基于抽象语义表示的汉语疑问句的标注与分析

闫培艺[1]，李斌[1]，黄彤[1]，霍凯蕊[1]，陈瑾[1]，曲维光[2]
1. 南京师范大学　文学院，江苏　南京
2. 南京师范大学　计算机科学与技术学院，江苏　南京
ypyheta@gmail.com; libin.njnu@gmail.com; iwanttardis@163.com;
kairui.huo.nj@gmail.com, chenn_jin@163.com; wgqu@njnu.edu.cn

摘要

疑问句的句法语义分析在搜索引擎、信息抽取和问答系统等领域有着广泛的应用。计算语言学多采取问句分类和句法分析相结合的方式来处理疑问句，精度和效率还不理想。而疑问句的语言学研究成果丰富，比如疑问句的结构类型、疑问焦点和疑问代词的非疑问用法等，但缺乏系统的形式化表示。本文致力于解决这一难题，采用基于图结构的汉语句子语义的整体表示方法—中文抽象语义表示（CAMR）来标注疑问句的语义结构，将疑问焦点和整句语义一体化表示出来。然后选取了宾州中文树库CTB8.0网络媒体语料、小学语文教材以及《小王子》中文译本的2万句语料中共计2,071句疑问句，统计了疑问句的主要特点。统计表明，各种疑问代词都可以通过疑问概念amr-unknown和语义关系的组合来表示，能够完整地表示出疑问句的关键信息、疑问焦点和语义结构。最后，根据疑问代词所关联的语义关系，统计了疑问焦点的概率分布，其中原因、修饰语和受事的占比最高分别占26.53%、16.73%以及16.44%。基于抽象语义表示的疑问句标注与分析可以为汉语疑问句研究提供基础理论与资源。

关键词： 疑问句；抽象语义表示；语义角色；中文信息处理

Chinese Interrogative Sentences Annotation and Analysis Based on the Abstract Meaning Representation

Peiyi Yan[1], Bin Li[1], Tong Huang[1], Kairui Huo[1], Jin Chen[1], Weiguang Qu[2]
1. School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu
2. School of Computer Science and Technology, Nanjing Normal University, Nanjing, Jiangsu
ypyheta@gmail.com; libin.njnu@gmail.com; iwanttardis@163.com
kairui.huo.nj@gmail.com; chenn_jin@163.com; wgqu@njnu.edu.cn

Abstract

The syntactic and semantic analysis of interrogative sentences has a wide application in the fields of search engines, information extraction and question answering systems. The NLP systems usually use a combination of classification and syntactic analysis to process interrogative sentences, with poor accuracy and efficiency. The interrogative sentence has rich linguistic research results, such as interrogative sentence structure types, etc., but it lacks systematic formal representation. We use Chinese Abstract Semantic Representation (CAMR) based on graph structure to annotate. The data comes from Penn Chinese Treebank 8.0, Chinese textbooks for elementary schools, and the Chinese translation of *Little Prince*, for a total of 2071 sentences. All kinds of interrogative words are represented by the combination of the interrogative concept—amr-unknown and the semantic relationship, which can represent the key information of

the interrogative sentence, the question focus and the semantic structure of the interrogative sentence. Finally, we calculate the probability distribution of the focus, of which the cause, modifier, and argument accounted for the highest proportion, respectively accounting for 26.53%, 16.73%, and 16.44%. Interrogative sentences annotating and analysis based on abstract semantic representation provides a better theory and resources for the study of Chinese interrogative sentences.

Keywords: interrogative sentences , abstract meaning representation , semantic roles , Chinese information processing

1 引言

随着人工智能的发展，自动问答(Sankar et al., 2019)、对话机器人(冯升, 2014)等系统成为了研究热门(Sankar et al., 2019)，其中疑问句的自动理解是自然语言处理中一项非常基础而复杂的任务。而现阶段疑问句的自动分析则主要采用问句分类(Madabushi et al., 2016)、句型识别(Maredia et al., 2017)、疑问焦点语义角色标注(彭洪保 等, 2009)等方法，精度和效率不理想。同时，随着聊天机器人(Hancock et al., 2019)、智能问答(Fan et al., 2019)等系统的发展，疑问句的自动分析越来越重要，这就需要从整体结构上把握疑问句的语义，为自动句法分析提供基础。

然而，传统的疑问句分析存在三个问题。首先，疑问句表示需要将问句分类和依存分析分别进行建模计算后再进行组合，效率较为低下。其次，现有问句分类方法难以解决一句多问的情况。例如图7例句“谁知道怎么赢？”是特指疑问句且拥有两个疑问焦点，当下分析方法难以清楚表示此类疑问句结构。最后，目前标注体系缺乏对省略、指代消解、小句关系等语言现象的有效表示方法，因此难以完整地表示疑问句的语义结构。

在语言学领域，疑问句相关研究集中在疑问句的结构类型等方面。而汉语疑问句以其结构复杂、形式多样等特点备受关注，如邵敬敏 (1996)、闫亚平 (2019)、赵睿艺 (2019)等，但是在形式化表示方面的研究较少，对计算没有直接帮助。

