

# 篇章标注在医学领域问答系统中的应用<sup>1\*</sup>

王宇昕, 李素建, 王荀

北京大学计算语言学教育部重点实验室 北京 100871

E-mail: {wangyuxin, lisujian, xunwang}@pku.edu.cn

**摘要:** 问答系统是自然语言处理领域中的一个研究热点, 针对特定领域的问答系统也越来越受到重视, 但相关研究主要利用了浅层的语言处理技术。为了适应深层的语言理解, 本文针对医学领域, 提出了一套篇章标注体系, 同时考虑了篇章单位的内容和语义逻辑关联, 并给出了基本的内容标签集合和关系标签集合。同时, 我们将篇章标注应用在面向医学领域的问答系统中, 通过具体示例说明如何辅助问答系统进行答案的抽取, 实验结果表明我们提出的篇章标注体系可以较好地应用在医学领域问答系统中。

**关键词:** 篇章标注; 问答系统; 医学问答

中图分类号: TP391

文献标识码: A

## Applying Discourse Annotation in Medical Question Answering System

WangYuxin, LiSujian, WangXun

Key Laboratory of Computational Linguistics (Peking University), Ministry of Education, Beijing 100871

E-mail: {wangyuxin, lisujian, xunwang}@pku.edu.cn

**Abstract:** Question answering has been hot in the field of natural language processing and how to develop a domain specific QA system has drawn much attention. Prior researches on QA mainly adopt the shallow NLP techniques and seldom explore the results of discourse parsing. To better understand language, this paper proposes a set of discourse annotation tags which cares for both content and logical association of discourse units. At the same time, we apply our discourse annotation results in a medical domain specific QA system, and exemplify how the discourse tags help to extract the final answers. Experimental results also verify that our discourse annotation can better improve the medical QA system.

**Key words:** Discourse Annotation; Question Answering System; Medical Domain Specific QA

### 1 引言

早在半个世纪前, 人工智能刚被提出时, 问答系统就已成为热门话题。现在, 人们希望问答系统返回精确、简练的答案的同时, 也对特定领域问答系统的专业性提出了更高要求, 而相关研究仍处在主要利用浅层语言处理技术的阶段。比如, 回答事实性问题时, 根据疑问词确定问题类型, 根据关键词从文档中检索并提取候选句子; 而基于模式匹配的方法考虑了部分表层语法信息, 在时间、地点等问题上有较好效果, 但对复杂问题收效甚微。问答系统广泛采用的自然语言处理技术以浅层的句法分析为主, 基本没有考虑篇章分析结果。篇章分析对理解句子与段落间的深层关系有非常大的帮助, 而篇章标注是将其结构化, 通过规范的标签标注体系呈现篇章组成单位之间的语义关系。为了适应深层的语言理解, 本文针对医学领域, 提出一套篇章标注体系, 使用篇章单位的内容和语义逻辑关联相结合的方法, 并给出

---

<sup>1\*</sup> 本项目受到国家自然科学基金项目(编号: 61273278)、国家社会科学项目(编号: 12&ZD227), 国家科技支撑计划子课题项目(编号: 2011BAH10B04-03)和国家 863 计划(编号: 2012AA011101)的资助。

了基本的内容标签集合和关系标签集合,改进医学领域问答系统性能,探讨篇章标注在医学领域问答系统中的应用。

## 2 问答系统与篇章标注的概况

### 2.1 问答系统概况

问答系统,就是一种针对用户以自然语言提出的问题,从相关语料库中查询并返回准确、简洁回答的系统。一般的问答系统,首先在问题库中查询是否存在相同或相似的问题,如果有就直接返回答案。如果没有,就进行问题分析,包括词性标注、句法分析、问题分类等,一方面提取关键词,并扩展构成检索式进行文档检索,另一方面根据不同的问题类型,从检索返回的结果中抽取最相关的词语、句子或段落,整理成最终的答案。国际信息检索领域最权威的评测会议 TREC,设有针对英文问答系统的评测比赛,每年都吸引着全球最知名的问答系统研究机构参与。

