中文病理文本的结构化处理方法研究

陈德华 冯洁莹 乐嘉锦 潘 乔
(东华大学计算机科学与技术学院 上海 200051)

摘 要 病理文本作为一类重要的非结构化临床文档,对临床诊断至关重要。针对具体的中文病理文本数据,提出一种简单有效结构化处理方法。首先对中文病理历史文本数据进行预处理,包括数据清洗、短句切分及主干提取等步骤,从中提取出各个样本所对应的文本信息;然后通过短句聚类和统计参数筛选实现样本描述模板的提取;最后利用模板对病理文本进行即时结构化处理,得到最终的结构化处理结果。实验证明,该方法对同类文本可以达到很好的结构化效果;同时提取的模板会被定期优化以适应最新的数据结构化需求。

关键词 中文病理文本,结构化,短句聚类,模板提取
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.10.051

Research on Structured Method for Chinese Pathological Text

CHEN De-hua FENG Jie-ying LE Jia-jin PAN Piao
(School of Computer Science and Technology, Donghua University, Shanghai 200051, China)

Abstract Pathological text as an important kind of unstructured clinical documents, is essential to clinical diagnosis. For the specific Chinese pathological text, this paper put forward a simple and effective structured approach. Firstly the Chinese pathological texts are preprocessed, including data cleaning, clauses split and trunk extraction, in order to extract the corresponding information of each sample. Then each sample's final template information is extracted by the way of clauses clustering and statistical parameters filtering. Finally, the templates are used for immediate pathological text structuring process, and the structured results are obtained. Experiments show that the proposed method can achieve satisfactory structured results for similar pathological texts, and the extracted templates will be regularly optimized to meet the needs of the latest text structuring.

Keywords Chinese pathological text, Structuring, Clauses clustering, Template extraction

1 引言

随着国内各大医院信息化建设进程的不断发展,电子病历、医护工作站、实验室信息系统、医学影像传输和存储系统、放射科信息系统、手术麻醉管理系统、病理管理系统、心电图生理管理系统等临床信息系统已被引入到医院中。这些临床信息系统的引入已使医院中积累了丰富的临床数据资源。病理报告作为重要的非结构化临床文档资源,是临床医生对患者疾病状况进行判断的重要依据。

目前,如何从非结构化的病理文本数据中获取有用的信息成为当前的研究热点之一。已有的医疗文本数据结构化处理分为两类:1)前结构化处理,即数据按照直接结构化输入,主要方式是设计规范的病历系统[1-4],这种方式保证了数据的可靠性和完整性;2)后结构化处理,即对已有的医疗文本数据利用自然语言处理技术进行结构化处理。

近年来,在医疗文本数据的结构化处理方面,国内外均有许多相关的研究工作。国外的工作主要集中在英文的医疗文本数据处理。例如,Apostolova E[5]构建了一个英文病理报告的语义切分系统,实现了病理报告的自动语义切分;Maria Skeppstedt 等[6]采用 CRF 模型从瑞典临床文档中识别出临床实体类型;Guergana K Savova 等研发了基于 openNLP 的 cTAKES 系统[7],该系统也可实现临床自由文本的信息提取;Carol Friedman 等设计并实现了 MedLEE 系统[8],自动将医疗文档编码成带修饰符的描述方式,产生结构化的输出。这些研究结合语言自身的特性和医疗文本的书写特性,充分利用已有的临床文档体系结构、医学术语标准和自然语言处理技术,取得了不错的成果。

国内对非结构化医疗文本处理的研究起步较晚。大多数的研究工作是在借鉴国外已有技术的基础上,对中文医疗文本的结构化处理开展相关的研究。例如,孔晓风等[9]利用中科院分词系统 ICTCLAS 和 UMLS 语义网络结构,对消化内科内窥镜检查报告进行结构化处理。但由于中文语言自身的

到稿日期:2015-09-29 返修日期:2016-01-07 本文受上海市科委科技创新行动计划:基于"互联网+"技术的多病种多中心临床大数据行业应用(15511106900)资助。

