

基于抽象语义标注的《左传》语料库构建初探

芦靖雅¹, 陈瑾¹, 张艺璇¹, 常博林¹, 许智星¹, 李斌^{1,2}, 王东波³

(1.南京师范大学文学院, 江苏南京 210097)

(2.南京师范大学语言大数据与计算人文研究中心, 江苏南京)

(3.南京农业大学信息工程学院, 江苏南京 210095)

摘要

抽象语义表示 (Abstract Meaning Representation, AMR) 是一种基于句子的语义表示方法, 其核心是单根有向无环图。因其在语义表示方面的简单便捷的优势进而发展出了中文AMR、韩文AMR、西班牙语AMR等。同时构建了对应的标注库, 相应的评测和解析都取得了不错的成绩。古汉语语料库的建设有分词和词性标注语料库, 但仍缺少古汉语语义库。为此, 本文基于一部优秀的上古汉语文献——《左传》, 提出了一种为古汉语服务的抽象语义表示方法, 制定抽象语义标注规范, 初步建成《左传》1500句抽象语义标注语料库, 以期为古汉语的语义表示和语料库建设提供参考。

关键词: 抽象语义表示; 语义分析; 《左传》; 语料库; 古汉语信息处理

A Preliminary Investigation on construction of ZuoZhuan corpus based on Abstract Meaning Representation

Jingya Lu¹, Jin Chen¹, Yixuan Zhang¹, Bolin Chang¹, Zhixing Xu¹, Bin Li^{1,2}, Dongbo Wang³

(1.School of Chinese Language and Literature, Nanjing Normal University, Nanjing Jiangsu 210097, China)

(2.Center of Language Big Data and Computational Humanities of Nanjing Normal University, Nanjing Jiangsu, China)

(3.Institute of technology, Nanjing Agricultural University, Nanjing Jiangsu 210095, China)

Abstract

Abstract Meaning Representation (AMR) proposes a semantic representation method that is easy to read and process, and its core is a rooted acyclic directed graph. AMR has developed Chinese AMR, Korean AMR, Spanish AMR, etc. because of its advantages in semantic representation. At the same time, the corresponding semantic library is constructed, and the corresponding evaluation and analysis have achieved good results. The construction of the ancient Chinese corpus has a corpus of word segmentation and part-of-speech labeling, but there is still a lack of ancient Chinese semantic library. Therefore, based on an excellent ancient Chinese document - ZuoZhuan, this paper proposes an abstract meaning representation method serving ancient Chinese, establishing a set of rules for abstract semantic annotation, we have built a preliminary corpus of 1,500 sentences from ZuoZhuan, aiming to provide reference for semantic representation and corpus construction in Classical Chinese.

Keywords: Abstract Meaning Representation, Semantic analysis, ZuoZhuan, Corpus, Ancient Chinese Information Processing

项目号: 国家社科基金项目(18BYY127), 古籍工作重点课题 (22GJK006), 江苏省社科基金项目 (20JYB004)。

1 引言

大型人工标注语料库的构建对于推进自然语言处理 (Natural Language Processing, NLP) 技术至关重要。英文宾州树库(Marcus et al., 1993)的建成为许多在句法树库上训练的解析器的存在提供了可能, 这些解析器的准确率约为90%(Charniak and Johnson, 2005)。相比之下, 由于英语中的语义资源较为分散, 语义层面的标注语料库和解析器发展较为缓慢。为了解决这种困境, 抽象语义表示 (Banarescu等人, 2013) 应运而生, 该理论将句子的词语抽象为概念和关系, 将句子结构抽象为单根有向图。此外, AMR产生的图结构能直接为计算机所理解这一优势也使得该理论扩展到了其他语言方面, 例如现代汉语、韩语和西班牙语等。现代汉语方面, 李斌等 (2017) 构建了CAMR的标注规范, 并基于规范对中文《小王子》进行标注, 中英标注一致性达到了0.83的Smatch值(李斌et al., 2017a)。Migueles-Abraira等 (2018) 在用AMR标注规范标注西班牙语时, 发现西班牙语存在七种无法用AMR规则表示的句式, 故采用新增标签或新增标注规则的方法完善针对西班牙语的标注规范, 为建设西班牙语的抽象语义库提供指导(Migueles-Abraira et al., 2018)。韩语AMR构建中, 研究者对韩语中多样的系词提出了新的标注方案(Choe et al., 2019)。

