文章编号：2096-1472(2020)-04-15-04
DOI:10.19644/j.cnki.issn2096-1472.2020.04.004

中文分词模型在中医病症语义理解中的研究与应用

许林涛，叶欣欣，裴成飞，吴荣士

(安徽理工大学，安徽 淮南 232000)
✉1194663015@qq.com;xxye999@163.com;1138664088@qq.com;Rongshi_Wu@163.com

摘　要：中医临床记录的病症内容是中医医师进行诊断的重要依据。由于中文表达形式的多样性与复杂性，如何从这些病症内容中进行标准化四诊信息的提取对于中医证候分析具有重要的研究价值。本文在充分分析各种中文分词算法的基础上，选择将最大正向匹配分词算法应用于中医临床病症内容中的四诊信息语义理解，构建的中医四诊语义模型在100个实际病例的四诊信息提取，再对最大分词数进行变量控制，得出最大分词数为5时得出的准确率和召回率最高。

关键词：中文分词；证候分析；四诊信息
中图分类号：TP311　**文献标识码**：A

Research and Application of Chinese Word Segmentation Model in Semantic Understanding of TCM Diseases

XU Lintao,YE Xinxin,PEI Chengfei,WU Rongshi

(*Anhui University of Science & Technology*,*Huainan* 232000,*China*)
✉1194663015@qq.com;xxye999@163.com;1138664088@qq.com;Rongshi_Wu@163.com

Abstract:TCM clinical record of the disease content is an essential basis for the diagnosis of TCM physicians.Due to the diversity and complexity of Chinese expressions,how to extract standardized four-diagnosis information from the contents of these conditions has important research value for TCM syndrome analysis.Based on the full analysis of various Chinese word segmentation algorithms,this paper chooses to apply the maximum forward matching word segmentation algorithm to the semantic interpretation of the four-diagnosis information in the clinical symptoms of traditional Chinese medicine.This research conducts the extraction of four-diagnosis information of 100 actual cases based on the constructed traditional Chinese medicine four-diagnosis information diagnostic model. Then the variable control is performed on the maximum number of word segmentation,and the high accuracy and recall rate are obtained when the maximum number of word segmentation is five.

Keywords:chinese word segmentation;syndrome analysis;four consultation information

1 引言(Introduction)

中医提倡以“以证迁方”为基础，实现对症下药。“证”是指证候，即通过方与证的关系，达到推荐名医名方的作用[1]。证候在中医中通常指的是在诊断过程中，具有潜在联系的一组病症和体征。如完谷不化、小便频数、夜频尿多、全身肿胀、舌淡、苔白等是肾阳虚的证候。大部分中医在诊断过程中会通过‘望’‘闻’‘问’和‘切’将病人的病症和体征用描述性的文字记录下来，凭此记录为病人开处方。由于中文表达形式的多元性和复杂性，加上中医医师在记录病症时通常用古文的形式，如何从这些病症内容中进行标准化四诊信息的提取对于中医证候分析具有重要的研究价值。

随着自然语言处理技术的不断提高，中文分词算法也被广泛应用于中医领域，对中医的证候分析有重要的研究价值。张千、王庆玮等人[2]对传统的特征提取方法和最新的深度学习在文本挖掘方面的技术做了综述；郭德海、郑光[3]等人利用文本挖掘技术总结了慢性咳嗽的中医诊治规律；王丽颖、郑光[4]等人使用文本挖掘技术探索高血压常见中医证候即常用方剂。本文在充分分析各种中文分词算法的基础上，选择将最大正向匹配分词算法为核心，构建了中医四诊语义模型应用于中医临床病症内容中的四诊信息语义理解。

2 中医四诊语义模型(Semantic model of TCM four diagnosis)

2.1 中文分词技术

中文分词技术[5]是自然语言处理中的一项核心技术，英文中已经将词和词之间用逗号或者空格分开，而中文对词定义

的边线很难划分。在汉语中以字为最小单位，但是词的数量和不同词在不同语境下的语义也是不一样的。因此在理解中文文本内容时，中文分词是一个不可或缺的一个步骤。将一段文本转化为词的表示，就是中文分词。

当前主流的中文分词算法分别为：基于词典的中文分词算法、基于统计模型的中文分词算法和基于语义理解的中文分词算法[6]。

2.1.1 基于词典的中文分词算法

基于词典的中文分词算法又称基于字符串匹配分词算法，它是按照一定的规律将一段中文文本与已经定义的“词典”中的词条进行匹配，若在词典中找到某个字符串，则可以分成一个词。这种算法的好坏与词典和匹配规则有着密切的联系，也和扫描的方向相关。又根据扫描方向的不同，分为最大正向匹配算法、最大逆向匹配算法和双向最大匹配算法。

