文言信息的自动抽取：基于统计和规则的尝试[1]

虞宁翌[2]，饶高琦[1,2]，荀恩东[2]

（1 北京语言大学语言科学院，北京市 100083；2 北京语言大学信息科学学院，北京市 100083）

摘要：文言信息的自动抽取有利于语言监测和语料库构建。同时本文的计算研究也验证了语言学界关于汉语文白系统连续性的自省结论。本文将从混合语料中标注文言文的问题视为短文本分类的问题进行处理。使用基于规则和基于统计的方法对文言文、白话文本进行分类。基于规则的方法中，本文考虑文言常用虚词和句式的影响。在基于统计的分类方法中，本文对 N-gram、朴素贝叶斯、最大熵、决策树模型的性能进行了研究。结果表明监测虚词系统的一元语言模型的 F 值达到了 0.98。

关键词：文言标注，文本分类，规则模型，统计模型

An Attempt to Ancient Chinese Extraction based on Statistical and Rule based Methods

YU Ningyi[2], RAO Gaoqi[1,2], XUN Endong[2]

(1Faculty of Language Sciences, Beijing Language and Culture University; 2College of Information Sciences, Beijing Language and Culture University, Beijing 100083)

Abstract: The automatic extraction of ancient Chinese benefits language monitoring and corpus construction. The computational research in this paper also help to confirm the conclusion on Chinese evolution as a continuum. This paper regards the ancient Chinese tagging in mixed corpus as a task of short text classification . We research both rule and statistic based methods. For rule based methods, the paper considers the effect from function words and constructions in ancient Chinese. For statistical methods, we conduct experiments on N-gram, Naive Bayes, Maximum Entropy, and Decision Tree. The unigram model over performs others in F value of 0.98.

Key Words: ancient Chinese tagging, text classification, rule based model, statistic based model

1 引言

中国语言由古代文言文到现代白话文经过了近三千年的发展演变。排除字形的变化，语言本身在词汇、语法和篇章层面都产生了巨大变化，但却不失其连续性。这一特点使得在大时间跨度上研究汉语特征变化成为重要课题。对书面语进行文言/白话标注有助于对语言进行历时性的描写，分析语言风格，了解汉语书面语的发展情况。同时也方便对文言、白话混杂语料的分类和加工。

传统的语言学自省的方法有其固有的主观、高成本和缓慢的局限性。在文言、白话分类标注这一问题中引入自然语言处理的成熟方法和模型，结合专家自省知识，则有助于克服以上问题。本文研究中发现的特征和方法反过来又可以深化对汉语演变作为一个连续统的认识，因而具有一定的理论价值。本文的研究在一定程度上验证了王力先生提出的观点，即文言与白话的分野不在词汇与句式，而是虚词系统[1]。

在语料库构建的实践中，我们遭遇了文言文语料和白话文语料混合的情况。给语料库的科学平衡构建带来了一定困难。在语言生活的调研工作中，社会大众的文言使用情况是重要的调查目标。在现代书面语写作中文言、白话夹杂的现象也比比皆是，这给句法语义分析带来很大困扰。因而在大规模语料中通过计算手段自动标注文言文/白话文也具有重要的实践

[1] **基金项目：**国家自然科学基金项目（61300081，61170162）；国家高技术研究发展计划(863 计划) 2015AA015409；国家社会科学重大基金（12&ZD173）

和工程价值。本文研究发现使用用字的统计特征可以实现对文言文较为精确的标注。

文章的组织结构如下：第二节介绍简述了现有相关研究；第三节描述了语料和测试集的构建；第四节描述了基于规则的方法；第五节描述基于统计的方法；第六节是结论与展望。

2 研究现状

经过调研，与本文研究方向相同的研究工作并不多，相关的研究方向有汉语年代划分、用字特征、语言风格、中文文本分类等方向。语言的发展是一个有序，缓慢，逐步演变的过程。社会语言学的理论揭示：语言是在稳态中变化，在变化中保持稳态。稳态不同于静态。自然语言处理通常关注共时语料，也即一个时间切片上的语言数据。大规模语料库亦少对时间信息进行标注。而实际上，语言是不断发展变化的。语料数据亦有其时效性。这不仅表现在词汇短语的分布上，也表现在语义乃至语言风格上[2,3]。

石毓智对汉语发展的双音化趋势和动补结构进行了探究。汉语的双音化并非一蹴而就，是自从汉代以来逐渐发展的[3]。胡裕树在 1981 年对双音化情况进行过统计，不计词类区别，在 3000 个最常用的词中，75%是双音节词。还有大量未列入的双音节常用词。总体上说，双音词占汉语词汇的 80%以上[4]。吕叔湘在 1961 年提出，在很多情况下，单音节词只有加上一个音节（词缀）才能独立成词或作为句子成分[5]。

