计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20160611
53(12):2669-2680,2016

基于依存句法分析的病理报告结构化处理方法

田驰远　陈德华　王　梅　乐嘉锦
(东华大学计算机科学与技术学院　上海　201620)
(chiyuantian@163.com)

Structured Processing for Pathological Reports Based on Dependency Parsing

Tian Chiyuan, Chen Dehua, Wang Mei, and Le Jiajin
(*College of Computer Science and Technology, Donghua University, Shanghai* 201620)

Abstract　Most of pathological reports are unstructured texts which can not be directly analyzed by computers. The current researches on structured texts mainly focus on the information extraction. However, the syntactic features of pathological reports are particular, which makes it more difficult to extract information relations. To solve this problem, a novel method of structuralizing pathological reports based on syntactic and semantic features is proposed in this paper. First of all, we construct a synonym lexicon by using neural network language models to eliminate the phenomenon of synonymy. Then the dependency trees are generated based on the preprocessed pathological reports to extract medical examination indices. Meanwhile, we use short-sentence segmentation and annotation as optimized strategies to simplify the structure of dependency trees, which makes the grammatical relations of medical texts clearer and improves the quality of the structured results. Finally the key-value pairs of medical examination indices can be extracted from pathological reports in Chinese, and the structured texts can be generated automatically. Experimental results based on real pathological report data sets show that the performance of the proposed method on medical indices and values extraction achieves 82.91% and 79.11% of accuracy, which provides a solid foundation for related studies in the future.

Key words　medical data; pathological reports; dependency parsing; text structured processing; neural network language model

摘　要　病理检查报告中的文本通常为非结构化数据,不利于计算机自动分析和处理.目前文本结构化主要采用信息关系抽取方法,然而病理检查报告所具有的语义特殊性,给中文信息关系抽取带来了挑战.为解决上述问题,设计了一种针对病理检查报告的结构化方法,首先通过神经网络语言模型获得病理报告中的同义词表,合并一义多词现象;在此基础上,生成病理检查报告文本的依存关系树,并提出切分短句和信息标注的剪裁策略,以简化初始生成的依存关系树结构,从而使语法关系更加清晰,提高结

收稿日期:2016-08-16;**修回日期**:2016-10-24
基金项目:上海市科技创新行动计划项目(15511106900);上海市科技发展基金项目(16JC1400802);中央高校基本科研业务费东华大学励志计划项目(B201312);上海市信息化发展专项资金项目(XX-XXFZ-01-14-6349)
This work was supported by the Shanghai Innovation Action Project of Science and Technology (15511106900), the Science and Technology Development Foundation of Shanghai (16JC1400802), the DHU Distinguished Young Professor Program of Fundamental Research Funds for the Central Universities (B201312), and the Shanghai Specific Fund Project for Informatization Development (XX-XXFZ-01-14-6349).
通信作者:陈德华(chendehua@dhu.edu.cn)

构化结果的准确度;进而,利用依存句法分析结果从中文检查报告中提取指标及对应指标值,并自动生成结构化模板.实验采用医生真实使用的医疗病理检查报告进行验证,其结果表明:该方法在指标词和对应指标值提取任务中的准确率可以分别达到82.91%和79.11%,为相关研究打下了基础.

关键词 医疗数据;病理报告;依存句法分析;文本结构化处理;神经网络语言模型

中图法分类号 TP391

随着信息化建设的快速发展,目前我国医疗数据急速增长,积累了大量电子临床数据资源,为医疗大数据的分析和挖掘提供了基础.然而当前医疗文档大多是以自然语言描述的非结构化文本,由于自然语言与机器语言之间存在巨大鸿沟,导致用计算机直接处理和分析非结构化文本的效率较低,也影响了分析结果的质量.为了能有效利用现有技术成熟的分析工具对医疗文档进行数据分析和数据挖掘,从而提高医疗数据价值,对非结构化数据进行结构化就成为了该领域学者关注的重点.

病理检查报告是诊断病理学中的重要临床文档,医生将活检样本送往病理科检查,然后凭借自身经验对检查结果作出判断并将影像描述、临床诊断、诊断意见等内容以自然语言形式记录在报告中.这些文档包含的信息往往是临床医生进行疾病诊断的重要依据,也决定了病人将要接受的治疗方案.检查报告的结构化目标是发现其中包含的关键指标*key*,以及对应指标值*value*,最终形成key-value形式的结构化模板.表1给出了甲状腺超声检查报告中的1个实例,其结构化结果由13个key-value形式的二元组组成,其中诸如甲状腺大小、形态、边界等关键指标及其对应指标值是病理诊断的关键内容.从上例可以看出,将甲状腺超声检查报告中的所有描述转化成结构化模板,可保留报告中的核心信息,并建立简明规整的结构,方便读取和查询,同时也有利于借助R软件或SPSS(statistical product and service solutions)软件等现有数据挖掘工具对结构化指标与诊断结果进行关联分析,挖掘出大量与患者密切相关的医疗知识,从而辅助医生进行诊断.

Table 1 Example of Structured Thyroid Ultrasound Report

表1 甲状腺超声检查报告结构化信息举例

Example 1	Indices	Structured Key-Value Pairs
甲状腺左右叶大小及形态正常,峡部厚度正常,边界清楚,表面光滑,包膜完整,内部呈密集中等回声,回声分布均匀,右侧甲状腺内可见1个低回声结节,大小约5.1 mm×4.4 mm×5.0 mm,与包膜接触面积0~25%,其内血供程度低,血流模式为混合型.	(甲状腺)大小、形态、峡部厚度、边界、表面、包膜、内部回声、回声分布、低回声数量、(结节)大小、与包膜接触面积、血供程度、血流模式	(甲状腺大小,正常)、(甲状腺形态,正常)、(甲状腺峡部厚度,正常)、(甲状腺边界,正常)、(甲状腺表面,光滑)、(甲状腺包膜,完整)、(内部回声,密集中等)、(回声分布,均匀)、(低回声结节数量,1个)、(结节大小,5.1 mm×4.4 mm×5.0 mm)、(与包膜接触面积,0~25%)、(结节血供程度,低)、(结节血流模式,混合型)

目前,在非结构化文本转化为结构化数据领域已存在大量的研究工作,如自动问答系统[1]、关键词抽取[2]和自动摘要[3]等,而针对中文医学自然语言处理的研究相对较少,主要研究方向集中于实体识别和信息抽取,其研究对象往往是结构化或半结构化文本,于是如何对文本进行结构化便成了关键步骤.目前文本结构化技术大多采用基于规则的处理方式,但由于医疗文本中不同组织器官所具有的属性不同,且描述不同病种所使用的指标词也不同,又由于基于规则结构化方法的可扩展性较差,所以若想制定出一种适用所有病理检查报告的结构化规则十分困难.除了上述基于规则结构化方法外,还可以通过句法语义特征和词性特征识别语义,从而进行实体关系抽取和结构化处理,有效减少人工阅读工作量.Socher等人[4]提出了一种基于依存关系树识别语义的方法,利用循环神经网络将句子成分抽象为语序和句法信息,从而得到句子的语义信息.但是病理检查报告在语义特征上具有其特殊性,医生通常采用名词、形容词或名词性短语对指标进行描述,句中的谓语往往不以动词形式出现.而传统的依存句法分析方法以动词作为核心词支配其他句子成分,可见现有句法分析方法对于病理检查报告的结构化并不适用.

