Open Access Library Journal
2021, Volume 8, e7821
ISSN Online: 2333-9721
ISSN Print: 2333-9705

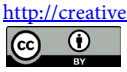
The Exploration of the Approach to Data Preparation for Chinese Text Analysis Based on R Language

Jiang Li

Shenzhen Institute of Information Technology, Shenzhen, China
Email: platojiang@126.com

How to cite this paper: Li, J. (2021) The Exploration of the Approach to Data Preparation for Chinese Text Analysis Based on R Language. *Open Access Library Journal*, **8**: e7821.
https://doi.org/10.4236/oalib.1107821

Received: August 2, 2021
Accepted: August 31, 2021
Published: September 3, 2021



Abstract

This paper explores how to prepare data for analyzing the Chinese texts with R language based on the theory of Welbers, particularly comparing the R package Rwordseg with jiebaR to see the results of Chinese text segmentation at the step of preprocessing.

Subject Areas

Big Data Search and Mining, Complex Network Models

Keywords

Data Preparation, Text Analysis, R Language, Chinese Text Segmentation

1. R语言的起源及优势

R 语言起源于上世纪六七十年代由美国贝尔实验室开发的 S 语言。到 90 年代 Ross Ihka 和 Robert Gentleman 在 S 语言基础上开发出 R 语言。取名为 R 的原因是两位开发者的名字的首字母均为 R。R 语言可以用来编程，进行统计学分析，还可以绘图。R 语言作为一种开源软件，拥有众多针对解决不同实际问题的安装包，比如用于制图的 ggplot2 包，用于统计分析的 Car，Power Analysis 包等。王家钺认为，数据可视化(制图)能力是 R 的巨大优势之一；相比 SPSS，它的统计包功能更有助于激发想象力和创造力。(王家钺，2019) [1] 由于 R 语言的诸多优势，越来越多的人文类研究，尤其是文本分析研究，采用 R 语言来进行。此外可以充分利用丰富的 R 语言程序包来得出不同角度方向的结果。根据 Welbers 的初步统计，R 语言程序库里有大约 50 个程序包用于文本分析，且不同语言包能自由切换或联合使用(Welbers 等，2017) [2]。

因此，掌握 R 语言，学会使用文本分析程序包，能为文本分析研究提供强大的工具。

2. Welbers的数据准备框架

Welbers 等认为“数据准备是任何数据分析的起点”。文本分析也无法绕开数据准备的工作。如表 1 所示，他把数据准备分为 5 个按先后顺序执行的步骤：文本导入，字符串操作，预处理，文档词条矩阵创建，矩阵筛选和加权，并且列出各步骤可用的 R 语言程序包(Welbers 等，2017) [2]。

其中导入文本和字符串操作是文本分析的常规操作，而从预处理开始的步骤是数据准备的难点。预处理包括多种处理方法，如标识化(tokenization)、标准化(将所有英文字符转为小写以及去除词缀，便于统一处理)、去除停用词等。Welbers (2017) [2]所列举的预处理方法，对于中文文本来说，去除停用词只须去除中文的停用词列表即可，但是标识化比较困难。标识化中很重要的操作是分词处理。Welbers 承认对于中文这种词分界不明显的语言，标识化操作须使用模式词典(Welbers 等，2017) [2]。标准化也主要是指英文的处理，因为中文文本不存在全部大写转换成小写以及去除词缀的处理环节。

王建红等在对马克思的《资本论》进行情感分析前，对其英文文本进行了数据准备处理。他手动删除了干扰性的脚注、章节目录、表格、计算公式等，去除大约十万余单词、数字、字符等，最终得到精简的英文文本(王建红等，2020) [4]。手动处理的工作比较繁琐，但在处理后，文本分析的效率和精确度大幅提高。参照 Welbers 的观点，王建红等的删除操作，属于字符串操作环节。Welbers 等认为“R 语言程序包大多内置了很多文本分析要用到的字符串操作应用，所以手动的，低级别的字符串操作，通常认为是没有必要的”(Welbers 等，2017) [2]。字符串操作应充分发挥程序包中的函数功能，如孟诗琼等分析汽车数据时，采用的是字符串函数与正则表达式相结合，这样充分发挥 R 语言程序包的优势，全面提升效率(孟诗琼等，2015) [5]。

