# 面向中医古籍的隐式关系抽取方法研究

马月坤[1,2,3,4]，　冯烨琛[1]

(1. 华北理工大学 人工智能学院　河北 唐山 063210；2. 河北省工业智能感知重点实验室　河北 唐山 063210；3. 北京科技大学 计算机与通信工程学院　北京 100083；4. 材料领域知识工程北京市重点实验室　北京 100083)

**摘要**：自然语言种类丰富、形式灵活多变的特征使得隐式关系抽取成为目前关系抽取领域中富有难度和挑战性的任务之一。通过引入构式语法理论和依存句法分析两种认知语言学范畴的理论技术，构建了一种面向中医古籍中隐式关系的抽取方法。首先利用构式语法理论制定文本构式化策略、分析并定义出 8 种构式特征与 5 种构式类型，并使用 CART(classification and regression tree，CART)分类模型完成文本分类；其次对其中 4 类构式使用依存句法分析技术构建句法树，通过分析句法树中的特定结构，制定医学类实体间的关系三元组抽取规则，实现隐式关系抽取；最后在经典中医古籍《黄帝内经》数据集上进行测试，实验结果表明了方法的有效性。

**关键词**：关系抽取；中医古籍；隐式关系；构式语法理论；依存句法分析

**中图分类号**：TP391.1　　**文献标志码**：A　　**文章编号**：1671-6841(2024)02-0034-09

**DOI**：10.13705/j.issn.1671-6841.2022332

# Research on Traditional Chinese Medical Text Implicit Relation Extraction Method

MA Yuekun[1,2,3,4], FENG Yechen[1]

(1. *College of Artificial Intelligence*, *North China University of Science and Technology*, *Tangshan* 063210, *China*; 2. *Hebei Provincial Key Laboratory of Industrial Intelligent Perception*, *Tangshan* 063210, *China*; 3. *School of Computer & Communication Engineering*, *University of Science & Technology Beijing*, *Beijing* 100083, *China*; 4. *Beijing Key Laboratory of Knowledge Engineering for Materials Science*, *Beijing* 100083, *China*)

**Abstract**: The variety of natural languages and their flexible forms are very rich, which makes implicit relation extraction one of the difficult and challenging tasks in the field of relation extraction. Two theoretical techniques were introduced in the field of cognitive linguistics, namely constructive grammar theory and dependent syntactic analysis, to construct a method for extracting implicit relations in traditional Chinese medical texts. Firstly, the constructive grammar theory was used to formulate a text structuring strategy, analyze and define eight constructive features and five constructive types, and the CART classification model was used to classify the text. Secondly, the dependent syntactic analysis technique was used to construct a syntactic tree for four constructs, and by analyzing the specific structure of the syntactic tree, the extraction rules of the relational triad between medical entities were formulated to realize the implicit relation extraction. Finally, the tests were conducted on the dataset of the classic traditional Chinese medical text, i.e., *Huangdi Neijing*, and the experimental results showed the effectiveness of the method.

**Key words**: relation extraction; traditional Chinese medical; implicit relation; constructive grammar theory; dependent syntactic analysis

**收稿日期**：2022-11-20

**基金项目**：河北省三三三人才项目(A201803082)。

**第一作者**：冯烨琛(1997—)，男，硕士研究生，主要从事自然语言处理研究，E-mail：fyc617774894@163.com。

**通信作者**：马月坤(1976—)，女，教授，主要从事自然语言处理、知识工程等研究，E-mail：mayuekun@163.com。

## 0　引言

在人工智能技术蓬勃发展的时代,促进中医的智能化发展成为国家战略,中医知识的智能获取是促进中医领域智能化的重要基础。利用自然语言处理技术对中医领域文本进行处理,实现中医领域知识抽取成为当前研究热点。关系抽取是知识抽取的重要任务之一,旨在已标注实体及实体类型的句子中确定实体间的关系类别[1]。

关系抽取主要分为显式关系抽取和隐式关系抽取两类。当前相关研究大多都集中于显式关系抽取而很少关注隐式关系抽取[2],主要原因在于自然语言种类丰富、形式灵活多变,文本中蕴含的隐式关系类型繁多、分布特点各异,但它们的共同点都体现在语句中缺少代表具体关系类型的词汇或语义成分,例如在“厥头痛,面若肿起而烦心。”一句中,可以通过语义推断出“厥头痛”这种疾病与“面若肿起”“烦心”两个症状产生了一种病症关系,然而在表层字词中没有明确表示“疾病-症状”关系的语义成分。这种语言现象导致目前的显式关系抽取方法无法很好地检测出隐式关系,不能直接用于解决隐式关系抽取问题。在这种情况下,有些专家学者利用额外的知识或不同的系统来协助推理隐式关系。万常选等[2]设计了三种基于协陪义动词的隐式实体关系推理规则,借助语句内部已知的显式关系对隐式关系进行推理,在新闻和旅游领域的语料上进行实验,并取得了良好的效果。缪峰等[3]通过归一化处理显示因果事件中进行的事件动作,形成事件方向和标准的匹配因果事件对集合,并利用此集合、根据事件相似度从新闻语料中抽取隐式因果关系。文献[4]提出结合马尔可夫逻辑网和一阶逻辑方法提取维基百科中的实体关系,采用一阶逻辑范式发现隐式关系。文献[5]提出了基于规则的方法识别阿拉伯人名和组织实体之间“属于”类型的隐式关系。由此可见,面对不同领域的文本,采用额外知识进行隐式关系抽取的方式也会有所不同,很难使用通用的方法实现统一建模。