针对上述问题,本文在传统句法分析和信息关系抽取技术的基础上,根据医疗病理检查报告特有的语法特征,提出了一种基于依存句法分析的医疗指标结构化方法,从病理检查报告中抽取某一器官组织或病症的属性描述,随后生成依存关系树并按

照句子的语义特征形成 key-value 形式的结构化数据.实验表明:本文提出的结构化方法能够较好地针对不同组织器官的检查报告,指标词及对应指标值提取的准确率分别可达 82.91%和 79.11%,接近基于规则方法.

1 相关工作

早期对于文本关键信息抽取的研究大多采用基于启发式规则方法,其优势在于实现简单且准确率高,但获取规则是一个非常复杂的过程,且完全依赖开发人员的知识和经验,若要提高分析结果的质量,必须增加人工阅读量,且其健壮性和可移植性较差,若文档结构不适合当前启发式规则,就不得不对已有规则进行修改.在语料库构建技术越发成熟后,人们开始采用基于统计的句法分析方式,该方法采用统计学的处理技术从大规模语料库中获取语言分析所需的知识,能在减少人工规则制定的同时尽可能使语言接近真实规律[5].依存文法是由法国语言学家 Tesnière[6]于 1959 年提出的一种信息抽取方法,通过分析词语之间的依存关系揭示其句法结构,并主张句子中核心动词是支配其他词语的中心成分,而它本身却不受其他任何词语的支配,所有受支配词语都以某种依存关系从属于支配词.依存关系反映的是句中词语的语义修饰关系,它可以无视句中词的位置关系,获取长距离搭配的信息[7].

在针对非结构化中文自然语言的信息抽取研究中,其主要研究对象是命名实体之间的关系抽取,郭喜跃等人[8]提出了一种基于句法特征、语义特征的实体关系抽取方法,融入了依存句法关系、核心谓词、语义角色标注等特征,实验结果表明该方法的 $F1$ 值与传统方法相比有明显提升;甘丽新等人[9]提出了一种基于句法语义特征的实体关系抽取方法,将 2 个实体各自的依存句法关系组合,获取依存句法关系组合特征,并利用依存句法分析和词性标注获取最近句法依赖动词特征;Li 等人[10]提出了一种基于位置语义特征的实体关系抽取方法,利用位置特征的可计算性与可操作性以及语义特征的可理解性,将词语位置信息增益与基于 HowNet 语义计算结果整合在一起;在医疗领域也存在信息关系抽取方面的相关研究,Uzuner 等人[11]以句子为单位识别电子病历实体关系,并训练了 6 个支持向量机分类器实现疾病、症状、检查和治疗之间的关系识别,其结果表明词汇特征在关系识别中发挥了重要作用;Chen 等人[12]从医学文献和电子病历中分析疾病和药品实体的共现来发现二者间的关联关系,获取疾病和药品的潜在医疗知识.上述关系抽取的结果一般以二元组或三元组的形式出现,这种键值对的表示形式与本文所要提取的指标词及指标值模板相似,然而上述方法主要关注医疗领域特定实体如疾病、治疗等之间关联关系的知识发现,针对适用于不同病理检查报告的通用、自动的结构化方法目前研究较为少见.

目前,随着机器学习领域研究的发展,人们逐渐开始尝试借助机器学习方法对自然语言进行处理,而后,深度学习也从语音识别和图像处理扩展到自然语言处理领域.词向量是将深度学习方法引入自然语言处理的关键因素,而从神经网络语言模型中获取词向量已成为研究热点,许多训练词向量的工具也随之产生.Blunsom 等人[13]提出一种基于组合范畴语法方法将句子训练成高维词向量,并在向量空间中获取语义信息;Vulić等人[14]提出了一种利用 skip-gram 模型从双语词向量中训练新单词表示的方式,并将其应用于机器翻译领域;本文利用 Word2vec 词向量训练工具[15]得到病理检查报告中高频词的词向量,并通过余弦相似度合并同义词.Word2vec 是 Google 在 2013 年开发的 1 款基于深度学习的开源工具,采用“输入层-隐含层-输出层”结构的 3 层神经网络,其核心架构包括连续词袋模型(continuous bag-of-words model, CBOW)和 skip-gram 模型.其中连续词袋模型是将相邻的词向量直接加到隐含层,并用隐含层预测中间词的概率;而 skip-gram 模型通过中间词来预测周围词的概率.

文本标注方法在图像检索领域中已得到广泛使用,Tariq 等人[16]通过抽取图像所在网页中的文本信息对图像添加标注,并将文本检索与图像检索相结合,有效提高了图像检索的效率和准确性.近年来,文本标注方法也逐渐应用于自然语言处理方面,而且对于提高计算机处理自然语言的准确率起到了很大作用;Araki 等人[17]提出了基于词袋相似模型的文本标注方法,并将其用于自动问答系统中的文本检索,对于提高文本排序准确度起到关键作用.

2 系统框架

本文提出了一种基于依存句法分析的病理检查报告结构化方法,具体流程如下:1)针对病理报告中频繁出现的同一指标多种描述情况进行预处理,利用神经网络模型求出词向量,在此基础上计算余弦

相似度找出同义词,规范病理检查报告的文本表述,同时切分短句并引入词语信息标注方法简化句子结构,降低依存关系树的高度,从而使语法关系更加清晰,提高结构化结果的准确度;2)利用依存句法分析得到每个短句的依存关系树,利用所得语义特征和词性特征提取指标及对应指标值,便可将非结构化文本转化成 key-value 形式的结构化模板;3)将标注信息还原,同时修正噪声数据. 根据实现功能的不同,整个结构化过程可以划分成图 1 所示的 3 个模块:预处理模块、结构化模块、后处理模块.

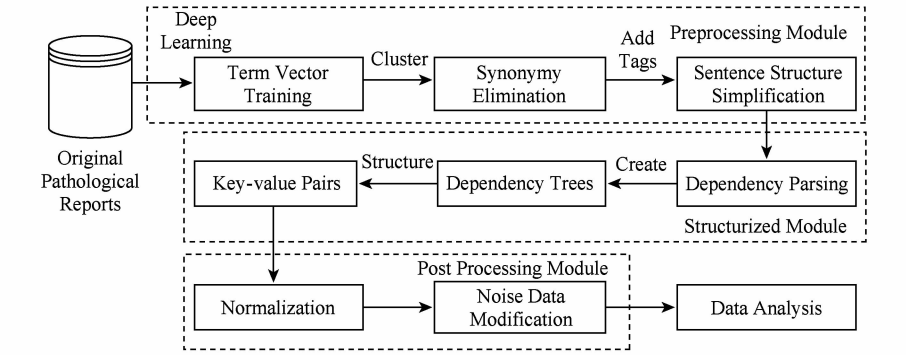


Fig. 1 Structured processing procedure of pathological reports.

