面向精准医疗伦理的信息安全标准体系构建方法研究

邢玉艳 刘 耀
（中国科学技术信息研究所，北京 100038）

摘要：以精准医疗伦理领域为例，提出一种基于数据驱动的标准体系自动构建方法。其中包括构建标准体系模型、自动抽取标准体系概念与关系、生成标准体系结构。通过选取精准医疗中的个人隐私领域进行实验，分别形成个人隐私领域中单个标准和单类标准，并且正确率达到86.99%，召回率达到65.53%。实验证明，基于数据驱动的标准体系自动构建方法的有效性和有用性。

关键词：精准医疗伦理；模型构建；自动抽取；结构生成；个人隐私
中图分类号：G350 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2019.06.014

Research on the Construction Method of Information Security Standard System for Precise Medical Ethics

XING Yuyan, LIU Yao
(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: Taking the field of precision medical ethics as an example, this paper proposes an automatic construction method based on data-driven standard system. This system includes the construction of the standard system model, the automatic extraction of the standard system concept and relationship, and the generation of the standard architecture. Finally, the experiments are conducted by selecting the personal privacy field in precision medicine, and the individual standards and single-class standards in the personal privacy field are generated separately. The correct rate reached 86.99% and the recall rate reached 65.53%, which proved the validity and usefulness of the method.

Keywords: precise medical ethics, model construction, automatic extraction, structure generation, personal privacy

0 引言

精准医疗作为全新的医学模式，可能会带来疾病诊断、诊疗和健康保健方面的革命而造福人类。精准医疗的发展给人类带来福利的同时，也会带来突出的医学伦理问题。而出现任何伦理的问题，都会对个人和社会带来巨大损失，同时也会阻碍精准医学的发展，因此要清醒地认识并积极应对[1]，为其制定相应的规范和标准体系。

标准体系是在一定范围内的标准按其内在联系形成的科学有机整体，是编制标准、修订计划的依据。标准体系包含了宏观标准体系和微观

作者简介：邢玉艳（1991—），女，中国科学技术信息研究所硕士研究生，主要研究方向：知识工程与知识发现；**刘耀**（1972—），男，中国科学技术信息研究所研究员，博士，主要研究方向：知识工程与知识发现（通信作者）。

基金项目：国重点研发计划项目“精准医疗伦理、政策法规框架研究”之子课题“构建安全、可靠的面向生物医学大数据的、跨系统样本和数据共享的保障体系”（2017YFC0910101）。
收稿日期：2019 年 9 月 23 日。

标准体系两种，其中宏观标准体系是指某领域所有标准构建的体系结构，微观标准体系是指某个标准的体系结构[2]。根据《标准体系构建原则和要求》[3]，目前通用的标准体系构建方法是确定目标、调查研究、分析整理、编制体系表、动态维护更新等部分。无论是宏观标准体系还是微观标准体系，若要进行标准体系构建，标准工作者就需要依据规范的方法进行大量的资料整理与搜集，从海量资源中提炼出大量的概念、关系、结构，耗费了大量的人力、物力，但是也难以找全标准体系中包含的各方面内容，其广度和深度都难以达到理想状态。为解决这一问题，本研究提出了一种基于数据驱动的标准体系构建方法，利用概念自动获取、关系自动抽取、结构表示等技术，实现标准体系的自动构建。

1　信息安全标准体系模型构建

标准体系模型是标准体系构建的基础，同时也需要一定的理论支撑。在标准化领域，经常运用的是霍尔三维模型。该模型是美国系统工程专家A.D.HALL[4]于1969年提出的一种系统工程方法论。霍尔三维模型是将系统工程整个活动过程分为前后紧密衔接的7个阶段和7个步骤，同时还考虑了为完成这些阶段和步骤所需要的各种专业知识和技能，形成由时间维、逻辑维和知识维所组成的三维空间结构。

本研究将精准医疗伦理的标准体系模型的构建分成5个阶段，包括精准医疗领域概念获取、医学伦理领域概念获取、信息安全领域概念获取、三个领域概念关系获取、领域知识获取。将这5个阶段分成了3个维度，分别是概念维、关系维和知识维。标准体系模型如图1所示。其中，精准医疗领域概念获取采用《2018−2023年中国精准医疗行业深度分析及发展前景预测报告》作为模型构建的依据；医学伦理领域概念获取是借鉴大学医学专业教材《医学伦理学》第五版；信息安全领域概念获取是借鉴全国信息安全标准委员会发布的290个标准；领域知识获取是根据所构建的检索式，抽取同时出现上述概念和关系的句子与段落。模型部分展示如图2所示。

