表 4-2 多个繁体字共存使用的样例 28
表 4-3 《毛诗引得》与《诗经》分词语料词长词频统计 29
表 4-4 简单特征模板说明 36
表 4-5 增加一个特征知识的复杂特征模板 37
表 4-6 《毛诗引得》领域词表的词汇长度分布 38
表 4-7 语料以 9:1 训练在三类特征中的性能比较 39
表 4-8 基于规则分词方法与机器学习模型分词性能对比 39
表 5-1 标注字音的常用训释方式 43
表 5-2 训诂术语以及相应的正则表达式形式 44
表 6-1 训诂术语正则匹配“引文”和“注”的测评结果 60
表 6-2 相似度计算匹配的测评结果 62

万方数据

《诗经》与其注疏文献的句子对齐研究

摘 要

《诗经》的注疏文献颇多，“十三经”各注释本中以阮刻本《十三经注疏》最为完善，其冠列经史子集的经部之首，故本文研究选用阮刻本《十三经注疏》中的《毛诗正义》作为《诗经》注疏文献的研究对象。

本文以《毛诗正义》《毛诗引得》等作为资源，借鉴已有的《孟子》《论语》《左传》等先秦典籍的研究，实现从更深的层面对《诗经》进行加工和处理，对《诗经》进行较深层次的加工和处理，尝试探索基于引得领域词表的《诗经》自动分词，采用正则表达式匹配和相似度计算等方法探索《诗经》与其注疏文献《毛诗正义》的句子对齐研究得到了较好的对齐效果。本文主要是从以下三方面进行句子对齐研究：

首先，对《毛诗正义》的行文结构和训诂语言进行分析。总结归纳《毛诗正义》的篇章划分特征，为后续实现篇和章的对齐提供依据。《毛诗正义》中不同的训释对象和训释内容，一般会使用不同的具体术语，通过归纳总结训诂术语在术语、格式、内容、方法和体例这五个方面的特征和规律，找出训诂的规律，利用正则表达式提炼加工，构建训诂具体术语模式库，用以辅助探索本文的句子对齐。

其次，采用条件随机场的机器学习方法探究了《诗经》的自动分词。《诗经》自动分词是基于《诗经》手工分词的语料，采用《广韵》字表和统计分析相结合的方法，得到 23 组融合不同特征知识的特征模板，训练产生机器学习分词模型。并对每个分词模型进行性能测试，分析发现词性特征对《诗经》分词效果的影响最大，且分词模型的调和平均值 F 值最高可达到 97.42%。

再次，采用正则表达式匹配和相似度计算相结合的方法进行《诗经》与《毛诗正义》的句子对齐研究。正则表达式匹配实验是利用训诂术语模式库，采用 python 对实验语料进行模式库的多次匹配，逐步进行效果测试，可获得较好的匹配效果。Gensim 是一种自然语言处理库，可自动从文本中提取语义信息，本文基于 gensim 结合词频权重向量（TF-IDF），构建 LSI 主题模型进行相似度计算，结合正则表达式匹配结果，共同探索《诗经》与《毛诗正义》的句子对齐研究，并对对齐实验结果进行效果测评。

利用训诂术语正则匹配进行句子对齐，准确率 P 最高可到 86.92%，调和平均值 F 可达 84.61%，该方法结合了文本结构特点及训诂术语语言特征，实验效果较佳。采用 gensim 相似度计算的方法结合了语义特征，实验得到的句子匹配结果也较理想。经-注-疏匹配的准确率可达到 89.92%，经-疏匹配的准确率有 86.14%，调和平均值 F 最

万方数据

高可达85.6%。由此可看出，训诂术语正则匹配和gensim相似度匹配的方法均可用于探索本文的句子对齐，且对齐实验效果较好。此外上述两种方法是从不同的角度探究《诗经》与《毛诗正义》的句子，后续研究可结合使用。

本文的创新点有两个：一是对《毛诗正义》的行文结构和训诂语言特点进行了归纳总结，构建了较为完整的行文结构体系和训诂术语模式库，可计算机化便于后人在此基础上进行相关研究。第二是探索得到一种从多角度采用多种词表，构建机器学习模型实现《诗经》自动分词的方法，并构建了《诗经》分词语料库。进而探索了一种基于训诂术语模式库的正则表达式匹配和相似度计算相结合的方法实现了《诗经》与《毛诗正义》的句子对齐，并取得了较好的对齐效果。

但本文仍有不足之处。首先训诂术语模式库，虽尽可能归纳全，但仍可扩大《诗经》的注疏文献范围实现进一步完善。其次，句子对齐实验中未对《毛诗正义》“疏”部分中各专家言论进行细致划分，后续研究可分析归纳各家言论所长。

关键字：诗经；毛诗正义；句子对齐；训诂；正则表达式；相似度计算

万方数据

The Study of the Sentence Alignment between Shijing with its Commentations and Annotations

ABSTRACT

There are quite a few notes on the annotation of the Shijing, but the most perfect one is the notes of the "*the Thirteen Cassics*", it is the head of the history subsets. Therefore, this paper uses *Mao Shi Zheng Yi* by Ruan Yuan in block-printed edition of *the Thirteen Classics* commentary, which is subject investigated in *the Book of Songs*'s Commentary Literature.

In this paper, we take the *Mao Shi Zheng Yi* and *Mao Shi Index* as resources. This article learns from previous studies on Pre-Qin Classics such as *Mencius, analects of Confucius*and *Master Zuo's spring and Autumn Annal*, which made a deeper level processing of Shijing, by trying to explore the automatic word segmentation of Shijing, which is based on the domain word list in *Mao Shi Index*. We use regular expression matching and similarity calculation to explore the sentence alignment of *the Book of Songs* and its annotation literature *Mao Shi Zheng Yi*, it has a good alignment effect. This paper mainly studies sentence alignment from the following three aspects:

First, we analyze the textual structure features and exegesis language of *Mao Shi Zheng Yi*. This paper sumarizes the textual classification featuers of *Mao Shi Zheng Yi*, and it provides a basis for the alignment of article and chapters.In the *Mao Shi Zheng Yi*, the different training objects and the content of the training generally use different specific terms. By summarizing the characteristics and laws of the five aspects of exegesis terms in terms, formats, contents, methods and styles, finding out the laws of exegesis. We use regular expressions to refine processing, build an exegesis specific term pattern library to assist in the exploration of sentence alignment in this article.

Secondly, this paper studies the automatic word segmentation of *The Book of S ongs* by machine learning method of conditional random fields. The automatic segm entation of *the Book of Songs* is based on the corpus of the manual word segmentatio n of *the Book of Songs*. The method of combining the Guang Yun and statistical analysis w as used to get 23 sets of feature. Based on the corpus of the manual word segmentation o f The Book of Songs, the method of combining the Guang Yun and statistical analysis wa s used to get 23 sets of feature template which fuse different characteristic knowledge and t hen producing machine learning segmentation model by training. template which fuse diffe

万方数据

rent characteristic knowledge and then producing machine learning segmentation model b y training. The performance of each word segmentation model is tested. It is found that the word character has the greatest impact on the effect of the word segmentati on of Shijing, and the highest F value of the harmonic mean value of the participle model can reach 97.42%.

Thirdly, this paper studies the sentence alignment of *Shijing* and *Mao Shi Zheng Yi* by combining regular expression matching and similarity calculation. The regular expression matching experiment is based on the exegesis specific term pattern library by using Python to carry out multiple pattern matching for experimental corpus, and test the effect step by step to obtain better matching effect. Gensim is a Natural Language Processing library, which can automatically extract semantic information from text. Based on Gensim, this paper combined with the word frequency weight vector (TF-IDF). This paper constructs LSI theme model for similarity calculation and combines the results of regular expression matching to explore the sentence alignment of *Shijing* and *Mao Shi Zheng Yi.*

Using terminologies of regular matching for sentence alignment, the highest accuracy rate of P can reach 86.92% and the harmonic mean value of F can reach 84.61%. This method combines the text structure features and the terminologies of linguistic features, the experimental effect is better. Gensim similarity calculation method is combined with semantic features. The matching result of the sentence using Gensim similarity is actually ideal. The accuracy rate of the matching between the classics notes and commentaries is 89.92%, and the accuracy of the matching between the classics and notes is 86.14%, and the harmonic mean value of F can reach 85.6%. It shows that the method of regular matching and gensim similarity matching can be used to explore sentence alignment in this paper, and the alignment experiment is effective. In addition, the above two methods exploring the sentence alignment of *Shijing* and *Mao Shi Zheng Yi* from different perspectives, which can be combined for subsequent research.

There are two innovative points in this paper. First, the paper summarizes the text structure features and the terminologies of linguistic features of *Mao Shi Zheng Yi*, building a relatively complete system of text structure and the exegesis specific term pattern library. It can be implemented on the computer, which is convenient for future generations to do related research on this basis. Secondly, the paper explores a method to realize automatic segmentation of *the Book of Songs* by using multiple words vocabularies from different angles, and it obtains the word corpus of *The Book of Songs* that fuses the expert vocabulary

万方数据

knowledge of the *Mao Shi Index*. this paper studies the sentence alignment of *Shijing* and *Mao Shi Zheng Yi* by combining regular expression matching and similarity calculation, which is based on the exegesis specific term pattern library. This method of sentence alignment botains a better alignment effect.

However, there are still some shortcomings in this paper. First of all, the exegesis specific term pattern library can be further improved by expanding the scope of the annotation of the *Shijing*, although it has been summarized as complete as possible. Secondly, there is no detailed division of the opinions of various experts in the commentaries of *Mao Shi Zheng Yi* in the experiment of sentence alignment. Subsequent studies can analyze and summarize the expressions of each expert.

KEY WORD: *Shijing*; *Mao Shi Zheng Yi*; sentence alignment; regular express; similarity calculation

万方数据

第一章 绪 论

1.1 研究背景

《诗经》作为中国诗歌类型典籍的鼻祖及典型代表，从古至今对其的相关研究不胜枚举。已有的《诗经》的相关研究主要是针对其内容、风格等方面。随着信息技术的发展以及人文计算的手段和方法的成熟，对《诗经》进行深入的内容挖掘已成为可能。

在先秦典籍的信息处理方面，已有不少的典籍可实现完整的信息处理过程，包括前期的自动分词、命名实体、句子对齐等相关的研究，研究者们不仅对先秦典籍本身进行研究，还结合其注疏文献等各类典籍进行研究。借鉴已有的先秦典籍相关研究的方法和思路，探究《诗经》与其注疏文献的句子对齐研究，将填补《诗经》在这部分研究的缺口，且要想让更多的后人深入了解和充分利用《诗经》的精华部分，对《诗经》的内容挖掘研究必不可少。

《诗经》是中国古代第一部诗歌总集，奠定了中国两千年来的诗歌创作和发展的基础，《诗经》位居第一位在古文经学派的“五经”中，其同时兼有经学经典、文学经典和文化、思想经典等多重身份，是中国古代政治伦理学、美育、博物学的教科书。

目前对《诗经》及其注疏文献的研究，更多的是从语法学、训诂学、修辞学、美学、诗学观念、文学理论等角度研究分析，从信息处理角度进行《诗经》研究的极少，已有的研究是针对《论语》《孟子》《左传》。

1.2 研究目的与意义

1.2.1 研究目的

本研究以《诗经》及其注疏文献《毛诗正义》为研究对象，借鉴已有的《孟子》《论语》《左传》等先秦典籍的研究，实现从更深的层面对《诗经》进行加工和处理，，尝试探索出一条适用于先秦诗歌体的句子对齐的新方法、新思路。

先秦文献的句子对齐研究通常是利用训诂术语或词表，采用基于规则的方法，或者利用动态规划和相似度计算等方法进行探索。《毛诗正义》训诂语言方面的研究已有不少，但分布不集中，散落在各类文献中，由于不同的研究者是从不同的角度进行研究，面对星罗棋布的文献，对其进行分类系统地梳理是很有必要的。

本文搜罗各类相关的论文，从训诂术语的术语、格式、内容、方法和体例这五个方面对《毛诗正义》的训诂术语特点进行归纳总结，力图构建较为完整的《毛诗正义》

万方数据

训诂术语模式库，用以辅助探索本文的句子对齐研究。再结合句子对齐的结果，构建基于注疏内容的词表，并将其与引得词头和手工分词的结果进行对比，进而获得一种效果更佳的用于《诗经》自动分词的方法。

1.2.2 研究意义

实现《诗经》与其注疏文献的句子对齐，对实现《诗经》进一步的命名实体识别和内容挖掘等奠定了数据基础。此外，本研究可用于构建基于《诗经》对齐研究结果的检索平台，即以这个句子对齐和注释对齐的数据库为基础，就可以设计一个快捷高效的检索工具，凭借这个检索工具，先秦文献的研究者可以很方便的检索到文献中每一句话在多部注疏文献中的注、疏、正义等等，研究者探索古籍文献的视野将会更加开阔。

《毛诗正义》训诂术语模式库的构建对类似于《诗经》的先秦诗歌体典籍的训诂术语研究或对齐研究有借鉴意义，本文从多角度对《毛诗正义》的训诂术语特点进行归纳总结，这种探索模式同样可适用于其他典籍的相关研究。为更全面地利用《毛诗正义》文献语言特点，本文在训诂术语模式库的基础上，利用兼具语义相似度的 gensim 相似度计算方法，探索《诗经》与其注疏文献的句子对齐研究，并获得了较好的对齐效果。

1.3 研究内容

本文主要从几方面探索《诗经》与其注疏文献句子对齐研究。

（1）《毛诗正义》行文结构分析。进行文献调研，确定《诗经》在发展过程中产生的不同版本的注疏文献，了解它们的背景、版本、作者、体例等，比较每个版本的差异，从而确定本文采用的《诗经》注疏文献是清嘉庆二十一年阮元校刻《十三经注疏》(简称阮刻本)中的《毛诗正义》。分析归纳《毛诗正义》的篇章布局特点，为后续界定篇和章的范围，实现章的对齐做铺垫。此外，还需对每章中的句子语言特点进行分析归纳，进而实现每个句子与其注文的对齐。

（2）《毛诗正义》训诂语言分析。《毛诗正义》训诂术语特征为进行系统的梳理，因此，本文针对其训诂术语从术语、格式、内容、方法和体例五个方面对其进行整理归纳，从而构建《毛诗正义》训诂术语模式库，为后续进行正则匹配实验做铺垫。

（3）《诗经》自动分词。句子对齐实验需对文本进行分词处理，为此文本利用《毛诗引得》领域词表和手工分词的语料，基于统计分析的方法，搭建 23 组基于不同特征知识的特征模板，用机器学习的方法实现自动分词，分词效果更好。

万方数据

（4）句子对齐实验。基于训诂术语模式库，利用训诂术语正则匹配和相似度计算相结合的方法探索本文的句子对齐实验，采用常规的进行效果测评的P、R、F三个指标，校验对齐效果，以验证本文方法用于实现《诗经》与其注疏文献的句子对齐是有效可行的。

1.4 创新点

（1）本文研究的是《诗经》与其注疏文献《毛诗正义》的句子对齐。文中对《毛诗正义》的行文结构及训诂语言特点进行了深入细致的分析，结合语料本身的特性，对《毛诗正义》的篇章布局、文本结构都进行了探索分析，并从训诂术语的术语、格式、内容、方法和体例五个方面对《毛诗正义》的训诂特点进行剖析，形成了一个较为完整的《毛诗正义》行文结构分析体系，并构建了结合《毛诗正义》训诂特点的训诂术语模式库，可计算机化便于后人在此基础上进行相关研究。

（2）本文探究《诗经》与其注疏文献《毛诗正义》的句子对齐，基于《毛诗引得》领域词表和《广韵》字表，从多角度构建了机器学习模型，实现了《诗经》的自动分词并构建了《诗经》分词语料库。此外，采用正则表达式匹配和相似度计算匹配相结合的方法，利用训诂术语模式库和 gensim 相似度计算方法，取得较好的句子对齐效果，此方法不仅适用于《诗经》典籍与其注疏文献的句子对齐研究，且对类似《诗经》的诗歌体典籍的对齐研究起辅助作用。

1.5 研究思路和研究方法

1.5.1 研究思路

本文首先搜集对比了《诗经》不同版本注疏文献的差异，从版本来源、注释手段、保存的完整性等多个角度比较筛选，最终确定本文研究所用的《诗经》注疏文献版本；此外，本研究最初得到的《毛诗正义》的数据是简体版，为进行后续的研究，需采用相关的算法进行简繁体转换，并对得到的繁体版《毛诗正义》进行繁体字的校对，以便用于后续的对齐研究。接下来，对选定的《诗经》注疏文献《毛诗正义》进行行文特点分析，从其行文结构和训诂语言两方面分析，并构建《毛诗正义》训诂术语模式库，用于后续进行正则匹配，并将正则匹配结果结合相似度计算的方法一并用于研究《诗经》与《毛诗正义》的句子对齐，从而探索出一种适用于《毛诗正义》的句子对齐方法，并尝试将得到的对齐结果，用于《诗经》基于注释信息的分词研究。

万方数据

1.5.2 研究方法

本文研究的是《诗经》与其注疏文献《毛诗正义》的句子对齐研究。首先从语料本身出发，从其篇章、训诂语言等角度，总结了《毛诗正义》的文本结构特点。

（1）文献调研法。进行文献调研，确定《诗经》在发展过程中产生的不同版本的注疏文献，了解它们的背景、版本、作者、体例等，比较每个版本的差异，从而确定本文采用的《诗经》注疏文献是清嘉庆二十一年阮元校刻《十三经注疏》(简称阮刻本)中的《毛诗正义》。

（2）统计分析法。在探究《诗经》自动分词中，结合统计分析的方法确认特征知识，从而构建不同特征知识下的特征模板。

（3）实验建模。通过正则表达式实验和句子相似度匹配实验，探索《诗经》与《毛诗正义》的句子对齐。

万方数据

第二章 研究综述

秦始皇的焚书和楚汉的连年战争，使得《诗经》的文化传承遭到严重破坏，朝代的交替及语言自身的演变，先秦典籍所使用的语言到汉代已有不少的差异，因此，后人学《诗》自汉以来需借助于注释。

毛亨于西汉创作的《毛诗训诂传》，是中国历史最悠久的《诗经》典籍，又被称作《毛传》；郑玄在东汉末年为《毛传》作笺，有被后人称之为《毛诗笺》或《郑笺》，；而后生于梁末陈初，经历了陈、隋、唐三朝的陆德明又为《毛诗》作了音义，后人称之为《毛诗音》；唐太宗时期，孔颖达被皇帝任命对《毛传》《郑笺》进行深入地疏解，进而创作了《毛诗正义》，其又被后人称为《毛诗注疏》，后人将其列为《十三经注疏》之一[1]。

《毛诗正义》主要有四个部分，分别包括：《毛诗》经文、注文、《正义》以及《毛诗释文》，其中经文采用的是《毛诗定本》，出自于颜师古；注文则是《毛传》和《郑笺》；《正义》是由孔颖达主持编撰的版本[2]。每部注释文献承上启下，均对之前采用的典籍进行了注解，并收录了该时代有代表性学者的相关注释。《毛诗正义》将经文、注文和释音融合，在文字、音读、疏义三个方面实现了统一，这是后续编排注疏文献的惯用标准体例，可供阅读者同时获知《诗经》相关的各类相关注释，这无疑为阅读和可以提供了极大地便利。

2.1 《诗经》注疏文献版本流变

曹诣珍(2006)[3]全面地考察了《毛诗注疏》版本流变，并进行了清晰的梳理，构建其版本流变图(如图 1)：

[1] 江山丽.《诗经》及其注疏若干语法现象比较[D]. 湖南：吉首大学,2010:2.
[2] 韩宏韬. 《毛诗正义》单疏本考[J]. 河池学院学报(哲学社会科学版), 2006(06):44-48.
[3] 曹诣珍.《毛诗注疏》版本流变考略[J].古籍整理研究学刊, 2006(03):70-72+39.

万方数据

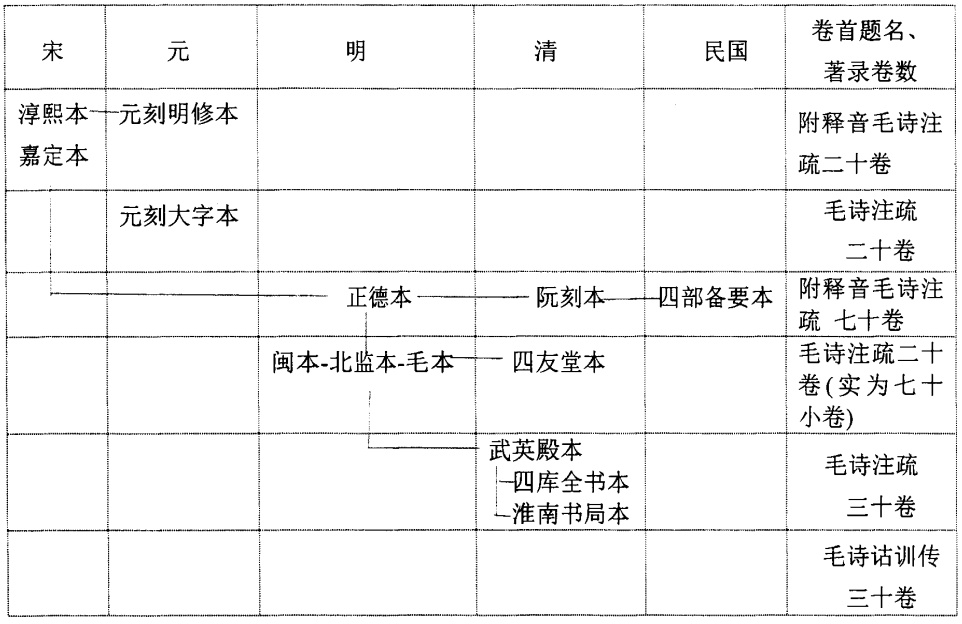


图 2-1 《毛诗注疏》版本流变

由此可见，《毛诗注疏》的版本流变清晰可查，版本内容逐渐丰富，经、传、笺、音义、疏的排版，已有最初的单行本，经、传别行，发展成为经、传同行，以疏合注，经、注、疏、释文合刻。

目前而言，阮元组织校刻的《十三经注疏》版本最为全面，其系统地对前人的相关《诗经》注释文献进行了梳理，也是后世研究《诗经》注疏文献不可缺少的素材，故本文就采用这个版本作为基准展开研究。北京大学对《毛诗正义》进行过一番整理疏通，收录了阮刻本及孙诒让版本的《校勘记》，形成了相应的整理本，该版本是繁体字，为此，本文研究将此本作为繁体字校对的底本，对获取的《毛诗正义》数据进行简繁体字校对[1]。

《四库全书总目》认为《毛诗正义》的涵盖范围广，内容很是丰富，其将《毛诗义疏》以及《毛诗术义》作为全文注释的底本，这两本著作都是当时有代表性的《诗经》注释文献。”因此，《毛诗正义》对研究《诗经》具有十分重要的作用[2]。

2.2 《诗经》及其注疏文献的相关研究

前人对《诗经》及其注疏文献的研究，更多的是从语法学、训诂学、修辞学、美学、诗学观念、文学理论等角度研究分析。

[1] 龚抗雲, 李传书, 胡渐逵等, 北大整理本_毛诗正义(全三册)_汉毛亨传;郑玄笺;唐孔颖达疏[M], 2000:1-11.
[2] 谈春蓉, 李锐.《毛诗正义》的注疏成就[J]. 文学教育(下), 2007(08):32-33.