在中文语义建设方面, 不少学者也构建了一些汉语句子语义库。例如山西大学的中文FrameNet(郝晓燕et al., 2007)、清华大学的语义依存树库(尤et al., 2003)等。在国外语义理论的基础上, 实际结合了中文的特点, 建立中文的语义分析方法。在古汉语方面的语料库建设主要包括三个方面: 基础的生文本资源库、词性标注语料库、词义标注语料库。基础的生文本资源库, 如国家语委古代汉语语料库和北京大学CCL语料库中的古代汉语语料库(俞士汶et al., 2002), 都收录了大量的文献语料供检索查找。词性标注语料库, 包括台湾“中研院”古代汉语标注语料库 (“Academia Sinica” Ancient Chinese Corpus, ASACC) 和英国谢菲尔德大学历时汉语语料库 (The Sheffield Corpus of Chinese, SCC)。台湾“中研院”古代汉语标注语料库收录了上古到近古文献共137部, 采用多级制的词性标记集, 对文献进行词性标记处理(Huang and Chen, 1992)。英国谢菲尔德大学历时汉语语料库收录了上古到近古文献共40部, 设计了Word Class、Category、Tag Label的三层标签结构, 对古汉语中的文本进行了词性标注(Hu et al., 2007)。词义标注语料库, 北京语言大学的古汉语词义标注语料库收录整理了200多个常用的古汉语单音节词语, 并设计了古汉语多义词的词义划分原则, 对多义词所在的语料进行了词义标注(舒蕾et al., 2022)。古汉语语料库的建设探索从最基础的生文本语料库发展到词义标注语料库, 在构建方法和语料处理方面都有很大的发展, 古代汉语信息处理也在此基础上逐渐由文本识别发展到词义消歧等方面。但句子级别语义语料库的缺乏, 不利于古汉语句法语义在自然信息处理方面的发展, 延缓了古籍文献信息处理发展的脚步。

在CAMR语料库和标注规范的基础上, 本文探索了《左传》的AMR表示方式。由于CAMR是在现代汉语的基础上完成标注的, 虽然其中有涉及谚语、俗语等, 但其标注规范并未完全适配上古汉语的标注。古汉语句子级别的语义库需要在对古汉语文本进行语义标注的基础上构建, 和现代汉语不同, 古代汉语中词汇量小, 句式多变。变化的句式是否会对语义产生影响, 又该如何标注相同语义的不同句式, 需要在标注中提出一个合理的解决方案。句子中各个结构的语义关系并不是单一的, 并存的几种语义也有着主次之分(袁本良, 2003)。《左传隐公元年》中“公赐之食”这一例, 既包含动词“赐”的及物性关系, 也包含“公”和“食”的领属关系。在进行语义标注时, 选择能体现句子语义表达的方法即为重要。基于构建古汉语句子级别语义库的目的, 本文将会标注《左传》文本中的句子的语义, 讨论并解决标注过程中发现的问题, 并调整修改标注细则, 最终形成《左传》的标注规范。未来还会形成一个较大规模的《左传》抽象语义标注语料库, 希望能够促进古汉语在抽象语义方面的研究。

2 抽象语义语料库建设方式

2.1 语料选取原则

上古汉语是指殷商时期到五胡乱华以前的古汉语, 在语音、语法和词汇上有独特的特点。《左传》是中国历史上第一部编年体史书, 是儒家的经典著作之一, 也是为历代儒家学子所重点研习的著作之一。《左传》主要记载了周朝王室衰微的过程, 以及王朝内的礼仪规范、典章制度、社会风俗、历法时令、神话传说等各类能反映当时当世生活的内容。其所记录的事件既有场面宏大的战争事件, 也有精细的人物描写。描述战争事件的语言庄重严谨, 刻画人物细节的语言生动细腻。综合来看, 《左传》中的语言既包含严谨的书面语, 也包括生动活泼的口

语，是一部非常具有语言研究价值的文献。对《左传》中的语言现象的研究包括如下几个方面：其一，研究《左传》的语言风格。《左传》的叙述语言丰富多变，其中的外交辞令更是历代学者研究的重点。有学者从语用学角度探讨了简洁的外交辞令背后所反映的“言内之意”和“言外之意”(武惠华, 1994)。其二，研究《左传》中的句法现象。例如《左传》“使字句”研究(李佐丰, 1989)。从句法、语义、语用三个层面探讨《左传》中的省略现象(张景霓, 2000)。

