文章编号：1003-0077(2020)04-0021-09

基于关系对齐的汉语虚词抽象语义表示与分析

戴玉玲[1]，戴茹冰[1]，冯敏萱[1]，李 斌[1,2]，曲维光[3]

(1. 南京师范大学 文学院，江苏 南京 210097；
2. 哈佛大学 计量社会科学研究所，美国 剑桥 02138；
3. 南京师范大学 计算机科学与技术学院，江苏 南京 210023)

摘 要：虚词具有丰富的语法意义，对句子理解起着不可或缺的作用。虚词的语言学研究成果丰富，但缺乏形式化表示，无法直接被计算机利用。为了表示虚词的句法语义信息，该文首先在抽象语义表示(abstract meaning representation，AMR)这种基于概念图的语义表示方法的基础上，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。其次，选取了语言规范的人教版小学语文课本 8 587 句作为语料，进行 AMR 的标注。然后，针对语料中 24 801 个虚词实例进行统计，发现介词、连词、结构助词对应概念间的关系，占虚词总数的 58.80%；而语气词和体助词表示概念，占 41.20%。这表明 AMR 可以动态地描写出虚词功能，为整句句法语义分析提供更好的理论与资源。

关键词：虚词；抽象语义表示；关系对齐；语言知识库

中图分类号：TP391 **文献标识码**：A

Representation and Analysis of Abstract Meaning of Chinese Function Words Based on Relation Alignment

DAI Yuling[1], DAI Rubing[1], FENG Minxuan[1], LI Bin[1,2], QU Weiguang[3]

(1. School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
2. Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA;
3. School of Computer Science and Technology, Nanjing Normal University, Nanjing, Jiangsu 210023, China)

Abstract: Function words have rich grammatical meanings and are crucial to sentence comprehension. The existing linguistic researches on function words cannot be directly adopted by computational linguistics due to lack of formal representation. In this paper, to represent their syntactic and semantic information, we align words and conceptual relations in the abstract meaning representation (AMR) based on concept graphs, so that function words correspond to nodes or arcs between conceptual nodes. Then, 8,587 sentences from PEP primary school Chinese textbooks are selected for AMR annotation. Among the total 24,801 tokens of function words in this corpus, 58.80% are prepositions, conjunctions and structural auxiliaries which are correspond to relations between concepts, and 41.20% are modals and aspects which express concepts. This shows that AMR represents function words dynamically, providing better theory and resources for the syntactic and semantic analysis of whole sentences.

Keywords: function words; abstract meaning representation; relation alignment; linguistic knowledge base

0 引言

虚词从传统上来说，是指没有实际的词汇意义，且无法单独充当句法成分的词。目前在中文自然语言处理中可利用的虚词信息不多，常用虚词约有二三百个[1]，数量有限且使用频繁，对句子的语义理解起着至关重要的作用[2]。若仅在句中简单增减或变

收稿日期：2019-09-16 **定稿日期**：2019-10-16

基金项目：国家社会科学基金(18BYY127)；国家自然科学基金(61772278)；江苏高校哲学社会科学优秀创新团队建设项目(2017STD006)；国家语委科研项目(YB135-61)

换虚词,便能得到与原意大相径庭的句子,例如,“张三把李四说了一通”和“张三被李四说了一通”[3]。这两句的施事和受事完全相反,语义上可谓失之毫厘,谬以千里。近年研究也发现,虚词并不是一种赘余信息,其可以表达概念,表示时体意义和情态意义等[4]。因此,在中文自然语言处理领域,尤其是句法语义的分析中,需要加入虚词特征。

但传统语言学倾向于对虚词“逐个击破”,即主要对某个特定的虚词进行深入研究,充分挖掘其用法;或是比较相似的虚词,探究它们在句义表达中的差异。如在《现代汉语虚词散论》[5]中,对虚词“比”进行了细致描写,对语气词“啊”“吧”“呢”“吗”进行了比较分析。这些从虚词特殊的语言现象中归纳出来的语言学知识硕果累累,但由于缺乏形式化的表示,难以被计算机利用。

目前在中文自然语言处理中的应用如自动文摘、信息检索中,普遍做法是将虚词列入停用词表,过滤其信息[6]。但随着自动句法语义分析的发展,这种做法逐渐受到挑战[7]。昝红英等建设了《现代汉语虚词用法知识库》(the Chinese function word usage knowledge base,CFKB),收录了虚词的用法、使用规则和语料库[8],基于此知识库,还研究了介词[9]和助词[10]等用法的自动识别技术。抽象语义表示(abstract meaning representation,AMR)是一种基于概念图[11]的语义表示方法,其将句子语义抽象为一个单根有向无环图,有效保留了句子的语义信息。本文根据汉语的特点对其进行了改进,在整句语义分析的基础上,保留了句子中的虚词特征,探究了虚词新的表示方法,形成了中文的抽象语义表示(Chinese AMR, CAMR)。通过适中颗粒度的标签标注,首次将虚词特征纳入语义分析中,构建了句法语义一体化表示方法。

