汉语语义构词的资源建设与计算评估

王悦[1,2], 刘扬[1,2]*, 梁启亮[1,3], 王涵思[1,2]
[1]北京大学计算语言学教育部重点实验室，北京100871
[2]北京大学计算机学院，北京100871
[3]北京大学信息科学技术学院，北京100871
{wyy209,liuyang}@pku.edu.cn
lql_eecs@qq.com; whs1900014165@163.com

摘要

汉语是一种意合型语言，汉语中语素的构词方式与规律是描述、理解词义的重要因素。关于语素构词的方式，语言学界有语法构词与语义构词这两种观点，其中，语义构词对语素间关系的表达更为深入。本文采取语义构词的路线，基于语言学视角，考虑汉语构词特点，提出了一套面向计算的语义构词结构体系，通过随机森林自动标注与人工校验相结合的方式，构建汉语语义构词知识库，并在词义生成的任务上对该资源进行计算评估。实验取得了良好的结果，基于语义构词知识库的词义生成BLEU值达25.07，较此前的语法构词提升了3.17%，初步验证了这种知识表示方法的有效性。该知识表示方法与资源建设将为人文领域和信息处理等多方面的应用提供新的思路与方案。

关键词： 汉语语素；汉语语义构词；资源建设；词义生成

Construction of Chinese Semantic Word-Formation and its Computing Applications

Yue Wang[1,2], Yang Liu[1,2]*, Qiliang Liang[1,3], Hansi Wang[1,2]
[1]Key Lab of Computational Linguistics (MOE), Peking University, Beijing 100871
[2]School of Computer Science, Peking University, Beijing 100871
[3]School of Electronics Engineering and Computer Science, Peking University, Beijing 100871
{wyy209,liuyang}@pku.edu.cn
lql_eecs@qq.com; whs1900014165@163.com

Abstract

Chinese is a paratactic language, where the ways and rules of its word-formation play an important role in describing and understanding the meanings of words. There are two perspectives on morphemes and word-formation in linguistics: grammatical word-formation and semantic word-formation, with the latter indicating a deeper relationship between morphemes. In this paper, following the perspective of semantic word-formation, we propose a set of computing-oriented semantic word-formation labels based on characteristics of Chinese, build a Chinese semantic word-formation knowledge-base by combining random forest automatic labeling and manual verification, and evaluate the resource on the task of definition generation. Experimental results show that definitions generated from the semantic word-formation knowledge-base achieve a BLEU value of 25.07, which is 3.17% higher than previous grammatical

*通讯作者
基金项目：国家自然基金项目（62036001）、国家社科基金项目（18ZDA295）

word-formation approach. These findings confirm the effectiveness of our knowledge representation and resource construction, which may provide new insights into and solutions for a variety of tasks in humanities and computing applications.

Keywords: Chinese morphemes , Chinese semantic word-formation , Resource construction , Definition generation

1 引言

汉语是一种意合语言，普遍认为，汉语遵循从语素、词、短语到句子的层级结构。其中，语素作为汉语中最小的音义结合体，是构词的基本单位(朱德熙, 1982; 尹斌庸, 1984)。汉语中的语素在构词中非常活跃，处于重要的地位(徐枢, 1990)，语素义的组合在一定程度上能体现词义(符淮青, 1981)。汉语的词汇系统庞大，且不断产生新词、新义，但作为构词基本单元的语素，其数量与意义是相对稳定的，高达93%的汉语语素均为单音节语素，且87%以上的语素在构词时会保持意义不变(苑春法与黄昌宁, 1998)。而且，相对于句法分析和理解方面的工作(Xue and Palmer, 2003; You and Liu, 2005)，当前，词法分析方面的研究和资源建设还比较欠缺。在NLP领域，此前对词义进行表征的主流方法是利用词间信息(Zheng et al., 2013; Gui et al., 2018)，即通过目标词的上下文环境对词义进行表征。注意到语素构词的重要性和相关研究与开发的欠缺，我们希望将语言学观点引入到计算中，从语义构词的角度对汉语的词义进行表征。

关于语素构词的方式，在语言学界，主要有语法构词和语义构词这两种不同观点。

语法构词的观点认为，汉语构词的原则和汉语造句的原则基本一致，可以用主谓、动宾等句法结构标签对构词结构进行分类(陆志韦, 1964; 郭绍虞, 1979; 赵元任, 1980)。董秀芳(2011)认为，现代汉语中二字词的前身是古代汉语中单字词的自由句法组合，考虑形式上的构造性，也支持语法构词的看法。语法构词的结构体系较为简单，便于标注及计算处理(傅爱平, 2003)，因此，此前对语素构词的研究更多地关注语法构词(康司辰等, 2020)。

语义构词的观点认为,字与字之间是按语义关系组成词(刘叔新, 1990)，使用施事、受事等语义标签来分析构词语素之间的关系(朱彦, 2003)。周荐(2003)指出，在语素组合时，起决定作用的因素是它们是否能在意义、习惯上相搭配。相对而言，语义构词更符合汉语社团的思维模式(徐通锵, 1997)，对词内语素间关系的描述更为深刻(朱彦, 2003)，具有天然优势，采取该观点研究语素构词会带来一些额外的收益。但语义构词的结构较为复杂，且目前尚没有相对清晰的标准(傅爱平, 2003)，因此，当前对语义构词知识表示的研究还比较欠缺，从该角度研究语素构词也是一项有挑战性的工作。

注意到这些情况，我们希望在语法构词结构的基础上，提出一套新的语义构词结构体系；从语素出发，通过自动标注与人工校验相结合的方式构建语义构词知识库；并通过词义生成任务，对该知识表示的有效性进行计算评估。语义构词的体系规范、标注方法与资源建设，将对人文领域和信息处理等多方面的应用提供新的思路与方法。

