文章编号：2095-2163(2019)04-0021-07 中图分类号：TP311 文献标志码：A

基于深度学习的甲状腺病史结构化研究与实现

骆轶姝[1,2]，申舒心[1]，陈德华[1]
（1 东华大学 计算机科学与技术学院，上海 201620；2 东华大学 资产管理处，上海 201620）

摘 要：甲状腺病史作为一类重要的非结构化文档，对医疗诊断至关重要。针对具体的甲状腺病史数据，提出一种基于深度学习的甲状腺病史结构化处理方法。首先，构建专业词库和病史本体，使用专业词库指导分词，基于本体结构完成结构化输出；其次，通过使用实体识别技术，完成对分词结果标签的预测；最后，使用标签抽取和词库匹配两种方法对病史数据进行信息抽取，并将结构化结果以 RDF 进行存储。实验结果表明该方法的准确率和泛化性较传统方法有明显提升。

关键词：甲状腺；病史；深度学习；实体识别

Research and implementation of structuring medical record of thyroid disease based on deep learning

LUO Yishu[1,2], SHEN Shuxin[1], CHEN Dehua[1]
(1 School of Computer Science and Technology, Donghua University, Shanghai 201620, China;
2 Asset Management Office, Donghua University, Shanghai 201620, China)

【Abstract】 As an important class of unstructured documents, the medical record of thyroid disease is critical to disease diagnosis. According to the specific medical record of thyroid disease data, a method for structuring medical record of thyroid disease based on deep learning is proposed. Firstly, professional dictionary and medical record ontology are constructed. Thus, word segmentation is realized by using professional dictionary and structured output is completed based on ontology structure. secondly, entity recognition technology is employed to complete prediction of segmentation result label; finally, label extraction and dictionary matching are used to extract information from the medical record data, and the structured results are stored in RDF. The results of experiments show that the accuracy and generalization of the method are significantly improved compared with the traditional methods.

【Key words】 thyroid; medical record; deep learning; entity recognition

0 引 言

随着医学信息化水平的不断提高，逐渐积累了越来越丰富的非结构化临床诊疗数据。如何有效利用这些数据已然成为目前智慧医疗领域备受关注的重点研究课题。

甲状腺疾病是内分泌科常见疾病之一。甲状腺病史作为非结构化临床诊疗数据资源，为医生诊断患者疾病提供了重要依据。但甲状腺病史结构化主要面临以下难点：适用于通用数据集的传统分词方法难以对医学领域的专业知识进行准确分词；对于传统的信息抽取方法，当应用在非标准缩写、术语以及拼写错误和不完整句子上时，难以兼顾模型的泛化性和准确性；传统的结构化输出难以为结构化数据的存储、分析、检索起到便捷支持作用。

针对上述问题，本文结合甲状腺病史数据的具体特点，提出一种基于深度学习的甲状腺病史结构化处理方法，以期为中文临床诊疗数据结构化提供参考。

1 方法

甲状腺病史完整的结构化工作包含 3 个模块，分别是：预处理模块、实体识别模块以及信息抽取模块，如图 1 所示。其中，预处理模块和此过程中构建的基础专业词库是整个框架的基础，预处理的水平直接决定了实体识别模型的效果；实体识别模块在预处理模块输出的数据集上训练得到一个模型，该模型作为标注工具用以指导结构化；信息抽取模块依赖于实体识别模型的标注结果和本体构建的结果，最终模块会将结构化文本通过 RDF（Resource Description Framework）以一种“树型”结构进行存储。

