非结构化病理文本的结构化信息抽取方法*

张盈利 夏小玲

(东华大学 上海 201620)

〔摘要〕 介绍病理文本数据结构和概念层次结构，以非结构化的病理文本为对象，首先对非序病理文本的结构进行分析，其次利用模式匹配对病理文本予以模式提取和泛化，最后从分词序列中抽取结构化信息，实验表明该方法能够获得较高的准确率和召回率。

〔关键词〕 病理文本；模式匹配；模式提取；结构化信息

〔中图分类号〕R-056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2016.04.012

Method of Structured Information Extraction from Unstructured Pathological Texts *ZHANG Ying-li, XIA Xiao-ling, Donghua University, Shanghai 201620, China*

〔**Abstract**〕 The paper introduces the data structure and conceptual hierarchy of pathological texts. Based on unstructured pathological texts, at first, it analyzes the structure of pathological texts other than prefaces. Next, it conducts pattern extraction and generalization of the pathological texts by pattern matching. At last, it extracts structured information from the participle sequence. As proved by the experiment, this method can achieve a high accuracy and recall rate.

〔**Keywords**〕 Pathology text; Pattern matching; Pattern extraction; Structured information

1 引言

非结构化文本（Unstructured Text）[1-2]是一类面向领域的应用型文本，具有较强的领域特征。病理文本作为医学领域特有的文本形式，具有独特的结构和书写规范。随着医疗服务的快速发展，以文本形式呈现的病理文本的应用越来越广泛，从中抽取结构化信息具有较高的研究价值和可观的应用前景。目前的非结构化文本信息抽取方法[3-7]主要是面向多领域的，适用性比较强，但抽取结果的准确性和完整性普遍较低，无法满足医疗数据分析的要求。为了获得有分析价值的抽取结果，必须提高准确性和完整性。针对上述问题，本文提出了关于病理文本、基于模式匹配[8]的结构化信息抽取方法。结合病理文本的结构特点和病理报告书写规范，分析病理文本信息的特点，根据其特点，利用模式匹配对病理文本进行模式提取。

2 方法

本文提出的关于病理文本、基于模式匹配的结

〔修回日期〕 2015-12-18

〔作者简介〕 张盈利，在读硕士研究生。

〔基金项目〕 上海市信息化发展专项资金项目“基于瑞金医院的临床大数据平台建设及深度应用”（项目编号：20140314）。

构化信息抽取方法主要包括4部分，分别是：病理文本数据预处理、模式提取、模式泛化和信息抽取，其具体流程，见图1。其中，病理文本数据预处理模块主要进行子句切分、标本名提取以及对子句分词；模式提取模块根据现有模式从子句序列中提取新模式；模式泛化模块把多个模式泛化为一个模式，降低模式选取的复杂度，提高适用性。

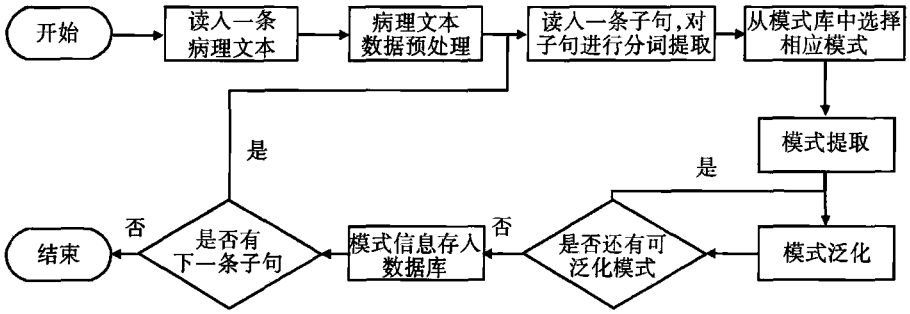

图1 结构化信息抽取系统流程

3 病理文本数据

3.1 病理文本数据结构

本文所处理的病理数据来自医院的真实病理报告，其中病理文本在形式上是病理科医生采用自然语言描述的文本数据。主要研究病理报告的肉眼观察描述信息，病理报告的样例数据，见表1。