图 1 病理检查报告结构化处理过程

2.1 预处理模块

一义多词在自然语言文本中是普遍现象,且在病理检查报告中尤为突出,所以设置预处理模块的主要作用是消除文本中的一义多词现象. 举例来说,"甲状腺左右叶大小正常"和"双侧甲状腺大小未见异常"是甲状腺超声检查报告中经常出现的 2 种描述,所要表达的含义都是该病人 2 侧甲状腺的大小在正常范围内,这 2 种表述的句法结构分别如图 2 所示(依存关系树的概念将在 4.1 节中详细介绍),前者用形容词"正常"作为谓语描述甲状腺的大小而后者采用动宾短语"未见异常"表达了相同的意思. 另外,在这组描述中用于表示位置信息的词也不同,前者使用了"左右叶"而后者用了"双侧".

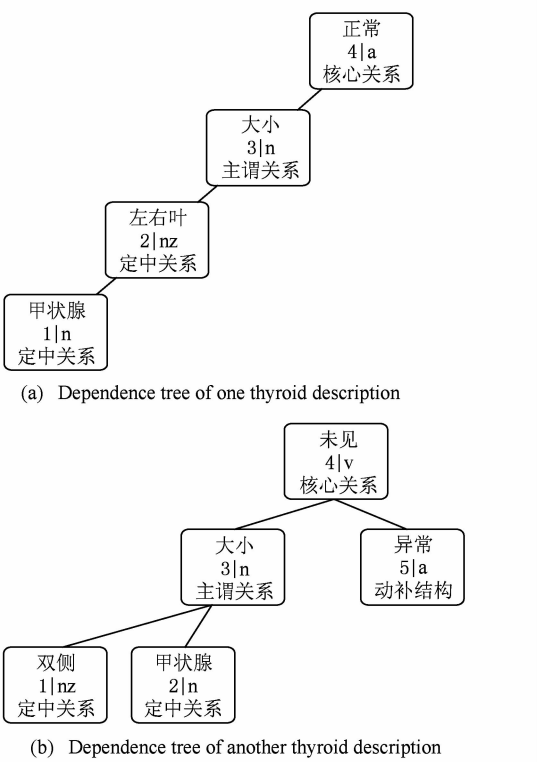


Fig. 2 Comparison between two different dependency trees expressing the same meaning.

图 2 语义相同表述不同的依存关系树对比

由此可见,中文自然语言的复杂性导致了句法分析难度的增加,所以,针对这种情况,本文在进行文本结构化处理之前设计了预处理模块,利用 Word2vec 工具训练得到词向量后计算其余弦相似度,合并余弦值大于某个阈值的词向量,从而消除一义多词现象,规范病理检查报告中的文字表述,提高结构化模块处理的准确性.

此外,在消除一义多词后,预处理模块还需要对病理检查报告中的句子结构进行了简化,将长句切分成若干短句,同时为了避免在切分短句的过程中丢失语义信息,在预处理模块中将对每个短句所描述的器官或组织等关键信息进行标注,在保留原始信息描述对象的同时也起到了本文 4.2 节中提到的简化依存关系树的作用.

2.2 结构化模块

关键指标的自动发掘和对应指标值的提取是病理检查报告结构化的关键步骤,也是整个结构化过程的核心模块,本文借助依存句法分析方法实现了针对病理检查报告的结构化模块. 依存句法分析是通过分析词与词之间的依存关系来揭示其句法结构,依存句法分析的结果可用简洁的依存关系树结

构表示,能够直接反映词与词之间的支配和依赖关系,这种支配关系有利于检测出指标及对应指标值之间的关系.

结构化模块的输入是经过预处理后的病理检查报告文本,随后对所输入的短句进行依存句法分析,可以得到词与词之间的依存关系树,通过句法分析和词性分析获取指标词 *key* 及对应指标值 *value*,最终将自然语言描述的病理检查报告转化成 key-value 键值对形式的结构化模板.

2.3 后处理模块

为了能够进一步优化结构化结果,所以在结构化模块之后设计了后处理模块. 当结构化结果输入该模块后,首先将模板中含有剪裁策略标注的特殊字符还原为对应的汉语表述;其次,借助停用词词典去除结果中的停用词,规范结构化模板的文字表述. 此外,该模块通过人工辅助审查的方式修正结果中包含的噪声数据,进而分析出预处理和结构化算法中存在的不足,既优化了结构化结果的质量,也起到了优化算法的作用,使本文的结构化算法能够适用于更广泛的情况,提高了算法的可扩展性.

3 一义多词消除算法

本文借助 Word2vec 中的神经网络语言模型训练病理检查报告中高频词汇的词向量,同时通过词向量间的余弦相似度对词向量进行聚类,得到文本中的同义词集,最终利用同义词集合并病理报告中的同义表述,消除一义多词现象.

3.1 词向量训练

在训练词向量之前,需要对原始文本进行分词操作. 本文借助 HanLP 汉语语言处理工具[18]对病理检查报告文本进行分词,其分词模块采用 Aho-Corasick 自动机[19]结合双数组 Trie 树[20]的极速多模式匹配算法,其分词速度可达到每秒 1 400 万字. 接着,将分词结果输入 Word2vec 工具训练得到病理检查报告中所有高频词的词向量. 在生成词向量的模块中,Word2vec 采用 Distributed representation 方法,该方法最早是由 Collobert 等人[21]提出,其基本思想是利用"输入层-隐含层-输出层"结构的 3 层神经网络模型将词表征为 k 维实数向量. 在早期的词向量研究方法中,词向量通常以 One-hot representation 形式表示,在获取包含文档中所有词的词汇表后,每个词向量的维度与词汇表的大小相同,向量的分量由 0 或 1 表示,若某个词在词汇表中的位置为 k,那么该词向量第 k 维为 1,其他维度为 0. 可见 One-hot representation 词向量表示方法虽然简单,但很容易造成数据稀疏与维数灾难,可扩展性较差,也无法有效反映词与词之间的语义相关性. 而 Distributed representation 方法则是利用神经网络将这些高维词向量转化成低维向量.

Word2vec 中的 3 层神经网络如图 3 所示,采用了层次化 Log-Bilinear 语言模型中的连续词袋模型,其基本实现思想是根据上下文预测周边单词出现的概率. 以预测词 w_t 出现概率为例,其计算公式如下:

$$p(w_t|context)=p(w_{t-n},w_{t-n+1},\cdots,w_{t-2},w_{t-1}),$$

其中,词 w_t 的上下文 *context* 是取 w_t 前 n 个词,以 One-hot representation 方式表示成词向量,并组合成 $|V|\times n$ 的矩阵 $\boldsymbol{C}$,其中 V 是文本所有词的集合,$|V|$ 是该集合的大小.