传统语法研究视角下的《左传》语言研究已经有许多成果，但仍缺乏从计算语言学视角出发的研究。所以本文选择杨伯峻《春秋左传注》(中华书局, 1981)版的《左传》作为研究语料，从抽象语义视角揭示《左传》中词与词之间的语义关系，并尝试从中总结规律。

2.2 标注体系

语义的自动分析是计算语言学研究的一大难点，为解决这一难点，早期语言学家们从句法研究中发现语义能更加准确地解释语言现象。于是，特斯尼耶尔提出了依存语法，将句法和语义结合起来进行研究。依存语法将一个句子当作是句子中词汇之间的依存关系网络，在依存语法中每个词都是语法上的核心(Tesnière, 2015)。在分析句子结构时，依存语法更强调个别词汇之间的依存关系，即从属词的语义附加在它所依附的支配词的语义上。例如，“饭被我吃了”“我把饭吃了”“我吃了饭”这三个语序不同但是语义相同的句子，就能够通过依存语法清晰地进行分析。但是依存句法只研究句子中词语之间的依存关系，没有涉及到词语本身的词义，因此不能完整表达句子语义。尽管存在一些局限性，但依存语法可提供一些不同的分析工具和不同的翻译策略，适用于自然语言的语义表达和计算体系等领域。

随后，研究者们相继提出了范畴语法、框架语义学等相关的理论。在依存语法和格语法的发展下，菲尔墨提出了框架语义学(Frame Semantics)。该理论认为框架是认知结构中的一种关系结构，它包括一个中心概念和一组与该概念相关联的知识、事件、情境以及该概念与其他概念之间的关系。在语言中，一个句子的意义可以被看作是一系列的框架，其中每个框架都表示一个事件或情境。每一类动词都具备一定的论元体系，体系中包含若干个论元，分别用ARG0, ARG1...ARGn对论元进行编号(Fillmore and others, 2006)。框架语义学提供了一种描述一个文本中的各种事件或情境的方式。同时，它还能够帮助我们识别并处理文本中的一些隐含信息，比如说暗示和解释。这种理论在机器翻译、问答系统、信息提取和自然语言生成等领域得到了广泛地应用。现代汉语的研究在依存语法等理论的影响下也逐渐深入发展。但在面对灵活多变的中文句式时，由于汉英双语语法的不一致，产生了诸多问题，在依存语法理论中，其投射的方法无法完全应用于汉语句式中。从语言学和句子深层语义理解的角度对汉语中的非投射现象进行归纳解释，并解决其无法投射的现象，对自动语义分析有着重要的作用(郑丽娟 et al., 2014)。

计算语义发展到21世纪时，Banarescu等(2013)学者提出了Abstract Meaning Representation, AMR语义标注体系。AMR将句子中的词语抽象为概念和关系，并以单根有向无环图的形式具象地表示句子的语义结构，结合了PropBank的标注框架集，同时兼顾句法语义和词汇语义，它可以将人类语言表达的意义映射为计算机可以理解的形式。AMR通过预先设定的标注规则对自然语言句子进行编码，其中节点表示概念，边表示概念之间的关系(Banarescu et al., 2013)。其便于计算机理解的图结构、能够跨语言处理的优势，使其在问答系统、信息抽取、机器翻译等方面都有不错的表现。随后，结合中文特点，卜丽君(2017)在借鉴AMR的框架和理论的基础上建立了CAMR语料库，构建了CAMR标记集，并进行了两阶段的语料标注工作。利用CAMR语料库分析汉语语义特点，从语义角度对汉语的结构特征进行总结(卜丽君, 2017)。闻媛(2018)在此基础上解决了AMR语义表示和原句词语未对齐的问题，构建了一个概念关系对齐的CAMR语义库(闻媛, 2018)。

AMR所采用的单根有向无环图能更好地揭示论元共享与回指等语言现象。同时，从英文AMR发展而来的CAMR，同样也证明了AMR在跨语言表达方面的能力，可以对不同语言的句子进行语义表示，为跨语言自然语言处理提供了一种有用的工具。AMR可以用于文本理解、自动摘要、翻译、问答等多个领域。同时，它也可以用于生成自然语言文本，有效提高了自然语言生成的效率。基于AMR的强大理论和应用价值，本文选取了AMR作为语义分析的理论和方法。