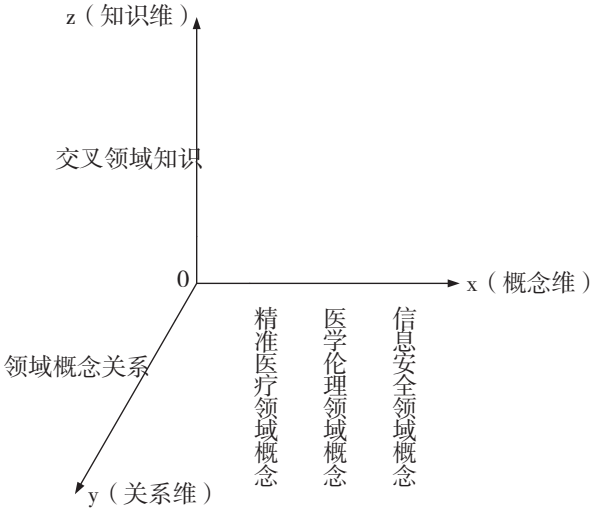

图1　标准体系模型图

2　信息安全标准体系构建方法

2.1　概念获取

概念词自动获取的方法有多种，其中包括基于规则的方法、基于机器学习的方法和基于深度学习的方法。基于规则的方法需要大量的人工，目前已经很少使用。基于机器学习和基于深度学习的方法是目前比较受欢迎的。其中，Zheng在命名体识别任务中使用CRF模型，选取的特征有词性、词语的TFIDF值，准确率达79.63%，召回率达73.54%，但是需要选择出对任务有帮助的特征，并将其转化成被机器学习的特征向量[4]。Collobert[5]提出了采用神经网络搭建概念获取模型，将词向量输入到CNN+CRF模型中，在CONLL2003数据集上取得了89.59的F值。在相同的数据集上，Huang[6]等提出LSTM+CRF模型，得出85.19的F值。总体来说，基于深度学习的方法，输入词向量就可以达到很好的效果，本文也将采用该方法进行后续实验。

标准体系概念自动获取是信息安全标准体系构建的关键，因此首先要对标准体系中所需要的关键词进行分析并进行人工分类。

2.1.1　标准体系概念词分析

标准体系一般包括标准体系框架和标准明细表，在这里所提到的标准体系是指标准体系框架

图 2　模型部分展示图

结构。标准体系框架结构是某个领域内的所有标准按照一定的层级结构划分的有机整体，在这个整体结构中，涉及内容范围广泛，每个分支都代表一个方面，每个方面又会细分很多小的不同的方面，这些小的不同的方面是由领域的概念词所组成，一般情况下为名词或者是名词性词组。

标准体系框架中概念词可能会来源于该领域已有的标准文本，也可能来源于研究性论文、政策文本等，这取决于要制定的标准体系的类型，如果是修改和完善之前的标准体系，那么体系中的结构点就会来源于已有标准，如果是新增性的标准体系，那么其来源相对来说就会比较广泛，可能是相关领域的标准、国家政策文本、研究性论文等。本文所研究的精准医疗伦理的信息安全标准体系，就属于新增性，在构建体系的过程中，就会搜集大量的相关领域的文本。本文所涉及的领域是精准医疗领域、信息安全领域以及医学伦理领域。

2.1.2　标准体系概念词获取

BiLSTM−CRF的命名实体识别模型作为一个序列标注模型，主要由Embedding层（主要有词向量、字向量）、双向LSTM层以及CRF层构成[7]。输入序列输入X后，通过向量表将每个字符映射成相应的向量，将其作为初始向量输入到神经网络模型中；双向LSTM层采用softmax函数得到概率分布矩阵；最后通过CRF层模型确定一个概率最高的序列路径，对应到每个字符作为最后标签。其整体结构图如图 3 所示。