因此，本文尝试通过一种新的语义表示方法——抽象语义表示（Abstract Meaning Representation，AMR）来描写汉语疑问句，解决疑问句的疑问焦点、疑问结构、省略等问题，形成一个完整的疑问句语义表示体系，来服务于汉语疑问句理论和自动分析研究。本文通过2000多句真实语料的标注，测试了抽象语义表示的形式化表征能力，并统计分析出疑问句在疑问焦点和疑问结构上的分布特点。

全文结构如下：第1节梳理了疑问句的理论以及形式化表示的研究脉络。第2节总结了使用抽象语义表示标注汉语各类疑问句的特点，介绍了数据来源和标注方法。第3节统计了疑问概念标签amr-unknown的语义关系，分析了疑问代词的语义功能特点。第4节是结论和未来工作。

2 相关工作

疑问句是人们在日常生活中经常使用的一种句型，也是问答系统、搜索引擎、信息抽取领域中的主要使用句型。从传统语法时期就受到国内外语言学界的关注，相关研究不断进行。

2.1 疑问句的理论研究

传统语法时期，疑问句的研究主要围绕分类和表达效果展开，如Curme et al. (1931)、Jespersen (1933)。从语法角度根据表层结构将其分为一般、特殊、选择以及附加疑问句，认为疑问句除了表示询问等情感外，还有寒暄等语用含义。这些研究以描写为主，虽比如Nesfield (1911)也提到了变换（transformation），但未能触及到疑问句在句法语义层面的内在规律。此时期值得一提的是疑问代词，其研究成果较多，主要集中在指示代词和疑问代词的对比分析方面(Diessel, 2003)。结构主义语言学强调句子在语法研究中的重要性。布拉格学派提出了主位的概念，认为主位是一个句子的话题。主位的提出和疑问焦点的相关理论在某种程度上是一样的。Vachek et al. (1968)还提出了标记性（markedness）理论，最开始用来分析音位的区别性特征，后来人们也用来分析疑问句标记。

以Chomsky为代表的生成语法学派最有代表性的成果是对疑问句语序生成机制的分析。英语疑问句通常把系动词、助动词及疑问词置于句首，这和汉语保持原位不一样。生成学派将小

句的根设置为一个CP，英语助动词和疑问词在疑问句中从原位移入CP的C位；而在肯定句中，这个C由that充当。Chomsky (1973)针对特殊疑问句提出了wh-移位说，但该学派只关注句法层面疑问句的生成机制，不关注语义层面的表示。Baker (1970)认为疑问句在本质上是在生成时包含了一个疑问成分[+Q]。系统功能语法认为言语功能通过语气选择体现在合乎语法规律的小句中。Halliday et al. (2014)认为对一个语言项目进行分类时，应该按照精密度的阶，由一般逐步趋向特殊，对每一个选择点上的可选项给以近似值。

国内疑问句的研究历来属于语气范畴。马建忠 (2010)把语气分为传信和传疑。陆俭明 (1982)标志着疑问句的研究从宏观分类转向微观描写。吕叔湘 (1985)把疑问语气分为“询问、反诘、测度”三种，并将疑问句分为特指问和是非问两类，对疑问句的形式与功能关系等进行了讨论。在疑问句分类方面，王力 (1985)把疑问句分为：叙述句、描写句和判断句。黄伯荣 (1985)提出疑问句类型有特指问、是非问、正反问和选择问四类。邵敬敏 (1996)第一次将语法、语义、语用三个平面的理论运用到汉语疑问句的研究中，标志着汉语疑问句研究进入了新阶段。此后，疑问句理论研究成果也越来越多。在疑问代词方面，黎锦熙 (1992)认为有些疑问代词有“不定称”和“虚指”的用法，还有邵敬敏 等 (1989)、刘月华 (1985)等文的研究。

通过对国内外疑问句理论研究的梳理，可看出国外侧重于通过疑问句的形式探究疑问句本质，不断完善其生成机制。国内虽对疑问句进行了细致描写，比如分类体系等，这些有助于学科语言教学和句法理论研究，但对于疑问句的语义结构问题涉及较少，未能从整体上刻画疑问句的语义。

2.2 疑问句的形式化表示研究

随着疑问句理论的不断发展，国内外不断有学者尝试对疑问句进行表示，大致分为两类，一类是建立疑问句语料库，确定标注体系，另一类是一般语料库附带对疑问句标注方法的简单说明。

首先是疑问句语料库，国外比较著名的是Clark et al. (2004)从TRC 评测语料中抽取了1171句以what开头的疑问句，主要标注了词性信息。Judge et al. (2006)构建了一个含有4000句疑问句的语料库，数据主要来源于TREC跟踪测试集，以期生成的句法分析树对问答系统有所帮助。Myers (2007)针对法语wh-疑问句中不同句法结构可以表示相同语义的特点，建立了法语疑问句语料库。Mrozinski et al. (2008)提供了一个关于提问“为什么”疑问句的语料库，695句语料来源于维基百科。还使用Amazon Mechanical Turk框架收集了问句的匹配答案。Sidi et al. (2011)构建了马来语疑问知识语料库，以期完善马来语语法和语义规则。

