基于NLP的中医医案文本快速结构化方法

肖晓霞[1]，刘明婷[2]，杨冯天赐[3]，刘鉴建县[4]，杨阳[5]，石月[6]
1. 湖南中医药大学信息科学与工程学院，湖南 长沙 410208；
2. 湖南大学信息科学与工程学院，湖南 长沙 410082；
3. 湘潭大学化学学院，湖南 湘潭 411105；
4. 湖南泽塔科技有限公司，湖南 长沙 410012；
5. 东北林业大学工程技术学院，黑龙江 哈尔滨 150040；
6. 北京瑞迪弘欣科贸有限公司，北京 100071

摘要
中医医案是中医医生学习临床经验的重要文献资料，对中医医案进行结构化处理有利于采用机器学习等方法总结临床经验，加速中医传承。为了实现中医医案快速结构化，提出了一种基于自然语言处理的中医医案文本快速结构化方法。将《中国现代名中医医案精粹》作为结构化对象，采用光学字符识别技术识别医案截图的文本，同时对文本做初步结构化。构建简单症状词典，采用结合词典的改进的N-gram模型获取医案文本中的症状、体征等词，并在结构化过程中更新词典，实现了对4 754份文本医案的结构化。随机选取666份医案文本对最终模型进行测试，其F1值达到82.99%。

关键词
N-gram模型；自然语言处理；中医医案；中文分词；光学字符识别

中图分类号：TP391.1 文献标志码：A doi: 10.11959/j.issn.2096-0271.2022025

A fast text structuring methodology of TCM medical records based on NLP

XIAO Xiaoxia[1], LIU Mingting[2], YANG Fengtianci[3], LIU Jianjianxian[4], YANG Yang[5], SHI Yue[6]
1. School of Informatics, Hunan University of Chinese Medicine, Changsha 410208, China
2. College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
3. The College of Chemistry of Xiangtan University, Xiangtan 411105, China
4. Hunan Zeta Technology Co., Ltd., Changsha 410012, China
5. College of Engineering and Technology, Northeast Forestry University, Harbin 150040, China
6. Beijing Ruidi Hongxin Science and Trade Co., Ltd., Beijing 100071, China

Abstract
Traditional Chinese medicine (TCM) medical records are the most valuable documents for TCM doctors to learn clinical experience. The structured TCM medical records are conducive to extract the clinic knowledge based on machine learning and other methods, which can accelerate the inheritance of TCM. A fast text structuring methodology of TCM medical

records based on natural language processing (NLP) was proposed to structure the clinic cases. *Essence of Chinese Modern Famous Chinese Medical Records* was selected as the medical record structuring objects, and the text in the screenshots of the medical records was recognized by optical character recognition (OCR) and the text was initially structured. A simple symptom dictionary was constructed, and the improved N-gram model combined with the dictionary was used to recognize the symptoms, signs and other words in the text, and the dictionary was updated in the structuring process. At last, 4 754 text medical records were structured. The final model was test on 666 medical records selected randomly from the corpus, and its F1 value reached 82.99%.

Key words
N-gram model, NLP, TCM medical records, Chinese word segmentation, OCR

0 引言

中医医案是中医历代医家临床过程的记录，往往采用叙述的方式记录病人的症状、体征和理法方药，是历代医家综合运用中医理法方药解决临床问题的经验总结，是中医知识传承的载体。但医案浩如烟海，若能够将医案中的症状、体征、证、方提取出来，并结构化为独立可用的数据单元，才能利用现代数据科学技术构建“症状（体征）–证–方”的关系，才能更高效地总结海量医案中的诊疗经验，更有利于中医传承[1]。

目前，医案资料大多以书籍的形式存在，基本都有对应的电子书籍，但电子书籍也是以扫描版本为主，而非可计算机直接识别的文字。人工整理和结构化医案费时费力，直接采用自然语言处理结构化图片文字也不可能，但可以先采用光学字符识别技术将图片式医案转化为计算机文字，再用自然语言技术来处理。

1 医案结构化现状

医案的描述一般包括病人姓氏、年龄、性别、症状、体征、证名、治则或治法、病因、方剂名、汤药名、中草药名、西药品名等，这些都是采用自然语言形式描述的，要将医案结构化，就需要将这些信息提取出来作为一个独立的数据单元。这些信息提取中难度最大的就是症状、体征和现代医案中生化指标信息的提取，由于中文语句中没有词的间隔符，信息提取之前往往需要对文本进行词语的切分并将其识别为目标对象，对应的技术有中文分词和命名实体识别技术。目前中文分词和命名实体识别主要有基于词典、基于规则、基于统计以及规则与统计相结合的方法[2]。

基于词典的方法要求词典涵盖所有需要抽取的实体，并且随着数据量的增大，匹配速度会大幅度降低，对未登录词（即自然语言处理中的未被词典收录的词）的补充较难实现[3-4]，缺乏自学能力。由于人类语言的灵活性和多变性，基于规则的实体抽取也很难有一个通用的方法。基于统计的机器学习方法、深度学习方法是目前发展比较快、应用比较广的中文自然语言处理方法，如隐马尔可夫模型（hidden Markov model，HMM）、最大熵（maximum entropy，ME）模型、条件随机场（conditional random field，CRF）模型、长短期记忆（long short-time memory，LSTM）网络等[5]。由于基于统计的机器学习方法和深度学习方法需要对所处理的文本进行标注，短时间内无法完成，并且标注的方