2.2　关系自动获取

获取到标准体系概念后，下一步就要识别概念之间的关系，也就是实体关系抽取。实体关系抽取的主要任务是从句子中自动抽取概念之间的关系，这也是知识结构化的重要任务之一。概念关系的抽取主要包括基于规则的、有监督、弱监督、无监督的方法。Leek等[8]首次在关系抽取中使用HMM，完成了从生物学的文献中抽取出基因名字和其对应位置信息的任务；Ray等[9]结合句子的短语结构分析信息利用HMM做信息抽取，取得了较好的效果。实验证明，HMM在关系抽取任务上有一定的有效性，与其他方法相比也有一定的优越性。但是，也存在HMM结构确定困难等问题。董静等[10]结合中文语料库的特点，将中文实体关系划分为包含实体关系和非包含实体关系，分别利用不同的句法特征，而其他词汇等特征完全相同，在CRF模型框架下，以ACE 2007语料作为实验数据，取得较好的抽取结果。

支持向量机是Cortes和Vapnik于 1995 年首先提出的，它是建立在统计学习理论（SLT）基础之上的一种新型的机器学习算法，根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷，以期获得最好的推广能力。目前，使用最多的是基于有监督的方法，将实体关系抽取任务转化成分类问题，因此，一般的分类方法都可以用到实体抽取任务上。常见的分类算法有：

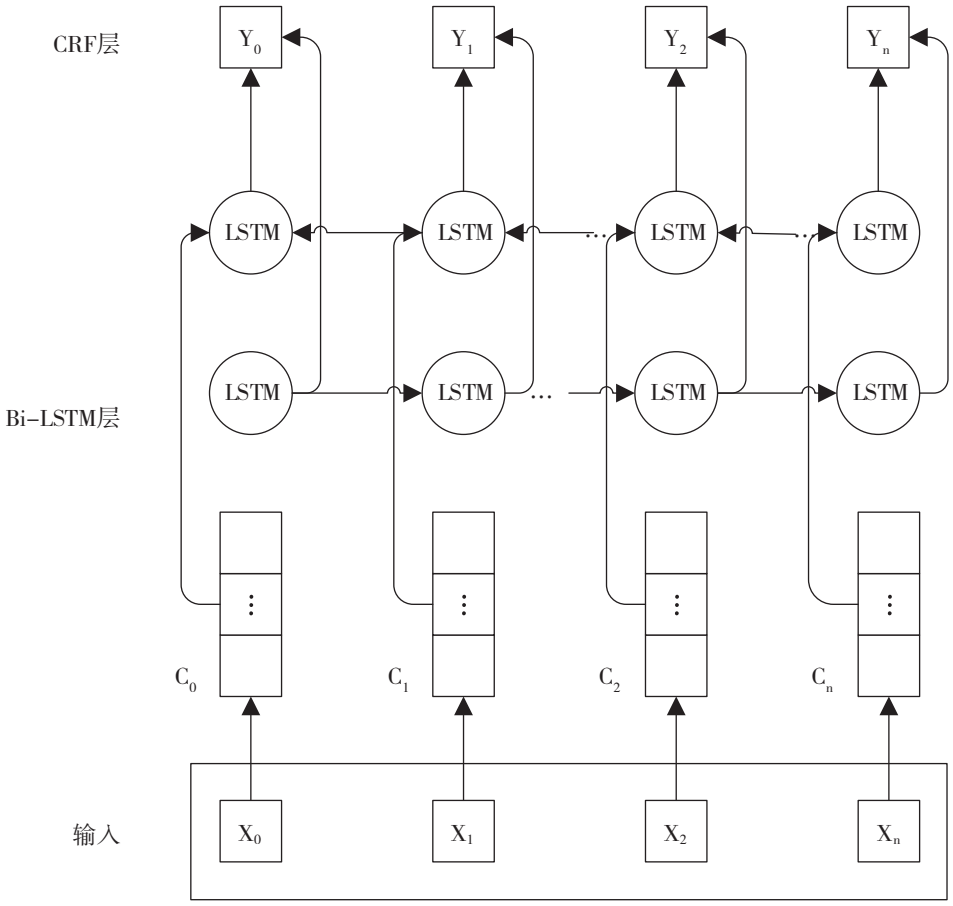

图 3　Bi-LSTM-CRF 模型

SVM、KNN、朴素贝叶斯、决策树等。SVM分类器理论框架完善、通用性和鲁棒性强、计算简单，而且还具有较强的抗噪声能力和较高的分类正确率[11]。SVM分类算法不需要无穷大样本数量，也不局限于解决线性问题，也可以通过核函数处理非线性问题，因此本研究将采用SVM算法进行实体关系的抽取。