万方数据

罗荣华(2004)[1]采用对比、归纳、统计等方法，从句法、语义、语用三方面总结分析《毛诗诂训传》、《毛诗笺》、《毛诗正义》的语法观和句法观的特点。该文不仅对“三家注”的句式句型进行分析，还从语义关系以及修辞手法等角度对部分动宾关系进行了具体分析，例如使动关系、意动关系、为动关系等，这有助于理解“三家注”的文本结构。

向熹(2013)[2]从词汇语义、语音、文字及语法等方面回顾和梳理了两千年来特别是20世纪的《诗经》语言研究。提出虽原则上疏不破注，可实际上《毛诗正义》的许多解释不同于《毛传》《郑笺》，但其得到后世学者的赞同。此外，刘挺颂(2007)[3]收集整理了20世纪80年代至2007年间《毛诗正义》的相关论文近40篇，从多个角度对《毛诗正义》的诗经学方面的成就进行总结分析，主要是从文学的角度包括其理论和批评，以及整个诗经的发展史的角度。

韩宏韬(2007)[4]研究提出《毛诗正义》在编撰方面具有官学性质、偏重北学，作者分布呈现主导性和群体性。分析了《毛诗正义》的版本和体例，这有助于理解“正义”的流变和其体例变化。此外，该文从《正义》对《诗经》汉学的统一、发展和批评三方面研究了《正义》的经学成就，并对其解经过程中的情志、比兴、语境三方面分析《毛诗正义》的文学思想。

2.3 先秦文献句子对齐的相关研究

目前针对先秦文献信息处理方面的相关研究成果不多，已有的研究对象主要是《左传》《论语》《孟子》，针对《诗经》的信息处理研究甚少。

陈小荷、冯敏宣、徐润华等编著的《先秦文献信息处理》(2013)[5]一书，较为全面的阐述了先秦文献信息处理的特点和方法，针对以下主题做了较为细致的研究：注疏文献处理、词语切分、词性标注、古今字和通假字资源建设、词汇概貌、词义消歧、专名标注、版本异文发现等等诸多方面。

经典古籍与注疏文献主要是在句子层面和注释角度这两方面进行对齐研究。其中，句子对齐是指将原典籍中的经文句子与引用了典籍的注疏文献的对应句子实现对照的一系列过程；注释对齐则主要是针对注释的词汇包括字、词甚至是长度较长的短语，将这些注释词汇的注释内容与被注释内容实现对应的过程。通常将注疏文献中出现的经典古籍中的句子称作是引文，即引用原典籍的内容[5]。

[1] 罗荣华.《诗经》三家注的语法观及其发展[D]. 宁夏：宁夏大学, 2004:6-60.
[2] 向熹.《诗经》语言研究古今谈[J]. 汉语史研究集刊, 2013(16):1-29.
[3] 刘挺颂. 20世纪80年代以来《毛诗正义》研究述略[J]. 贵州教育学院学报, 2007(06):79-82.
[4] 韩宏韬.《毛诗正义》研究[D]. 济南：山东大学, 2007:11-197.
[5] 陈小荷, 冯敏宣, 徐润华. 先秦文献信息处理[M]. 2013:16-82.

万方数据

马创新(2012)[1]构建了《论语》及其注疏文献的对齐语料库。该文梳理了《论语》具有代表性的注疏文献的主要源流，精选其中的四部注疏文献作为其建库的基本材料，首先在句子层面构建对齐关系，此外，还分析注疏文献中每条注释的指向，在注释语句和注释对象间建立对齐关系。此对齐语料库是通过三层倒排表的形式展示，在《论语》与其四部注疏文献间建立多向关联，使得注疏文献能通过经文这个枢纽建立联系，在这些著作间建立相互联系的关系网。

在实现句子对齐的过程中，句读划分上的不一致，将使得在进行典籍与注疏文献对齐的过程中出现非一一匹配的现象，这给对齐处理过程带来了一定的难度。马创新(2012)[2]采用动态规划的方法实现句子对齐，动态规划方法适用于解决对齐过程中的多种匹配现象。该方法在《论语》注疏文献的句子对齐实验中的测试结果理想。

梁社会(2013)[3]实现的是《孟子》注疏文献的平行对齐。首先通过引入异体字、简繁体字转换字表，预先对引文中的字形进行转换，并采用基于范围检查的句子对齐算法，该算法需先总结注疏文献的对齐规则，依靠规则和上下文实现对齐。还采用字符串相似度的句子对齐算法，针对匹配过程中一对多的情况，将经典古籍中无法对齐的小句与注疏文献中的单个小句进行余弦相似度计算，设定一个核定的阈值，当相似度的结果高于设定的阈值，则可以认为进行匹配的这两部分句子存在对齐关系。笔者总结不同类型词语的训释方式并构建正则表达式模式库，利用该正则表达式模式库匹配注疏文献实现注释对齐。

由此可见，要实现注疏文献的句子对齐，可先对注疏文献的文档结构进行分析，总结归纳行文结构特点，先采用基于规则的方法实现初步的对齐，对于一对多或多对一的状况，可采用动态规划或相似度计算等方法实现句子对齐。此方法在《论语》和《孟子》注疏文献的对齐研究中有效且效果良好。

注疏文献句子对齐的结果可用于经典古籍分词处理。徐润华(2012)[4]充分利用文献自有的特征，不论是在词汇、篇章布局还是其隐含的语义信息方面，注疏对齐后可产生注疏词表，利用该词表从新的角度探索《左传》的自动分词。自动对齐包含两方面的内容：一类是从注疏文献中找到与原典籍相关的句子，二类是将原典籍词汇在注疏文献中的注释内容匹配上。在实现自动对齐后，可利用子串查找和匹配的方法，从注疏文献中将经文中出现的词汇抽取出来，构建注疏词表。由于注疏词表是直接地利用已经实现了注疏文献对齐的结果，因此可省略例如机器学习分词方法中前期的语料

[1] 马创新, 陈小荷,曲维光, 陆鹏飞.《论语》与其注疏文献对齐语料库的构建[J]. 现代教育技术, 2012(07):109-113.
[2] 马创新, 陈小荷,曲维光. 注疏文献中的注释语句自动分析[J]. 计算机科学, 2012(10):220-223.
[3] 梁社会.《孟子》及其注疏的信息处理[D]. 南京：南京师范大学, 2013:10-30.
[4] 徐润华, 陈小荷.一种利用注疏的《左传》分词新方法[J]. 中文信息学报, 2012(02):13-17+45.

万方数据

分词标注过程。此外，将注疏词表用于分词处理，其分词的效果也可反向证明该注疏文献对齐工作的重要性。目前已有的注疏对齐研究，为将对齐结果的准确度提高，需进行相似度计算，可从注、疏中将含有经典古籍的词汇抽取出来，将其与注、疏中相应的内容匹配，以此实现字、词或短语上的注疏对齐，抽取出的词汇可用于构建注疏词表，从而实现注疏对齐。

2.4 相似度计算研究

2.4.1 相似度计算相关研究

自然语言处理研究中，不可或缺的研究领域是相似度计算。其主要是利用文本自身的特点，包括其语义关系、语法结果、语用等方面，从这些角度出发，探索文本的相似度。但目前针对文本相似度可行的通用的方法式从语义的角度出发，涵盖相似度涉及的多个层面，主要是从词、组块、句子等多个角度，层层叠加交融互通，相互间互为支持和联系，但又有着各自独特的相似度计算的方法。为此，从相似度计算的层面角度，可将相似度计算分为从词语、组块、句子和文本这四个方面的相似度计算，处理对象的内容逐渐丰富。

主要是从规则和统计两个角度进行词语层面的相似度计算。规则的方法是利用语义词典，在概念层次上实现的，其中的概念是词汇在语义方面的描述。这类的相似度方法主要还是依靠词表等实现。统计的方法则是在规模较大的语料的基础上，通过对语料进行训练，进而发现判断进行相似度计算的两个对象间是否拥有一定数量的相关词，进而判断两者之间是否存在有较高的相似度。

组块相似度其实和词语相似度的应用原理是相似的，都是利用词语在语义层面的特征进行相似度计算。但句子层面的相似度则是在前者相似度实现的基础上，从词汇、语义、编辑距离、关键词、语境框架等多个角度和切入点，探索句子的相似度。这些探索方法形式各异，利用统计或多种方法相融合，从不同的角度进行句子相似度计算，但每种方法都有各自的优缺点。

从向量角度出发探索相似度的方法是文本层面普遍应用的方法，该方法是将需要进行相似度计算的相关文本通过向量转化方法转化成空间向量，之后对比已转化成的向量空间之间的相似度值来确定对应的文本之间的相似度关系，向量相似度计算中最常用的是余弦系数。在文本的相似度计算中，基于统计的方法相较于语义的方法可取得更好的实验效果，但在词语层面、组块层面和句子层面进行相似度计算时，则是从语义角度出发的相似度计算方法相对于统计方法可取得更好的实验效果，随着研究的深入，各层面的相似度计算研究都取得了相应的进展。

万方数据

2.4.2 词语相似度计算相关研究

词语相似度计算已逐步延伸发展，从最初的从字面角度探索的相似度计算方法，后期的研究已发展到从词素角度考虑进行相似度计算。但前者的实现方式更简单，并不依赖其他的外在资源，利用研究对象本身的特征即可实现，但该方法无法区分在字形上相似的词汇的相似度关系。

Redaetal(1989)[1]构建了联机医学文献分析和检索系统，在该系统中构建了对应的语义图，通过比对图中两节点间的距离长短来判断对应的语义距离，节点的实际距离越短则语义相似度越接近。该系统总共有九个层次，分布着一万多个节点，供相关人士进行专业论文的检索。该系统语义距离的测定是利用节点间的变数来确定，这种语义距离的测定思想是极为简单的，但却可以取得十分出色的结果，该系统可谓是成功的。

Agirre 等人[2]采用概念层次数的思路，基于 Word Net 的语义词典进行相似度计算，其度量依据是路径长度和节点深度，当路径长度相同时，节点的层次越接近底部则语义距离越小。Resnik.P.(1999)[3]采用类似的基于词汇概念层次的路径和深度衡量相似度，区别在于其概念层为网状，如果两个词汇之间有路径存在，则将最原始的祖先节点所在的信息量最大值作为计算相似度的衡量指标。王斌(1999)借助《同义词林》[4]，以节点间的路径长度作为衡量语义距离的标准进行汉语词语的相似度计算。词条的结构是树状的分布于《同义词林》中，且按规模的大小分为三类，级别的增长将使得在语义方面的刻画越细致，整个体系呈现一个网状结构。刘群等人[5]基于“部分相似度加权平均得到整体相似度”的思想，针对《知网》的词汇提出一种计算两概念语义表达间相似度的计算方法。根据上下位关系间的语义距离转换计算两个义原的相似度。从国内外研究情况可知，大体有两类计算词语间语义相似度的方法，一类是根据上下位和同义关系，依赖结构层次关系形成的语义词典，采用概念距离方法计算。一类是利用词汇的上下文信息的概率分布，依靠大规模语料库统计，参照计算词汇间语义相似度。语义词典的一般结构是一个或多个树状的层次结构，利用语义词典计算相似度无法包括未登录词，故而需要考虑语义词典的扩展问题。 易丽萍等[6]提出了一种基于知网、面向语义、可扩展的汉语词语相似度计算的新方法，解决了未登陆词之间

[1] 王利局. 基于语义分析树核的句子相似度计算[D].大连：大连理工大学,2008:4-5.
[2] Chatterjee N. A statistical approach for similarity measurement between sentences for EBMT[C]//Proceedings of Symposium on Translation Support Systems STRANS-2001. 2001.
[3] ResnikP. Semantic Similarity in a Taxonomy: An Information-based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research, 1999(11):96-130.
[4] 梅家驹, 竺一鸣, 高蕴琦等.同义词词林[M]. 上海:上海辞书出版社出版, 1983:1-263.
[5] 刘群, 李素建. 基于《知网》的词汇语义相似度计算[J].中文计算语言学,2002(02):59-76.
[6] 易丽萍, 竹勇, 雷小春. 知网在词语相似度计算方面的应用[J]. 信息技术与信息化, 2005(01):24-26.

万方数据

及未登陆词与登陆词之间的相似度计算问题。但该方法对语义词典的依赖性较大，其计算效果受语义词典的词汇量、组织方式、概念的表达方式、更新速度等多种因素的影响。此外，词汇相似性受人为主观因素影响较大，因此需先对词汇间的语义距离进行计算，再利用相应的转换公式将其转化为对应的相似度值，这个数值越大则说明相似度越小。

采用统计的方法在词语层面依靠语义进行相似度计算，其实是依据进行相似度计算的对象处在相同的上下文环境中，且两者的语义相似这样的假设上开展的。语料的数量规模要求较高，并以上下文信息作为在此层面上的相似度计算的依据 [1]，实现精确地度量词语间的语义相似度。关毅等[2]利用统计技术计算词语间的相似度。但采用统计的方法进行相似度计算，其实验效果与训练所采用的语料库关联较大，且整个实验产生的计算量较大，计算方法也相对较复杂。此外，数据稀疏性的大小以及数据噪声的程度对实验结果的干扰会较大。

2.4.3 组块相似度计算相关研究

组块分析不要求得到完全的句法分析树，只需识别某些相对简单的但在句子组成中意义重要的结构成分，这些被识别的结构即是组块。组块分析的结果是每个组块对应的子树集合，加上缺失的依附关系就可构成完整的句法树。借助于组块分析将可简化句法分析，便于句法分析大规模地应用于文本处理中。Abeny 最早于 1991 年提出了组块概念[3]，此外，句法块相关概念的引入，可从识别块、块内结构分析、块间关系分析三个阶段进行句法分析。识别块可识别句中的所有块；块内结构分析可将句法结构赋予块内部结构；块间关系分析则可组合各块成句法分析树。

组块的划分可准确定位子问题，采用合适的语言模型及策略进行分析，块内结构分析和块间关系分析可大大提高分析效率，降低分析难度。组块相似度分析方法主要有两类，一类是依靠词语相似度分析，采用句子相似度方法计算。其中，李素建[4]利用《同义词林》及《知网》，定量计算词语相关度和相似度，结合句子相似度，并赋予较高的权重于中心词来计算组块相似度。。骆正华等[5]探究语块相似度计算，甚少考虑中心语及修饰词的作用以及组块类型差异上的特点，正是由于组块类型的差异导

[1] Miller G A, Fellbaum C. Semantic of English. In Lexical and Conceptual Semantics[J]. Amsterdam, 1991:197-229.
[2] 关毅, 王晓龙. 基于统计的汉语词汇间语义相似度计算[A]. 哈尔滨工业大学计算机科学与技术学院、清华大学智能技术与系统国家重点实验室.语言计算与基于内容的文本处理——全国第七届计算语言学联合学术会议论文集[C].哈尔滨工业大学计算机科学与技术学院、清华大学智能技术与系统国家重点实验室,2003:7.
[3] Abney S. Part-of-Speech Tagging and Partial Parsing. In: Ken Church, Steve Young, Gerrit Bloothooft(eds.), Corpus-Based Methods in Language and Speech[J]. An ELSNET volume, Kluwer Academic Publishers, 1997:26-28.
[4] 李素健. 汉语组块计算的若干研究[D]. 北京:中国科学院研究生院(计算技术研究所), 2002:1-24.
[5] 骆正华,樊孝忠,刘林,龚永罡.基于 E-Chunk 的问句实例分析系统[J].北京理工大学学报,2005(01):63-66.

万方数据

致组块中各部分对整体的贡献不同，为此并不适宜用句子相似度衡量组块相似度。夏天等[1]充分考虑组块类型差异上的特征，对中心词以及修饰词给予不一样的权重，选取上述两者相似度值的平均值数作为组块层面的相似度。此方法区别对待组块成分并考虑组块类型差异，具有合理性。但该方法权重设定的主观性较强，算法的优越性不突出，且很难区分界定中心词和修饰词。

2.4.4 句子相似度计算国外相关研究

国外的句子相似度计算较多，主要集中于英文句子中，但仍可借鉴有效算法和研究方法。。Nirenburg[2]等采用基于词组合法的相似度程度的度量机制的三重匹配，即将切块、重组等及整句级匹配的两类串匹配方法。罚分制是上述两种方法的核心，其总罚分由句中各单词罚分值汇总所得。Carroll 和 Sumita 均提出了“距离”或“相似度”两者比较的衡量机制。Carroll 采用三角形法在两个测试点间进行句子相似度计算，该两测点分别是被比句子的长度和内容，并用“费用”来表示误匹配。。Sumit 等提出词间距离及其相应权重的乘积和可度量语言片段的距离的相似度计算方法，该方法基于单个词。Niladri[3]提出了一种采用多重回归技术确定相关计算系数的相似度方法，其基于线性模型。在该方法中，句子的相似性或非相似性均由非相似成分叠加作用构成。Chiris 等人[4]采用隐含语义索引的方法计算句子的相似度。Lambros[5]-[6]等提出通过兼容句子的内容计算和在表层方面的表层结构计算句子相似度的方法。该相似度计算系统，采用具有分级结构的动态规划技术，在长度不一的两个句子间用动态规划算法进行语句相似度的计算。Bär D.等人[7]先将句子进行分词，并将词汇从构成成分的角度分成构成与非构成成分这两种类型，这是基于词汇本身的内在属性。采用不同的方法对两种不同类型的词汇采用不同的处理方式，前者进行相似度计算，后者则用其他模型计算，最终将两种类型计算结果进行加权组合，得到完整句子的相似度。

[1] 夏天. 中文信息处理中的相似度计算研究与应用[D]. 北京: 北京理工大学信息科学技术学院, 2005:1-35.
[2] Nirenburg, S. C. Domashnev and D.J.Grannes. "Two Approaches to Matching in example-Based Machine Translation."[J]. Proceedings of TMI-93, Kyoto, 1993:13-17.
[3] Chatterjee N. A statistical approach for similarity measurement between sentences for EBMT[C]. Proceedings of Symposium on Translation Support Systems STRANS-2001. 2001:30.
[4] Ding C H Q. A similarity-based probability model for latent semantic indexing[C]//Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.ACM,1999: 58-65.
[5] Lambros Cranias,Harris Papageorgiou,Stelios Piperdis. A Matching Technique in Example-Based Machine Translation[J].In Proceedings of 1994 COLING,1994:100-104.
[6] Yang Q X,Yuan S S,Zhao L,et al. Faster Algorithm of String Comparison. Pattern Analysis and Applications[J].2003,6(2): 122-133.
[7]Bär D. A Composite Model for Computing Similarity Between Texts[J]. 2013: 23-30.

万方数据

2.4.5 句子相似度计算国内相关研究

尽管中国的中文方面的句子相似度计算相较于国外起步更晚，但中文信息处理技术和手段等的发展，推动中文汉语句子对齐算法的发展，产生了适用于不同场景的中文句子相似度计算方法。主要有基于词语相似度、结合结构信息相似度、基于语义相似度以及基于向量空间的四类汉语句子相似度计算方法。(1)基于词语相似度的句子相似度计算。龙昊等人[1]在句子相似度计算中采用传统的词汇比对的方法。陈利人等人[2]则从语义和语法功能两方面的异同上对汉语句子进行句型的划分，提取句子骨干部分以及语义构成部分，计算不同构成部分的相似度，进而提出一种用句子成分模式的相似度代替句子相似度的计算方法。张民等人[3]考虑词汇共现现象，结合词汇共现和词汇间位置关系，进而发现句子中词汇的构成方式在一定程度上影响相似度计算的结果，其是通过对上述位置相关度进行比较分析，可改善句子层面的相似度计算效果。鉴于进行对比的两个句子中的单词数量不同，且句子长度的差异较大，为此，两句子的各单词间可能存在相关关系，这种相关度可通过加权处理其相对位置实现。该方法不仅多角度地综合考虑了同级词汇的相关性、词汇共现现象、后向叠加评价值，而且采用递归类算法，对句子级的相似度进行计算，此外还可实现“亚句子”级的相似度计算。(2)结合句子结构信息的相似度计算方法。李彬等[4]提出一种依靠语义层面的依存关系的中文句子相似度计算方法，该方法结合了句子结构信息。穗志方、俞士汶[5]响应在句子相似度计算以及定义两方面的特殊要求，提出骨架依存分析方法，该方法依靠骨架依存树这种在语句层面的相似度计算模型，折中地考虑定义与计算两方面的因素进行句子的相似度计算。杨思春等[6]在进行句子相似度计算时，借鉴语法功能角度的句型模式，其需含有部分格语义。该模式与语法层面的句型相似，但相较于抽象的句型要更详细，其不仅可以在句子层面上表达基本特征语义，而且可放映骨架的主要成分构成。因此，当需要从句子的定义和计算角度进行相似度计算时，可根据进行对比的句子的句型模式进行直接比较，用以衡量两个句子的相似度差异。也就是用模

[1] 龙昊, 蔡东风. 一种基于词匹配的句子相似度计算方法[A]. Northeastern University、Tsinghua University、Chinese Information Processing Society of China、Chinese Languages Computer Society, USA.Advances in Computation of Oriental Languages--Proceedings of the 20th International Conference on Computer Processing of Oriental Languages[C].Northeastern University、Tsinghua University、Chinese Information Processing Society of China、Chinese Languages Computer Society, USA:,2003:7.

[2] 陈利人, 陈群秀. 基于实例的日汉机器翻译方法中的句子相似度计算研究[C]. //全国第三届计算语言学联合学术会议论文集.北京:天津医科大学, 1995: 257-263.

[3] 张民, 周明. 一种汉语句子间相似度的度量算法及其实现[C].//全国第三届计算语言学联合学术会议论文集.北京:哈尔滨工业大学%天津医科大学, 1995: 152-158..

[4] 李彬, 刘挺, 秦兵等. 基于语义依存的汉语句子相似度计算[J]. 计算机应用研, 2003(12):15-17.

[5] 穗志方, 俞士汶. 基于骨架依存树的语句相似度计算模型[C]. //1998 中文信息处理国际会议论文集(ICCIP'98), 1998: 458-465.

[6] 杨思春, 程节华. 一种基于模式的汉语句子相似度计算方法[J]. 微型机与应用, 2001(8):52-53.

万方数据

式相似度的判断表示句子层面的相似度判断。(3)基于语义相似度计算方法。陈群秀等人[1]结合句子语义相关性和事件框架的相关性，充分融合得到了一种在句子层面的依赖事件框架进行相似度计算的方法。车万翔等人[2]改进了编辑距离，并采用 How Net 进行语义距离程度上的判定，其从词义的角度进行考量，规避了仅从字符串或词汇匹配角度片面地对编辑距离进行衡量的方法。侯红旭等人[3]采用词对齐的匹配方法，在匹配的数量、匹配的长度两个角度考量，利用汉蒙翻译软件进行试验操作，进而将得分最高的一项作为该平台效果最好的参考译文。 (4)基于向量空间的句子相似度计算方法。刘挺等人[4]使用了一种结合了语义信息，采用 VSM 并利用 TF-IDF 进行权重权重的设定，结合上述向量计算方法，对向量各维度进行赋值，此方法可实现特定类型句子效果较佳的相似度计算，例如在问句这一句型方面进行相似度计算。

[1] 梁晗，陈群秀，吴平博. 基于事件框架的信息抽取系统[J].中文信息学报, 2006, 20(2): 40-46.
[2] 车万翔，刘挺，秦兵. 基于改进编辑距离的中文相似句子检索[J]. 高技术通讯, 2004(07): 15-19.
[3] 侯宏旭，刘群，那顺乌日图. 基于实例的汉蒙机器翻译[J]. 中文信息学报, 2007(04): 65-72.
[4] 秦兵，刘挺，王洋等. 基于常问问题集的中文问答系统研究[J]. 哈尔滨工业大学学报, 2003, 35(10):1179-1182.