基金项目：上海市经信委人工智能创新发展专项资金项目（RX-RJJC-08-16-0483，2017-RGZN-01004）。

作者简介：骆轶姝（1974-），女，博士，副教授，主要研究方向：数据库、数据仓库与智慧医疗；申舒心（1994-），男，硕士研究生，主要研究方向：自然语言处理。

收稿日期：2019-04-29

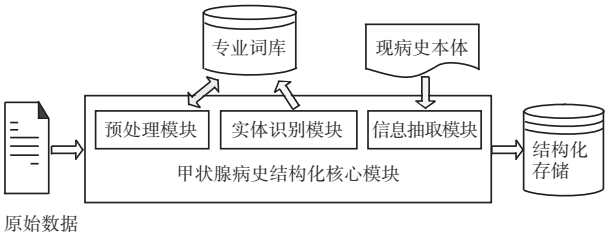

图 1 总体框架

Fig. 1 Overall framework

1.1 专业词库构建

构建专业词库主要目的在于数据预处理过程中指导原始数据分词和结构化过程中基于词库匹配进行实体抽取。使用专业词库指导文本分词则旨在避免通用分词工具对专业数据进行误分、错分;基于词库进行信息抽取的核心思想是指结合领域知识和抽取目标信息建立的字符串标识匹配与定位。词库的最初构建来源于多个专业词表,包括:ICD-10 疾病标准[1]、2017 年国家医保药品目录、ICD-9-CM[2](手术操作编码)、某三甲医院收费明细与收费标准和中华医学会内分泌分会发表的 2008《甲状腺疾病诊治指南》[3]。标准词表及其对应的实体类型和举例详见表 1。

表 1 实体举例表

Tab. 1 Examples of entity table

标准词表	实体	举例
《甲状腺疾病诊治指南》、《甲状腺疾病的诊断及个体化治疗》	症状	胸闷/心慌/心悸/手抖/手颤…
医院收费明细与收费标准	检查	CT/超声/甲状腺功能检查/MRI…
ICD-9-CM、2017 年国家医保药品目录	治疗	甲强龙注射液/丙基硫氧嘧啶…
ICD-10	疾病	甲亢/甲减/甲状腺结节/甲状腺炎…

1.2 病史本体构建

由于甲状腺病史文本表达形式多样且内容繁杂,相较于传统的句子模板,通过使用构建甲状腺病史本体的方法,对甲状腺病史文本数据进行一定程度的抽象概括,更适用于当前结构化任务。相关研究表明,基于描述逻辑和规则的本体可以进一步表述数据的语义,本体基于逻辑的知识表示形式可以有效提高知识的语义表述能力,相应的逻辑推理算法可以改进知识的发现能力和解释能力[4-5]。考虑到对结构化结果的存储、分析、检索的便捷支持需要,本文使用决策七步法构建病史本体,使用自左向右的方法构建甲状腺病史本体中的类和类之间的关系,并采用软件 Protégé 完成本体模型的构建,继而使用 RDF[6] 语言描述构建的本体模型。

(1)确定本体的专业领域和范畴。本文以医学领域为特定的研究领域,构建甲状腺病史本体:通过一套明确的体系规范甲状腺病史数据中的词汇,使数据中的术语得到统一,能够被其它领域认可;基于本体结构,使用词典匹配和实体识别标签抽取实现甲状腺病史的结构化。

(2)考虑复用现有本体的可能性。本文的原始数据来自于上海市某三甲医院的真实临床采集得到,本体的结构依据病史的内容和记录格式,且由于医生的个人习惯原因,病史的记录规则相对比较灵活,且构建本体的目的是为实现病史的结构化,目前也尚未见到可以复用或是具有参考价值的公开本体。

(3)列出本体中的重要术语。通过与标准词表进行匹配构建基础专业词库,通过专家纠错和使用实体识别算法扩充专业词库。专业词库的专业术语样本见表 2。

表 2 专业术语表

Tab. 2 Professional vocabulary list

类型	专业术语
症状	畏寒发热、心悸、心慌、声音改变、声音嘶哑…
检查	CT、B 超、甲状腺功能检查、MRI、PET…
治疗	云克片、优甲乐片、环磷酰胺片、甲巯咪唑片、小金丸…
疾病	甲亢、甲减、甲状腺结节、甲状腺炎…