1 相关工作

1.1 国内外虚词研究

1.1.1 国外虚词研究

国外对于虚词的理论研究主要集中在语音和语法层面。语音层面上,Zec[12]研究了塞尔维亚语中的虚词音韵,发现音韵特征不同的虚词在句法分布上有重叠关系。查莫罗语中也发现音韵特征可以决定特定代词的出现位置[13]。语法层面上,主要从历时演变的角度来考察。Hacking[14]研究了保加利亚语、马其顿语和俄语中的小品词,认为其是语法化的一个有效例证。Heine 等[15]发现英语中的人称代词在历时演变中具有特殊现象,由此对语法化理论进行拓展。

在应用方面,Bailey[16]根据虚词出现频率构建模型推断作者的写作风格。Marečcek 等[17]通过实验证明,在无监督的依存分析中,虚词有一系列的依存词,加入其特征后显著提高了分析的效果。Tang 等[18]在类比推理和段落识别的任务中,发现虚词会改善词嵌入模型的效果。

在虚词的信息处理资源方面,Reich 等[19]构建了常见的口语词汇表,其中包括了数量稳定的虚词。在词汇级的语料库如英国国家语料库(British National Corpus)①中收录了虚词及其例句,但在句法及语义级的语料库中虚词信息被忽略[20]。

1.1.2 国内虚词研究

现代汉语虚词的研究可以从句法、语义、语用三个层面来概括。

在句法层面上,大量的研究专注于虚词的用法和语法功能,如《现代汉语八百词》《现代汉语虚词释例》《现代汉语虚词词典》《现代汉语虚词散论》等,主要通过比较来剖析具体的虚词。此外,石毓智等[21]从历时角度出发,从实词虚化的过程中探究语法化的动因和机制。

在语义层面上,由于学界对虚词是否有语义概念争执不清,关于虚词语义的研究较少。孙中一[22]提出虚词是有概念义的,郭锐[23]提出了针对虚词的语义结构分析法,此后也出现了针对具体词如“只”[24]“就”[25]的语义研究。

在语用层面上,张斌、范开泰[26]在《现代汉语虚词研究系列丛书》中详细介绍了介词、副词、连词、助词、语气词的语用功能,知网中也不乏针对具体词如“都”[27]的语用研究。

但这些研究都存在共通的问题,即现代汉语的虚实分类的标准不统一,这也就导致了虚实的归类不一,虚词内部也存在划界问题[2]。再者,专著和研究浩如烟海,但都是面向人的,缺乏形式化的表示。这些问题在一定程度上阻碍了虚词相关的自动分析。因此,需要将传统的虚词研究成果加以形式化,促进汉语句法语义自动分析技术的发展。

① https://www.english-corpora.org/bnc/

1.2　短语结构语法和依存语法的虚词表示

面向中文自然语言处理的汉语虚词资源主要有两大来源：汉语句子级的资源库及北京大学现代汉语虚词知识库。前者对虚词信息的保留度不高。短语结构语法和依存语法是目前构建句法库和语义库的主要语法理论。汉语的短语结构库有宾州中文树库、清华中文树库、国家语委中文树库及北大中文树库等[28]。图 1 以清华中文树库的标注体系为例，对“他在剧院看了演出”构建短语结构树。

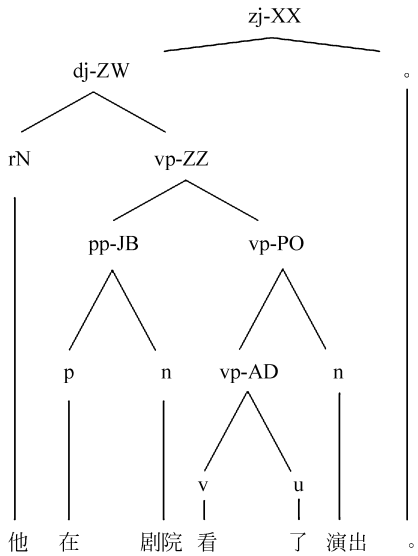

图 1　“他在剧院看了演出”清华树库标注示例

图 1 中，p 为介词，u 为助词，pp 为介词短语，vp 为动词短语，JB 为介宾结构，AD 为附加结构。这种标注方式保留了虚词的词性、语法结构、成分关系，且虚词全部标注在节点上。但在基于短语结构的语义角色分析中，会对虚词进行剪枝处理[29]，舍弃其信息，并对各实词的语义角色进行标注。这一点不利于挖掘虚词的用法，以及把握各语义角色之间的关系。