目前的问答系统一般分为基于知识库的问答系统、基于传统检索的问答系统和基于原始文本的问答系统。基于知识库的问答系统,需要相关领域的大规模数据库,擅长回答领域专业知识,但知识库成本较高。基于传统检索的问答系统,在传统搜索引擎的基础上,从大量网页文本中返回给用户需要的内容,所以也被称为智能搜索引擎,它返回的答案通常较多,更像是一个搜索引擎。基于原始文本的问答系统,从未经人工处理的原始语料中获取答案,不需要建立知识库,但对自然语言处理的要求比较高。

问答系统的发展与自然语言处理技术的发展密切相关。对于简单的回答时间、地点等问题,问答系统已较为成熟,但针对医学这样的特殊领域,由于医学文本知识结构和医学问题描述方式的复杂性,使得医学领域问答系统还不完善,相关研究还停留在起步阶段。

### 2.2 篇章标注概况

篇章标注通常是指对篇章文本组成部分间的语义关系的标注。最小语义单元为句子、小句或短语,一般只考虑相邻成分间的语义关系,自底向上层层推进,构成树状结构的语义关系图。有时过于复杂,也会采用图模型进行表示。

现在的篇章标注一般基于 RST 理论,它的观点是,文档整体由一个个语义单元以及连接它们的不同语义关系构成。在 RST 基础上发展而成的 RST-DT (RST-Discourse Treebank) 语料库 (Carlson 等, 2001),就是对 385 篇 Wall Street Journal 文章进行标注得到。而应用颇为广泛的 PDTB (Penn Discourse Treebank) 语料库 (Prasad 等, 2006),对上百万字的 Wall Street Journal 文章进行标注,使用了更加简洁的标签设置,以及更加严谨的层次化结构标签体系。另外的还有基于框架语义学的 FrameNet 语料库和基于图模型的 Graphbank 等。

## 3 篇章标注体系的设置

以往篇章标注系统中更注重的是语言单位之间的修辞关系,反应了语言的语义逻辑。而我们将内容标签和关系标签分别作为篇章标注中独立的体系。文档的重要内容用内容标签进行标识,而文档单元之间的逻辑关系用关系标签进行标识。内容标签既有通用标签,又有根据知识库语料可扩展的特殊标签,保证了内容标签的灵活性。关系标签采用语义逻辑中的常用关系标签,保证了稳定性。同时,在保证清晰标识文档内容和关系的基础上,尽可能简洁地使用标签。

在实际标注过程中，我们利用自己开发的篇章标注软件，实现单个成分的划分、内容标签的标注和关系标签的标注。下面就对标签选择和标注软件的应用展开介绍。

### 3.1 内容标签的选择

内容标签用来说明文档单个成分的内容意义。以皮肤病领域为例，基于我们收集的皮肤病学语料，我们分别使用通用范围、医学领域、皮肤病领域这三个层次的内容标签对语料进行内容标注，如表 1 所示。其中，通用范围和医学领域的内容标签，可以扩展到医学领域其他方面的问答。

表 1：内容标签

适用范围	内容标签
通用范围	定义、功能、分类、数据等
医学领域	病理、病因、诊断、治疗等
皮肤病领域	股癣、湿疹、恶性黑素瘤等

典型的语料如图 1 所示。我们可以看到，医学文献结构比较规范，领域专有名词经常出现在标题中或用方括号加以标识。这段文本的自动内容标注结果如图 2 所示，“#”后面是内容标签。在自动标注的基础上，我们会进行人工校对，根据文档本身人工添加、修改内容标签。另外，一段文本可以同时存在多个内容标签。

19.1 【诊断】 儿童头上有断发的鳞屑斑、“黄癣”、“菌鞘”发或“黑点”发，病发真菌镜检有发外，发内孢子菌丝或鹿角菌丝即可诊断	
19.2 真菌培养可确诊	
20.1 【鉴别诊断】 与头皮的脂溢性皮炎、银屑病头皮脓疱疮和脓肿等相鉴别	19.1-19.2 #诊断
20.2 真菌镜检和培养是关键	20.1-20.2 #鉴别诊断
21.1 【治疗】 采用综合治疗方法	21.1 #治疗