接着是一般语料库中的疑问句标注，宾州树库选取了华尔街日报的真实语料，着重标注了句子中的短语结构和短语功能(Marcus et al., 1993)。布拉格依存树库主要由形态层（morphological level）、句法层(analytical level)和语义层(tectogrammatical level)构成(Alena et al., 2000)。这两个大型语料库数据丰富，但是都没有为疑问句设计系统的表示方案，对其处理较为简单。

国内关于疑问句形式化表示的研究发展比较缓慢，比较著名的是山西大学彭洪保（2010）的基于汉语框架网的疑问句语义角色标注语料库，其语料主要来源于山西旅游景点，共计3011句疑问句。该语料库提出了一种根据疑问句目标词共现率来判别疑问句所属框架的方法。李茹 等 (2009)的小型疑问句语料库包含1566句关于旅游景点五台山的疑问句，主要根据焦点进行了疑问句类别统计。

关于疑问句分类体系，国内较为著名的是哈尔滨工业大学的分类体系。文勖 等 (2006)在UIUC(毛先领 等, 2012)的基础上，根据汉语特点将疑问句分为人物、地点、数字、时间、实体、描述、未知七大类，以及根据实际情况又定义了60小类。在一般语料库中，也没有对疑问句的标注方法进行补充说明，比如哈尔滨工业大学依存语料库、清华大学语义依存网络语料库等。下面以哈工大的依存库为例，对“谁想去公园啊？”进行标注示例：

哈工大语义依存分析不像以往简单进行语义角色标注等浅层语义分析，而是通过依存结构将词汇之间的语义关系表示出来。在图1中，Aft表示感事，dCont表示操作的客事，Dir表示趋向，mPunc表示标点标记，mTone表示语气标记。句子的基本架构较为清晰，但对于疑问信息的表示还不够明确。例如，我们需要根据“谁”来确定疑问焦点，但是“谁”也有无疑而问的情况，例如“谁也做不出来。”同时，“啊”的意义也比较多样，仅根据mTone也难以判断其疑问含义。疑问句最重要的就是清楚知道该句到底在问什么，就是我们所说的疑问焦点是什么。该句是特指疑问句，那么疑问代词就是疑问焦点，我们需要将其标注出来，点明其语义关系，才有

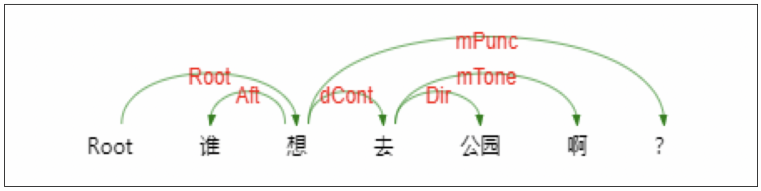

图 1: “谁想去公园啊？”的语义依存树分析

利于计算机的自动分析，上图并没有标识出疑问焦点。再者该分析也忽略了“去”“想”和“谁”的论元共享关系，不利于把握完整的深层语义。

随着自然语言处理的发展，国内外学者越来越重视疑问句的形式表示。国外集中在词性标注等方面；而国内关注分类等研究。总体而言，这些研究对于疑问句整体语义表示研究涉及较少，且研究重点较为分散，不利于计算和自动分析，也不利于系统研究。作为自然语言处理界新兴的句子语义表示方法，抽象语义表示能够更为完整地表示整句的语义结构和疑问结构信息。因此本文将基于抽象语义表示来标注汉语疑问句，系统介绍其标注方法，统计疑问焦点的语义关系等相关信息，以期对疑问句的研究和自动语义分析起到一定作用。

2.3 抽象语义表示研究

抽象语义表示（AMR）是一种新兴的完整的句子语义表示方法。它将句子中的词语抽象为概念，分析概念之间的语义关系，并将这些语义关系抽象为带有语义关系标签的有向弧，把句子语义抽象为一个单根有向无环图(Banarescu et al., 2013)。AMR将句子中词语抽象为概念，用图结构来表示概念以及概念之间的关系，拥有新增、删除、替换的抽象机制(Bos, 2016)。利用这一机制，AMR可突破表层句法结构的差异，将深层的语义结构统一表示出来。

AMR为英语制定，李斌 等 (2017)针对汉语特有的语法特点完善标注体系，形成了中文抽象语义表示（CAMR）。在CAMR标注体系中概念的编号不再由标注器随机分配，而是先对句子进行分词，根据词语序列分配相应编号。下面以“谁想去公园啊？”为例，对改进后的CAMR标注方法进行简要展示。

谁 [1] 想 [2] 去 [3] 公园 [4] 啊 [5]？ [6]	谓词信息
x2/想-01	想:01
:arg0 x1/amr-unknown	arg0:people described
:arg1 x3/去-01	arg1:thoughts of arg0
:arg1 x4/公园	
:arg0 x1/amr-unknown	
:mode x5_x6/interrogative	

图 2: “谁想去公园啊？”的CAMR表示

“谁”在该特指疑问句中是**疑问焦点**，是理解语义的关键，用**核心语义关系arg0（原型施事）和疑问概念amr-unknown**共同来表示，并且使用**关系mode和概念interrogative**点明了疑问语气类型。相对于哈工大的依存库来说，CAMR兼顾了“想-01”、“去-01”和“谁”的论元共享关系，语义结构表示较为完整。并且分词对应编号实现了语义图中的概念与原句词语的对齐。