表1 病理报告样例

病理号	年龄	性别	巨检	诊断
2012-138255	56	男	右下叶11*13*8cm，胸膜局部增厚、粘连。基底段管壁浸润型肿块4.5*2.5*2cm，灰白，质硬，边界不清，侵犯支气管壁，管腔大部分阻塞，距支气管切端0.8cm，前基底段另见肿块2*1.5*1.5cm，灰白，质硬，界不清，紧贴胸膜，余肺支气管通畅，轻度气肿	所送标本经检验均为良性

3.2 概念层次结构

医生书写病理报告时，一般都会按照病理报告规范书写，例如，描述的对象一般都是送检标本中的不同标本。然后针对具体的一个标本，描述其具体的指标和指标值，见图2。

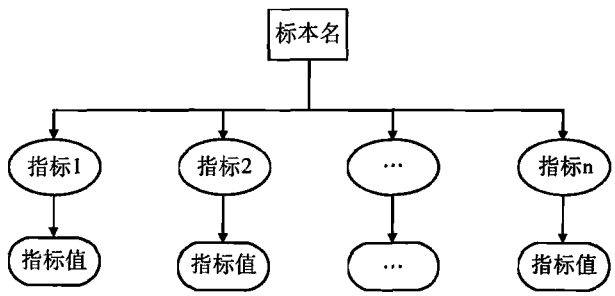

图2 病理概念层次结构

以表1中的病理报告为例，该例中的标本有右下叶、肿块等，从中还可以看出，右下叶指标有大小、胸膜；肿块指标有类型、位置、大小、颜色、质地等。在对某一标本的描述中，医生会遵循一定的模式。本文提出的方法就是从病理报告中提取标本的描述模式，从而尽可能精确地抽取文本信息。例如，在表1中的病理报告巨检字段包括两个标本：右下叶和肿块。其描述模式前者为［标本名］，大小：［数值］，胸膜：［描述］。肿块标本描述模式为［标本名］，位置：［描述］，类型：［描述］，大小：［数值］，颜色：［描述］，质地：［描述］，边界：［描述］。标本描述模式可以分为两部分，一个是标本名，位于描述模式最前方；另一个就是指标名和指标值的映射关系，对每组映射的描述通常用“，”隔开。

4 病理文本数据预处理

4.1 标本归类

在病理文本数据的预处理过程中，首先切分子句，对标本进行分类，根据病理文本的描述特点把文本切分成子句，一个子句对应一个标本名；其次，提取标本名；最后，对所有子句进行中文分词。标本归类是根据病理报告的特点和书写规范实现的。首先，对病理报告描述进行分析。经分析，可得出病理标本子句切分规则。(1) 文本切割。在遍历病理文本时，如遇“见”、“可见”、“未见”、“送”、“另送”、“另见”、“NO. X”、“找到”等字眼或其前后方包含“,”、“;”、“。”等，对文本进行切割。(2) 断句标点。在遍历病理文本时，如遇到“;”、“。”等表示断句的标点符号，对文本进行切割。以表1中数据为例，分割之后的示意结果如下所示：(1) 右下叶11 * 13 * 8cm，胸膜局部增厚、粘连。(2) 基底段管壁浸润型肿块4.5 * 2.5 * 2cm，灰白，质硬，边界不清，侵犯支气管壁，管腔大部分阻塞，距支气管切端0.8cm。(3) 前基底段另见肿块2 * 1.5 * 1.5cm，灰白，质硬，界不清，紧贴胸膜，余肺支气管通畅，轻度气肿。