本文结构如下：在引言中，介绍了语素构词的重要性和不同的路径方式；在第2节中，从汉语语义构词的理论探讨和相关语言资源研发等角度出发，对前人的相关工作做梳理和评述；在第3节中，结合汉语语素构词的特点，提出了一套面向理解和计算的汉语语义构词结构标签；在第4节中，基于上述标签集，对汉语中的二字词通过随机森林算法做结构标签的自动标注并进行人工校对，构建汉语语义构词知识库；在第5节中，在词义生成任务上对该知识表示的有效性进行计算评估，分析实验结果，并在新词数据集上做进一步的验证；在结语部分，总结本文方法与数据成果，并展望后续可能的研究领域与方向。

2 相关工作

2.1 汉语语义构词理论探讨

汉语语义构词关注的是构词语素之间的关系。在当代词汇语义学的研究中，生成词库理论(Pustejovsky, 1995)受到了极大的关注，被誉为是精细的形式化分析手段(Geeraerts, 2010)。基于该理论，词项的表征包括论元结构、事件结构、物性结构和词汇类型结构这四个层面。其中，论元结构为概念整合网络中的动词心理空间提供明晰的语义框架，物性结构为名词心理空间提供明晰的知识框架(王笑, 2017)。

论元结构采用语义角色（也称论元角色）来实现描述。它多以谓词性成分为核心，描述与中心谓词相关的语言成分在事件中所扮演的参与者角色，是语言学家对句子中有关结构成分之间意义关系的一种分类方式，可以为计算机处理语言提供较为充分的语义知识。但从经验上来说，想要系统、一致地给所有动词的全部配项标明语义格几乎是不可能的(Dowty, 1991)，因此，语义角色标签的划分至今仍然没有相对明晰的标准。现有的汉语语义角色标签划分方式有多种，它们多关注于句子级语义分析:袁毓林基于计算需求(2002)提出一种了包含4个层级、17种论元角色的结构体系，并描述了这些语义角色的定义和句法特征;刘茂福和胡慧君(2013)在袁毓林标签的基础上做了归并调整;鲁川(2010)将语义角色分为中枢语义角色和周边语义角色，并提出26种中枢语义角色和26种周边语义角色;宋衡等(2023)在鲁川研究的基础上，对周边语义角色做了一些调整，并将其分为主要周边语义角色和辅助周边语义角色两类，提出了一套包含60种语义角色的分类体系。在构词方面，顾阳和沈阳(2001)将论元结构引入构词理论研究中，朱彦(2003)利用论元结构，深入语义底层对4263个复合词进行分析研究。但由于构词与造句的差异，即词结构短小、无法容纳多个（或内部结构复杂的）论元，且词的上下文语境有不确定性，句子级的语义角色体系难以直接迁移到构词上。另外，汉语中的一些词不包含谓词性成分，无法用论元结构描述，单纯的论元结构在词内语义结构分析中带有一定的局限性。

物性结构是关于词汇本体知识的描述体系。与论元结构不同，它以名词为核心，使用各种物性角色从多个方面说明名词与其相关事物、事件和属性的关系。现有的物性角色标签划分方式有多种：Pustejovsky(1995)提出形式角色、构成角色、施成角色、功用角色这四种基础物性角色，后来又补充提出了规约化属性(Pustejovsky and Jezek, 2008)。袁毓林(2013)根据汉语名词在真实文本中的搭配情况，提出了10种物性角色；张念歆和宋作艳(2015)在研究形名组合定中式复合词时，在基础的四类物性角色上细分了小类；宋作艳等在研究名名(宋作艳等, 2015)、动名(宋作艳, 2022)复合词的物性结构时，在四种物性角色的基础上将规约化属性引入汉语，并在五个物性角色大类下细分小类。由于前人的研究通常只局限在特定的词性结构上，物性结构标签的划分至今同样没有相对明晰的标准，且这些研究多停留在理论语言学领域，面向计算应用的研究还比较匮乏。与论元结构的缺陷类似，汉语中的一些词也不包含体词性中心，无法用物性结构描述，单纯的物性结构在词内语义结构分析中也带有一定的局限性。

在词法分析中，还有一些其它结构不能被论元结构或物性结构覆盖。其中，包括属于单纯词、前（后）缀式和并列式等结构的词。前两类的词中至少有一个语素不表义，因此找不出论元或物性角色，后者则是因为两个语素的地位平等，找不到中心语素，所以也无法使用论元结构或物性结构来描述。在名+名联合式复合词上，王晗(2013)研究了《现代汉语词典》中的1130个此类复合词，并按前、后语素的意义关系把它们分为同义联合、相关联合和反义联合这三类，按语素和词的关系分为重合类、组合类、融合类这三类。刘明珠(2020)在同义联合、相关联合和反义联合这三类的基础上，补充了远义并列类，在语素和词的关系上则分为互注、选取、融合、合取这四类。而对于动+动的连动式，李可胜和满海霞(2013)提出了毗邻、聚合和加合这三种事件结构。

2.2 汉语构词相关语言资源研发

从中文信息处理的实践来看，此前的语义分析多集中在句子级别，对语素构词的研究还比较欠缺。在语言知识工程方面，目前具有较大影响的几项典型资源如下：

苑春法和黄昌宁(1998)的“汉语语素数据库”以语素描写和构词分析为核心，覆盖了6763个常用汉字的17470个语素项信息，包括语素义、语法类、构词方式等信息，并对这些语素项构成的43097个二字复合词进行了构词结构分析和语素项的绑定，并初步总结了汉语语素构词的规律。但在语素项上，仅仅形成了一个离散的集合而没有形成关联体系，缺乏面向整个语言系统的意义关联，难以满足现实的计算需求。

