

# 基于自然语言处理的医学实体识别与标签提取

赵君珂,张振宇,蔡开裕  
(国防科技大学,湖南长沙 410073)

**摘要:**随着信息化建设的快速发展,数据产生了爆炸式的增长,医院每天也同样产生大量的医疗记录与数据。其中大部分内容是非结构化数据,具有真实性、主观性和不规范性,不利于解读和处理。由于医疗数据是以非结构化的文本形式存储的,因此无法直接通过计算机直接处理和分析,不仅效率低下,分析质量也无法保证。目前的信息抽取研究中使用的方法的可扩展性都较差,具有一些局限性,故自动化程度不高。文中通过自然语言处理中的规则描述语言方法,对数据中非结构化的医学命名实体进行识别,并通过语义分析进行标签提取,使非结构化的数据结构化,让数据中的描述更为准确、统一。优化了目前信息抽取方法中存在的可扩展性差的缺点,能够根据情况适应不同的情景。

**关键词:**自然语言处理;医学数据;非结构化;实体识别;标签提取

中图分类号:TP31

文献标识码:A

文章编号:1673-629X(2019)09-0018-06

doi:10.3969/j.issn.1673-629X.2019.09.004

## Medical Entity Recognition and Label Extraction Based on Natural Language Processing

ZHAO Jun-ke, ZHANG Zhen-yu, CAI Kai-yu  
(National University of Defense Technology, Changsha 410073, China)

**Abstract:** With the rapid development of information construction, data has exploded. Hospitals also produce a large number of medical records and data every day. Most of them are unstructured data with authenticity, subjectivity and irregularity, which is not conducive to interpretation and processing. Since medical data is stored in the form of unstructured text, it cannot be directly processed and analyzed by computer, which is not only inefficient, but also cannot guarantee the quality of analysis. At present, the methods used in information extraction research have poor scalability and some limitations, so the degree of automation is not high. We recognize unstructured medical named entities in data by rule description language method in natural language processing, and extract labels by semantic analysis, so that unstructured data can be structured to make the description of data more accurate and unified. It also optimizes the shortcomings of poor scalability in current information extraction methods, and can adapt to different scenarios according to the situation.

**Key words:** natural language processing; medical data; unstructured; entity identification; label extraction

## 0 引言

在信息化发展的大数据智能时代,各行各业面临着新的机遇和挑战。医疗大数据作为新的焦点领域,也得到了各界的广泛关注。如今,通过信息系统可以方便快捷地收集病人各方面的就诊信息。医院各业务系统中积累了大量的医疗数据,而这些数据存在异构、分布式、碎片化等特点<sup>[1]</sup>。

随着医疗信息系统建设的进步与互联网的广泛应用,电子病例检查报告渐渐兴起,传统纸质的手写报告逐渐退出舞台。这一现象使得医疗数据的管理更加方便、快捷<sup>[2]</sup>。病人的医疗记录是医生手动通过信息系

统录入的,而其中的内容则大多是非结构化数据。由于以医生较为熟悉的方式来描述诊断与检查结果,能够让医生在信息录入时更加迅速、准确、方便,所以目前的医疗数据文档,尤其是症状描述部分大多是以医生的口头语言进行描述的非结构化数据。

由于医疗数据是以非结构化的文本形式存储的,因此无法直接通过计算机直接处理和分析,不仅效率低下,分析质量也无法保证。目前的信息抽取研究中使用方法的可扩展性都较差,具有一些局限性,故自动化程度不高。为了能够有效地通过现有的分析方法对医学病案数据进行分析 and 信息挖掘,从而更好地利用

收稿日期:2018-10-15

修回日期:2019-02-18

网络出版时间:2019-03-28

基金项目:国家自然科学基金(61572514);长沙市科技局项目(K1705007)

作者简介:赵君珂(1994-),男(壮族),硕士研究生,研究方向为计算机科学;蔡开裕,副教授,研究方向为网络安全。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20190327.1633.074.html>