4.2 标本名的提取

由于医学病理报告的书写规范要求，对于标本名的描述一般出现在语句的前端。根据这样的特性，提出如下算法来提取标本名，算法如下：(1) 输入。标本描述语句。(2) 算法。①对输入的标本描述语句，根据标点符号切分成子句，放入集合中。②将集合前端的子句利用IKAnalyzer进行中文分词。③利用正则匹配的方式过滤掉分词结果中对量词的描述，如将“右下叶11 * 13 * 8cm”中的“11 * 13 * 8cm”过滤掉。输出该子句中剩余的部分，作为该标本的标本名。(3) 输出。该标本名与描述文本的键值对。经过标本名提取算法的处理，标本名提取之后的结果如下：输入的标本为上述分割之后的子句，经过标本名提取算法之后的输出结果：(1) <右下叶，右下叶11 * 13 * 8cm，胸膜局部增厚、粘连。>(2) <肿块，基底段管壁浸润型肿块4.5 * 2.5 * 2cm，灰白，质硬，边界不清，侵犯支气管壁，管腔大部分阻塞，距支气管切端0.8cm。>(3) <肿块，前基底段另见肿块2 * 1.5 * 1.5cm，灰白，质硬，界不清，紧贴胸膜，余肺支气管通畅，轻度气肿。>

5 模式提取及泛化

5.1 模式提取

本文采用样本学习的方法构建初始模式库。用户根据样本中待抽取信息对应的分词序列，按顺序提取指标名放入模式中，对没有指标名与其相对应的指标值，用户自己定义指标名，放入模式中，然后对相同标本名的模版进行模式泛化，最后将泛化后模式保存到模式库中。初始模式库构建完毕后，依据以下方法可自动从描述语句中提取新模式，假设当前子句S的分词序列为：WS = {key/, /w_1/w_2/... /w_n/。}，其中w_i表示分词序列中的第i个词，且分词序列保留所有标点符号，key是该子句的标本名。对任一模式P，其关键词也是key。模式提取算法如下：

输入：子句分词序列WS、初始模式库模式Pf，且WS. key = Pf. key。

算法：Step1： 令i = 1，w_i是WS的第i个词，Pf [j] 是模式Pf的第j个指标。

Step2：for (int j = 0; j < Pf. length; i + +)
{if (w_i = Pf [j]) {转 Step6。}}
else 转 Step3。

Step3：switch (w_i) {
case ‘% 型’：temp = ” 类型”；转 Step4；
case ‘距%’：转 Step5；
case ‘余肺’：转 Step8；
... break;}

Step4：for (int j = 0; j < Pf. length; i + +)
{ if (temp = Pf [j]) { 转 Step6。}
else { temp 放入 PN 中；转 Step7。}}

Step5： temp = w_i + w_{i+1}，转 Step4。

Step6： 把 Pf［j］放入新模式 PN 中。

Step7：顺序读取分词序列中下一个标点符号后第1个词，若为空，返回新模式 PN；若不为空，转 Step2。

Step8：返回新模式 PN。

输出：新模式 PN。

上述算法中，Pf 是初始模式库中关键字与分词序列标本名相同的模版，PN 是与分词序列相对应提取的新模式。其中，根据模式匹配算法思想，建立分词序列与模式的对应关系，为无指标名的指标值生成对应的指标名，最终建立与分词序列相对应的新模式。

5.2 模式泛化

为满足抽取模式具有较强通用性的要求，对上述抽取模式要进行优化和泛化等操作。参考逆向最短编辑距离泛化方法[9]，本文根据泛化模式 X 和模式 Y 的标本名是否相同，提出了两种模式泛化路径，见图3。