亢世勇(2004)的“汉语义类信息库”覆盖了6763个常见汉字的17430个字位（可理解为语素）的释义和词性，并与《同义词词林》（以下简称《词林》）中的语义分类体系进行了绑定。在此基础上，继续对52366个二字词中的每个字进行义类标注和简单释义，建立了“汉语语义构词信息库”。这两项工作对字位和二字词进行了归类并形成了积极的意义关联，其归类以现有《词林》为标准，存在着语素义与词义的因果参照问题，结构的合理性有待商榷。

吉志薇和冯敏萱(2015)提取了《现代汉语词典》（以下简称《现汉》）中的2268个词素并标注了每个词素所属的义类，构建了“词素-义类数据库”。在此基础上，标注和统计了8984个二字词的词素意义和词素间的词化意义（可理解为语素义和构词结构），构建了二字词语义描写

体系，并应用于二字未登录词的理解。目前其收录的二字词均包含前50高频词素，采样不均衡且数据规模过小，难以满足全局数据上的计算需求。

刘扬等(2018)研究开发了“汉语概念词典”，提取和编码《现汉》中全部8514个汉字的20855个语素释义，每个语素义具有一个唯一的编码，如“雄$_{1_05_01}$”代表“雄”字的某个语素释义；以这些全局信息为依据，进而采用“同义语素集”来表征“语素概念”，建立了“语素概念体系”。在此基础上，对《现汉》中41474个二字词的全部52108个义项赋予唯一的词条编码，进一步描述这些词的语法构词结构，实现了语法结构下的语素与其语素义的严格绑定，以此来诱导和表达汉语词义(陈龙等, 2019)，并在词义生成等任务上实现了应用(康司辰等, 2020; Zheng et al., 2021a; Zheng et al., 2021b)。其构词结构标签共分16种，分别为定中、联合、述宾、状中、单纯、连谓、后缀、述补、主谓、重叠、方位、介宾、名量、数量、前缀与复量(郑画等, 2022)。“汉语概念词典”注重表征构词语素与语素义的绑定，采取的是语法构词而并非语义构词，这些系统性的工作也为进一步提炼和标注语义构词标签提供了条件和可能。

3 面向计算的汉语语义构词结构体系

基于语义构词的技术路线和计算需求的现实考虑，我们首先从单音节语素构成的现代汉语二字词入手，在此前工作的基础上，构建了一套包含论元角色、物性角色和其它标签的汉语语义构词结构体系。其具体情况如下：

3.1 汉语语义构词的论元结构

对于谓词性中心的汉语二字词，在袁毓林(2002; 2013)等研究的基础上，考虑到二字词长度过短无法同时出现多个（或内部结构复杂的）论元，以及尽量消除在缺乏上下文情况下可能出现的结构歧义，我们归并了受事与对象等形式相近的论元结构，并加入了内容与事量，提出了一套包含14种论元角色的论元角色结构体系。其定义如表1所示：

论元角色	论元角色描述	示例
施事	自主性动作、行为的施行者	**人**造　**雷**鸣
感事	非自主性心理感觉的主体	**头**疼　**心**酸
主事	性质、状态或变化性事件的主体	**年**轻　**身**故
受事	动作、行为的承受者	吃**饭**　**皮**试
结果	动作、事件造成的影响、产物	扩**大**　录**像**
系事	在事件中与主事对应，表达主事的属性、类别	有**名**　当**官**
内容	言语行为、信息传递或心理活动的内容	讲**课**　求**婚**
工具	动作、行为所凭借的工具	**枪**毙　**珠**算
材料	动作、行为所用的材料，事件过程中所凭借和消耗的物品	**水**解　**铁**打
方式	动作、行为所使用的方式、方法	**周**游　**上**调
原因	动作、行为发生的原因	**仇**杀　**惊**醒
时间	动作、行为所发生的时间	**日**用　**春**训
空间	动作、行为所发生的空间	**家**访　**野**营
事量	事件所涉及的数量、频率、幅度	**多**疑　三**思**

Table 1: 论元角色与示例（示例中字体加粗部分为论元角色）

3.2 汉语语义构词的物性结构

对于体词性中心的汉语二字词，此前研究多局限于一种词性的组合。我们综合魏雪和袁毓林(2013)、宋作艳(2022)、张念歆和宋作艳(2015)等在名名、动名、形名组合方面的研究，结合汉语语素构词的特点，提出了一套包含17种物性角色的物性角色结构体系。其定义如表2所示：

3.3 汉语语义构词的其它结构

论元角色与物性角色适用于构词的前、后语素都表义且具有明确中心的情况，显然，它们无法覆盖全部的汉语二字词。对于前、后语素地位均等或至少一个语素不表义的汉语二字词，我们另设了单独的语义标签。其定义如表3所示：