2012 年和 2013 年，Mihalcea 等和 Popescu 等[6,7]提出了时代消歧和时代检测两个任务及其基线。前者使用多种 Welch 测试，Run 测试，最小二乘、Ratio、斯皮尔曼和 Kendall 测试等统计方法来判断重要词语（尤其是政治相关词语）在近两百年的 Google N-gram Corpus 的分布，以判断其是否随机。由此来进行历史时期飞分割。Mihalcea 等提出的时代消歧任务在是在词语中挑选出具有时代区分力的词语。

在历时语料的建设方面，北京语言大学建立的现代汉语词汇历时检索系统，使用了《贵州日报》《福建日报》和《人民日报》共计 8 亿字、4.7 亿词，并提供在线检索[2]。时间跨度为 1949-2013 年[8]。北语汉语语料库 BCC 的文学频道则收集了时间跨度约为一百年的文学语料 24 亿字，并提供在线检索[9][3]。

3 测试集与统计基线

3.1 单句测试集

单句测试集包括 1372 句文言文和 1538 句白话文，共有 2900 句，文言文和白话文的数量大致平衡。文言文部分选用了《论语》中的单句。《论语》形成于我国春秋时期，是最早的语录体文集，记录了孔子及其弟子的言行。《论语》作为儒家经典文学，有悠久的历史，其中没有白话文成分，是典型的文言作品。《论语》有较为成熟的句读，易于程序切分为单句，方便使用。白话文部分采集了《人民日报》。《人民日报》是我国第一大报，使用了典范的现代汉语白话文，用字用词十分规范。

测试集的句子长度保持在 5-100 字之间。若句长小于 5 个字，句子中可判断特征不明显，实际可判断力过差，会降低测试结果的有效性。考虑单句的实际情况，句长超过 100 字的现象并不常见。古汉语的平均句长通常小于现代汉语的平均句长。若采用大量特殊的过长现代汉语作为测试集，可能影响标注，再则缺乏效力。

测试集样例：

文言文：〈文〉有朋自远方来，不亦乐乎？

〈文〉孝弟也者，其为仁之本与！

---

[2] http://nlp.blcu.edu.cn/historical20%computing

[3] http://bcc.blcu.edu.cn/index.php?corpus=1

白话文：〈白〉那么，增收的原因何在？
〈白〉移风易俗，提倡健康文明的生活习惯！

3.2 段落测试集

段落测试集包括 1050 段古汉语和 1050 段现代汉语，共 2100 段，古汉语和现代汉语的数量持平。文言文部分选用了《古文观止》和《全唐文》的段落。《古文观止》是历代文言散文总集，清康熙年间编纂。《全唐文》是唐代及五代十国的文言散文，清嘉庆年间编纂。两者均为清代中期前编纂，是典范的文言文作品，不包含白话文。而且其段落长度适中，适宜被选做段落测试集。

白话文部分选用了《人民日报》和《王朔文集》的段落。《人民日报》的段落中包含较多阿拉伯数字和字母，在文言文中没有阿拉伯数字和字母，因而不宜在测试集中使用，在寻找纯汉字段落之外，我们还引入了《王朔文集》。王朔从 20 世纪 80 年代开始写作，作品内容为典型的当代白话文。《现代汉语》等教材亦多使用其内容做例句、例文。

测试集中，段落长度基本保持在 100-300 字之间。段落测试集格式与单句测试集相同。段落测试集仅用于对单句测试集结果的补充验证。

3.3 报章体测试集

报章体测试集包括 1000 段梁启超的作品。梁启超是报章体文学的代表人物，他的作品文白相间，在近代中国具有很大的影响力。测试集被同时标为文言文和白话文两种形式，用于测试。段落长度基本保持在 100-300 字之间。该测试集并不用于测试方法性能，仅用于研究报章体文学的用字特性。

测试集举例如：

<白>|<文>　四曰厉国耻。务使吾国民知我国在世界上之位置，知东西列强待我国之政策，鉴观既往，熟察现在，以图将来。内其国而外诸邦，一以天演学物竞天择、优胜劣败之公例，疾呼而棒喝之，以冀同胞之一悟。

3.4 评测标准

测试集中每个单句或段落之前有文言文和白话文的区别标注。将测试集中的一个条目通过文言文和白话文的判别模型，通过字与字之间关联性的 log 值叠加，得到文言文与白话文得分。比较得分的高低，机械地将测试语句标注为古汉语或现代汉语，然后和原语句标注情况进行比较，分别获得白话文和文言文的正确率 P、召回率 R 和 F 值。