依存语法是通过分析句子中各语法成分之间的依存关系揭示其句法关系的。它重点关注各成分之间的关系，认为虚词附属于实词[30]。汉语的依存库主要有哈尔滨工业大学中文依存树库。图 2 以哈工大[31]的依存库为例，对上述例句构建依存树。

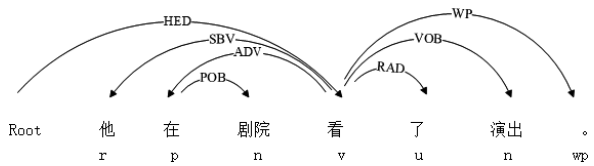

图 2　“他在剧院看了演出”句法依存样例

图 2 中，p 为介词，u 为助词，POB 为介宾关系，RAD 表示右附加关系。句法依存保留了虚词的词性及依存关系，虚词同样标注在节点上。至于语义依存分析，分为浅层分析和深层分析，浅层分析是指在依存语法基础上的语义角色的标注；深层分析是指借助词汇的语义框架，分析句子各个语言单位之间的语义角色关联，并将语义关联以依存结构呈现。例如对上文的例句进行语义依存分析，结果如图 3 所示。

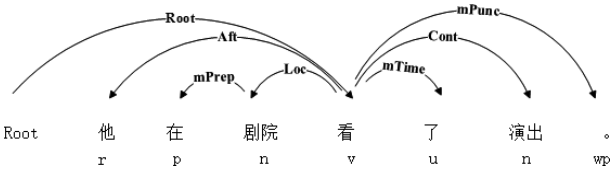

图 3　哈尔滨工业大学语义依存样例

图 3 中，mPrep 表示介词，mTime 表示时间。语义依存在具有直接语义关联的实词之间建立依存弧，而虚词被弱化，作为辅助标记存在，且句法和语义的依存弧的指向并不一致。不管是浅层分析还是深层分析，虚词信息基本被忽略。此外，虚词的语法意义涉及到两个实词之间的关系居多，而目前表示方式基本都将虚词处理为类似实词的节点，这一点也有待商榷。从语法、语义和表示方式来看，短语结构语法和依存语法对虚词的处理可综合成表 1①。

表 1　短语结构语法与依存语法中的虚词信息

理论载体	词性	语法结构	成分关系	语法意义	表示方式
短语结构语法	＋	＋	＋	－	节点
句法依存	＋	－	＋	－	节点
语义角色标注	＋	＋	＋	－	节点
语义依存	＋	－	＋	－	节点
CAMR	－	－	＋	＋	节点和弧

1.3　现代汉语虚词用法知识库

为了促进虚词的形式化表示，丰富计算机可利用的虚词信息，北京大学参考了《现代汉语八百词》等传统虚词著作，于 2013 年构建了 CFKB 知识库，该知识库集《现代汉语虚词用法规则库》《现代汉语虚词用法词典》《现代汉语虚词标注语料库》于一体，共收录虚词 2 401 个，用法共计 4 337 个，语义共计

① 表格内的“＋”表示该理论可以描述对应信息，“－”表示无法描述对应信息。

2 982个,针对虚词用法设计了4 696条用法规则。知识库中使用了“词类”“体宾谓宾”“否定前后”等多条字段描述虚词,光描述介词的属性字段就多达49个[9]。各个虚词的用法切分也较为细致,例如,词典中收录了“把”的13种用法。不过,这些信息还较为稀疏,例如,介词仅有“格标记”字段记录相关语义信息。如“把”的格标记表示为“施”“受”“处”。此外,其语料库基于《人民日报》1998年1月及2000年1~6月的分词与词性标注语料[32],标注了虚词在词典中对应的编号。这种标注方法简单易行,在虚词的用法识别过程中可以发挥良好的效果[33]。不过,若是要将虚词信息加入整句语义分析中,还需要句子级别的句法语义库的支持。

2 CAMR简介

抽象语义表示在2013年由Banarescu等[34]学者提出,目前已经涌现了许多关于AMR的研究与应用。其基于PropBank命题库,将句子语义抽象为一个单根有向无环图。其中,各节点表示句子中实词的概念,有向弧表示概念之间的关系。相较于短语结构语法和依存语法,AMR可以根据句子的语义增删概念节点,还原句子的隐含信息,且允许回指,允许论元共享。因此,AMR灵活性更强,能够清楚地表示出一个句子的语义。但AMR会在语料预处理过程中过滤英文中的意义较虚的词,例如,a、an等[11]。