法及文本的领域特点也会使算法无法泛化到其他领域。除此之外，由于深度学习涉及大量的高维稀疏矩阵运算，需要特殊计算硬件来加速[6]。

医案结构化过程中最大的工作量就是对医案中症状、体征命名实体的识别，但目前并没有专门针对中医医案症状、体征命名实体识别的技术，也没有公开的用于中医医案症状、体征命名实体识别的词典和通用的语料库，因此涉及的中医药词典和语料都需要研究者自行构建。例如，张帆等人[7]构建了中医领域词典，对600份医案进行了人工标注，之后采用层叠隐马尔可夫模型结合中医词典的方法对600份医案进行处理，F1值为94.14%；李明浩等人[8]在对492份医案中2 069条规范症状进行标注的基础上，采用LSTM-CRF对这些医案中的症状进行识别，F1值为78%。

中医临床命名实体识别研究随着技术的发展不断进步，但由于中医领域特点及研究起步较晚，症状命名实体识别要么需要大量人工语料标注，要么其F1值不高。为了找到合适的快速结构化医案文本的方法，本文在搜狗细胞词库中下载了与中医诊断、症状、中药等相关的词典近30部，共收集约17万个词条。尽管这些词典词条丰富，但要结构化的医案中的大量症状、体征未包含在其中，因此采用上述词典结合jieba库的分词效果不佳，长词基本无法识别，对未登录词的识别准确率也不高；尝试采用FuzzyWuzzy库进行模糊字符串匹配，准确率有所提高，但运行速率太低，整个实验从开始运行到完成花费将近7 h。鉴于此，本文采用无须人工标注语料的基于统计的N-gram模型结合词典来完成症状、体征命名实体的识别。

2 医案文本采集及预处理

2.1 医案来源

医案选择由董建华、王永炎两位院士主编的人民卫生出版社出版的《中国现代名中医医案精粹》丛书第1至第6集（以下简称名中医医案丛书）作为研究对象。整套丛书共收录了434位全国三批名老中医的医案，其中不少医案由名老中医自行整理，并分析其机理，探讨用方用药奥秘[9]。对这些医案进行结构化并深入研究对名老中医知识的传承是大有裨益的，并且此研究方法还可以推广到其他非结构化医案的研究。

2.2 医案编排特点

名中医医案丛书对医案编排的基本规范为：第一段以姓氏、性别、年龄独立成段，大部分医案有主诉及病史、诊查、辩证、治则或治法、处方、几诊等部分。同时也发现部分医案辩证和治法融合在一起；部分医案有辩证但缺失治则或治法；部分与针灸相关的医案用操作一词替代治法等。为了在医案图片识别过程中对医案进行初步结构化，针对上述问题，收集了医案中的同义词或者对应的结构词，见**表1**。

2.3 医案文本采集及初步结构化

本文的数据采集处理对象为网络下载的加密扫描版PDF书籍，加密的PDF一般无法直接进行文字转换，需先将PDF书籍切割成医案图片进行Base64编码后，再使用光学字符识别（optical character

recognition，OCR）技术转化为计算机能够处理的文字，由于百度AI开放平台的OCR的准确率高达99%，本文采用百度的HTTP在线接口将图片转换为文字，校验无误后录入数据库。

在名中医医案丛书中，医案之后都有按语，医案文本长短不一，且本次研究只关注医案结构化不考虑按语，因此采用人工方式只对医案进行截图，确保每个医案图片可通过OCR获得正确的文本。为了方便后期处理，将每份医案截图保存到相应文件夹中，并对其编码。医案文件夹的编码规则为集号+该医案在书籍中的顺序，医案图片编码规则为集号+该医案在书籍中的顺序_该医案图片总数_目前该图片顺序。例如名中医医案丛书第2集中的第808个医案需要截图2张，则需要创建名为2808的文件夹，文件夹中将依次存放编号分别为2808_2_1和2808_2_2的两张图片。在识别过程中，这种编码可以按文件名从小到大的顺序识别并获得各个医案，并且能够很好地标识该医案的出处，方便后期对识别所得医案文本进行修订。

本文对4 902份医案截取了7 287张图片，并将医案图片用Base64转码，再将50份医案分为一组，采用OCR识别图片中的文字。在识别过程中，根据医案编排特点和分割关键字对应表对识别的字符串做切割，得到初步结构化的医案文本，并录入数据库。经人工核对，除去批量录入数据库时出错、信息不全的医案数据，最后整理出有效的医案数据共4 754例，结果如**图1**所示。

表1 分割关键词对应表示例

分割关键词	同义词或对应结构词
女	女孩、女士、妇、妇女、小姐
初诊	主诉及病史、诊查
治法	治则、操作
处方	方药

医案中患者姓名、治则或治法、处方等的编排基本一致，在文本识别过程就做了结构化。为了保证采集的数据都能溯源，数据库中还保存了原文、原文出处及处理的图片信息等，由此获得的初步结构化内容包括患者的姓名、性别、年龄、主诉及病史、诊查、辩证、治法、处方、医生、医案来源、原文、对应图片信息等。主诉及病史、诊查的文本基本采用非结构化的自然语言描述，其中包含大量症状、体征的描述，下一步的工作就是集中结构化此部分内容。