医学病案数据,如何有效地将医学数据作结构化处理就成为了一个值得研究、探索的问题。

文本结构化实际上可以认为是隶属于信息抽取技术,是该技术的一个发展方向。在针对信息抽取技术的研究过程中,研究者们通常采取基于规则和基于统计的方法<sup>[3]</sup>进行抽取,但这些方法都有不少缺点。首先框架较紧,自由度不高;其次机器学习能力不强,需要人工手动进行处理、辅助,因此存在不小的局限性,自动化程度也不尽如人意。

文中的目的是通过自然语言处理对医疗领域的数据进行处理,再结合整合基于词典、规则、机器学习、自然语言处理多种方法的关键字、语义关系提取算法,使得医生对病人情况的描述更为标准、统一,一定程度上克服了目前存在的可扩展性差的缺点,能够根据情况适应不同的情景。同时,通过打标签的方法,让用户以可视化维度的方式更全面地理解病人,从而更好地描述一个病人的各种属性。

1 非结构化医学命名实体识别

在对非结构化医学命名实体的识别过程中,首先涉及的是自然语言处理技术<sup>[4]</sup>。自然语言处理(natural language processing, NLP),有时也称为计算语言学或自然语言理解(NLU),是人工智能领域与计算机科学领域中的一个重要研究方向。自然语言处理所使用的基础工具是计算机,它对人们日常使用的具有各种表示形式的语言进行分析与处理,是语言信息处理中一个重要的研究领域<sup>[5]</sup>。

自然语言处理系统主要的核心部分是语言分析器,主要用于语法研究和语法分析<sup>[6]</sup>。而在进行语法研究和语句分析时主要是区分语义、句法、语用分析几个模块。

文中在自然语言处理方面使用了jieba分词<sup>[7]</sup>作为基础,将医疗数据进行初步的分割。但jieba分词是一个基础的分词软件,无法满足特定情形下的特定需求,如对医学命名实体的识别问题,就无法完全依赖jieba分词进行识别。因此,在经过了初步的处理后,还需要用其他的方法对数据做进一步的分析。

文中的实体识别是针对非结构化的医疗信息数据,通过对其中的实体(如症状、疾病、术语、药品)进行识别,以为下一步标签提取提供基础。

目前生物医学名称识别技术分为三大类:基于词典的方法、基于规则的方法和机器学习方法<sup>[8-9]</sup>。然而,基于字典的方法往往会漏掉字典中未提到的未定义的术语。基于规则的方法需要从文本中识别术语的规则,并且由此产生的规则并不是在所有情况下都是有效的。

1.1 规则描述语言

在后续分析中,使用了歧义切分校正的方法,属于意境语义分词<sup>[10]</sup>。在分词中,不同的字或单词在不同语句环境下会产生不同的结论,不可能提出一种能够适用全部情况的规则集,而是需要对一个个词进行研究分析,并且逐步进行补充和完善。因此需要设计规则描述语言(RDL),用以创建和保存歧义切分校正规则。

规则描述语言是产生式规则描述的工程化与具体化的产物,用以描述汉语分词、分析和生成规则。规则描述语言既具有产生式的一般形式,同时也对产生式中的各个部分定义了具体的规则与实现。规则区分为简单规则、复合规则、标号规则和控制规则。规则的结构式采用多层次的树形结构,如图1所示。