第二十二届中国计算语言学大会论文集，第456页-第467页，哈尔滨，中国，2023年8月3日至5日。

物性角色	物性角色描述	示例
材料	物性角色是制成核心名词的材料	**木**板　**纸**钱
数目	物性角色是核心名词表示事物的数量	**单**亲　**七**彩
整体	物性角色是核心名词所属的整体	**果**皮　**羊**毛
领属	物性角色是核心名词的领主	**沙**俄　**人**情
成分	物性角色是核心名词的组成成分	**雨**点　**字**幕
上位	物性角色是核心名词的上位概念，核心名词是其中的一种	鲤**鱼**　氧**气**
外形	物性角色是核心名词的外在表现	蒜**黄**　**方**阵
评价	物性角色是对核心名词的主观评价	**真**品　**奸**商
单位	物性角色是作为核心名词的单位的量词	云**朵**　马**匹**
时间	物性角色是核心名词所表示事物所处的时间	**唐**诗　**早**饭
空间	物性角色是核心名词所表示事物所处的空间	**山**寨　**壁**画
方位	物性角色是方位词，整体词义指核心名词的某个方位	江**南**　**后**面
用途	物性角色是核心名词所表示事物用来做的事情	**玩**具　**鱼**网
用法	物性角色是使用核心名词所表示事物的方法	**挂**钟　**吊**灯
职能	物性角色是核心名词所表示事物（多为人）所从事的工作	**农**民　**教**师
施成	物性角色是核心动词所表示是事物的产生方式	**烤**鸭　**配**方
状态	物性角色是核心名词所表示事物的所处状态或常规活动	**沸**水　**飞**鸟

Table 2: 物性角色与示例（示例中字体加粗部分为物性角色）

其它结构	结构描述	示例
重合	词义与前、后语素义均相同或相近	错误　奶奶
组合	词义与前、后语素义均有关，前、后语素义不同且地位平等	左右　花草
顺序	词义与前、后语素义均有关，前、后语素是先后发生的动作或事件	签收　判断
偏义	词义仅与其中一个语素义有关	老**师**　**灿**烂
单纯	这个词是单独的语素	葡萄　端木

Table 3: 其它语义结构与用例（偏义结构示例中字体加粗部分为中心语素）

对于论元结构和物性结构无法覆盖汉语二字词的情况，第一种可能是，前、后语素均表义且地位相等，在语法结构上，这些词通常为并列结构类型。在语义结构上，则根据词义与语素义、前语素义与后语素义之间的依赖关系，可以进一步把它们分为重合、组合和顺序结构；第二种可能是，二字词中有且仅有一个语素与词义关系较强，这类词既包含“老师”“兔子”这样前（后）缀式的词，也包含因一个语素义脱落导致词义偏向另一个语素义的联合式的词，如“国家”“灿烂”等；第三种可能是，词义与前、后语素义均没有关联，比如一些单纯词，包括“沙发”“端木”“葡萄”等。在语言学中，通常把这些词视为独立语素。我们重点关注在现代汉语中占绝大多数的单字语素，为了计算上的形式一致性，对于这些单纯词，视构成该词的前、后语素为空语素，标注单纯结构。

4　汉语语义构词的资源建设

4.1　面向计算的汉语语义构词知识表示

表4展示了“汉语概念词典”中既有的语法构词描述信息。其中，对于多义词的不同义项，视为不同的词条分别标注。示例中的前（后）语素义指该词中前（后）语素的语义，用构词语素和语素义的绑定来做表达。我们希望从这些信息入手，通过自动标注与人工校验相结合的方式构建汉语语义构词知识库。王洪君(2000)曾指出，现代汉语中绝大多数的双音节复合词可以用句法结构的形式理解，因此，语法结构信息的应用有助于达成语义结构的识别，对语义结构的自动标注有极大的帮助。我们要开展的资源建设工作是将既有的语法结构信息拓展为语义结构信息，该信息包括语义构词结构和中心语素位置，其中，语义构词结构指前、后语素之间的语义关系，中心语素位置指在词义表征中占核心地位的语素位置。

第二十二届中国计算语言学大会论文集，第456页-第467页，哈尔滨，中国，2023年8月3日至5日。

词	上天$_1$	上天$_2$	上天$_3$
词义	上升到天空	用作婉辞，指人死亡	迷信者指主宰自然和人类的天
语法结构	述宾	述宾	定中
前语素	上$_{2_14_01}$	上$_{2_14_02}$	上$_{1_06_01}$
前语素义	由低处到高处	到；去	位置在高处的
后语素	天$_{1_12_01}$	天$_{1_12_11}$	天$_{1_12_10}$
后语素义	天空	迷信者指神佛仙人所住的地方	迷信者指自然界的主宰者
例句	人造卫星～	（无）	～保佑

Table 4: “汉语概念词典”中既有的语法构词描述信息

我们采用自动标注与人工校验相结合的方式对每个词的语义构词结构进行标注。标注工作由三名标注人员完成，初始抽取6000个词条，由两名标注人员独立标注语义构词结构和中心语素位置；依据这些标注信息，采用随机森林对“汉语概念词典”中的其余二字词的语义构词结构和中心语素位置进行自动标注；最后，由第三人对全部标注结果进行人工校验，作为最终的标注结果。

4.2 基于CART决策树和随机森林的语义构词结构自动标注

由于人工标注难以实现较大的样本覆盖，考虑到语义构词结构的分类多达33类，部分结构存在样本偏少、难以有效覆盖的问题，这种情况下，主流的深度学习算法无法有效捕捉这些特征。借鉴魏雪和袁毓林(2013)等对名名组合词构建释义模板的工作思路，该方法类似于人工构建决策树，且有较好的可解释性，因此，我们采用CART决策树和随机森林算法对汉语语义构词结构进行初步标注。