2.3 语料预处理

标注《左传》之前，需要将纸质版《左传》中的内容录入到计算机中，并进行批量预处理。

理。预处理包括以下几个内容：

首先，需要删除重复性的内容和注释。《左传》文本正文的每一章节的经文部分是对传文的概括归纳，和传文有重合，将每一章节的经文部分删除，使标注内容更加连贯流畅。而注释是对传文部分的解释说明，形式上不成文，语义上不连贯，所以选择删除注释，主要对《左传》经文部分进行语义标注。这样进行预处理之后的文本内容更加适合本文所要构建的语料库的定位。

其次，根据标注需求，对文本进行分词处理。抽象语义标注的工作是以词为单位进行的，所以需要对话料进行分词处理。依据南京师范大学计算语言研究组所制定的分词标准对纯文本进行处理，最终得到分词后的文本。在分词之后的文本中，词与词之间通过空格来分隔，而标注平台中词之间的区分也是通过空格来判定的，所以该分词之后的文本非常便于导入标注平台进行标注。经过预处理的语料共包括9000个句子，范围从隐公元年到昭公三十二年。

2.4 人工标注平台

现有语料的标注是在李斌等（2017）开发的CAMR专用的融合语义图与原句对齐的一体化人工标注平台CAMR Annokit上进行的(李斌et al., 2017b)。该平台内含有CAMR配套的谓词框架词典、概念标签等，都以表格的方式置于标注平台内。但该平台内的谓词框架词典中的字体全部为简体字，与《左传》文本中的繁体字不相适应；并且由于概念标签是基于现代汉语标注所设计的，所以还需要对平台内的资源进行整理。

首先，将标注平台内的谓词框架词典的表格下载并转化为繁体。其次，补充《左传》中常见的谓词。包括《左传》中常使用但现代汉语中不常用的谓词，以及这类谓词的论元结构。最后，需要对已有的谓词的论元框架进行修改，目的是使该谓词论元更加符合《左传》中的动词语义。以上的工作都是用具有强大功能的通用电子表格Excel软件完成，构建完成之后统一导入到MySQL数据库管理系统中，形成一个新的配套《左传》的词典数据表。数据表部分内容见下表：

序号	词语 ¹	语义编号	论元
1	爱	01	arg0:love giver; arg1:thing, person loved
2	逆	02	arg0:the people to meet someone; arg1:the people be greeted
3	书	01	arg0:agent; arg1:the thing be written
4	军	01	arg0:army; arg1:location
5	言	01	arg0:speaker; arg1:content spoken; arg2:listener

¹ 为清晰展示采用简体，实际的谓词词典中动词都为繁体。

表 1. 《左传》谓词词典

其次，需要调整标注平台中的语义关系标签。在对中文《小王子》标注过程中构建的CAMR中的语义关系标签是基于英文AMR的语义关系标签发展而来的，在保留了基本的表示施事论元的“arg0”、表示受事论元的“arg1”等核心语义关系标签的情况下，还根据中文语言表达特点增加了部分新的非核心角色关系关系(Li et al.,)。其中，中文个体量词“cunit”、句末无意义语气“smood”两个标签都在《左传》标注中被大量使用。

序号	非核心语义角色关系	中文说明
1	aspect	体
2	cunit	中文个体量词
3	dcopy	对象不一致的借用
4	refer	对象一致的借用
5	smood	句末无意义语气
6	tense	时态

表 2. CAMR新增语义角色关系

最后，对导入的语料进行语义标注。标注原则为一句一标、全面精确。即要标注完整的一

句话，并且对每句话中的词语都要标注到位。

3 语料标注

3.1 人工标注过程

标注的原则是基于AMR以及CAMR的标注规范来对《左传》语料进行标注。AMR和CAMR为《左传》抽象语义表示的标注提供了基本的标注原则和标注方式。标注过程和标注规范的设计流程如下：（1）对语料进行预处理之后，将语料分批导入CAMR标注平台，将语料前1000句试标导入原标注平台；同时新建一个新的标注平台，将完整的语料导入新的标注平台。（2）标注1000句语料，并根据语料特点涉及标注规范，补充修改标注平台内的资源。（3）在新的标注平台上进行正式标注，同时根据需要不断改进标注规范。（4）使用完整的标注规范修正全部语料，并由两个标注人员进行核对。