万方数据

第三章 《毛诗正义》文本分析

《毛诗正义》兼容了毛传以及郑玄的郑笺这两家学说，并且是以《义疏》《述义》作为疏解文意的底本，在638年，唐太宗时期，孔颖达被任命主持编撰。《毛诗正义》涵盖了丰富的训诂知识，其有助于本文采用正则表达式的方法进行对齐研究，为此，探究《毛诗正义》的训诂语言特征很有裨益。

3.1 《毛诗正义》行文结构

要实现《诗经》与《毛诗正义》的句子对齐研究并不是一蹴而就的，需要进行诸多的预处理工作。实际上，尽管《毛诗正义》的电子善本在网络时代可以很容易的获取到，并且可以利用相关的技术被计算机识别，存储成所需的格式，经过这样一系列的处理，其文本的质量本身是比较高的，但由于其在行文方式上的差异，以及句读等的不一致，在进行研究前，仍需对获取到的电子善本进行数据预处理。数据预处理主要有以下几方面：《毛诗正义》文本的格式化及编码处理，不同版本语料间的数据校对，《毛诗正义》文本的全文断句，注疏文本的全文对齐处理等等。

在具体介绍本研究的预处理工作前，先探讨《毛诗正义》的行文结构，或称文本组织方式。通过观察可发现，其在行文结构上属于“夹引夹议”文体，其经、注、疏合一，先引用原始典籍中的一个或多个小句，其后紧跟注文；再引用原始典籍中的内容，之后再紧跟注文；一小段篇幅的原始典籍注释完之后才会出现“疏”，并且“疏”是独立存在的一段或多段内容，以“[疏]”字样作为起始的标注，其先疏释原始典籍的内容，再疏释相应的注文[1]。从体例上看，《毛诗正义》全文均对《毛传》《郑笺》以及经文进行阐释，但是《正义》是其主要部分，而仅对经文和注文标注相应经、注的起止，在进行内容疏解前，孔颖达为申明所疏释的对象，经常选取经文的首尾几字以及注文的首尾多字，例如经常用“.....至.....”这样的句式来标明经文和注文的起止句。以《诗经》中的《殷武》这首诗为例，比如下例中的《诗经》原文以及《毛诗正义》注疏文献对原文的相应注释内容的具体行文结构。

《诗经》原文：

有截其所，汤孫之緒。

《毛诗正义》：

[1] 马创新. 注疏文献的结构化知识表示[D].南京：南京师范大学, 2014:7-32.

万方数据

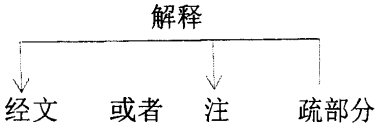
有截其所，湯孫之緒。箋云：緒，業也。所，猶處也。高宗所伐之處，國邑皆服其罪，更自勑整截然齊壹，是乃湯孫大甲之等功業。○處，昌慮反。下同。

[疏]“撻彼”至“之緒”。…… 眾者，深入之意，故爲深也。“裒，聚”，…… 大武丁之伐與湯同”。

通过上例的具体行文结构可知，《毛诗正义》主要是由两部分组成，一部分是引用原典籍经文中内容，也就是引文；一部分是对典籍中的内容进行注释的部分，也就是所谓的注疏，可对注疏这部分进行进一步细分，划分为“注”、“疏”两部分。《毛诗正义》是先有注文后有疏。

另外，《毛诗正义》电子善本存储结构方面，引用原典籍的内容与“注”的内容是被放置在相同的段落中，两者没有明显的分界，为此，并非每部分引文的后面都有“注”。“疏”则是单独的一部分，其对引文与“注”两部分的内容都有进行相应的疏解，篇幅小的则只有一段，篇幅大的则会存在数段，但“疏”一般都会有，但不排除存在少数的例外情况。“疏”部分的内容相当丰富，有孔颖达《正义》方面的注释，是以“正义曰”作为点明此部分内容的出处；郑玄所标注的注释，是以“郑”或者“郑曰”等字样来注明；《毛传》对应的注释则是以“传”或者“传曰”字样进行标明。此外，在《毛诗正义》的全文中，还申引观点鲜明的训诂学家的对相关的内容的理解，例如许慎。

《毛诗正义》的基本行文结构可大致表示为下图的模式：



3.2 训诂语言分析

《毛诗训诂传》是我国现存创作时间最早、保存最完整、对后世的影响力最大的《诗经》随文训释著作，它作为古文经学训诂的鼻祖，首创故、训、传三体合一的训释体例，对后世训诂学的产生和发展具有深远影响。它所包含的训诂条例多有规律可循，这使其在训诂学方面的研究上具有极高的科研价值。《郑笺》为《毛传》作笺，是对《毛传》的一种补充和订正，基本没有进行删减。《毛诗正义》则是对《毛传》《郑笺》两者共同做的注疏，并创造性地在训诂方面有了诸多的独到见解。

训诂术语的意义与它在使用过程中的形式和功能是密不可分的，较常见的训诂术语主要有以下两类：一类属于抽象的术语，它们是近现代训诂学家总结出来的说明训诂方式的术语，如形训、音训和义训等；另一类属于具体的术语，它们是从古代学习

万方数据

者们所沿用的训释字词的习惯用语[1]。不同的训释对象和训释内容，一般会使用不同的具体术语，对这些具体术语进行总结归类，找出训诂的规律，不仅可指导训诂实践，对于本文实现对齐研究提供了帮助。

本文从训诂术语的角度出发，从其术语、格式、内容、方法和体例这五个方面对《毛诗正义》的训诂语言进行分析，归纳总结已成规范体例的训诂具体术语模式，构建训诂术语模式库，并用正则表达式提炼加工，将其用于探究《诗经》与《毛诗正义》的句子对齐研究。本文探究的训诂术语，是指对《毛诗正义》中的各部分结构成分进行解释说明而使用的专门用语，有特定的含义和较为固定的用法。如：曰、言、指等等，普遍应用于训诂条例中。

3.2.1 训诂术语及训诂格式

《毛诗正义》中的训诂术语根据不同的术语功能有着不同的训诂格式，可将其分为五类训诂术语。第一类是释义或辨析文意之术语，如：为、谓、谓之、之言、者，也、犹、犹言、犹云、亦、称、斥、指、对文、散文、文势、曰、言等。第二类是注音或假借之术语，如：读为、读当如、读曰、读当为、读如、xx 声也、xx 之声等。第三类修辞之术语，如：互言、重言、以兴、喻、以喻、此喻、互文、比、兴、变文、韵句、省文等。第四类是语法之术语，如：辞、语辞、叹辞、语助、xx 之辞、为义、xx 之称、状、貌、然、不为义、xx 之名等。第五类是校勘之术语，如：xx，当作 xx；未闻、xx，当为 xx；文同、衍、误、谬、异文、未详、断章等。

3.2.1.1 曰、为、谓之、之谓

“曰、为、谓之、之谓”都是对事物名称的定义。其中，“曰”和“为”是进行定义确定，辨析词语含义的作用，可翻译成“叫做；称为”，其主要是对名词性质的词汇进行注解，但也有少量被用于对动词性质的词汇注解，“曰”可相当少地应用于注释形容词词性的词汇。当“曰”是多个包含两个同时被使用时，主要是对词汇的意义或语义进行辨析；或者对《诗经》原典籍中出现的相近的词汇进行辨析；或者是对相关的有一定关联的事务进行辨析，当处在同一个语义层面上的词汇，可以进行对应的特征区分，同理，这种用法也适用于“为”。一般的训诂格式是“x 曰 y”，“x 为 y”，x 为被训释的词，y 为用于训释的词。“曰、为”可并用，可见两术语用法基本相同。

要想对事物进行称呼，可用“谓之、之谓”，也就是把某种事物叫做何种名称，而不是直接对事物进行定义。“谓之”意为“把……叫做，称为”，一般的训诂格式是

[1] 方一新. 训诂学概论[M]. 南京: 江苏教育出版社. 2008:3-170.

万方数据

“x 谓之 y”，被释词位于“谓之”的前面和后面的情况均存在，但被释词均为名词。“之谓”表示“.....的意思；指的是”，其训诂格式可以是“x 之谓 y”，被释词可放在“之谓”的前面或后面，也可以是“x，y 之谓”这样的训诂格式，x 是被注释的词汇，y 是用于注释的词汇，且“之谓”通常被置于末尾。

3.2.1.2 谓、言、称、斥、指

“谓、言”是跟随文本的语义，结合上下文环境进行注释，意思为“指的是；意思是”，用于进行语境的梳理，此外还可用于注释词语、短语或者是句子。其一般的训诂格式是“x 谓 y”，“x，谓 y”，“x 言 y”，“x，言 y”，其中 x 是被注释的对象，y 是用于训释的内容。“谓”可与“者”一起使用，形成“x 者，谓 y”格式，“者”之前是被注释的对象，“谓”之后是用于训释的内容，“言”亦有这样的用法，同样能够和“者”连用。由此可见，“谓、言”具有相近的性质，可以用于引出被注释的词语，进而进行句子或者词汇的注释。

“称”的训诂格式有：“y 称 x”，“x，y 之称”，其中 x 是被注释的词，y 是用于训释的词，用以点明事物的称呼或命名。“斥”和“指”在意义上是很相近的，表示指明被注释的词所指向的具体对象，含义是“指的是谁”，“斥”的训诂格式基本是“x，斥 y”，x 是被注释的词，y 是用于训释的词。通过分析《毛诗正义》可发现：“指”相对于“斥”，前者的指示对象范围更广，可以涵盖各种普通的人和事物，但“斥”一般是用来对社会地位更高、更尊贵的人进行身份的点明。

3.2.1.3 犹、犹言、犹云、亦、亦称

“犹、犹言、犹云”意为“如同；等于说；和.....一样”，可注释同义词或近义词，点明通假字，也可解释句意。一般的训诂形式是“x 犹 y（也）”，“x 犹云 y”，“x 犹言 y”，“x，犹言 y”，“x，犹 y”，“x，犹云 y”，x 是被注释的对象，y 是用于训释的内容。“犹”可以被用作比喻修辞，在这种用法上，其与“犹如”是一样的。

“亦”是用于注释词汇，对被注释的词汇在上文中同样有出现在词语意义方面相近或者是可相关联的内容时进行点明“亦”与“犹”在上下文的承上启下的作用上的用法是一样的，并没有差异[1]。“亦”和“称”可以搭配，构成“x 亦称 y”的训诂格式，x 是被训释的词，y 是用于训释的内容。

3.2.1.4 读为、读曰、读如、读当如、读当为

“读为、读曰、读如、读当如、读当为”都是可以被用作指代说明进行音调标注或者是有通假的字，但具体的使用场景，不同的训诂学专家对其有着不同的见解。但

[1] 陈炳哲.《毛传》、《郑笺》训诂术语比较研究[D]. 北京：首都师范大学, 2005:1-67.

万方数据

本文不探讨其具体用处，而是针对其使用情况，确定被释对象与训释内容之间关系的考究。“读曰”用以标明通假和解释古今字，可表示为“x 读曰 y”的训诂格式，x 是被释词，y 是训释词。

“读当为”是“读为”的一种转变表现形式的方法，这两种表达的意义是相同的，用法也类似，都是对通假字等进行标注。训诂格式可表示为“x 读（当）为 y”，x 是被释词，y 是训释内容。“读当如”视为“读如”的变体，用法相近，主要用于标明被释对象的读音和释义，也可指出被释词相关的语境及点明通假。一般以“x 读（当）如 y”格式表示，x 是被释对象，y 是训释内容。

3.2.1.5 辞、貌、叹辞、语助、然、意

“貌”意为“.....的意思；.....的样子”，放在形容词的后面或者是副词的后面，可表示成“x，y 之貌”，“x，y 貌”的训诂格式，x 是被注释的词，y 是用于训释的词，对被注释的词所处的状态进行点明，用于训释的词主要是对当时被注释的词的外形或面貌等方面进行描述。“貌”还可与“者”及“为”搭配，表示成“x 者 y 之貌”，“x 为 y 之貌”的训诂格式。

“然”经常被放在形容词词性、副词词性或具有动词词性的词汇后面，用来对当时的存在的事物或产生的相关的动作所处的状态进行描述[1]。“然”与“貌”相似，都可标注形容词，通常置于被注释的词或者短语的后面，对被注释词的状态进行点明。

“意”多用于表明所处状态的形容词或副词，训诂格式有：“x，y 意”，“x，y 之意”，x 是被注释的词，y 是用于训释的内容。“意”是可与“貌”及“声”互补的术语。

“辞、叹词、助语”多为训释虚词，训诂格式有：“x，辞也”，“x，y 辞”，“x，叹辞（也）”，“x，y 之辞”，“x，y 嗟叹辞也”，“x，y 之辞（也）”，其中，x 是被注释的词，y 是用于训释的内容。上述格式主要是用于点明被注释释的虚词的意义或语法功能。

3.2.2 训诂内容

《毛诗正义》的训诂内容十分丰富，有辨正用字、辨明文字异形、解释词汇的语义或意义、辨析词义、分析语法、申述修辞、串讲文意、校订旧文、考订读音、辨别剖析音义关系、考稽典故、引史证诗等方面。

3.2.2.1 辨正用字、辨明文字异形

[1] [元]卢以纬著，王克仲集注.《助语辞集注》[M]. 北京：中华书局, 1988:1-186.

万方数据

古人训释经传，凡举文字扞格难通者，均作辨析。文字随时代的变迁，形体发生了变化，从而产生了相同的字但表现出不同的形体，相同的词存在着不一样的字的现象。为此，训诂学家把不同形体的字和同词异字加以会通，说明某字即某字[1]。

“读曰、读当曰、读当为、读如、读为”等训诂术语，通常被用来对通假字进行点明，并标注出对应的正字，用以训释古今字。“x，古 y 字”“x，古文 y”，“x，今 y 字”，这组可用于辨析古今字。例如：《毛诗正义》对《小雅・鹿鸣》：“視民不恌”一句，疏部分的内容：[疏].....此云“視民不恌”，謂以先王之德音示下民，.....故鄭辨之：“視，古示字也”。

“之言”，“之为言”对通假字进行标明，用声音相同或相近的词语训释。例如：《大雅・灵台》：“於論鼓鍾”，箋云：論之言倫也。[疏].....〇正義曰：以倫理之字宜爲倫，故曰“論之言倫傳”。“論”和“倫”是通假字。

3.2.2.2 解释词义、辨析词义、串讲文意

《毛诗正义》中的大部分训诂术语是用于对文中的词汇进行注释，并对上下文语境下的相关内容兼并说明。“者，也”，“为”，“谓”，“言”，“谓之”，“犹”，“亦”，“犹言”，“之谓”，“称”，“斥”，“指”，“曰”等主要用于释义。

“对文”，“别名”,“散文”，“通言”，“异名”主要用于对词汇的含义进行辨析。当经文和传存在相同的意思时是相对使用的，意义有别，但单独使用时，意义相同。为此，孔颖达首先提出了“对文”和“散文”这两个概念，它们主要是对文中出现的同义词进行辨析，两术语配合使用，当一组同义词被用作“对文”，则可区分出同中之异；如果用作“散文”，则可确定两者的共同意义，故“对文”则别，“散文”则同。

“别言”，“倒文”，“文势”，“文次”主要用于文本的上下文语境以及相关词汇的意义进行训释，对相关词汇在文中出现的顺序进行注释。“文势”是指上下文所处的语境，文势观是孔颖达的语义观，同时也是《毛诗正义》解析语义方法的理论基础。《诗经》重章叠句现象突出，在对词语进行注释时，要系统地考察词语在整个经文中出现状态等情况，进而联系上下文比较分析，释义伴随对应的具体的诗文来确定。“文次”是文章中不同内容布局的次序，不仅要使得全文中各部分在词义上没用冲突，具有一致性，还不能够违背常规，要遵循客观进程，或者遵从相应的思维表达方面的习惯。“倒文”则是语义顺序的颠倒，也可称作“文倒”、“倒言”、“倒”，用来对语序进行注释，可对引用原典籍的内容、注文等对应的语序进行注明[2]。

[1] 张艳.《毛传》、《郑笺》对《诗经》训诂之比较[D]. 兰州: 兰州大学, 2007.45-47.
[2] 党代莉.《毛诗正义》训诂术语研究[D]. 兰州: 西北师范大学, 2011:1-55.

万方数据

3.2.2.3 分析语法

《毛诗正义》的训诂术语考究词语的词性，涵盖了对不同词性词语的判定，并对实词和虚词进行区分。在进行注释的过程中，对省略的句子成分进行补充。通过对语序进行变换，如动词和宾语之间的顺序，中心词和修饰词之间的顺序等，并点明结构，便于进行句子分析，以助理解。

“辞”，“禁辞”，“叹辞”，“语助”，“应辞”，“助句辞”可以用来表示助词，可指明被释词是虚词，用以区分实词和虚词。“为义”，“于义无取”等术语可表示虚词的概念，“为义”则表示实词的概念。

“……之称”，“……之辞”，“……之名”这三个术语通常用于解释实词，可交替使用。“貌”，“状”，“然”则主要对动词词性、象声词词性和形容词词性等词汇的后面，用以表示“……的样子”。“体”，“用”两个术语专门用以对名词的解释，“体”是从事物的形状、实体、质地的角度解释，“用”则是从用途、作用方面进行解释。

3.2.2.4 申述修辞

“互言（互辞）”，“此喻”，“重言”，“兴”，“连言之”，“喻”，“反言之”，“以兴”，“以喻”，等都是可用于注明修辞方式的训诂术语。关于赋比兴，其中的兴体，大多以“兴”、“犹”、“若”、“譬喻”、“如”、“喻”等术语说明，用“以兴”术语的较少。

《毛诗正义》中涉及的修辞方式很多，有“变文”，“兼言”，“互文”，“反语”，“省文”，“协句”，“重复”，“韵句”，“倒文”，“共文”，“并言”，“足句”，“便文”，“甚言”，“婉文”，“直言”，“反明”，“长读”，“指言”，“连言”，“举言”，“逆言”，等。“变文”指的是文句前后相对之处意义相同而用词有异，以避重复，前后词的词性词义大致相同，形似对仗。“互文”较为隐晦，是对同一内容交互说明，以点明词义。“倒文”是颠倒语序，以协韵。“反语”顾名思义是用反义释词。“连言”简略提及作者行文过程，原本的用意是对其中一种事物，但附带的在某种缘由下，对另一种事物进行略带[1]。

3.2.2.5 校订旧文

“衍”，“未详”，“误”，“文同”，“谬”，“误衍”，“异文”，“未闻”，“断章”均可用以校勘文句。“衍”指在传递抄送过程中、刊刻过程中出现误差而增

[1] 刘卫宁.《毛诗故训传》、《毛诗笺》与《诗集传》训诂比较研究[D]. 广州：暨南大学, 2005:2-40.

万方数据

加产生的多出来的文字。“误”及“误衍”则指抄写、刊刻错了的字，或者是引文错误。“谬”则是指出言语的错误的术语。

“异文”产生的原因是版本的差异导致对相同的一种事物却有着不一样的说法，或者出现在不同的书籍中，来源的出处不一致而导致的差异。“文同”则是指同一个事物出现在同一本书中，但却是在不同版本或者是不同的书中。“未闻”和“未详”则是治学过程中呈现出的一种态度，但对某种看法或者事物不确定时，暂不对这部分内容进行解释，或者是表明不确定，用存疑的方式对待。

“当为”和“当作”主要用于揭示因为音调相近或者是字的形态上字形相近而造成的文字讹误，还可用于指明假借。构成“x 当为（作）y”的训诂格式，x 是被注释的词，y 是用于训释的词。

“或作”和“作”可用来校勘文字的异同，文字不同而记载的事物相同，彼此音读有相通之处。构成“x 或作 y”，x 是被释词，y 是训释词。

3.2.2.6 考订读音、辨析音义关系

“读为、读当为、读曰、读当如、读如”这些术语可用以拟音或者表明假借字。“古音 xx 同”，“xx 声同”，“读声相似”三个术语是用古音或方言来指明因为发音相同或发音相近而可以通用、假借的字。“某古文作某”，“某今某字”，“某古某字”，“古之某字或作某”都是用于说明古今字的术语。[1]。

在辨析音义关系方面，“破字”和“易字”都是表示修改了的字，是针对《毛传》《郑笺》在对词语进行注释上存在的差异进行说明。“字虽异，音实同”，或作“字虽异，音义同”可用以指明异体字或假借字，需要谨慎区分。“古字通用”是讲文字通假的训诂术语。通假是指两个字在音调上相同或者音调相近，但所包含的意义却不相同，但由于音调相近或者是音调相同而被借用时，产生了假借义[错误!未定义书签。]。

3.2.2.7 考稽典故、引史证诗

由于名物、制度、风俗习惯等的不明，古书难懂。故致释名物、出典、制度、故实、古语等是《毛诗正义》不可或缺的训释内容，涉及的范围很广，涵盖服饰、礼法、马制、货币制、税法等方面。《毛传》的在每一篇的开头，都有一个简短的小序用于对每首诗的题目进行注解，主要是对整个诗篇的主旨，所处的时代背景，创作的作者等进行阐释。小序是一次全面系统的概括，对于全文中出现的每一首诗，这是后人研究《诗经》不可或缺的素材，后世虽然对小序存在异议，甚至是抨击，但仅只是对小序进行修改，并没有完全抛弃小序。《郑笺》和《正义》都是借鉴了小序，进行地独

[1] 陆德明.《经典释文》[M]. 上海：上海古籍出版社,1984:3-30.

万方数据

立研究，重新对《诗经》的相关内容进行注释。在注释词语时，训诂学家会结合当时的时代背景，旁引史实，以助理解。

3.2.3 训诂方法

在训诂方法上，《毛诗正义》除了用传统的“三训”法来对词语的意义进行探究，还从语法和修辞的角度通解训诂，形成了在语法和修辞两个角度进行注释的方法，提出并实践了“文势”说，依据语境解释词义的训诂方法。

3.2.3.1 重视结合典籍注释字词

《毛诗正义》广征博引，在注释中所征引的著作包括了经、史、子、集各个门类，多达 292 种[1]。其中引用礼学方面的书籍最多[2]。《毛诗正义》中多处依照《诗序》来揭示主题，用以注解《诗经》。《毛诗正义》在结合典籍注释字词时，不仅通过征引典籍直接作注，也就是用经文注解经文，并添上自身的注解，而且还会引用《诗经》以外的其他经文或经书的注进行注释，例如《公羊传》等。此外，《毛诗正义》还会引用字书、辞书来解释，诸如《尔雅》《广雅》等。其中，引用《尔雅》及注数量最多，对后世语言学家对雅学的研究产生很大的影响。

3.2.3.2 因形求义-形训

形训是根据对汉字形体结构对字义进行解释分析的方法。根据字形来判读字义最直接的用法就是对词的本义或者是最原始的出处进行探究[3]。《说文解字》是现存的最早使用字形探索字义的方法。形训的方式主观性很强，不同的训诂学家有着不同看法，为此，这种形训的训诂方法在《毛诗正义》中出现的次数更少，其出现较多的是根据声音和词义进行训释。

3.2.3.3 因声求义-声训

声训是通过声音的线索进行训诂，利用声音（同音字）或相近的（双声叠韵字）词，解释词义的方法。“因声求义”不仅要追求语源，还要从语音入手，突破文字形体的束缚，探求本字。洪诚先生曾指出：“不懂得因声求义，就不懂得前人在常见的字下注音的意义，就必然不能对所注的词义有正确的理解[4]。”

声训的方法主要是用于探求语源和说明假借。推求语源即指寻求一个词或是词所代表的名物得名的由来，并非对词义的直接解释，而是通过探求语源，使得对词或词

[1] 韩宏韬.《毛诗正义》研究[D]. 济南：山东大学, 2007:2-91.
[2] 高華娟. 《毛詩正義》注釋研究[D]. 重庆：西南大学, 2013:1-69.
[3] 楊琳.《訓詁方法新探》[M]. 北京：商務印寿館, 2011:3-272.
[4] 洪诚.《训话学》[M]. 南京：江苏古籍出版社, 1984:1-40.