朱昶胜等在研究股票网评论文本时，提到了中文文本挖掘。他将中文文本挖掘分成七个步骤：中文分词、合并同义词、去停用词、TF-IDF (词频逆文本)、生成矩阵、特征降维(主要是指删除占比例小的词语)、生成模型(朱昶胜等，2018) [6]。按 Welbers 的分法，朱昶胜等所列的挖掘步骤中，中文分词、合并同义词、去停用词等 3 步属于预处理环节，生成矩阵等同于文档词条矩阵创建环节，词频逆文本和特征降维等 2 个步骤属于矩阵筛选和加权的环节。

表 1. 数据准备步骤及推荐的 R 程序包[3]

数据准备步骤	可用 R 语言程序包
导入文本	readtext, jsonlite, XML, antiword, readxl, pdftools
字符串操作	stringi, stringr
预处理	quanteda, stringi, tokenizers, snowballC, tm, etc.
文档词条矩阵创建	quanteda, tm, tidytext, Matrix
矩阵筛选和加权	quanteda, tm, tidytext, Matrix

这前 6 个步骤全部可归入 Welbers 的数据准备部分。最后的步骤生成模型，依据 Welbers 的分类，应属于数据分析部分。此外，按照朱昶胜等所述，合并同义词是指“将意思相似或相同的词语用同一个词语代替，从而实现降维和提高文本处理的准确性”(朱昶胜等，2018) [6]。笔者认为，有些网络文本研究者只想分析对某一产品的积极评价，比如汽车销售网的评价文本，通过合并诸如“好”、“优秀”等同义词，可减少处理的词汇量，提升计算分析效率。但文学作品等经典文本中，表达近似意思的不同词汇，却可能会成为分析作者风格的有力证据。所以笔者认为合并同义词作为中文文本挖掘步骤之一的适用范围有待商榷。

3. 数据准备的难点：中文分词

如前所述，Welbers 所列的方法主要是针对英文文本这种单词有明显分界的语言。那么中文词汇分界不明显，其数据准备必然面临很大挑战。对中文文本进行分词属于数据准备的预处理环节，中文文本数据准备的难点就出现这个环节。

(一) 中文分词的技术发展

根据唐琳等的最新研究，中文分词模型算法的发展有三个阶段：基于匹配的词典分词、基于标注的机器学习算法、和基于理解的深度学习算法。词典分词法又称为机械分词法，而机器学习算法和深度学习算法统称为统计分词法(唐琳等，2020) [7]。梁喜涛等认为词典分词是“按照一定的策略把待切分的字符串与分词词典中的词进行比对，如果在词典中找到待切分的字符串则匹配成功”，统计分词是“根据相邻字的紧密结合程度来进行分词”(梁喜涛等，2015) [8]。

为解决中文分词问题，研发人员在不同平台上开发了多种中文分词程序。R 语言平台上，中文分词程序比较有代表性的是 Rwordseg [9]和 jiebaR [10]程序包。

吴丹露等曾对 rmmseg4j 包、RsegWord 包和 Rwordseg 包等 3 个分词程序包进行比较，检验对“大数据时代政府服务解读”，“基于 R 软件可视化及主题”两个例句进行分词的效果。结果发现 Rwordseg 分词正确率最高，不受新加入词库的新词干扰，且中英文混合词也能识别(吴丹露等，2015) [11]。Rwordseg 进行中文分词的操作简单，载入 Rwordseg 后，输入 segmentCN (“”)的函数命令，注意双引号里放入中文字符串，就可得到分词结果。

Rwordseg 的帮助文档，说明其分词算法模型是 HMM，即隐马尔可夫模型，是一种机器学习算法，属于统计分词法。Rwordseg 可以使用 setAnalyzer 函数，选择其他算法模型来分词，比如“fmm”，“coreNLP”，以及“jiebaR”。FMM 是正向最大匹配法，属于机械分词法。coreNLP 是斯坦福自然语言处理小组开发的 java 程序，支持中文分词，R 通过 Rjava 支持对其调用。

jiebaR 则是与 Rwordseg 一样，可以实现分词的 R 程序包。Rwordseg 可以调用 jiebaR 来分词。在 Rwordseg 的函数 setAnalyzer 帮助文档中，特别说明 Rwordseg 默认的 HMM 分词处理器是由内部 R 代码来执行的，目前仍然在完善之中。帮助文档建议，如要提升分词效果，选择 fmm，coreNLP，jiebaR。