(4)定义类和类的层次。通过实体识别算法构建词典和实体标签,然而大都属于专业术语,且这些实体(见表 2)的分布是散乱的,关系不明确,仅仅是信息抽取,很难达到预期结构化的效果。因此本文提出基于原始数据的记录结构,依据标签,将这些实体进行归类。类的顺序按照病史的数据结构自顶向下地逐级排序,依次是时间、地点、诱因、症状、检查、治疗、效果、入院情况和疾病;类的层次结构通常采用自左向右的方法加以确定,即先确定父类,再确定子类。将这种关系定义为 part-of 关系。

(5)定义类的属性。在第(4)步的过程中,通过提取部分术语定义了类和类之间的关系,然而简单的类名无法体现具体的知识,本体的具体知识通过定义类特有的属性来体现。本文提出将现有的属性分为 2 种,即:数据型属性和对象型属性。两者间的区别就在于实例的不同。其中,数据型属性是指实例中具有文字、字符串、数字和日期的属性,包括:时间、地点、诱因、症状、效果、入院情况和疾病的属性;

对象型属性是指实例中包含另一个子类的属性，即该属性不是具体的属性值，而是另一个父类下的一个子类，包括：检查和治疗两种属性，检查的属性是某种检查项目、接下去才是检查内容，治疗下是某种治疗方式、紧接着才是治疗内容。本文将数据型属性定义为 instance－of 关系，对象型属性定义为 attribute-of 关系。

在实体抽取的所有实体中，基本上形成了 3 种关系，见表 3。

表 3　本体关系表

Tab. 3　Relationship of ontology

关系	描述
Part-of	本体-类
attribute-of	类-属性
instance-of	属性-属性值

(6)本体决策。本体在使用前需要经过 3 个步骤进行验证，来证明本文构建的本体是否符合实际需求。首先经过逻辑推理证明本体构建逻辑无误；其次，本文基于本体结构，使用实体识别技术进行信息抽取，构建本体；最后，经由专家验证该本体的正确性。

(7)创建实例。本文使用本体的主要目的是为结构化数据的存储、分析、检索提供便捷支持，本体结构如图 2 所示。

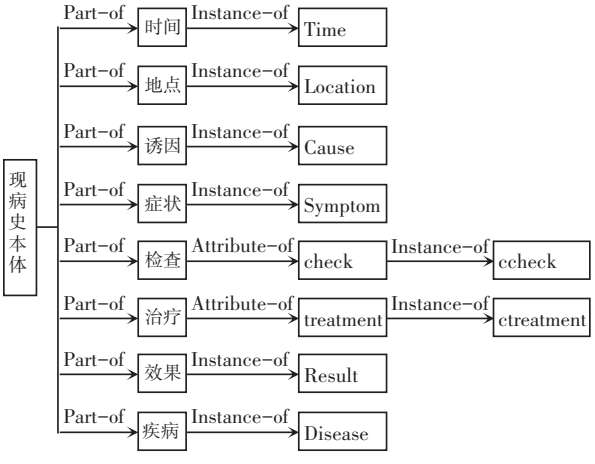

图 2　甲状腺现病史本体结构图

Fig. 2　Ontology structure of medical record of thyroid disease

1.3　数据预处理

(1)数据标准化。甲状腺疾病现病史文本数据是由医生手动录入，而不同的医生有不同的输入习惯，这主要体现在标点以及特殊符号使用上的不统一与不规范，导致分词效果并不理想。同时存在比较严重的错别字。故而在预处理过程中需要对标点符号进行规范化，并对错别字做出修改。标准化样例见表 4。