图 1：内容标注语料

图 2：内容标注结果

图 3 是标注软件的内容标注界面。选择文本后点击上方的按钮完成内容标签标注。由于内容标签的灵活性，我们也可以任意添加新的内容标签。

文件	操作							
	分段分句							
	添加标签	内容标注	功能	分类	数据	病理	病因	诊断
		关系标注						
95	19.1 【诊断】	儿童头上有断发的鳞屑斑、“黄癣”、“菌鞘”发或“黑点”发，病发真菌						
96		发外，发内孢子菌丝或鹿角菌丝即可诊断						
97	19.2	真菌培养可确诊						
98	20.1 【鉴别诊断】	与头皮的脂溢性皮炎、银屑病头皮脓疱疮和脓肿等相鉴别						
99	20.2	真菌镜检和培养是关键						
100	21.1 【治疗】	采用综合治疗方法						

图 3：内容标注界面

### 3.2 关系标签的选择

关系标签用来描述文档相邻成分之间的逻辑关系。关系标签体系采用李素建(李素建等, 2013)提出的汉语篇章关系体系, 包括附加、对照、因果、目的、评价、总分、顺承、并列等 26 种关系。在医学领域中, 许多复杂的关系很少出现。我们统计语料中出现的关系标签, 按照频率从高到低排序, 前 9 名如表 2 所示。这 9 种篇章关系占了所有关系的 84.61%, 其中“解释说明”排名第一, 是因为我们把语料中小标题与它下属段落的关系统一标成“解释说明”, 同时, 规范的医学文献中小标题的形式十分常见。另外, “分述”可以认为是没有总起句的“总分”关系。

表 2: 出现频率最高的 9 种关系标签

关系标签	释义	频率
解释说明	进一步解释	24.85%
并列	二者彼此独立, 同等地位	16.86%
顺承	时间上先后、逻辑发展或者空间顺延	16.27%
因果	原因	7.69%
分述	从不同角度叙述	6.51%
条件	条件	4.14%
结论	结论	3.25%
总分	总分	2.66%
递进	主要内容和描述对象相同, 语义上累加新的信息	2.37%

语料上的标注结果如图 4 所示, 在内容标注基础上实现自动的关系标注。“#”后面是已有的内容标签, 根据医学领域专有名词中, “治疗”和“手术治疗”、“化学治疗”、“免疫疗法”、“放射疗法”之间的层级关系, 我们给这段文本标识为“总分”关系。在自动标注的基础上, 我们同样会进行人工校对, 实现精准的关系标签标注。

```

153.1-157.1 总分 #恶性黑色素瘤 #治疗
  153.1
  154.1-157.1
    154.1-154.3 #手术治疗
      154.1
      154.2
      154.3
    155.1-155.2 #化学治疗
      155.1
      155.2
    156.1-156.2 #免疫疗法
      156.1
      156.2
    157.1 #放射疗法

```

图 4: 关系标注结果

图 5 是标注软件的关系标注界面。选择文本后点击上方的按钮完成关系标签标注。已标注的关系标签还可以在界面左侧以树状图的方式实时展现。

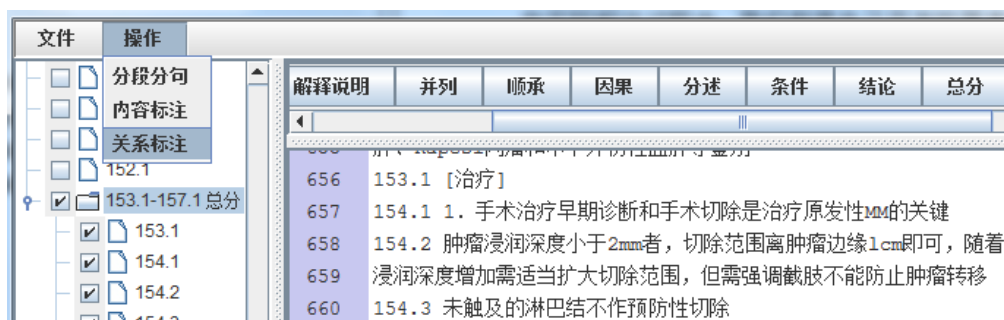


图 5：关系标注界面

## 4 基于篇章标注的医学领域问答

篇章标注在需要依赖篇章关系的任务方面能发挥显著作用，问答系统就是典型应用。问答系统一般由以下几部分组成：问题分类、关键词提取与扩展、信息检索、答案抽取。在没有篇章标注时，信息检索和答案抽取一般停留在浅层的词语匹配。而内容标签和关系标签揭示了文本的深层语义关系，可以辅助问答系统获取更精确的答案。