CAMR是在AMR的基础上对中文句子的抽象语义表示,其根据汉语的特点对其进行了优化,增加了aspect(体)、cunit(中文特殊量词)、perspective(方面)等关系标签。此外,还设置了10个表示复句的概念标签,如causation(因果关系),以适应汉语多样的复句关系[35]。

为了保留汉语虚词中的丰富信息,CAMR还增加了虚词信息的标注,根据语法特点将其以两种方式加以表示[36]:对于只对应实词间的关系意义的虚词,可以看作是语义关系的实例,因此与语义关系标签一同标注在有向弧上;对于表示句子的体意义和语气意义的虚词,将其处理为概念节点。CAMR保留虚词的特征,构建了句法语义一体化的标注平台。

2.1 CAMR中的虚词表示

CAMR将虚词标注在实词概念间的有向弧及概念节点上,保留了虚词的成分关系及语法意义。在如图4所示的CAMR图中,xn(n∈N*)为概念标签的编号,arg0(原型施事)、arg1(原型受事)、aspect(体)、location(处所)为关系标签。介词“在”引出动作的地点,对应了处所词“剧院”和谓词“看”的location,因此将其标注在二者之间的有向弧上,作为处所关系的实例。“了”表示“看”的动作已经完成,表示的是体意义,因此将其处理为概念,并使用aspect标签将句子的体信息表示出来。

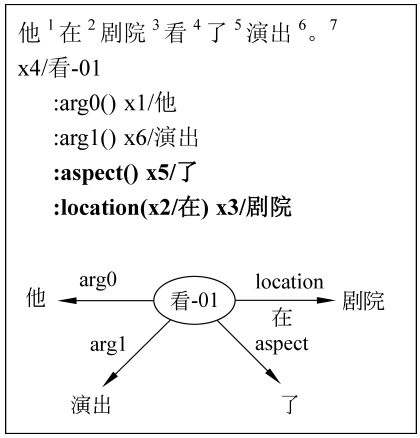

图4 CAMR标注及图结构示例

CAMR使用5个核心语义关系标签、44个非核心语义关系标签对概念及概念间的关系进行标注,颗粒度较为细致。因此可以挖掘出虚词较为完整的属性。例如,“把”在CAMR语料库中对应的语义关系共有施事、受事、工具、话题四种。表2列出了CAMR中部分关系标签及含义[35]。

表2 CAMR部分关系标签①及含义

关系标签	含义	关系标签	含义
:arg0	原型施事	:location	地点
:arg1	原型受事	:manner	方式
:arg2	工具等	:mod	修饰
:arg3	出发点等	:part-of	部分
:beneficiary	受益	:poss	领属
:cause	起因	:purpose	目的
:direction	方向	:source	源
:instrument	工具	:time	时间

CAMR保留虚词特征,主要出于以下三点考虑:①首先,句子的体信息和语气信息对句义贡献很大,对它们加以标注可以有效减少句子在表示过

① 由于CAMR中语义角色标签较多,共49个,碍于幅面限制,这里仅呈现下文出现的标签并加以解释。

程中的语义损失；②再者，在后续构建篇章语料库时，就可以提取上下文中虚词特征，通过这些信息可以快速将句子衔接起来，大大提高 CAMR 连句成篇的效率，为之后的篇章语料库的建设以及篇章的语义理解打下良好的基础；③第三，根据虚词本身的语法特征，将对应实词关系的虚词标注在实词间的有向弧上，将表示整句的体和语气概念的虚词表示为节点，符合虚词的语法属性，有助于发现中文虚词多样的语法意义，不论对虚词的知识库建设还是理论研究来说都是珍贵的资源。

2.2　语料来源与标注

本文以 2001 年人教版一至六年级的语文课本作为语料（以下简称小学语料）并对其进行标注，抽取其中的虚词加以分析。选择小学语料作为本文的研究对象主要出于以下两点考虑：首先，小学语文课文都经过了专家精心编写和审核，内容权威，表述规范；第二，文本中的虚词丰富，足以囊括常见的虚词，且具有代表性。

共有 3 名语言学研究生参与了此项小学语料的 CAMR 标注。其中，前 260 句由其中两人同时标注，标注一致性的 smatch 值[37]达 80%。在经过相互讨论、及时调整不一致的判断后，对剩余语料进行均分标注，再由第三人对所有语料进行整体校对。

3　虚词的统计与分析

小学语料共含句子 8 696 个，除去诗词和空句外，有效标注的句子达到 8 587 个。本文从中抽取了虚词信息。由于 CAMR 不标注词性信息，本文手工标注了抽取出的虚词词性，统计介词、连词、助词、语气词①的分布情况，并从语法形式和语法意义两个方面切入，简单进行分析。