3 医案症状体征数据提取

3.1 N-gram模型

N-gram模型是一种基于统计的语言模型，可用于分词。给定一个句子w，

ID	姓名	年龄	性别	主诉及病史	辩证	治法	处方	医生	医案来源	原文	图片ID
1	王某	2	女	1983年6月16	阴虚生热;伤及经络	滋阴清热;通经活络	鳖甲,10g青蒿5g白薇5(	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	王某,女,2岁半初诊:1983	00001_3_1.PNG;(
2	高某	6	女	1974年3月7日	脑髓不充;肝风未熄	平肝熄风;滋益肝肾	生紫贝齿60g生紫石英(	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	高某,女,6岁。初诊:1974	00002_1_1.PNG;
3	郭某	50	女	1983年6月9日	心气不足;心阴亏损	养心安神	茯神,12g丹参10g生龙(	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	郭某,女,50岁初诊:1983	00003_1_1.PNG;
5	刘某	34	男	1951年5月7日	病在半表半里之间	拟以截疟之法治之	常山9g草果9g柴胡4.5(	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	例一刘某,男,34岁初诊:1	00004_2_1.PNG;(
7	孙某	18	男	1953年6月10	诊属每日疟;由于阴	拟滋阴截疟开举;扶	青蒿4.5g常山9g条苓9(	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	例二孙某,男,18岁。初诊	00005_1_1.PNG;
8	王某	49	女	1983年9月22	属阴血不足;肝阳上	拟先以轻剂平肝熄风	珍珠母,30g白僵蚕10g(	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	王某,女,49岁。初诊:198	00006_2_1.PNG;(
9	冯某	43	女	1983年6月30	肝气郁结;肝风内动	疏肝解郁;熄风定志	合欢花,10g夜交藤15g(	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	冯某,女,43岁初诊:1983	00007_1_1.PNG;
10	张某	43	女	1983年8月1日	病发于暑;暑温湿热	清利湿热	鸡苏散30g,开水冲泡,滑(	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	张某,女,43岁初诊:1983	00008_1_1.PNG;
11	安某	50	女	1983年6月20	肝胃不和;痰热互结	宽胸理气;涤痰开结	炒川连,5g清半夏5g全(	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	安某,女,50岁。初诊:198	00009_1_1.PNG;
12	骆某	54	女	1983年9月21	中风之体;肝肾阴亏	不变但可去活血化(	菟蔚子15g桃仁10g红(	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	骆某,女,54岁。初诊:198	00010_1_1.PNG;

图1 医案文本初步结构化结果部分截图

$w=\omega_1\omega_2\omega_m$表示句子由$m$个有序的词组成，$P(w)$表示句子出现的概率，N-gram模型可用于计算句子概率。在现实中句子是多样的，即使将互联网上的文本作为语料库，也不能穷尽所有的句子形式，单个句子的出现频次多为1，句子重复出现的概率低而导致数据稀疏，因此直接计算$P(w)$是非常难的[10]。考虑到句子由词构成，词是有限的，$P(w)$可以由$P(\omega_1,\omega_2,\cdots,\omega_m)$表示，假设词$\omega_i$的出现只与该词前面$N-1$个词相关，则$P(w)$的计算就可以转化为如下计算式：

$$P(\omega_1,\omega_2,\cdots,\omega_m)=\prod_{i=1}^{m}P(\omega_i|\omega_{i-N}\cdots\omega_{i-3}\omega_{i-2}\omega_{i-1}) \quad (1)$$

式（1）就是N-gram模型，当N很大时，模型的参数空间过大，会出现数据稀疏和词表维度过高的问题。N-gram模型中的ω_i可以是词也可以是字，将ω_i用于分词时，为了提高低频词分词效果，ω_i的粒度为字。若$\omega_i\omega_{i+1}\omega_{i+2}$是一个词，则其出现的概率和$P(\omega_i\omega_{i+1})P(\omega_{i+2})$或$P(\omega_i)P(\omega_{i+1}\omega_{i+2})$相似，一个词的凝固度可定义为该词出现概率与该词中其他组合概率比值的最小值，具体见式（2），本文根据词的凝固度对医案进行分词并识别新词。

$$\min\left(\frac{P(\omega_i\omega_{i+1}\omega_{i+2})}{P(\omega_i\omega_{i+1})P(\omega_{i+2})},\frac{P(\omega_i\omega_{i+1}\omega_{i+2})}{P(\omega_i)P(\omega_{i+1}\omega_{i+2})}\right) \quad (2)$$

定义预处理后的语料库为corpus，语料中的字用ω_i表示，n表示切分词的最大长度，模型识别出的词都被保存在词库VG中。本文采用的N-gram模型的具体步骤如下。

（1）对corpus中的字按1到n的顺序切分，并统计各个片段的频次，根据式（2）计算切分片段的内部凝固度。多次实验选取合适的阈值，将凝固度高于阈值且字数大于2的切分片段加入VG中。

（2）根据步骤（1）中的凝固度对句子进行切分并统计频次。切分方法是若存在两个片段的凝固度低于某个片段，则从此处切分。如存在$\frac{P(\omega_i\omega_{i+1}\omega_{i+2})}{P(\omega_i\omega_{i+1})P(\omega_{i+2})}<a$，则从$\omega_{i+1}$处切分。其中，$a$为一个给定的阈值，通过实验确定。

（3）对步骤（2）中的切分片段进行检测，若切分的片段在VG中或部分在VG中，则保留切分片段，筛选出高频片段并加入VG。

3.2 正向最大匹配法

正向最大匹配法是最基础的基于词典的中文分词算法，其算法流程如**图2**所示，MaxLen为分词词典中最长词条所包含的汉字个数。应用此算法之前需要先确定一个分词词典。

例如，待分词文本为$s1$=｛“舌”，“脉”“为”，“舌”，“红”，“苔”，“黄”，“腻”，“脉”，“弦”｝，对应分词词典为dict[]=｛“舌红”，“舌红苔黄腻”，“脉弦”｝。根据**图2**进行分词，从$s1[1]$开始，取长度为5的字符串w为“舌脉为舌红”，扫描dict[]，发现w不在dict[]中，因此去掉“红”，继续扫描“舌脉为舌”是否在词典中，如此重复上述过程直到剩下的部分是dict[]中的词或单字，并加入$s2$中。最终$s2$的结果为“舌/脉/为/舌红苔黄腻/脉弦”。