在医学领域问答中，答案经常包含医学领域专有的命名实体，而问题和答案之间也经常具有一些固定的逻辑关系。借助文档重要内容含义的标注，以及文档单元之间的结构关系标注，我们可以更有效地获得更为精确的答案。

我们假设，问题  $Q$  经过关键词提取与扩展后得到的检索 query 为  $Key=\{Key_1, Key_2, \dots\}$ ，检索返回的文本按照分数从高到低依次为  $Answer=\{Answer_1, Answer_2, \dots\}$ 。对某一个  $Answer_i$ ，它具有内容标签  $Content_i=\{Content_{i1}, Content_{i2}, \dots\}$ ，以及与相邻文本构成关系  $Relation_i=\{Relation_{i1}, Relation_{i2}, \dots\}$ 。如果问题  $Q$  本身包含内容标签信息  $Content_q=\{Content_{q1}, Content_{q2}, \dots\}$ ，以及关系标签信息  $Relation_q=\{Relation_{q1}, Relation_{q2}, \dots\}$ ，那么，根据  $Content_i$  与  $Content_q$  的匹配程度，以及  $Relation_i$  与  $Relation_q$  的匹配程度，我们可以调整  $Answer_i$  在  $Answer$  中的排序，更准确地选择答案。最后我们选取的答案，一般是检索结果中排序靠前，且内容标签、关系标签匹配程度较高的文本段落。下面用实例说明如何结合篇章标注进行答案的选择。

### 4.1 内容标签辅助答案抽取过程

在答案抽取模块中，一般的问答系统会根据信息检索得到的文本段落，选取可能性最大的答案。有时，遇到较为复杂的问题时，还需要基于出现问题关键字的上下文，返回文本片段，或者返回多个近似的候选答案。在医学领域中，专有名词针对性明确，如果需要返回的答案明确属于一种专有名词类型，那么答案抽取工作将变得非常简单。所以，我们构建词典存储“词/短语-内容标签”对，进行查询并参与答案抽取过程。几个典型的“词/短语-内容标签”对如表 3 所示。

表 3：“词/短语-内容标签”对

词/短语	内容标签
什么病	病名
怎么引起	病因

引起哪些	临床表现/症状
怎么治	治疗
怎么防止	预防

Q1: 股癣是怎么引起的?

A1: 为最常见的皮肤癣菌病男性患病率明显高于女性, 致病菌大多为红色毛癣菌(94%)、其次为絮状表皮癣菌(3%)、须癣毛癣菌(1%)等。发病与温暖潮湿、肥胖或局部潮湿多汗有关。

A1\*: 由于奇痒不断搔抓, 可引起渗液和结痂, 甚至红肿化脓, 反复搔抓使皮肤呈苔藓样变。

问题 Q1 是一个典型的例子。直接检索 Q1 得到文档中最接近的段落为 A1\*, 因为 A1\* 描述的是“股癣”, 同时又包含“引起”这一关键词。但事实上, A1\* 的内容为股癣引起的临床表现, 该文本段落的内容标签为“股癣”、“临床表现”。根据我们的方法, 从“词/短语-内容标签”词典里找到 Q1 中“怎么引起”对应的内容标签为“病因”。信息检索返回的段落中, A1 的内容标签就是“股癣”、“病因”, 它也确实是该问题的正确答案。

我们利用内容标签参与答案抽取过程, 而非作为最关键的答案抽取标准, 是为了防止错误的内容标签导致抽取错误的答案。当符合内容标签的候选答案在检索结果中排序非常靠后时, 我们并不盲目选取该候选答案作为最终答案。