陈德华(1976-),博士,副教授,主要研究方向为数据库、数据仓库与智慧医疗,E-mail:chendehua@dhu.edu.cn;冯洁莹(1991-),硕士生,主要研究方向为数据库、大数据与智慧医疗;乐嘉锦(1951-),教授,博士生导师,主要研究方向为数据库与数据仓库、大数据;潘 乔(1977-),副教授,硕士生导师,主要研究方向为大数据和网络安全。

特殊性，借鉴国外技术受到很多的制约。一方面，中文的病理报告结构化处理需要分词，但现有的中文分词工具（如中科院的 NLPIR[10]、复旦的 FNLP[11]、斯坦福的 NLTK[12] 等）在处理专业性强的中文病理报告上并不能取得令人满意的分词结果；另一方面，与英文不同，中文没有所谓的词根/前缀，不能直接根据词本身进行分类，需要语义等来辅助判别。此外，受国内长期手写病历习惯的影响，医生在病理文本的语言组织上相对较随意，需要进行规范化操作才能较好地套用现有的医学规范。

针对上述问题，本文结合中文病理文本数据的具体特点，提出一种简单有效的中文病理文本结构化处理方法。该方法首先对中文病理历史文本数据进行数据预处理操作，包括数据清洗、短句切分、主干信息提取，从原始数据中提取出各个样本所对应的文本信息；然后通过短句聚类和统计参数 IDF 值、C-value 值[13] 的综合筛选实现样本描述模板的提取；最后利用模板对病理文本进行即时结构化处理，得到最终的结构化处理结果。

本文第 2 节给出了相关定义和问题阐述；第 3 节详细阐述了病理文本结构化处理方法的基本思路和具体处理过程；第 4 节给出了实验结果及分析；最后总结全文。

2 问题阐述

病理文本数据是医生用自然语言书写的文本数据。一般在病理报告中，医生根据病理切片分别会从肉眼所见和镜下所见对患者的疾病病理进行描述，肉眼所见的数据样例如表 1 所列。

表 1 病理报告的肉眼所见描述样例

病理号	肉眼所见
2013-00141	降结肠肠段切除标本，长 8cm，周径 3cm，距一侧切端 3cm，另一切端 1.5cm 处见一隆起型肿块，大小 3×2×0.8cm，绕肠管 1/2 周，切面灰白质硬，找到肠旁淋巴结 10 枚，直径 0.5～0.8cm。

在语义上，病理文本是对样本及其指标的描述，由若干句独立的短句组成。其中，每个短句一般会包含一个样本。每个样本又会有若干个指标。样本及指标的具体定义如下。

定义 1 样本名指人体组织、器官、身体部位等名词。

定义 2 指标名指带出某一样本的大小、形状、颜色、质地、位置等信息的名词。

定义 3 指标值指与指标名一一对应的描述性信息。

以表 1 中的数据为例，该病理报告中包含“降结肠”和“淋巴结”两个样本名。而降结肠样本则包含“长”、“周径”、“距一侧切端”、“另一切端”等多个指标名。其中，指标名“长”对应的指标值为“8cm”，周径为“3cm”。可以明显看出，短句中存在如图 1 所示的短句三层结构。

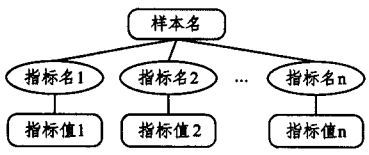

图 1 短句三层结构

针对上述的中文病理文本数据，本文的病理文本结构化处理方法利用自然语言处理、数据挖掘等相关技术，实现病理文本的结构化处理，从非结构化的病理文本数据中提取出样本、指标及其指标值等，得到结构化的输出结果。

3 病理文本结构化处理方法

病理文本结构化处理涉及到自然语言处理的分词、正则表达式等，在处理的过程中还会用到数据挖掘中的聚类算法。本节将从基本思路和主要的处理阶段来详细阐述病理文本结构化处理方法。