万方数据

所表达的名物有更准确深入的理解，这与一般的词语解释相比，推求语源是更深层次的词语解释。

可用声同或音近的词直训，或者利用义界中心词来解释语源。义界中心词是指与被释词音近义通的关键词。

就声音而言，同音相训、双声相训、叠韵相训、双声叠韵相训都是声训直训的几种方式。同音相训是训释字与被释字读音相同。双声相训是以声母相同的字为训。叠声相训是以相同韵部的字相训。双声叠韵相训则是以声母韵部皆相同的字为训。

就字形而言，声训的方式有异字相训和同字相训两类。其中，异字相训又可分为形体完全不同者和形体部分相同者。形体部分相同者利用形声字的声符来进行训释。

3.2.3.4 义训

“凡是不以声音为训，不以字形为训的，都可归入义训之中。[1]”义训的方式多样，有如下七种：

（1）以今字释古字。诗为古字，字义较简单，后辗转引申，一字多义的现象相当普遍，故于本字加偏旁作区别，使文字各有专职，形成了古今字或区别字。例如：《周南·兔置》：：“赳赳武夫，公侯干城。”傅：“干，扞也。”其中，“干”为古字，“扞”为今字。

（2）以今语释古语，以通语释方言。古字因时地衍移变迁，或读音稍变，或意义少别，于是另造一字，新造字与原字便是转注关系。例如：《大雅·大明》》“大邦有子，倪天之妹。”傅：“倪，磐也。”磐，今语；倪，古语。

（3）递训。乙训甲，再以丙训乙，一训接一训，以后面的训释补充前面的训释。例如：《小雅·小弁》“弁彼鸒斯，歸飛提提。”傅：“鸒，卑居。卑居，雅鳥也。”

（4）反训。反训实为兼言正反，反言证之。例如：《小雅·車攻》：“徒御不警，大危不盈。”傅：“不警，警也。不盈，盈也。”

（5）列举为训。列举被释词所包含的内容训释。例如：《秦風·權輿》“於我乎，每食四簋。”傅“四簋，黍、稷、稻、粱。”也可用词组或句子解释被释字、词。例如：《大雅·皇矣》：“临冲闲闲，崇墉言言。”传：“闲闲，动摇；言言，高大也。”

（6）引书为训，以经解经，为读经最善之法。可分为明引和暗引两种，明引指明出处，暗引不说出处。

（7）推因。推求语源，根据被释词语音线索，探索其音义的由来，关联由同一语源派生的同源词。这也是声训的一种方式。例如：《大雅·行葦》：“舍矢既均。序

[1] 路广正.《训话学通论》[M]. 天津: 天津古籍出版社, 1996:141.

万方数据

賓以賢。”笺：“舍之言释也。”“舍”有“放弃”的意思，“释”有“放下，解脱”的意思。“舍”和“释”属于声母双声为训，构成同源关系。

3.2.4 训诂体例

《毛传》兼有故、训、传三种体例，除了对字和词进行解释外，还进行义理方面的注解。每一章的开头都对其采用的体势进行注明，并在每一句的后面进行注解，并且列举相关的典故，辅助解析文章的意义，并对每章的内容进行辨析，在每一篇的末尾，都标有章节的名称以及每章中含有的句子的数目。《郑笺》对《诗经》原典籍的内容进行注释，且对《毛传》进行注解。郑玄先注释《诗序》时，毛亨没有提及的内容进行补充，注解不清楚的地方剖析细致辨明，在注解的简要的地方详细地进行解释，观点不一致的地方没有进行篡改，而是单独表达自己的观点。故《毛诗正义》于每篇的末尾都沿袭了《毛传》的主要思想，对每章每句的都进行了出处的标注，以“传曰”、“笺曰”等字眼来表示。孔颖达探讨了篇、章、句作为诗形式要素的起源、演变和含义。篇名和章名是一直都存在的，在诗产生的时候就有了。“句”古代被称之为“言”，“句”这个说法是在秦汉以来儒学们注释经典产生的。“句”又可称为“局”，可视为是将言语分界，将字联结构成整体形成的“句”。“章”是总括各句意思，其包含在诗体中，辨明情义。“篇”则是铺陈情义，全面表述。

《毛诗正义》中《诗经》原文的句子，以四字句数量最多，不少于二字，有字数为三、字数为五、字数为六、字数为七、字数为八，字数以八个字为最高的上限，字数的变化取决于所处的情景的不同而发生变化。纵观《诗经》的篇章布局，一句不能成章，至少二句，多则不限，多的可有三十八句[1]。篇的大小因表达的内容和体制的不同而有大小之分。

《毛诗正义》对全文篇章进行划分，拆以风雅颂各部分，每部分内包含多章，细分各篇，以大标题概括。例如：《甫田之什·賓之初筵》篇末后谓：“《賓之初筵》五章，章十四句。”后谓：“《甫田之什》十篇，三十九章，二百九十六句。”用以概述《甫田之什》篇章句情况。

[1] 杨金花.《毛诗正义》研究[D]. 石家庄：河北大学, 2007:12-116.

万方数据

此页不缺内容

万方数据

第四章 语料库构建

本文探究的是《诗经》与其注疏文献《毛诗正义》的句子对齐研究，其对齐结果可用于构建基于《诗经》的诗歌体典籍的句子对齐语料库，该语料库融合相关分词词表、分词语料、分词结果以及对齐实验数据等。为实现《诗经》与《毛诗正义》的句子对齐，需先进行一系列的数据预处理工作，其中包括对各类语料的数据预处理、分词词表构建、分词模板构建等，还需规范对齐实验及自动分词实验的标准，包括对齐结果展示的外观布局、标注的规范等。此外，句子对齐需要进行文本分词，为此，需进行自动分词处理。

4.1 数据预处理

为实现本文的句子对齐研究，构建对齐语料库，需进行一系列的数据预处理。数据预处理包括对获取的各类语料进行数据校对，为后续分词实验和对齐实验构建相应的词表，以及进行语料标注。

4.1.1 数据校对

本文获取的《诗经》和《毛诗正义》的语料版本不一致。《诗经》的原始语料是繁体的，来源于民国 63 年，艺文印书馆出版的朱熹《诗集传》。《毛诗正义》则源于阮刻本《十三经注疏》，通过网上爬虫获取的简体版数据。自动分词实验是基于经过手工分词得到的《诗经》语料，总字数是 38443，据此获得了《诗经》词表，总词数是 4046。该语料是南京师范大学陈小荷研究团队的成果，由具有古文知识背景的研究生手工切分完成，其原始文本来自香港中文大学的汉达文库[1]。北大整理本《毛诗正义》[2]仍是以十三经注疏作为基础，并施加标点，虽然学术界对其存在争议，但本文仅将北大整理本作为繁体校对的标本，不进行句读校正。

为统一本文研究语料的一致性，均使用繁体版的数据进行研究，为此需进行简繁体转换以及繁体字校对工作。此外，实现句子对齐研究需要进行文本分词，本文利用《毛诗引得》领域词表，采用机器学习的方法对文本进行自动对齐。需对《毛诗引得》的数据进行校对处理，使之与纸质版数据相符。

[1] 汉达文库[EB/OL]. [2005-04-13]. http://www.chant.org/.(CHANT-Chinese Ancient Texts.
[2] 龚抗雲，李传书，胡渐逵等. 北大整理本_毛诗正义(全三册)_汉毛亨传;郑玄笺;唐孔颖达疏[M]. 2000:9-11.

万方数据

4.1.1.1 繁体字校对

由于获取的《毛诗正义》数据是简体版，科学研究应遵从原经，使用繁体数据，因此需要利用程序将其转换成繁体版。但一个简体字通常对应两个或多个繁体字，这就给本文的句子对齐研究带来了不便。为此，转化后的繁体版《毛诗正义》需与其他语料进行繁体字校对。本文选取北大整理本《毛诗正义》作为校对的标本，将《诗经》《毛诗正义》及手工分词的《诗经》三份语料与北大整理本进行比对，将各版本繁体字不一致之处进行校正。使得《诗经》与繁体版《毛诗正义》中引用《诗经》经文的内容在繁体字上不存在差异，便于后续利用相关的算法进行句子对齐研究。

在校对过程中，对校正的简繁体字进行记录会发现，“于”与“於”这两个字在《毛诗正义》中并存，依具体的情境而定，并非都用“於”。例如：“于彼新田，于此菑畝。”用的是“于”，“旄在地已屬之於干旄”则是用“於”。如下表 4-1 和表 4-2 所示，部分校对记录样例：

表 4-1 繁体字校对记录

繁体字转化后《毛诗正义》繁体字	北大整理本《毛诗正义》繁体字	校正前《毛诗正义》繁体字句子样例
瀋	沈	瀋尺遂反
鏇	旋	還音鏇
戍	戍	遣將帥及戍役
彆	别	雖三章三輩彆行
雲	云	《玉藻》雲
鞦	秋	當至鞦末始平
徵	征	將使我出徵伐
禦	御	禦夫也
於	于	於彼牧矣
廄	厩	仍有在廄供用者
閤	合	季春乃閤纍牛騰馬遊牝於牧
餘	餘	旗音餘
茲	兹	御夫則茲益憔悴
遙	遥	慈遙反
製	制	凡製軍
韏	千	萬二韏五百人為軍
眾	衆	有軍眾者畫異物

表 4-2 多个繁体字共存使用的样例

繁体字	样例	繁体字	样例
于	于彼新田，于此菑畝	於	旄在地已屬之於干旄
礼	礼记	禮	《魚麗》，美萬物盛多，能備禮也。
至	至也	緻	無不攻緻
鱉、鼈	鳥獸魚鱉皆得其所然	鰲	如是，則鳥獸魚鰲各得其所然也。

万方数据

由表 4-1 及附录 1 可看出，简繁体字转换后的《毛诗正义》语料，与其他语料存在多处繁体字不一现象，可见不同版本间繁体字的差异，故进行繁体字校正是很有必要的工作。由表 4-2 可发现，《毛诗正义》中还存在针对表达同一意义的字，使用了其繁体字的不同表现形式，需依具体的语境区分。

4.1.1.2 引得数据校对

《毛诗引得》领域词表是从《毛诗引得》中手工录入的引得词头，总词数是 3143。鉴于《毛诗引得》领域词表与训练语料的数据来源不一致，需进行数据预处理，校对数据中的繁体字。为方便后续的研究，校对以手工分词语料中的繁体字为准，也就在进行引得数据校对时，将《毛诗引得》手工录入词头语料与本文涉及的各类语料均进行统一的繁体字校对，且将获取的《毛诗引得》词头与纸质版的《毛诗引得》词头进行比对，确保数据无误。

《毛诗引得》是由哈佛燕京学社引得编纂处的专家针对《诗经》典籍联合索引标注产生的，故从《毛诗引得》中抽取得到的领域词表融入了专家的词汇知识，其分词的切分原则和本文所采用的南京师范大学手工分词训练语料存在差异，但更具权威性。古代汉语以单字词居多，但典籍分词后更有研究价值的是长词（多字词）。

通过引得数据校对发现，《毛诗引得》中异体字有 113 个。剔除异体字后，对《毛诗引得》词头和《诗经》手工分词语料进行词长词频统计，以便进行更细致的比对。如下表 4-3 所示：

表 4-3 《毛诗引得》与《诗经》分词语料词长词频统计

	引得词频	占总词数比例	手工分词词频	占总词数比例
总词数	3143	-	4046	-
单字词	2830	90.04%	2429	60.03%
多字词	313	9.96%	1617	39.97%
双字词	239	7.60%	1554	38.41%
三字词	20	0.64%	31	0.77%
四字词	50	1.59%	31	0.77%
五字词	2	9.96%	1	0.02%
六字词	2	0.06%	0	0

由表 4-3 可看出，《毛诗引得》以单字词为主，单字词与多字词的比例高达 9：1，多字词中以双字词为主，词长最多为六字，其多字词基本是以诗名和大标题构成；《诗经》手工分词的语料，单字词与多字词比例近 6：4，多字词中以双字词为主，词长最多为五字。

万方数据

此外，本文还需构建多个词表以满足后续自动分词实验以及句子对齐实验的语料需求。利用手工分词语料，对其进行词频统计并去重处理后，可得到基于手工分词语料的词表，该词表是后续自动分词和句子对齐研究的分词词表底本。以相同的处理方式，可得到基于《毛诗引得》词头语料的《毛诗引得》领域词表，这实际上是一个融合了引得专家词汇知识的分词词表。

4.1.2 语料标注

本文进行句子对齐研究，基于总结归纳出的文本结构特点，先进行篇和章的对齐处理。《毛诗引得》是较为规范的经-注-疏结构，其一段经文后紧跟“注”，后紧接着另一段经文和“注”；换行以“[疏]”字样开头，对前述的经文和“注”部分的内容进行疏解，“[疏]”之后的内容就是经文和“注”对应的“疏”。

需对文本进行标注以便区分经和“注”。纵览《毛诗正义》全文，其篇章的边界可归纳出一定的格式特征，但存在个别篇章的上下文格式特征不全，需要标记补足。此外，在每段经文和相应的“注”之间添加“/”标识，并换行处理，使得每段经文与对应的“注”都是单独成段成行，可被视为一个整体，以便后续利用程序识别切分。如下图 4-1 和图 4-2 所示，以《蟋蟀》的部分句子为例，比对数据校对和语料标注前后《毛诗正义》行文结构。

《诗经》原文：

蟋蟀在堂，歲聿其逝。今我不樂，日月其邁。……良士蹶蹶。

蟋蟀在堂，歲聿其逝。今我不樂，日月其邁。邁，行也。無已大康，職思其外。外，礼樂之外。箋云：外謂國外至四境。o礼樂，此一樂字音嶽。好樂無荒，良士蹶蹶。蹶蹶，動而敏於事。o蹶，俱衛反。

[疏]傳"蹶蹶"至"於事"。o正義曰：《釋詁》云："蹶，動也。"《釋訓》云："蹶蹶，敏也。"

图 4-1 数据校对和语料标注前的《毛诗正义》行文格式

蟋蟀在堂，歲聿其逝。今我不樂，日月其邁。/邁，行也。
無已大康，職思其外。/外，礼樂之外。箋云：外謂國外至四境。o礼樂，此一樂字音嶽。
好樂無荒，良士蹶蹶。/蹶蹶，動而敏於事。o蹶，俱衛反。

[疏]傳"蹶蹶"至"於事"。o正義曰：《釋詁》云："蹶，動也"。《釋訓》云："蹶蹶，敏也"。

图 4-2 数据校对和语料标注后的《毛诗正义》行文格式

万方数据

由上述两个图可发现，《毛诗正义》的经文内容引用的是《诗经》原文的内容，“注”注释的经文内容是《诗经》原文的一个或多个句子，多个句子的经文和“注”的内容对应着同一段“疏”，数据校对处理已可初步区分经文和“注”。

4.2 对齐规范设计

所谓对齐规范设计就是对《毛诗正义》与《诗经》的对齐结果展现的外在布局，以及自动对齐实验前的标注工作进行一个统一的规范设计，以期使展现的对齐结果更雅观明了。

4.2.1 外观布局

本文研究的是《诗经》与《毛诗正义》的句子对齐，所谓句子对齐就是实现《毛诗正义》中引用《诗经》原文句子的每句经文都能与之相应的“注”和“疏”的内容对齐。

通过结构分析发现，《毛诗正义》每章中一般有一个到三个“引文”（引用《诗经》原文内容）句子，每个句子由两个至三个小句构成。因此，接下来对已对齐的每章中的句子，利用对齐算法或结合正则表达式等方法实现小句中的句子对齐。例如“約軧錯衡，八鸞鶬鶬。以假以享，我受命溥將。自天降康，豐年穰穰。”这段是已实现了每章中的句子对齐的“引文”内容，其中包含三个句子（以句号划分），六个小句（以逗号划分）。这部分内容，最终，应采用相关算法实现六个小句的内容与相应的“注”和“疏”部分的内容实现匹配对齐，这样才算实现这一部分内容的句子对齐。

《毛诗正义》中“疏”部分的内容比较繁杂。“疏”是对经文和“注”的注解，孔颖达在作“疏”时，若“注”部分的内容已经足够注释经文时，将只对“注”的部分注释。以“[疏]传‘某某’”开头，点明这段“疏”是对这部分“传”的内容进行的注释。比如《诗经》和《毛诗正义》由如下内容。

《诗经》原文之《生民之什·泂酌》：

泂酌彼行潦，挹彼注兹，可以濯罍。豈弟君子，民之攸歸。

《毛诗正义》：

泂酌彼行潦，挹彼注兹，可以濯罍。濯，滌也。罍，祭器。○罍音雷。滌，徒歷反。豈弟君子，民之攸歸。

[疏]傳“濯，滌。罍，祭器”。○正義曰：《説文》云：“滌，洗也”。“濯，浣也”。……《卷耳》云：“我姑酌彼金罍”。則饗燕亦有罍。以此論祭事，故言祭耳。

万方数据

本文要实现的句子对齐，是指先将《诗经》中的“泂酌彼行潦，挹彼注兹，可以濯罍。豈弟君子，民之攸歸。”这两个句子与《毛诗正义》中相应的“引文”对应起来，再将已对齐的“引文”内容与相应的“疏”内容对应起来，在本文称之为“篇和章的对齐处理”。“引文”中一般都包含多个小句，需将“注”和“疏”中的内容与已对齐的“引文”中的每个小句实现对齐，本文称之为“小句对齐处理”。其对齐内容的排版是一小句“引文”对应相应的“注”和“疏”的内容，并将“引文”、“注”和“疏”单独成行，平行显示。即使存在“注”或“疏”的内容缺失状况，也以同样的方式显示。在每部分的“注”和“疏”前加识别标记，以便区分“注”和“疏”。在对齐处理中，“疏”的内容繁杂，其中不止有孔颖达《正义》的内容，还有其他训诂学家的注释，但本文对齐过程不进行注释来源区分，只要是注释同一经文和“注”的“疏”内容都一视同仁，均视为其对应的“疏”。

4.2.2 标注规范

《毛诗正义》句子对齐实验有一系列的标注工作，为此需要制定一套准则对其进行规范，以便正确无误的显示所对齐的内容。

《诗经》总共有 305 首诗，其中有八首诗的标题重复出现，其中《柏舟》《黄鳥》《谷風》《莆田》《杕杜》《無衣》各出现两次，《揚之水》《羔裘》各出现三次。为此，在句子对齐实验前，有必要对《诗经》的 305 首诗按顺序进行序号标记，在匹配时会先匹配标题，这样可以避免内容匹配错误，确保《毛诗正义》匹配的是无误的诗标题及诗内容，且顺序标记也有助于比对每首诗的匹配情况。如图 4-3 所示，展示的是《诗经》原文的内容的顺序标记，图 4-4 和图 4-5 则是对《行露》该首诗，进行经-注-疏初步匹配的对齐格式展示，其中图 4-4 展示的是《行露》诗名与诗名的“注”以及对应的“疏”的对齐结果，图 4-5 则是诗内容的对齐结果。

万方数据

1 1.《關雎》
2 關關雎鳩、在河之洲。窈窕淑女、君子好逑。參差荇菜、窈淑女、寤寐求之。求之不得、寤寐思服。悠哉悠哉、輾荇菜、左右采之。窈窕淑女、琴瑟友之。參差荇菜、左右女、鍾鼓樂之。
3 2.《葛覃》
4 葛之覃兮、施于中谷。維葉萋萋、黃鳥于飛。集于灌木、之覃兮、施于中谷。維葉莫莫、是刈是濩。爲絺爲綌、服師氏、言告言歸。薄汙我私、薄澣我衣。害澣害否、歸寧
5 3.《卷耳》
6 采采卷耳、不盈頃筐。嗟我懷人、寘彼周行。陟彼崔嵬、姑酌彼金罍、維以不永懷。陟彼高岡、我馬玄黃。我姑酌不永傷。陟彼砠矣、我馬瘏矣、我僕痛矣、云何吁矣。
7 4.《樛木》
8 南有樛木、葛藟纍之。樂只君子、福履綏之。南有樛木、只君子、福履將之。南有樛木、葛藟縈之。樂只君子、福
9 5.《螽斯》
10 螽斯羽、詵詵兮。宜爾子孫、振振兮。螽斯羽、薨薨兮。

图 4-3 《诗经》顺序标注

104 >《行露》.16
召伯聽訟也。衰亂之俗微，貞信之教興，彊暴之男不能侵陵貞女也。衰亂之俗微，貞信之教興者，此殷之末世，周之盛德，當文王與紂之時。
[疏]"《行露》三章，一章三句，二章章六句"至"貞女"。○正義曰：作《行露》詩者，言召伯聽斷男女室家之訟也。由文王之時，被化日久，衰亂之俗已微，貞信之教乃興，是故彊暴之男不能侵陵貞女也。男雖侵陵，貞女不從，是以貞女被訟，而召伯聽斷之。《鄭志》張逸問："《行露》召伯聽訟，察民之意化耳，何訟乎？"荅曰："實訟之辭也。"民被化久矣，故能有訟。問者見貞信之教興，怪不當有訟，故云察民之意而化之

图 4-4 《行露》诗名经-注-疏初步匹配结果

105 厭浥行露，豈不夙夜？謂行多露！
興也。厭浥，濕意也。行，道也。豈不，言有是也。箋云：夙，早。夜，莫也。厭浥然濕，道中始有露，謂二月中嫁取時也。言我豈不知當早夜成昏礼與？謂道中之露大多，故不行耳。今彊暴之男，以此多露之時，礼不足而彊來，不度時之可否，故云然。《周礼》仲春之月，令會男女之無夫家者，行事必以昏昕。○厭，於葉反，徐於十反，又於立反，沈又於占反。浥，本又作"挹"，同於及反，又於賄反。莫，本又作"暮"，同忙故反，又亡博反。《小星》詩同。與音餘。大音泰，舊吐賀反。彊，其丈反，下"彊委"同；沈其常反。度，待洛反。否，方九反。令，力政反，後不音者放此。昏昕，許巾反，至礼用昕，親迎用昏。
[疏]"厭浥"至"多露"。○毛以爲厭浥然而濕，道中有露之時，行人豈不欲早夜而行也。有是可以早夜而行之道，所以不行者，以爲道中之露多，懼早夜之濡己，故不行耳。以興彊暴之男，今來求己，我豈不欲與汝爲室家乎？有是欲與汝爲室家之道

图 4-5 《行露》诗内容经-注-疏初步匹配结果

万方数据

由图 4-3 可看出，对《诗经》的每首诗都进行了顺序标记。由图 4-4 可看出，《毛诗正义》中不止是对诗的经文有“注”和“疏”，对每首诗的标题都有相应的“注”和“疏”，有些对诗名的“注”和“疏”的篇幅还很长，用以交代这首诗的诗旨等情况。显示的格式是：诗名.序号标记，换行显示诗名的“注”，再换行显示诗名的“疏”。图 4-5 则展示的是《诗经》原文内容诗句在《毛诗正义》中的经-注-疏匹配格式，每部分内容单独成行，便于一目了然区分。