自2013年标注规范公开发布以来，AMR语料标注工作不断推进。目前AMR已经有近五万句的英文语料库，语料内容来自新闻等领域。CAMR也公布了中文《小王子》语料库，还有向LDC提交的10000句对齐版的标注语料[1]，语料内容除CTB8.0外，还兼顾语文课本、微博等领域的数据。在自动分析方面，F值达到了61%(吴泰中 等, 2019)。本文主要基于CAMR对汉语疑问句进行标注。

[1]https://catalog.ldc.upenn.edu/LDC2019T07

3 数据来源及标注

3.1 数据来源

本文语料主要是从已经标注过的语料中抽取出来的疑问句：来源一是CTB8.0版的10149句网络媒体语料，其中疑问句1215句；二是2001年人教版一到六年级的语文课本中的8696句语料(戴玉玲 等, 2020)，其中疑问句692句，三是和英文《小王子》句对齐的中文小王子1563句，其中疑问句164句，共计2017句疑问句。

3.2 CAMR表示疑问句的特点

通过1.2节的梳理，我们可以发现：以往的疑问句形式化表示没有完整的标注体系，研究重点集中在分类和语义角色标注上。如果要理清疑问句的句子语义结构，这些是不够的。

CAMR的标注体系在AMR的基础上，根据汉语特点进行了优化，形成了一套较为完整的疑问句标注方法，具体特色如下：一是**设置虚节点标签**。CAMR使用xn(n∈N)的形式表示虚节点，n是根据输入的原始句子（基于分词结果）序列分配的有序编号。若为人工添加，则由系统随机分配。这样一来就实现了概念、关系与词的对齐。特别地，对于部分形式意义较为凝固的构式成分，CAMR将其整体作为一个谓词标注或只标注其表层义。AMR中的虚节点标签由概念单词的首字母表示，对于首字母相同的概念，不容易区分。二是**标注疑问语气**。语气信息对句子语义影响很大，尤其在书面汉语中。汉语没有严格意义上的形态变化，语气词和语法意义之间是多对多的关系，是否添加标点符号“？”、是否具有语气词等都会使整句的情感和语义发生变化。三是**既可以从整体上理解疑问句的深层语义结构，又能清晰把握疑问焦点的语义关系**。以往的疑问句研究多集中在分类和浅层语义分析上，CAMR允许根据句子语义增删概念节点，允许论元共享，如图2所示。它可以通过图结构清晰而完整地将整句语义表示出来。再加上疑问概念amr-unknown与不同语义关系的搭配使用设置，我们可以清楚地知道句子的疑问焦点是什么、具有什么样的语义关系，以及疑问焦点的对齐信息。

3.3 数据标注

CAMR中的语义关系分为两种：核心和非核心语义角色关系。用形如“argx(x∈[0,4])”的5个标签来表示核心关系，用如“cause（起因）”等48个语义标签来表示非核心语义角色关系。表1列出了CAMR表示疑问句常用的语义关系标签以及含义。

关系标签	含义	关系标签	含义
:arg0	原型施事	:degree	程度
:arg1	原型受事	:location	地点
:arg2	工具等	:manner	方式
:arg3	出发点等	:mod	修饰
:arg4	终点等	:poss	领属
:cause	起因	:quant	数字
:domain	陈述	:time	时间

表 1: 常用语义关系标签以及含义

处理疑问句时，除了常规的标注操作外，需要特别注意的是对疑问语气和疑问代词的处理。上表中的关系标签mode在CAMR中对应imperative祈使、interrogative疑问、expressive感叹和judgement判断四种语气概念，即用mode和interrogative共同表示疑问语气，将其标注在整句的根（root）上，若遇到有多个分句的长句，其中最后一个分句有语气的，语气则标注在此分句的根上。

主要标注对象有标点符号“？”、疑问语气词“吗”等。当句子中只有“？”或者疑问语气词时，疑问语气由“？”或者疑问语气词单独承担；当两者一起出现时，疑问语气由其共同承担。但当一个句子有多种语气时，如“他为什么这样呢！”既有疑问又有感叹，此时由“呢”承担疑问语气，由“！”承担感叹语气，将这两种语气都表示出来。最后，疑问代词“谁”、“什么”等使用概念标签amr-unknown搭配不同的语义关系标签来表示。

本文的疑问句标注借鉴现代汉语传统的分类体系(邵敬敏, 1996)——将疑问句分为是非疑问句、选择疑问句（包含正反疑问句）和特指疑问句三大类，同时也兼顾了其他一些特殊的疑问句结构，各类疑问句使用的主要关系及概念标签如表2。

类别	关系标签	概念标签
是非疑问句	:mode（语气）	interrogative（疑问）
选择疑问句（包含正反疑问句）	:mode（语气）、:opx（并列/选择）、:polarity（极性）	interrogative（疑问）、or（选择）
特指疑问句	:mode（语气）	interrogative（疑问）、amr-unknown（疑问代词）