该算法的一个弊端是在算法开始前需预设一个匹配词长的初始值，初始值一般是词库中最长词的长度，如果这个词长初始值过大，在查找短词时，就会导致很多无效匹配；如果词长初始值过小，就不能进行有效的切分，这就会导致算法的效率降低[11]。

3.3 词典构建

结构化医案过程中需要尽可能保留其原始样貌，保证数据的真实和完整，因此提取临床症状、体征时不会对其做任何规范化处理。中医临床中采用自然语言描述的医案症状描述多样，如发烧就有大热、壮热、微热等描述，口渴有口渴欲饮、口渴

不欲饮等描述，为了满足后期智能诊断需求，这些症状都需作为命名实体提取。

根据《中医诊断学(新世纪第4版)》[12]及《诊断学基础(第2版)》[13]中对症状术语最小粒度的界定，以及尽可能保持数据的真实和完整的原则，整理出待处理的语料、停用词表、高频症状短语词库、西医临床诊断关键词库、中医关键词库、中医布尔类型关键词库。词库构成简述如下。

● 停用词表：由语料中非症状体征的字词构成，如患者、入院后、家属、出示、来诊时等。

● 高频症状短语词库：根据医案的结构特点，使用正则表达式将“诊查”字段与“辩证”字段之间的症状信息提取出来，并统计词频。将人工核验后为最小症状提取单元且词频大于或等于5的症状短语收录到高频症状短语库中，如烦躁不安、恶心呕吐、不思饮食、心悸气短、形体瘦削等。

● 西医临床诊断关键词库：医案中含有一部分西医诊断信息，考虑到直接删除该部分信息会导致疾病的诊断依据不完整、偏离专业诊断方向，故建立西医临床诊断关键词库，用于提取体温、尿糖、血压、黄疸指数、血清胆红素、血小板计数、白细胞计数、麝香草酸/草酚浊度试验、孕二醇测定等信息。

● 中医关键词库：医案中存在大量名词–形容词的搭配，对于同一名词，可能会出现多个形容词与之构成不同的症状短语，可将这些名词收录整理为提取语料中的症状信息的辅助工具。这些关键字有舌苔、舌质、形体、面色、二便、口、四肢等。

● 中医布尔类型关键词库：中医布尔类型关键词是指不可拆分的、在疾病描述中只有出现与否两种状态的症状短语，它在本研究医案中出现的频率低。若中医布尔类型关键词出现在语料中，则可直接提取。手足心热、午后潮热、头晕目眩、角弓反张、张口抬肩、少气懒言、潮热盗汗、形寒肢冷等都属于布尔类型关键词，对于某一患者来讲，只有是否出现该症状两种情况。

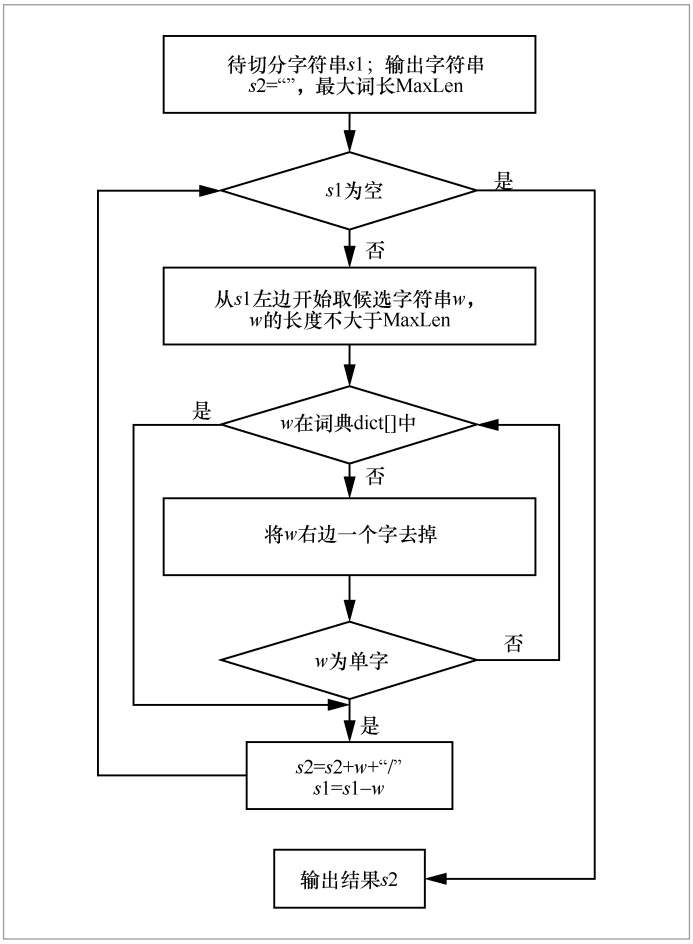


图 2 正向最大匹配算法流程

在训练模型的过程中，为了提高提取的准确度，除上述词典外，还补充了前缀修饰词库以及西医需特殊处理的关键词库，分别见**表2**与**表3**。

3.4 症状体征提取

医案结构化的一个重要工作就是将医案中的症状短语提取出来，对于一个没有足够大的短语词典以及没有人工标注的医案文本数据集，采用的提取方式是非常受限的。本文采用结合规则、词典和