正确率 P = 提取出的正确信息条数 / 提取出的信息条数

召回率 R = 提取出的正确信息条数 / 样本中的信息条数

F 值=正确率 * 召回率 * 2 / (正确率 + 召回率)

3.5 基线 0

将测试集的结果全部判断为文言文或白话文。当全部判断为白话文时，白话文的正确率约为 0.529，召回率为 1，F 值约为 0.692；当全部判断为文言文时，文言文的正确率约为 0.471，召回率为 1，F 值约为 0.641。

在以下的实验中，测试集与训练语料均没有交叠。

4 基于规则的方法

4.1 用字特征

汉语在漫长的演变历史中存在双音化现象，也即越古老的文本中，越多的词语为单音节词，而越现代的则越多使用多音节词（双音为主）。在大多数情况下，古代的单音节词在现代汉语的译文中都以双音节词的形式出现。所以，在通常情况下，现代汉语的句长长于古汉语。以论语为例，原文总字数为 21475 字，某译文总字数为 29725。原文总字数约占译文总字数的 72.2%。

随着语言的演变，常见字集的内容出现了明显的转移。比如：文言文中常见的指示代词“斯”、“彼”等，在白话文中逐渐被“这”、“那”等所取代；文言文中常用的人称代词“尔”、“其”等，在白话文中表示为“你”、“他”等。常见字的出现情况对古汉语、现代汉语的区分可以起到一定的参考价值[1]。

通常认为，实词往往具有鲜明的时代特征。但是在本文任务中，实词需要谨慎对待。很多实词，如“经济”、“民主”、“国家”等，看似可以成为白话文的特征词，实则在其历史可追溯到中古乃至上古，只是其含义与今日不同罢了[10]。因而实词反而不适合作为判别特征来使用。

4.2 句式分析

在文言文中，特殊句式主要有四种，分别为：判断句、被动句、倒装句、省略句。有些句式可以用结句式直接表示出来，例如：判断句“……者，……也”、“……也”等，被动句“……见……于”、“为……所”等。还有一些无法用结句式直接表示出来，例如：倒装句、省略句。

在现代汉语中，特殊句式有六种，分别为：把字句、被字句、连动句、兼语句、判断句，存现句。其中，把字句、被字句可以直接由“把”字、“被”字判断，其他句式的判断很难形式化。但是，由于白话文中“把”字、“被”字不仅仅是介词，还会出现在其他词语里，所以仅凭“把”字、“被”字很难确定是否是把字句、被字句。文言文的特殊句式对文言文、白话文的区分可以起到的参考价值相对较大[11,12]。因此本文在基于规则的方法中使用文言句式来进行分析。

4.3 基于规则的实验

选取常见的古汉语 24 个虚词：之、乎、者、也、耶、矣、哉、於、吾、汝、尔、而、何、乃、其、且、若、所、为、焉、以、因、于、则。但是我们注意到，许多现代汉语的词中也包含有这些虚词。又因为测试集本身不做分词处理，因此我们从现代汉语词典文件中匹配含有该虚词的现汉词语，形成一个排歧词表。对于测试集句子，匹配到该虚词，且又不是排歧词表中的词语，则虚词数加 1。匹配结束后，返回该句虚词总数。

构造句式函数，将测试句输入。匹配测试句中是否出现下列句式：以“也”作为结尾，“……者，……也”，“为……所”，“无乃……于”。若出现一次句式，则句式数加 1。匹配结束后，返回该句句式总数。

规则方法（基线 1）：对于测试集中的句子，通过虚词函数和句式函数，若其中一个函数的返回结果大于 0，则输出句子为古汉语，反之，输出句子为现代汉语。

经过测评，白话文的判断正确率约为 0.821，召回率约为 0.458，F 值约为 0.588；文言文判断的正确率约为 0.594，召回率约为 0.888，F 值约为 0.712。

由测评结果可知，通过虚词和句式规则测评后，白话文判断的正确率较高，但是召回率不足，文言文判断的正确率不足，但是召回率较高。出现这种现象的原因主要有：1.文言文

中的常用虚词在白话文中仍有大量运用，且还是作为虚词运用。2.文言文中的词语在 bhw 中仍有运用。3.文言文中存在不包含虚词的单句。

这从一个侧面上反映了现代汉语和古汉语之间没有明确分界的事实。

4.4 基于规则的扩充实验

在基于规则的实验中，我们进行两方面的扩充：1.虚词。2.句式。

在虚词的扩充情况中，不仅仅考虑虚词是否存在，而是将虚词出现的次数与句长联系起来。虚词内容与上一实验相同，虚词出现次数通过虚词出现的次数减去含虚词的白话文词语（排歧词表内容）出现的个数得到，然后除以句子长度。