#### 4.2 关系标签辅助答案抽取过程

在英文问答中, 问题经常包含“why”、“how”等虚词。同样, 在中文问题里也经常出现“为什么”、“有哪些”等表达疑问的词/短语。通常的信息检索过程中, 这些词/短语的深层语义并不会参与检索。如果问题和答案、答案和答案之间存在明确关系, 这种关系又能被特定的词/短语描述, 那么我们可以利用这种关系调整信息检索的结果, 帮助抽取更为准确的答案。所以, 我们构建词典存储“词/短语-关系标签”对, 进行查询并参与答案抽取过程。几个典型的“词/短语-关系标签”对如表 4 所示。

表 4: “词/短语-关系标签”对

词/短语	关系标签
为什么	因果
有哪些	总分/分述
怎么发展	顺承/递进
怎么导致	因果
类似	并列

Q2: 皮肤免疫系统的细胞成分有哪些?

A2: 角质形成细胞、淋巴细胞、朗格汉斯细胞、内皮细胞、肥大细胞、巨噬细胞、真皮成纤维细胞。

A2\*: (一)皮肤免疫系统的细胞成分

问题 Q2 是一个典型的例子。根据 Q2 的关键词“皮肤”、“免疫系统”、“细胞成分”检索得到文档中最接近的段落为 A2\*，其次是 A2 等。再从“词/短语-关系标签”词典里找到 Q2 中“有哪些”对应的关系标签为“总分”。文档中，和 A2\*和 A2 构成总分关系，A2\*是“总”部分，A2 是它的“分”部分，所以 A2 也应该作为正确答案被抽取出来，而且更接近准确答案。这样，我们抽取答案 A2，或将 A2 与 A2\*一并抽取。

## 5 实验

基于内容标签和关系标签的方法，我们通过实验说明，篇章标注能帮助提升医学领域问答系统的性能。以下分别展示我们的问答示例和实验结果。

### 5.1 两类问题的问答示例

问答系统通常包含问题分类过程。问题分类是指根据答案类型或问题语义信息，对问题进行类别划分。针对不同类别的问题，采用不同的答案抽取方法，从而更有效地得到准确答案。基于答案类型的问题分类体系，一般将问题分为任务、地点、数字、时间、实体、描述及未知七个大大类。在医学领域问答中，实体类问题和描述类问题出现频率最高，也是相对困难的两类问题。内容标签和关系标签相结合，在一定程度上能有效帮助处理这两类问题。

#### 5.1.1 实体类问题

在医学领域中，实体类问题通常表现为询问病症名称、药物名称、生物学术语等实体类名词。例如 Q2 就是询问生物学术语的实体类问题。针对实体类问题，我们在答案抽取过程中应该把实体类名词作为答案的可能性提高。

Q3: 身体痒死了，一到秋冬季比较明显，一着急特别痒，热了也痒，请问是什么病？

A3: 过敏性皮肤病，或胆碱能性荨麻疹。

问题 Q3 是典型的医学领域实体类问题，对问题抽取关键词后，检索“痒”、“秋冬”、“着急”等，并倾向于寻找实体类名词。同时，从“词/短语-内容标签”词典里找到“什么病”对应的内容标签为“病名”。在返回的许多结果中，抽取最有可能的病名，得到答案 A3。

#### 5.1.2 描述类问题

描述类问题是最常见的一类问题。在医学领域中，描述类问题通常表现为询问病因、病理、治疗过程、生物学原理等描述性内容。同时，由于描述类问题需要返回的段落经常存在逻辑关系，所以这也是最困难的一类问题。例如 Q1 就是询问病因的描述类问题。针对描述类问题，如果有明确的关系标签，我们在答案抽取中应该把它们权重提高。

Q4: 皮肤的吸收能力有差别吗？为什么？

A4: 一般吸收能力是阴囊>前额>大腿屈侧>上臂屈侧>前臂>掌跖。[S1]由于角质层厚薄不一，不同部位的皮肤吸收能力有很大差异。[S2]

问题 Q4 是典型的医学领域描述类问题。Q4 的第一问只需要进行普通检索，得到可能性最大的文本段落为[S1]，它的内容标签为“皮肤”、“吸收能力”，显然就是我们想要的答案。Q4 的第二问，是询问[S1]的原因。从“词/短语-关系标签”词典里找到“为什么”对应