4.3 自动分词

本文的句子对齐研究需对文本进行分词处理。为此本节基于《诗经》手工分词语料和《毛诗引得》领域词表，利用机器学习的方法对《毛诗正义》进行分词处理。

4.3.1 自动分词的流程

本节探究的是《诗经》的自动分词，其自动分词流程如图 4-6 所示。

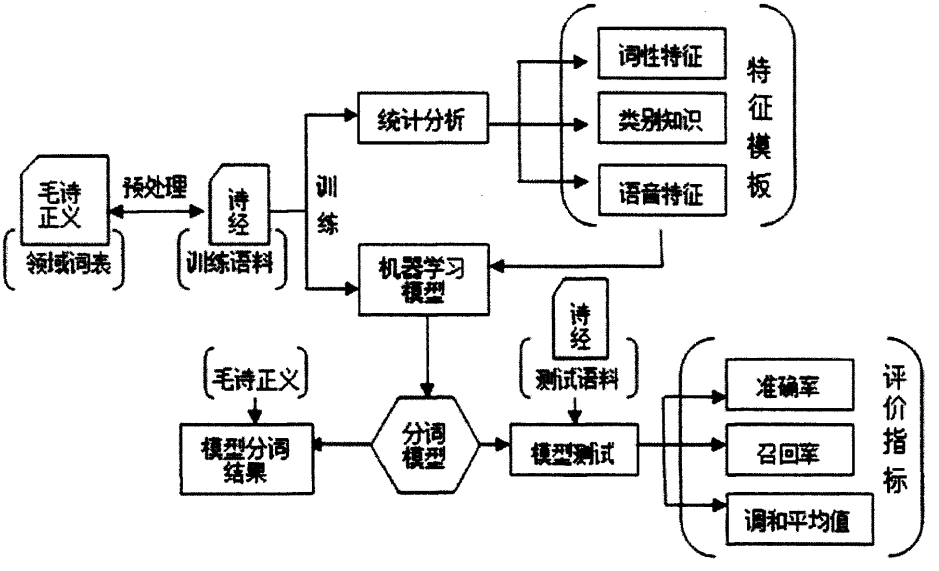


图 4-6 《诗经》自动分词流程

由图 4-6 可知，结合了三类特征知识构建特征模板，并对所有的特征模板进行分词效果的测试，从中选出效果最好的分词模型，对《毛诗正义》语料进行 100%训练语料的封闭测试，从而得到分词结果。

《自动分词》的训练语料是经过手工分词得到的《诗经》语料，据此获得了《诗经》词表。《诗经》手工分词语料的部分样例，如图 4-7 所示，该手工分词语料进行了词性标注，《诗经》手工分词语料的词性标记集合见附录 2 中的表 2-1。

万方数据

關關/s 雎鳩/n ，/w 在/v 河/ns 之/u 洲/n 。/w
窈窕/a 淑/a 女/n ，/w 君子/n 好/a 逑/n 。/w
參差/q 荇菜/n ，/w 左/zn 右/zn 流/v 之/u 。/w
窈窕/a 淑/a 女/n ，/w 寤/v 寐/v 求/v 之/u 。/w
求/v 之/r 不/d 得/v ，/w 寤/v 寐/v 思/v 服/v 。/w
悠/a 哉/y 悠/a 哉/y ！/w
輾轉/v 反/v 側/v 。/w
參差/a 荇菜/n ，/w 左/zn 右/zn 采/v 之/r 。/w
窈窕/a 淑/a 女/n ，/w 琴/n 瑟/n 友/v 之/r 。/w
參差/a 荇菜/n ，/w 左/zn 右/zn 芼/v 之/r 。/w
窈窕/a 淑/a 女/n ，/w 鐘/n 鼓/n 樂/vs 之/r 。/w
葛覃/n
葛/n 之/u 覃/v 兮/y ，/w 施/v 于/p 中谷/n ，/w
維/y 葉/n 萋萋/a 。/w
黃鳥/n 于/u 飛/v ，/w 集/v 于/p 灌木/n ，/w 其/r
鳴/v 喈喈/s 。/w
葛/n 之/u 覃/v 兮/y ，/w 施/v 于/p 中谷/n ，/w
維/y 葉/n 莫莫/a 。/w
是/u 刈/v 是/u 濩/v ，/w 爲/v 絺/n 爲/v 綌/n
，/w 服/v 之/r 無/d 斁/a 。/w

图 4-7 《诗经》手工分词语料部分样例

由图 4-7 可看出，自动分词是基于《诗经》手工分词语料和《毛诗引得》领域词表，结合《广韵》字表，利用统计分析的方法构建特征模板，采用条件随机场（CRFs）的机器学习方法搭建《诗经》自动分词模型，并对不同特征组合下的分词模型采用准确率、召回率、调和平均值指标进行性能测试，从中选出性能最佳的分词模板对《毛诗正义》文本进行封闭测试，从而实现《毛诗正义》的自动分词。

4.3.2 机器学习模型的简介和特征模板的构建

4.3.2.1 CRFs 机器学习模型

条件随机场是一种基于统计的序列标记和分割的机器学习方法，通过已标记的数据，构建特征模板进而训练产生得到更好的实验效果[1]。

CRFs 的特征分为两类：Unigram 特征，以“U”开头；Bigram 特征，以“B”开头，均可用特征模板(template)表示。本文所用的特征模板为 Unigram 特征，特征模板其实是指针对条件随机场模型在训练过程中融合所需识别序列单位的词性、窗口字和词、类别知识、语音特征（反切、韵、调）构成的特征组合。特征模板根据是否融入

[1] John Lafferty, Andrew McCallum and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. InProc .ICML, 2001:282-289.

万方数据

特征知识可分为简单特征模板、复杂特征模板两种。《诗经》分词主要涉及的特征是字，从字的观察角度看，结合其长度特征，本文所用的简单特征模板如表 4-4 所示。

表 4-4 简单特征模板说明

编号	CRF++模板形式	模板的具体语义
1	U00:%x[-2,0]	当前字的前两个字
2	U01:%x[-1,0]	当前字的前一个字
3	U02:%x[0,0]	当前字
4	U03:%x[1,0]	当前字的后一个字
5	U04:%x[2,0]	当前字的后两个字
6	U05:%x[-1,0]/%x[0,0]	前一个字到当前字的转移概率
7	U06:%x[0,0]/%x[1,0]	当前字到后一个字的转移概率
8	U07:%x[-1,0]/%x[1,0]	前一个字到后一个字的转移概率

4.3.2.1 特征模板构建

为构建《诗经》自动分词机器学习模型，对《诗经》训练语料进行词长词频统计，统计结果显示《诗经》文本以单字词为主，多字词中以两字词和三字词居多，因此在简单特征模板中主要利用了当前字左右两个窗口的特征。以简单特征模板作为整个分词模型的训练过程中的基准线，后续的复杂特征模板是在简单特征模板的基础上产生的。

条件随机场最大的优势是不仅可以融入当前字左右的各种特征知识，还可以结合当前字本身的特征以此形成最有效的特征模板。本文针对《诗经》典籍，从统计的角度，结合《广韵》字表，确定模型训练过程使用的内在特征，具体如下：

（1）词性特征。古代汉语并没有在词语词之间设置明显分隔符用以区分不同的词，并非像英语类字母型的语言以空格进行分隔。在古代，仅地名、人名或者是特有名词存在多字词，其他状况下基本是一个汉字就是一个词。词类划分的主要依据就是词性，词类则是语言学中词的语法分类，词性作为词条的一个重要特性，可以提供当前词与其上下位词的有效讯息，也是将词条与句法连接的重要纽带。

（2）字本身的类别知识。在《诗经》的词汇组成中，涉及到不同的字符，将这些字符进行一个初步的分类，对于确定《诗经》中的字是否构成词具有直接的影响。本文借鉴邢福义先生的“小句”理论，将《诗经》典籍中出现的字符分为古汉字、小句句中的标点符号、小句句尾的标点符号。具体例子如：“關”、“，”、“。”。

（3）字的语音特征。读音作为汉字呈现形式的重要载体，其在由字构成词，尤其是在构成象声词、叠音词上特征明显，可借助《广韵》字表，选取字构成词的特征时，将汉字的反切、韵、调作为特征知识添加到复杂特征模板中。本文采用的《广韵》字

万方数据

表是从中国国际广播出版社 2011 年出版的严学窘先生编著的《国学经典导读：广韵》[1]一书中获取的电子版。比如：關的语音特征分别是“古還”，“删”，“平”。

基于上述三类特征知识，结合简单特征模板，构成复杂特征模板，相对于简单特征模板，复杂特征模板利用当前字左右三个窗口及其相应的转移概率，相对而言更复杂。只增加了一个特征知识的特征模板如表 4-5 所示，产生了 12 个特征，其他复杂特征模板与此模式相同。因此，在简单特征模板的 8 个特征的基础上，形成了最终的 44 个复杂特征模板。

表 4-5 增加一个特征知识的复杂特征模板

编号	CRF++模板形式	模板的具体语义
1	U10:%x[-2,1]	当前字的前两个字
2	U11:%x[-1,1]	当前字的前一个字
3	U12:%x[0,1]	当前字
4	U13:%x[1,1]	当前字的后一个字
5	U14:%x[2,1]	当前字的后两个字
6	U15:%x[-2,1]/%x[-1,1]	前两个字到前一个字的转移概率
7	U16:%x[-1,1]/%x[0,1]	前一个字到当前字的转移概率
8	U17:%x[0,1]/%x[1,1]	当前字到后一个字的转移概率
9	U18:%x[1,1]/%x[2,1]	后一个字到后两个字的转移概率
10	U20:%x[-2,1]/%x[-1,1]/%x[0,1]	前两个字和前一个字到当前字的转移概率
11	U21:%x[-1,1]/%x[0,1]/%x[1,1]	前一个字和当前字到后一个字的转移概率
12	U22:%x[0,1]/%x[1,1]/%x[2,1]	当前字和后一个字到后两个字的转移概率

4.3.3 机器学习分词实验与结果分析

4.3.3.1 基于《毛诗引得》领域词表的标注集合确定

条件随机场的标注方式有多种，常用的有 4-tag 和 6-tag，因此在模型训练前，确定标注集合是一项不可或缺的基础工作，可参照以下公式并结合《毛诗引得》领域词表的词汇分布完成。具体公式如下：

$$L = \frac{1}{N}\sum_{i=j}^{k} iN_i \qquad (1)$$

其中，L 表示当 i≤k 时，《诗经》语料词汇平均加权后的长度情况，Ni 代表整个训练语料《诗经》中长度为 i 的词汇出现的次数，k 和 j 代表的是整个训练语料中最长和最短的词汇的具体长度，N 表示整个训练语料的词汇总数量。经过词长频次统计，《毛诗引得》领域词表中词汇长度分布如表 4-6 所示。

[1] 严学窘. 国学经典导读广韵[M]. 北京：中国国际广播出版社, 2011:20-75.

万方数据

表 4-6 《毛诗引得》领域词表的词汇长度分布

编号	词长	词数	样例
1	1	2830	鳳
2	2	239	泮水
3	3	20	雨無正
4	4	50	十畝之間
5	5	2	節南山之什
6	6	2	閔予小子之什

上表中的长度统计主要是依据汉字的个数计算的。从表 3 中可看出，《毛诗引得》领域词表的词汇主要是单字词和两字词，三字词和四字词相对较少。在公式（1）和表 3 的基础上，本文确定使用四词位（4-tag）的标注集合，用 W 表示，具体为 W = {B，M，E}，B 表示一个词的初始字，M 表示一个词的中间字，E 表示一个词的结束字，如果是单字词，统一标注为 BE。

4.3.3.2 基于特征模板的语料标注

本节的自动分词实验，是基于条件随机场的机器学习方法，利用 python 运行 CRF++工具包[1]实现。为更细致深入地探究各项内在特征知识对《诗经》分词效果的影响力，本文将词性特征、类别特征、语音特征（反切、韵、调）三项进行排列组合，得到 23 组不同特征组合下的分词模板，具体见附录 2 中的附录 2-2。针对所有的分词模板进行两类实验，第一类是十重交叉实验，即把整体语料随机均分为 10 份，从中依次抽取 1 份作为测试语料，剩下的 9 份作为训练语料，重复 10 次操作，可以在每个特征模板下训练产生 10 个分词模型。第二类是整体实验，即对所有的分词模板，将整体语料按照 9:1 的训练和测试比例实验，每个特征模板可训练产生一个分词模型，并对不同的分词模型在相应地 10%开放测试语料上进行分词效果的性能测试，因此，确定性能测试的评价指标是十分必要的。

（1）评价指标

按照 SIGHAN 的测评标准，用分词的准确率 P，召回率 R，调和平均值 F 对机器学习模型的分词性能进行测评，各项测评指标的具体计算公式如下所示。

$$P = \frac{\text{正确识别的词语数量}}{\text{系统识别的总的词语数量}} \quad (2)$$

$$R = \frac{\text{正确识别的词语数量}}{\text{测试语料中总的词语数量}} \quad (3)$$

[1] CRF++. [EB/OL]. [2017-2-15]. https://sourceforge.net/projects/crfpp/.

万方数据

$$F=\frac{2*P*R}{P+R} \quad (4)$$

（2）实验结果

对机器学习训练产生的所有分词模型的性能测试结果分析，从两方面展开，先重点分析，再整体分析。重点分析是从所有分词模型中选取十重交叉训练出的简单特征模型、只增加词性特征的复杂特征模型、三类特征征齐全的复杂特征模型各 10 个，以及三类特征整体语料以 9:1 训练和测试后的模型各一个，对这些分词模型的性能进行分析。具体性能见表 4-7 三种特征模板上的分词模型性能。

表 4-7 语料以 9:1 训练在三类特征中的性能比较

特征模板类型	准确率 P	召回率 R	调和平均值 F
简单特征模板	94.80%	92.67%	93.72%
只添加词性特征模板	97.03%	97.29%	97.16%
三类特征齐全模板	97.50%	97.27%	97.39%

表 4-7 是整体语料按 9:1 训练和测试实验性能结果。简单特征模型的调和平均值 F 是 93.72%，增加词性特征之后 F 值是 97.16%，提高了 3.44%，在特征齐全下的 F 值是 97.39%，也能看出添加特征可提高分词模型的效果，但在所有特征中，词性特征对分词效果的影响最深。比如：“臨衝 閑閑，崇 墉 言 言，執 訊 連 連，攸 馘 安安。”是在简单模板下的分词结果，“臨衝 閑閑，崇 墉 言言，執 訊 連連，攸 馘 安安。”是只添加了词性特征模板的分词结果，“臨衝 閑閑，崇 墉 言言，執 訊 連連，攸 馘 安安。”是三类特征齐全模板下的分词结果。从分词结果看，添加了词性特征的模板识别了句中的全部叠词，提高了分词效果。但该句在三类特征齐全模板下的分词结果与只添加词性特征模板的分词结果一样，由此可看出，语音特征和类别特征其实对《诗经》分词影响不大，对分词结果影响最大的是词性特征。

为更凸显本文针对《诗经》分词的机器学习模型方法的有效性，对比基于规则的《诗经》自动分词方法，结果见表 4-8。此外，为更全面的进行分词模型效果的整体性能分析，现将 23 组特征组合按整体语料以 9:1 训练和测试得到的分词模型性能结果中的调和平均值 F 以折线图和柱状图结合的形式展示，具体见下图 4-8。

表 4-8 基于规则分词方法与机器学习模型分词性能对比

《诗经》分词方法	精确率 P	召回率 R	调和平均值 F
基于规则分词	86.89%	82.24%	84.50%
简单特征机器学习模型	94.80%	92.67%	93.72%
复杂特征机器学习模型	97.50%	97.27%	97.39%

万方数据

用基于规则的方法，采用《毛诗引得》领域词表逆向最大匹配的方法切分《诗经》语料，可将其视为本文的基础实验。为测试分词效果，将该分词结果与《诗经》手工分词的结果比对，计算其 P、R 和 F 值。从表 7 可看出，机器学习模型的分词方法相对基于规则的分词方法在分词性能上得到了明显的提升，F 值提升了 9.22%。且本节在构建机器学习模型时，增加三类特征形成复杂特征模板，使得分词性能得到进一步的提升，调和平均值 F 相对简单特征提升了 3.67%，准确率 P 提升了 2.7%，召回率 R 提升的最多，达到 4.6%。由此可见，在训练机器学习模型时，增加有效的特征将很大程度的提高机器学习模型的分词性能。

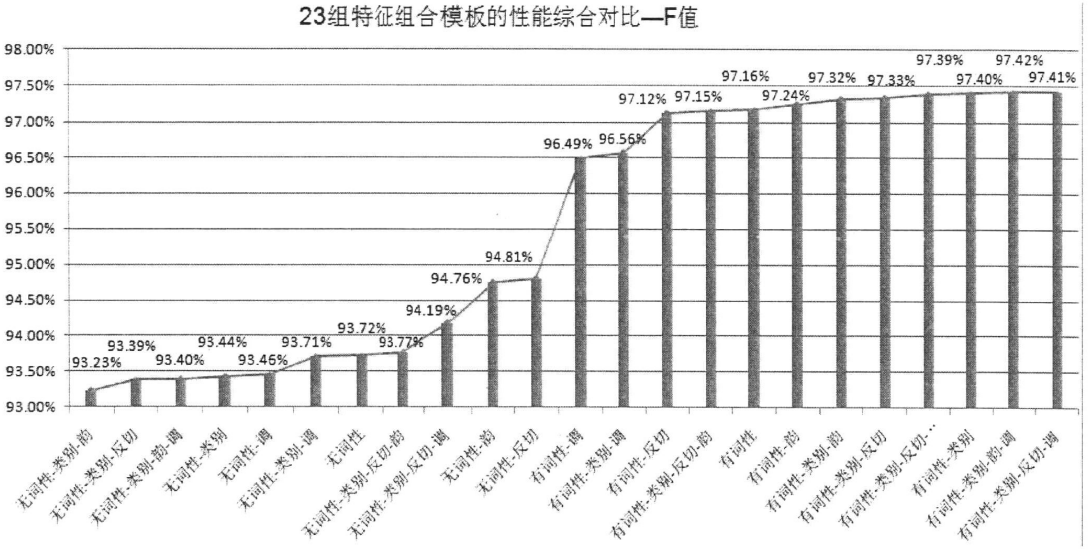


图 4- 8　23 组特征组合模板的性能综合比较

图 4-8 中的横坐标对应的是 23 组特征组合模板。由图 4-8 可看出调和平均值 F 值最大为 97.42%，对应的特征组合是“有词性-类别-反切-调”，其次为 97.41%，对应的特征组合是“有词性-类别-韵-调”，三项特征齐全即“有词性-类别-反切-韵-调”的特征组合下的 F 值为 97.39%，由此可见，语音特征中的反切、韵、调三项进行任意排列组合得到的各项分词模型性能差异不大，缘于语音特征都是从《广韵》字表中程序自动匹配获取的，每个汉字所对应的反切、韵、调应一致，故其单一的组合与综合组合对分词效果的影响几乎相近。由图中，可明显的看出，但凡特征组合模板中添加了词性特征，其 F 值相比同等条件下无词性的分词模型性能要优良很多，因此也可得到词性特征对《诗经》自动分词性能的影响很大的结论。

最后以全部的《诗经》手工分词语料作为训练数据，用分词效果最好的特征模板对《毛诗正义》语料进行封闭测试，可实现《毛诗正义》的分词处理。

万方数据

第五章 对齐实验

5.1 对齐算法的总体设计

所谓《诗经》与其注疏文献的句子对齐，就是将《诗经》中的句子，与在注疏文献中出现的相同的内容实现匹配对应[1]。

《毛诗正义》的行文结构特征可用于初步实现文本的篇章对齐，使得整个文本可呈现分区分块状态，该对齐结果可以视为是将全部的文本分为了两部分，一部分是“引文”，一部分是“注”和“疏”，将之称为是注释部分。接下来将利用采用训诂术语正则表达式匹配与句子相似度计算方法相结合的方式探索《诗经》与《毛诗正义》的“引文”和“注释”内容的对齐。

训诂方法的科学性对训诂学弥足重要，训诂学要“论其法式，明其义例”[2]。也就是说需要从训诂的方法中总结归纳出规律性的东西，并用它来实践[3]。因此，本文总结归纳《毛诗正义》中的训诂术语，并找出训诂的规律，将其用抽象的正则表达式进行具体描述，并构建训诂术语模式库，采用正则匹配的方式进行初步的匹配。此外，还利用 gensim 进行相似度匹配算法对匹配结果进行完善。本文句子对齐具体的实验流程如图 5-1 所示。

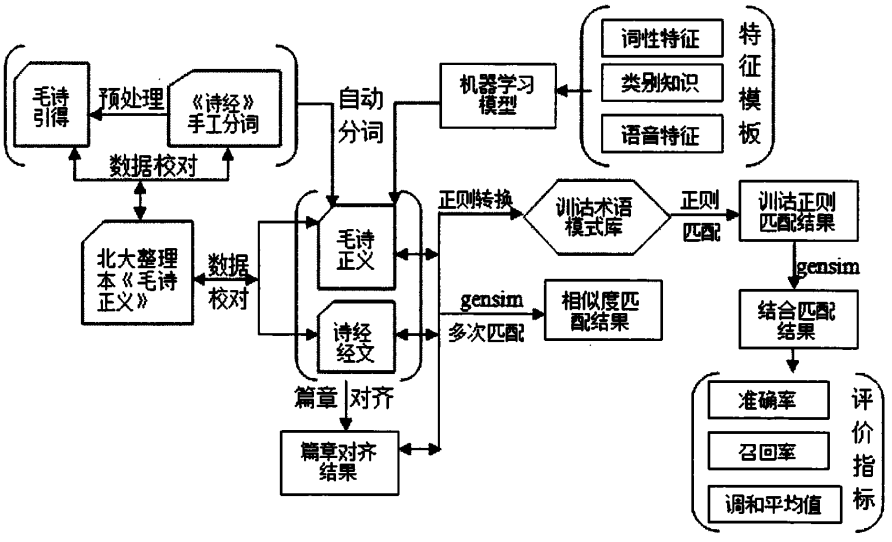


图 5-1 句子对齐实验流程

[1] 马创新，陈小荷，曲维光. 注疏文献中的注释语句自动分析[J]. 计算机科学, 2012(10):220-223.

[2]赵振铎. 《训话学纲要》[M]. 西安：陕西人民出版社, 1987:6-26.

[3] 马树杉. 试论孔颖达的训诂方法(一)[J]. 常州工业技术学院学报, 1989(03):28-31.