在句式的扩充情况中，将原来的 4 种句式扩充为 26 种句式，包括：句首的“夫”、“若夫”、“且夫”、“今夫”、“孰”、“吾”；标点前的“也”、“矣”、“焉”、“乎”、“诸”、“邪”、“哉”、“之”、“耶”、“曰”；以及固定搭配“如……何”、“若……何”、“奈……何”、“何以……为”、“何……之有”、“……者，……也”、“为……所”、“问于”、“之以”、“无乃……于”。对测试集语句进行匹配以考察其是否满足句式。

优化规则：在测试中，若满足句式或者虚词频率大于阈值 t，就判断句子为文言文，否则，为白话文。本文对虚词频率的阈值 t 进行了对比实验。

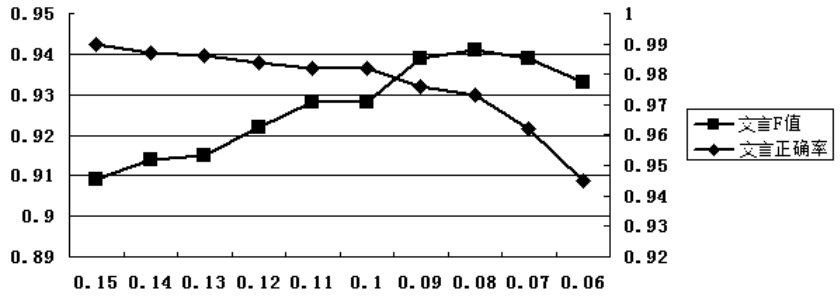

图 1　古汉语正确率和 F 值对比图

图 1 是文言文正确率和 F 值在虚词频率的阈值 t 改变情况下的变化情况。横坐标为虚词频率的阈值，主纵坐标为文言文 F 值，次纵坐标为正确率。由图可知，文言文正确率随 t 值减小，古汉语的 F 值在 t=0.08 的情况下最高达到 0.941。

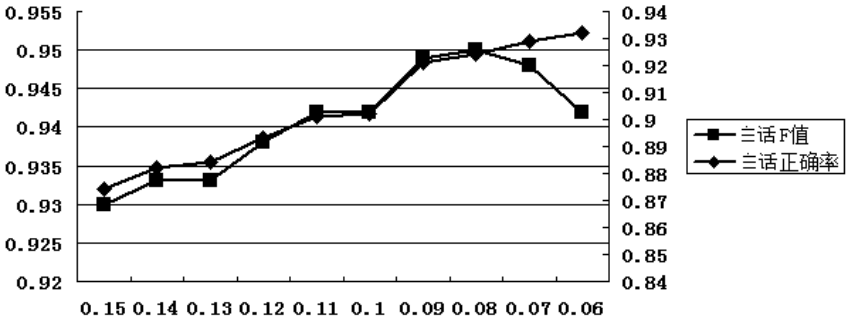

图 2　现代汉语正确率和 F 值对比图

图 2 是白话文正确率和 F 值在虚词频率的阈值 t 改变情况下的对比图。横坐标为虚词频率的阈值，主纵坐标为白话文 F 值，次纵坐标为白话文正确率。由图 2 可知，白话文正确率随 t 值减小而增大。白话文的 F 值在 t=0.08 的情况下最高达到 0.95。

综上所述，当虚词频率的阈值 t 为 0.08 时，扩充规则模型最优。由上一实验可知，虚词本身的存在对现代汉语的影响比较大，但是白话文的句长普遍长于文言文，且白话文虚词数

少于文言文的虚词数。所以，虚词数除以句长得到的虚词频率在白话文中会远远小于文言文，因此 t 值可以发挥其分类作用。图 3 为两种规则方法和基线 0 的 F 值比较。