3.1 基本思路

本文提出的中文病理文本结构化处理方法主要包括数据预处理、模板提取和模板应用等 3 个阶段，其处理流程如图 2 所示。

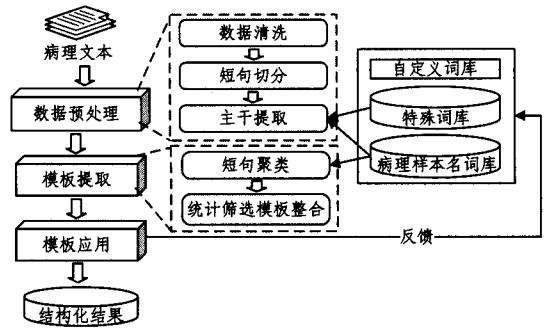

图 2 结构化处理流程

其中，数据预处理包括数据清洗、短句切分和主干提取 3 个步骤，经过此阶段的处理，原始的病理文本将转换为由样本名和指标名表示的短句集；第二阶段为模板提取，该阶段包括短句聚类和统计筛选两个步骤，经过此阶段的处理，每个样本都将对应维护一个模板文件；最后是模板应用阶段，即对新的病理文本，匹配其对应的模板并套用，产生结构化的输出。除上述 3 大阶段外，本方法还将维护两个自定义词库，分别为特殊词库和病理样本名词库 DIC，词库的具体内容将在下文详细介绍。

3.2 数据预处理

数据预处理阶段包括数据清洗、短句切分和主干提取 3 个步骤。

1）数据清洗

由于病理数据是医生在诊断过程中输入的，因此不规范输入无可避免。数据清洗的任务就是将数据规范化，提高数据利用率，主要包括去掉无关或冗余的标点符号、统一中英文标点符号和全角半角数字字母、补全缩写的中英文医学术语、纠正常见拼写错误等。为简单起见，可以采用字符串匹配替换的方法，本文采用的数据清洗方案的部分内容如表 2 所列。

表 2 数据清洗方案的部分内容

类别	匹配目标	替换内容
标点、数字、字母	1|2|3|4|5|6|7|8|9|0	1|2|3|4|5|6|7|8|9|0
	A|B|C|D|…|X|Y|Z	A|B|C|D|…|X|Y|Z
	a|b|c|d|…|x|y|z	a|b|c|d|…|x|y|z
	,|;|:|“|”|×|（|）等	,|;|:|“|”|*|（|）等
医学术语	LN|Ln|ln	淋巴结
	Ca|CA|ca	癌
	…	…
	左甲|右甲|左叶|右叶|左乳|右乳|左室|右室|乳癌…	左甲状腺|右甲状腺|左肺叶|右肺叶|左乳房|右乳房|左心室|右心室|乳腺癌…
拼写错误	距一切断|未及|成…	距一切端|未见|呈…

2)短句切分

为提取每个样本的模板信息，需要对经过清洗的数据按样本名进行短句切分。考虑到医生书写的病理文本遵循一定的书写规范，本文采用简单的正则匹配进行切分。

3)主干提取

本文提出的短句三层结构，每个短句一般会包含一个样本，针对这个样本，又会有若干个指标对其描述。主干提取的任务就是从经过分词处理的短句中提取出候选样本名和若干候选指标名。在中文的书写习惯中，主语一般位于句首，因此本文采用病理样本名词库匹配的方法，从短句的第一个文本片段中提取出候选样本名；再结合特殊词库以及数量词和停用词等，从每个文本片段中筛选出候选指标名。

此外，考虑到数据的特殊性，直接采用现有的分词工具不能取得令人满意的结果，因此在主干提取中采用词典和分词工具相结合的方法，即将自定义词库作为扩展词库加载到分词工具中，最大限度地减少分词错误。提取出来的候选样本名和候选指标名可能由一个或多个字词(由空格隔开)组成。

扩展词库的自定义词库由特殊词库和病理样本名词库DIC两部分组成，词库部分内容如表3所列。本实验整理了4个通用的特殊词库，包括颜色库、形状库、质地库和状态库；病理样本名词库DIC整理自网络，主要包括人体组织、器官、身体部位等名词。