本文共抽取了有向弧上的信息共 17 611 条，其中虚词词例数 14 583 例，词型数 65 例；概念标签中的虚词词例数 10 218 例，词型数 173 例，共获得虚词总词例数 24 801 例，总词型数 238 例。表 3 给出了虚词各词类在小学语料中的频次分布。可以看出，助词词例数最多，为 13 250 例。其中，8 249 例助词出现在了有向弧上，5 001 例标为了概念节点。也即，助词既可以对应到概念间的关系，也可以直接表达概念。其次是语气词，为 5 217 例，只能标为概念节点，无法对应概念之间的关系；而连词和介词较少，分别有 2 629 例和 3 705 例，都只能出现在有向弧上，即其只能对应到概念之间的关系，而不单独表达概念。接下来本文将逐个分析每个词类的具体信息。

表 3　小学语料虚词词型及词例分布

标注位置	助词		语气词		连词		介词	
	词型	词例	词型	词例	词型	词例	词型	词例
弧	4	8 249	/	/	131	2 629	61	3 705
概念	3	5 001	39	5 217	/	/	/	/
总计	7	13 250	39	5 217	131	2 629	61	3 705

3.1　表示概念的虚词

第一类虚词可以表示概念，在 CAMR 中表示为概念节点，和关系标签合用标注出其语法信息。在小学语料中，标注为概念节点的虚词有体助词和语气词，约占总虚词词例数的 41.20%。这表明在语料中，接近半数的虚词都可以表示概念：体助词表明句子的体信息，语气词表明句子的语气信息。

3.1.1　体助词

体助词和关系标签 aspect（体）合用，在 CAMR 中表示句子的体信息。体助词包括“了”“着”“过”。关系标签 aspect 在小学语料中共计 5 451 例。表 4 列出了 aspect 标签下各虚词的频次及频率。

表 4　小学语料各体助词频次及频率

体助词	频次	频率/%
了	3 349	61.43
着	1 487	27.28
过	165	3.03
总计	5 001	91.74

可以看出，体助词“了”“着”“过”占 aspect 总频次的 91.74%，其他的 8.26%主要是由时间副词“曾

① 由于传统语言学中虚词分类不一，出于语言信息处理的方便，本文采用邵敬敏《现代汉语通论》的观点，包括介词、连词、助词、语气词。

经”“已经”等提供。这说明要想表示句子的具体意义,表达动作的完成或改变,优先使用体助词。“了”的使用频率最高,占总频次的61.43%,而表示经历过某种动作的“过”仅占3.03%,与“了”差异悬殊。

3.1.2 语气词

CAMR将句子语气统一由规定的概念标签和语义关系mode共同表示。语气概念标签包括imperative(祈使语气)、expressive(感叹语气)、interrogative(疑问语气)、judgement(判断语气)。其中,前三种语气主要单独使用标点符号或与“吗”“呢”等具体语气词合用,表示相应的语气概念,而后一种则主要由虚词框架“是……的”表示,此处暂不考虑虚词框架的情况。图5给出了虚词和标点符号表示语气概念的频次分布。

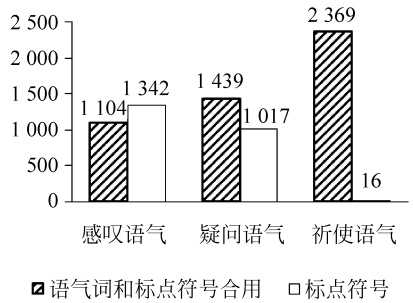

图5 小学语料表语气概念的语气词及标点符号频次分布

如图5所示,在这三种语气概念中,虚词的参与程度都很高,感叹语气中的虚词占总频次45.22%,疑问语气中的虚词占比达到60.91%,而祈使语气中的虚词占比最高,达总频次的99.35%。通过统计,共24种语气词表示感叹,如“呀”“吧”“吗”等10种语气词表示疑问。“吧”等4种语气词表示祈使。这说明,一种语气词可表示多种语法意义,一种语法意义也可以用多种语气词表示。

表5列出了小学语料中频次最高的10种语气词及其在三种语气下的分布。在标点符号判断句子语气不明朗或符号缺失的情况下,虚词便能成为一个重要判定指标。例如,“!”既可表达感叹语气,也可表达祈使语气,但如果句子中出现了“吧”,则可判断出该句一般为祈使句。因此,该统计结果可为自动判定句子语气提供参考。