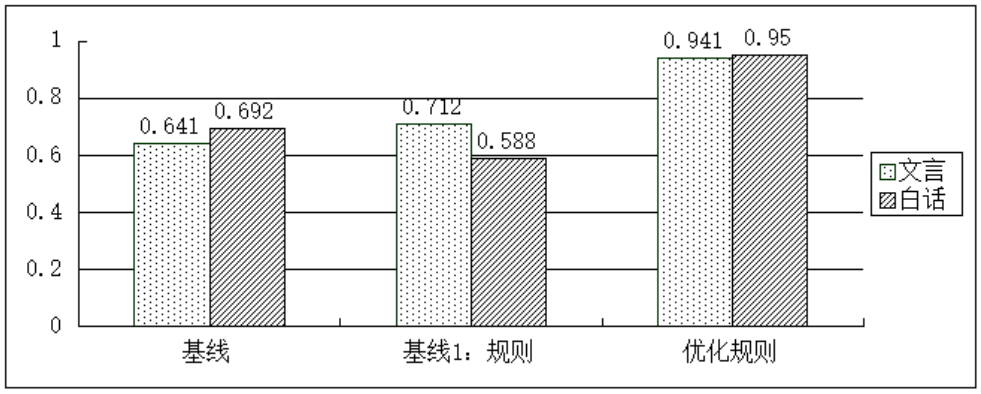

图 3　基线 0、规则和优化规则的 F 值

5 基于统计的方法

5.1　N-gram 语言模型

本文在 BCC 语料库古汉语频道选取清代中期以前的文言文语料 1 亿 5 千万字（gbk 编码下约 300M）和 2000 年前后的人民日报语料 1 亿 5 千万字（gbk 编码下约 300M）。我们使用 Cambridge-CMU language toolkit 实现了语言模型[13]。

选用单句测试集，在测试的过程中，将测试语句在一元、二元、三元状况下频率的 log 值相加作为分数。将在文言模型和白话模型中得到的分数对比。将句子标记为得分较高的模型。如：

1.有朋自远方来，不亦乐乎？

白话分值：-36.470589

文言分值：-33.058824

文言分值高于白话分值，则标记为文言。

将标记结果与测试语句人工标注结果对比，得到模型的正确率、召回率和 F 值。从中选取 F 值最高，且大小适中的模型为最优模型。本文认为 F 值越大，模型测试的结果越好。

图 4 是三元和二元语言模型的训练语料规模大小对标注 F 值的影响。

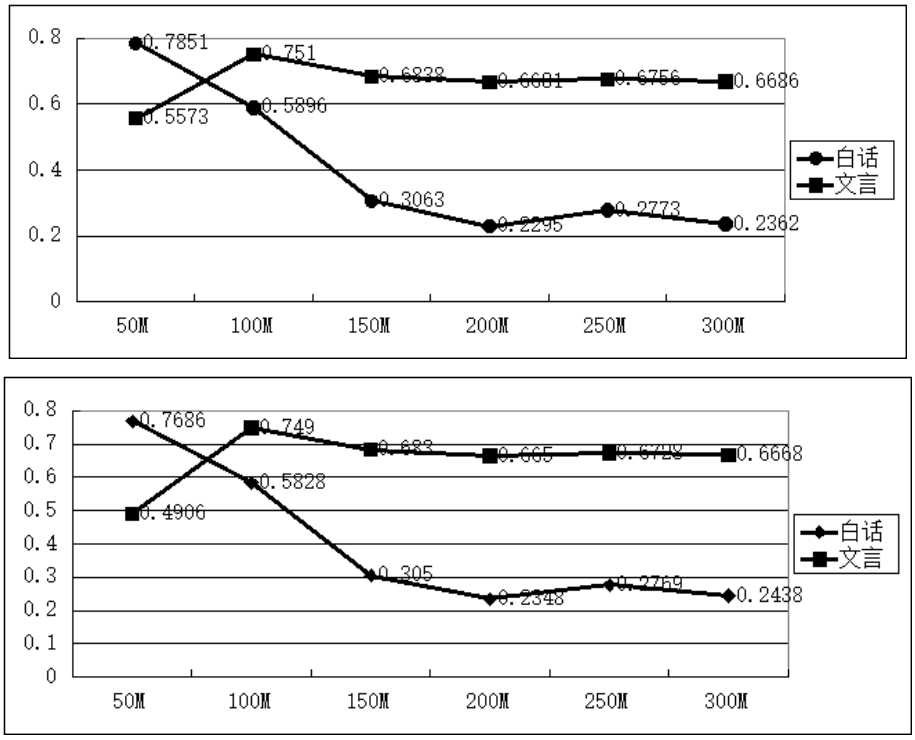

图 4　三元模型（上）和二元模型的对比

在三元模型中，白话模型测试结果的 F 值随模型语料的增大呈振荡下降趋势。在模型为 50M 时，F 值最大，约为 0.7851。在模型为 100M 时，F 值降为约 0.590。当模型为 150M-300M 时，F 值保持在 0.3-0.2 左右。

文言模型测试结果的 F 值当模型为 100M 时最大，最大值约为 0.751,。当模型为 50M 时，F 值最小，约为 0.557。当模型为 150M-300M 时，F 值保持在 0.67 左右。

在二元模型中，现汉模型测试结果的 F 值随模型语料的增大所呈现的趋势与三元模型相仿。古汉模型测试结果的 F 值当模型为 100M 时最大，最大值约为 0.749,。当模型为 50M 时，F 值最小，约为 0.491。当模型为 150M-300M 时，F 值保持在 0.67 左右。