综上所述,隐式关系抽取是目前关系抽取领域中一项富有挑战性的工作。中医古籍文本中蕴含了大量的隐式关系,本文借助认知语言学中的理论制定面向中医古籍文本隐式关系抽取的策略。

## 1　整体框架

本文通过引入构式语法理论和依存句法分析两种认知语言学范畴的理论技术,构建了一种面向中医古籍的隐式关系抽取方法。本文方法依顺序分为两个步骤。

(1) 文本分类。本文结合构式语法理论分析得到中医古籍文本的构式特征,并利用构式特征确定构式类型,最后结合构式特征训练 CART 分类模型,并实现将文本按照构式类型进行分类。

(2) 隐式关系抽取。本文在文本分类工作的基础上,针对不同构式类型的文本,使用依存句法分析技术分别构建句法树,结合句法树结构设计相应的隐式关系抽取规则,实现对多种隐式关系的抽取。

## 2　基于构式语法理论的文本分类策略

Goldberg[6]提出构式语法理论并给出定义:构式被理解为包含形式、语义及话语功能的配对。构式包括语素、词、习语、抽象的短语格式和句型传统语法理论中的各级语言单位。构式具有符号特征,每个构式都是形式与意义的匹配(或形式与功能的匹配),即有自身的话语功能和意义。结合中医古籍的文本特征,则可以理解为蕴含隐式关系的句子会具备一定的句式结构。如果将此种结构定义为构式,就可以实现对同种句式结构的识别。同时,间接实现同类隐式关系的归类,即如果多个句子中包含的隐式关系类型相同,那么这些句子所属的构式类型也将会相同。

### 2.1　中医古籍的构式化策略

文献[6]还明确提出任何具有高频度的表达式都预备了一个构式。文献[7-8]认为任何高频率的构造或语句,不管是否具有规律性,都可被视为构式,例如 happy birth day 这类表达式就是构式。因此针对中医古籍构建的构式,其构式特征应具备以下特点。

(1) 在句中仅作为一个结构存在,无实际语义。构式强调的是句子的结构具备某种固有特征,其中有实际语义的成分是可以被随意替换的,因此构式特征应是一种固定的结构特征。

(2) 高频率出现。构式语法理论证明了高频率出现的表达式可以成为构式,高频率的特征同时也能体现某种现象具有广泛性、具备研究价值的特点。

(3) 能够维持句子结构完整性。句子结构完整是要求构式中各种要素、成分的完整,例如标点符号的缺失同样是一种对句子结构完整性的破坏。

以《黄帝内经》文本为例,孙琴等[9]对其中的虚词进行了计量分析,统计出虚词单字种类数占全书

单字种类数的20.73%,虚词频次数占全书总字数的26.87%,同时还举例分析了不同种类的虚词在句子中的作用。因此将高频率出现的虚词或结构成分作为构式特征进行处理,可以在不受语义的影响下实现文本分类。

综上所述,本文选取“,”“;”“者”“曰”“故”“其”“于”“则”8个字符共同作为中医古籍文本的构式特征。由于本文关注中医古籍文本中的医学类知识,而不考虑其他类型知识之间的隐式关系,因此本文结合上文选取的构式特征以及中医古籍文本的数据特点,归纳出中医古籍中包含的构式类型主要有5种,分别命名为简单二分型构式、复杂二分型构式、带则构式、多段流程型构式和无效构式,如表1所示。

**表1** 中医古籍中包含的构式类型

**Table 1** The types of constructions contained in traditional Chinese medical texts

| 构式类型 | 描述 | 例句 | 标签 |
|---|---|---|---|
| 简单二分型 | 仅有一个逗号将句子分为两部分 | 心脉急甚,为瘈瘲 | 1 |
| 复杂二分型 | 句子结构为“总-分”式,逗号数量大于1 | 皮寒热者,不可附席,毛发焦,鼻槁腊,不得汗,取之三阳之络 | 2 |
| 带则构式 | 以连词“则”为关键分割点,将句中成分分割为含义不同的两部分 | 邪在肺则皮肤痛,寒热,上气喘,汗出,欬动肩背 | 3 |
| 多段流程型 | 介词“于”在句中频繁出现,句式整体形成排比句 | 厥阴根于大敦,结于玉英,络于膻中 | 4 |
| 无效构式 | 不包含隐式关系,也没有上述构式的结构特征,不再参与分类后的流程 | 黄帝曰:善哉 | 0 |

8种构式特征与5种构式类型具有一定的对应关系,具体如下。

(1) 是否拥有字符“,”是区分无效构式与其他构式类型的重要特征之一。中医古籍文本中短句占有很大比重,并使用逗号来分隔不同的语义成分,实体间代表显式关系的语义成分也因此被这种分隔形式省略,仅能读取到隐式关系,所以当一个句子中不存在逗号时,往往可以认为其结构过于简略而不存在隐式关系,进而被判定为无效构式,同时该句也将不再参与分类之后的流程。句中逗号数量同样会影响分类结果,例如数量等于1时则必定不是复杂二分型构式;数量大于1则必定不是简单二分型构式。