表5 小学语料中频次前十的语气词比例及分布

语气词	感叹语气/%	疑问语气/%	祈使语气/%	总频次
吧	4.81	3.76	91.43	2 473
呢	20.17	79.83	0.00	843

续表

语气词	感叹语气/%	疑问语气/%	祈使语气/%	总频次
吗	0.00	100.00	0.00	546
啊	95.00	5.00	0.00	360
呀	77.89	20.19	1.92	104
嘛	43.06	0.00	56.94	72
啦	83.82	16.18	0.00	68
哇	80.00	20.00	0.00	30
哪	100.00	0.00	0.00	25
罢	95.83	4.17	0.00	24

3.2 对应到关系的虚词

第二类虚词可以对应到概念间的关系,与关系标签共同标注在CAMR的有向弧上,包括介词、连词和结构助词“的”“地”“得”“之”,约占总虚词词例数的58.80%。这些虚词都能够提供必要的语法意义,将处于概念节点上的实词联系起来。

3.2.1 结构助词

在助词中,上文提及的体助词较特殊,表达句子的体意义时,由专门的关系标签aspect表示,除去这些体助词之外,还有8 249个结构助词,与对应的关系标签共同标注在有向弧上,表示各实词概念之间的关系。在小学语料中,结构助词有四种,为“的”“地”“得”“之”。

尽管小学语料中结构助词只有4种,但所涉及的概念之间的关系却很多样。频次前十位的概念间关系及其对应频次如表6所示。在4种结构助词中,“之”由于带有古体色彩,极少在小学语料中使用,主要对应概念间的mod(修饰)关系。“地”的使用频次较低,也主要用于修饰关系中。“得”主要集中在manner(方式)和cause-of(cause的反关系,表示“由于”)。“的”频次占绝对优势,共2 391次,占总频次的28.99%,且对应的关系较多较杂,共涉及语义关系35种,可用于arg0-of(施事的反关系)、poss(领属)、mod等关系中。而新华词典中,“的”的助词用法仅有修饰、领属两种,统计结果也可用于补充字典的相关释义。

表6 小学语料结构助词对应的频次前十的语义关系

语义关系	的	地	得	之
:arg0-of	2 391	10	33	2
:poss	1 244	0	0	5

续表

语义关系	的	地	得	之
:manner	914	0	102	0
:mod	748	4	0	14
:arg1-of	585	2	12	3
:location	438	0	0	1
:part-of	322	0	0	1
:time	101	5	0	0
:cause-of	3	0	99	0
:arg1	49	2	19	2

3.2.2　介词

在小学语料中，标在有向弧上的介词共有 61 种，共计 3 705 个，而相对应的关系标签仅有 38 种。碍于幅面限制，本文只选取在有向弧上出现频次最高的语义关系和虚词。表 7 中，行标签表示频次前十位的虚词，列标签表示频次前十位的语义关系，表格中的数字则表示虚词标为对应语义关系的频次。从表 7 中的行标签来看，介词在句子中关联的语义关系主要可分为两种：一种是标为 arg1、arg2、arg3 的核心语义角色，另一种则是外围语义角色，如 instrument（工具）。这两种语义角色与介词关联的频次不相上下。但整体来看，即使频次前十的虚词，也只能对应部分语义角色，即特定介词有特定语法意义，在表示关系时有选择性。如介词“在”频次最高，且基本都对应 location（地点），共 431 个实例，而标注为 beneficiary（受益）频次为 0，表明在语料中，“在”一般引出“地点”，而不与受益格的实词搭配。介词“把”和“被”对应的关系标签主要是核心语义角色 arg0、arg1 等，表明这一对介词常引出核心语义角色，而基本与外围语义角色无关。根据该统计，可为下一步构建介词框架提供参考。如“在……中”一般引出的是地点。

表 7　小学语料中频次前十的介词分布

介词	:arg1	:location	:arg2	:source	:arg0	:time	:direction	:arg3	:beneficiary	:instrument	:purpose
在	204	431	67	1	20	64	18	2	0	0	2
把	591	0	10	0	28	0	0	0	0	1	0
从	28	5	25	153	7	8	0	1	77	1	45
给	45	0	68	0	8	0	1	0	0	0	6
对	39	0	12	0	0	0	5	101	11	0	1
向	44	0	34	0	1	0	62	3	16	0	1
被	35	1	3	0	113	0	0	2	0	0	0
到	48	13	26	0	0	0	0	10	0	0	0
用	6	0	9	0	8	0	0	0	0	93	0
为	24	0	18	0	2	0	0	0	41	0	8