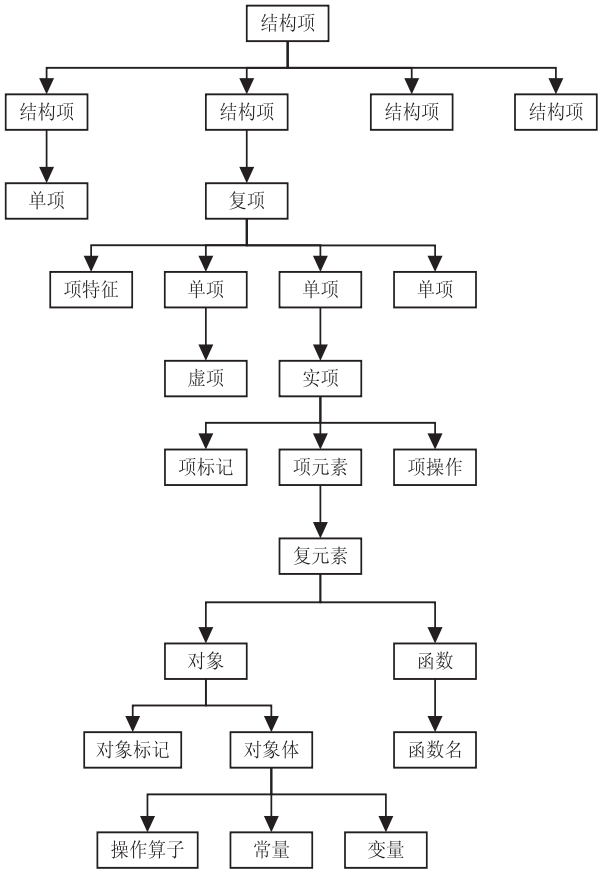


图1 规则结构式的结构树

在实际操作环境中,结构式可以用来表达分析语句的一个片段,其中结构项与语句中的“词”相对应,而有些词是单项,有些词为复项。项元素是一个词的属性或属性集的统称;项标记用来区分结构项之间的相对位置,它用项标记符以及相互间的搭配来表达某一结构项的确切位置,使得分析器能够进行准确的测试或操作;而项操作则是指对每个结构项自身及其属性的操作。

歧义切分校正中有许多规则,如1+1歧义切分、

2+1歧义切分、组合歧义切分等,例如:

(1)“ $AD_1 + VV_1$ ”校正。其中  $AD_1$  是时态副词,  $VV_1$  是一般动词,其规则为:

$\sim * VV_1 + DIS(AD_1 | VV_1) + \sim * VV_1 :: - + AD_1 + VV_1 + -$

其含义为:如果在句子中没有谓语,且有  $(AD_1 | VV_1)$  类固有歧义切分,则切分为  $AD_1 + VV_1$ 。

(2)“ $NN_1 + SF_1$ ”合成校正。其中  $SF_1$  为名词前缀。其校正规则为:

$NN_1 + '字' + VV * :: C(NN_1 + '字') + -$

其含义为:若一般名词+‘字’后面紧接着动词  $VV *$ ,则把‘字’作为名词后缀处理,  $C$  为合成函数。

根据以上规则,针对医学方面的需求加以改进,得到的算法如下:

$NN_1 + '征' + (VV * / AD * ) :: C(NN_1 + '征') + -$   
 $QQ_1 + '(可)' + (AD * ) + NN_1 :: NN_1 + -$

以第一个算法为例,对应的情况为:“Auspitz 征是一种银屑病特征。”

由于满足 Auspitz+‘征’+动词,因此把第一个征作为名词后缀处理,得到 Auspitz 征这个关键词;而同时第二个特征的‘征’并不满足这个规则,因此虽然根据分词可能还能得到特征,但是一般未必将其归纳为关键词。

通过分析句子中的语法结构,判断是否有谓语主语等词,从而对句子的形态进行判断、切分。因此这个歧义切分校正不仅能很好地识别出一些新词语,并且还能定位关键词,适当解决提取出大量无意义的词语的问题。

## 1.2 实体识别

文中使用的识别方法是机器学习方法。机器学习方法通常需要标准注释的训练数据集。大多数机器学习方法趋向于数据驱动,面向应用领域和精度、召回率和  $F_1$  值通常用于评估性能的识别。

目前,医学命名实体识别系统的最佳  $F_1$  值不如一般目标识别系统的结果。为此尝试了多种方法来改善性能,通过组合不同的方法并提出混合方法,进行机器学习后的处理,并添加生物医学领域知识。可以使用基于机器学习的命名实体识别系统来消除不正确的疾病和症状名称引起的基于字典匹配的术语识别错误。