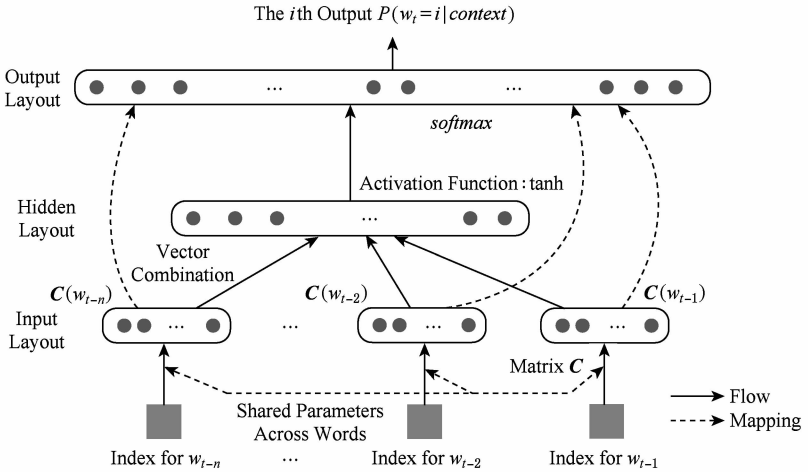


Fig. 3 Three-tier neural network architecture of Word2vec.

图 3 Word2vec 3 层神经网络结构

将矩阵 $\boldsymbol{C}$ 中的每个行向量 $\boldsymbol{C}(w_{t-n})$,$\boldsymbol{C}(w_{t-n+1})$,…,$\boldsymbol{C}(w_{t-1})$作为输入层结点,并将其首尾接拼,形成 $n\times n$ 维的向量记为 $\boldsymbol{x}$ 传入隐含层.在隐含层以 tanh 作为激活函数,得到 1 个 $|V|\times 1$ 的向量 $\boldsymbol{y}$,$\boldsymbol{y}$ 中的每个元素 y_i 表示下一个词 w_i 的未归一化概率.最后使用函数 $softmax$ 对向量 $\boldsymbol{y}$ 进行归一化,最终得到向量 $\boldsymbol{y}'$,其计算公式如下:

$$\boldsymbol{y}'=\boldsymbol{b}+\boldsymbol{W}\boldsymbol{x}+\boldsymbol{U}\tanh(\boldsymbol{d}+\boldsymbol{H}\boldsymbol{x}),$$

其中,输入向量 $\boldsymbol{x}$ 为上下文语境对应词向量的拼接向量;矩阵 $\boldsymbol{W}$ 用于表示输出层和输入层是否存在联系,通常为零矩阵即没有直接联系;矩阵 $\boldsymbol{U}$ 表示从隐含层到输出层各词的权重;$\boldsymbol{H}$ 为输入层到隐含层的权重矩阵;$\boldsymbol{b}$ 为隐含层到输出层的偏置向量;$\boldsymbol{d}$ 是输入层到隐含层的偏置向量.容易看出,本文方法通过语言模型建模,并且利用了上下文信息进而获得向量空间中的词向量表示,使语义信息更加丰富.

3.2 余弦相似度

一般而言,若词向量训练算法选取得当,生成的词向量可形成 1 个具有语义特征的词向量空间,每个向量是空间中的点,2 点之间的距离可视为词与词之间的语义相似性.词向量之间的距离可以通过欧氏距离、切比雪夫距离等公式计算,也可以利用向量之间的余弦值进行比较.本文在预处理模块中利用词向量训练工具 Word2vec 获取病理检查报告的词向量,并采用余弦值比较词与词之间的语义相关性.对于 2 个 n 维向量 $\boldsymbol{A}$ 和 $\boldsymbol{B}$,其中 $\boldsymbol{A}=(a_1,a_2,\cdots,a_n)$,$\boldsymbol{B}=(b_1,b_2,\cdots,b_n)$,2 个向量的余弦值为

$$\cos\theta=\frac{\sum_{i=1}^{n}(a_i\times b_i)}{\sqrt{\sum_{i=1}^{n}{a_i}^2}\times\sqrt{\sum_{i=1}^{n}{b_i}^2}}=\frac{\boldsymbol{A}\cdot\boldsymbol{B}}{|\boldsymbol{A}|\times|\boldsymbol{B}|}.$$

若求得的余弦值越接近 1,就表明 2 个向量之间的夹角越接近 0,也就表明 2 个向量越相似.所以本文在获取同义词表时将余弦值大于某个阈值的词归为一类,并将其中出现次数最多的词作为类别名称,用于替换病理报告中的其他同义词.

4 指标信息提取算法

本文利用 HanLP 汉语语言处理工具[18]对病理检查报告进行依存句法分析,HanLP 中的依存句法分析模块是根据词语本身、词性、后缀以及 2 词间的距离等信息,利用最大熵模型求出任意 2 个词之间可能性最大的依存关系及其概率,并由此确定该词在依存关系树中的结点位置以及与父结点之间关系,最终使用最小生成树算法得到整棵依存关系树.

4.1 依存句法分析

以甲状腺超声检查报告中的文本为例,“甲状腺左右叶大小及形态正常”是一句对甲状腺情况的影像描述,其依存关系树如图 4 所示.从图 4 中可以看出依存关系树的根结点指向每句话的核心词,其他各个结点代表句中的 1 个成分,且每个结点包含 4 项信息,分别是词原型、词所在句中位置、词性以及依存关系,其中依存关系表示结点中的词与其父结点词之间的语法关系.病理检查报告中通常以名词或形容词作为谓语,而谓语是一句话的核心成分,所以图 4 中的句法分析结果显示,形容词“正常”作为依存关系树的根结点.根据这一特性可知,病理检查报告中以名词或形容词作为核心词时,依存关系树的根结点往往是指标词或指标值.由于谓语通常直接由主语支配,句中“大小”一词与根结点之间形成主谓关系,由此判断核心词“正常”描述的对象是“大小”,于是得到 1 组 key-value 二元组:(大小,正常).而定中关系作为修饰成分,可以和指标词进行合并,最终确定这组指标词与指标值为(甲状腺左右叶大小,正常).从这个例子中可以看出利用依存关系树提取指标的基本思想,其具体实现思路将在 4.3 节中进行详述.

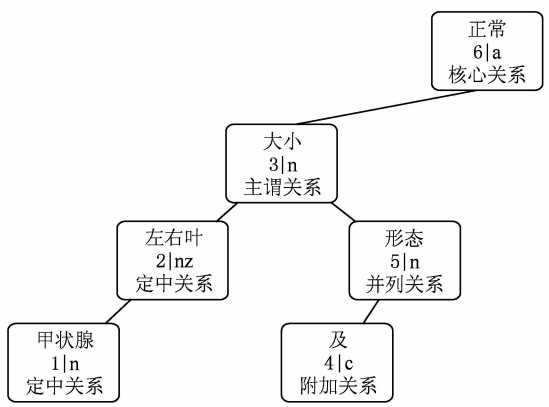


Fig. 4 One example of dependency tree.