3.2.3　连词

在小学语料中，标在有向弧上的连词共 131 种，共计 2 629 个，对应的语义角色共 14 种，共涉及 10 种复句关系。其中关系标签 op2、op3 等与概念标签 and（并列）合用表示并列关系，而 arg1、arg2 等与概念标签 contrast（转折）、temporal（承接）、condition（条件），causation（因果）合用表示转折关系、承接关系、条件关系、因果关系。其中 arg1 表示复句的前接句，arg2、arg3 等表示后继句。与介词处理方式相同，本文只选取在语料中频次最高的复句类型及对应的连词。表 8 中的行标签表示频次前十位的虚词，列标签表示频次前五位的复句关系，表格中的数字则表示虚词标为对应语义关系的频次。

前十的高频连词中，“和”“而”对应了 temporal、and 复句关系，说明这两个连词既可对应顺承关系，也可对应并列关系。其他八个连词只能对应单一的复句关系，表示单一的语法意义。例如，“可是”只在 contrast 标签的 arg2 下出现，只能对应转折复句的后继句；“如果”只能对应条件复句的前接句。数据表明，连词的语法意义较为单一，倾向用于特定复句类型中。因此，连词是判定复句的有效依据，保留该信息可提高整句语义分析的效果。小学语料中标注的总复句数为 695 例，其中以转折复句为主，有 406 例，约占总复句的 58.42%，其次是顺承和条件复句，分别

为89例和80例，约占12.81%和11.51%，而并列句数量较少。笔者推测连词的频繁使用可能是小学语料为避免平铺直叙而采取的做法，这也可为文本风格分析提供依据。

表8 小学语料频次前十的连词对应的复句类型及频次分布

连词	转折		承接				条件	因果	并列			总计
	:arg1	:arg2	:arg1	:arg2	:arg3	:arg4	:arg1	:arg1	:op2	:op3	:op4	
可是	0	104	0	0	0	0	0	0	0	0	0	104
却	0	99	0	0	0	0	0	0	0	0	0	99
和	0	0	0	42	0	1	0	0	10	22	7	82
如果	0	0	0	0	0	0	80	0	0	0	0	80
因为	0	0	0	0	0	0	0	70	0	0	0	70
但	0	74	0	0	0	0	0	0	0	0	0	74
但是	0	60	0	0	0	0	0	0	0	0	0	60
然后	0	0	0	25	14	7	0	0	0	0	0	46
虽然	43	0	0	0	0	0	0	0	0	0	0	43
而	0	26	0	0	0	0	0	0	0	9	2	37
总计	406		89				80	70	50			695

4 总结与展望

本文梳理了短语结构语法和依存文法中虚词的表示方法，发现在以这二者为代表的文法中，倾向于忽略虚词的信息，且将虚词直接与实词一同放在节点上，这一做法难以表现出虚词的语法特点：多数虚词表示的是实词间的关系。此外，北京大学CFKB知识库在识别虚词用法的任务中成效显著，尚缺少整句的一体化句法语义表示。出于保留和挖掘虚词特征、提高整句语义分析效果的目的，本文在AMR的基础上，根据汉语特点进行改进，保留了虚词特征，将对应实词间关系的虚词表示在实词间的有向弧上；将表示概念的虚词表示在概念节点上，完善了句法语义一体化表示的CAMR。基于CAMR标注了小学语料，抽取其中的虚词信息，发现虚词的语法形式与其表达的语法意义有密不可分的关系，要想充分理解句子的意义，虚词提供的这些信息必不可少。

未来的工作中，在扩大中文AMR语料规模的基础上，挖掘更多的虚词知识，并和已有的虚词词典和资源进行义项链接，使得虚词知识库更加完善。同时，在语义自动分析过程中加入虚词在各种语义关系下的概率信息，以统计和规则相结合的方法，提升语义自动分析效果。

参考文献

[1] 陆俭明. 现代汉语语法研究教程[M]. 北京：北京大学出版社，2015：186.

[2] 齐沪扬，张谊生，陈昌来. 现代汉语虚词研究综述[M]. 安徽：安徽教育出版社，2002：2-16.

[3] 马真. 现代汉语虚词研究方法论[M]. 北京：商务印书馆，2004：1-2.

[4] 郝斌. 虚词的语义及其翻译[J]. 中国俄语教学，2006，25(3)：24-28.

[5] 陆俭明，马真. 现代汉语虚词散论[M]. 北京：语文出版社，1999：179-253.

[6] 宗成庆. 统计自然语言处理[M]. 北京：清华大学出版社，2013：696.

[7] Rosenberg D. Stop，words[J]. Representations，2014，127(1)：83-92.

[8] 张坤丽，眷红英，柴玉梅. 现代汉语虚词用法知识库建设综述[J]. 中文信息学报，2015，29(3)：1-8.

[9] 张腾飞. 介词用法识别及其在信息抽取中的应用研究[D]. 郑州：郑州大学硕士学位论文，2013.