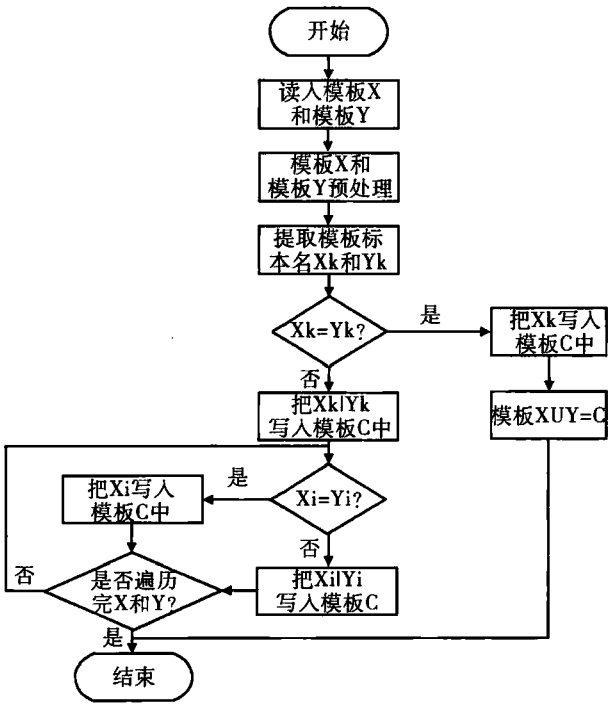

图3 模式泛化路径

其中，Xk 和 Yk 分别是模式 X 和模式 Y 的标本名，Xi 和 Yi 分别是模式 X 和模式 Y 的指标名。标本名不同时，根据待泛化模式 X 和 Y 的编辑距离矩阵 H_{mn} 构造正向最短编辑距离路径 L，在构造路径过程中执行保留、删除、合并等操作，得到新模式 C，本文采用如下公式所示编辑距离计算公式构造编辑距离矩阵：

$$H(i,j) = \min\{h(i,j) + H(i-1,j-1), 1 + H(i-1,j), 1 + H(i,j-1)\}$$

其中，当 Xi 和 Yj 不同时，$h(i,j) = 1$；当 Xi 和 Yj 相同时，$h(i,j) = 0$；Xi 和 Yj 分别是模式 X 和模式 Y 的第 i 个和第 j 个指标名。根据编辑距离矩阵中正向最短编辑距离的模式泛化过程如下：

Step1：记 i=1，j=1。

Step2：若 i=m 且 j=n，返回模式 C；否则转 Step3。

Step3：记 $\min\{h(i,j) + H(i-1,j-1), 1 + H(i-1,j), 1 + H(i,j-1)\}$ 对应下标 i 和 j′，记 $\Delta H_{ij} = H(i,j) - H(i',j')$。

Step4：若 $i' = i-1$ 且 $j' = j-1$ 转 Step 7；否则，转 Step 5。

Step5：若 $i' = i-1$ 且 $j' = j$，转 Step 8；否则，转 Step 6。

Step6：若 $i' = i$ 且 $j' = j-1$，转 Step 9。

Step7：若 $\Delta H_{ij} = 0$，Xi 加入 C 中；若 $\Delta H_{ij} = 1$，(Xi | Yj) 加入 C 中；转 Step 2。

Step8：若 $\Delta H_{ij} = 0$，Xi 加入 C 中；若 $\Delta H_{ij} = 1$，(Xi) 加入 C 中；转 Step 2。

Step9：若 $\Delta H_{ij} = 0$，Yj 加入 C 中；若 $\Delta H_{ij} = 1$，(Yj) 加入 C 中；转 Step 2。

如下两个模式：

X｛a/b/c/d/e｝、Y｛a/b/g/d/e｝

泛化过程，见图4。

	a	b	c	d	e
a	0	1	2	3	4
b	1	0	1	2	3
g	2	1	1	2	3
d	3	2	2	1	2
e	4	3	3	2	1

图4 基于正向最短编辑距离的模式泛化

根据正向最短编辑距离泛化后的抽取模式如下：C {a/b/ (c | g) /d/e}。

6 实验结果及评价

6.1 模式提取结果展示

本文实验用的数据源为某三甲医院除去敏感信息之后的真实医学病理报告数据集。通过对历史病理数据的模版提取及泛化，最后得到的病理标本模式的整理结果，见表2。其中，肺包括右下叶、右上叶、右中叶、左下叶、左上叶、左中叶、左肺、右肺。通过表2可以发现，提取出的标本模式的关键字覆盖了绝大部分的病理指标，并且通过模式泛化和评价指标的独立，使模式与病理标本的契合度更高，有效提高信息抽取的完整性。