表3 自定义词库部分内容

特殊词库	颜色库	灰白乳白色、乳白色、亮黄色、半透明、淡咖啡色、咖啡色、咖啡、土黄色、墨绿灰白、墨绿色…
	形状库	三角形、不等息肉样、不等隆起样、不规则型、不规则结节状、中央型、中央形、中央溃疡形、中央瘢痕样、中央细乳头状…
	质地库	质地偏硬、质稍偏硬、质韧偏硬、质偏硬、偏硬、质地偏韧、质嫩偏韧、质实偏韧、质软偏韧、质偏韧…
	状态库	已做冰冻、已全程剖开、已冰冻取材、已分离、已切开、已切碎、已切除、已剖开、已剖、已剥离…
病理样本名词库DIC		扁桃体、拇指、支气管、新生物、松果体、桡骨、横结肠、淋巴结、溃疡、滑膜…

3.3 模板提取

模板提取包括短句聚类和统计筛选两个步骤。在短句聚类部分，采用基于词典的文本聚类，并构造了一种文本相似度计算方法；在统计筛选部分，采用基于统计参数IDF值和C-value值的模板提取方法。

3.3.1 文本相似度计算

Jaccard相似系数是衡量两个集合相似度的一种指标，定义为两个集合A和B的交集元素在A，B的并集中所占的比例，用符号$J(A,B)$表示如下：

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|} \tag{1}$$

实际应用中，可将Jaccard相似系数用于衡量样本的相似度。

在本文的数据中，短句的样本名sn和指标名列表sn_list可能由一个或多个字词(由空格隔开)组成，病理样本名词库DIC中的词$word$可能是一个字或词。结合数据特点和Jaccard相似系数的计算思想，本文构造了一种样本名相似度和指标名相似度的计算方法。

样本名相似度计算方法如下：

$$SimilaritySN(sn,word)=\begin{cases}\frac{1}{C_{sn}} & ,若\ sn\ 包含\ word\\ 0 & ,若\ sn\ 不包含\ word\end{cases} \tag{2}$$

其中，C_{sn}是sn包含DIC中的词的数目。

对某一短句样本名sn的指标名列表sn_list和病理样本名词$word$的指标名列表$word_list$，指标名相似度计算如下：

$$SimilarityLIST(sn\_list,word\_list)=\frac{C_{overlap}}{C_{total}} \tag{3}$$

其中，$C_{overlap}$是两个列表中重复的指标名数目，C_{total}是两个列表中指标名去重后的总数。

3.3.2 基于词典的文本聚类

结合上述文本相似度计算方法，本文提出了一种基于词典的文本聚类。

算法1 基于词典的文本聚类

输入：短句集CLAUSES；
　　病理样本名词库DIC；
输出：聚类结果文件ClusterResult；
　　中间文件Matched和Unmatched；

```
1. CLAUSES=preliCluster(CLAUSES);
2. Array.init(CLAUSES,DIC);
3. for each dic in DIC
4.     for each c in CLAUSES
5.         if(Array(c.sn,dic.word)==1){
6.             Update(Matched,ClusterResult);
7.             CLAUSES.remove(c);
8.         }
9. for each c in CLAUSES{
10.    dic'=findBest();
11.    Update(Matched,ClusterResult);
12.    CLAUSES.remove(c);
13.}
14. for each c in CLAUSES
15.    Update(Unmatched,ClusterResult);
16. return ClusterResult,Matched,Unmatched.
```

上述算法中，第1步是对*CLAUSES*进行初步分类，即将完全相同的候选样本名分为同一类；第2步是初始化样本名相似度数组，即两两计算每个短句的候选样本名和病理样本名词库DIC中的词条的样本名相似度；第3－8步优先处理样本名相似度为1的短句，先在*Matched*中更新当前*dic*词条对应的指标名列表信息，并更新*ClusterResult*，再将当前已经处理完的*c*从*CLAUSES*集合中删除；第9－13步处理样本名相似度值在区间(0,1)中的短句，在DIC中找到与当前*c*的样本名相似度值最大且不为0的词*dic'*，然后在*Matched*中更新当前词条*dic'*对应的指标名列表信息，更新*ClusterResult*，并将当前*c*从*CLAUSES*集合中删除；第14,15步对余下的短句直接采用初步分类的结果，在*Unmatched*中更新当前每个词条对应的指标名列表信息，更新*ClusterResult*；最后返回聚类结果文件*ClusterResult*和中间文件*Matched*、*Unmatched*。