(2) 分号字符“;”本身体现了一种排比的句型,例如文本在具体描述“五脏”概念的相关内容时,对心、肝、脾、肺、肾的描写会具备完全相同的句式结构并使用分号分隔,因此包含了分号的句子往往也包含了隐式关系。

(3) 字符“者”是复杂二分型构式的判定条件之一。当它的上文出现一个总结式的实体名词,下文出现多个相同类型的实体名词且类型与前文名词的类型不同,这就构成了明显的总-分式结构,即符合复杂二分型构式的结构。隐式关系往往存在总-分式结构中。

(4) 字符“曰”是复杂二分型构式的判定条件之一。当它的上文出现数量词,下文紧跟一个总结式的实体名词,则该实体与后续再出现的实体会构成总-分式结构,而由于存在数量词,句中往往还会存在分号,组成多个符合复杂二分型构式的总-分式结构。

(5) 字符“故”是复杂二分型构式的判定条件之一。当它下文紧跟一个总结式的实体名词,则该实体与后续再出现的实体或前文出现的实体会构成总-分式结构,两种情况的区别在于“故”处于句首还是句尾。“故”处于句首时可能会与“者”同时出现,处于句尾时可能会与“曰”同时出现。

(6) 字符“其”是复杂二分型构式的判定条件之一。该字符所指代的主语必定在句首出现。当有多个“其”出现时,句子中往往还会出现同数量的分号以形成排比句型。在各个被分隔的短句中,“其”及其指代的主语与其他实体会构成总-分式结构。

(7) 字符“于”是多段流程型构式的主要判定条件。当“于”上文紧挨一个动词,下文紧接一个实体名词,则构成了一种流程式的句型,“动词+于+名词”的组合也会在句中多次出现,例如描写人体经脉的走向,那么经脉与其途经的身体部位就会包含隐式关系。

(8) 字符“则”是带则构式的主要判定条件。“则”作为连词在中医古籍中代表一种因果关系,并将代表因与果的实体分隔,由于连词本身并无实际语义,因此两个实体间的关系就是隐式因果关系。

确定以上构式特征后,须对这些特征进行数字化处理,以实现对文本的特征提取,如算法1所示。

**算法1** 中医古籍文本特征提取算法。

输入:原始文本中以句号为结尾的一条语句$S$。

输出:该句的构式特征数组 $a$,其中数组 $a=[num1, num2, num3, num4, num5, num6, num7, num8]$。

首先将数组 $a$ 中所有元素初始化为 0,

FOR $x$ in $S$://从句子第一个字符开始依次读取和判断,直至句尾最后的句号结束,$x$ 表示当前被实施判断的字符,

IF $x$ ==“则”:$num1$ += 1;//统计“则”数量

IF $x$ ==“;”:$num2$ += 1;//统计“;”数量

IF $x$ ==“,”:$num3$ += 1;//统计“,”数量

IF $x$ ==“曰”:$num4$ += 1;//统计“曰”数量

IF $x$ ==“者”:$num5$ += 1;//统计“者”数量

IF $x$ ==“其”:$num6$ += 1;//统计“其”数量

IF $x$ ==“故”:$num7$ += 1;//统计“故”数量

IF $x$ ==“于”:$num8$ += 1;//统计“于”数量

IF $x$ ==“。”:break　//当遍历到句号则结束当前句的读取。

经由算法 1 可得到一句话的构式特征数组,当获取到全文所有句子的构式特征数组后,输入 CART 分类模型中进行训练,可实现将全部文本划分为 5 种构式类型,完成文本分类。

## 3　基于依存句法分析的隐式关系抽取

依存句法分析是针对给定的句子序列,应用某一依存语法体系对自然语言进行自动分析,构建句子对应的依存句法树的一种方法[10]。这种句法树描述各个词语之间的依存关系,指出词语之间在句法上的搭配关系,这种搭配关系是和语义相关联的[11]。因此可以理解为隐式关系是实体间语义关联的体现,通过分析特定实体对的搭配关系就可以确定其中的隐式关系。

### 3.1　构建句法树

本文使用了哈工大社会计算与信息检索研究中心研制的语言技术平台(language technology platform,LTP)[12]中的依存句法分析模块构建句法树。该模块集合了分词、词性标注、命名实体识别和依存句法分析 4 项功能,为了减小误差传递,构建出合理的中医古文句法树,实验引入已完成实体标注的语料,并通过调整代码以实现跳过模块中的分词、词性标注、命名实体识别功能,仅保留依存句法分析功能的效果。依存句法分析过程中涉及的依存关系标签如表 2 所示。

### 3.2　隐式关系抽取策略

本文关注于中医古籍中的医学类知识,不考虑其他类型知识之间的隐式关系,因此本文结合中医古籍文本的数据特点,归纳出隐式关系抽取过程涉及 5 类实体,类型名称分别为体征、疾病、症状、病因和生理。本文仅抽取一条语句内部实体之间的关系,不考虑分别属于两个句子的实体之间是否有包含关系。在此基础上总结出包含隐式关系的实体对组合形式,其语句所属构式类型以及相应的隐式关系如表 3 所示。