由此可见，Rwordseg 的开发维护人员是认为 jiebaR 的分词效果强于 Rwordseg 的。

R-project 文档介绍，jiebaR 支持四种分词模式，包括最大概率法、隐马尔可夫模型、混合模型、索引模型。相比 Rwordseg 程序包，jiebaR 除了 HMM 模式增加了另外三种分词模式。最大概率法是将分词结果中最大概率的选项确定为最终结果，因此属于统计分词法。混合模型是将最大概率法和隐马尔可夫模式结合进行分词。索引模型是在运行混合模型进行分词后，再到词库中寻找匹配词，结合了词典分词法和统计分词法。四种模式都有各自优点，但其核心是最大概率法。此外，jiebaR 还能利用搜狗细胞词库，这样大大提升其实用性。

确认两种程序包哪种分词效果好，只须通过对同一个文本样本进行分词即可验证。Rwordseg 和 jiebaR 是比较流行的 R 程序包，使用者众多，得到广泛认可。如仅对普通的文本进行分词，两个程序包的算法模型没有得到充分挑战，分词区分度不会太高。所以笔者选取一段新浪网最新报道以及《红楼梦》的一段文字来对比分词效果。前者是新闻，有很多新词、专有名词；后者是古典作品，文白混杂，对分词软件的技术挑战很大，预测两个软件包会有不同的分词结果，分词效果会有明显的区分。

(二) Rwordseg 与 jiebaR 程序包的新闻报道分词效果对比

新闻报道的分词对比实验样本，选取的是一篇关于民法典的新闻片段。这篇新闻报道的标题是《民法典正式施行！婚姻法继承法合同法等废止，2021 年你的生活将有这些大不同》，所选取的片段如下：

“民法典将正式施行，现行婚姻法、继承法、民法通则、收养法、担保法、合同法、物权法、侵权责任法、民法总则同时废止。”(45 字) (杨杰，2021) [12]。

比较表 2 中 Rwordseg 和 jiebaR 的结果，能发现除了“侵权责任法”拆分成“侵权”和“责任法”，“民法总则”拆分成“民法”和“总则”之外，jiebaR 准确识别出其他所有词。jiebaR 未能识别出来的两个专有名词，拆分后并没有影响读者的理解。Rwordseg 则未能正确识别出所有的法律法规名，并且“民法典”拆成“民”和“法典”，“继承法”拆成“继”和“承法”，“收养法”拆成“收”和“养法”，“担保法”拆成“担”和“保法”是明

表 2. 新浪网关于民法典等新闻报道的分词结果对比

Rwordseg 分词结果	【1】“民”“法典”“将”“正式”“施行” 【6】“现行”“婚姻”“法”“继”“承法” 【11】“民法”“通则”“收”“养法”“担” 【16】“保法”“合同”“法”“物权”“法” 【21】“侵权”“责任法”“民法”“总则”“同时” 【26】“废止”
jiebaR 分词结果	【1】“民法典”“将”“正式”“施行” 【5】“现行”“婚姻法”“继承法”“民法通则” 【9】“收养法”“担保法”“合同法”“物权法” 【13】“侵权”“责任法”“民法”“总则” 【17】“同时”“废止”

显的错误，影响了读者的理解。

新闻报道通常要求及时和最新，以上的分词结果证明在新词、专有名词上面，jiebaR 比 Rwordseg 分词效果要好，更适合新闻报道的文本数据预处理。

(三) Rwordseg 与 jiebaR 程序包的经典文本分词效果对比

经典文本的样本，选取《红楼梦》程乙本校注版第一百二十回贾雨村和甄士隐的一段对话：

雨村听了，虽不能全然明白，却也十知四五，便点头叹道：“原来如此，下愚不知！但那宝玉既有如此的来历，又何以情迷至此，复又豁悟如此？还要请教。士隐笑道：“此事说来，先生未必尽解。太虚幻境，既是真如福地。两番阅册，原始要终之道，历历生平，如何不悟？仙草归真，焉有‘通灵’之不复之理呢？”(123 字) [13]。

在使用 Rwordseg 和 jiebaR 执行分词后，标点符号空格等干扰信息都自动删除。总共发现有 11 处不同的分词，共 36 个字，占整个样本的比重为 29.3%。其他均为相同的分词，共 87 个字，占比 70.7%。本文重点分析不同的分词结果项，所以将识别出来的相同词删去。两个软件包不同分词结果项如表 3：