万方数据

由图 5-1 可看出，本文在进行数据校对和文本自动分词后，对《毛诗正义》先利用文本结构特征进行了篇章的对齐处理，再利用正则匹配和相似度计算两种方式探索《诗经》与《毛诗正义》的句子对齐，均生成了初步的匹配结果，利用 gensim 相似度计算结果针对正则匹配的结果再进行相似度计算，得到混合的匹配结果，并用测评指标对匹配结果进行测评，验证本文句子对齐方法的有效性。

5.2 篇和章的对齐

要实现《诗经》与《毛诗正义》的句子对齐，即经文、“注”和“疏”的对齐，需要将上述三部分内容先切分开，为此，首先应利用《毛诗正义》的文本结构特点，先进行“篇和章的对齐”。

以《诗经》305 首诗的诗名作为全文“篇”的总领标识，为此，需要对每首诗即每章的排列顺序进行序号标记，以免受重复诗名的干扰，避免诗内容匹配不一致。借助已标注的“/”标识，可将经文和“注”隔开，使其单独成行。这部分实际上已经实现了“引文”（引用《诗经》经文的句子）与“注”的对齐，只是这部分的“引文”句子较多，其中可能包括两个到三个长句子，多个小句子，需后续进一步细分。

利用“[疏]”标识，可将每章中的“疏”抽取出，与经文和“注”分隔开，但只能将整个“疏”内容一并与其相对应的经文和“注”对应上。“篇和章对齐”的结果以诗名为每篇的标识，并在诗名后标记篇序，“引文”、“注”和“疏”单独成行，其结果如图 5-2 所示。

104 >《行露》.16
召伯聽訟也。衰亂之俗微，貞信之教興，彊暴之男不能侵陵貞女也。
，當文王與紂之時。
[疏]"《行露》三章，一章三句，二章章六句"至"貞女"。○正義曰：
王之時，被化日久，衰亂之俗已微，貞信之教乃興，是故彊暴之男
，而召伯聽斷之。《鄭志》張逸問："《行露》召伯聽訟，察民之意
故能有訟。問者見貞信之教興，怪不當有訟，故云察民之意而化之，
化日久，貞女不從，男女故相與訟。如是民被化日久，所以得有彊
人。經三章，下二章陳男女對訟之辭。首章言所以有訟，由女不從
殷之末世，故有衰亂之俗；周之盛德，故有貞信之教。指其人當文
世，周之盛德邪？"當文王與紂之事，此其文也。
105 厭浥行露，豈不夙夜？謂行多露！
興也。厭浥，濕意也。行，道也。豈不，言有是也。箋云：夙，早。
也。言我豈不知當早夜成昏礼與？謂道中之露大多，故不行耳。今
否，故云然。《周礼》仲春之月，令會男女之無夫家者，行事必以
反。浥，本又作"挹"，同於及反，又於賄反。莫，本又作"暮"，同
舊吐賀反。彊，其丈反，下"彊委"同；沈其常反。度，待洛反。否，
，至礼用昕，親迎用昏。
[疏]"厭浥"至"多露"。○毛以爲厭浥然而濕，道中有露之時，行人豈
者，以爲道中之露多，懼早夜之濡己，故不行耳。以興彊暴之男，

图 5-2 “篇和章的对齐”结果

万方数据

由图 5-2 可看出，“篇和章的对齐”是将《行露》这篇诗的整体以一个统一的标识区分，如图中的“<”标记；每篇中的诗名、诗名的“注”以及诗名的“疏”作为一个整体，都是对诗名的注释，但每部分单独成行；诗内容则以《毛诗正义》中每章内容的行文结构作为划分标准，每章视为一个整体，内含每章的经文、每章的“注”、每章的“疏”，并单独成行。

5.3 训诂术语正则匹配实验

所谓正则表达式匹配就是利用总结出的训诂术语的规律，利用正则表达式对其进行具体描述并搭建训诂术语模式库，进而采用 Python 正则匹配的方法基于模式库匹配《毛诗正义》全文。

5.3.1 训诂术语与正则表达式

在正则表达式匹配实验前，需要对训诂术语进行正则表达式的转换。如训诂术语：“x 者，y 也”，可用“(.+)者，.+也”进行具体描述，进而将该模式套用到语料全文，从而得到更全面的内容。

正则表达式由原子和运算符组成，以一定的规则将不同部分以运算符连接而成 [1]。在训诂著作中，标注字音方面的训释方法主要有譬况、反切、直音、如字、标四声、叶音等几类。采用的训释方式如表 5-1 所示：

表 5-1 标注字音的常用训释方式

训释方法	训释术语		正则表达式
譬况法	注释中包含术语“闭口、笼口、横口、急气、缓气、内言、外言、长言、短言”等描述性术语		(.),.*(闭口|笼口|横口|急气|缓气|内言|外言|长言|短言)
反切法	A，BC 反	A，BC 切	(.)，..[反切翻纽]
	A，BC 翻	A，BC 纽	
直音法	A，音 B	A，声 B	(.)，[音声].+
	B 音	A，音 B	(.)，.音
如字	如字	A 如字	(.)，(如字|依字读)
	A 音如字	A，依字读	(.)音?如字
标四声	A，x 声	A，x 读	(.)，.+[声读]

训诂学的主要任务是解释词义，其释义的方法多种多样，针对专有名词、动词、形容词等不同类型的词语既有不同的专释术语，也有通用的训诂术语，因后续正则匹配实验的需要，将所归纳的训诂术语按注释对象的词性进行划分，从而得到归纳的所有训诂术语的正则表达形式。如表 5-2 列出了常见的训诂术语和相应的正则表

[1] Behrouz A.Forouzan, Richard F.Gilberg 著.彭松虎译. UNIX 和 shell 程序设计权威教程[M]. 北京：清华大学出版社. 2003:40-60.

万方数据

达形式。此外，还针对《毛诗正义》的训诂术语特征，总结出了《毛诗正义》训诂术语模式库，见附录 3-1。

表 5-2 训诂术语以及相应的正则表达式形式

类别	训诂术语	正则表达式
人名、物名	x,名 y　　x, 字 y x, y 名(也)　　x, 姓 y	(.+)，[名字姓].+ (.+)，.{1,5}名也?
动词	x, y 之 x, 为 y	(.+)，.+之 (.+)，为.+
形容词	x, y 貌　　x, y 之状 x, y 的样子　x, y 意也	(.+)，.+(貌|之状|的样子|意也)
代词	x, 身也　　x, 自称之词 x, 指 y 之辞　　x,指 y 辞也	(.+)，(身也|自称之词) (.+)，指.+(之辞|辞也)
助词	x, y 辞也　　x,y 辞 x，y 语助(也)　　x, 发声也	(.+)，.+(辞也|辞|语助也|语助) (.+)，发声也
叹词	x, y 叹辞(也)　x, y 之声 x，y 嗟叹辞也	(.+)，(嗟叹辞也|叹辞)[，。] (.+)，.+之声[，。]
象声词	x, y 声(也)	(.+)，.+(声也|声)
通用术语	x 者，y　　x 者，y 也　　x,谓 y x, y 也　　x 曰 y x 为 y x 谓之 y	(.+)者，.+也 (.+)者，.+ (.+)，.+也 (.+)，谓.+ (.+)(谓之|为|曰).+

表 5-1 和表 5-2 的“训诂术语”一列中，用 A、B、C 表示单个字；用 x、y 表示单音节词或多音节词；同时 A，x 也表示的是训释对象，B、C、y 也表示训释内容的主体部分。在“正则表达式”一列中，所有的正则表达式都使用了分组保存运算符，能够与各个正则表达式中的第一组模式相匹配的字符串就是训释语句中的训释对象，都暂时保存在文档中，可以使用“向后引用”原子来引用它，然后就可以判断这个训释对象是否出现在与训释语句相同文本块的引文中。

5.3.2 正则表达式匹配实验

5.3.2.1 文本句子切分

正则表达式匹配是在“篇和章的对齐”的基础上实施的进一步处理。训诂术语模式库中每项正则表达式都是根据《毛诗正义》文本的训诂术语特征总结归纳的，为提高匹配的精度，训诂术语正则表达形式对标点均进行了限制约束，在正则匹配时会区分标点，故在进行训诂术语正则匹配前，需对文本进行句子切分处理。句子切分应区分经文、“注”和“疏”的内容，将每部分分别进行句子切分。

《毛诗正义》的很多标点都是“，”，将整个文本分成了很多小句。以“。”结尾的句子通常句子字数较多，其中包含了多个以“，”或“、”等划分的小句。但由

万方数据

于训诂术语模式库中的正则表达式含有“，”和“。”等标点符号，是其进行正则匹配重要的一部分，为避免打乱正则匹配的模式，本文采用的是最简单的以“。”作为句子切分的标准。但由于文本中还存在“！？；”等标点符号，其断句作用类似于“。”。故先将这些符号统一处理转化成“。”，此操作可减少字数在几十字，甚至上百字的句子出现的频率，从而将句子切分的更细，正则匹配时可涵盖到更多的句子。句子切分处理后，文本将由不同长度的句子组成，为后续匹配做语料铺垫。以《关雎》中的一小段内容为例。

《关雎》注文：

“荇，衡猛反，本亦作荇，沈有並反。左右，王申毛如字；鄭上音佐，下音佑。“接余”音餘，本或作“荽荼”，非。”

句子切分后，分为四个句子：

“荇，衡猛反，本亦作荇，沈有並反

左右，王申毛如字

鄭上音佐，下音佑

“接余”音餘，本或作“荽荼”，非

由此可看出，以“。”切分，存在将两个双引号切分在两个句子中的情况，这种情况难以避免，但并不影响后续匹配实验的进行。

5.3.2.2 正则表达式匹配

正则表达式是一种强大、快捷、高效的文本处理工具，描述了一种字符串匹配的模式，可以用来检查一个串是否含有某个子串、将匹配的子串进行替换以及从某个串中取出符合某个条件的子串等，其具有字符串所不具备的强大和灵活的表达能力，能够准确地表达出复杂的特征[1]。

《毛诗正义》训诂术语模式库中的正则表达式匹配的内容包括前后两项，一般匹配的前项是被释对象，后项是训释内容，均以贪婪模式对符合条件的内容进行匹配，所谓贪婪模式即尝试匹配尽可能多的字符。贪婪模式胜在抽取的全面，但容易抽取的内容太杂，为此后期还需进行内容筛查。

本文分别对“注”和“疏”分两个阶段进行正则表达式匹配，并进行多重交叉匹配，力求匹配到更全面的内容。“注”匹配阶段，将“注”与“引文”进行匹配，抽取出符合正则表达形式且包含“引文”内容的“注”句子，将匹配结果分别保存，得到三个文档：“注”和“引文”的匹配结果；匹配的“引文”；匹配的“注”。“疏”

[1] 秦元坤. 正则表达式匹配中的 DFA 优化技术研究[D]. 北京：清华大学, 2008: 9-50.

万方数据

匹配阶段，第一：将“疏”与“引文”进行匹配，将“疏”中符合正则表达式的“引文”内容输出，同样可得到三个类似的文档；第二：将“疏”与“注”进行匹配，匹配结果保存为三个文档；第三：将“注”和“引文”两部分内容视为整体，将其与“疏”进行匹配，匹配结果保存为三个文档。

上述一次匹配出的句子中，有些句子的长度很长，其中还包含着模式中的内容，为此可以将一次匹配的结果输入，利用训诂术语模式库进行二次匹配，从而得到再次匹配的结果，将匹配结果保存到文档中。

5.3.3 正则表达式匹配结果及分析

正则表达式匹配得到多组匹配结果。其中“注”匹配的结果，以“關關雎鳩，在河之洲。”一句的匹配结果为例，如图 5-3 所示。为了更好的进行结果的对比分析，将经文、“注”和匹配的结果放在一个文档。从匹配的结果中可看出，只是抽取出了部分内容，比如“水中可居者曰洲”被抽取出了，但“關關，和聲也”却没有被匹配上，由此可反查是否训诂术语模式库中的正则表达式不够齐全，加以完善补充，以期得到更全的匹配结果。以这种方式反复验证补充，经过多次匹配将能得到较好的匹配结果。由此可看出，正则表达式匹配是可以进行回溯革新的，通过匹配结果反观匹配过程，检验匹配过程中可完善之处，进而改善得到更佳的结果。

万方数据

1 >關關雎鳩，在河之洲。
2 >興也。關關，和聲也。雎鳩，王雎也。鳥摯而有別洲。后妃說樂君子之德。無不和諧。又不淫其色。雎之有別焉。然後可以風化天下。夫婦有別則父子臣敬。君臣敬則朝廷正。朝廷正則王化成。箋云。王雎之鳥。雌雄情意至然而有別。雎，七胥反。鳩，有至別者。洲音州。興，虚應反，沈許甑反。案，意有不盡。故題曰興。他皆放此。摯本亦作鷙。音。下同。說音悦。樂音洛。諧，户皆反。朝，直遥。
3 水中可居者——洲
4 摯之——至也
5 雌雄情意至——有別
6 案——之
7 案——興是譬諭
8 案，興——譬諭之名
9 故題——興
10 摯本——作鷙
11 >窈窕淑女，君子好逑。

图 5-3 “注”一次匹配结果

5.4 相似度算法对齐

上节利用训诂术语模式库进行了正则表达式匹配，从“注”和“疏”中将部分与经文相关的注释内容抽取出来了，但并没有将大部分的内容抽取出来，且抽出来的成分较杂。为此本节从另一个角度出发，利用 gensim 计算相似度探索《诗经》与《毛诗正义》的句子对齐。

5.4.1 gensim 简介

句子相似度计算的方法有很多，本文采用 gensim 进行相似度计算。gensim 是 Python 的自然语言处理库，gensim 可分析纯文本文档获取语义结构，可检索语义相近的文档，还可扩展和统计语义信息，可将文档根据 TF-IDF（词频与逆文档频率算法），LDA（Latent semantic analysis，浅层语义分析），LSI（Latent semantic indexing，浅层语义索引）等模型转化成向量模式，以便进行进一步处理。此外，gensim 可实现 word2vec（词向量的训练与生成模板）、doc2vec（篇章向量的训练和生成模板）功能，

万方数据

将单词转化成词向量，可将其用于后续的相似度计算中。gensim 进行无监督的学习，并不需要人工标注的附加信息，能够自动地从文本中提取语义信息。

LSI 是通过 SVD（奇异值分解）的方式降维，将词项和文本映射在一个新的空间，从而解决一词多义，一义多词问题。LSI 本质上是把每个特征映射到一个更低维度的子空间。LSA 是通过“矢量语义空间”来提取文档与词中的“概念”，进而分析文档与词之间的关系，其与 LSI 原理相似。

5.4.2 相似度计算实验

5.4.2.1 预处理文本

gensim 基于一个原始文本的语料集合 corpus，本文实验采用《毛诗正义》语料，可从中抽取部分或全部的语料作为 corpus。

采用 gensim 进行相似度计算前，需先对文本进行一些预处理工作，例如分词处理、去停用词、去低频词等，从而将文档中原始的字符文本转换成 gensim 模型所能理解的稀疏向量。此处采用本文提出的构建特征模板的机器学习分词方法用于此处的文本分词，该分词结果相对 jieba 效果要更好，可将结果导入 gensim 用于后续处理。由于目前没有针对古汉语繁体字的停用词表，且本文的语料是繁体字，因此没有进行去停用词操作。但进行了词频统计，去除了低频词，只保留词频大于 1 的词汇。

利用 gensim 提供的 API 建立语料特征的索引字典，并将文本特征的原始表达转换成词袋模型对应的稀疏向量的过程。即 bag-of-words（词袋）可将分词的结果以字典的形式保存，每次调用的时候，只需对字典进行调用即可，给字典中的每个词汇分配一个 id，使得每个词具有独特的识别标识。对字典类型的数据进行二维向量的转换，其中 id 是向量的一维数据，词的频率计数是向量的二维数据。进行向量转换后，每个句子都表示成了二维矩阵的形式，便于后续进行相似度计算。

5.4.2.2 主题向量转换及相似度计算

对文本向量的转换是 gensim 的核心，通过挖掘语料中隐藏的语义结构特征，可变换得到一个高效整洁的文本向量。可利用 tf-idf 对二维矩阵的句子中 id 标注的特征词汇转换成 tf-idf 词频权重向量，计算特征词汇在句子中权重大小，tfidf 向量中，用 corpus 中每个词的稀缺性的权重代替词频。

基于 bow 向量或 tf-idf 词频权重向量，采用 gensim 内置的主题模型，如 LDA、LSI、RP、HDP 等，生成相应的主题向量，进而构建相似度矩阵。

以《殷武》最后一章的“引文”、“注”和“疏”为例，其内容如下所示：

万方数据

《诗经》原文：

“陟彼景山，松柏丸丸。是斷是遷，方斫是虔。松桷有梃，旅楹有閑。寢成孔安。”

《毛诗正义》注：

丸丸，易直也。遷，徙。.....沈音倫，理也。

《毛诗正义》疏：

[疏]“陟彼”至“孔安”。○毛以爲，.....詩人之意主美寢也。

将《毛诗正义》含有上述《诗经》原文内容的“引文”，与其“注”和“疏”进行相似度计算。将“注”和“疏”视为corpus，保留“引文”在corpus中出现的词频大于1的词汇，对剩余的词汇进行主题向量转换和相似度计算。统计分析知：corpus分成了70个句子，获取到118个特征词，“引文”中的特征词有21个，如图5-4和图5-5所示：

```
{'又': 77, '故': 66, '閑': 15, '路寢': 28, '方': 6, '于': 37, '丸丸':
4, '教': 53, '人': 82, '者': 43, '作': 75, '始': 106, '云': 29, '彼':
0, '義': 96, '經': 105, '孔': 16, '殷': 110, '宗': 62, '椽': 99,
'安': 18, '棲': 74, '宜': 104, '梃': 14, '擇': 79, '斷': 8, '知':
100, '其': 56, '乙': 117, '傳': 95, '音': 68, '立': 111, '角': 73,
'所': 55, '有': 12, '昇': 41, '未': 107, '焉': 59, '施': 57, '貌':
25, '長': 26, '然': 86, '爲': 47, '釋': 98, '室': 90, '也': 22, '復':
67, '楹': 13, '陳': 27, '連': 76, '居': 51, '宫': 103, '注': 70,
'廟': 118, '大': 83, '上': 46, '高': 60, '亦': 101, '陟': 2, '王':
50, '美': 91, '旅': 11, '不': 58, '取': 45, '列': 89, '則': 102,
'而': 39, '小': 113, '松柏': 3, '明': 108, '易': 21, '箋': 33, '得':
54, '既': 49, '政': 52, '反': 72, '斫': 5, '衆': 34, '以': 35,
'前王': 63, '椹': 30, '材': 38, '脩': 65, '能': 92, '敬': 24, '庚':
114, '遷': 9, '揄': 80, '虔': 7, '直': 20, '辛': 116, '曰': 97, '至':
81, '耳': 78, '徙': 23, '此': 94, '山': 85, '寢': 19, '是': 10, '盤':
115, '工': 88, '寢廟': 61, '甚': 48, '文': 71, '木': 44, '景山': 1,
'崩': 112, '與': 36, '同': 69, '正': 40, '無': 87, '之': 31, '謂':
32, '弟': 109, '廢': 64, '桷': 42, '言': 93, '成': 17, '事': 84}
```

图 5-4 corpus特征词及词频

```
{'旅': 14, '孔': 17, '斫': 8, '有': 13, '方': 9, '寢': 20, '虔': 10,
'楹': 15, '是': 7, '松柏': 3, '陟': 2, '安': 19, '景山': 1, '梃': 12,
'丸丸': 4, '松桷': 11, '斷': 5, '遷': 6, '閑': 16, '成': 18, '彼': 0}
```

图 5-5 “引文”特征词及词频

```
[['陟', '彼', '景山'], ['松柏', '丸丸'], ['是', '斷', '是', '遷'],
['方', '斫', '是', '虔'], ['松桷', '有', '梃'], ['旅', '楹', '有',
'閑'], ['寢', '成', '孔', '安']]

[[(0, 1), (2, 1), (4, 1)], [(1, 1), (3, 1)], [(5, 1), (8, 1), (10,
2)], [(6, 1), (7, 1), (9, 1), (10, 1)], [(11, 1), (12, 1)], [(12,
1), (13, 1), (14, 1), (15, 1)], [(16, 1), (17, 1), (18, 1), (19, 1)]]
```

图 5-6 “引文”特征词二维矩阵转换

万方数据

上述图 5-4 和图 5-5 中，每个词后的数字代表的是每个特征词的特有 id，是每个词的专有属性，进行 bow 向量或 tf-idf 向量转换后，可依据需求转换成主题向量进行相似度计算。图 5-6 是特征词进行二维矩阵转换的结果，其中第一维是每个特征词的 id 标识，第二维是特征词在句子中的词频。将上述的二维矩阵转成 tf-idf 向量，效果如图 5-7 所示：

```
[(0, 0.5405604418232821), (2, 0.5948925990192759), (4, 0.5948925990192759)]
.....
[(1, 0.6778675031920496), (3, 0.7351840913105891)]
.....
[(5, 0.39660925928302954), (8, 0.46206652903541234), (10, 0.7932185185660591)]
.....
[(6, 0.40701032969040846), (7, 0.5236381586672796), (9, 0.5679139980048379), (10, 0.4874621638476231)]
.....
[(11, 0.7296268053402812), (12, 0.6838455417189876)]
.....
[(12, 0.4706417516891375), (13, 0.5394157063716362), (14, 0.5394157063716362), (15, 0.44334832017969544)]
.....
[(16, 0.5383693279890951), (17, 0.45106041783061007), (18, 0.37072869066029074), (19, 0.6076703086940747)]
```

图 5-7 “引文”tf-idf 词频权重向量

“引文”中有 7 个句子，为此有 7 组 tfidf 向量矩阵，其第一维是特征词的 id，第二维是词频权重值。

5.4.3 相似度算法对齐实验及结果分析

基于 tf-idf 词频权重向量，训练可得到一个 LSI 模型。LSI 最核心的意义是将训练文档向量组成的矩阵 SVD 分解，并进行一个秩为 2 的近似 SVD 分解，可实现将文档映射到一个二维的 topic 空间中，如图 5-8 所示，将 tf-idf 向量通过 gensim 转换成 LSI 主题向量。

```
tfidf = models.TfidfModel(corpus)
corpus_tfidf = tfidf[corpus]

lsi = models.LsiModel(corpus_tfidf,id2word=dictionary_corpus,num_topics=100)
corpus_lsi = lsi[corpus_tfidf]
index = similarities.MatrixSimilarity(lsi[corpus])
```

图 5-8 主题向量转换代码

进行 LSI 主题向量转换后，可计算特征词的相似度，依据相似度值的高低，设定阈值或筛选条件选定目标对象。《殷武》为例，如图 5-9 所示，LSI 模型可按相关度高低进行主题的输出，如图 5-9 所示：

万方数据

```
topic #0(2.210): -0.243*"之" + -0.214*"也" + -0.200*"焉" + -0.186*"斫" + -0.183*"易" + -0.183*"以" + -0.157*"椹" + -0.148*"有" + -0.148*"是"
topic #1(1.775): -0.335*"貌" + -0.292*"長" + -0.291*"擬" + -0.236*"爲" + 0.232*"路廋" + -0.203*"閒" + 0.201*"廋" + -0.194*"旅" + -0.172*"撮"
topic #2(1.753): -0.279*"反" + 0.253*"有" + -0.217*"椹" + -0.199*"斫" + -0.181*"易" + 0.176*"貌" + 0.176*"廋蒯" + 0.174*"政" + 0.173*"脩" +
topic #3(1.647): -0.811*"同" + -0.249*"注" + -0.173*"反" + -0.142*"音" + -0.127*"擬" + -0.127*"廋" + -0.105*"安" + -0.105*"路廋" + 0.091*"易
topic #4(1.620): -0.253*"安" + 0.241*"裹蟲" + -0.238*"廋" + 0.235*"脩" + -0.232*"成" + -0.213*"路廋" + 0.199*"同" + -0.178*"居" + -0.176*"王
topic #5(1.548): -0.580*"度" + -0.530*"敬" + 0.182*"直" + 0.169*"同" + 0.166*"易" + 0.140*"丸丸" + -0.136*"貌" + -0.108*"篹" + -0.108*"擬" +
topic #6(1.517): -0.460*"立" + -0.460*"崩" + -0.388*"小" + -0.293*"弟" + -0.241*"乙" + -0.197*"辛" + 0.152*"同" + -0.152*"庚" + -0.152*"盤"
topic #7(1.497): -0.367*"陳" + 0.306*"反" + -0.300*"旅" + -0.261*"同" + 0.216*"陟" + 0.212*"椹" + -0.202*"敬" + 0.181*"貌" + -0.180*"列" + 0
topic #8(1.481): 0.576*"音" + 0.229*"陳" + 0.220*"旅" + 0.215*"角" + -0.205*"同" + 0.205*"反" + -0.188*"徙" + -0.183*"貌" + 0.179*"也" + -0.
topic #9(1.476): 0.475*"音" + -0.282*"椹" + 0.250*"斷" + -0.238*"旅" + -0.221*"反" + -0.219*"陳" + 0.198*"徙" + 0.188*"是" + 0.168*"遷" + -0
topic #10(1.402): 0.312*"徙" + -0.301*"易" + 0.276*"椹" + 0.264*"音" + -0.256*"反" + 0.218*"旅" + 0.210*"謂" + 0.198*"篹" + -0.175*"丸丸" +
topic #11(1.366): -0.378*"掄" + -0.365*"擇" + -0.314*"徙" + -0.207*"注" + -0.194*"陳" + -0.189*"也" + 0.189*"爲" + -0.183*"云" + -0.177*"有"
topic #12(1.353): -0.281*"有" + -0.248*"丸丸" + 0.246*"爲" + 0.230*"掄" + 0.228*"云" + 0.212*"擇" + -0.184*"至" + -0.184*"彼" + -0.180*"陟"
topic #13(1.332): 0.234*"陟" + -0.231*"裹" + -0.218*"曰" + 0.215*"衆" + -0.187*"篹" + 0.187*"旅" + 0.178*"孔" + 0.178*"反" + 0.176*"安" + 0.
topic #14(1.315): 0.485*"徙" + 0.239*"反" + 0.223*"所" + 0.195*"得" + -0.185*"廋" + -0.174*"經" + 0.174*"遷" + 0.172*"其" + -0.163*"廉" + 0.
topic #15(1.280): -0.327*"至" + -0.272*"陟" + -0.219*"傳" + 0.201*"反" + -0.190*"彼" + 0.174*"易" + 0.167*"直" + 0.167*"而" + 0.163*"耳" + 0
```