表 2: 各类疑问句的基本关系及概念标签

3.3.1 是非疑问句

对于是非疑问句，CAMR使用关系标签mode和表示疑问的概念标签interrogative共同描写句子的疑问语气。

男孩 [1] 被 [2] 找到 [3] 了 [4] 吗 [5] ？ [6]
x3/找到-01
:arg0(x2/被) x8/person
:arg1 x3/男孩
:aspect x4/了
:mode x5_x6/interrogative

图 3: “男孩被找到了吗”的CAMR表示

图3例子中“？”和“吗”一起承担了疑问语气，用“_”连接分词编号。“被找到”表示被动，因此增加了虚节点person来引出“找到”的行为施事，其标签编号由系统随机分配。再者，增加了词语和概念关系的对齐信息，使得虚词对应于概念节点或节点之间的关系弧上(Li et al., 2019)，“被”字引出施事，标注在了实词“男孩”和“找到”之间的有向弧上。另外AMR不标注体，CAMR根据汉语特点增加了关系标签aspect用于标注助词“着”、“了”等。

另外，是非疑问句中经常出现的“是不是”、“是否”等副词成分，如“他是否收集蝴蝶标本呀？”。这些副词是对事件的真实性进行发问，本质上也属于是非疑问句的范畴。所以CAMR在处理这些成分时，也会将其抽象表示为关系mode和疑问概念interrogative。

3.3.2 选择疑问句

CAMR 在处理选择疑问句时，会将表示选择概念的“或者”“还是”等替换为概念or。同时，搭配关系标签operator x，即opx，一起使用。另外在正反疑问句中，使用关系polarity和概念“-”表示否定概念。

你 [1] 喝 [2] 茶水 [3] **还是** [4] 咖啡 [5] ？ [6]
x2/喝-01
:arg0 x1/你
:arg1 x4/or
:op1 x3/茶水
:op2 x5/咖啡
:mode x6/interrogative

你 [1] 走 [2] 不 [3] 走 [4] 啊 [5] ？ [6]
x10/or
:op1 x2/走-01
:arg0 x1/你
:op2 x4/走-01
:arg0 x1/你
:polarity x3/-
:mode x5_x6 /interrogative

图 4: 选择（包含正反）疑问句的CAMR表示

在图4左例中，“还是”被等价替换为or，关系标签op1和op2对选择项进行了说明。右边例子中的选择项“走”和“不走”属于正反两种情况，将“不走”中的否定项“不”等价替换为“-”。

3.3.3 特指疑问句

在特指疑问句中，会将“什么”、“怎么”等疑问代词抽象为概念amr-unknown。

谁 1 帮 2 了 3 窝（我）4 这么 5 大 6 的 7 忙 8？ 9	你们 1 发现 2 了 3 **谁 4 的 5** 玩具 6？ 7
x2_x8/帮忙-01	x2/发现-01
:aspect x3/了	:arg0 x1/你们
:arg0 x1/amr-unknown	:arg1 x6/玩具
:arg1 x4/我	**:poss(x5/的) x4/amr-unknown**
:degree x6/大	:aspect x3/了
:degree x5/这么	**:mode x7/interrogative**
:mode x9/interrogative	

图 5: 特指疑问句的CAMR表示

图5左例中，“帮忙”是一个离合词，CAMR把“帮”和“忙”连接合并处理，且可将“窝”更正为正确的概念“我”。但是在比如哈工大的语义依存分析体系中，“帮”和“窝（我）”的关系则无法显示出来。在右边的例子中，CAMR使用关系标签poss表示“谁”和“玩具”之间的领属关系，“的”作为语义比较虚的词语，将其标注在“谁”和“玩具”之间的关系上。

3.3.4 其他疑问句的处理

一是“非疑问句+疑问小句”类附加问结构。该结构通常是由一个陈述小句，加逗号（也可不加），最后加上一个“是吧”、“是吗”等疑问小句组成。因为CAMR表示的是句子深层结构的抽象语义，所以语序对其标注没有影响。所以“是吗”等疑问小句本质还是对前面陈述句所表达事实的质疑，如图6左侧例子。

二是“难道”类反问结构。在CAMR中，关系标签mod（modifier），用来表示一般的修饰关系，用来表示衔接上下文的关系词，如“难道”“又”“再”等，如图6右侧例子。

大家 1 找到 2 他 3 了 4，5 **是 6 吗 7**？ 8	**难道** 1 女孩 2 发现 3 他 4 了 5？ 6
x2/找到-01	x3/发现-01
:arg0 x1/大家	:arg0 x2/女孩
:arg1 x4/他	:arg1 x4/他
:aspect x4/了	**:mod x1/难道**
:mode x6_x7_x8/interrogative	:aspect x5/了
	:mode x6 /interrogative

图 6: 附加问和反问类疑问句的CAMR表示

图6中的“是吗”是附着在陈述小句“大家找到他了”上，是对“大家找到他了”这个特定事实的质疑，所以将“是吗？”一起抽象为表示疑问语气的关系mode和概念interrogative，该结构的疑问句语义在本质上与是非疑问句无异。

三是间接问句。疑问短语可以单独成句，也可以作为一个结构成分出现在另一个句子中，通常是充当宾语。疑问短语做宾语有两种类型，一是全句为陈述句，如“你了解这是为什么。”这时宾语已经失去了疑问性质和功能。故不关注该类用法。二是全句为疑问句，如图7左侧例子。