表 2 前缀修饰词库示例

词库	内容
前缀修饰词库	无、不、多、右、左、未、有、稍、微、频、两、较、二、少、易、出、欲……

表 3 西医需特殊处理的关键词库示例

词库	内容
西医特殊处理的关键词库	次 / 分、CT、血检、血象、妇检、尿检、血常规、B 超检查、胃镜、病理检查、肝功能检查、胃黏膜活检、大便检查、尿常规……

N-gram新词发现的算法提取医案中的症状、体征命名实体。定义整个医案文本为S，算法具体步骤如下。

第一部分：语料预处理及高频短语提取

（1）准备语料：将第2.3节中获得的每个医案的诊查文本以标点符号为分隔符，分行存储在按文献顺序编号的单独的TXT类型的文件中。将**图1**所示文件中的主诉及病史中的文本汇总为一个TXT文件，命名为TXTcorpus。

（2）使用N-gram模型处理TXTcorpus：用第3.1中介绍的N-gram模型处理TXTcorpus，并将高频词加入对应的词典。

第二部分：对每个医案文本文件中的文本行进行处理

（3）去掉停用词：采用过滤停用词的方式过滤语料中的非症状词。

（4）识别已登录词：先用高频症状短语词库和预先准备的西医临床诊断关键词等词典提取语料中的症状单元，将其输出到相应文件中，并将其从语料中删除；若语料剩余长度为0，则读取下一个文件并进行处理。

（5）清洗剩余语料：对剩余语料再次进行停用词处理，若处理后的语料长度为0，则读取下一个文件并进行处理。

（6）识别未登录词：采用N-gram模型识别未登录词，将其输出到对应文件中，并将其更新到症状词典中后删除该词。若语料剩余长度不为0，则进行人工处理。

（7）人工处理：手动处理剩余语料，若剩余的语料为无意义的字词，则将其加入停用词表；若为症状短语，则将其手动输出到对应的文件中，并分析该词未被识别的原因，将其收录到对应的词典。

重复（3）~（7），直至所有语料处理完毕，算法结束。

3.5 实验结果及分析

本文采集到的、需要处理的名中医医案丛书的文本字数见**表4**。实验采用词典、N-gram模型、词典+N-gram模型3种方式提取医案中的症状、体征命名实体，3种方式的提取结果与医案数成比例增长，结果如**图3**所示。由于本次医案结构化的目的是为后期数据挖掘和术语规范化提供真实数据，提取对象多为包含症状描述的症状短语，其中有的短语有症状程度的描述，如口渴欲饮、口渴不欲饮；有的是多个症状同时出现时的常用描述，如寒热往来、眩晕目糊、失眠惊惕。因此，随着医案数的增加，症状短语数不断增加，而且数量可观。

为了能客观准确地描述3种方案的优劣，从4 754份医案中随机抽取666份医案组成最终的测试样本空间，剩下的4 088份医案则进行模型训练及相关处理。随机抽取的666份医案经人工处理得到5 439个症状、体征命名实体，将其作为实验评价的参考标准。不同方案提取的结果采用准确率P、召回率R、F1值3个指标进行评测，计

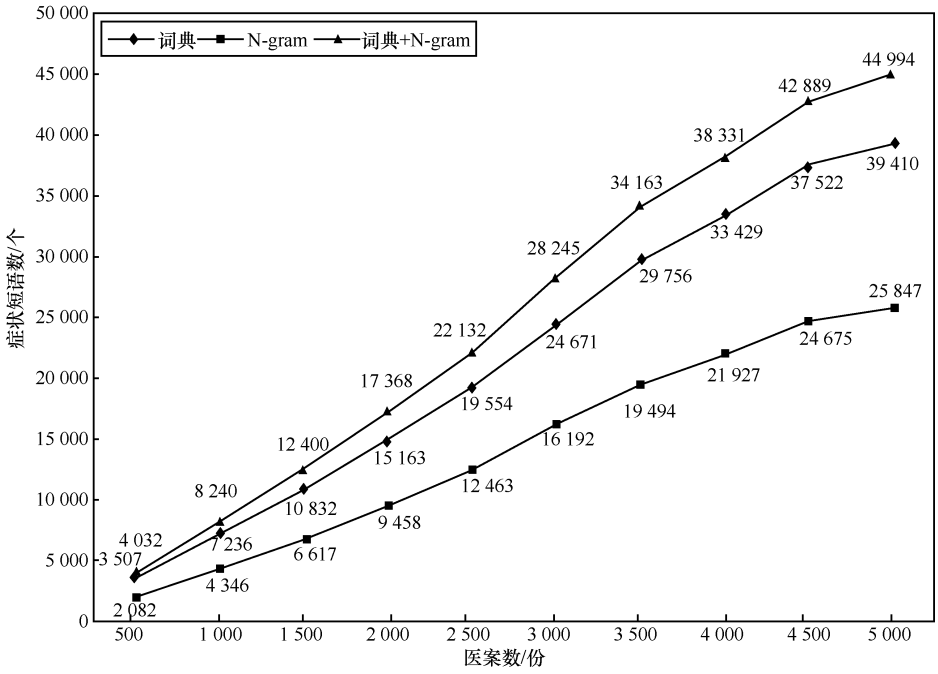


图 3　3 种方案提取的症状短语数与医案数的关系

算式如下[14]。

$$P=\frac{\text{切分结果中正确分词数}}{\text{切分结果中所有分词数}}\times 100\% \quad (3)$$

$$R=\frac{\text{切分结果中正确分词数}}{\text{切分结果中所有分词数}}\times 100\% \quad (4)$$

$$\mathrm{F1}=\frac{2\times P\times R}{P+R}\times 100\% \quad (5)$$

3种方案的实验结果见**表5**，其中正确症状总数为实验结果中与标准输出完全一致的症状短语个数，分词总数为实验提取的症状、体征命名实体总数，有效医案总数为提取命名实体数不为0的医案数，随机抽取的666份医案中都包含有症状或症状短语，因此人工处理的有效医案总数为666份，分词总数为5 439个。从**表5**可以看出，单独采用N-gram模型提取症状的结果最差，有28份医案无法识别。