本文将信息安全标准体系构建中的关系分为5类，即推进关系、融合关系、阻碍关系、包含关系、因果关系，并根据设定的概念关系进行关系特征的选取，经过分析本文用到的概念特征有概念类别、概念相邻词、概念词间的词性标注、概念词的上下文词、句法依存分析。

2.3　结构生成

标准体系结构的生成主要包括标准体系节点的表示和标准体系的结构表示。标准体系节点表示是对所选目标文本中的标题进行向量化，标准体系的结构表示是对所选文本中挑选出的标题下的文本进行向量表示。

2.3.1　标准体系节点表示

目前已有的网络表示学习算法[12]各有优劣。本文将采用2018年在第二十五届国际人工智能联合会议上，Pan[13]等提出的TriDNR模型。该模型提出一种新的用于深度网络表示学习的神经网络模型，加强了网络结构层次、节点内容层次、节点标签层次。TriDNR模型如图4所示。从图4可以看出，该模型分为两层，节点关系建模和节点标签与上下文建模，其中上层的节点关系建模中采用的Deepwalk算法[14]。该算法随机游走均匀地选取网络节点（词语），同时生成固定长度的随机游走序列，该序列可以看成是句子，最后应用Skip-Gram模型预测上下文节点，并将该层的随机排序传入到下一层中。

2.3.2　标准体系结构表示

本文的层级分为两种，一种是与概念直接相连的层级，另一种是上层的标题层级，对于前者将概念作为词，概念连接随机游走路径作为句子，利用Doc2vec算法计算该层级向量，后者则采用同一层级取平均的方法。Doc2vec是Le[15]在Word2vec的基础上提出的一种将文本表示成向量的方法，通过分布式学习的方法，对不同长度的文本片段进行采样，获取固定长度的特征表示。Doc2vec属于无监督算法，其优点就是可以较好地处理没有太多标记数据的任务。Doc2vec算法的模型如图 5 所示，将文本中的段落映射到向量空间中，用D的一列进行表示，与此同时，将每个词要映射到向量空间中，用矩阵W来表示，然后将前面得到的段落向量和词向量相加，作为下一个词的输入。

3　实验结果与分析

3.1　实验数据

本实验以精准医疗伦理中的“个人隐私安全”领域进行标准体系的生成，用于结构生成的语料分为两大部分：一是某一个具体标准下的规范性引用文件和参考文献；二是检索到与个人隐私安全相关的标准、政策、法规。对这两部分的语料经过人工去重后进行本研究后续的实验。

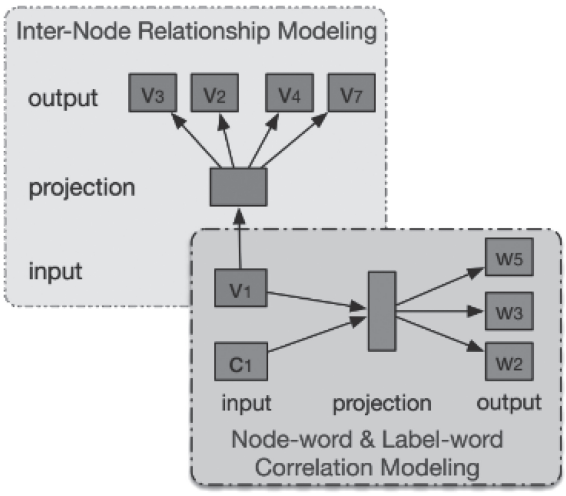

图 4　TriDNR模型

3.2　实验流程

信息安全标准体系中包含多个领域、多个方面，每个方面又由多个标准所组成的。为了验证本文方法的有效性，在体系中挑选目前热门且急需解决的个人隐私安全领域生成单个标准结构。通常标准有四级标题、五级标题甚至六级标题，本文旨在以说明方法为目的，所生成的标准结构到三级标题。以下是具体的生成步骤。

（1）选取参照标准。为了验证本文的方法，在个人隐私安全领域中选取目前已经发布的标准作为参照标准，用新生成的标准结构和参照标准结构进行对比。

（2）收集资源。根据选取的参照标准，找到对应的规范性引用文件和参考文献列表，对列表中的资源进行检索，获取能够下载的资源，同时在限定领域中检索其他类似标准并进行下载。