表 4　标准化样例表

Tab. 4　Standardized sample

类别	匹配目标	替换内容
标点及特殊符号	,丨;丨:丨"丨"丨///丨(丨)丨——等	,丨;丨:丨"丨"丨。丨(丨)丨－等
拼写错误	未及丨成	未见丨呈

(2)文本分词。为确保实体识别模型的顺利训练，本文依赖基于标准词表构建的专业词库对病史文本进行精准分词。针对现有的中文分词工具对专业性较高的医学文本存在错误分词的问题，建立专业词库，提高分词准确度。专业词库包括症状、疾病、检查和治疗四个子库，初始化来源于几个专业数据集。另外，分词模块中需要对训练集加上标注，专家团队对 13 类实体进行标注，产生 21 种标签用于模型的监督学习。本文基于病史数据内容对甲状腺病史分词后的数据设计标签见表 5。

表 5　实体标签表

Tab. 5　Lable of entity

标签名称	标签定义	样例
时间	time	2012-3-15、半年前等
地点	location	东方医院、瑞金医院等
诱因	cause	劳累、感冒等
否定词	negative	无、未出现等
肯定词	positive	感觉、发现等
临床表现	symptom	心慌、手抖等
检查项目	check	B 超、甲功等
检查内容	ccheck	边界清晰、回声均匀等
治疗方式	treatment	用药、手术等
治疗内容	ctreatment	心得安等
治疗效果	result	左颈肿块略缩小等
入院情况	situation	二便无殊、精神可等
疾病诊断	disease	左甲肿块、甲亢等

本文为模型设计了 13 种标注，对应不同的语义内容，这些标注包含了一定的实体信息。表 5 对部分语义内容进行了详细分类。为避免在结构化过程中的稀疏存储，本体定义没有做到细致的属性划分，这些标签最终会有助于定义结构化内容的属性。

1.4　实体识别

在专业词库构建过程中，通过使用实体识别技术对专业词库进行扩充和更新；在结构化过程中，通过使用实体识别技术对给定文本进行标签预测。本

文使用 Bi-LSTM[7-10] 作为实体识别的主体,该模型在 LSTM 的基础上加入逆向传播过程,使得网络可以同时利用上下文中的语义特征。另外,由于 Bi-LSTM 的各输出之间没有相互影响,仅仅获得独立的最大概率标签,造成 Bi-LSTM 的输出中可能存在非法标签问题,即 B-cause 后连接 I-time,本文为该模型添加 CRF[11] 的后处理层来适应多变的输出。CRF 中的转移特征会分析输出标签之间的顺序,以获得最优的标签序列。Bi-LSTM-CRF 网络结构如图 3 所示。

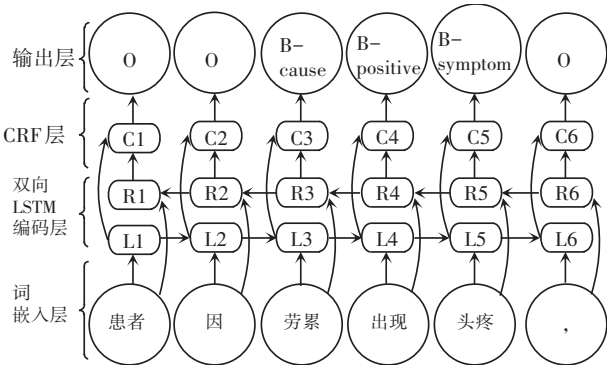

图 3 Bi-LSTM-CRF 模型结构图

Fig. 3 Model structure of Bi-LSTM-CRF

图 3 中,第一层为词嵌入层(Word Embedding layer),主要是将基于自定义词典分词后的序列文本数据转化为词向量序列,并将向量序列输入模型进行训练。第二层为 Bi-LSTM 编码层,通过 LTSM 的正向推导和反向传播对序列文本数据中的各个词进行独立分类,获取标注信息。第三层为 CRF 层,通过使用 CRF 中的条件转移矩阵从已获得标注信息的分词中选取合法标注,获得最优标注序列。第四层为输出层,在给定目标语句的情况下,通过深度学习模型可以对目标语句自动进行单词的语义标注。