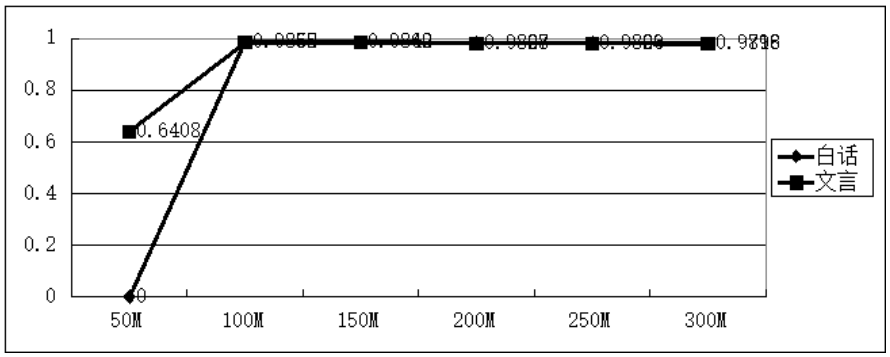

图 5　一元模型的表现

而在一元语言模型则呈现了巨大差异，白话模型测试结果的 F 值在模型为 50M 时为 0，测试集没有判断为现汉的结果，也即在较小的训练集上，文言文和白话文的用字差异无法得到体现。当模型为 100M 时，F 值最大，约为 0.985。当模型为 150M-300M 时，F 值基本不变，保持在 0.98 左右。图 6 为文言、白话在三元、二元、一元模型下最好 F 值的对比。

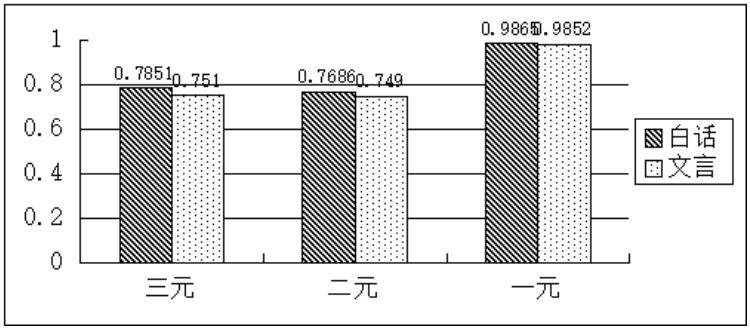

图 6 诸语言模型的标注表现（F 值）

其中，50M 的现汉模型在三元、二元情况下最优，其余情况下，均为 100M 的古汉、现汉模型最优。三元模型最好的 F 值，古汉约为 0.751，现汉约为 0.785。二元模型最好的 F 值，古汉约为 0.749，现汉约为 0.769。一元模型最好的 F 值，古汉约为 0.985，现汉约为 0.986。

经对比，在各模型不同元数下的标注结果中，一元状况下 100M 古汉现汉对比模型的标注结果最优。在接下来的实验中，主要针对 100M 模型进行测试、标注和优化。

5.2 段落测试实验

用段落测试集测试 100M 语言模型，以检测单句测试集中句子长度对于模型标注的偏差是否具有有限性。

图 7 为 100M 文白对比模型通过段落测试集后，在一元、二元、三元情况下的测试结果。

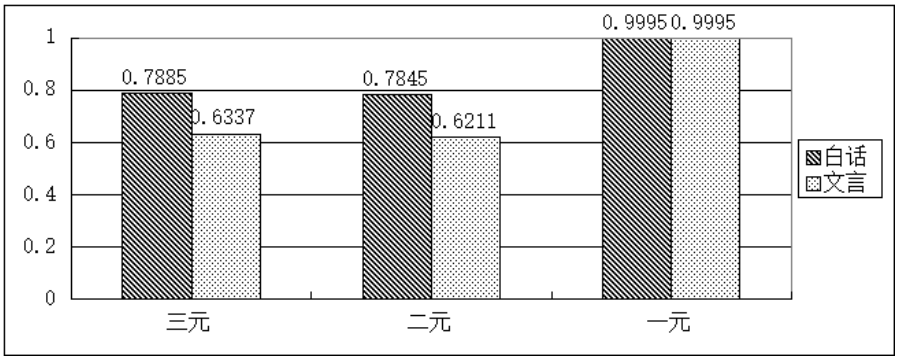

图 7　段落测试集测试结果

100M 模型经过段落测试集测试，测试结果大致与在单句测试集中相似。在一元模型中，古汉、现汉识别的 F 值大于 0.999，测试结果略优于单句测试集。由此可见，100M 模型一元情况下测试结果优秀不是偶然情况。