3.3.3 统计参数计算

本文在模板提取中需要用到两个统计参数：IDF和C-value。IDF即信息检索中最常用到的“逆文本频率指数”，

C-value最早由 Katerina 和 Sophia[14] 提出用于解决嵌套术语的抽取问题。同时，指标名的统计参数计算范围应该是它所对应的短句集的子类，而不是整个数据集。

计算指标名 w 的 IDF 值的公式[15]如下：

$$IDF(w)=\log(\frac{D}{D_w}) \tag{4}$$

其中，D 为子类中的全部短句数，D_w 为子类中有 w 出现的短句数。

计算指标名 w 的 $C\text{-}value$ 值的计算公式[13]如下：

$$C\text{-}value(w)=\begin{cases}\log_2|w|\cdot f(w), & w\text{不被嵌套在其他指标名中}\\ \log_2|w|\cdot(f(w)-\frac{1}{P(T_w)}\sum_{b\in T_w}f(b)), & w\text{被嵌套在其他指标名中}\end{cases} \tag{5}$$

其中，$|w|$ 为 w 中的字数，$f(w)$ 为 w 在子类中出现的次数，T_w 为子类中包含指标名 w 的所有候选指标名的集合，$P(T_w)$ 为集合中候选指标名的数目。

3.3.4 基于统计参数计算的模板提取

基于上述统计参数计算，本文提出的模板提取过程描述如下。

算法 2 基于统计参数计算的模板提取

输入：中间文件 Matched 或 Unmatched；

输出：模板文件；

1. 对中间文件 Matched 或 Unmatched 中的每个样本 r：
2. 找到 r 对应的短句子集；
3. 对 r 中的每个指标名：
4. 计算其在对应短句子集中的 TDF 和 C-value 值；
5. 根据阈值筛选指标名，将其整合入对应的模板；
6. 返回最终的模板文件.

因指标名的统计参数计算范围是它所对应的短句集的子类，而不是整个数据集，所以需要经过第 2 步的处理，找到当前样本 r 在短句集中对应的短句子集。经过上述过程的处理，可以对每个样本得到对应的模板信息。

3.4 模板应用

对新的一句病理文本，先将其切分为短句，再用短句的第一个文本片段和模板库中的样本名匹配，找到相应的模板。若有可匹配的模板，则直接调用对应的模板对短句进行结构化，并返回结果；若匹配不到，则对短句进行简单处理，如用特殊词库、数量词等进行匹配，并返回结果。数据流程如图 3 所示。

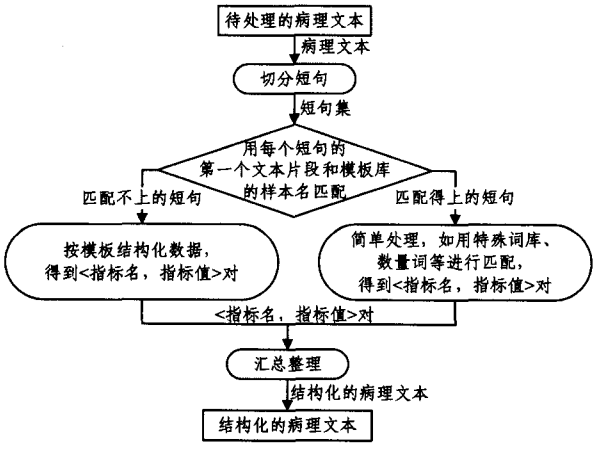

图 3 模板应用的数据流图

同时可以将匹配不到的模板或者模板处理效果不好的病理文本加入到待处理库中，当待处理库中的病理文本累积到一定量之后，再次通过结构化处理流程来优化模板。

4 实验结果

4.1 处理结果

实验数据集来自经过隐私处理的真实数据集的 8 万条历史病理文本记录，随机选择其中 7 万条作为训练数据集，剩下 1 万条作为测试数据集。实验的硬件和软件环境为：CPU 主频 3.10GHz，内存为 12GB；Windows 7 系统，实现语言为 Java，编程环境为 Eclipse J2EE。