1.5 信息抽取

在甲状腺病史结构化过程中,本文主要选取 2 种方法用于信息抽取研究,即在不同部分的数据使用不同的方法:对于描述相对多样化或是过度依赖上下文语义的实体使用实体识别标签抽取,见表 6。

除此以外,对于症状、检查、治疗、疾病这四类描述相对较为规范、固定的实体使用词库匹配的方法进行信息抽取。

最后结合病史本体结构,将通过上述两种方法获得的信息实现结构化输出。

表 6 实体类别-数据特点表

Tab. 6 Entity category - data characteristics table

实体类别	数据特点
就诊地点	取值范围过大
治疗效果、诱因	需结合语义
时间、入院情况	格式多样化
治疗内容、检查记录	数据结构复杂

1.6 结构化数据存储

RDF 数据模型本质上是一个图结构模型,由主语、谓词和对象组成,底层使用 XML/RDF 语言实现。由于医学专业知识具有数据库量大和增长快的特点,本文构建的现病史本体也需要以 RDF 的形式存储。常用的单节点 RDF 数据库不能满足存储现病史本体实例的需求,使用传统的关系型数据库存储又面临信息冗余高和查询性能低的问题,所以研究和构建分布式的具备图存储功能的本体存储系统是一个可行的方法。资源描述框架模型如图 4 所示。

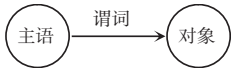

图 4 资源描述框架模型图

Fig. 4 Model of RDF

2 实验

2.1 实验数据

病史是病历中的一部分,通常包括现病史、既往史、家族史、个人史和婚育史。其中,现病史记述患者发病后的全过程,即发生、发展、演变和诊治的经过,具有数据量最大、内容最多和记录结构最复杂的特点。本文选取上海某三甲医院从 2005~2015 年十余年间、共 9 386 条甲状腺病史中的现病史数据内容作为实验数据。

现病史从内容上大致分为 4 个部分,也就是:疾病发生、病情发展、治疗经过和入院情况。其中,疾病发生主要包括:起病时间、临床症状和起病诱因;病情发展主要包括:病程中主要症状的变化、新出现症状以及伴随症状;治疗经过是指:本次就诊前已经接受过的诊断检查及其结果,治疗所用药物的名称、剂量、给药途径、疗程及疗效;入院情况是指:医生从患者病后的精神、体力状态、饮食情况、睡眠与大小便等方面,对病人得出全身情况的评价。甲状腺病史样例数据可表述如下。

2015-3 患者因劳累出现消瘦、乏力、无多汗、心慌、无手抖等症状，至徐汇区中心医院查甲状腺功能提示甲亢，给予赛治最初 20 mg，bid，口服 2 周后改用 10 mg，bid，3 周后复查甲状腺功能后改用 5 mg，bid，口服半月后复查甲状腺功能 FT3、FT4 较前升高，1 月前（2015-6-20）患者自诉双眼突出逐渐明显，并出现右眼复视，视力下降，2015-7-3 随至复旦大学附属耳鼻喉医院查双眼 CT 提示双侧甲状腺相关性眼病，查甲状腺功能提示 FT3 6.41 pmol/L，FT4 16.68 pmol/L，TSH 0.006 5 uIU/ml。今为求进一步诊治，门诊以“甲状腺相关性眼病”收住院。发病以来，患者神志清楚，精神一般，双眼突出，畏光流泪，无明显充血水肿，右眼有复视，无呕血、黑便，无腹痛，胃纳可，二便可，夜眠佳，近期未见明显体重下降。

2.2　实验与结果

（1）参数设置。本文通过平均实验结果来确定最优的参数组合，实验中采用的可调参数设置见表 7。

表 7　参数设置表

Tab. 7　Parameter settings

参数名称	取值
字向量维度	300
每块样本数量大小	32
LSTM 隐层单元限制	160
丢弃率	0.5
学习率	0.001
迭代次数	3 500

（2）评估标准。对于实体关系抽取结果的评价，本文针对全部实体分别计算准确率（*precision*）、召回率（*recall*）和 F_1 值。对应数学公式可顺次表示如下：

$$precision = \frac{TP}{TP + FP}; \tag{1}$$

$$recall = \frac{TP}{TP + FN}; \tag{2}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall}. \tag{3}$$