决策树的分类特征为：$features = \{fm, mor_1, mor_2, sim\}$，这是为每个待标词输入决策树的特征集。在该特征集中，1)fm为语法构词结构；2)$mor_i = \{m_{i1}, m_{i2}, \ldots, m_{i11}(i = 1, 2)\}$为第$i$个语素的嵌入向量表示。我们使用语素概念体系层级结构中的语素概念路径信息对每个语素进行向量表征：m_{ij}表示第i个语素在该体系的第j层下的子节点序号；3)为了衡量不同语素对词义贡献的大小，增加相似度信息$sim = \{s_{1*}, s_{2*}, s_{12}, \frac{s_{1*}}{s_{2*}}\}$。其中，$s_{1*}, s_{2*}, s_{12}$分别表示前语素与词、后语素与词、前语素与后语素之间的相似度，$\frac{s_{1*}}{s_{2*}}$则是用于衡量前、后语素对词义贡献的比重，这主要是考虑不同的语义结构对语素义贡献的侧重不同。比如“猪獾”和“鲤鱼”，它们的语法结构均为定中，前、后语素都属于动物类，但语义结构明显不同，“猪獾”是“外形像猪的獾”，侧重于后语素，而“鲤鱼”是“品种属于鲤的鱼”，侧重于前语素。注意到决策树算法每次只选择一个特征进行划分的特点，因此，把前、后语素对词义贡献的比重单独作为一个维度，便于决策树提取这一特征。

我们将语法构词结构、语素嵌入向量和相似度信息拼接起来，作为每个待分类词的特征集输入决策树。为了避免过学习，另进行了决策树剪枝和语素嵌入向量降维。决策树剪枝涉及树的最大深度、每个节点的最小样本数等因素，并将最大深度和最小叶节点样本数均设为9；语素嵌入向量降维则实验性截取每个语素嵌入向量的前n个维度，这主要考虑语素概念体系中偏底层节点的粒度过细，特征不明显，易对计算造成干扰。我们集成采用随机分类特征生成的100棵CART决策树，构成随机森林，随机森林仅对语素嵌入向量降维、不做剪枝。然后将决策树算法和随机森林算法的分类结果进行对比，以此选择最优模型。为了充分利用数据，使用10折交叉验证来衡量分类的准确率。

我们用上述算法对汉语的语义构词结构进行自动标注的分类，分类结果如表5所示，其中，语素嵌入向量为7维的随机森林效果最好，33分类下的准确率达74.26%。增大或减小向量的维数均导致准确率下降：维度较大时，向量的后端会出现大量零值，失去了语素义表征的价值；而维度较小时，意义粒度的表征过粗，无法表达相对精确的语素义。此外，为了验证语法结构信息和相似度信息在算法中的作用，我们也做了消融实验，在随机森林算法上分别去掉这些信息。结果表明，在最优的7维语素嵌入向量上，去掉语法构词结构信息，分类的准确率下降了28.86%；去掉相似度信息，分类的准确率下降了1.50%。这验证了在分类算法中采纳语法构词结构信息和相似度信息的重要性。

模型/语素嵌入向量维度	11	8	7	6
随机基准模型	3.18	3.00	2.82	3.10
决策树（不剪枝）	70.65	70.53	71.07	72.50
决策树（剪枝）	73.35	73.52	74.02	73.98
随机森林	73.52	73.72	**74.26**	73.63
随机森林w/o fm	55.92	56.00	55.40	54.80
随机森林w/o sim	72.91	73.05	72.76	72.68

Table 5: 不同向量维度下决策树算法与随机森林算法的分类准确率

4.3 汉语语义构词知识库构建

使用效果最好的随机森林算法，我们对未经人工标注的46108个二字词条进行自动标注，并在自动标注的基础上进行人工校验。此外，考虑到一些词的语义结构相同但中心语素位置不同，除了结构信息外，根据中心语素位置，我们给出了用于词义表达的语素义序列。语素义序列的生成规则为：对使用论元结构的词，将核心谓词排列在前，论元排列在后；对使用物性结构的词，将核心名词排列在前，物性角色排列在后；对偏义式的词，将与词义关联较强的语素排列在前，与词义关联较弱的语素排列在后；对并列式与单纯词，语素义依然按照在词中出现的顺序排列。表6给出了新的语义构词知识表示的示例：

词	结构	语素义序列	语素义1	语素义2
植树	受事	<植$_{1_04_01}$，树$_{1_04_01}$>	栽种	木本植物的通称
谣传$_2$	受事	<传$_{1_08_03}$，谣$_{1_03_02}$>	传播	谣言
竹器	材料	<器$_{1_05_01}$，竹$_{1_02_01}$>	器具	竹子
花束	单位	<花$_{1_18_02}$，束$_{1_05_02}$>	可供观赏的植物	用于捆在一起的东西
灿烂	偏义	<灿$_{1_01_01}$，烂$_{0_00_00}$>	光彩耀眼	<空语素>
诞生	重合	<诞$_{1_02_01}$，生$_{1_10_01}$>	诞生	生育；出生

Table 6: 语义构词知识表示示例

我们在“汉语概念词典”的基础上，将原有的语法构词结构替换为上述语义构词结构与语素义序列信息，构建了汉语语义构词知识库，知识库中的词相关信息包括：词，语义构词结构、语素义、语素义序列、词义和例句。我们的知识库涵盖了41474个二字词的52108个义项，基本实现了对《现汉》中二字词的覆盖。

5 汉语语义构词的计算评估

在语义构词思路和资源建设的有效性验证方面，Noraset(2017)提出的词义生成任务是一种评估知识表示质量的恰当且自然的方式，并有直观、良好的可解释性。该任务的目标是依据给定的词相关信息，由机器自动生成针对该词的新的释义文本，此前也被用于生成词向量的质量评估。需要指出的是，我们做语义构词结构自动标注并不使用词的释义文本信息，且最终的词义表示知识中也不直接包含词的任何释义文本信息，用词义生成任务来做计算评估是相当严苛的一项考验。