（3）资源预处理。将收集到的资源进行预处理，处理成需要用到的格式和需要保留的文本，将不同类型的资源进行统一，最终得到json格式的文本。

（4）句子向量表示。利用Doc2vec算法计算所选文本的句子向量，其中用概念节点表示向量作为该算法的预训练向量。

（5）标题向量表示。其中三级标题的向量是三级标题下所对应的句子向量，二级标题、一级标题、题目节点向量分别是下一级标题的平均值。

（6）排序筛选。分层次利用Textrank算法进行排序，选择新结构中需要加入的节点。

（7）生成新标准结构。将筛选出的章节节点按照层次等级进行整合，最终得到新标准结构。

3.3　结果分析

3.3.1　单个标准生成结果

本次实验生成了 3 个标准的结构，其中包括“个人信息安全规范”“健康医疗信息安全指南”“个人信息去标识化指南”。由于篇幅原因，在这里给出其中一个标准的具体结构。“健康医疗信息安全指南”生成的新结构及对比如表 1 所示。从新生成的结构中可以看出，生成的一级标

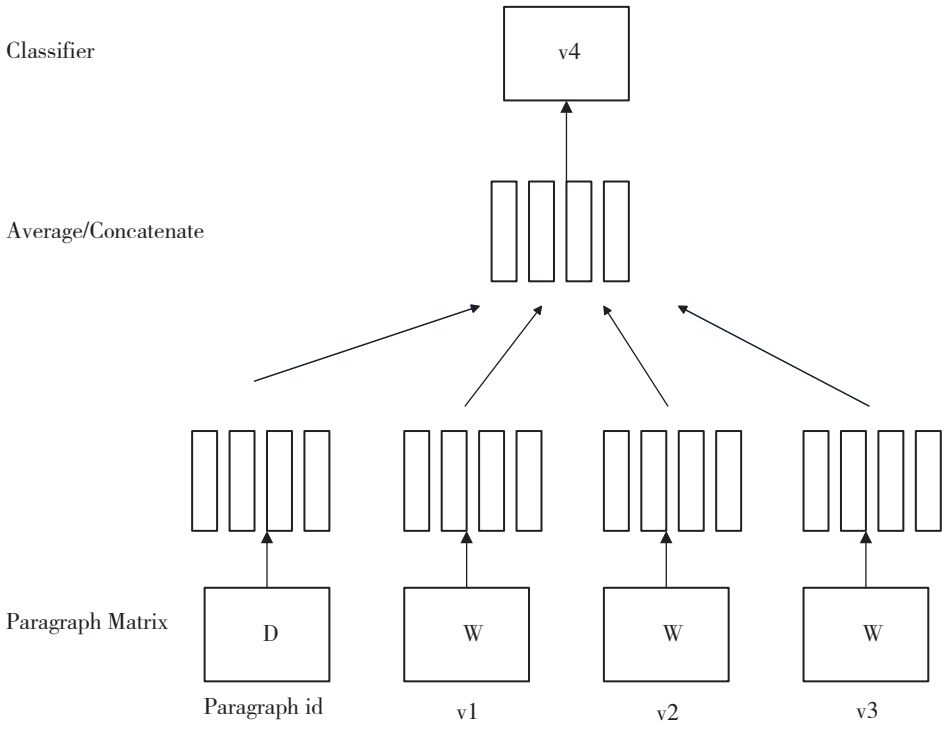

图 5　Doc2vec 模型图

题：健康医疗大数据、个人信息的使用、个人信息安全事件处置、去标识化概述、法律责任，基本都是与健康医疗信息安全相关的内容，可以为相关研究者提供一定的支持。而一些二级标题的名称与原结构中标题并不完全匹配，但是所要展现的内容则是更加细节的，比如，安全框架中的实施方法中就包含了新结构中的去标识化，数据使用环境中就包含新结构中的个人信息查询、更正、删除等操作，这就需要研究者根据实际需求进行筛选。

在对结构中的一级标题和二级标题进行比对的同时，计算结构的正确率（Precision）与召回率（Recall）。正确率是正确标题数目与生成的标题总数目比值，召回率是生成的标题中含有原结构标题的数目与原结构中所有标题总数的比值。正确率与召回率的算法是将一级标题和二级标题同时进行统计的，具体结果如表 2 所示。