在实体识别中,使用基于熵扩展的术语抽取<sup>[11]</sup>思想,设计算法如下:

$$\text{Entropy}(S) = - \sum_{x \in X} P(x_s | S) \log_2 P(x_s | S)$$

$$P(x_s | S) = \frac{N(x_s)}{N(S)}$$

其中,  $X$  表示关键字  $S$  周围出现的词语集合;  $x_s$  表示  $S$  周围出现的词语  $x$  与  $S$  共同出现时的字符串组

合;  $P(x_s | S)$  表示当关键字  $S$  出现时,  $x$  作为其邻接词语的条件概率,采用极大似然估计计算。Entropy( $S$ ) 值越小,说明关键字  $S$  周围出现的词语越稳定,  $x_s$  越可能是一个包含关键字  $S$  的关键词。

在提取关键词并且完成最终诊断后,还对关键词进行了后续的判断分类,例如疟疾、面容、心脏病、巩膜、神志、脓性分泌物等词,通过诊断的结果反推过程中出现的关键词,判断其对诊断是否起到了作用以及关联性,属于阳性还是阴性等。然后再通过这个方法,结合数据中的样例,计算  $F_1$  值得到关键词识别的准确率,对文中系统进行性能评估并改进不足之处,解决之前工作中存在的关键词不准确的问题。

## 1.3 识别流程

实体识别大致分为 5 个步骤,如图 2 所示。

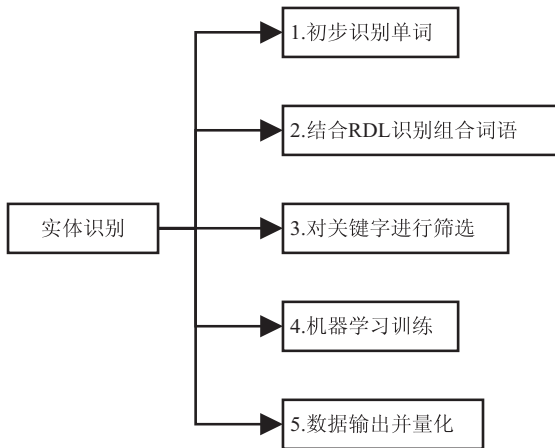


图 2 实体识别过程

初步识别单词是通过简单的算法、分析对比语料库等识别单个关键词;第二步结合 RDL 识别则是通过上面所述的方法,进一步识别形式各异的关键词;第三步是通过人工方法,针对专业领域设定规则,对识别出来的关键词进行筛选;第四步是通过机器学习训练,使系统掌握筛选的规则;最后将所得的数据进行量化,便于多个数据之间的横向比较,以进行标签提取。

## 2 医学命名实体标签提取

所谓医学术语,是指在医学活动中通过长期的实践形成的,具有明显的领域特色的专业语言。与其他专业术语一样,医学术语作为医务人员间的共同语言,在本学科间的相互沟通中发挥重要作用。如口头上常说的“脑血管意外”、“半身不遂”等各种症状,在医学术语上则称为“脑卒中”。脑卒中是严重危害人类健康的脑血管疾病之一,该病具有高发病率、高死亡率及高致残率等特点<sup>[12]</sup>。尽管这三者的意思是一样的,但作为医学期刊来说,脑血管意外、半身不遂等描述就属于通俗语言,是不专业、不规范的用词<sup>[13]</sup>。这些不规

范用词给数据分析带来了不便与麻烦。

因此,要提取出医疗数据中的标签则需要将数据中的医学术语规范化、结构化、专业化。而通过对医疗数据的标签提取之后,还要对标签进行描述,将标签作为键,而描述作为值,形成“标签-描述”的“键-值”对,即标签向量。但是通过对标签的描述进行初步提取之后,得到的标签描述值不只一个,导致其空间维数较高,无法确定其权重,因此还要对标签向量进行描述词的筛选以降低其维数。