图 5-9 LSI 相关度前 15 的 topics

```
[('徙', 0.31235157425060212), ('易', -0.30098257200987749), ('椹',
0.27618199510474051), ('音', 0.26445525690208849), ('反',
-0.25561642205647472), ('旅', 0.21849078143794515), ('謂',
0.20959227258769619), ('篹', 0.19831443058289827), ('丸丸',
-0.17476656266244631), ('釋', 0.1646350264260327), ('傳',
-0.15447530639643145), ('擇', -0.1447032052022042), ('遷',
0.1386662464060297), ('直', -0.13861467342296901), ('掄',
-0.13831058684898853), ('斷', 0.13742069908925508), ('其',
-0.13439693195462021), ('衆', 0.12660125959264115), ('故',
-0.12329450208120005), ('宜', -0.11570813784822176), ('以',
-0.11548742997194517), ('文', 0.11510585807180856), ('長',
-0.10453131904906685), ('未', -0.10270375841118486), ('敬',
-0.097387861129336908), ('之', 0.095009438298086057), ('又',
0.094389730525137341), ('曰', 0.094026786349223429), ('棧',
0.092985363720461303), ('正', 0.092006089985335854)]
```

图 5-10 相关度排第十的 topic

图 5-9 中的 topic 相关度最高是 2.21，LSI 以 topic 的值越高表明相关度越大。以每个特征词的相关度综合所得。图 5-10 是详细地展示排序第十的 topic 所在的句子相关度状况。由此可看出 LSI 的值有正负之分，这是进行 SVD 转换导致的。

依照 LSI 模型设置一个阈值可将相关度更高的句子抽取出来，进而实现该篇中“引文”与“注”和“疏”的匹配词汇的抽取，进而引申将相关的句子提取出，进而将“引文”中的句子和“注”及“疏”中的内容对应上，形成一个整体的对齐模式。

万方数据

此页不缺内容

万方数据

第六章 对齐实验效果测评

本文为探究《诗经》与其注疏文献《毛诗正义》的句子对齐，通过构建训诂术语模式库，采用训诂术语正则匹配的方式初步进行探索。gensim 可考虑词语的语义关系进行句子相似度计算，为此，本文还将正则匹配结果与 gensim 计算句子相似度的方法相结合，进而综合比较两种方法的差异。

6.1 测评方法

6.1.1 测试语料

对齐实验效果的评价对象是指本文采用不同方法实现了《诗经》与《毛诗正义》的句子对齐，在不同实验方式下得到的最终对齐结果。进而在此基础上，依靠通用的效果测评指标 P、R、F 来评价不同方法下的对齐效果。

采用正则表达式匹配方法实现本文的句子对齐主要是结合行文结构特点，利用归纳得到的训诂术语模式库，采用正则表达式进行多次匹配，从而达到句子对齐的目的。在本实验中进行匹配是将《毛诗正义》分“注”和“疏”两部分，利用训诂术语模式库对“注”和“疏”的内容分别进行匹配，每组匹配都可得到前后两部分内容，其中一部分内容包含的是与《诗经》原文相近的内容，也是《毛诗正义》引用《诗经》原文的内容。如下图 6-1 所示。

万方数据

莫莫——成就
莫莫，成就之——貌
箋云：成就者——其可采用
箋云：成就——其可采用之時
箋云：成就者，其可采用之——時
o莫，美博——反
o莫——美博
濩——煮
濩，煮之——也
精曰絺，粗——綌
精曰絺，粗曰——綌
斁，厭——也
古者王后織玄紞，公侯夫人紘綖——卿
古者王后織玄紞，公侯夫人紘綖，卿之内子大帶，大夫命婦成祭服，士妻朝服，庶士以下各衣其——夫
箋云：服，整——也
女在父母之家，未知將所適，故習之以絺綌煩辱之事—乃能整治
女在父母之家，未知將所適，故習之以絺綌煩辱之事，乃能整治之無厭倦，—其性貞專
女在父母之家，未知將所適，故習之以絺綌煩辱之事，乃能整治之無厭倦，是其性貞—專
o艾，本—作"刈"，魚廢反

图 6-1　正则表达式匹配“注”一次匹配的前后两部分内容

从图 6-1 可看出，借助训诂术语模式库的正则匹配，可将《毛诗正义》中符合模式的所有的内容都匹配出来，当然会存在重复匹配，匹配不精确的现象，为此需要在上述匹配结果的基础上，进行多次匹配，每次得到的匹配结果会进行预估，计算其在《诗经》原文中出现的概率，以此判定是否进行再次匹配，从而得到更准确的匹配结果，并将最终的匹配结果用于效果测评。

采用 Gensim 相似度计算方法进行句子对齐，则是利用 gensim 算法对《毛诗正义》中“注”和“疏”的内容分别进行匹配，可直接得到相似度的大小，进而得到匹配结果。其匹配结果也会有两部分，其中一部分与《诗经》原文的内容关联度更大，可从中抽取得到词表。匹配结果如图 6-2 所示。

万方数据

汙，煩也——汙
o汙音烏-汙
衣，謂褘衣以下至褖衣——衣
褘音煇，王后六服，一曰褘衣—衣
私，燕服也—私
其餘則私也—私
寧，安也——寧
父母在，則有時歸寧耳——寧
父母在，則有時歸寧耳——父母
私服宜澣，公服宜否—澣
害，何也——害
o害，户葛反，下同—害
父母在，則有時歸寧耳——歸
私服宜澣，公服宜否—否
箋云：我之衣服，今者何所當見浣乎？何所當否乎？
言常自潔清，以事君子—否
否，方九反—否
卷耳，苓耳也——卷耳

图 6-2 gensim 相似度匹配“注”一次匹配的前后两部分内容

从图 6-2 可看出，初次进行匹配的结果还比较粗糙，还需要对匹配到的句子较长的内容进行再次匹配，尽可能多地从中抽取出《诗经》原文中出现的内容，这样的匹配效果将更好。

6.1.2 测试方法

采用 SIGHAN 的测评标准，测评指标计算的前提是需要有一个已经标注好的参照数据集。为此，本文针对不同对齐方法实验得到的最终结果中，抽取部分数据进行人工标注，以便得到标准测试集。

在《毛诗正义》实验语料中，标注用逗号进行划分的“注”部分内容，所有被切分的“注”部分数据的句子总数达到 17954 行，从中选取 3000 行进行标注，用作测试数据的标准数据集，将抽取出的用作标注的数据集标注的方式，是将数据集中含有《诗经》原文中的词汇标注出，已标注的数据集如图 6-3 所示。

万方数据

o【皦】，本又作"皎"，古曉反
【壙】，苦晃反
【别】，彼列反
【留】，大夫氏
【子嗟】，字也
丘中墝埆之處，盡有麻、麥、草、木
，乃彼子嗟之所治
箋云：子嗟放逐於朝，去治卑賤之職
而有功，所在則治理，所以爲賢
o【墝】，本亦作"墽"，苦交反
【埆】，苦角反，又音學
本或作"遠"，此從孫義而誤耳
【施施】，難進之意
箋云：【施施】，舒行，伺間獨來見
己之貌
o【將】，王申毛如字，鄭七良反，
下同
【施】，如字
【伺】音司
【間】音閑，又如字
【子國】，子嗟父

图 6-3 已标注的标准数据集示例

从图 6-3 可看出，用方括号“【】”将出现在《毛诗正义》中的《诗经》原文中的词汇标注出。将抽取出的 3000 行未标注的数据作为测试数据，用不同的对齐方法对上述数据进行对齐实验，将得到的结果，与已标注数据中正确匹配的数据进行对比，从而评价相应对齐方法的对齐效果及方法的有效性。

6.1.3 测评指标

本文为探究《诗经》与《毛诗正义》的句子对齐，进行了多组实验，单独使用训诂术语正则匹配对齐实验，仅采用 gensim 相似度计算对齐实验，也有将训诂术语正则匹配方法与相似度计算方法相结合的对齐实验。为考察不同方式在实现句子对齐上的对齐效果，采用准确率 P、召回率 R、调和平均值 F 来衡量本文对齐实验的效果。

针对句子对齐实验的准确率 P 指的是在对齐中，匹配正确的句子数在测试语料中的总句子数。召回率 R 则是匹配正确的句子数占所有匹配正确的句子数的比例，调和平均值 F 则是 P 和 R 的调和平均值。通过考察在句子对齐实验过程中，通过统计匹配到的词数或句子数的比例来衡量方法的有效性，从而有针对性的进行改进。

万方数据

$$P = \frac{\text{正确对齐的句子数}}{\text{分析的句子数}} \times 100\% \quad (1)$$

$$R = \frac{\text{正确对齐的句子数}}{\text{实际存在的句子数}} \times 100\% \quad (2)$$

$$F = \frac{2PR}{P+R} \times 100\% \quad (3)$$

召回率的计算需基于手工分析的对齐句子作为标准答案。准确率则是以匹配正确的句子数与分析的句子数的比值来衡量。

6.2 训诂术语正则匹配实验测评及结果分析

本文进行了训诂术语正则匹配实验和 gensim 相似度计算实验，实验的效果不一。现通过 P 来测算不同方法对齐的效果。

正则匹配实验主要是依靠构建的训诂术语模式库，依据模式库对《毛诗正义》中的训诂条例进行抽取，从中获取的到和《诗经》经文相关的词汇，往回追溯将其所在的句子找到，从而将“引文”中的句子与“注”和“疏”的相应句子对齐对应起来。

正则匹配实验进行了两类实验，分别是“引文”与“注”的匹配，“引文”和“疏”的匹配。实验的匹配结果可分为前后两部分，即将正则表达式的前后两部分的匹配结果分开，正则表达式的前部视为第一部分，通常是被释对象；正则表达式的后部视为第二部分，通常是训释内容。由于匹配结果的前后两部都有可能存在“引文”中的词汇，因此将前后两部分匹配内容均进行抽取效果的测试。鉴于训诂术语正则表达形式中含有标点符号，导致匹配结果将前后的标点切分开，易出现单个字或多字词的前后带有一个标点符号，这就给效果的测试带来了干扰，为此，测试前需对匹配结果进行一定的处理，将字或词前后的标点符号去除后，再进行词频统计及效果测试。

对正则匹配实验的实验结果进行测试，将匹配正确的词抽取出，计算其在所有匹配内容的概率，即为抽取效果的准确率。还应将匹配正确的词回溯到《毛诗正义》语料的句子中，从而确定在其句子对齐的准确率。

“引文”和“注”部分匹配结果展示如图 6-4 至 6-6 所示：

万方数据

水中可居者　　洲↓
摯之　　至也↓
雌雄情意至　　有別↓
案　　之↓
案　　興是譬諭↓
案，興　譬諭之名↓
故題　　興↓
摯本　　作鷙↓
怨耦　　仇↓
本　　作仇，音同↓
鄭云“怨耦　　仇”↓
以色　　妒↓
本　　作荅↓
服　　思↓
卧而不周　　輾↓
輾，本　作展↓
鄭云“不周　　輾”↓
萋萋　　茂盛↓

图 6-4　“引文”和“注”第一次匹配结果

>關關雎鳩，在河之洲。↓
>興也。關關，和聲也。雎鳩，王雎也。鳥摯而有別。水中可居者曰洲。后妃說樂君子之德。無不和諧。又不淫其色。慎固幽深。若關雎之有別焉。然後可以風化天下。夫婦有別則父子親。父子親則君臣敬。君臣敬則朝廷正。朝廷正則王化成。箋云。摯之言至也。謂王雎之鳥。雌雄情意至然而有別。雎，七胥反。鳩，九尤反。鳥之有至別者。洲音州。興，虛應反，沈許甑反。案，興是譬諭之名。意有不盡。故題曰興。他皆放此。摯本亦作鷙。音至。別，彼竭反。下同。說音悅。樂音洛。諧，户皆反。朝，直遥反。廷，徒佞反。↓
水中可居者　　洲↓
摯之　　至也↓
雌雄情意至　　有別↓
案　　之↓
案　　興是譬諭↓
案，興　譬諭之名↓
故題　　興↓
摯本　　作鷙↓
>窈窕淑女，君子好逑。↓

图 6-5　“引文”和“注”第一次匹配结果配“引文”内容

万方数据

水中可居者↓	
摯之↓	仇↓
雌雄情意至↓	作仇，音同↓
案↓	仇”↓
案↓	妒↓
案，興↓	作荟↓
故題↓	思↓
摯本↓	輾↓
怨耦↓	作展↓
本↓	輾”↓
鄭云“怨耦↓	茂盛↓
以色↓	婦人↓
本↓	婦人之所有事↓
服↓	婦人之所有事也↓
卧而不周↓	葛延蔓於谷中↓
輾，本↓	和聲↓
鄭云“不周↓	鳥↓
萋萋↓	成就↓
	成就之↓
	成就↓

图 6-6 匹配“注”的第一部分和第二部分内容

图 6-4 展示的是“引文”和“注”进行正则匹配后的匹配结果，第一列是正则表达式前部的匹配结果，第二列是正则表达式后部的匹配结果。图 6-5 则详细的展示了进行匹配的“引文”是“關關雎鳩，在河之洲。”，其对应的“注”内容是“興也。關關，和聲也。.....廷，徒佞反。”，将两部分内容进行训诂术语正则表达式匹配后，产生的结果就是图中的下方的结果，从图中可见，该段内容中只有“洲”抽取正确，其他抽取的内容虽然符合训诂术语正则表达式的模式，但其实并非是“引文”中的内容。图 6-6 展示的是“引文”和“注”正则匹配第一部分和第二部分的匹配内容的汇总。

对上述的匹配结果的前后两部分内容都进行抽取效果的测试，将匹配结果的第一部分和第二部分都进行词频统计得到词频词表，将该词表与抽取出的词例数比对，如果抽取的词汇是原文中的词则认为抽取正确，准确率的计算则是抽取正确的词数占所有的总词数的比例，召回率则是看其在已标注的数据集中正确匹配的比例。测试结果如表 6-1 所示：

万方数据

表 6-1 训诂术语正则匹配“引文”和“注”的测评结果

	准确率 P	召回率 R	调和平均值 F
“引文”和注匹配第一部分	86.92%	82.30%	84.61%
“引文”和注匹配第二部分	84.14%	80.7%	82.42%

利用正则表达式匹配的方法对《诗经》和《毛诗正义》进行句子对齐时，选择以“注”部分的内容进行对齐实验的效果测评。从表 6-1 中可看出，“引文”和“注”匹配得到的第一部分的准确率 P 达到 86.92%，调和平均值 F 是 84.61%，匹配得到的第二部分的准确率 P 相较于第一部分要低 2.78%，这与训诂术语模式库中前后两部分匹配的对象相关，如果是与《诗经》经文相关度更高的内容则可以得到更高的准确率。但匹配得到的两部分的调和平均值均在 82%之上，这样的对齐效果在古文领域是可观的。

6.3 gensim 相似度计算实验测评及结果分析

gensim 相似度计算实验是对已实现“篇和章对齐”的语料，在每篇中将“引文”与“注”和“疏”进行相似度计算，从“注”和“疏”中抽取出相似度关联密切的句子，形成配对关系，实现句子对齐。

本文的 gensim 相似度计算实验有两类，一种是将“注”视为 corpus，将“引文”与“注”进行相似度计算；另一种是将“注”和“疏”视为 corpus，将“引文”与“注”和“疏”进行匹配，得到相应的匹配结果，从匹配结果中抽取出含有“引文”中的词汇，比对实验抽出的所有句子数，计算准确率，从而测试匹配效果。

以“引文”与“注”和“疏”的匹配结果为例，分析 gensim 相似度计算方法的效果。匹配结果如图 6-7 所示：

万方数据

1	雎鳩，王雎也，鳥摯而有別 雎鳩↓
2	○正義曰：毛以爲關關然聲音和美者，是雎鳩也 雎鳩↓
3	此雎鳩之鳥，雖雌雄情至，猶能自別，退在河中之洲，不乘匹而相隨也，以興情至，性行和諧者，是后妃也 雎鳩↓
4	“雎鳩，王雎也”，《釋鳥》文 雎鳩↓
5	陸機《疏》云：“雎鳩，大小如鴟，深目，目上骨露，幽州人謂之鶖雎鳩↓
6	俗本云“雎鳩，王雎之鳥”者，誤也 雎鳩↓
7	又解以“在河之洲”爲喻之意，言后妃雖悦樂君子，不淫其色，能謹慎貞固，居在幽閒深宫之内，不妄淫褻君子，若雎鳩之有別，故以興焉 雎鳩↓
8	此雎鳩之鳥，雖雌雄情至，猶能自別，退在河中之洲，不乘匹而相隨也，以興情至，性行和諧者，是后妃也 河↓
9	又解以“在河之洲”爲喻之意，言后妃雖悦樂君子，不淫其色，能謹慎貞固，居在幽閒深宫之内，不妄淫褻君子，若雎鳩之有別，故以興焉 河↓
10	水中可居者曰洲 洲↓
11	洲音州 洲↓
12	此雎鳩之鳥，雖雌雄情至，猶能自別，退在河中之洲，不乘匹而相隨也，以興情至，性行和諧者，是后妃也 洲↓
13	“水中可居者曰洲”，《釋水》文也 洲↓
14	《釋水》又曰“小洲曰渚”，“小渚曰沚”，“小沚曰坻” 洲↓
15	“江有渚”，傳曰：“渚，小洲也” 洲↓

图 6-7 “引文”与“注”和“疏”匹配结果

1	食	240↓
2	螽斯	8↓
3	雝	22↓
4	淇水	4↓
5	寤	27↓
6	埤	4↓
7	童	20↓
8	功	40↓
9	薖	23↓
10	刈	7↓
11	政	29↓
12	嘒	12↓
13	肅心	1↓
14	大姒	10↓
15	阿衡	9↓
16	語語	1↓
17	蓬	2↓
18	圃	8↓
19	徐國	4↓
20	被	40↓

图 6-8 “引文”与“注”和“疏”匹配得到的词频词表

万方数据

由图 6-7 可看出，gensim 相似度计算，将注疏含有“引文”中字或词的句子都抽取出来了，但句子切分的原因，可能会存在句子较长，有些对同一个字或词注释的长句子被切分开了，后续可针对这个问题进行改进。图 6-8 则是对匹配后得到的 token 进行词频统计得到的词表，该词表中的词是注疏中含有的“引文”的词汇。

Gensim 相似度计算匹配的结果如表 6-2 所示：

表 6-2 相似度计算匹配的测评结果

	准确率 P	召回率 R	调和平均值 F
经-注-疏匹配	89.92%	81.30%	85.60%
经-疏匹配	86.14%	82.7%	84.42%

由表 6-2 可看出，利用 gensim 相似度计算得到的句子匹配结果其实还理想。经-注-疏匹配的准确率可达到 89.92%，经-疏匹配的准确率有 86.14%。同时，调和平均值前者可达到 85.6%，后者是 84.42%。由此可看出，利用 gensim 相似度匹配的方法探索本文的句子对齐效果较佳。

正则表达式匹配和 gensim 相似度计算匹配这两种的方法其实是可以结合使用的，上述两种方法是从不同的角度实现《诗经》与《毛诗正义》的句子对齐，前者结合了《毛诗正义》文本结构特点和训诂语言特点，后者则在语义层面考虑到了词汇间的相关性，这都是对于实现句子对齐涉及的重要的方面，从不同的角度综合考虑，结合这两种方法，将能取长补短，在一定程度上增强对齐实验的效果。

万方数据

第七章 总结与展望

7.1 本文总结

本文进行《诗经》与其注疏文献《毛诗正义》句子对齐研究。力图通过文献梳理，归纳整理《毛诗正义》训诂术语的特点，构建训诂术语模式库，将其用于后续进行正则表达式匹配实验，将正则匹配结果与句子相似度计算的方法相结合，进而探索出一种用于《诗经》与其注疏文献的句子对齐方法。具体来说，取得了如下研究成果：

（1）进行文献调研，花费大量时间甄别了《诗经》及其注疏文献的版本，搜集、整理和校对了其相应的电子版本，并以北大整理本为标本，对不同数据来源的语料进行了校对工作，得到了《诗经》相关的宝贵资源。

（2）分析了《毛诗正义》注疏文献的注疏结构，界定了注疏平行对齐的形式化定义，并构建了《毛诗正义》训诂术语模式库，提出了基于训诂正则表达式匹配和相似度计算相结合的方法用于探索《诗经》与其注疏文献《毛诗正义》的句子对齐。

（3）指定了《诗经》分词规范，分别采用基于规则的自动分词方法，基于《毛诗引得》领域词表构建分词模型的机器学习自动分词方法。其中，对自动分词实验进行分词效果测评，采用机器学习的分词方法 F 值可达 97.39%，基于规则的自动分词方法的 F 值 84.5%，显然机器学习的分词效果明显更好。

（4）句子对齐结果，生成了基于《毛诗正义》注疏信息的分词词表，可用于后续进行《诗经》等先秦诗歌体文献的自动分词研究。基于条件随机场的机器学习自动分词后期进行《毛诗引得》领域词表校正，将能得到融合了引得专家词汇知识的《诗经》分词语料。

7.2 存在的不足

本文总结了《毛诗正义》的文本结构特征及其训诂术语特征，总结归纳出了《毛诗正义》训诂术语模式库。基于训诂术语模式库，采用训诂术语正则匹配和 gensim 相似度计算两种方法对文本进行匹配，抽取得到匹配的内容，从而得到一种探索《诗经》与其注疏文献《毛诗正义》的句子对齐方法。对齐实验产生的注疏词表后期可进一步用于《诗经》典籍的自动分词研究。但本文在研究过程中存在不足之处，具体如下：

（1）本文的句子对齐实现的是将《毛诗正义》中“注”和“疏”的内容与“引文”相对于的内容对应起来，但并没有对“疏”里面具体的内容来源进行区分。实际上，“疏”中涵盖了多位训诂学家的注释内容，未来可以将《毛诗正义》不同训诂学家的言论进行归纳汇总，分析各家言论所长。

万方数据

（2）本文所用的训诂术语模式库有进一步完善的可能，这样对于利用训诂术语模式库的正则表达式匹配结果将更好，此外，将训诂术语正则匹配方法和 gensim 相似度计算方法结合使用，将能更好地从词汇特征的不同角度进行考虑，从而获得更好的对齐，这部分后续研究可以补充完成。

万方数据

参考文献

[1] 江山丽.《诗经》及其注疏若干语法现象比较[D]. 湖南：吉首大学,2010:2.

[2] 韩宏韬. 《毛诗正义》单疏本考[J]. 河池学院学报(哲学社会科学版), 2006(06):44-48.

[3] 曹诣珍.《毛诗注疏》版本流变考略[J].古籍整理研究学刊, 2006(03):70-72+39.

[4] 龚抗雲, 李传书, 胡渐逵等, 北大整理本_毛诗正义(全三册)_汉毛亨传;郑玄笺;唐孔颖达疏[M], 2000:1-11.