5.1 词义生成的基础模型

当前的词义生成模型依据注入特征通常分为三类：第一种是基于预训练词向量的，包括Noraset(2017)的SG模型，由于注入的特征单一，其生成效果较差，且无法区分多义词的不同义项；第二种是在词向量的基础上追加语料，包括Gadetsky(2018)基于AdaGram和注意力机制的模型和Ishiwatari(2019)的LOG-CaD模型，在词向量之外增加了上下文向量信息，该方法的有效性严重依赖上下文质量；第三种是基于知识库的算法，包括基于HowNet义原的AAM、SAAM模型(Yang et al., 2020)和基于“汉语概念词典”中语法构词结构与语素义标注

的DeFT模型(Zheng et al., 2021a)。在这种情况下，DeFT模型也是验证新的语义构词知识表示的适合的比较基准。

5.2 基于语义构词的改进模型

DeFT模型的输入序列是(w, fm, mor_1, mor_2, C)。其中，w是词形信息，fm是构词结构，mor_1与mor_2分别为前、后语素的语素义，C是例句信息。使用的特征包括如下五项：词向量w、字向量$ch = (ch_1, ch_2)$、语法结构fm、语素义向量$mor = (mor_1, mor_2)$、例句向量C。该模型中使用的语法结构是单向的，在我们构建的知识库中，语义结构是双向的：如“受事”标签既有可能表示前语素是后语素的受事，也可能表示后语素是前语素的受事，二者的支配关系由语素义序列决定。语素义序列是按照语素义的重要性排列的，并非词中的原始语素顺序，因此，在特征的注入上有两种考虑：第一种是依据模型结构调整数据格式，将语素义按照原始的语素顺序排列，这种方法简单快捷，但对少量语义标签难以判断前、后语素的语义关系，造成相关信息的混淆与损失；第二种是对模型进行改进以适应数据格式，对于语素义，按此前界定的语素义序列直接做输入，如果语素义序列中后语素在前，则对字向量信息也进行反向处理。改进后的模型结构如图1所示：

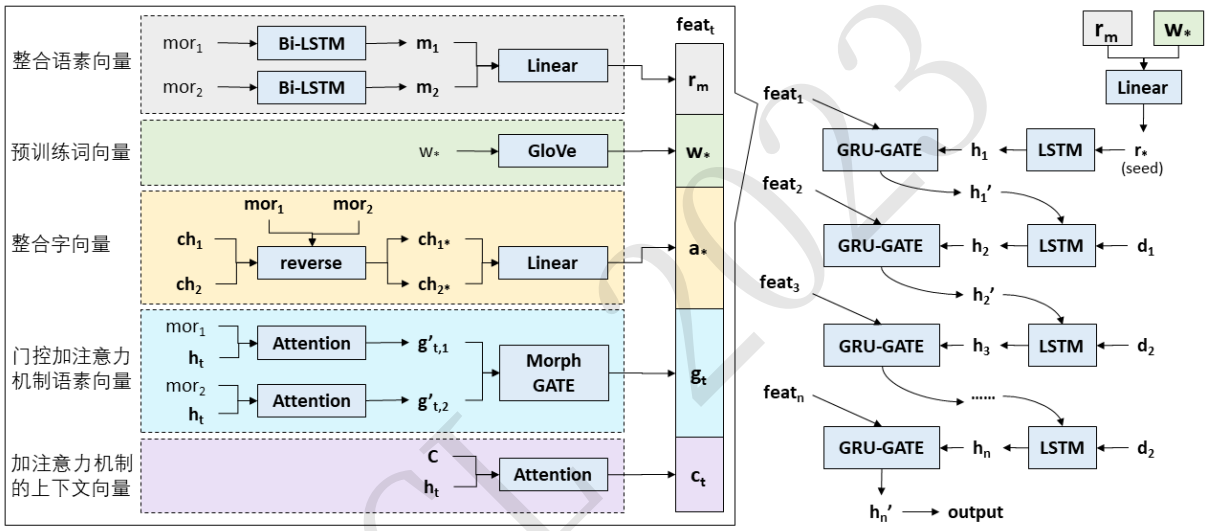

Figure 1: 改进后的DeFT+reverse模型结构图

其中，Linear是对每个构词结构表达特异性的线性层，MorphGATE通过一个线性层和一个sigmoid层对两个语素向量做权重分配，字向量和词向量使用预训练的fastText(Bojanowskiet al., 2017)。对改进后的模型，mor_1与mor_2是按语素义序列规约的语素义向量，$ch1$与$ch2$分别是词中的第一、第二个字向量，reverse函数判断字向量与语素义向量是否匹配，如不匹配则切换两个字向量的顺序。

依据“汉语概念词典”中既有的语法构词结构信息与新构建知识库中的语义构词结构及中心语素位置信息，按模型要求输入的数据格式做统一处理，我们分别建立服务于词义生成的语法构词数据集和语义构词数据集，将它们按8:1:1的比例分为训练集、测试集和验证集，并确保每个二字词在不同构词结构下出现在同样功能的集子里，且多义词的所有义项也都出现在同样功能的集子里。

在参数设置上，使用fastText词向量对词信息进行初始化，词向量的维度为300，Bi-LSTM的隐层大小为300，训练批次大小为64，随机失活率0.2。优化器使用Adam，学习率初始化为0.001。在每个epoch结束时计算验证集上的BLEU值，若出现连续6个epoch无提升则学习率乘3；为避免过学习，若连续12个epoch效果无提升则停止训练，保存验证集上BLEU值最高的作为最终模型。

第二十二届中国计算语言学大会论文集，第456页-第467页，哈尔滨，中国，2023年8月3日至5日。

5.3 实验结果与分析

我们分别在三种模型与输入的组合上进行词义生成实验：第一种是使用语法构词结构的DeFT，以此作为基准模型；第二种是使用语义构词结构的DeFT模型，不考虑语素义序列的规约；第三种是使用语义构词结构的DeFT+reverse改进模型，考虑语素义序列的规约。利用这三种模型和不同的构词结构信息进行训练，并在相同的测试集上进行测试。词义生成的示例如表7所示：