其中，*TP* 表示本类别中正确识别的样本数；*FP* 表示本类中标注错误的样本数；*FN* 表示原本属于本类的标注，却错误地标注为别的种类的标签的样本数。F_1 值可以加权调和平均模型的准确率和召回率，能综合地表征一个模型的优劣。

（3）实验结果。实验在现病史共定义 13 类特征实体，21 种标签，通过混淆矩阵计算出各类实体的准确率 *P*、召回率 *R* 以及 F_1 值。实验结果见表 8。

表 8　标签准确率表

Tab. 8　Accuracy of lable

Name	*P*	*R*	F_1
B-time	0.76	0.83	0.79
B-location	0.72	0.86	0.78
B-cause	0.98	0.96	0.97
B-negative	0.99	0.99	0.99
B-positive	0.87	0.89	0.88
B-symptom	0.97	0.94	0.96
B-check	0.73	0.78	0.75
B-ccheck	0.81	0.77	0.79
B-treatment	0.57	0.73	0.64
B-ctreatment	0.71	0.66	0.68
B-result	0.69	0.85	0.77
B-situation	0.88	0.97	0.92
B-disease	0.74	0.89	0.81
I-time	0.75	0.94	0.83
I-location	0.94	0.98	0.96
I-cause	0.86	0.89	0.87
I-ccheck	0.72	0.71	0.71
I-ctreatment	0.71	0.94	0.81
I-result	0.55	1	0.71
I-situation	0.95	0.91	0.93
O	0.88	0.97	0.92

甲状腺病史中现病史将识别结果绘制混淆矩阵，如图 5 所示。

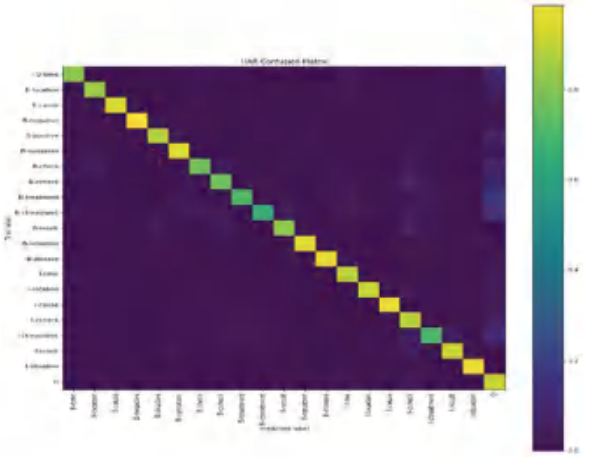

图 5　混淆矩阵图

Fig. 5　Confusion matrix

图 5 中，颜色越亮代表该标签预测的准确率越

高,混淆矩阵的横轴表示预测结果,纵轴表示真实标记,可以看到在 O 标记上,本文模型出现的偏差比较明显,但依然保持在较高的水准,这是因为 O 标记总体数据样本占据的比例最大、也相对更为分散。另外,文本的模型在时间点、肯定词、否定词等关键实体的识别上也达到了较高的准确率,这对本文结构化过程中的按时间节点分段,按肯定词、否定词分句有较大影响。