**表 2**　依存关系标签

**Table 2**　Dependency tags

| 关系类型 | 标签 | 例句 |
|---|---|---|
| 关系类型 | SBV | 心脉急甚,为瘈瘲(心脉急甚<-为) |
| 动宾关系 | VOB | 取之三阳之络(取->三阳之络) |
| 间宾关系 | IOB | 取之三阳之络(取->之) |
| 动补结构 | CMP | 厥阴根于大敦(根->于) |
| 并列关系 | COO | 取腋与膺(腋->膺) |
| 介宾关系 | POB | 厥阴根于大敦(于->大敦) |
| 左附加关系 | LAD | 诸急者多寒(诸<-急) |
| 右附加关系 | RAD | 诸急者多寒(急->者) |
| 独立结构 | IS | 两个单句在结构上彼此独立 |
| 核心关系 | HED | 指整个句子的核心 |

**表 3**　实体对、构式类型和隐式关系对应表

**Table 3**　Correspondence table of entity pairs, conformational types and implicit relations

| 实体对 | 构式类型标签 | 隐式关系 |
|---|---|---|
| 体征-疾病 | 1 | 诊断为 |
| 体征-体征 | 1 | 推测 |
| 疾病-生理 | 1 | 发病于 |
| 疾病-疾病 | 1 | 包含 |
| 疾病-症状 | 2 | 病症 |
| 疾病-生理 | 2 | 治位 |
| 病因-症状 | 3 | 引发 |
| 体征-症状 | 3 | 致使 |
| 生理-生理 | 4 | 途经 |

从表 3 中可以看出,部分实体对对应一个以上的构式类型标签,使得隐式关系应当由构式类型和实体对组合形式两者共同决定,例如“疾病-生理”实体对在简单二分型构式和复杂二分型构式中分别包含“发病于”和“治位”两种不同的隐式关系。因此本文分别针对 4 类构式,结合依存句法分析树的结构制定相应的隐式关系抽取规则。依存句法分析过程中涉及的词性和实体标签如表 4 所示。

**例 1**　“痈发于嗌中,名曰猛疽。”的词性标注和依存句法分析如图 1 所示。

在例 1 中句子仅由一个逗号分隔为两部分,前半部分包含疾病实体“痈”和生理实体“嗌中”,阐述

**表 4** 词性及实体标注标签

**Table 4** Lexical and entity labeling tags

| 标签名 | 含义 | 标签名 | 含义 |
|---|---|---|---|
| n/TZ | 名词/体征 | v | 动词 |
| n/JB | 名词/疾病 | r | 代词 |
| n/ZZ | 名词/症状 | c | 连词 |
| n/BY | 名词/病因 | p | 介词 |
| n/SL | 名词/生理 | wp | 标点 |

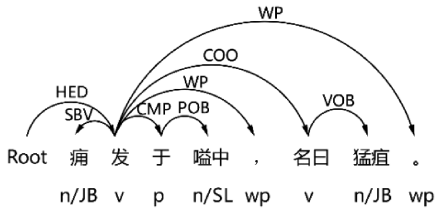


**图 1** 例 1 的词性标注和依存句法分析图

**Figure 1** Lexical annotation and dependency syntactic analysis diagram of example 1

了痈病发作的部位为嗌中;后半部分包含疾病实体“猛疽”,阐述了在前半句的情况下的疾病命名为“猛疽”。句中虽然包含字符“曰”,但其结构不为总-分式,按照表 1 中的定义符合简单二分型构式的特征。在图 1 中的句法树结构中“痈”与该句的 HED 成分“发”发生了 SBV 主谓关系,“发”与“名曰”发生了 COO 并列关系,“名曰”与“猛疽”发生了 VOB 动宾关系,由于动词间存在并列关系,使得“痈”与“猛疽”并未产生直接的语义关联;“嗌中”与“于”发生了 POB 介宾关系,使得“嗌中”与“猛疽”成为不同的动词宾语,同样无法产生直接的语义关联。然而根据语义可知,“痈”是一种涵盖范围较广的疾病名,当它在“嗌中”发作时,便有了更为具体的“猛疽”的病名,当其在其他部位发作时,也会有其他病名与之相对应,因此句中包含了隐式关系“包含”“发病于”及其三元组〈痈,包含,猛疽〉〈猛疽,发病于,嗌中〉,而“痈”与“嗌中”则不包含上述关系,不能仅依靠实体标签选取实体来填充三元组。综上所述,为了获得以上两种隐式关系的三元组,制定以下规则。

**规则 1** 抽取位于 SBV 边的终止节点中成分 A 和位于 VOB 边的终止节点中成分 B,如果 A、B 均为携带标签的名词且标签类型均为“n/JB”,则生成〈A,包含,B〉三元组。

**规则 2** 当规则 1 中的步骤全部完成后,若存在标签类型为“n/SL”的成分 C,且 B 与 C 不为同一动词的宾语,则生成〈B,发病于,C〉三元组。

**规则 3** 当句中存在两种语义成分属于同一 IS 边的初始、终止两个节点,如果此两种成分均为携带标签的名词且标签类型均为“n/TZ”,则生成〈初始节点的实体 A,推测,终止节点的实体 B〉三元组。

**规则 4** 抽取位于 SBV 边的终止节点的成分 A 和位于 VOB 边的终止节点的成分 B,如果 A、B 均为携带标签的名词且 A 标签类型为“n/TZ”、B 标签类型为“n/JB”,则生成〈A,诊断为,B〉三元组。

**规则 5** 当句中存在两种语义成分属于同一 IS 边的初始、终止两个节点,如果此两种成分均为携带标签的名词且处于初始节点的成分标签类型为“n/JB”、处于终止节点的成分标签类型为“n/ZZ”,则生成〈初始节点的实体 A,病症,终止节点的实体 B〉三元组,若还存在实体 C 与 B 同属于 COO 边,则生成〈A,病症,C〉三元组。