在表 3 结果中，存在有两种分词结果都可判定为正确，但其中一种更优的情况。两种都可判定为正确的分词比较情况如表 4：

依据明显正确、明显错误、更优的三种分类，对分词结果统计分析如表 5：

根据表 5 数据，通过计算可以得出以下结果：

1) 不同分词结果项中，jiebaR 三个统计项目表现更好。jiebaR 明显正确项多于 Rwordseg，更优项远多于 Rwordseg，且明显错误项少于 Rwordseg。

表 3. Rwordseg 和 jiebaR 对同一红楼梦选段的不同分词结果项

Rwordseg 分词结果	【1】“雨”“村听”【2】“便点”“头”【3】“原来”“如此”【4】“那”“宝玉”【5】“既”“有”【6】“说”“来”【7】“太虚”“幻境”【8】“真如”【9】“原始”“要终”【10】“焉”“有”【11】“通灵”“之”“不复”“之理”
jiebaR 分词结果	【1】“雨村”“听”【2】“便”“点头”【3】“原来如此”【4】“那宝玉”【5】“既有”【6】“说来”【7】“太虚幻境”【8】“真”“如”【9】“原始要终”【10】“焉有”【11】“通”“灵”“之”“不”“复”“之”“理”

表 4. Rwordseg 和 jiebaR 对同一红楼梦选段的分词更优项

R 程序包更优项	Rwordseg 分词	jiebaR 分词
jiebaR	“原来”“如此”	“原来如此”
Rwordseg	“那”“宝玉”	“那宝玉”
jiebaR	“既”“有”	“既有”
jiebaR	“说”“来”	“说来”
jiebaR	“太虚”“幻境”	“太虚幻境”
jiebaR	“焉”“有”	“焉有”
字数小计	3	14

表 5. Rwordseg 和 jiebaR 对同一红楼梦选段的不同分词正、误、更优项统计

R 程序包		明显正确	明显错误	更优
	项目	【8】【11】	【1】【2】【9】	【4】
Rwordseg	数量小计	**2**	**3**	**1**
	字数小计	**9**	**10**	**3**
	字数小计	**10**	**9**	**14**
jiebaR	数量小计	**3**	**2**	**5**
	项目	【1】【2】【9】	【8】【11】	【3】【5】【6】【7】【10】

2) jiebaR 样本总体分词准确率高于 Rwordseg。如果默认两种软件包同样的分词项为正确项，并且不管分词是否更优，仅计算准确分词的比率，那么准确率 = (样本总字数 − 分词明显错误项字数)/样本总字数。总字数按 123 字计，jiebaR 准确分词字数为 123 − 9 = 114 字，因此 jiebaR 的分词准确率约为 92.7%；Rwordseg 准确分词字数为 123 − 10 = 113 字，准确率约为 91.9%。两个软件都有满意的分词准确率，虽然二者差距不大，但 jiebaR 略高。

3) jiebaR 更优项比率高于 Rwordseg。按占整个文本字数(123 字)比重计，jiebaR 更优项为 14 字，比重为 11.4%，Rwordseg 更优项为 3 字，比重为 2.4%；按占各自准确分词字数比重计，jiebaR (准确分词字数 114 字)更优项为 12.3%，Rwordseg (准确分词字数为 113 字)为 2.7%。jiebaR 最优项占总字数比率以及准确识别字数比率都要超过 Rwordseg 软件包 9~10 个百分点，表明分词的质量远好于 Rwordseg。

此外，比较分词项具体分词内容，还可发现 jiebaR 能正确识别出 4 字词，比如第【7】处“太虚幻境”和第【9】处“原始要终”。第【4】处虽然 jiebaR 将“那宝玉”识别为一个词，Rwordseg 识别为“那”和“宝玉”更优，但侧面也证明 jiebaR 识别三字词的能力。这会是 jiebaR 分词的一大优势。相比之下，Rwordseg 没有识别出 1 个超过两个字的词。

尽管 jiebaR 综合表现优秀，但也有两处明显分词错误。分词结果第【8】处“真如福地”中的“真如”，意指“永恒真理”，是个完整的词，却拆分成了“真”和“如”。第【11】处“通灵之不复之理”全部拆为单个字，至少“通灵”应该能识别出，但没有做到。如果分析体量大的中文经典文本，这样的分词错误会影响最终的文本分析结果，可能会得出错误的结论。而这两处分词，Rwordseg 得出了正确的结果，准确识别出“真如”、“通灵”、“不复”等。所以为提升分析的结果准确性，建议在分析经典文本时，尝试分别用 jiebaR 和 Rwordseg 进行分词，再比较二者的分词结果项，择优选用。