图 4 依存关系树举例

医疗病理报告中一般以单句出现,在汉语自然语言中,单句的组成成份主要有 6 种:主语、谓语、宾语、定语、状语和补语,其中核心词一般是句中的谓语.图 4 所示的依存关系树将词之间具有语法关系的结点用边相连,单句中词与词之间最常出现的语法关系有 5 种:主谓关系、动宾关系、定中关系、状中关系及中补关系.可以清楚地看出,依存关系树结构不仅反映了词之间的依赖关系,而且给出了每个词

的词性以及不同依赖关系的类型,这为判断词与词之间的语义关系提供了良好的基础.之后便可根据词在句中的语法关系及其词性,提取关键信息.

由于本文针对的是病理检查报告,其中涉及许多医学领域的专业术语,所以本文在执行句法分析时增加了医疗领域词库,这是为了尽量避免在分词、词性标注及句法分析中发生错误,从而直接影响到指标提取的准确性.

4.2 剪裁策略

对于汉语中的长句而言,完整的依存关系树不仅结构复杂,算法运行时间也十分冗长,同时对复杂的树结构进行分析会引入大量噪声,影响句法分析结果的质量.所以本文根据病理检查报告特征,提出了一种简化依存关系树结构的方法,目的是在进行依存句法分析之前尽可能过滤无用信息,并让大多数有效信息更容易被机器获取.

中文自然语言中,一句语义完整的陈述句往往以句号结尾,句中的逗号起停顿作用,以逗号分隔的短句之间往往存在着语义上的关联.但病理检查报告的特点在于以逗号分隔的短句之间往往是独立的,它们单独成句且能够表述完整的语义,相邻短句之间不存在语义上的关联.另外,病理检查报告中会重复出现指标所描述的某些组织器官名称,这些名称的多次出现会使依存关系树的结构变得复杂,而且对于指标词的识别没有辅助作用.鉴于这些特性,本文提出了一种剪裁策略,预先构建组织器官名称与字符标注对照表,按照对照表中的信息将文本中的组织器官名称替换为特殊字符标注.

值得注意的是,切分短句虽然能够在保留语义的情况下有效降低依存关系树的高度,但是存在混淆原句信息的可能性.以超声检查报告为例,病人在1次超声检查中可能包含多个部位,如甲状腺、肝脏、肺部等,而且同一次检查的影像描述及检查结果都会记录在同一份病理报告中,当切分短句后,各短句的描述对象会有一定程度的缺失,可能导致结构化过程中出现组织器官与指标不匹配的情况.所以在标注特殊字符时不仅需要替换文本中的组织器官名称,还需要对切分后的每个短句进行标注,以确保在以短句为单位分析语义时信息不会丢失.在汉语表述中,通常以逗号分隔的2句短语所描述的对象具有一致性,基于这样的语义特征,本文制定了一种标注规则:检测当前短句中是否存在组织器官的关键词,若存在则将组织器官名称对应的特殊字符放在短句句首,并将相应名称删除;若不存在则以与前一短句相同的特殊字符进行标注.

表2给出了本文的信息标注对照表.按照表2的对应关系以及上述的病理检查报告特点,可以将"双侧甲状腺外形欠规则,包膜光整,实质内未见异常结节回声,甲状腺实质血供稍增多"这句甲状腺超声检查报告中的影像描述转化为:"@T@LR外形欠规则,@T@LR包膜光整,@N@LR实质内未见异常结节回声,@T@LR实质血供稍增多".可见,标注结果与原始文本所要表达的语义相同,故这种特殊字符标注方法具有可行性.

Table 2 Examples of Thyroid Ultrasound Report Tag Sets

表2 甲状腺超声检查报告字符标注对照表举例

Category	Name	Tag
Location	左侧	@LL
	右侧	@RR
	双侧	@LR
Tissue and Organ	甲状腺	@T
	结节	@N
	团块	
	密度影	
⋮	⋮	⋮

4.3 指标信息检测

本文的目标是通过依存句法分析和词性特征提取病理报告中的医疗指标信息及其对应指标值.根据之前对于病理检查报告语义特征以及词性特征的分析可知,句中各组成成分之间有明确的语义关系,通过对这些语义特征的分析可以提取出指标词 key 及对应指标值 $value$,提取步骤可分为3步:

步骤1.判断核心词的词性;

步骤2.遍历子树,根据核心词词性寻找与之相关的语义关系;

步骤3.根据依存句法分析得到的语义特征,提取指标词 $key_i(i\in\mathbb{N}_+)$ 或指标值 $value$,形成二元组 $(key_i,value)$.

利用语义特征提取指标时,以下5种语义关系可以指示指标词 key 与指标值 $value$ 在句中的关系:主谓关系、动宾关系、定中关系、动补关系、并列关系.在寻找这5种语义关系时应遵循4条规则:

规则1. 若核心词为动词,当其孩子结点中存在与之成主谓关系或动宾关系的名词结点时,从语义角度来看主语往往是整句话描述的对象,而宾语是动词的直接对象,由此可以确定二者分别对应为指标词 key 和指标值 $value$;

规则 2. 若核心词是名词或形容词,这在汉语自然语言中属于名词或形容词充当动词的情况,所以核心词是整句话描述的关键信息,可以将其确定为指标值 *value*,与之成主谓关系的词便是指标词 *key*;

规则 3. 由于定语起到修饰作用,所以可将形成定中关系的名词与形容词进行合并,组成 1 个指标词 *key* 或指标值 *value*;

规则 4. 由于并列关系成分在句中起到相同的作用,若判定其中一词是指标词 *key*,那么另一个也可视作指标词 *key*,同理若其中一词是指标值 *value* 则另一个也是指标值 *value*.

根据上述规则,可以得到基于依存句法分析提取指标算法,算法 1 和算法 2 的伪代码如下:

算法 1. 指标提取主程序.

输入:依存关系树邻接表;每个结点是 1 个四元组:(*ID*,*LEMMA*,*POSTAG*,*DEPREL*),其中 *ID* 表示词在原句中的位置,*LEMMA* 是词本身,*POSTAG* 表示词性(n 表示名词,v 表示动词,a 表示形容词),*DEPREL* 表示结点与其父结点之间的依存关系;邻接表表头包含所有结点的 *ID*;

输出:二元组(key_i,*value*),其中 $key_i(i\in\mathbb{N}_+)$ 是指标词,*value* 是指标值.

① CASE WHEN *root*→*POSTAG*='n'
② 提取 *root*→*LEMMA* 为指标词 key_i;
③ 对每个 *root* 的孩子结点而言
④ 将所有 *DEPREL*='并列关系'结点提取为新指标词 $key_j(j\in\mathbb{N}_+)$;
⑤ 将所有 *DEPREL*='定中关系'结点提取为指标值 *value*,并调用算法 2;
⑥ CASE WHEN *root*→*POSTAG*='v'
⑦ 对每个 *root* 的孩子结点而言
⑧ 将所有 *DEPREL*='主谓关系'结点提取为指标词 key_i;
⑨ 将所有 *DEPREL*='动宾关系'or'补关系'点提取为指标值 *value*,并调用算法 2;
⑩ CASE WHEN *root*→*POSTAG*='a'
⑪ 提取 *root*→*LEMMA* 为指标值 *value*;
⑫ 对每个 *root* 的孩子结点而言
⑬ 将所有 *DEPREL*='主谓关系'结点提取为指标词 key_i,并调用算法 2.