2.1.2 基于统计模型的中文分词算法

基于统计模型的中文分词算法是根据统计中文文本的词频进行分词，若在文本中出现同一个词的频率越高，则构成一个词的可能性就越大。这个算法不使用“词典”，只会对分词的中文文本中相邻的字之间进行一个词频统计来计算他们同时出现的概率，概率越大，说明构成词的可能性越大，通常会设定一个阈值来控制这个概率。

2.1.3 基于语义理解的中文分词算法

该算法的基本思想是借助大量的语义和语法知识来训练模型。在分词的过程中，利用这些训练好的模型来对文本进行语义、语法分析和歧义识别。但由于汉语的歧义性和复杂性，将文本语义转化为机器可识别的语言难度较大。

2.2 最大正向匹配分词算法

最大正向匹配算法是自然语言处理中最常见的一种算法，其主要思路是将一段待分词的文本数据，根据用户所设定的最大分词长度来循环遍历，与“词典”中的词进行匹配，得到匹配的结果就是所要的分词结果。

最大正向匹配算法的步骤如下：

步骤1：根据自定义设置的最大分次数W，将待分词的文本s1中从左向右取出W个字符，查看这W个字符是否在词典中。如果在词典中就直接输出，如果不在词典中则将W最后一个字去掉，如果剩下的W是个单字，也直接输出。去掉末尾字且不是单字，继续循环查看是否在词典中。

步骤2：继上述的一次输出分词结果后，继续将文本(s1−w)中从左向右取出W个字符，重复步骤1操作，直到s1为空结束。

步骤3：将上述分词结果统计，计算他们的准确率和召回率。

最大正向匹配分词算法流程图如图1所示。

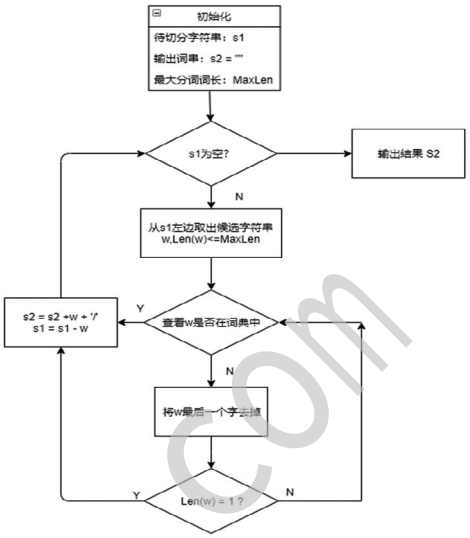

图1 最大匹配分词算法流程图

Fig.1 Maximum matching word segmentation algorithm flowchart

一般通过准确率(Precision)和召回率(Recall)来做为最大正向匹配分词算法的评价标准，其计算公式如下：

$$Precison = \frac{WordCount(CorrectR)}{WordCount(TrainSet)} = \frac{C}{X}$$

$$Recall = \frac{WordCount(CorrectR)}{WordCount(TestSet)} = \frac{C}{Y}$$

其中，X和Y分别表示训练数据集和测试数据集的词数，C表示正确匹配的词数。

2.3 中医四诊语义模型

本文将中文分词模型应用于中医证候分析中特征词的提取和分析，通过对病症内容得到的描述性文本信息的分词和同义词匹配构建了中医四诊语义模型。模型构建步骤如图2所示。

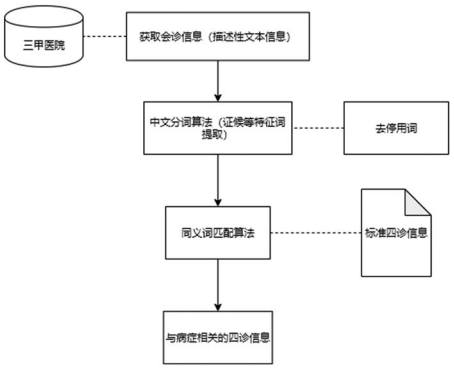

图2 中医四诊语义模型结构图

Fig.2 Structural diagram of the four models of TCM semantics

步骤1：将样本病历中描述性文本信息进行中文分词和去停用词。

步骤2：将得到的证候等特征词进行同义词匹配，排除文言文或同义不同词的影响，根据标准四诊信息得到与病症相关的四诊信息。

步骤3：调整最大分词数，重复步骤1和步骤2，分别得出分词结果。

步骤4：由上述产生的分词结果，计算不同的最大分词数的准确率和召回率，保留准确率和召回率最高的一组。

3　样本选择与特征提取(Sample selection and feature extraction)

3.1　样本选择

本次实验的数据来源为常州市中医院等十余所临床医院采集到的100例中医会诊记录。

3.2　特征提取

如何有效地从文本信息提取出样本信息特征，从而为证候分析提供重要的数据基础，是本文的研究重点。在充分分析现有的样本病历的基础上，采用四诊信息的方式进行特征提取是一个非常有效地方法，通过提取与病症相关的四诊信息来进行证候分析，更能抓住病人的病症和机理，从而达到对症下药的效果。具体的特征提取方法如下：