词	**理财**	**雨具**
词义	管理财物或财务	防雨的用具
语法结构	述宾	定中
语义结构	受事	用途
语素义序列	<理$_{1_07_04}$，财$_{1_02_01}$>	<具$_{1_03_01}$，雨$_{2_02_01}$>
前语素义	管理；办理	从云层中降向地面的水
后语素义	钱和物资的总称	用具
例句	当家～｜～之道	（无）
DeFT-语法结构	对资产的资产	可以做的用具
DeFT-语义结构	对一定的资金、资金等	防雨的用具
DeFT+reverse-语义结构	管理资金	防雨的用具

Table 7: 词义生成结果示例（中间六行为词的相关信息，后三行为藉此产生的词义生成结果）

对上述三种模型的生成结果进行分析：在“理财”一词中，语义结构为“受事”。一般而言，“受事”标签下的核心谓词位置不能完全确定，在不考虑语素义序列的规约时，仅使用“受事”标签无法区分核心谓词及其论元，难以捕捉有效的语义特征。考虑了语素义序列的规约后，生成结果得到了进一步的提升。在“雨具”一词中，语义结构为“用途”，“用途”标签下的核心名词几乎全部为后语素，语素义序列的规约对生成结果的影响较小。相比较而言，语法结构里的“定中”标签过于宽泛，该标签下词的语义结构存在极大差异，难以捕捉到有效的语义特征，导致生成的词义缺乏限制、过于笼统。

为定量评价词义生成的结果，使用自动评测指标BLEU进行评估，结果如表8所示。易见，在其它注入信息相同的情况下，仅把语法构词结构替换成语义构词结构，BLEU值就能获得1.56%的提升，即便不考虑语素义序列信息带来语义关系的混淆和损失。对模型进行改进，增加语素义序列信息，BLEU值达到25.07，较语法构词结构提升了3.17%。这些，初步验证了语义构词思路和资源建设的优势。

模型	**BLEU**
DeFT-语法结构*	24.30
DeFT-语义结构	24.68(+1.56%)
DeFT+reverse-语义结构	**25.07(+3.17%)**

Table 8: 词义生成结果评估（*为基准模型，最佳结果加粗表示）

5.4 在新词上的推广探讨

为了验证语义构词知识表示的可推广性，我们继续在新词上进行词义生成评估。新词的特点和难度在于它催生了新的词形、词义，并可能衍生出新的语素义界定，这也为新词、新义的理解和计算带来了挑战。我们在郑画等(2022)构建的新词数据集的基础上，结合词义生成任务的需求和新的语义构词资源做改进：使用全局数据上重新训练过的随机森林，对所有新词的语义构词结构进行自动标注；考虑原数据集中的新词释义与《现汉》中词的释义在长度上的差异，使用GPT3.5对新词释义进行适当简化；同时，考虑GPT简化释义与《现汉》释义在行文

风格上的差异，请一名标注人员参照《现汉》给出新词的人工释义。在改进后的新词数据集上，我们使用上述三种模型进行词义生成评估，其结果如表9所示：

模型	BLEU(GPT简化释义)	BLEU(人工释义)
DeFT-语法结构*	13.76	18.59
DeFT-语义结构	14.98(+8.87%)	20.50(+10.27%)
DeFT+reverse-语义结构	**15.98(+13.89%)**	**22.19(+19.37%)**

Table 9: 新词的词义生成结果评估（*为基准模型，最佳结果加粗表示）

实验结果表明，考虑了语素义序列信息的改进模型在GPT简化释义、人工释义上的BLEU值分别为15.98、22.19，相较于语法结构信息分别大幅提升了13.89%、19.37%，此试验结果也符合主实验中的总体趋势，且提升幅度十分显著。这进一步验证了语义构词路线的优势和资源建设的有效性，表明本文的知识表示与标注算法可以进一步推广到新词上。

但另一方面，对比主实验中的生成结果，上述三个模型在新词上的BLEU值均有所降低。我们猜测，导致这一现象的原因可能有如下三个方面：1)新词中存在谐音应用和新语素义衍生等问题，使得相关语素无法在《现汉》中找到对应而被标为“空语素”。如“美眉”是“妹妹”的谐音而不是“美丽的眉毛”，“潮妈”中的“潮”指“新潮的”，但在目前的《现汉》中，“潮”字并没有这些语素义界定。事实上，在新词数据集中，有高达11.1%的词中会出现“空语素”的情况，有效语素义表征的缺失导致这些词的词义生成结果不够理想。例如“潮妈”，三个模型的词义生成结果均体现了“妈”的语素义，而曲解或忽略了“潮”的语素义；2)大量新词中存在转喻、隐喻等非字面义的情况(陈龙等, 2019)。如“草根”指“社会中的中低收入群体”而非“草本植物的根部”，“孩奴”指“因为孩子的养育成本而感到经济压力的父母”而非“孩子的奴隶”，这种低语义透明度的状态削弱了词和词义之间的直接联系。因此，目前生成的词义往往倾向于选择词的字面义，例如“孩奴”，三个模型的生成结果均接近于字面义，无法体现非字面义；3)GPT简化释义的初始来源为中文维基百科，与训练集中的《现汉》词的释义风格不同，这种风格差异会导致BLEU值的降低。例如“辅警”的GPT简化释义为“辅助警察，是协助正规警察提供额外警察力量的人员”，而在《现汉》中，与其结构相同、意义相近的“巡警”释义为“巡逻、维持治安的警察”，二者风格有较大差异。目前最优模型生成的结果是“指辅助工作的警察”，这与《现汉》更为相似，而与GPT简化释义存在风格差异，倾向导致BLEU值偏低。使用人工释义降低释义风格的影响后，最优模型的BLEU值达22.19，较主实验仅降低了2.88。