算法 2. 深度遍历子树提取算法.

输入:依存关系树结点 *node*;

输出:指标词或指标值集合.

① 从 *node* 开始深度遍历其子树
② CASE WHEN *root*→*POSTAG*='n'
③ 将所有子树中 *DEPREL*='定中关系'的结点按照遍历顺序与指标值 *value* 合并;
④ CASE WHEN *root*→*POSTAG*='v'
⑤ 将所有子树中 *DEPREL*='并列关系'的结点按照遍历顺序与指标值 *value* 合并;
⑥ CASE WHEN *root*→*POSTAG*='a'
⑦ 将所有子树中 *DEPREL*='定中关系'的结点按照遍历顺序与所有指标词 key_i 合并;
⑧ 将所有子树中 *DEPREL*='并列关系'的结点提取为新指标词 key_j.

结合病理检查报告的描述特征可知,结构化结果中指标词与指标值的对应关系可能是一对一或多对一关系,这是由于医生常常将多个表述相同的指标合并在一起,所以当处理完每个依存关系树后会生成若干指标词 key_i 和 1 个指标值 *value*,若结果中存在多个指标词的情况,则这些指标词 key_i 所对应的指标值均为 *value*.

5 结构化结果优化算法

为了进一步优化依存句法分析结构化结果,本文设计了后处理模块,其主要功能是在规范结构化模板中的文字表述的同时,通过分析噪声数据产生的原因优化算法.其优化算法流程如图 5 所示.

结构化结果优化算法的输入是经过依存句法分析得到的结构化模板,算法首先根据剪裁策略中制定的字符标注对照表还原模板中含有的特殊字符.随后利用停用词词典去除停用词,从而规范模板中文字的表述;接着利用在预处理模块中生成的同义词词典修正指标词和指标值的错误表述,同时配合人工校验方式删除结构化结果中的多余信息,从而去除噪声数据,提高结构化结果的正确性;最终将后处理得到的错误表述和噪声数据作为优化预处理和依存句法分析算法的依据,由于不同病理检查报告的文字特征存在差异,也存在某些特殊的表述方式,而这些差异往往导致了噪声数据的产生,所以在修复噪声数据时可以分析得到不同文档的特点,并将针对这些特殊表述的文本结构化方法加入算法,从而提高结构化的准确率,增强依存句法分析的适用性和可扩展性.

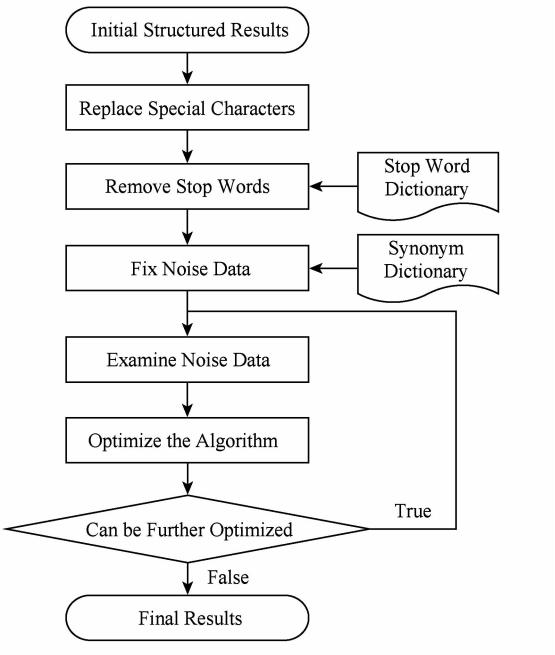


Fig. 5 Procedure of post processing module.

图 5 后处理模块流程

6 实验设置和结果分析

本文的实验数据来自某三甲医院提供的真实病理检查报告.为了使实验结果不失一般性,我们选择样本较多的甲状腺超声检查和胸部 CT 平扫检查数据作为本次实验的测试数据集,2 个数据集的记录数分别为 470 193 条以及 405 559 条.

在预处理模块中,根据多次实验得出的结果,当 2 个词的余弦相似度阈值设为 0.65 时,所得同义词表中单词个数较多且相似度较高,所以本文在获取同义词表时将余弦值大于 0.65 的词归为一类.此外,本次实验将词向量的维度设置为 100 维,此时训练词向量的时间复杂度最小,而且相对于其他维度的词向量而言,100 维的词向量在向量差、向量连接以及向量乘等特征中同样具有较好的分类性能表现.类似地,在选取训练窗口大小时,经过多次实验可知,当选取窗口大小为 8 时训练词向量的时间复杂度较低,同时分类性能较好,故将训练词向量时将上下文的选词个数设置为 8 个词.

为了能够更好地获取统计信息,我们从 2 个数据集中均随机选取了 4 000 条记录作为样本进行分析,并以手工判定的方式获取准确率 P、召回率 R 及 $F1$ 度量值,进而得到统计结果.准确率和召回率是广泛应用于信息检索和统计学分类领域的 2 个度量值,常用于评价结果的质量.在本文实验中,准确率表示在所有依存句法分析得到的结构化结果中,表述正确的指标词及对应指标值所占的比例,其计算公式如下:

$$\text{准确率 } P=\frac{\text{正确识别的指标及对应指标值个数}}{\text{结构化结果的记录总数}}.$$

召回率表示已提取指标词或指标值与原病理检查报告中包含的所有指标数量的比率,反映了结构化结果是否覆盖到病理检查报告中包含的绝大多数指标词或指标值,其计算公式可表示为

$$\text{召回率 } R=\frac{\text{正确识别的指标及对应指标值个数}}{\text{原病理报告中包含指标的总数}}.$$

$F1$ 值是准确率和召回率的调和平均值,可以综合反映文本结构化结果的好坏,计算公式如下:

$$F1=\frac{2\times\text{准确率}\times\text{召回率}}{\text{准确率}+\text{召回率}}.$$

6.1 剪裁策略效果分析

为了测试本文在预处理阶段提出的依存关系树剪裁策略的效果,本文从依存关系树的高度、词性类型占比以及依存关系类型占比这 3 个方面对甲状腺超声检查报告的原始文本、切分后文本以及剪裁后的文本进行了统计,利用 HanLP 的统计模块得到上述信息的统计结果.从表 3 可以看出,当按照短句进行依存句法分析后,依存关系树的层数明显下降,由 3.21 下降到 1.13,且表 5 所示的依存关系分布更加集中,定中关系与主谓关系所占比例明显上升,有利于算法提取指标词及其对应指标值在句中的位置.由本文 4.2 节可知,剪裁策略的主要目的是将复杂的专业术语替换为简单的特殊符号,并不会产生语序或语义的变化,所以表 3 中剪裁后文本与切分后文本的句子数量是不变的.