3.2 特殊语言现象的标注方法

CAMR提出了许多标注现代汉语的标注准则，例如对现代汉语中的复句关系构建标签并给出了相应的标注示例。在对现代汉语的标注和古代汉语的标注对比中发现：首先，现代汉语的许多句式都与古代汉语有同源关系，例如现代汉语中代词、名词的回指现象，在古代汉语中也存在大量的回指现象。其次，现代汉语的标注和古代汉语有所区别。在CAMR标注规范中，有对轻动词的省略的规定，最典型的例子例如“回答说”。CAMR规定这类依附在核心谓词之后的轻动词采取直接略去的方式，在标注中和“说”之前的动词系连在一起。

```

他1 回答2 说3 :4 “5 小明6 喜欢7 草莓8 ”9 。10
x2_x3/回答-01。
:arg0() x1/他。
:arg2() x7/喜欢-01。
:arg1() x8/草莓。
:arg0() x17/person。
:name() x6/小明。

```

图 1. “回答说”标注示例

但是在古代汉语中，“V+曰”的情况大量存在。由于“曰”在句子中有提示其后内容的作用，无法将其作为轻动词直接省去。这类现象会在下文详细讨论。

在对《左传》标注中发现有3类较难使用现有标注规范解决的现象：零形回指、使动用法、“V+曰”等。在接下来的文章中，将会对以上句法现象进行分析讨论，并给出相应的标注示例。

3.3 零形回指

回指（Anaphora）指篇章小句中某指称表达式的具体指涉需要参考语境中其他指称表达式才能完成释义。被参考的表达式称为先行词（Antecedent），需要参考其他表达式的表达叫做回指语/照应语（Anaphor）。根据回指语和先行语是否在同一篇章小句这一指标，可以将回指分为句内回指（Sentenital Anaphora）和篇章回指（Discourse Anaphora）。另外，根据先行语的词性，可以分为名词回指（Fully lexical NP）、代词回指（Fully lexical NP）和零形回指（Zero Form）。《左传》中绝大多数的句子都为主语位置的零形回指，指称距离 $1 \leq D \leq 2$ 的范围内，主要用于标识话题的连续性（[华建光 and 朱淑华, 2018](#)）。CAMR标注规范中只对简单的代词回指做出了说明：代词和其指代的名词使用相同的概念编号进行回指；当句子中出现零形回指时，将零形回指也指向先行词。

```

我1 问2 病人3 :4 “5 你6 为什么7 讨厌8 自己9 ,10 想11 自杀12 ?13 ”14
x2/问-01。
:arg0() x1/我。
:arg2() x3/病人。
:arg1() x18/and。
:op1() x8/讨厌-01。
:arg0() x6/x3。
:arg1() x9/x3。
:op2() x11/想-02。
:arg0() x3/病人。
:arg1() x12/自杀-01。
:arg0() x3/病人。
:cause() x7/amr-unknown。
:mode() x13/interrogative。
    
```

图 2. CAMR零形回指标注示例

《左传》中的零形回指句子大多用于强调话题的连续性、主语动作的连贯性。若按照CAMR的标注方法会使得抽象语义结构过于冗长，存在过多“and”标签表示的并列结构。因此，在经过对零形回指现象的标注和分析后，将零形回指分为以下两种类型：

```

秋1 ,2 狐突3 適4 下國5 ,6 遇7 大子8 。9
:top(/) x4/ 適 01。
:arg0(/) x12/ person。
:name(/) x3/ 狐突。
:arg0-of(/) x7/ 遇 01。
:arg1(/) x16/ person。
:name(/) x8/ 大子。
:arg1(/) x5/ 下國。
:time(/) x1/ 秋。
    
```

图 3. 《左传》零形回指标注示例 (1)