的关系标签为“因果”。文档中，和[S1]构成因果关系的“因”段落内容为[S2]。[S1]和[S2]共同构成了答案 A4。

## 5.2 实验结果

由于问答系统的评测较为复杂，本文只考虑准确率作为评测指标：

$$\text{准确率} = \frac{\text{答对的问题数}}{\text{问题的总数}}$$

其中，每个问题只返回可能性最大的答案。我们的题目来源于百度知道皮肤科类别下的已回答问题，以及医学领域专家撰写的问题，共 100 个，其中包含 29 个实体类问题和 42 个描述类问题，占有问题的 71%。同时，我们与普通检索的方式进行比较，分别统计普通检索、结合内容标签、结合关系标签以及结合内容和关系标签四种方法下的准确率，如表 5、表 6 所示。

表 5：实体类问题评测结果

方法	普通检索	+内容标签	+关系标签	+内容标签+关系标签
答对的问题数	12	21	14	23
准确率	41.38%	72.41%	48.28%	79.31%

表 6：描述类问题评测结果

方法	普通检索	+内容标签	+关系标签	+内容标签+关系标签
答对的问题数	17	26	23	30
准确率	40.48%	61.90%	54.76%	71.43%

可以看出，内容标签和关系标签分别提高了答案准确率，共同作用时准确率达到最高。准确率达到最高时，实体类问题比描述类问题结果更好，是因为描述类问题一般较为复杂，实体类问题相对简单。对这两类问题，内容标签的作用均大于关系标签，我们认为这是由于医学领域专业词汇较多且含义固定所导致，好的内容标签在答案抽取时能发挥很大作用。另外，加入关系标签后，描述类问题的提升效果较实体类问题更优，这也验证了描述类问题更依赖问题和答案之间的逻辑关系这一论断。

## 6 总结和展望

我们在医学领域语料上进行实验，从实验结果来看，这套篇章标注体系可以较好地应用在医学领域问答系统中。内容标签和关系标签对答案抽取的辅助作用也较为明显。在以后的研究中，我们会尝试将篇章标注体系应用到问答系统中的问题分析、关键词扩展、信息检索等环节。比如，我们可以建立关系标签的索引，将问题中隐含的关系标签作为关键词加入检索式，直接返回相关度较高且语义关系精确的答案。这有待我们进一步完善。

## 参 考 文 献

- [1] Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski. "Building a discourse-tagged corpus in the framework of rhetorical structure theory." Proceedings of the Second SIGdial Workshop on Discourse and



- Dialogue-Volume 16. Association for Computational Linguistics, 2001.
- [2] The Penn Discourse TreeBank 1.0 Annotation Manual. The PDTB Research Group. March 29, 2006
  - [3] Verberne, Susan, et al. "Discourse-based answering of why-questions." *Traitement Automatique des Langues, Discours et document: traitements automatiques* 47.2 (2007): 21-41.
  - [4] Kim, Sanghee, Rob Bracewell, and Ken Wallace. "From discourse analysis to answering design questions." *Proceedings of EKAW 2004 Workshop on the Application of Language and Semantic Technologies to support Knowledge Management Processes*. 2004.
  - [5] Chai, Joyce Y., and Rong Jin. "Discourse structure for context question answering." *Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004* (pp. 23-30).
  - [6] Sun, Mingyu, and Joyce Y. Chai. "Discourse processing for context question answering based on linguistic knowledge." *Knowledge-Based Systems* 20.6 (2007): 511-526.
  - [7] 李素建, 王荀, 王宇昕. 内容标签和关系标签相结合的汉语篇章标注规范. 第十四届汉语词汇语义学研讨会 2013.
  - [8] 刘宝艳. 面向生物医学领域的问答系统的研究与实现. *MS thesis*. 大连理工大学, 2007.
  - [9] 毛先领, 李晓明. 问答系统研究综述. *计算机科学与探索* 6.3 (2012): 193-207.
  - [10] 余正涛, 樊孝忠, 郭剑毅, 耿增民. 基于潜在语义分析的汉语问答系统答案提取. *计算机学报* 29.10 (2006): 1889-1893.