Table 3 Statistics of Thyroid Ultrasound Sample Data

表 3 甲状腺超声检查报告样本数据统计

Statistic Items	Original Text	Segmented Text	Annotated Text
Number of Sentences	4 000	28 629	28 629
Number of Words in Per Sentence	35.89	4.15	4.59
Average Height of Dependence Trees	3.21	1.13	1.16

另外,通过统计可以发现,指标词以及指标值的词性分布相对集中,其中 90%以上的指标词是名词词性,而指标值中名词占 2/3,其余词性大多为形容词和数量词,因此词性特征对指标信息检测而言十分重要.表 4 显示的分别是原始文本、按短句切分后文本以及执行剪裁策略后文本中包含词性类型的分

布.从列2和列3数据可以看出,当执行剪裁策略后,文本中标注的特殊字符在依存句法分析时会被识别为标点,于是标点符号在文中的占比大幅度提升,而名词占比随之下降.将专业名词转化为特殊符号的好处在于可以一定程度上增加词与词的分隔标记,减少歧义,提升分词的准确率.表5列举出了6种数量最多的语义关系,按出现次数从高到低分别是定中关系、主谓关系、核心关系、状中结构、并列关系和动宾关系.另外,表5中列2和列3数值没有发生明显变化,可见标注特殊字符对于语义的影响不大.

Table 4 Statistics of High-Frequency Part-of-Speech Types

表4 高频词性类型占比统计 %

Part of Speech	Original Text	Segmented Text	Annotated Text
Noun	40.75	46.01	29.91
Verb	8.41	9.7	9.44
Adjective	15.62	4.01	3.39
Numeral	2.59	3.04	2.75
Symbol	0.46	0.38	21.83

Table 5 Statistics of High-Frequency Semantic Relations

表5 高频语义关系占比统计 %

Semantic Relation	Original Text	Segmented Text	Annotated Text
定中关系	26.61	41.17	45.05
主谓关系	17.23	25.73	26.03
核心关系	2.74	12.99	11.74
状中结构	6.25	6.38	4.45
并列关系	19.64	4.12	4.36
动宾关系	3.53	2.16	1.9

为了进一步说明剪裁策略对于依存句法分析结果的影响,本文以甲状腺超声检查报告文本作为数据集,对其结构化的准确率和召回率进行了统计.如表6所示,在未使用剪裁策略时,结构化的准确率低于55%,相比使用剪裁策略时的准确率低了近15个百分点,召回率也低了16%,可见剪裁策略的使用可以有效减少分词和词性判断中出现的歧义情况,从而提高结构化结果的质量.

Table 6 Comparison of Annotated and Unannotated Text

表6 剪裁策略使用前后对比结果 %

Extraction Method	Index Identification			Value Identification		
	P	*R*	*F*1	*P*	*R*	*F*1
Unannotated Text	54.75	61.9	58.11	53.83	59.55	56.55
Annotated Text	71.06	87.61	78.47	70.79	86.16	77.72

6.2 后处理效果分析

为了提高算法的健壮性,本文针对结构化结果提出了后处理方法,在本节中将对后处理模块的效果进行分析.本节将从准确率、召回率和*F*1值这3个方面进行分析,同时,为了分析后处理方法对于算法可扩展性的影响,故选取了甲状腺超声和胸部CT平扫2个检查报告作为测试集.从表7和表8中的数据可以看出,后处理方法能够对2个数据集的结构化结果都起到优化作用,平均准确率提高了近10个百分点,甲状腺超声检查报告的结构化准确率更是从71.06%上升至82.45%,提升了近12%.此外,后处理模块对于结构化方法的召回率也有一定的影响,将指标词和指标值的召回率平均提升了约5%.

Table 7 Thyroid Ultrasound Report Optimization Results

表7 甲状腺超声检查报告后处理前后实验结果 %

Extraction Method	Index Identification			Value Identification		
	P	*R*	*F*1	*P*	*R*	*F*1
Dependency Parsing	71.06	84.43	77.17	70.79	81.37	75.71
Optimized Strategy	82.45	87.61	84.95	79.11	86.16	82.48

Table 8 Chest CT Scan Report Optimization Results

表8 胸部CT平扫检查报告后处理前后实验结果 %

Extraction Method	Index Identification			Value Identification		
	P	*R*	*F*1	*P*	*R*	*F*1
Dependency Parsing	73.16	84.25	78.31	69.03	82.81	75.29
Optimized Strategy	82.91	89.33	86.00	78.73	87.92	83.07

从表7、表8中还可以看出,若对不同文本进行后处理,这些文本结构化的准确率均提升到同一水准,可见不同数据集的后处理过程能够相互影响,也可以提高其他数据集结构化的准确率.从这组实验中可以看出,后处理方法在提高结构化质量中能够起到很大程度的作用,而且也提升了本文方法的可扩展性.

6.3 对比实验及分析

本节将依存句法分析结构化结果与基于人工制定规则结构化结果进行了对比,对比结果如表9和表10所示.基于人工规则的结构化方法一般是指通过关键字信息定位所要结构化的文本范围,然后通过人工阅读方式分析文本中的句式模式特征,并由此编写关系抽取算法将非结构化文本转化为结构化数据[22].本节采用的基于人工规则结构化方法是指利用现有医疗知识库,从文本中定位指标词*key*可能出现的位置,而后通过人工发现文本中包含标点

符号、数字、特殊字符及停用词等句式的特征,凭经验归纳出指标词 *key* 与指标值 *value* 在文本中的关系,从而编写出结构化算法.上述结构化方法具有较高的准确率和召回率,所以现常用于衡量其他结构化方法.考虑到该方法需要耗费大量人力阅读文本,故不再用于实际生产之中.此外,为了验证本文方法在多样化病理报告中有较强的适用性,本文选取甲状腺超声检查报告和胸部 CT 平扫检查报告 2 种医疗文档作为实验数据集.

Table 9 Comparison Results on Thyroid Ultrasound Reports

表 9 甲状腺超声检查报告结构化对比实验结果 %

Extraction Method	Index Identification			Value Identification		
	P	*R*	*F*1	*P*	*R*	*F*1
Artificial Rule	85.56	94.18	89.66	83.31	93.65	88.18
Dependency Parsing	82.45	87.61	84.95	79.11	86.16	82.48

Table 10 Comparison Results on Chest CT Scan Reports

表 10 胸部 CT 平扫检查报告结构化对比实验结果 %

Extraction Method	Index Identification			Value Identification		
	P	*R*	*F*1	*P*	*R*	*F*1
Artificial Rule	88.17	95.27	91.58	86.76	94.78	90.59
Dependency Parsing	82.91	89.33	86.00	78.73	87.92	83.07

从表 9 和表 10 中可以看出,针对本文选取的 2 个测试集,基于人工规则方法在提取指标词和指标值的准确率可以达到 85%,而召回率最高可达到 95%,可见通过基于人工规则方法能够准确地提取结构化信息,而且几乎能够覆盖所有指标.本文提出的结构化方法在 2 个数据集样本上的指标词识别准确率均可达到 82%以上,其对应指标值的准确率可达到 79%,且召回率均可达到 86%以上,可见本文方法在准确率和召回率上都能接近基于人工规则方法.虽然本文方法未能在准确率和召回率上超过基于人工规则方法,但是基于人工规则方法需要消耗大量的人力资源阅读文本,而且如果不同文本之间的语言表述存在差异,那么针对不同文本需要制定不同提取规则,可见该方法的可移植性较差.相比之下,基于依存关系的结构化方法能够省去大量人工阅读的工作量,接近 90%的召回率表明依存句法分析能够识别大部分指标词,而且能够适用于不同检查报告中的不同句式结构,可以在很大程度上实现自动化提取的目标.此外,当病理检查报告中出现新词时,若采用基于人工规则的提取方法则无法识别这些新词,但依存句法分析结构化方法仍然可以通过句法特征将其识别为指标关键字,最终转化为 key-value 形式的结构化数据.