(1) 在存在零形回指的句子中只有两个动词时，采用反关系标签进行标注。上图所示的句子较为简单，包含两个小句，只有两个动词“适”和“遇”。为使标注简便采用“arg0-of”标注，也更加符合原句的语义。

```

蘇子1 叛2 王3 卽4 狄5 ,6 又7 不8 能9 於10 狄11 ,12 狄人13 伐14 之15 ,16 王17
不18 救19 ,20 故21 滅22 。23 |。
:top() x23/causation。
:arg1() x25/and。
:op1() x2/ 叛。
:arg0() x27/person。
:name() x1/ 蘇子。
:arg0-of() x4/ 卽。
:arg1() x31/country。
:name() x5/ 狄。
:arg1() x3/ 王。
:op2() x9/ 能 01。
:arg0() x36/person。
:name() x1/ 蘇子。
:accompanier(x10/於) x37/country。
:name() x11/x5。
:polarity() x8/-。
:mod() x7/ 又。
:op3() x14/ 伐 01。
:arg0() x13/ 狄人。
:arg1() x15/x1。
:op4() x19/ 救 01。
:arg0() x17/x3。
:polarity() x18/-。
:arg2(x21/故) x22/ 滅 01。
    
```

图 4. 《左传》零形回指标注示例 (2)

(2) 当同一主语下存在多个动词时，使用反关系会导致小句顺序混乱。这种情况下采用CAMR的回指规范，使用“and”等标签连接各个小句，并在每个小句的核心角色下标注出主

语。

3.4 使动用法

词类活用是指某些词通过临时改变其基本语法功能去充当其他词类或基本未改变而用法比较特殊的现象。现代汉语和古代汉语都有词类活用现象，但因为古代汉语以单音节词为主，词语数量较少，一个词除了其基本用法之外，还能产生临时用法，用以完整表达句子(方加胜, 2011)。词类活用的基本分类为使动用法、意动用法、名词作动词、名词作状语。其中，在《左传》的抽象语义标注中，动词的使动用法是标注过程中的难点。其原因在于：第一，由于时代发展词性的变化，《左传》中的一些名词词语发展到今天固定了动词的词性，例如“祸”，其本义为灾难，属于名词。但是在新华字典中，其动词词义“危害”也成为了基本词义。如果没有专门查阅古汉语词典，就会将“祸”作为动词直接标注为根节点，导致标注错误。第二，活用作动词的词类包括名词、动词、形容词。动词和形容词本身包含论元结构，能够作为语义核心，活用作动词后可以直接增加动词“使”，使用其论元表示句子语义。但当活用的词类是名词时，仅仅增加动词“使”会造成语义为“使+N+N”的结构，语义表达错误。如果要增加动词“使”，就还需要增加一个动词，确保其符合“使+N+V+N”的语法结构。动词本身就具有自己的论元，使动等活动用法会在其基础上引起原本配价的变化(李永, 2008)。过多增加动词会增加标注的复杂性，谓词论元数量过多也会影响原始的语义。

为了能够使句子的抽象语义表达完整，本文选择新增一个谓词论元“make[arg0:action; arg1:causer; arg2:People who suffer from the action of word meaning]”来表示使动用法。“arg0:action”表示活用作动词的词语所表达的动作，“arg1:causer”表示该动作的施事，“arg2:People who suffer from the action of word meaning”表示受到该动作影响的受事。

```

無極1 謂2 令尹3 曰4： 5 “ 6 吾7 幾8 禍9 子10。 ” 11 《昭公·二十七年》。
:top() x2_x4 / 謂 02。
:arg0() x40 / person。
:name() x1 / 無極。
:arg1() x42 / person。
:name() x3 / 令尹。
:arg2() x44 / make 01。
:arg0() x9 / 禍。
:mod() x8 / 幾。
:arg1() x7 / x1。
:arg2() x10 / x3。
    
```

图 5. “祸”使动用法

3.5 “V+曰”

《左传》中记录了许多对话，动词“曰”使用频繁。当句子中仅存在“曰”时，可以将“曰”标记为根结点，按照其论元结构进行后续标注。但《左传》中也有许多“V+曰”的形式，其中，“曰”在这个结构中起到提示下文的作用，“V”则提示着说话者是以怎样的方式说话。