四是自问自答类的设问句。自问和自答是设问句不可分割的一个整体，可以看出发问者其实是无疑而问，如图7右侧例子。采用multi-sentence（多句关系）概念标签来处理多个句子之间的关系，与关系标签sntx（x∈N*）配合使用。

在这一节中，我们对是非、选择（包含正反）、特指这三大类疑问句的标注方法进行了举例说明，同时也对一些特殊的疑问句结构进行了标注展示。CAMR既可以处理常规的疑问句标注，表达出深层的语义结构，也可以较好地表示一些无疑而问等特殊的疑问句表达。

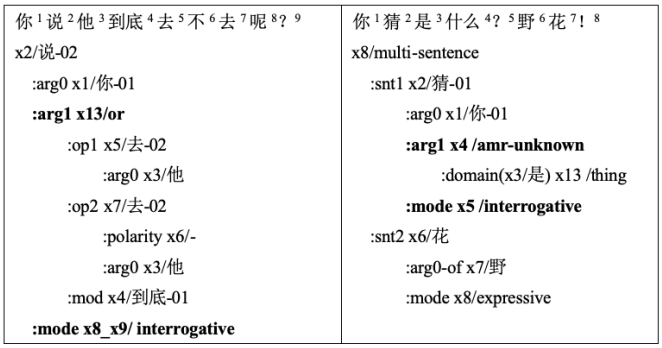

图 7: 间接问句和设问句的CAMR表示

4 统计分析

虽然CAMR无需借助分类系统分析疑问句的语义结构，但我们也可以利用表2相关标签统计出三大类疑问句的占比情况，如表3。从表中可以看出，特指疑问句的占比最高，达51.71%，选择疑问句最少，只有4.73%。

类别	次数/比例
是非疑问句	994/43.56%
选择疑问句（包含正反疑问句）	108/4.73%
特指疑问句	1180/51.71%

表 3: 各类疑问句的比例分布

4.1 特指疑问句的疑问焦点

CAMR允许根据句子语义增删概念节点，允许论元共享，既可以通过图结构清晰而完整地将整个句子深层语义表示出来，又可以通过语义关系和疑问概念amr-unknown搭配使用等把握疑问焦点信息，这对于我们准确理解疑问句非常有帮助。吕叔湘 (1985)指出“回答问话，一般不用全句，只要针对疑问焦点，用一个词或短语就够了”。对于疑问句来说，我们需要清楚的就是疑问句是针对什么提出疑问，疑问语义中心在哪里，即疑问焦点在哪里(唐燕玲 等, 2009)，这对于计算机自动分析是非常重要的。是非疑问句是对整个句子的客观事实提出疑问，那么疑问焦点就落在了整句的语义上；选择疑问句有选择项，那么opx关系标签所对应的概念标签就是我们需要关注的疑问焦点语义项。

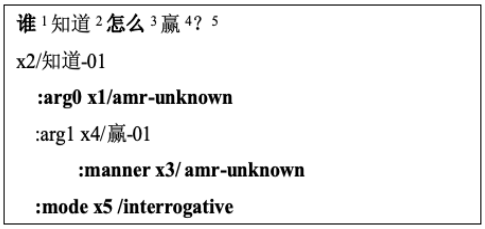

图 8: “谁知道怎么赢？”的CAMR表示

但是特指疑问句比较特殊，具有不一样的构成要素——疑问代词，比如“怎么”、“什么”、“哪里”等。疑问代词作为句法功能和意义的结合，是特指疑问句的疑问焦点(唐燕玲 等, 2009)。林裕文 (1985)也指出“特指是对准疑问代词回答的”。再加上有的特指疑问句不止一个疑问焦点，仅从疑问句分类角度难以准确把握完整的语义信息，如图8所示，该句有“谁”和“怎么”两个疑问焦点，分别具有arg0（原型施事）和manner（方式）两种语义关系，传统计算研究方法难以直接处理。针对特指疑问句要素特点，CAMR使用疑问概念amr-unknown，同时搭

配各种语义关系来共同表示疑问焦点信息。疑问代词的不同使用方法可能会有不同的语义关系，下面将通过统计数据详细分析疑问代词语义角色的分布特点，总结疑问代词的语义功能特点。

4.2 疑问概念amr-unknown的语义关系特点

本文对2071句疑问句中的1410个疑问代词所对应的1410个概念amr-unknown的语义关系信息进行了统计，不同语义关系的使用分布情况如表4。

语义关系（含义）	次数/比例	语义关系（含义）	次数/比例
:cause（起因）	**373/26.53%**	:source（源）	10/0.71%
:mod（修饰）	**236/16.73%**	:purpose（目的）	8/0.57%
:arg1（原型受事）	**232/16.44%**	:degree（程度）	6/0.43%
:arg0（原型施事）	168/11.90%	:value（值）	3/0.21%
:manner（方式）	134/9.50%	:destination（目的地）	2/0.14%
:quant（数字）	58/4.11%	:beneficiary（受益者）	2/0.14%
:arg2（间接宾语、工具等）	52/3.68%	:day（天）	2/0.14%
:opx（选择项）	38/2.70%	:arg3（出发点、收益者等）	1/0.07%
:time（时间）	26/1.83%	:frequency（频率）	1/0.07%
:domain（陈述）	21/1.48%	:direction（方向）	1/0.07%
:poss（领属）	19/1.35%	:topic（话题）	1/0.07%
:location（处所）	16/1.13%	合计	1410/100%