经过训练数据集的训练，可得各个样本名对应的模板。图 4 展示了几个典型样本的模板正确率、召回率和 F 值。此外，每个样本在数据中被描述的次数不同，导致其对应模板的正确率也随之变化，如图 5 所示。

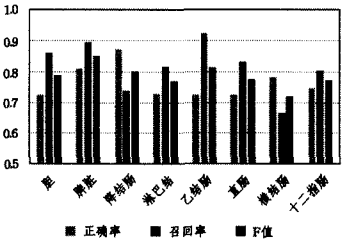

图 4 几个典型样本的模板的正确率、召回率和 F 值

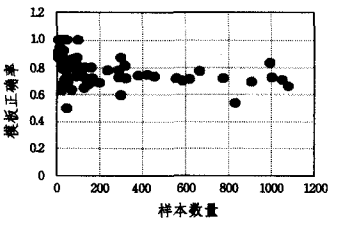

图 5 样本数量与模板的正确率

从图 4 和图 5 可以看出，模板 F 值在 80%上下浮动。同时可以看到在样本数量变少时，模板正确率开始不稳定，这是因为随着样本数量的减少，对应模板的训练集也变小，指标名提取的误差随之增大。同时观察模板数据可以发现，部分分词碎片在大多数模板中高频出现，而这些分词碎片并不是我们期望的模板中的成分，这也导致了整体模板正确率偏低，这一现象在样本数量小时表现尤为明显。从图 4 还可以看出，模板的召回率与正确率接近，且召回率普遍偏高，而上文提到的分词碎片也是在提高召回率时带来的副作用。

用训练好的系统对测试数据集进行结构化处理，经计算，平均每条记录结构化后的正确率为 82.8%。

4.2 实验对比

将本文提出的结构化处理过程中的文本聚类算法分别替换成改进的 KMEANS 算法[16]和 RPCL 算法[17]，在相同的实验环境下实现并应用于实验数据集，得到如图 6 所示结果。从中可以看出，3 种聚类算法的聚类正确率都在可接受的范围内，而本文的方法的聚类正确率和模板 F 值要比其他两种高，导致这一结果最主要的原因是病理数据的特殊性。改进的 KMEANS 基于“距离最远的样本点最不可能分到同一个簇中”这一事实，采用最大距离法依次选取 k 个初始簇中心，同时也对应重新构造迭代中的簇中心计算公式和测度函数；

RPCL 算法在初始簇中心数大于实际类别数的情况下，可以在聚类过程中不断修正获胜中心和次胜中心的权重，使得在迭代过程中，将权值接近于数据集的初始簇中心收敛到实际的类中心，并推离远离数据集的簇中心。数据的类别越明显，以上两种算法的效果越好，适用于社交文本或不同领域文本的聚类，但病理文本的聚类是对一个领域内的文本的更细粒度的划分，数据的类别特征并不明显，因此需要一种更有针对性的文本聚类算法。此外，3 种聚类方法的模板正确率相差不大，但是聚类结果的正确与否会大大影响模板召回率。

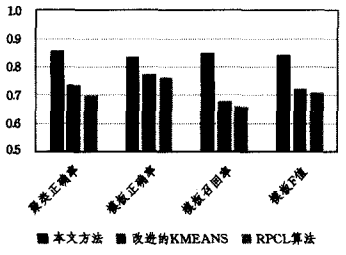

图 6　不同聚类方法对结果的影响

同时，通过对本文提出的文本相似度计算与最常用的余弦相似度计算的对比可以看出，采用余弦相似度，需要先将样本数据表示成数值，再对数值进行计算，每一步转换都会带来误差，而本文提出的方法直接从中文字词入手，减少了不必要的误差，如图 7 所示。