表 4　文本总字数

书籍序号	文本字数/千字
1	1 165
2	1 126
3	1 203
4	1 382
5	1 050
6	947
总计	6 873

表 5　3 种方案实验结果

方案	正确症状总数/个	分词总数/个	有效医案总数/份
手工标准数据	5 439	5 439	666
词典	4 280	5 273	664
N-gram模型	1 878	3 400	638
词典+N-gram模型	4 669	5 812	664

根据**表5**计算*P*、*R*和F1值，结果见**表6**。其中，词典+N-gram模型的F1值为82.99%，词典方案的F1值为79.91%。本次实验中N-gram模型中的*N*取值为5，但通过回溯仍可提取字数超过5的长短语。实验中采用的词典除少数是人工添加的外，大量的短语来源于非人工处理：一部分采用正则表达式提取得到，一部分是训练N-gram模型时获得的未登录高频词。

本文采用的N-gram模型仅使用词的凝固度来分词，不需要对语料进行标注，节省了人力。由于提取的命名实体包含症状短语，需要保留症状的常用描述方式及非数值描述的程度描述部分，因此没有考虑词的自由度，进而“口渴”“口渴欲饮”“口渴不欲饮”等词都能被提取。使用N-gram模型训练时，采用的语料由所有医案组成，这些医案按照书籍编辑顺序保存为一个文档，并且只保留中文字符，这种做法简化了算法，但在语料不充足的情况下，很容易造成数据稀疏问题。本文采用的N-gram模型还需从数据结构、数据稀疏问题的平滑技术等方面进行优化，才能获得更高性能。

表 6　3 种提取方案的效果

方案	P	R	F1 值
词典	81.17%	78.69%	79.91%
N-gram	55.24%	34.53%	42.50%
词典 +N-gram	80.33%	85.84%	82.99%

4 医案数据结构化

完成症状、体征信息的提取后，可以根据所建词典对医案进行结构化。根据临床信息分析需求、文献来源需求统计分析医案结构，医案结构化数据由医案ID、姓名、年龄、性别、是否婚配、初诊时间、主诉及病史、诊查、辩证、治法、处方、其他诊次、医生、医案来源、原文和图片ID等字段组成，其中诊查部分完全结构化为症状、体征数据，数据之间用逗号分隔，结构化后的医案如**图4**和**图5**所示。本次医案采集数据总计4 754条，每页存储15条，共317页。在医案结构化过程中，应尽量保持医案数据原貌，不对医案中的术语进行标准化处理，其目的是

ID	姓名	年龄	性别	是否婚配	初诊时间	主诉及病史	诊查	辩证	治法	图片id
1	王某	2	女	未知	1983年6月16日	主诉及病史(其父代诉):午后…	体温37.5℃,神志尚清,答话含糊,面色不荣,精神不…	阴虚生热;伤及经络	滋阴清热;通经	00001_3_1.PNG…
2	高某	6	女	未知	1974年3月7日	主诉及病史:患儿于出生后4天…	舌质红,脉象弦细,脑发育不全,	脑髓不充;肝风未熄	平肝熄风;滋益	00002_1_1.PNG;
3	郭某	50	女	未知	1983年6月9日	主诉及病史:素罹"风湿性心脏…	精神比较紧张,面色不华,恐惧心理甚浓,整夜不成…	心气不足;心阴亏损;胆虚不眠	养心安神	00003_1_1.PNG;
5	刘某	34	男	未知	1951年5月7日	主诉及病史:患间日疟4个月余…	舌苔白腻,脉象弦滑,	病在半表半里之间	拟以截疟之法	00004_2_1.PNG…
7	孙某	18	男	未知	1953年6月10日	主诉及病史:发病月余,每夜先…	舌质红,舌苔微薄白,脉象弦细而数,	诊属每日疟;由于阴分已伤	拟滋阴截疟并	00005_1_1.PNG;
8	王某	49	女	未知	1983年9月22日	主诉及病史:左侧偏头痛已20…	左下眼睑肌肉痉挛,眼裂变小,左手有时小颤动,睡…	属阴血不足;肝阳上亢;发为…	拟先以轻剂平	00006_2_1.PNG…
9	冯某	43	女	未知	1983年6月30日	主诉及病史:因夫丧逐渐发生…	伸舌颤动,舌尖,伸出,舌质润,苔薄白,脉弦缓无力,…	肝气郁结;肝风内动	疏肝解郁;熄风	00007_1_1.PNG;
10	张某	43	女	未知	1983年8月1日	无	高热10余日,体温38~40℃,多汗口渴,便秘,尿现尿…	病发于暑;暑温湿热下注膀胱	清利湿热	00008_1_1.PNG;
11	安某	50	女	未知	1983年6月20日	主诉及病史:近40天来,胸骨后…	面色无华,舌质淡红、舌苔薄黄,脉弦滑略数,食管…	肝胃不和;痰热互结	宽胸理气;涤痰	00009_1_1.PNG;
12	骆某	54	女	未知	1983年9月21日	主诉及病史:1982年10月中风…	舌质暗红,舌面干光,脉沉弦细数,	中风之体;肝肾阴亏;津不上…	不变但可去活	00010_1_1.PNG;
13	朱某	37	女	未知	1983年11月3日	主诉及病史:素有神经衰弱,于…	表情淡漠,神清,太息,纳可,二便正常,心肺及各项…	肝郁气逆;心脾不和;病属脏…	治拟疏肝降逆;	00011_2_1.PNG…
14	张某	32	男	未知	1983年7月19日	主诉及病史:患者于1982年6…	神清,表情淡漠,失语,鼻唇沟右侧变浅,右侧歪斜,…	肝阳上亢;风痰痹阻脉络;发…	祛风通络;化痰	00012_2_1.PNG…
15	连某	29	女	已婚	1978年8月12日	主诉及病史:去岁殒胎,下血量…	经期,腹痛如引,舌边嫩红,脉象细数,	此属肝肾两损、血热血瘀	治拟补肝益肾;	00013_2_1.PNG…
16	鱼…	24	女	未知	1952年	主诉及病史:妊娠逋近7个月,…	正值发作,入视其状,四肢抽搐有力,唇紫,面青,少…	余退而语其夫:此子痫也,乃…		00014_2_1.PNG…
17	张某	24	女	未知	1972年9月13日	主诉及病史:素体尚健,于10数…	腹胀痛不欲按,胁,不思食,泛恶,口苦,舌红、苔薄…	诊为热入血室	拟清热透邪;和	00015_1_1.PNG;