表 4: 疑问概念amr-unknown的语义关系分布

从表4可以看出，在本次统计中，疑问概念amr-unknown各类语义关系有23种，总共出现了1410次，但分布不平衡，使用频率较高的前三大类依次是cause、mod以及arg1，分别用来提问原因、修饰成分以及原型受事，分别占比26.53%、16.73%以及16.44%。在出现的4种核心语义关系中，概念amr-unknown为受事的语义关系最常见。在出现的4种核心语义关系中，概念amr-unknown为受事的语义关系最常见，施事、间接宾语次之。非核心语义关系有19种，种类比较多，且出现总次数是核心语义关系的两倍左右，达67.87%。这些不同的语义关系代表的是说话人不同的提问对象，弄清疑问代词的不同语义关系是什么，是我们把握特指疑问句语义重点所在，也是问答系统提高回答准确率的关键所在。

4.3 小结

通过对2071句疑问句的标注，我们可以看出CAMR可以完整而清晰地表示出汉语疑问句的整体结构。以往处理疑问句的方法，比如问句分类、依存分析等，很难完整表示出疑问句结构的深层语义。通过对1410个疑问概念amr-unknown的语义角色种类进行统计分析，发现cause、mod以及arg1的语义关系使用最为频繁。在CAMR的标注体系下，处理疑问句有一套完整的标注体系，无需设置分类标签，通过语义关系标签就可以知道句子的疑问焦点是什么，位置在哪里，从而准确把握整句的语义结构。

5 结论及未来工作

随着自然语言处理领域的不断发展，其中以问答系统最为突出，疑问句的形式化表示越来越受到各界学者的重视，但是由于汉语疑问句形式多样，结构复杂，目前还没有比较完整的标注体系可以很好地表示汉语疑问句的整体结构。本文首先梳理了国内外疑问句的相关理论与计算研究。接着使用改进之后的CAMR体系针对2071句汉语疑问句，对不同结构类型疑问句的标注方法进行了说明。最后对1410个疑问概念amr-unknown的语义关系种类进行了统计分析，发现其非核心语义角色的使用频率最高。这一标注体系不需要进行疑问句分类，就可以更好地描写疑问代词的功能，把握其语义关系，对问答系统作出正确回答有很大的帮助。

在未来工作中，我们会扩大汉语疑问句的语料规模，丰富语料类型，关注口语化的疑问句表达，进而继续完善CAMR标注体系，推动相关理论研究。最后，希望通过标注语料库进行机器学习，不断提高CAMR语义自动分析效果，推进疑问句的自动分析和应用。

致谢

本文得到以下基金项目的支持：国家社科基金项目（18BYY127）；国家自然科学基金（61772278）；江苏省高校哲学社会科学优秀创新团队建设项目，在此一并感谢。

参考文献

Alena Böhmová, Jan Hajic, Eva Hajicová, Barbora Hladká. The Prague Dependency Treebank: A Three-Level Annotation Scenario[A]. In Treebanks: Building and Using Parsed Corpora[C], Amsterdam: Kluwer, 2000:103-127.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, Michael Auli. ELI5: Long Form Question Answering[J]. arXiv: Computation and Language, 2019: 3558–3567.

Angel Maredia, Kara Schechtman, Sarah Ita Levitan, Julia Hirschberg. Comparing Approaches for Automatic Question Identification.[C]// Proceedings of the 6th Joint Conference on Lexical and Computational Semantics, Vancouver, Canada: Association for Computational Linguistics, 2017: 110-114.

Baker, Carl L. Notes on the Description of English Questions: The Role of an Abstract Question Morpheme[J]. Foundations of language, 1970: 197-219.

Bin Li, Yuan Wen, Li Song, Weiguang Qu, Nianwen Xue. Building a Chinese AMR Bank with Concept and Relation Alignments[J]. Linguistic Issues in Language Technology, 2019, Vol 18.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, Jason Weston. Learning from Dialogue after Deployment: Feed Yourself, Chatbot![C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy: Association for Computational Linguistics, 2019: 3667-3684.

Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, Yoshua Bengio. Do Neural Dialog Systems Use the Conversation History Effectively? An Empirical Study[J]. arXiv: Computation and Language, 2019.

Chomsky, Noam. Conditions on Transformations[A]. Andersen, Stephen and Paul Kiparsky. A Festschrift for Morris Halle[C], New York: Holt, Rinehart and Winston, 1973:232-286.

Curme, George Oliver. A Grammar of the English Language in Three Volumes. Vol. 3. [M]. Berlin：Indogermanische Forschungen,1931.

Diessel, Holger. The Relationship between Demonstratives and Interrogatives[J]. Studies in Language, 2003, Vol 27.3: 635-655.