虽然本文提出的基于依存句法分析结构化方法的准确率还有待提高,但是目前针对医疗文本结构化的研究较少,且本文方法能有效减少人工阅读大量文本的工作,虽然在后处理中仍需要人工参与校验,但这与基于人工规则进行文本结构化的方法相比,其工作量有了大幅降低,大大减少了人工参与的比重,且很容易扩展到其他医疗文档的结构化过程中,应用范围更广泛,给医疗指标结构化提供了新的思路.

7 结 论

本文针对病理检查报告的结构化进行了研究:

1) 利用神经网络语言模型尽可能地消除一义多词现象;

2) 为了提升依存句法分析结果的准确性,提出了切分短句与标注关键信息的剪裁策略;

3) 根据病理报告文本的依存关系特征,提出了一种有效的指标提取方法.在实际数据集上的实验结果验证了本文方法的有效性.然而,基于依存句法分析提取指标词及对应指标值的准确性还有待进一步提升.针对这个问题,未来的工作将尝试深度学习技术,自动获取更加准确的自然语言语义特征.

参 考 文 献

[1] Zhao Shiqi, Wang Haifeng, Li Chao, et al. Automatically generating questions from queries for community-based question answering [C] //Proc of the 5th Int Joint Conf on Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2011: 929-937

[2] Tsolmon B, Lee K. An event extraction model based on timeline and user analysis in latent dirichlet allocation [C] // Proc of the 37th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2014: 1187-1190

[3] Wan Xiaojun, Yang Jianwu. Multi-document summarization using cluster-based link analysis [C] //Proc of the 31st Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2008: 299-306

[4] Socher R, Karpathy A, Le Q V, et al. Grounded compositional semantics for finding and describing images with sentences [C] //Transactions of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2014: 207-218

[5] Wen Xu, Zhang Yu, Liu Ting, et al. Syntactic structure parsing based Chinese question classification [J]. Journal of Chinese Information Processing, 2006, 20(2): 33-39 (in Chinese)
(文勖, 张宇, 刘挺, 等. 基于句法结构分析的中文问题分类[J]. 中文信息学报, 2006, 20(2): 33-39)

[6] Tesnière L. Eléments De Syntaxe Structurale [M]. Paris: Librairie Klincksieck, 1959

[7] Hu Baoshun, Wang Daling, Yu Ge, et al. An answer extraction algorithm based on syntax structure feature parsing and classification [J]. Chinese Journal of Computers, 2008, 32(4): 662-676 (in Chinese)
(胡宝顺, 王大玲, 于戈, 等. 基于句法结构特征分析及分类技术的答案提取算法[J]. 计算机学报, 2008, 32(4): 662-676)

[8] Guo Xiyue, He Tingting, Hu Xiaohua, et al. Chinese named entity relation extraction based on syntactic and semantic features [J]. Journal of Chinese Information Processing, 2014, 28(6): 183-186 (in Chinese)
(郭喜跃, 何婷婷, 胡小华, 等. 基于句法语义特征的中文实体关系抽取[J]. 中文信息学报, 2014, 28(6): 183-186)

[9] Gan Lixin, Wan Changxuan, Liu Dexi, et al. Chinese named entity relation extraction based on syntactic and semantic features [J]. Journal of Computer Research and Development, 2016, 53(2): 284-302 (in Chinese)
(甘丽新, 万常选, 刘德喜, 等. 基于句法语义特征的中文实体关系抽取[J]. 计算机研究与发展, 2016, 53(2): 284-302)

[10] Li Haiguang, Wu Xindong, Li Zhao, et al. A relation extraction method of Chinese named entities based on location and semantic features [J]. Applied Intelligence, 2013, 38: 1-15

[11] Uzuner O, Mailoa J, Ryan R, et al. Semantic relations for problem-oriented medical records [J]. Artificial Intelligence in Medicine, 2010, 50(2): 63-73

[12] Chen E S, Hripcsak G, Xu H, et al. Automated acquisition of disease drug knowledge from biomedical and clinical documents: An initial study [J]. Journal of the American Medical Informatics Association, 2008, 15(1): 87-98

[13] Blunsom P, Hermann K M. The role of syntax in vector space models of compositional semantics [C] //Proc of the 51st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2013: 894-904

[14] Vulić I, Moens M. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings [C] //Proc of the 38th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2015: 363-372

[15] Danielfrg. Word2vec [CP/OL]. San Francisco: GitHub, (2015-12-11) [2016-04-07]. https://github.com/danielfrg/word2vec

[16] Tariq A, Foroosh H. Feature-independent context estimation for automatic image annotation [C] //Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1958-1965

[17] Araki J, Callan J. An annotation similarity model in passage ranking for historical fact validation [C] //Proc of the 37th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2014: 1111-1114

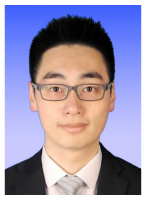
[18] Hankcs. HanLP [CP/OL]. San Francisco: GitHub, (2015-07-12) [2016-10-16]. https://github.com/hankcs/HanLP/releases

[19] Aho A V, Corasick M J. Efficient string matching: An aid to bibliographic search [J]. Communications of the ACM, 1975, 18(6): 333-340

[20] Aoe J. An efficient digital search algorithm by using a double-array structure [J]. IEEE Trans on Software Engineering, 1989, 15(9): 1066-1077

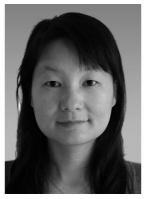
[21] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch [J]. Journal of Machine Learning Research, 2011(12): 2493-2537

[22] Buchanan B G, Shortliffe E H. Rule-based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project [M]. Boston: Addison Wesley, 1984



Tian Chiyuan, born in 1990. Master candidate. His main research interests include natural language processing and database.

Chen Dehua, born in 1976. PhD and associate professor. His main research interests include database, data warehouse, big data and deep learning.



Wang Mei, born in 1980. PhD and professor. Member of China Computer Federation. Her main research interests include database, image semantic analysis and information retrieval (wangmei@dhu.edu.cn).

Le Jiajin, born in 1951. Professor and PhD supervisor. Member of China Computer Federation. His main research interests include database and data warehouse, software engineering theory and practice (lejiajin@dhu.edu.cn).