随机选取一样例做实体识别,识别结果展示如图 6 所示。

图 6 实体识别结果展示图

Fig. 6 Entity recognition result display

通过实体识别后的数据就可以进行结构化处理,结构化结果中的一个样例如图 7 所示。

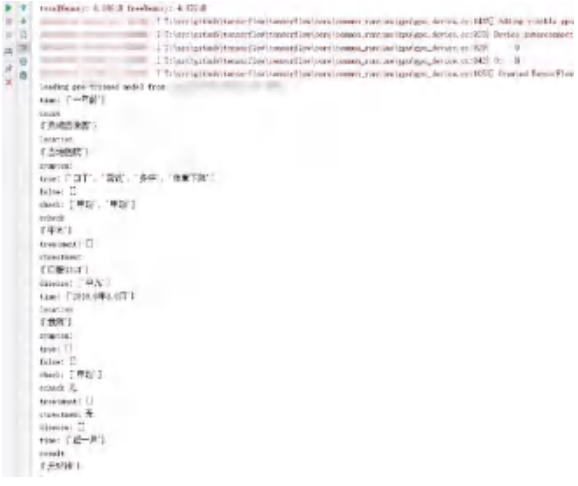

图 7 结构化结果展示图

Fig. 7 Structured results display

图 7 中,通过{}及[]不同的括号来区分不同方法得到的结构化信息,{}为实体识别的内容,[]为词库匹配的内容。

(4)结构化存储。将最终的结构化数据以资源描述框架的形式进行存储。结构化存储借助 python 第三方扩展(rdflib),以 XML 形式进行 RDF 序列化存储,最终对每个时间段内的内容都生成一个 XML 文件。由于文本限制,只截取一条完整病史数据的结构化结果的起始部分内容,序列化的一个样本如图 8 所示。

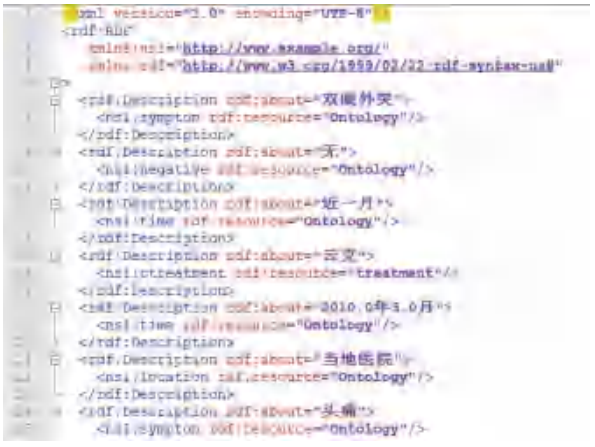

图 8 RDF 存储样本

Fig. 8 Storage sample of RDF

3 结束语

本文结合现有自然语言处理技术和甲状腺病史的数据特征,提出了一种甲状腺病史结构化处理方法。首先,构建专业词库和病史本体,分别用于指导分词和实现结构化输出;其次,对原始数据进行预处理,并将预处理后的数据进行实体识别,实现对分词结果的标签预测;最后,基于病史本体结构,使用标签抽取和词库匹配两种方法,实现对甲状腺病史的结构化,并通过 RDF 将结构化结果进行存储。

参考文献

[1] SUNDARARAJAN V, HENDERSON T, PERRY C, et al. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality[J]. Journal of Clinical Epidemiology, 2004, 57(12):1288-1294.

[2] DEYO R A. Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: A response[J]. Journal of Clinical Epidemiology, 1993, 46(10):1081-1082.

[3] 中华医学会内分泌学分会《中国甲状腺疾病诊治指南》编写组. 中国甲状腺疾病诊治指南[J]. 中华内科杂志, 2007, 47(10):867-868.

[4] MAEDCHE A. Ontology learning for the semantic Web[M]// Ontology learning for the semantic Web. Boston, MA: Springer, 2002:117-147.

[5] 杜文华. 本体构建方法比较研究[J]. 情报杂志, 2005(10):24-25.

[6] GIBBINS N. Resource description framework[J]. Serials Review, 2009, 27(1):58-61.

[7] QIN Ying, ZENG Yingfei. Research of clinical named entity recognition based on Bi-LSTM-CRF[J]. Journal of Shanghai Jiaotong University, 2018, 23(3):392-397.

(下转第 32 页)