步骤1：定义一个标准的四诊信息库。本文涉及的四诊信息的定义依据常州中医院申春悌制定的标准信息库，标准信息库部分定义如表1所示。

表1 标准信息库部分定义表

Tab.1 Standard information library part definition table

编码	望诊名称	编码	闻诊名称
SZ110100	昏仆	SZ210100	语声重浊
SZ110200	神昏	SZ210200	语言謇涩
SZ110300	精神萎靡	SZ210300	少气懒言
SZ120100	面色青	SZ220100	口臭
SZ120200	面色暗黄	SZ220200	尿臭
SZ120300	面色萎黄	SZ220300	汗臭
SZ130100	肥胖	SZ211900	喷嚏
SZ130200	消瘦	SZ212000	鼻鼾
SZ140200	口眼歪斜	SZ212100	干呕
SZ140300	颜面抽搐	SZ212200	肠鸣

(续表)

编码	问诊名称	编码	切诊名称
SZ310100	恶风寒	SZ410100	浮脉
SZ310200	畏寒	SZ410200	沉脉
SZ310300	寒战	SZ410300	迟脉
SZ310400	恶热	SZ410400	数脉
SZ310500	骨蒸	SZ410500	洪脉
SZ310600	壮热	SZ410600	细脉
SZ310700	潮热	SZ410700	虚脉
SZ310800	烘热	SZ410800	实脉
SZ310900	手背热	SZ420100	腹诊
SZ311000	身热夜甚	SZ420200	腹部硬满

步骤2：从采集到的样本信息，采用最大正向匹配算法、同义词匹配加上人工处理的方式，提取与病症相关的四诊信息。

由于样本信息是类似文言文的文本信息，以及中医们的口述信息，有些词会出现与四诊信息同义不同词的现象，利用同义词匹配可以排除这些影响。

步骤3：结合描述性的病历信息，给每个病症相关的四诊信息定义一个层级，一般分为无、轻、中、重四级，分别用1、2、3、4来进行特征表示，从而完成从病历文本信息的特征提取。

4　实验结果(Experimental results)

本文实验所涉及的数据集是100例中医会诊时的会诊记录，我们首先需要将这些会诊信息中关键信息提取出来，以得出该病人的具体患病信息。为了保护病人的隐私，将每个病例只取其会诊信息，并用病例1病例2来编号，部分会诊信息如表2所示。

表2 部分患者—会诊信息表

Tab.2 Partial patient-consultation information forms

病例	会诊信息
病例1	秋燥之季，风热之邪流行，恶风发热，汗出不畅，延今半月不退，伴喉痛作咳，咯痰不爽，舌偏赤，苔薄黄，脉浮数带滑。曾经输液及抗病毒治疗
病例2	患者入秋即发哮喘，冬令自行缓解，反复六载。发作先多喷嚏，随见胸闷、喘息，张口抬肩，呀呷有声，大汗，咯出粘痰方舒，用平喘药及喷雾剂吸入，恙虽轻而难至平缓。从未发热等等症。舌体偏红，苔薄黄腻，脉浮弦而促。童年有类似发作
病例3	喉蛾喉痛屡发3年且易外感作咳，热退五天喉痒干咳，舌红苔少脉细滑数。两颈淋巴结肿胀质硬多枚

在上述的会诊信息的基础上，实现最大匹配中文分词算法，并进行词性标注，去停用词、语气助词和其他一些与证候无关词性的词，得到最初的分词结果如表3所示（部分病例示例）。

表3 部分患者—会诊信息分词表

Tab.3 Partial patient-consultation information words table

病例	会诊信息（分词后）
病例1	秋燥/nr 季/n 风热/n 邪/a 流行/v 恶风/n 发热/v 汗/n出/v不畅/a 不/d 退/v 伴喉/n 痛作/v 咳/v 咯痰/v 不爽/a 舌/n偏/d 赤/vg 苔/ng薄/a 黄/n 脉/ng 浮数/n 带滑/v 输液/n 抗/v 病毒/n 治疗/v
病例2	患者/n 入秋/n 发/v 哮喘/n 冬令/n 自行/r 缓解/v 反复/v 六载/m 发作/vn 先多/d 喷嚏/v 随/v 见/v胸闷/n 喘息/v 张口/nr 抬肩/v 呀/y 呷/v 有声/d 大汗/n 咯出/v 粘/v 痰/v 平喘药/n 喷雾剂/nz 平缓/a 发热/ v 舌体/n 偏/d 红/a，苔/ng 薄/a 黄腻/n 脉浮/ng 弦/n 促/v 发作/vn
病例3	喉蛾/n 喉痛/n 作咳/v 热/a 喉/ng 痒/a 干咳/v 舌/n 红/a 苔/ng 少脉/n细滑数/n 颈/ng 淋巴结/n 肿胀/v 硬/a