4. 结语

依据 Welbers 的理论，文本分析之前的重要工作是数据准备，它按照先后顺序分为文本导入，字符串操作，预处理，文档词条矩阵创建以及对矩阵过滤和加权。在预处理阶段，中文由于是连续字符串，相比英文缺乏明显的词分界，中文分词恰好是数据准备预处理环节的重难点。

在 R 语言平台上，Rwordseg 和 jiebaR 是有代表性的两个分词程序。通过对新浪网的最新中文新闻样本和红楼梦的一段中文经典样本的分词效果比较，发现在两种文本的分词上，jiebaR 均要优于 Rwordseg。在有限的文本分词实验中，未发现 Rwordseg 能识别出大于两个字的汉语词，但是 jiebaR 可以识别出四字词，并且证明有识别三字词的能力。

因此在使用 R 程序包分析中文文本时，jiebaR 是首选的分词程序。具体到新闻报道和经典文本两类的分词，jiebaR 对新闻报道的分词没有明显错误，所以 jiebaR 更适合用于新闻报道文本的分词。对经典文本分词时，jiebaR 有两处明显错误，由此得出这可能会影响甚至误导经典作品全文或者大型中文语料库的分析结论。所以使用 jiebaR 分析经典文本时，最好能参考 Rwordseg 分词结果，来综合两种分词程序的优点。

尽管本文采用的分词语料具有一定的代表性，但是网络文本仅 45 字，经典文本为 123 字，因此根据这两个字数有限的样本得出的结论，还需在后续的研究中，通过大量的文本语料实验或统计分析，进一步验证。

最后，现阶段各种分词软件都不是完美的，需要在实践中找到各自的优缺点，扬长避短。同时还需要继续研究更好的分词技术，完善现有的分词软件，使研究人员分析更多类别的中文文本时，得到好的分词效果。

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] 王家钺. 基于 R 的语言学统计方法[M]. 北京: 外语教学与研究出版社, 2019: 26-130.

[2] Welbers, K., Van Atteveldt, W. and Benoit, K. (2017) Text Analysis in R. *Communication Methods and Measures*, **11**, 245-265. https://doi.org/10.1080/19312458.2017.1387238

[3] R Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

[4] 王建红, 冉莹雪. 《资本论》中的“性情马克思”——基于 R 语言 syuzhet 安装包的文本情感分析[J]. 海南广播电视大学学报, 2002, 79(2): 31-37.

[5] 孟诗琼, 孟诗瑶, 尹志. 基于 R 语言的汽车消费数据挖掘及可视化方法[J]. 宁波工程学院学报, 2015, 27(4): 17-23.

[6] 朱昶胜, 孙欣, 冯文芳. 基于 R 语言的网络舆情对股市影响研究[J]. 兰州理工大学学报, 2018, 44(4): 103-108.

[7] 唐琳, 郭崇慧, 陈静锋. 中文分词技术研究综述[J]. 数据分析与知识发现, 2020, 4(2/3): 1-17.

[8] 梁喜涛, 顾磊. 中文分词与词性标注研究[J]. 计算机技术与发展, 2015, 25(2): 175-180.

[9] Li, J. (2019) Rwordseg: Chinese Word Segmentation. R Package, Version 0.3-2. https://CRAN.R-project.org/package=Rwordseg

[10] Qin, W. and Wu, Y. (2019) jiebaR: Chinese Text Segmentation. R Package, Version 0.11. https://CRAN.R-project.org/package=jiebaR

[11] 吴丹露, 魏彤, 许家清. R 语言环境下的文本可视化及主题分析——以社会服务平台数据为例[J]. 宁波工程学院学报, 2015, 27(1): 19-25.

[12] 杨杰. 民法典正式施行！婚姻法继承法合同法等废止, 2021 年你的生活将有这些大不同[EB/OL]. https://news.sina.com.cn/c/2021-01-01/doc-iiznctke9619275.shtml, 2021-01-01.

[13] 曹雪芹著; 程伟元, 高鹗整理; 启功等评注. 红楼梦: 程乙本校注版[M]. 桂林: 广西师范大学出版社, 2017: 1-17.

Appendix (Abstract and Keywords in Chinese)

基于 R 语言的中文文本分析数据准备方法探讨

摘要：本文依据 Welbers 等的数据准备框架，探索如何利用 R 语言准备中文的文本分析数据，重点在于比较数据预处理环节，Rwordseg 和 jiebaR 程序包的中文分词效果。

关键词：数据准备，文本分析，R 语言，中文分词