信息处理流程如图3所示。

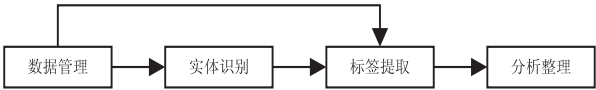


图3 信息处理流程

医疗数据保存在数据库中,主要包含病人的各种基本信息,以及病人在治疗过程中的检查、用药、医嘱等诊断信息,其中的检查信息则交由第一步实体识别处理。而标签的来源有两个,第一是数据中的基本信息,如年龄、性别、入院情况等;第二则是上一步实体识别得到的关键词。最后将标签提取出来后再进行上述的筛选、分析与处理。

2.1 标签提取

医疗数据标签提取有两个特点:一是标签是分类、层次化的,是树结构形状,例如:诊断结果的标签,检验结果的标签,用药的标签,患者主诉的标签,家族史、既往病史的标签;二是标签是标准化的,提取出的关键词整理映射成标准的医学术语和编码。将数据中的口语化语言描述,转换为医疗术语库中的专业化、规范化描述。

标签提取过程如图4所示。

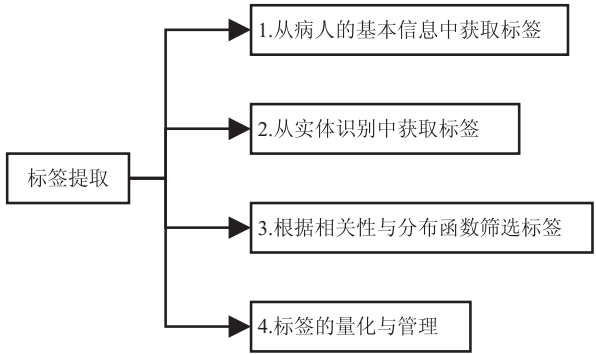


图4 标签提取过程

在实现过程中,首先是从数据库中病人的基本信息获取标签;然后再从实体识别中获得的关键字来获取所需标签;与此同时,由于同一个病人拥有的信息种类较多,因此需要对标签进行分析,进行相关性比较,构建分布图来判断标签是否影响症状、权值的大小等因素;最后,将得到的标签进行量化,构建对应的标签

库便于进行横向分析。

为了提取标签,参考了一种三元组的规则单元结构<sup>[14]</sup>,即 $[p, \Omega, T]$ ,其组成元素内涵如图5所示。

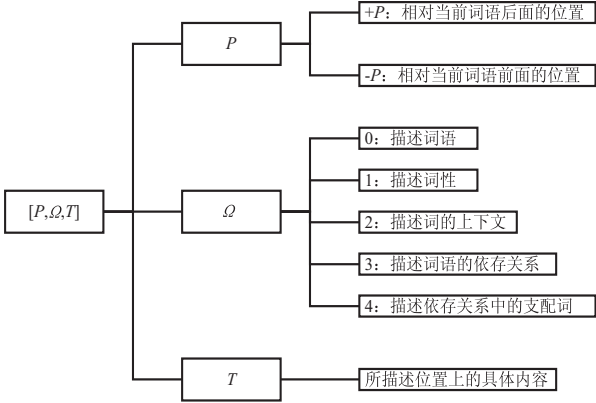


图5 标签提取规则单元结构

标签提取规则在相对位置  $p$ , 信息类型  $\Omega$  及其内容  $T$  三个方面进行了解释和设定。其中,与当前词语相关的其他词语的位置及其内容分别用  $p$  和  $T$  进行描述。 $+p$  表示相对当前词语的后面第  $p$  个位置, $-p$  表示相对当前词语的前面第  $p$  个位置。 $T$  表示所描述位置上的具体内容。信息类型  $\Omega$  从词法(词形,词性,上下文)、句法(依存关系,支配词)两个领域对对应位置所描述的信息类型进行了规定,这些类型分别用符号(0,1,2,3,4)表示。以此为基准,通过结合规则单元,构造出具有指定功能目标的标签提取规则模板,大大提高了可扩展性。