1 2 3 … 317 到第 1 页 确定 共 4754 条 15 条/页

图 4　医案文本结构化结果部分截图 1

ID	处方	其他诊次	医生	医案来源	原文
1	鳖甲,10g青蒿5g白…	二诊:6月27日。上方药服3剂,体温降至正…	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	王某,女,2岁半初诊:1983年6月16日。主诉及病史(其…
2	生紫贝齿60g生紫石…	诊:1974年3月7日主诉及病史:患儿于出生…	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	高某,女,6岁。初诊:1974年3月7日主诉及病史:患儿于…
3	茯神,12g丹参10g生…	二诊:6月12日。服上药1剂后,夜可入睡,连…	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	郭某,女,50岁初诊:1983年6月9日。主诉及病史:素罹"…
5	常山9g草果9g柴胡4…	诊:1951年5月7日。主诉及病史:患间日疟…	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	例一刘某,男,34岁初诊:1951年5月7日。主诉及病史:患…
7	青蒿4.5g常山9g条…	诊:1953年6月10日。主诉及病史:发病月…	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	例二孙某,男,18岁。初诊:1953年6月10日。主诉及病…
8	珍珠母,30g白僵蚕1…	二诊:8月29日。自服中药后,已1周未头痛,…	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	王某,女,49岁。初诊:1983年9月22日。主诉及病史:左…
9	合欢花,10g夜交藤1…	二诊:7月7日。服上药7剂,精神明显好转,…	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	冯某,女,43岁初诊:1983年6月30日。主诉及病史:因夫…
10	鸡苏散30g,开水冲…	诊:1983年8月1日主诉及病史(家属代诉):…	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	张某,女,43岁初诊:1983年8月1日主诉及病史(家属代…
11	炒川连,5g清半夏5g…	二诊:服上方药1剂,即觉胸骨后烧灼感减轻…	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	安某,女,50岁。初诊:1983年6月20日。主诉及病史:近…
12	苑蔚子15g桃仁10g…	诊:1983年9月21日。主诉及病史:1982年…	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	骆某,女,54岁。初诊:1983年9月21日。主诉及病史:19…
13	蕲蒲,10g灯心3g郁…	二诊:服药1周,胸膈堵闷减轻,未见哭笑失…	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	朱某,女,37岁。初诊:1983年11月3日。主诉及病史:素…
14	丹参,12g桃仁10g苑…	二诊:7月25日。服前药后,病情明显好转,…	何世英	[VISRIS.COM]中国现代名中医医案精粹（第1集）	张某,男,32岁初诊:1983年7月19日主诉及病史:患者于…
15	川续断,10g桑寄生1…	二诊:8月16日。服药两剂腹痛除,再三四…	哈荔田	[VISRIS.COM]中国现代名中医医案精粹（第1集）	连某,女,29岁,已婚。初诊:1978年8月12日。主诉及病…
16	秦当归,12g杭白芍2…	诊:1952年仲秋主诉及病史:妊娠逾近7个…	哈荔田	[VISRIS.COM]中国现代名中医医案精粹（第1集）	鱼场下坡王某之妻,24岁初诊:1952年仲秋主诉及病史:…
17	软柴胡,6g嫩青蒿9g…	二诊:9月17日。服上方药2剂,寒热发作已…	哈荔田	[VISRIS.COM]中国现代名中医医案精粹（第1集）	张某,女,24岁,未婚。初诊:1972年9月13日。主诉及病…

图 5　医案文本结构化结果部分截图 2

为后续医案数据挖掘、临床信息标准化等相关工作提供原始数据。

5 结束语

中医医案是中医临床经验的总结，为中医治学提供了关键的第一手实践资料，对于深化、传承和发展中医药具有非常积极的作用。系统地对中医医案进行结构化整理和研究，有助于中医传承和发展。本文提出的基于自然语言处理的中医医案文本快速结构化方法，可以迅速地对图片医案或文本医案进行结构化，在结构化过程中还能动态完善中医词库，尽可能最大限度地收集中医临床术语，并保持症状描述的完整性，为后期其他医案结构化、医案数据挖掘、医学知识总结、医学知识库构建、中医术语标准化等提供信息完整的数据支持。

从实验结果来看，所提方法还有很多改进空间，后期还可以以此为基础对医案进行标注，并采用神经网络进行新词发现研究，进一步提高医案结构化效率及自动化处理能力。数据采集过程中也需要提高自动化处理程度，对现有语言模型进行补充。

参考文献:

[1] 滕文静, 孙长岗, 李雁. 浅谈不同中医医案研究方法对临床思维建立的重要性[J]. 中华中医药杂志, 2018, 33(3): 811-815.
TENG W J, SUN C G, LI Y. Discussion on the importance about different research methods of traditional Chinese medicine cases to the establishment of clinical thinking[J]. China Journal of Traditional Chinese Medicine and Pharmacy, 2018, 33(3): 811-815.