Fatimah Sidi, Marzanah A. Jabar, Mohd Hasan Selamat, Abdul Azim Abdul Ghani, Md Nasir Sulaiman,Salmi Baharom. Malay Interrogative Knowledge Corpus[J]. American Journal of Economics and Business Administration, 2011, 3(1): 171-176.

Harish Tayyar Madabushi, Mark Lee. High Accuracy Rule-based Question Classification using Question Syntax and Semantics[C]// the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan: The COLING 2016 Organizing Committee, 2016: 1220-1230.

Jespersen, Otto. The System of Grammar[M]. London:G. Allen & Unwin ltd, 1933.

Joanna Mrozinski, Edward W. D. Whittaker, Sadaoki Furui. Collecting a Why-question Corpus for Development and Evaluation of an Automatic QA-system[C]//Proceedings of ACL-08: HLT, Columbus, Ohio: Association for Computational Linguistics, 2008: 443-451.

Johan Bos. Expressive Power of Abstract Meaning Representations[J]. Computational Linguistics, 2016(3): 527–535.

John Judge, Aoife Cahill, Josef van Genabith. Question Bank: Creating a Corpus of Parse-Annotated Questions[C]// Proceedings of the 21st International Conference on Computational

Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia: Association for Computational Linguistics, 2006: 497-504.

Josef Vachek. The Linguistic School of Prague[J]. Journal of the American Oriental Society, 1968: 369.

Banarescu L, Bonial C, Cai S, et al. Abstract Meaning Representation for Sembanking[C]// Linguistic Annotation Workshop and Interoperability with Discourse, Sofia, Bulgaria: Association for Computational Linguistics, 2013:178–186.

Lindsy Lee Myers. WH-interrogatives in Spoken French: A Corpus-based Analysis of their Form and Function[D]. Diss. 2007.

Michael Alexander Kirkwood Haliday. An Introduction to Functional Grammar[M]. London: Edward Arnold, 1985.

Mitchell Marcus, Mary Ann Marcinkiewicz, Beatrice Santorini Building a Large Annotated Corpus of English: the Penn Treebank [J]. Computational Linguistics, 1993: 313-330.

Nesfield, John Collinson. Idiom, Grammar, and Synthesis[M]. London: Macmillan and Co. Ltd, 1929.

Stephen Clark, Mark Steedman, James Curran. Object-Extraction and Question-Parsing Qsing CCG[C]// Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 2004:111-118.

戴玉玲, 戴茹冰, 冯敏萱, 李斌, 曲维光. 基于关系对齐的汉语虚词抽象语义表示与分析[J]. 中文信息学报, 2020, 34(04): 21-29.

冯升. 聊天机器人系统的对话理解研究与开发[D]. 北京邮电大学, 2014.

黄伯荣. 陈述句, 疑问句, 祈使句, 感叹句[M].上海:上海教育出版社,1985.

李斌, 闻媛, 宋丽, 卜丽君, 曲维光, 薛念文. 融合概念对齐信息的中文AMR语料库的构建[J]. 中文信息学报, 2017, 31(06): 93-102.

黎锦熙. 新著国语文法[M]. 北京: 商务印书馆, 1992.

李茹, 王文晶, 梁吉业, 宋小香, 刘海静, 由丽萍. 基于汉语框架网的旅游信息问答系统设计[J]. 中文信息学报, 2009, 23(02): 34-40.

林裕文. 谈疑问句[J]. 中国语文, 1985, (2): 91-98.

陆俭明. 由“非疑问句形式+呢”造成的疑问句[J]. 中国语文, 1982: 640.

刘月华. “怎么”与“为什么”[J]. 语言教学与研究, 1985(04): 130-139.

吕叔湘. 疑问. 否定. 肯定[J]. 中国语文, 1985(4): 274.

马建忠. 马氏文通[M]. 北京:商务印书馆, 2010.

毛先领, 李晓明. 问答系统研究综述[J]. 计算机科学与探索, 2012, 6(3): 193-207.

彭洪保, 李茹, 段建勇. 基于汉语框架网的问句语义角色自动标注研究[C]. 中国计算机语言学研究前沿进展（2007-2009）. 北京: 清华大学出版社, 2009:220-225.

彭洪保. 基于汉语框架网的问句语义角色标注研究[D]. 山西大学, 2010.

邵敬敏. 现代汉语疑问句研究[M]. 上海: 华东师范大学出版社, 1996.

邵敬敏, 赵秀凤. “什么”非疑问用法研究[J]. 语言教学与研究, 1989(1): 26-40.

唐燕玲, 石毓智. 疑问和焦点之关系[J]. 外国语(上海外国语大学学报), 2009, 32(01): 51-57.

王力. 中国现代语法[M]. 北京: 商务印书馆, 1985.

文勖, 张宇, 刘挺. 基于句法结构分析的中文问题分类[J]. 中文信息学报, 2006, 20(2): 33-39.

吴泰中, 顾敏, 周俊生, 曲维光, 李斌, 顾彦慧. 基于转移神经网络的中文AMR解析[J]. 中文信息学报,2019, 33(04): 1-11.

闫亚平. 汉语附加问句句法形式的浮现与发展[J]. 汉语学报, 2019(03): 21-29, 95.

赵睿艺. 现代汉语“疑问代词+V+不是V”构式研究[D]. 华中科技大学, 2019.