2.2 标签向量的筛选

文中针对标签描述词的筛选使用了特征提取的方法,该方法不仅能够降低标签向量的维数,筛选出多余的向量,以免影响计算结果,同时还能提高标签提取的速度和准确度。使用的特征提取方法为 CHI<sup>[15]</sup>,CHI 使用如下公式计算词  $\omega$  和标签  $t$  的相关性:

$$\chi^2(\omega, t) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}$$

其中,  $A$  为词  $\omega$  和标签  $t$  同时出现的次数;  $B$  为  $\omega$  出现而  $t$  没有出现的次数;  $C$  为  $\omega$  没有出现而  $t$  出现的次数;  $D$  为  $\omega$  和  $t$  都没有出现的次数。

取词  $\omega$  在多个标签类别中的  $\chi$  的最大值作为词  $\omega$  的特征值,将特征词按照特征值排序后保存起来。一般的标签提取过程中,特征值越大越能反映特征词的属性。

在医学领域当中,同样的病症可能有多种不同的反映,有的甚至十分罕见,而常见的病症类型即使是一般的医生也足以自主判断情况。因此,在这种条件下,特征值低的有可能反而是权值比较重的属性,可能是人们更加关心的问题。而在描述中出现的“发育正



常,营养中等”等信息都是正常指标,且多份电子医疗数据中的均可能出现,并非诊断的决定性因素,因此通过计算特征值,降低大量重复出现的描述权值,以便达到降低标签向量维度的目标。

3 实验结果

3.1 实验设计

由于文中选择医疗领域作为研究对象,对识别与提取进行的改善与优化均针对医疗领域,因此为了体现出该方法的针对性与专业性,测试语料选自军科院提供的卒中住院病历数据,共计 3 000 条住院病历,16 000 条诊断信息。住院病历主要包含病人的各种基本信息,诊断检查信息则是病人在治疗过程中的检查、用药、医嘱等情况。一个病人会对应多条诊断信息,信息通过数据库以文本形式保存。

实验数据如图 6 所示。

双肺纹理清晰,双肺散在见斑片状、条索状密度增高影; 右侧颈内静脉管腔内探及支架回声,紧贴血管壁,支架 与前片2016-04-05比较,双肺纹理稍增粗,双肺见 左室内径增大,室壁未见增厚,左室壁运动未见明显异常; 双肺下叶见少许条索影,界清,余双肺纹理清晰,肝 肝大小形态正常,包膜光滑,肝实质回声增强致密,分 双肾形态大小正常,包膜光滑,皮质回声均匀,锥体分 双肺野清晰,胸骨内可见条状高密度影;两肺上叶 双肺野清晰,胸骨内可见条状高密度影;两肺上叶 左室内径在正常范围,室壁未见增厚,左室壁运动未见 双侧颈总动脉、颈内颈外动脉形态结构及走行正常,管 左室内径正常,室壁未见增厚,左室壁运动未见明显异常
--

图 6 实验测试数据

实验通过将诊断信息输入,测试算法识别提取信息中的症状、疾病、术语、药品等各种命名实体的能力,输出提取出的关键词,主要对实体识别部分的可行性进行评估。在实验过程中,为了保证实验的准确性,同时使用基于互信息方法<sup>[16]</sup>及基于互信息与词语的共现方法<sup>[17]</sup>对数据进行测试,与文中方法进行对比,以验证该方法的有效性。

其中,基于互信息方法是基于大规模领域语料算其子串的内部结合强度,把内部结合强度超过预先设定阈值的子串抽取出来,完成术语候选的抽取;而词语共现则是指在某一语篇中词汇的使用与篇章的主题密切相关,同类词汇共同出现在同一语境中,以达到篇章连贯与衔接的目的,实现语篇的连贯功能。而文中方法则是结合了规则描述语言及基于熵扩展的术语抽取方法,针对医学领域对算法进行改进形成的方法。