[10] 刘秋慧，张坤丽，许鸿飞. 助词“的”的用法自动识别研究[J]. 北京大学学报(自然科学版)，2018，54(3)：466-474.

[11] Pourdamghani N，Gao Y，Hermjakob U，et al. Aligning English strings with abstract meaning representation graphs[C]//Proceedings of the 2014 Confer-

ence on Empirical Methods in Natural Language Processing，2014：425-429.
[12] Zec D. Prosodic differences among function words[J]. Phonology，2005，22(1)：77-112.
[13] Chung S. The syntax and prosody of weak pronouns in Chamorro[J]. Linguistic Inquiry，2003，34(4)：547-599.
[14] Hacking J. Grammaticalization theory and the particle bi/by in Bulgarian，Macedonian and Russian[J]. Canadian Slavonic Papers，1999，41(3-4)：415-430.
[15] Heine B，Song K. On the grammaticalization of personal pronouns[J]. Journal of Linguistics，2011，47(3)：587-630.
[16] Bailey B. A model for function word counts[J]. The Royal Statistical Society，1990，39(1)：107-114.
[17] Mareček D，Žabokrtský Z. Dealing with function words in unsupervised dependency parsing[C]//Proceedings of the 15th International Conference Computational Linguistics and Intelligent Processing，2014：250-261.
[18] Tang G，Rao G，Yu D. Can we neglect function words in word embedding? [C]//Proceedings of the 5th CCF Conference on Natural Language Processing and Chinese Computing，NLPCC 2016 and 24th International Conference on Computer Processing of Oriental Languages，2016：541-548.
[19] Reich C，Reich P. The construction of an orally based sight-word vocabulary list and its relationship to the vocabularies of beginning readers[J]. The Journal of Educational Research，1979，72(04)：198-204.
[20] Rapp R. Using word association norms to measure corpus representativeness [C]//Proceedings of the 15th International Conference Computational Linguistics and Intelligent Processing，2014：1-7.
[21] 石毓智，李讷. 汉语语法化的历程[M]. 北京：北京大学出版社，2001：139-144.
[22] 孙中一. 汉语中虚词类表达概念初探[J]. 逻辑与语言学习，1987(04)：26-28.
[23] 郭锐. 语义结构和汉语虚词语义分析[J]. 世界汉语教学，2008(04)：5-17.
[24] 殷何辉. 焦点敏感算子“只”的量级用法和非量级用法[J]. 语言教学与研究，2009(01)：49-56.
[25] 杨松柠. 虚词语义描写新方法探析——以副词“就”为例[J]. 现代语文(语言研究版)，2013(05)：81-83.
[26] 张斌，范开泰. 现代汉语虚词研究丛书[M]. 合肥：安徽教育出版社，2002：2-4.
[27] 吴义诚，周永. “都”的显域和隐域[J]. 当代语言学，2019，21(2)：159-180.
[28] 张引兵，宋继华，彭炜明，等. 短语结构树库向句式结构树库的自动转换研究[J]. 中文信息学报，2018，32(05)：31-41.
[29] 刘一韬. 基于汉语虚词用法的语义角色标注研究[D]. 郑州：郑州大学硕士学位论文，2015.
[30] Osborne F，Gerdes K. The status of function words in dependency grammar：A critique of universal dependencies (UD)[J]. Glossa-A Journal of General Linguistic，2019，4(1)：17-28.
[31] Che W，Li Z，Liu T. LTP：A Chinese language technology platform [C]//Proceedings of the Coling 2010：Demonstrations，2010：13-16.
[32] 眷红英，张坤丽，柴玉梅，等. 现代汉语虚词知识库的研究[J]. 中文信息学报，2007，21(05)：107-111.
[33] 眷红英，张腾飞，林爱英. 基于介词用法的时间信息抽取研究[J]. 计算机工程与设计，2013，34(7)：2750-2754.
[34] Banarescu L，Bonial C，Cai S，et al. Abstract meaning representation for sembanking[C]//Proceeding of the 7th Linguistic Annotation Workshop and Interoperability with Discourse，2013：178-186.
[35] 李斌，闻媛，宋丽，等. 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.
[36] 卜丽君. 基于AMR的中文句子语义标注及统计分析[D]. 南京：南京师范大学硕士学位论文，2017.
[37] Cai S，Knight K. Smatch：An evaluation metric for semantic feature structures[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics，2013：748-752.

戴玉玲(1996—)，硕士研究生，主要研究领域为计算语言学。
E-mail：aislingdai@163.com

冯敏萱(1978—)，通信作者，博士，副教授，主要研究领域为计算语言学。
E-mail：fennel_2006@163.com

戴茹冰(1989—)，博士研究生，主要研究领域为计算语言学。
E-mail：ice_dr@163.com