表2 病理标本模式提取结果

标本名	肺 | 楔切肺	肿块 | 结节 | 病灶	余肺
模式	{大小，胸膜}	{位置，类型，大小，直径，颜色，质地，边界，距胸膜，距支气管}	{评价}

6.2 评价指标

本文关注的主要是抽取结果的完整性和准确性，选取查全率和准确率[10-11]作为结构化信息抽取结果的评价指标，其计算公式如下：

$$\text{查全率 } R = I/T \quad (1)$$

$$\text{准确率 } P = I/W \quad (2)$$

其中，I表示所有正确抽取指标值数目，T表示所有待抽取信息指标值数目，W代表所有抽取到的指标值数目。通过对指标名添加项数目的设置，分别得出相应的准确率和查全率。对比准确率、召回率与指标名添加项之间的关系，发现：（1）指标名添加项数目超过10个时，召回率趋于稳定，约92%。（2）并不是指标名添加项越多，准确率越高。（3）当指标名添加项数目为12时，准确率最高达88%。由此可知，模式提取时若病理文本结构分析不充分，信息的准确性和完整性较低；病理文本结构分析充分时，能获得较高的查全率和准确率。

7 结语

本文针对病理文本数据的层次和结构特点，结合模式匹配算法思想提出了一种针对具体标本的结构化信息抽取方法。实验表明该方法能获得较高的召回率和准确率。在分析具体非结构化文本数据的层次和结构特点的基础上，提出了相适应的模式提取和泛化方法，获得了有分析价值的抽取结果，对非结构化数据转化为结构化数据的研究具有借鉴意义。

参考文献

1 邓世洲，王秀民，刘帆. 基于病种的结构化电子病历探讨 [J]. 医学信息学杂志，2012，33 (7)：11-14.

2 姚亮，陈耀龙，王琪，等. 病理报告的报告规范解读 [J]. 中国循证儿科杂志，2014，(3)：216-219.

3 李爱民，谭献海. 基于XML技术的非结构化数据到结构化数据转换的研究 [J]. 计算机应用，2012，21 (10)：12-15.

4 Imran R. Mansri, Sunita Sarawagi. Integrating Unstructured Data into Relation Databases [J]. JCDE, 2006, 3 (7): 29.

5 刘若中. 基于纯XML数据库和HL7的结构化电子病历研究与应用 [J]. 医学信息学杂志，2009，30 (9)：38-40.

6 Tao Peng, Lianying Sun, Hong Bao. Research of Unstructured Data Transformation Based on XML [J]. International Conference on Internet Technoeugy&Applieations, 2010, 20 (22): 1-4.

7 Ying Chen, Sophia Yat Mei Lee, Chu-Ren Huang et al. A Robust Web Personal Name Information Extraction System [J]. Expert Systems with Application, 2012, 39 (3): 2690-2699.

8 姚亚峰，蒋毅. 模式匹配算法及其优化 [J]. 南通职业大学学报，2011，(4)：98-100.

9 邵堃，杨春磊，钱立宾，等. 基于模式匹配的结构化信息抽取 [J]. 模式识别与人工智能，2014，(8)：758-768.

10 吴麒，陈兴蜀，谭骏，等. 基于权值优化的网页正文内容提取算法 [J]. 华南理工大学学报（自然科学版），2011，39 (4)：32-37.

11 向程冠，熊世桓. 一种基于特征树的Web碎片信息抽取算法 [J]. 兰州理工大学学报，2014，40 (1)：104-107.