**规则 6** 当规则 5 中的步骤全部完成后,若 A 存在 SBV 边、拥有 VOB 边的实体 D,且与 A 的 SBV 边存在公共节点,则生成〈A,治位,D〉三元组。

**规则 7** 当句中所有 IS 边拥有一个公共节点且该节点属于 HED 边,则将位于该节点上文且有 IS 关系的节点作为实体 A,将位于该节点下文且有 IS 关系的节点依顺序作为实体 B,生成等同于 B 个数的三元组。当 A 标签为“n/BY”则三元组形式为〈A,引发,B〉;当 A 标签为“n/TZ”则三元组形式为〈A,致使,B〉。

**规则 8** 抽取位于 SBV 边的终止节点的成分 A 和位于 POB 边的终止节点的成分 B,如果 A、B 均为携带标签的名词且标签类型均为“n/SL”,则生成〈A,途经,B〉三元组。若存在标签类型为“n/SL”的实体 C 且处于不同于 B 的 POB 边的终止节点,则生成〈A,途经,C〉三元组。

**例 2** “白色小理者,肺小。”的词性标注和依存句法分析如图 2 所示。

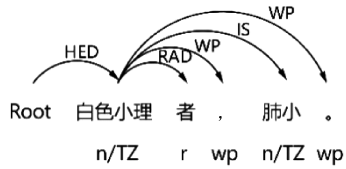


**图 2** 例 2 的词性标注和依存句法分析图

**Figure 2** Lexical annotation and dependency syntactic analysis diagram of example 2

在例 2 中句子仅由一个逗号分隔为两部分,前半部分包含体征实体“白色小理”,阐述了人的肤色偏白、皮肤纹理细小的外在体征;后半部分包含体征实体“肺小”,阐述了一种肺部体积小的内部体征。句中虽然包含字符“者”,但其结构不为总-分式,按照表 1 中的定义符合简单二分型构式的特征。从语

义上看,当一个人肤色偏白、皮肤纹理细小,就可以推断出此人肺部体积小,这是中医辨证手法“望闻问切”中“望”的方式,即通过外部体征推测人体内脏的状况。然而句中并未出现表示“推测”关系的语义成分,因此内外体征实体间包含的是隐式“推测”关系。同理,文本中还包含通过体征推断疾病的语句,如例 3 所示。

**例 3**　“心脉急甚,为瘈瘲;”的词性标注和依存句法分析如图 3 所示。

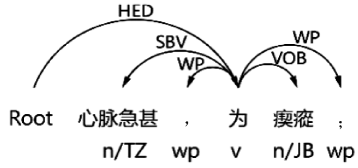


**图 3**　例 3 的词性标注和依存句法分析图

**Figure 3**　Lexical annotation and dependency syntactic analysis diagram of example 3

从结构上看,例 2 中不存在动词,因此把主要实体“白色小理”当作核心成分,而由一个逗号分隔出的实体“肺小”与主要实体发生 IS 独立关系,意味着句中产生“推测”关系的两个体征实体在结构上相互独立。例 3 中存在动词,体征“心脉急甚”与“为”发生了 SBV 主谓关系,“为”与疾病“瘈瘲”发生了 VOB 动宾关系,因此句中产生“推测”关系的体征实体与疾病实体构成了直接的主语与宾语的关系。为了加以区分,本文将体征实体与疾病实体的推测关系命名为“诊断为”关系。为了获得以上两种隐式关系的三元组,本文通过规则 3 抽取“推测”关系,通过规则 4 抽取“诊断为”关系。

**例 4**　“皮寒热者,不可附席,毛发焦,鼻槁腊,不得汗,取之三阳之络。”的词性标注和依存句法分析如图 4 所示。

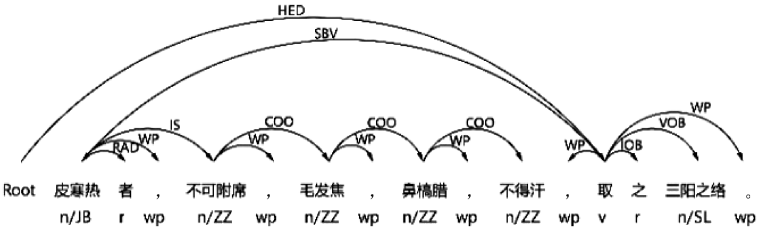


**图 4**　例 4 的词性标注和依存句法分析图

**Figure 4**　Lexical annotation and dependency syntactic analysis diagram of example 4

在例 4 中,句子由 6 个短句组合而成,句首由疾病名“皮寒热”总起,后续描述该病包含的“不可附席,毛发焦,鼻槁腊,不得汗”4 种症状,最后介绍该病的治疗部位“三阳之络”,句子整体结构为总-分式,按照表 1 中的定义符合复杂二分型构式的特征。从结构上看,疾病实体“皮寒热”与第一个症状实体“不可附席”间既不存在动词,又没有其他代表病症关系的语义成分,发生 IS 独立关系,包含隐式“病症”关系,各个症状实体由逗号分隔,发生 COO 并列关系,“皮寒热”与“取”发生 SBV 主谓关系,“取”与“三阳之络”发生 VOB 动宾关系,因此疾病实体“皮寒热”与生理实体“三阳之络”是直接的主语和宾语的关系,而从语义得知这是疾病实体与生理实体的“治位”关系,因此本文通过规则 5 抽取“病症”关系,通过规则 6 抽取“治位”关系。