为了评估文中构造的基于歧义切分校正的规则描述语言,以及基于熵扩展与机器学习的实体识别方法的性能,实验利用准确率、召回率以及  $F_1$  值为评价指标。

3.2 识别结果

实验识别结果如图 7 所示。

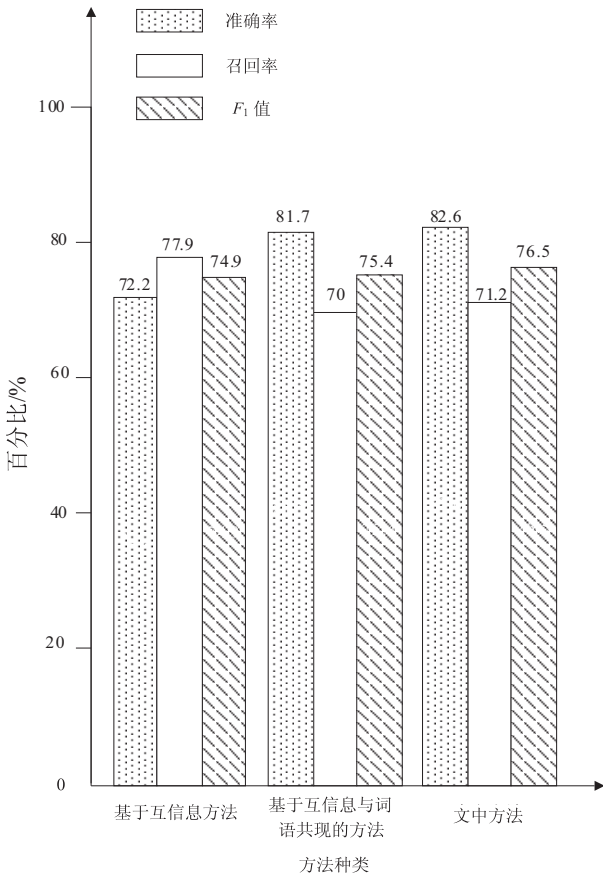


图 7 实验识别结果

实验结果表明,文中方法在针对医疗数据领域使用规则描述语言以及实体识别方法后,取得了一定的成效,识别效果有了一定的提高,识别与提取更为专业化、术语化,减少了无关词语的出现概率,使得准确率较高。虽然由于设定的规则无法概括医疗数据的各个方面,以及在向量筛选中舍弃了部分标签等因素,导致有部分关键词未能识别,召回率相对于互信息方法较低,但是总的  $F_1$  值仍然较高,证明该方法在医学领域术语抽取中是有效可行的。

4 结束语

通过歧义切分校正的自然语言处理方法,对医疗数据中非结构化的医学命名体进行了识别和关键字提取,并且用  $F_1$  值来检验识别的准确度;然后通过语义分析,将标签以及标签的描述值组合成键-值对,并且用 CHI 计算描述值的特征值。最终使得医疗数据中医生对病人情况的描述更为标准、统一。同时,通过打标签的方法,让用户以可视化维度的方式更全面地理解病人,从而更好地描述一个病人的各种属性以及状况。在标签提取过程中,经过提取规则以及特征提取的筛选,该方法相对基于统计、基于语言学的提取方

法,具有更好的扩展性,能够根据应用领域提供特定的提取方法,从而提高了提取效果。

参考文献:

[1] 席韩旭,李 维,计 虹. 基于临床数据中心的科研平台建设与实践[J]. 中国数字医学,2017,12(10):8-10.

[2] 刘一帆. 基于电子病历的科室临床数据中心的实现[D]. 广州:中山大学,2014.

[3] 何炎祥,罗楚威,胡彬尧. 基于 CRF 和规则相结合的地理命名实体识别方法[J]. 计算机应用与软件,2015,32(1):179-185.

[4] 王志勇,高 白,张少典,等. 病历智能分析系统的研究与实现[J]. 中国数字医学,2017,12(10):72-74.