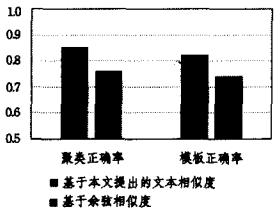

图 7　不同相似度计算方法对结果的影响

结束语　本文提出一种用于中文病理文本的结构化处理方法，该方法先从病理历史数据中提取出模板信息，然后将模板用于新的实时病理数据，从而将非结构化的中文病理文本结构化。实验证明，本文提出的方法在实际场合中具有很好的结构化处理效果，处理结果的平均正确率达 82.8%；同时本方法提取的模板会被定期优化，以适应最新的数据结构化需求。未来的工作将扩展本处理方法的兼容性，使其在更多的中文病理数据集上也能有很好的应用效果。

参 考 文 献

[1] Chen Jin-xiong. Electronic Medical Record and Electronic Medical Record System[J]. Chinese Medical Equipment Journal, 2010,31(10):1-7(in Chinese)
陈金雄. 电子病历与电子病历系统[J]. 医疗卫生装备，2010,31(10):1-7

[2] Gunnar E, Bente C, Line S. Developing Large-scale Electronic Patient Records Conforming to the openEHR Architecture[J]. Procedia Technology,2014,16:1281-1286

[3] Hiroshi T, Yasushi M, Takeo O, et al. A Japanese approach to establish an electronic patient record system in an intelligent hospital[J]. International Journal of Medical Informatics, 1998, 49(1):45-51

[4] Hiroshi T, Yasushi M, Shigeki K, et al. Architecture for networked electronic patient record systems[J]. International Journal of Medical Informatics,2000,60(2):161-167

[5] Emilia A, Channin D S, Demner-Fushman Dina, et al. Automatic segmentation of clinical texts[C]// Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Antalya,2009:5905-5908

[6] Maria S, Maria K, Nilsson G H, et al. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study[J]. Journal of Biomedical Informatics,2014,49(5):148-158

[7] Savova G K, Masanz J J, Ogren P V, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications[J]. American Medical Informatics Association,2010,17(5):507-513

[8] Carol F, Lyudmila S, Yves L, et al. Automated encoding of clinical documents based on natural language processing[J]. American Medical Informatics Association,2004,11(5):392-402

[9] Kong Xiao-feng, Li Ying, Li Hao-min, et al. Structurization of Digestive Endoscopy Report Based on NLP[J]. Chinese Journal of Medical Instrumentation,2008,32(5):348-351(in Chinese)
孔晓风，李莹，李昊旻，等. 基于自然语言处理技术的消化科内窥镜检查报告的结构化[J]. 中国医疗器械杂志，2008,32(5):348-351

[10] 张华平. NLPIR 汉语分词系统[OL]. http://ictclas.nlpir.org

[11] 邱锡鹏. 中文自然语言处理工具包 FNLP[OL]. https://github.com/xpqiu/fnlp

[12] Steven B, Ewan K, Loper Edward, et al. NLTK[OL]. http://www.nltk.org

[13] Mima H, Ananiadou S. An application and evalution of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese[J]. International Journal on Terminology,2001,6(2):175-194

[14] Frantzi K T, Ananiandou Sophia. Extracting nested collections [C]// Proceeding COLING'96 Proeeedings of the 16', Conference on Association Computational Linguistic. Stroudsburg, 1996:41-47

[15] 吴军. 数学之美[M]. 北京：人民邮电出版社，2012

[16] Zhai Dong-hai, Yu Jiang, Gao Fei, et al. K-means text clustering algorithm based on initial cluster centers selection according to maximum distance[J]. Application Research of Computers, 2014,31(3):713-719(in Chinese)
翟东海，鱼江，高飞，等. 最大距离法选取初始簇中心的 K-means 文本聚类算法的研究[J]. 计算机应用研究，2014,31(3):713-719

[17] He Hui, Chen Bo, Xu Wei-ran, et al. Short Text Feature Extraction and Clustering for Web Topic Mining[C]// Third International Conference on Semantics, Knowledge, and Grid, 2007: 382-385