**例 5**　“邪在肺则皮肤痛,寒热,上气喘,汗出,欻动肩背。”的词性标注和依存句法分析如图 5 所示。

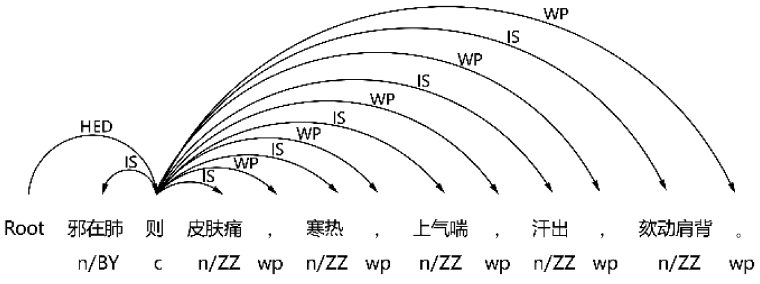


**图 5**　例 5 的词性标注和依存句法分析图

**Figure 5**　Lexical annotation and dependency syntactic analysis diagram of example 5

在例 5 中,句子由字符“则”分割为两部分,“则”之前描述了病因实体“邪在肺”,“则”之后描述了以多个逗号分隔的“皮肤痛”“寒热”“上气喘”“汗出”“欻动肩背”5 个症状实体,可被判定为带则

构式。“则”本身属于连词,并无实际语义,然而当它在“病因+则+症状”的结构中具备一种因果关系的抽象含义,这种含义和病因与症状之间的“致使”关系起到同样的作用,因此在这样的句式中病因实体与症状实体间存在隐式“致使”关系。

**例 6** “肺高,则上气,肩息咳。”的词性标注和依存句法分析如图 6 所示。

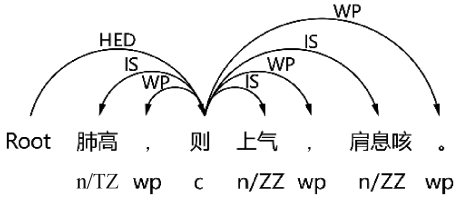


**图 6** 例 6 的词性标注和依存句法分析图

**Figure 6** Lexical annotation and dependency syntactic analysis diagram of example 6

“肺高”作为一种体征,会导致“上气”“肩息咳”症状的出现,句中“则”同样以因果关系的抽象含义代表“致使”关系,因此体征实体与症状实体间也存在隐式“致使”关系,为了加以区分,本文将病因实体与症状实体的致使关系命名为“引发”关系。从结构上看,例 5、例 6 均不包含动词,因此将“则”作为核心,体征、病因、症状三类实体均以独立结构的形式与“则”发生 IS 独立关系,因此“引发”关系和“致使”关系的抽取规则如规则 7 所示。

**例 7** “厥阴根于大敦,结于玉英,络于膻中。”的词性标注和依存句法分析如图 7 所示。

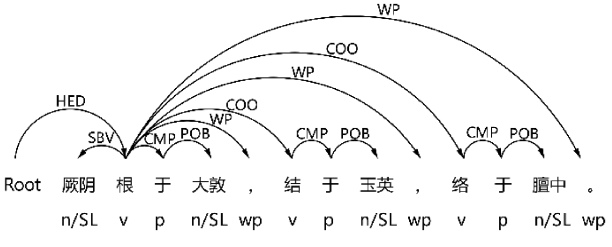


**图 7** 例 7 的词性标注和依存句法分析图

**Figure 7** Lexical annotation and dependency syntactic analysis diagram of example 7

例 7 的句首由生理实体“厥阴”作为唯一主语,句中出现了多个携带介词“于”的动宾短语,涉及的动词间均发生 COO 并列关系,且均与“于”构成 CMP 动补结构,以及存在与“于”发生 POB 介宾关系的生理实体,且 CMP 边与 POB 边的数量相等,因此例 7 属于多段流程型构式。从语义上看,生理实体“厥阴”是人体的一条经脉,其路径途经“大敦”、“玉英”、“膻中”三个代表人体具体部位的生理实体。因此多段流程型构式体现的是人体经脉的路径方向,并且利用“动词+于”的结构代替了表示走向的“途经”关系。本文使用规则 8 来实现抽取“途经”关系。

综上所述,本文通过整合规则 1~8 得到面向中医古籍的隐式关系抽取策略如算法 2 所示。

**算法 2** 基于依存句法分析的隐式关系抽取算法。

输入:① 原始语料中以句号结尾的语句 $S$ 及 LTP 输出的依存句法分析和词性标注结果 $C$;② 经算法 1 和 CART 分类树两个模块先后处理后得到的 $S$ 的构式类型标签 $b$。