由于会诊信息是类似文言文的描述性文本，分词后得出的词直接和标准的四诊信息进行匹配，准确率会大大降低。需要将分词后的结果进行同义词匹配，在和标准的四诊信息进行匹配。得到的最终的分词结果如表4所示。

表4 部分患者—会诊信息最终分词表

Tab.4 Partial patient-consultation information final segmentation table

病例	会诊信息（最终分词结果）
病例1	恶风/n 发热/v 汗/n 不畅/a 伴喉/n 痛作/v 咳/v 咯痰/v 不爽/a 舌/n 赤/vg 苔/ng薄/a 黄/n 脉/ng 浮数/n 带滑/v
病例2	气喘/n 冬令/n 喷嚏/v 胸闷/n 喘息/v 张口/nr 肩痛/n 大汗/n 咯出/v 粘/v 痰/v 发热/ v 舌体/n 红/a 苔/ng 薄/a 黄腻/n 脉浮/ng 弦/n 促/v
病例3	喉蛾/n 喉痛/n 作咳/v 热/a 喉/ng 痒/a 干咳/v 舌/n 红/a 苔/ng 少脉/n细滑数/n 颈/ng

最后与定义的标准四诊信息匹配可得出与病症相关的四诊信息，为中医进行后续的证候分析提供数据基础，如表5所示。

表5 部分患者—病症相关四诊信息表

Tab.5 Partial patient-illness related four diagnosis information form

病例	病症相关的四诊信息
病例1	恶风寒 恶寒发热 有汗 喉痛 咳嗽 咯痰 舌红 黄苔 浮脉 滑脉
病例2	气喘 喷嚏 胸闷痛 叹息 气喘 肩痛 多汗 咯痰 粘痰 舌红 黄苔 浮脉 弦脉 促脉
病例3	咽喉肿痛 咳嗽 恶寒发热 干咳 舌红 苔薄 细脉 滑脉 数脉

本文实验是通过Python实现了最大正向匹配分词算法，数据集是用txt格式来存储，通过Python程序读取。以20例病例作为测试病例，80例病例作为样本病例，经过多次调试最大分词数，分别计算他们的准确率和召回率，得出结果。结果对比发现组大分词数为5时，准确率和召回率最高，实验结果如表6所示。

表6 实验数据表

Tab.6 Experimental data table

最大词长	测试词数	样本词数	测试正确词数	样本正确词数	准确率	召回率
3	8	24	2	16	0.25	0.6667
4	7	23	3	19	0.4286	0.8261
5	7	21	5	19	0.7142	0.9048
6	6	20	3	17	0.5	0.85
7	6	20	3	17	0.5	0.85

经上述的实验得出，用词长为5的最大分词数和最大匹配分词算法，可以准确地得出该病例中会诊信息的特征词，即与病症相关的四诊变量，为后续证候分析提供数据基础。

5 结论(Conclusion)

本文是以100例病例的会诊信息为例，将语义分析应用到证候分析中，提取出病例的会诊信息中的特征词，与定义好的四诊信息匹配得出与病症相关的四诊信息，可以为中医的诊断提供更有效地数据基础。中医的证候分析具有重要的研究价值，而语义分析的应用，不仅局限于普通的分词匹配，还和标准的四诊信息进行比对替换，实现了证候名的统一，以更好地实现证候后续的挖掘和分析。

参考文献(References)

[1] 尹湘君,何庆勇,王阶,等.近40年血脂异常中医证候动态演变规律的研究[J].中华中医药杂志,2018(04):1523-1526.

[2] 张千,王庆玮,张悦,等.基于深度学习的文本特征提取研究综述[J].计算机技术与发展,2019(12):61-65.

[3] 郭德海,郑光,张洁,等.基于文本挖掘的慢性咳嗽中医诊治规律研究[J].中国中医药信息杂志,2019(10):101-104.

[4] 王丽颖,郑光,赵学尧.基于文本挖掘的高血压病中医辨证用药情况分析[J].世界中西医结合杂志,2018(04):462-465;470.

[5] 王梦鸽.基于深度学习中文分词的研究[D].西安邮电大学,2018.

[6] 张少聪.中医医疗辅助诊断系统研究与实现[D].电子科技大学,2018.

作者简介：

许林涛(1995-)，男，硕士生.研究领域：人工智能，数据挖掘.

叶欣欣(1996-)，女，硕士生.研究领域：隐私保护，数据挖掘.

裴成飞(1996-)，男，硕士生.研究领域：隐私保护，数据挖掘.

吴荣士(1995-)，男，硕士生.研究领域：隐私保护，数据挖掘.