用报章体测试集测试 100M 语言模型，以检测报章体文学的用字特征。

图 8 为 100M 古汉现汉对比模型通过报章体测试集后，在一元、二元、三元模型中的测试结果。

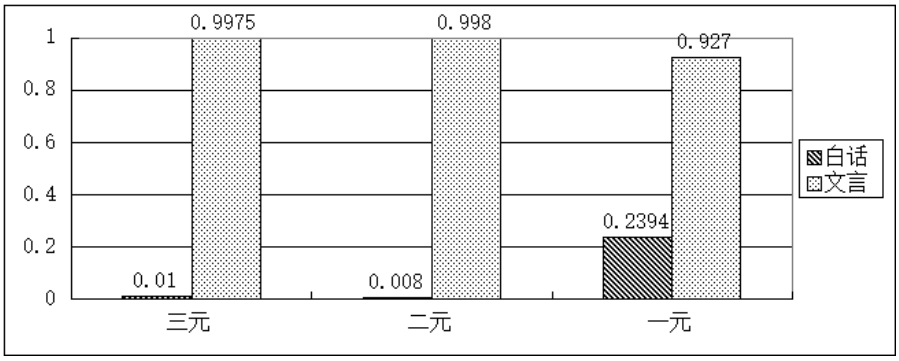

图 8　报章体测试集测试结果图

若报章体测试集被标记为文言文，在 100M 模型中被标注后，在一元、二元、三元模型中，F 值均在 0.9 以上。若报章体测试集被标记为现代汉语，F 值最小为 0.01，最大值为 0.239。由此可知，报章体大多会被模型识别为文言文。

据分析，报章体的主要句式基本与白话文相同，语法也与白话文类似。由于选用测试模型是基于字的统计模型，所以可以推测，报章体被判断为文言文的主要原因是大量使用文言文的基本用词。

5.3 基于机器学习的方法

本文还使用朴素贝叶斯、最大熵和决策树（ID3 算法）三种统计机器学习模型[4]进行了标注实验。我们选取 10M 古汉语单句语料和 10M 现代汉语单句语料。使用特征为标注、行号与字符串（基于字）。其中最大熵模型表现最好，F 值达到了 0.967 和 0.968。