输出:三元组 $T$。

```
IF (b==1)//当 S 为简单二分型构式
  IF(C 符合规则 1 的条件)
    使用规则 1 输出 T;
    IF(C 符合规则 2 的条件)
      使用规则 2 追加输出 T;
    END IF
  ELSE IF(C 符合规则 3 的条件)
    使用规则 3 输出 T;
  ELSE IF(C 符合规则 4 的条件)
    使用规则 4 输出 T;
  ELSE 跳过该句;
IF (b==2)//当 S 为复杂二分型构式
  IF(C 符合规则 5 的条件)
    使用规则 5 输出 T;
    IF(C 符合规则 6 的条件)
      使用规则 6 追加输出 T;
    END IF
  ELSE 跳过该句;
IF (b==3)//当 S 为带则构式
  IF(C 符合规则 7 的条件)
    使用规则 7 输出 T;
  ELSE 跳过该句;
IF (b==4)//当 S 为多段流程型构式
  IF(C 符合规则 8 的条件)
    使用规则 8 输出 T;
  ELSE 跳过该句;
```

## 4 实验结果及评测

本文选取经典中医古籍《黄帝内经》作为实验语料,并按照表 4 中的标签进行人工分词、词性标注和实体类型标注。同时,以句号或叹号为结尾的句子视作一条数据的方式对《黄帝内经》全文进行分

句处理,并使用基于构式语法理论的文本分类策略(即算法 1+CART 模型)划分构式类型,再使用基于依存句法分析的隐式关系抽取算法(即算法 2 以及规则 1~8)进行隐式关系抽取,最终得到实验结果如表 5 所示。

**表 5**　隐式关系抽取结果

**Table 5**　Implicit relation extraction results

| 实体对 | 隐式关系 | 三元组数量 |
|---|---|---|
| 体征-疾病 | 诊断为 | 60 |
| 体征-体征 | 推测 | 80 |
| 疾病-生理 | 发病于 | 16 |
| 疾病-疾病 | 包含 | 16 |
| 疾病-症状 | 病症 | 257 |
| 疾病-生理 | 治位 | 68 |
| 病因-症状 | 引发 | 52 |
| 体征-症状 | 致使 | 40 |
| 生理-生理 | 途经 | 357 |

由于隐式关系抽取任务无法采用通用的方法来实现统一建模,对于方法好坏的评测也未产生统一的评价指标,因此本文将从实际应用的角度进行评测。由于抽取出错误的隐式关系会严重影响知识抽取的质量,因此本文关注所提方法能否抽取出正确的隐式关系。

本文以"途经"关系的部分三元组为例,将相关实体、关系存储为 CSV 文件,使用 Cypher 语言的 LOAD CSV 语句将 CSV 文件的数据导入 Neo4j 数据库[13-14]中,利用 Neo4j 图数据库[15]可视化得到图 8 的知识图谱[16-17]。

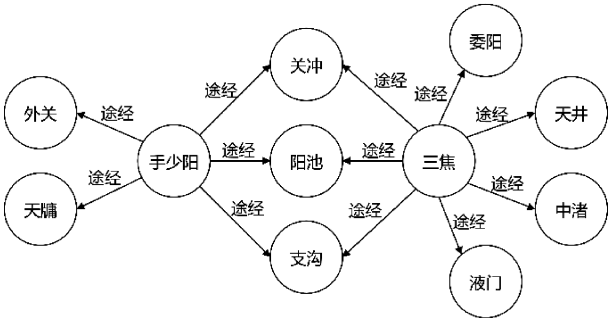


**图 8**　知识图谱示例

**Figure 8**　Example of knowledge graph

"手少阳"及其直系节点来源于《灵枢·根结》,"三焦"及其直系节点来源于《灵枢·本输》,二者所组成的图谱印证了中医学中的专有名词:三焦手少阳脉(《灵枢·经脉》《脉经·三焦手少阳经病证第十一》)。该图谱的正确代表了其中 12 个"途经"关系三元组的正确。

本文邀请三位中医学专家利用以上方式验证知识图谱的正确性,根据专家们的反馈并汇总得到隐式关系抽取结果如表 6 所示。

**表 6**　隐式关系抽取效果

**Table 6**　Implicit relation extraction effects

| 隐式关系 | 三元组总数 | 正确数量 | 正确率/(%) |
|---|---|---|---|
| 诊断为 | 60 | 60 | 100.0 |
| 推测 | 80 | 44 | 55.0 |
| 发病于 | 16 | 16 | 100.0 |
| 包含 | 16 | 16 | 100.0 |
| 病症 | 257 | 243 | 94.6 |
| 治位 | 68 | 64 | 94.1 |
| 引发 | 152 | 137 | 90.1 |
| 致使 | 40 | 40 | 100.0 |
| 途经 | 357 | 325 | 91.0 |
| 合计 | 1 046 | 945 | 90.3 |

从表 6 可知,本文方法在抽取大部分医学隐式关系时正确率较高,说明了本文提取的构式特征能够很好地适配中医古籍文本中的隐式关系在句法结构上具备的特征,由句法树分析制定的抽取规则也能较好地适配蕴含隐式关系的句法结构。而对于体征实体间的"推测"关系,部分语句在语义上具备多重限定条件(即在满足一定条件后才能从一种体征推测另一种体征),但这部分内容并未形成固定的句法结构,未能写成规则的抽取条件,从而导致结果正确率较低。从整体性能上看,本文提出的面向中医古籍的隐式关系抽取方法非常有效。