通过实验可以看出，生成的这 3 个标准的平均正确率达到 86.99%，召回率达到 65.53%。这就可以证明本文方法是具有有效性的，可以为标准体系构建者提供相应的帮助。在生成标准体系或单个标准的过程中，可以首先使用该方法进行自动构建，大致得出一个标准体系或者一个标准应当包含的子体系或者章节，然后依据系统提供的体系或者章节点进行修改，这样避免了标准工作者在前期工作中进行大量的重复工作，大大提高了标准工作者的工作效率。

3.3.2　单类标准生成结果

单个标准的生成证明了本文方法的有效性，但是要证明本文方法的有用性，需要生成固定的某一类标准。其中，某一类标准的生成是指类似内容的标准生成。本节以生成个人隐私安全领域中“个人信息安全规范”结构进行实验结果的展示。

在原标准中，主要将个人信息安全分为七部分，也就是一级标题，个人信息安全基本原则，个人信息的收集，个人信息的保存，个人信息的使用，个人信息的委托处理、共享、转让、公开披露，个人信息安全事件处置，组织的管理要求。按照本文的方法，利用个人信息安全相关的资源，得到新的一级标题如表 3 所示。

从表 3 新生成的一级标题中可以看出，新生

表 1 “健康医疗信息安全指南”结构对比

原结构	新结构
1 安全框架	1 健康医疗大数据
2 安全目标	2 健康医疗
2 使用或披露的原则	2 互联网医疗健康
2 实施方法	2 健康数据
2 安全措施集	2 医疗数据
1 数据使用场景分析	2 大数据标准管理
2 概述	1 个人信息的使用
2 用户类别	2 个人信息使用原则
2 数据类别	2 个人信息使用环境
2 数据使用目的	2 个人信息使用目的
2 数据使用环境	2 个人信息查询
2 场景基本分类与安全措施要点	2 个人信息更正
1 典型场景数据安全	2 个人信息删除
2 互联互通数据安全	2 个人信息主体撤回授权同意
2 远程医疗数据安全	2 个人信息主体注销账户
2 汇聚中心数据安全	1 个人信息安全事件处置
2 健康传感数据安全	2 个人信息安全事件措施
2 移动应用数据安全	2 安全事件告知
2 临床研究数据安全	1 去标识化概述
2 商保对接数据安全	2 去标识化目标
2 器械维护数据安全	2 去标识化原则
	2 重标识风险
	2 去标识化影响
	1 法律责任
	2 互联网信息服务提供者
	2 电信业务经营者
	2 个人信息保护

表 2 标准结构统计表

标准名称	正确率/%	召回率/%	F1
健康医疗信息安全指南	90.00	63.33	74.35
个人信息保护指南	87.65	75.00	80.83
个人信息去标识化指南	83.33	58.25	71.60
平均	86.99	65.53	75.59

成的一级标题中包含了原标准结构中该有的个人信息处理流程。将新生成的一级标题进行排序，选择与原一级标题重合的标题，生成二级标题，也就是某主题下又包含哪些子主题。从排序结果来看，个人信息采集、个人信息存储排序靠前，并且原结构有类似表达，可以为研究者提供一定的帮助。下面继续用本文方法生成这两个标题下的二级标题。二级标题结果如表 4 所示。

从表 4 中可以看出，新生成的二级标题可以覆盖原二级标题的一部分，同时又丰富了主题下的子主题，使结构更加全面。通过上面单类标准的实验，证明了本研究方法的有用性，也就是说在以后要生成某个相关标准或者相关标准体系，系统可以自动为研究者提供应当包含的部分。比如要生成某一领域下的术语标准，研究者只需设定资源的条件和范围，利用本文提出的方法即可得出该标准应当包含的章节，为标准制定者提供参考，然后再根据需求进行修改。

4 总结与展望

本文对标准体系自动构建的方法进行了详细介绍，其中包括标准体系模型的构建，该模型

表 3　一级标题对比表

原一级标题	新一级标题
个人信息安全基本原则	个人信息采集
个人信息的收集	个人信息存储
个人信息的保存	隐私政策
个人信息的使用	法律责任
个人信息的委托处理、共享、转让、公开披露	个人信息使用
个人信息安全事件处置	个人信息去标识
组织的管理要求	个人信息脱敏
	个人信息销毁
	个人信息共享
	个人信息加密
	信息鉴别
	信息匿名化
	告知同意