CAMR中类似的句式为“V+说”，CAMR标注规范将“说”处理为轻动词，并在标注中省略“说”。但“曰”并不能作为轻动词进行处理，因为“曰”是《左传》乃至古代汉语中使用量最多、动词用法最为确定的一个动词。“曰”主要的论元结构包括“曰[说话人；说话的内容]”，提示之后的说话者表达的内容，二者缺一不可。在“V+曰”的结构中，“曰”提示说话者的功能虽然被“V”承担了一部分，但其提示说话内容的功能依然保留。在这类句式下，又有一类特殊的结构：谓+曰。《说文解字》认为，“谓，报也。”，在《左传》中常以“说话人谓听话人：说话的内容”形式出现，论元结构是“谓[说话人，听话人，说话的内容]”。当“谓”和“曰”共同出现时，“曰”的提示说话内容的功能就被“谓”承担了部分。经过讨论，本文决定将“谓+曰”形式的标注规则设计如下：

```

    秦伯1 謂2 郤芮3 曰4 :5 “ 公子7 誰8 恃9 ? ” 1110 ” 1111 .
:top(/) x2 / 謂 .
    :arg0(/) x14 / person .
        :name(/) x1 / 秦伯 .
    :arg1(/) x16 / person .
        :name(/) x3 / 郤芮 .
    :arg2(x4/曰) x9 / 恃 .
        :arg0(/) x7 / 公子 .
        :arg1(/) x8 / amr-unknown .
        :mode(/) x10 / interrogative .
    
```

图 6. “谓+曰”标注示例

将“曰”放置在“arg2:说话的内容”的边上，这样既保留了“曰”的功能，也保留句子完整的抽象语义。

和“谓+曰”类似的结构还包括“言+曰”，动词“言”的论元同样也包括说话的内容。其他类型的“V+曰”，例如“对曰、辞曰”，都按照将“曰”作为根结点，使用manner标签标注“V”。

```

    管仲1 辭2 曰3 .
:top(/) x3 / 曰 .
        :arg0(/) x41 / person .
            :name(/) x1 / 管仲 .
        :manner(/) x2 / 辭 .
    
```

图 7. “辞+曰”标注示例

4 语料统计

目前，抽象语义标注语料库已有1500多句经过语义标注的语料，新增谓词论元框架70余个，加上已经存在并经过修改谓词论元框架共20000条。但考虑到该工作并未完成，由于古汉语词汇量小于现代汉语的词汇量，猜想最后成型的谓词论元框架总数应该小于目前的数量。

对试标的1500句进行初步统计，统计句子所包含的词数以及句子所包含的节点数的分布。从表3可以看出，1500句话中平均词数为13.53，词数为8的句子最多，最多有70个词语的句子。这是由于《左传》的句子大多精简，而词语数量较多的句子通常来自于人物对话，所以未能被分割。词语总数和节点总数相同，是由于在标注过程中没有增加或删除节点。

	平均数	中位数	众数	方差	最小值	最大值	总数
节点数	10.15	9	5	46.08	1	54	1455
词数	13.53	11	8	96.03	2	70	1455

表 3. 词数、节点数的统计数据

在已标注的1500句中，筛选出所有使用的“arg0-of”反关系标签的情况中有97%的句子都属于零形回指的用例，其余3%是反关系的其他用法。零形回指的出现导致“arg0-of”的使用大大增加。在零形回指的用例中，基本都使用了“arg0-of”进行标注，能够证明该方法的效度。

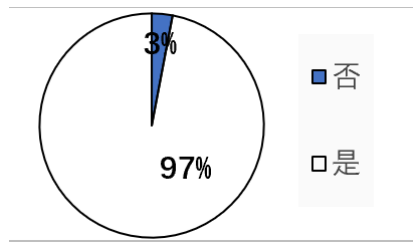


图 8. “arg-0 of”中零形回指占比

在已标注的全部用例中，一共有241例包含动词“曰”的句子，其中“V+曰”的数量共70例，包括“对曰、告曰、谓曰”等。“对曰”的占比最大，是由于《左传》中人物间的对话较多。详细的统计排列见下表：

V+曰	数量	占比	V+曰	数量	占比
对曰	31	0.44	徇曰	2	0.03
告曰	8	0.11	叹曰	2	0.03
谓曰	8	0.11	辞曰	2	0.03
谏曰	5	0.07	复命曰	1	0.01
言曰	4	0.06	赋曰	1	0.01
书曰	2	0.03	泣曰	1	0.01
问曰	2	0.03	反曰	1	0.01