## 5　结束语

本文通过分析中医古籍的文本特点,结合认知语言学中的构式语法理论和依存句法分析技术,制定了一种面向中医古籍中隐式关系的抽取方法。首先利用构式语法理论分析并寻找构式特征、制定分类策略,在此基础上使用 CART 分类树模型实现文本分类;其次按照文本类别使用依存句法分析技术分别构建句法树,通过制定规则抽取句法树中的特定节点并生成关系三元组,实现隐式关系抽取;最后在《黄帝内经》数据集上进行实验,并利用知识图谱直观地验证实验结果的正确性,验证了方法的有效性。

在未来的工作中将进一步考虑如何为构式语法理论等认知语言学理论建立数学模型,使其能够融入深度学习模型中,构建适用于中医古文、文言文等文体的关系抽取模型,从而更高效地完成隐式关系抽取任务,更好地促进中医古籍中的知识的学习与传承。

## 参考文献:


[1] 王传栋, 徐娇, 张永. 实体关系抽取综述[J]. 计算机工程与应用, 2020, 56(12): 25-36.
WANG C D, XU J, ZHANG Y. Survey of entity relation extraction[J]. Computer engineering and applications, 2020, 56(12): 25-36.

[2] 万常选, 甘丽新, 江腾蛟, 等. 基于协陪义动词的中文隐式实体关系抽取[J]. 计算机学报, 2019, 42(12): 2795-2820.
WAN C X, GAN L X, JIANG T J, et al. Chinese named entity implicit relation extraction based on company verbs [J]. Chinese journal of computers, 2019, 42(12): 2795-2820.

[3] 缪峰, 王萍, 李太勇. 基于事件动作方向的隐式因果关系抽取方法[J]. 计算机科学, 2022, 49(3): 276-280.
MIU F, WANG P, LI T Y. Implicit causality extraction method based on event action direction[J]. Computer science, 2022, 49(3): 276-280.

[4] YU X F, LAM W. An integrated probabilistic and logic approach to encyclopedia relation extraction with multiple features[C]//Proceedings of the 22nd International Conference on Computational Linguistics. Stroudsburg: ACL Press, 2008: 1065-1072.

[5] FEHRI H, HADDAR K, HAMADOU A B. Recognition and translation of Arabic named entities with NOOJ using a new representation model[C]//Proceedingsofthe 9th International Workshop on Finite State Methods and Natural Language Processing. Stroudsburg: ACL Press, 2011: 134-142.

[6] GOLDBERG A E. Constructions: a new theoretical approach to language[J]. Trends in cognitive sciences, 2003, 7(5): 219-224.

[7] DIEWALD G, POLITT K. Grammatical categories as paradigms in construction grammar[J]. The wealth and breadth of construction-based research, 2020, 34: 42-51.

[8] PARK C. Metonymy in grammar[J]. Functions of language, 2013, 20(1): 31-63.

[9] 孙琴, 李昌响, 晋永, 等.《黄帝内经·素问》中虚词英译规律探析[J]. 辽宁中医药大学学报, 2012, 14(6): 125-128.
SUN Q, LI C X, JIN Y, et al. Preliminary study on the principles of English translation of function words in huangdi's inner Classic·Plain questions[J]. Journal of Liaoning university of traditional Chinese medicine, 2012, 14(6): 125-128.

[10] 石翠. 依存句法分析研究综述[J]. 智能计算机与应用, 2013, 3(6): 47-49.
SHI C. Dependency parsing research[J]. Intelligent computer and applications, 2013, 3(6): 47-49.

[11] CHERNYSHOV A, KLIMOV V, BALANDINA A, et al. The application of transformer model architecture for the dependency parsing task[J]. Procedia computer science, 2021, 190: 142-145.

[12] CHE W X, LI Z H, LIU T. LTP: A Chinese language technology platform[EB/OL]. (2011-07-12)[2022-10-15]. https://aclanthology.org/C10-3004.

[13] TUCK D. A cancer graph: a lung cancer property graph database in Neo4j[J]. BMC research notes, 2022, 15(1): 45.

[14] 陈善达, 夏帅帅, 邓文祥, 等. 基于Neo4j的冠心病中医辨证论治知识图谱研究[J]. 中国医药导报, 2021, 18(21): 138-141.
CHEN S D, XIA S S, DENG W X, et al. Study on the knowledge map of Chinese medicine syndrome differentiation and treatment of coronary heart disease based on Neo4j[J]. China medical herald, 2021, 18(21): 138-141.

[15] ZAMINI M, REZA H, RABIEI M. A review of knowledge graph completion[J]. Information, 2022, 13(8): 396.

[16] 王昊奋, 漆桂林, 陈华钧. 知识图谱: 方法、实践与应用[M]. 北京: 电子工业出版社, 2019.
WANG H F, QI G L, CHEN H J. Knowledge graph [M]. Beijing: Publishing House of Electronics Industry, 2019.

[17] 昝红英, 窦华溢, 贾玉祥, 等. 基于多来源文本的中文医学知识图谱的构建[J]. 郑州大学学报(理学版), 2020, 52(2): 45-51.
ZAN H Y, DOU H Y, JIA Y X, et al. Construction of Chinese medical knowledge graph based on multi-source corpus[J]. Journal of Zhengzhou university (natural science edition), 2020, 52(2): 45-51.