表 4　二级标题对比表

一级标题	原二级标题	新二级标题
个人信息采集	收集个人信息的合法性 收集个人信息的最小必要 不强迫接受多项业务功能 收集个人信息时的授权同意 隐私政策 征得授权同意的例外	采集范围 知情同意 最小化原则 个人信息收集日志 个人信息风险评估
个人信息存储	个人信息保存时间最小化 去标识化处理 个人敏感信息的传输和存储 个人信息控制者停止运营	个人信息加密存储 个人信息匿名化 个人信息去标志化 个人信息备份 个人信息备份

是整个模型构建中的指导；概念、关系抽取过程中，分别采用BI-LSTM-CRF模型和支持向量机，选取句法语义特征进行实验，取得了良好的效果；标准体系结构生成过程中，采用TriDNR模型和Doc2vec模型进行实验，取得了良好的效果。最后选取个人隐私领域生成标准体系，分别形成单个标准和单类标准，最终得到结果的正确率达到86.99%，召回率达到65.53%。并且单个标准的实验采用回溯方法，与已发布的标准进行比对，验证了本文方法的有效性，单类标准的实验通过生成某一类的标准，验证了本文方法的有用性。利用本文方法生成的标准体系可以为相关研究人员在制定标准体系之前提供一个可以参考的框架与结构，缩短了研究人员大量收集相关材料的时间，大大地提高了工作效率。

在未来工作中，标准体系制定者若想制定新领域的标准体系或者标准，或者对已知标准体系进行更新，可以运用本文提出的方法限定资源后，进行生成或者筛查，这样大大提高了标准制定者的工作效率，进一步推动了标准化工作的智能化。当然，本文还有不足之处，下一步将会进一步扩大语料范围进行机器学习，并利用已有的知识库辅助概念与关系的标引，同时将生成的标准体系进行可视化展示。

参考文献

[1] POONEH S,郑君. 精准医疗伦理问题综述[J]. 中国医学伦理学,2018,31(9):1221-1223.

[2] 麦绿波. 标准体系的内涵和价值特性[J]. 国防技术基础,2010,5(12):3-7.

[3] 标准体系表编制原则和要求: GB/T 13016-2018 [S].

北京：中国标准出版社，2018.

[4] HALL A D. Three−dimensional morphology of systems engineering[C]. IEEE Transactionson Systems Science and Cybernetics. Dordrecht:Springer,1974:156−160.

[5] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12(8): 2493−2537.

[6] HUANG Z, XU W, YU K. Bidirectional LSTM−CRF models for sequence tagging[J]. arXiv,2015,6(5):1508.

[7] McNamee P, Mayfield J. Entity extraction without language−specific resources[C]// Proceedings of the 6th conference on natural language learning−Volume 20. Association for Computational Linguistics, 2002: 1−4.

[8] LEEK T R. Information extraction using hidden markov models[C]. Advances in neural information processing systems. Virginia:MUC,1998:140−144.

[9] RAY S,CRAVEN M. Representing sentence structure in hidden markov models for information extraction[C]// Proceedings of the 17th International Joint Conference on Artificial Intelligence. San Francisco:Morgan Kaufmann Publishers Inc,2001:1273−1279.

[10] 董静，孙乐，冯元勇，等. 中文实体关系抽取中的特征选择研究[J]. 中文信息学报，2007,21(4):80−85.

[11] 刘绍毓. 实体关系抽取关键技术研究[D]. 郑州：解放军信息工程大学，2015:23−25.

[12] 涂存超，杨成，刘知远，等. 网络表示学习综述[J]. 中国科学：信息科学，2017(8):32−48.

[13] PAN S, Wu J, Zhu X, et al. Tri−party deep network representation[J]. Network, 2016, 11(9): 12.

[14] PEROZZI B, Al−RFOU R, SKIENA S. Deepwalk: Online learning of social representations[C]// Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2014: 701−710.

[15] LE Q, MIKOLOV T. Distributed representations of sentences and documents[C]. International conference on machine learning. 2014: 1188−1196.