[2] 肖晓霞. 基于机器学习的中医临床症状数据元研究[D]. 长沙: 湖南中医药大学, 2018.
XIAO X X. Research on clinical symptom data elements of traditional Chinese medicine based on machine learning[D]. Changsha: Hunan University of Chinese Medicine, 2018.

[3] 翟凤文, 赫枫龄, 左万利. 字典与统计相结合的中文分词方法[J]. 小型微型计算机系统, 2006, 27(9): 1766-1771.
ZHAI F W, HE F L, ZUO W L. Chinese word segmentation based on dictionary and statistics[J]. Journal of Chinese Computer Systems, 2006, 27(9): 1766-1771.

[4] 蒋建洪, 赵嵩正, 罗玫. 词典与统计方法结合的中文分词模型研究及应用[J]. 计算机工程与设计, 2012, 33(1): 387-391.
JIANG J H, ZHAO S Z, LUO M. Analysis and application of Chinese word segmentation model which consist of dictionary and statistics method[J]. Computer Engineering and Design, 2012, 33(1): 387-391.

[5] 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(S1): 1-17.
TANG L, GUO C H, CHEN J F. Review of Chinese word segmentation studies[J]. Data Analysis and Knowledge Discovery, 2020, 4(S1): 1-17.

[6] 何晗. 自然语言处理入门[M]. 北京: 人民邮电出版社, 2019.
HE H. Introduction natural language processing[M]. Beijing: Posts &Telecom Press, 2019.

[7] 张帆, 刘晓峰, 孙燕. 中医医案文献自动分词研究[J]. 中国中医药信息杂志, 2015, 22(2): 38-41.
ZHANG F, LIU X F, SUN Y. Study on automatic word segmentation for traditional Chinese medical record literature[J]. Chinese Journal of Information on Traditional Chinese Medicine, 2015, 22(2): 38-41.

[8] 李明浩, 刘忠, 姚远哲. 基于LSTM-CRF的中医医案症状术语识别[J]. 计算机应用, 2018, 38(S2): 42-46.
LI M H, LIU Z, YAO Y Z. LSTM-CRF based symptom term recognition on traditional Chinese medical case[J]. Journal of Computer Applications, 2018, 38(S2): 42-46.

[9] 王永炎, 陶广正. 中国现代名中医医案精粹-第5集[M]. 北京: 人民卫生出版社, 2010.
WANG Y Y, TAO G Z. The essence of medical record from famous modern TCM doctor-episode 5[M]. Beijing: People's Medical Publishing House, 2010.

[10] MARTIN J H. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition[M]. Upper Saddle River: Pearson/Prentice Hall, 2009: 30-31.

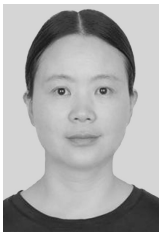
[11] 吴旭东. 正向最大匹配分词算法的分析与改进[J]. 科技传播, 2011, 3(20): 164-165.
WU X D. Analysis and improvement of forward maximum matching word segmentation algorithm[J]. Public Communication of Science & Technology, 2011, 3(20): 164-165.

[12] 李灿东. 中医诊断学(新世纪第4版)[M]. 北京: 中国中医药出版社, 2016.
LI C D. Diagnostics of traditional Chinese medicine (new century 4th edition)[M]. Beijing: China Press of Traditional chinese Medicine, 2016.

[13] 成战鹰, 王肖龙. 诊断学基础(第2版)[M]. 北京: 人民卫生出版社, 2016.
CHENG Z Y, WANG X L. Fundamentals of diagnostics (2nd edition)[M]. Beijing: People's Medical Publishing House, 2016.

[14] BIRD S, KLEIN E, LOPER E. Natural language processing with Python[M]. California: O'Reilly Media Inc., 2009.

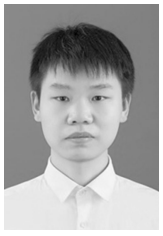
作者简介



肖晓霞(1981-), 女, 博士, 湖南中医药大学信息科学与工程学院副教授, 中国医药信息学会信息教育分会副秘书长, 主要研究方向为中医智能辅助诊断、智能数据分析、嵌入式系统。

刘明婷（1999- ），女，湖南大学信息科学与工程学院硕士生，曾获第二届全国中医药院校人工智能创新创业大赛二等奖，主要研究方向为人工智能、生物信息。

杨冯天赐（1999- ），男，湘潭大学化学学院硕士生，曾获第三届全国中医药大学生程序设计竞赛银奖，第十五届和第十六届湖南省大学生计算机程序设计竞赛三等奖，第四届团体程序设计天梯赛湖南省二等奖、全国三等奖，主要研究方向为机器学习。



刘鉴建县（1998- ），男，湖南泽塔科技有限公司Python开发工程师，主要研究方向为人工智能、机器学习。

杨阳（2000- ），女，东北林业大学工程技术学院硕士生，主要研究方向为人工智能、机器学习。

石月（1998- ），女，北京瑞迪弘欣科贸有限公司商务经理助理。

收稿日期： 2021-07-23

通信作者： 肖晓霞，amily_x@hnucm.edu.cn

基金项目： 国家重点研发计划基金资助项目（No.2017YFC1703300）；湖南中医药大学信息科学与工程学院学科开放基金项目（No.2018DK02）

Foundation Items: The National Key Research and Development Program of China (No.2017YFC1703300), Open Fund Program of School of Informatics, Hunan University of Chinese Medicine (No.2018DK02)