表 4. “V+曰”统计

5 结语与未来工作

古汉语文献语料库、分词和词性标注语料库都在评测方面取得了不错的成果，但语义评测仍是古文语言处理的短板。句子语义的自动分析处理的基础之一即为构建相应的句子语义语料库，本文研究则是古代汉语抽象语义标注语料库建设的基础。本文通过对《左传》的标注，最终建设一个古代汉语抽象语义语料库，该语料库能够为古文信息处理的工作提供语料资源，促进古汉语的自动语义分析发展。即使评测效果较好的CAMR标注规范仍不能满足《左传》的抽象语义表示，但不可否认的是，CAMR的标注规范为《左传》的标注提供了指导，其中基本的概念标签和基础的标注原则仍为《左传》所使用。本文所讨论的问题和已经形成的部分资源库只是最终工作的一部分，对《左传》的标注工作还未完成，标注规范仍在不断完善中。在未来的工作中，吸取本文研究过程中的经验教训，总结归纳并解决《左传》中不能为CAMR标注规范所解决的句式；逐渐完善《左传》的标注规范，最终形成一个系统全面，且能够基本涵盖上古汉语语义的标注原则；最终构建一个《左传》抽象语义标注语料库，为《左传》的相关研究提供语料资源。

参考文献

- Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. pages 178–186.
- Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180.
- Hyonsu Choe, Jiyeon Han, Hyejin Park, and Hansaem Kim. 2019. Copula and case-stacking annotations for korean amr. In *Proceedings of the first international workshop on designing meaning representations*, pages 128–135.
- Charles J Fillmore et al. 2006. Frame semantics. *Cognitive linguistics: Basic readings*, 34:373–400.

- Xiaoling Hu, Jamie McLaughlin, and Nigel Williamson. 2007. Syntactic positions of prepositional phrases in the history of chinese: using the developing sheffield corpus of chinese for diachronic linguistic studies. *Literary and linguistic computing*, 22(4):419–434.
- Chu-Ren Huang and Keh-jian Chen. 1992. A chinese corpus for linguistic research. In *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*.
- B. Li, W. Yuan, Q. U. Weiguang, L. Bu, and N. Xue. Annotating the little prince with chinese AMRs. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*.
- Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.
- Noelia Migueles-Abraira, Rodrigo Agerri, and Arantza Diaz de Ilarraza. 2018. Annotating abstract meaning representations for spanish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Lucien Tesnière. 2015. *Elements of structural syntax*. John Benjamins Publishing Company.
- 俞士汶, 段慧明, 朱学锋, and 孙斌. 2002. 北京大学现代汉语语料库基本加工规范. 中文信息学报, 16(5):51–66.
- 华建光 and 朱淑华. 2018. 《左传》篇章零形回指分布概况. 南京师范大学文学院学报, (1):10.
- 卜丽君. 2017. 基于AMR的中文句子语义标注及统计分析. Ph.D. thesis, 南京师范大学.
- 尤, 李涓子, and 王作英. 2003. 基于语义依存关系的汉语语料库的构建. 中文信息学报, 17(1):46–53.
- 张景霓. 2000. 论《左传》中的省略. 广西民族学院学报(哲学社会科学版), 22(5):107.
- 方加胜. 2011. 《左传》中的使动、意动用法研究. Ph.D. thesis, 南京师范大学.
- 李佐丰. 1989. 《左传》的“使字句”. 语文研究, (2):29–34.
- 李斌, 闻媛, 卜丽君, 曲维光, and 薛念文. 2017a. 英汉《小王子》抽象语义图结构的对比分析. 中文信息学报, 31(1):9.
- 李斌, 闻媛, 宋丽, 卜丽君, 曲维光, and 薛念文. 2017b. 融合概念对齐信息的中文amr语料库的构建. 中文信息学报, 31(6):10.
- 李永. 2008. 上古汉语动词配价分化的语义基础和句法机制. 古汉语研究, (2):47–51.
- 武惠华. 1994. 《左传》外交辞令探析. 中国人民大学学报, 8(4):47–54.
- 舒蕾, 郭懿鸾, 王慧萍, 张学涛, and 胡韧奋. 2022. 古汉语词义标注语料库的构建及应用研究. 中文信息学报, 36(5):21–30.
- 袁本良. 2003. 古汉语句法变换研究中的语义问题. 中国语文, (3):242–247.
- 郑丽娟, 邵艳秋, and 杨尔弘. 2014. 中文非投射语义依存现象分析研究. 中文信息学报, 28(6):41–47.
- 郝晓燕, 刘伟, 李茹, and 刘开瑛. 2007. 汉语框架语义知识库及软件描述体系. 中文信息学报, 21(5):96–100.
- 闻媛. 2018. 基于概念关系对齐的CAMR语义库构建及统计分析. Ph.D. thesis, 南京师范大学.