[5] 谈春蓉, 李锐.《毛诗正义》的注疏成就[J]. 文学教育(下), 2007(08):32-33.

[6] 罗荣华.《诗经》三家注的语法观及其发展[D]. 宁夏：宁夏大学, 2004:6-60.

[7] 向熹.《诗经》语言研究古今谈[J]. 汉语史研究集刊, 2013(16):1-29.

[8] 刘挺颂. 20 世纪 80 年代以来《毛诗正义》研究述略[J]. 贵州教育学院学报, 2007(06):79-82.

[9] 韩宏韬.《毛诗正义》研究[D]. 济南：山东大学, 2007:11-197.

[10] 陈小荷, 冯敏宣, 徐润华. 先秦文献信息处理[M]. 2013:16-82.

[11] 马创新, 陈小荷,曲维光, 陆鹏飞.《论语》与其注疏文献对齐语料库的构建[J]. 现代教育技术, 2012(07):109-113.

[12] 马创新, 陈小荷,曲维光. 注疏文献中的注释语句自动分析[J]. 计算机科学, 2012(10):220-223.

[13] 梁社会.《孟子》及其注疏的信息处理[D]. 南京：南京师范大学, 2013:10-30.

[14] 徐润华, 陈小荷.一种利用注疏的《左传》分词新方法[J]. 中文信息学报, 2012(02):13-17+45.

[15] 王利局. 基于语义分析树核的句子相似度计算[D].大连：大连理工大学,2008:4-5.

[16] Chatterjee N. A statistical approach for similarity measurement between sentences for EBMT[C]//Proceedings of Symposium on Translation Support Systems STRANS-2001. 2001:30.

[17] ResnikP. Semantic Similarity in a Taxonomy: An Information-based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research, 1999(11):96-130.

[18] 梅家驹, 竺一鸣, 高蕴琦等.同义词词林[M]. 上海:上海辞书出版社出版, 1983:1-263.

[19] 刘群, 李素建. 基于《知网》的词汇语义相似度计算[J].中文计算语言学,2002(02):59-76.

[20] 易丽萍, 竹勇, 雷小春. 知网在词语相似度计算方面的应用[J]. 信息技术与信息化, 2005(01):24-26.

[21] Miller G A, Fellbaum C. Semantic of English. In Lexical and Conceptual Semantics[J]. Amsterdam, 1991:197-229.

[22] 关毅, 王晓龙. 基于统计的汉语词汇间语义相似度计算[A]. 哈尔滨工业大学计算机科学与技术学院、清华大学智能技术与系统国家重点实验室.语言计算与基于内容的文本处理——全国

万方数据

第七届计算语言学联合学术会议论文集[C].哈尔滨工业大学计算机科学与技术学院、清华大学智能技术与系统国家重点实验室,2003:7.

[23] Abney S. Part-of-Speech Tagging and Partial Parsing. In: Ken Church, Steve Young, Gerrit Bloothooft(eds.), Corpus-Based Methods in Language and Speech[J]. An ELSNET volume, Kluwer Academic Publishers, 1997:26-28.

[24] 李素健. 汉语组块计算的若干研究[D]. 北京:中国科学院研究生院(计算技术研究所), 2002:1-24.

[25] 骆正华,樊孝忠,刘林,龚永罡.基于 E-Chunk 的问句实例分析系统[J].北京理工大学学报,2005(01):63-66.

[26] 夏天. 中文信息处理中的相似度计算研究与应用[D]. 北京: 北京理工大学信息科学技术学院, 2005:1-35.

[27] Nirenburg, S. C. Domashnev and D.J.Grannes. "Two Approaches to Matching in example-Based Machine Translation."[J]. Proceedings of TMI-93, Kyoto, 1993:13-17.

[28] Ding C H Q. A similarity-based probability model for latent semantic indexing[C]//Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.ACM,1999: 58-65.

[29] Lambros Cranias,Harris Papageorgiou,Stelios Piperdis. A Matching Technique in Example-Based Machine Translation[J].In Proceedings of 1994 COLING,1994:100-104.

[30] Yang Q X,Yuan S S,Zhao L,et al. Faster Algorithm of String Comparison. Pattern Analysis and Applications[J].2003,6(2): 122-133.

[31] Bär D. A Composite Model for Computing Similarity Between Texts[J]. 2013: 23-30.

[32] 龙昊, 蔡东风. 一种基于词匹配的句子相似度计算方法[A]. Northeastern University、Tsinghua University、Chinese Information Processing Society of China、Chinese Languages Computer Society, USA.Advances in Computation of Oriental Languages--Proceedings of the 20th International Conference on Computer Processing of Oriental Languages[C].Northeastern University、Tsinghua University、Chinese Information Processing Society of China、Chinese Languages Computer Society, USA:,2003:7.

[33] 陈利人, 陈群秀. 基于实例的日汉机器翻译方法中的句子相似度计算研究[C]. //全国第三届计算语言学联合学术会议论文集.北京:天津医科大学, 1995: 257-263.

[34] 张民, 周明. 一种汉语句子间相似度的度量算法及其实现[C].//全国第三届计算语言学联合学术会议论文集.北京:哈尔滨工业大学%天津医科大学, 1995: 152-158..

[35] 李彬, 刘挺, 秦兵等. 基于语义依存的汉语句子相似度计算[J]. 计算机应用研, 2003(12):15-17.

万方数据

[36] 穗志方, 俞士汶. 基于骨架依存树的语句相似度计算模型[C]. //1998 中文信息处理国际会议论文集(ICCIP’98), 1998: 458-465.

[37] 杨思春, 程节华. 一种基于模式的汉语句子相似度计算方法[J]. 微型机与应用, 2001(8):52-53.

[38] 梁晗, 陈群秀, 吴平博. 基于事件框架的信息抽取系统[J].中文信息学报, 2006, 20(2): 40-46.

[39] 车万翔, 刘挺, 秦兵. 基于改进编辑距离的中文相似句子检索[J]. 高技术通讯, 2004(07): 15-19.

[40] 侯宏旭, 刘群, 那顺乌日图. 基于实例的汉蒙机器翻译[J]. 中文信息学报, 2007(04): 65-72.

[41] 秦兵, 刘挺, 王洋等. 基于常问问题集的中文问答系统研究[J]. 哈尔滨工业大学学报, 2003, 35(10):1179-1182.

[42] 马创新. 注疏文献的结构化知识表示[D].南京：南京师范大学, 2014:7-32.

[43] 方一新. 训诂学概论[M]. 南京: 江苏教育出版社. 2008:3-170.

[44] 陈炳哲.《毛传》、《郑笺》训诂术语比较研究[D]. 北京: 首都师范大学, 2005:1-67.

[45] [元]卢以纬著, 王克仲集注.《助语辞集注》[M]. 北京：中华书局, 1988:1-186.

[46] 张艳.《毛传》、《郑笺》对《诗经》训诂之比较[D]. 兰州: 兰州大学, 2007.45-47.

[47] 党代莉.《毛诗正义》训诂术语研究[D]. 兰州: 西北师范大学, 2011:1-55.

[48] 刘卫宁.《毛诗故训传》、《毛诗笺》与《诗集传》训诂比较研究[D]. 广州: 暨南大学, 2005:2-40.

[49] 陆德明.《经典释文》[M]. 上海：上海古籍出版社,1984:3-30.

[50] 高華娟. 《毛詩正義》注釋研究[D]. 重庆: 西南大学, 2013:1-69.

[51] 楊琳.《訓詁方法新探》[M]. 北京: 商務印寿館, 2011:3-272.

[52] 洪诚.《训话学》[M]. 南京: 江苏古籍出版社, 1984:1-40.

[53] 路广正.《训话学通论》[M]. 天津: 天津古籍出版社, 1996:141.

[54] 杨金花.《毛诗正义》研究[D]. 石家庄: 河北大学, 2007:12-116.

[55] 汉达文库[EB/OL]. [2005-04-13]. http://www.chant.org/.(CHANT-Chinese Ancient Texts.

[56] John Lafferty, Andrew McCallum and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. InProc .ICML, 2001:282-289.

[57] 严学窘. 国学经典导读广韵[M]. 北京：中国国际广播出版社, 2011:20-75.

[58] CRF++. [EB/OL]. [2017-2-15]. https://sourceforge.net/projects/crfpp/.

[59] 赵振铎. 《训话学纲要》[M]. 西安: 陕西人民出版社, 1987:6-26.

[60] 马树杉. 试论孔颖达的训诂方法(一)[J]. 常州工业技术学院学报, 1989(03):28-31.

[61] Behrouz A.Forouzan, Richard F.Gilberg 著.彭松虎译. UNIX 和 shell 程序设计权威教程[M]. 北京：清华大学出版社. 2003:40-60.

[62] 秦元坤. 正则表达式匹配中的 DFA 优化技术研究[D]. 北京：清华大学, 2008: 9-50.

万方数据

此页不缺内容

万方数据

附录一 繁体字校对样例

附录 1-1 繁体字校对样例

繁体字转化后《毛诗正义》繁体字	北大整理本《毛诗正义》繁体字	校正前《毛诗正义》繁体字句子样例
齣	出	《齣車》《出車》
禮	礼	禮記
瀋	沈	瀋尺遂反
鏇	旋	還音鏇
戍	戍	遣將帥及戍役
彆	别	雖三章三輩彆行
雲	云	《玉藻》雲
鞦	秋	當至鞦末始平
徵	征	將使我出徵伐
禦	御	禦夫也
於	于	於彼牧矣
廄	厩	仍有在廄供用者
閤	合	季春乃閤纍牛騰馬遊牝於牧
纍	累	-
係	繫	累係在厩者
二	弍	掌御二車
乾	干	旄，乾旄
緻	致	緻也
鬍	胡	不旆旆
馀	餘	旗音馀
茲	兹	御夫則茲益憔悴
遙	遥	慈遙反
製	制	凡製軍
韁	千	萬二韁五百人為軍
眾	衆	有軍眾者畫異物
錄	録	此錄入《雅》
戴	載	軍吏戴旗
錶	表	以國錶地
麴	曲	《麴礼》
傢	家	以王傢之事多危難
間暇	閑暇	其間不得間暇跪處也
奚	鷰	許奚反
敕	勅	忡，敕中反
戶	户	降，戶江反
鼕	冬	則以鼕日平西戎也
睨	睆	有睨然其實
逾	踰	室家逾時則思
讬	託	讬有事以望君子
薑	姜	莊薑稱莊公為父母，與此同也
為	爲	以檀木為車

万方数据

繁体字转化后《毛诗正义》繁体字	北大整理本《毛诗正义》繁体字	校正前《毛诗正义》繁体字句子样例
蔔	卜	蔔筮偕止
閤	合	會，閤也
有	言	文、武並有者
筍	笋	寡婦之筍也
摺	折	草木不摺，不操斧斤
群	羣	諸侯不掩群
鱉	鼈	鳥獸魚鱉皆得其所然
捨	舍	捨人云
曆	歷	則魚麗曆於罟者
簿	薄	以簿爲魚笋
總	揔	北風，冬風之總名
總	總	-
蠟	蜡	蠟則作羅襦
黃	黄	季秋草木黃落
匝	迊	不得圍之使匝
塚	冢	直塚反
遍	徧	遍檢諸本
跡	迹	乃跡其所用亡之早晚
麵	面	笙入堂下，磬南北麵
答	荅	《鄭誌》答炅模云

万方数据

附录二 词性标记集合及特征模板

附录 2-1 《诗经》分词词性标记集合

词性标记	词性	词性标记	词性
a	形容词	s	拟声词
c	连词	t	时间词
d	副词	u	助词
f	方位词	v	动词
i	词缀	vs	使动用法
j	兼词	vw	为动用法
m	数词	vy	意动用法
n	普通名词	w	标点
nr	人名	x	其他语素、字
ns	地名	y	语气词
nx	专名	za	形容词作状语
p	介词	zn	名词作状语
q	量词	zv	动词作状语
r	代词	-	-

附录 2-2 机器学习模型中的 23 组特征组合模板

序号	特征模板	特征模板说明
1	无词性	最简单特征模板，语料不加词性
2	无词性-类别	类别特征（标注语料是字、词还是标点符号）
3	无词性-反切	《广韵》字表反切特征
4	无词性-韵	《广韵》字表韵特征
5	无词性-调	《广韵》字表调特征
6	无词性-类别-反切	类别和反切两个特征
7	无词性-类别-韵	类别和韵两个特征
8	无词性-类别-调	类别和调两个特征
9	无词性-类别-反切-调	类别、反切、调三个特征
10	无词性-类别-反切-韵	类别、反切、韵三个特征
11	无词性-类别-韵-调	类别、韵、调三个特征
12	有词性	语料添加词性特征
13	有词性-类别	词性和类别两个特征
14	有词性-反切	词性和反切两个特征
15	有词性-韵	词性和韵两个特征
16	有词性-调	词性和调两个特征
17	有词性-类别-反切	词性、类别和反切三个特征
18	有词性-类别-韵	词性、类别和韵三个特征
19	有词性-类别-调	词性、类别和调三个特征
20	有词性-类别-反切-韵	词性、类别、反切、韵四个特征
21	有词性-类别-反切-调	词性、类别、反切、调四个特征
22	有词性-类别-韵-调	词性、类别、韵、调四个特征
23	有词性-类别-反切-韵-调	词性、类别、反切、韵、调五个特征

万方数据

此页不缺内容

万方数据

附录三 《毛诗正义》训诂术语模式库

附录 3-1 《毛诗正义》训诂术语模式库

类别	训诂术语	正则表达式
人名、物名	x，名 y　x，字 y　x，姓 y	(.+)，[名字姓].+
	x，y 名(也)	(.+)，.{1,5}名也?
	x，y 之	(.+)，.+之
动词	x，为 y	(.+)，为.+
	x，故为 y	(.+)，故为.+
	x，y 貌　x，y 之貌　x，y 之状	(.+)，.+(貌|之貌|之状|的样子|意也|之称|之言)
	x，y 的样子　x，y 意也	
形容词	x，y 之称　x，y 之言	(.+)，(者|为).+之貌
	x 者 y 之貌　x 为 y 之貌	(.+)，是.+之别名
	x 是 y 之别名	
代词	x，身也　x，自称之词	(.+)，(身也|自称之词)
	x，指 y 之辞　x，指 y 辞也	(.+)，指.+(之辞|辞也)
助词	x，y 辞也　x，y 辞　x，y 之辞	(.+)，.+(辞也|辞|之辞|语助也|语助)
	x，y 语助(也)　x，发声也	(.+)，发声也
叹词	x，y 叹辞(也)　x，y 之声	(.+)，(嗟叹辞也|叹辞)[，。]
	x，y 嗟叹辞也	(.+)，.+之声[，。]
象声词	x，y 声(也)	(.+)，.+(声也|声)
通用术语	x 者，y 也	(.+)者，.+也
	x 者，y	(.+)者，.+
	x，y (也)	(.+)，.+也?
	x，y 者	(.+)，.+者
	x，谓 y(也)	(.+)，谓.+也?
	x 谓 y　x 谓之 y	(.+)(谓|谓之|曰|之曰|为|亦|亦称|亦当言|是|言|指|犹言|犹云|然而).+
	x 曰 y　x 之曰 y	(.+)[，。](谓之|言).+
	x 为 y　x 亦称 y	(.+)，(犹言|犹云).+
	x 亦 y　x 亦当言 y	.+曰(.+)
	x 是 y　x 言 y	
	x 指 y　x 犹言 y	
	x 犹云 y　x 然而 y	
	x,谓之 y	
	x，言 y　x。言 y	
	x 而 y 兮　x，犹 y	
	x 犹 y(也)	
	x，犹言 y　x，犹云 y	
	y 曰 x　x，z 曰 y	

万方数据

此页不缺内容

万方数据

附录四　自动分词-全特征模板样例

文	n	HZ	無分	平	文	B↓
王	n	HZ	雨方	平	陽	E↓
之	u	HZ	止而	平	之	BE↓
什	n	HZ	是執	入	侵A	BE↓
思	n	HZ	息兹	平	之	B↓
齊	n	HZ	徂奚	平	齊	E↓
思	y	HZ	息兹	平	之	BE↓
齊	a	HZ	徂奚	平	齊	BE↓
大	nr	HZ	徒蓋	去	泰	B↓
任	nr	HZ	如林	平	侵A	E↓
，	w	JZFH	JZFH	JZFH	JZFH	BE↓
文	nr	HZ	無分	平	文	B↓
王	nr	HZ	雨方	平	陽	E↓
之	u	HZ	止而	平	之	BE↓
母	n	HZ	古丸	平	桓	BE↓
。↓	w	JMFH	JMFH	JMFH	JMFH	BE↓
畏	v	HZ	於胃	去	微	BE↓
此	r	HZ	雌氏	上	支A	BE↓
簡	n	HZ	古限	上	山	B↓
書	n	HZ	傷魚	平	魚	E↓
。↓	w	JMFH	JMFH	JMFH	JMFH	BE↓
王	n	HZ	雨方	平	陽	BE↓
命	v	HZ	眉病	去	庚	BE↓
召	nr	HZ	直照	去	宵B	B↓
虎	nr	HZ	呼古	上	模	E↓
，	w	JZFH	JZFH	JZFH	JZFH	BE↓

万方数据

此页不缺内容

万方数据

附录五 正则对齐实验结果

38 〉葛之覃兮，施于中谷。維葉萋萋。↓
39 〉興也。覃，延也。葛所以爲絺綌。女功之事煩辱者。施，移也。中穀，穀中也
貌。箋云。葛者，婦人之所有事也。此因葛之性以興焉。興者，葛延蔓於谷中，
之家。形體浸浸日長大也。葉萋萋然。喻其容色美盛。施，毛以豉反。鄭如字，
兮反。蔓音萬。浸，子鴆反。長，丁丈反。↓
40 萋萋　　茂盛↓
41 葛者　　婦人↓
42 葛　　婦人之所有事↓
43 葛　　婦人之所有事也↓
44 興　　葛延蔓於谷中↓
45 〉黃鳥于飛，集于灌木。其鳴喈喈。↓
46 〉黃鳥，摶黍也。灌木，藂木也。喈喈，和聲之遠聞也。箋云。葛延蔓之時，則
因以興焉。飛集藂木。興女有嫁於君子之道。和聲之遠聞。興女有才美之稱達
古亂反。喈音皆。摶黍，徒端反，鳥名也。叢，才公反，俗作“藂”。一本作
反。聞音問，又如字。下同。稱，尺證反。↓
47 喈喈　　和聲↓
48 摶黍，徒端反　　鳥↓
49 〉葛之覃兮，施于中谷。維葉莫莫。↓
50 〉莫莫，成就之貌。箋云。成就者，其可采用之時。莫，美博反。↓
51 莫莫　　成就↓
52 莫莫　　成就之↓
53 莫莫　　成就↓
54 成就者　其可采用↓
55 成就　　其可采用之時↓
56 〉是刈是濩，爲絺爲綌，服之無斁。↓
57 〉濩，煮之也。精曰絺，粗曰綌。斁，厭也。古者王后織玄紞。公侯夫人紘綖。

万方数据

此页不缺内容

万方数据

附录六 相似度对齐实验结果

[疏]“關關”至“好逑” 好↓
后妃既有是德，又不妬忌，思得淑女以配君子，故窈窕然處幽閒貞專之善女，宜爲君子之好匹也 好↓
○鄭唯下二句爲異，言幽閒之善女謂三夫人、九嬪，既化后妃，亦不妬忌，故爲君子文王和好衆妾之怨耦者，使皆説樂也 好↓
今江東呼之爲鶚，好在江邊沚中，亦食魚” 好↓
○傳“窈窕”至“好匹” 好↓
又曰“后妃有關雎之德，是幽閒貞專之善女，宜爲君子之好匹”者，美后妃有思賢之心，故説賢女宜求之狀，揔言宜求爲君子好匹，則揔謂百二十人矣 好↓
○正義曰：下箋“三夫人、九嬪以下”，此直云“三夫人以下”，然則九嬪以下揔謂衆妾，三夫人以下唯兼九嬪耳，以其淑女和好衆妾，據尊者，故唯指九嬪以上也 好↓
所以得有怨者，以其職卑德小，不能無怨，故淑女和好之 好↓
見后妃和諧，能化羣下，雖有小怨，和好從化，亦所以明后妃之德也 好↓
窈窕，幽閒也 窈窕↓
后妃既有是德，又不妬忌，思得淑女以配君子，故窈窕然處幽閒貞專之善女，宜爲君子之好匹也 窈窕↓
○傳“窈窕”至“好匹” 窈窕↓
正義曰：窈窕者，謂淑女所居之宫形狀窈窕然，故箋言幽閒深宫是也 窈窕↓
傳知然者，以其淑女已爲善稱，則窈窕宜爲居處，故云幽閒，言其幽深而閒静也 窈窕↓
言后妃有關雎之德，是幽閒貞專之善女，宜爲君子之好匹 女↓
言后妃之德和諧，則幽閒處深宫貞專之善女，能爲君子和好衆妾之怨者 女↓
后妃既有是德，又不妬忌，思得淑女以配君子，故窈窕然處幽閒貞專之善女，宜爲君子之好匹也 女↓
○鄭唯下二句爲異，言幽閒之善女謂三夫人、九嬪，既化后妃，亦不妬忌，故爲君子文王和好衆妾之怨耦者，使皆説樂也 女↓
正義曰：窈窕者，謂淑女所居之宫形狀窈窕然，故箋言幽閒深宫是也 女↓

万方数据

此页不缺内容

万方数据

致 谢

又是一年毕业季，回首在南农三年的研究生生涯，收获颇多。南京是一座容易让人喜欢的城市啊！感恩在研究生生涯遇到了一群亲切温暖的老师们，在学习和生活上，给予我支持和帮助；一帮可以一起奋斗，互相鼓励支持的小伙伴们。

首先，要由衷感谢黄水清教授，因为您的信任，我才有机会接触到古文，进行古文方面的深入研究，不仅扩宽了自己的眼界，也在科研探究能力上有了更大的进步。同样地，要由衷感谢我的导师何琳教授以及指导老师王东波副教授。两位老师渊博的知识、严谨的学风、无私奉献的精神给了我深刻的影响。何老师不仅在学习上给予我耐心细致的指导，而且在思想和生活上给了我莫大的鼓励，犹记得我遇到挫折时，何老师总是耐心地开导我，是如此的善解人意。可以说，我取得的每一点成绩，凝聚着三位老师的心血。

我要深深感谢信息院给予我关怀和帮助的所有老师们，尤其是汤亚芬老师，您安排好学习外的各种琐事，和蔼耐心的对待学生，让学生能够安心的学习，专心的学术，您是学生们的坚强后盾。还有阎素兰老师，两次作为您的助教，您总是很热心的关心我的学习和生活，总是鼓励我，真的非常感谢您。

此外，我还要感谢一起努力奋斗的小伙伴们，尤其是同门的文静、舍友周惠敏、一起并肩作战的李艳、夏丽君、曾磊，大家一起参加研会，一起鼓励支持着完成毕业论文，奔波着找工作，还有图书馆学、情报学专业的其他同学们在这三年对我的包容和鼓励。三年的朝夕相处，培养了深刻的友谊，我终身难忘。

最后，感谢我的父母，感谢父母从小到大的栽培，在我失落、迷惘的时候给予我无条件的支持和鼓励，读书 20 载，即将步入社会，希望以后能有更多的机会陪伴爸爸妈妈。还要感谢支持我的家人们，你们的鼓励是我前进的莫大动力，希望以后一家人能永远快乐。

万方数据

此页不缺内容

万方数据

攻读学位期间已发表的学术论文

1. 王姗姗, 王东波, 黄水清, 何琳. 多维领域知识下的《诗经》自动分词研究[J]. 情报学报, 2018, 37(02):183-193.

万方数据