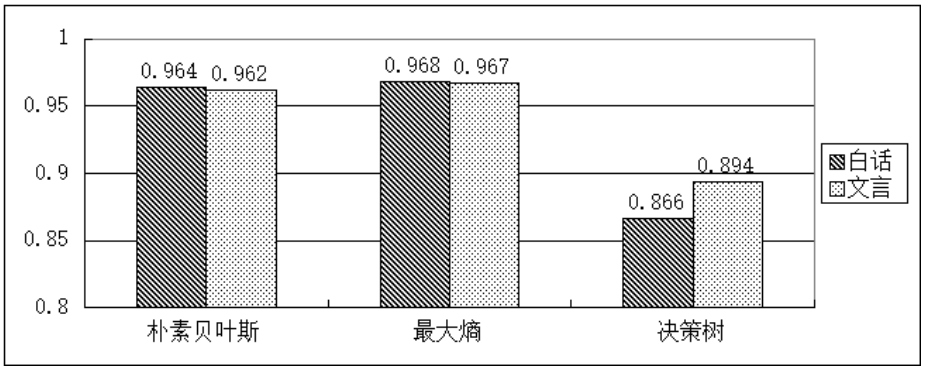

图 9　朴素贝叶斯、最大熵、决策树结果对比图

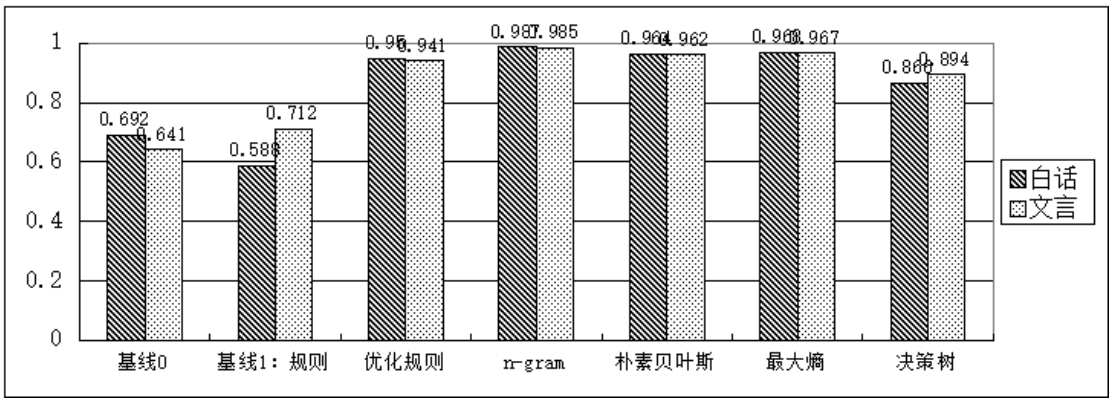

图 10　分类方法结果分析图

图 10 是本文所使用诸方法的测试结果。其中 ngram 为一元模型时的结果。基线 0 和基线 1 的结果相差不太大。基线 1 在古汉的标注中优于基线 0，在白话文的标注中弱于基线 0，这与规则的使用情况有关，因为基线 1 使用的规则主要是针对文言文特征的，而不考虑其对白话文特征的影响，所以对文言文的标注较为有利。由优化规则实验可以判断出，规则方法对本任务确有意义，但是规则本身的寻找和优化过程存在一定难度，需要进行大量实验，得

[4] 本部分的机器学习模型使用麻省大学的 MALLET 工具包实现[14]

到较为完善的规则库。

基于统计的模型的标注效果明显优于基线 0 和基线 1，由此可以确定基于统计的实验有其研究的意义，且可以得到了一个相对较好的结果。Unigram 模型的 F 值最高，达到 0.98 以上，是实验过程中构建的最为优秀的模型，且相较于朴素贝叶斯、最大熵和决策树三个机器学习模型，计算成本和时间成本都很低。

6 结论和展望

本文将文言文和白话文标注问题视作文本分类任务，通过基于规则和基于统计的方法进行标注。使用 26 种文言句式和 24 个文言虚词构成规则集，经由白话文词表进行消歧，取得了一定的效果。在统计方中，本文使用了 N-gram、朴素贝叶斯、决策树、最大熵算法等几种模型。实验发现基于统计的模型的标注效果明显优于基线，且 F 值普遍较高。其中一元语言模型取得了 0.98 的 F 值。

本文的结论支持了语言学家一直以来的直觉判断：即文言文的虚词使用是使之区分于白话文的主要标志，而非语法（或语序）。在语言演变过程中，最活跃的部分就是词汇[15]，而语法变化则相对缓慢。本文的工作也以计量的方式实证的证实了由文言文和白话文的分野主要集中在词汇层面这一判断。在这一现象中起主要作用的是虚词并少量动词（如“曰”）为代表的特征词汇。从一个侧面来说，我们工作实际描述了古代文言文到现代白话文作为一个连续统的存在性。

从本文标注任务的结果来看，民国时期的报章体更适合被视作文言文。

未来计划将规则方法和统计方法进行融合，并对更多时间段不同语体（如诗歌）进行测试，期待对这一问题给出更圆满的解决。

参考文献

[1] 王力著.中国语言学史[M].上海：复旦大学出版社,2007

[2] 张普.论语言的稳态[J].郑州大学学报(哲学社会科学版),2008(02)

[3] 张普.论语言的动态[J].长江学术,2008(01)

[4] 石毓智.汉语发展史上的双音化趋势和动补结构的诞生——语音变化对语法发展的影响[J].语言研究,2002(02)

[5] 胡裕树主编.现代汉语[M].上海：上海教育出版社,1981

[6] Mihalcea R, Nastase V. Word epoch disambiguation: Finding how words change over time[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. Association for Computational Linguistics, 2012: 259-263.

[7] Popescu O, Strapparava C. Behind the Times: Detecting Epoch Changes using Large Corpora[C]//International Joint Conference on Natural Language Processing. 2013: 347-355.

[8] 荀恩东,饶高琦,谢佳莉,黄志娥.现代汉语词汇历时检索系统的建设与应用[J].中文信息学报,2015(05)

[9] 饶高琦,臧娇娇,荀恩东.大数据视角下的语言实证工具：北语汉语语料库系统 BCC——以因果关系表达的语言模式研究为例[J]. 北京:北京市语言学年会,2014

[10] 金观涛,刘青峰著.观念史研究[M].法律出版社,2009

[11] 王力著.古代汉语[M].中华书局, 1964

[12] 王力著.汉语史稿[M].中华书局, 1980

[13] P.R. Clarkson and R. Rosenfeld. Statistical Language Modeling Using the CMU-Cambridge Toolkit From Proceedings ESCA Eurospeech, 1997.

[14] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit.", http://mallet.cs.umass.edu. 2002.

[15] 徐通锵,叶蜚声.语言学概论[M].北京：北京大学出版社,1981

作者简介：虞宁翌（1993--）女，学士，语料库建设与人文计算，Email: yuningyi58@126.com；饶高琦（1987--）男，博士研究生，语言规划与计算语言学，Email: raogaoqi-fj@163.com；荀恩东（1967--）男，通讯作者，教授，语言信息处理和教育技术