[5] 崔新华. 自然语言处理在信息检索中的应用研究[J]. 贵阳学院学报:自然科学版,2012,7(3):37-40.

[6] 贾媛媛. 自然语言处理中的语义消歧研究[J]. 淮南师范学院学报,2013,15(5):108-110.

[7] 迪丽达尔·迪力沙提. 自然语言处理中的中文自动分词技术[J]. 信息与电脑,2012(11):78-79.

[8] ZHU Fei,PATUMCHAROENPOL P,ZHANG C,et al. Bio-medical text mining and its applications in cancer research [J]. Journal of Biomedical Informatics,2013,46(2):200-211.

[9] YALA A,BARZILAY R,GRIFFIN M,et al. Using machine learning to parse breast pathology reports[J]. Breast Cancer

Research and Treatment,2017,161(2):203-211.

[10] 姚天顺,张桂平. 基于规则的汉语自动分词系统[J]. 中文信息学报,1990,4(1):37-43.

[11] 樊梦佳,段东圣,杜翠兰,等. 统计与规则相融合的领域术语抽取算法[J]. 计算机应用研究,2016,33(8):2282-2285.

[12] 罗 超,吴达军. 脑卒中患者综合医院焦虑/抑郁情绪测定评分的分析及治疗[J]. 湖南师范大学学报:医学版,2014,11(2):85-87.

[13] 马芳莲. 医学术语规范化的必要性:兼谈几个常用词的辨正[J]. 科技术语研究,2000,2(4):6-8.

[14] 余琦玮,肖 颖,林 静,等. 产品评论文本中特征词提取及其关联模型构建与应用[J]. 中国机械工程,2017,28(22):2714-2721.

[15] YANG Y,PEDERSEN J O. A comparative study on feature selection in text categorization[C]//Proceedings of the fourteenth international conference on machine learning. Nashville, Tennessee: Morgan Kaufmann Publishers, 1997:412-420.

[16] 张 锋,许 云,侯 艳,等. 基于互信息的中文术语抽取系统[J]. 计算机应用研究,2005,22(5):72-73.

[17] 吴海燕. 基于互信息与词语共现的领域术语自动抽取方法演技[J]. 重庆邮电大学学报:自然科学版,2013,25(5):690-693.

(上接第 12 页)

tion with denoising autoencoders[C]//Proceedings of the 2018 conference on empirical methods in natural language processing. [s. l. ]:ACL,2018:3922-3929.

[17] KAFFEE L A,ELSAHAR H,VOUGIOUKLIS P,et al. Learning to generate Wikipedia summaries for underserved languages from Wikidata[C]//Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics. [s. l. ]:ACL,2018:640-645.

[18] LIANG P,JORDAN M I,DAN K. Learning semantic correspondences with less supervision[C]//Joint conference of the meeting of the ACL and the international joint conference on natural language processing of the AFNLP. [s. l. ]:ACL,2009:91-99.

[19] CHEN D L,MOONEY R J. Learning to sportscast:a test of grounded language acquisition[C]//Proceedings of the 25th international conference on machine learning. Helsinki, Fin-

land;ACM,2008:128-135.

[20] NOVIKOVA J,DUŠEK O,RIESER V. The E2E dataset: new challenges for end-to-end generation[C]//Proceedings of the 18th annual SIGdial meeting on discourse and dialogue. [s. l. ]:ACL,2017:201-206.

[21] KONSTAS I,LAPATA M. Inducing Document Plans for Concept-to-text Generation[C]//Proceedings of the 2013 conference on empirical methods in natural language processing. [s. l. ]:ACL,2013:1503-1514.

[22] ANGELI G,LIANG P,KLEIN D. A simple domain-independent probabilistic approach to generation[C]//Proceedings of the 2010 conference on empirical methods in natural language processing. [s. l. ]:ACL,2010:502-512.

[23] DUSEK O,JURCICEK F. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings[C]// Proceedings of the 54th annual meeting of the association for computational linguistics. [s. l. ]:ACL,2016:45-51.