以上情况表明，面对历时变化的语言应用，现有词典中语素的语义空间划分存在一定的缺憾，无法反映并覆盖新词中可能衍生出的新语素义。注意到语义构词知识表示在新词上取得的显著提升，我们认为在语义构词标注工程的基础上，通过新的计算性手段的导入，有可能推测出新词衍生出的新语素义，该路径将为汉语的语言文字研究和词典编纂提供帮助。

6 结语

考虑汉语语素构词的特点，在前人研究的基础上，我们提出了一套面向计算的汉语语义构词结构体系；以此为指导，通过自动标注和人工校验相结合的方式构建了语义构词知识库；我们将新的知识表示应用于词义生成评估，取得了良好的效果：基于语义构词的词义生成BLEU值达25.07，较此前的语法构词提升了3.17%，显示了语义构词思路和资源建设的优势。同时，为了验证语义构词知识表示的可推广性，进一步将其应用于新词的词义生成，较语法构词的提升幅度也十分显著。

在后续工作中，我们计划将语义构词知识表示推广到汉语的多字词上，并利用资源建设成果进一步提升语义构词自动标注的准确性，以便更好地服务于人文领域和信息处理等多方面的应用，如词典编撰与浏览、汉语教育与研究、词向量训练及应用、词义消歧、语素义消歧、未登录词识别及语义预测等，为这些应用提供新的路径与方法。

参考文献

Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transaction of the Association for Computational Linguistics*, 135-146.

David Dowty. 1991. Thematic Proto-Role and Argument Selection. *Language*, 67(3):547-619.

Artyom Gadetsky, Ilya Yakubovskiy, Dmitry Vetrov. 2018. Conditional Generators of Words Definitions. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 266–271.

Dirk Geeraerts. 2010. *Theories of Lexical Semantics*. Oxford University Press, New York.

Tao Gui, Qi Zhang, Jingjing Gong, Minlong Peng, Di Liang, Keyu Ding and Xuanjing Huang. 2018. Transferring from formal newswire domain with hypernet for twitter POS tagging. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2540-2549.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, Masaru Kitsuregawa. 2019. Learning to describe unknown phrases with local and global contexts. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3467-3476.

Thanapon Noraset, Chen Liang, Larry Birnbaum, Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 3259-3266.

James Pustejovsky. 1995. *The Generative Lexicon*. The MIT Press, Mass.

James Pustejovsky, Elisabetta Jezek. 2008. Semantic Coercion in Language: Beyond Distributional Analysis. *Italia Journal of Linguistics*, 20(1):181-214.

Nianwen Xue, Martha Palmer. 2003. Annotating the Propositions in the Penn Chinese Treebank. *Proceedings of the Second SIGHAN Workshop*, 47-54.

Liner Yang, Cunliang Kong, Yun Chen, Yang Liu, Qinan Fan, Erhong Yang. 2020. Incorporating sememes into Chinese definition modeling. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 28:1669–1677.

Liping You, Kaiying Liu. 2005. Building Chinese FrameNet database. *Conference on Natural Language Processing and Knowledge Engineering. New York: IEEE*, 301-306.

Hua Zheng, Damai Dai, Lei Li, Tianyu Liu, Zhifang Sui, Baobao Chang, Yang Liu. 2021. Decompose, fuse and generate: A formation-informed method for chinese definition generation. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5524-5531.

Hua Zheng, Lei Li, Damai Dai, Deli Chen, Tianyu Liu, Xu Sun, Yang Liu. 2021. Leveraging Word-Formation Knowledge for Chinese Word Sense Disambiguation. *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021*, 918-923.

Xiaoqing Zheng, Hanyang Chen, Tianyu Xu. 2013. Deep learning for Chinese word segmentation and POS tagging. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 647-657.

陈龙, 饶琪, 刘扬. 2019. 汉语词的非字面义的表示与应用. 中国科学:信息科学: 49:1005-1018.

董秀芳. 2011. 词汇化：汉语双音词的衍生与发展. 商务印书馆, 北京.

符淮青. 1981. 词义和构成词的语素义的关系. 辞书研究, 1:98-110.

傅爱平. 2003. 汉语信息处理中单字的构词方式与合成词的识别和理解. 语言文字应用, (04):25-33.

葛本仪. 2001. 现代汉语词汇学. 山东人民出版社, 济南.

顾阳, 沈阳. 2001. 汉语合成复合词的构造过程. 中国语文, (02):122-133+191.

郭绍虞. 1979. 汉语语法修辞新探. 商务印书馆, 北京.

吉志薇, 冯敏萱. 2015. 面向普通未登录词理解的二字词语义构词研究. 中文信息学报, 29(05):63-68+83.

第二十二届中国计算语言学大会论文集，第456页-第467页，哈尔滨，中国，2023年8月3日至5日。

康司辰, 虞梦夏, 刘扬. 2020. 基于平行周遍原则的汉语未登录词的知识表示与预测. 中文信息学报, 34(08):23-31.

亢世勇, 李毅, 孙道功, 张楠. 2004. 汉语系统语料库的建设与词典编纂. *2004*年辞书与数字化研讨会论文集, 145-151.

李可胜, 满海霞. 2013. VP的有界性与连动式的事件结构. 现代外语, 36(02):127-134+218.

刘茂福, 胡慧君. 2013. 基于认知与计算的事件语义学研究. 科学出版社, 北京.

刘明珠. 2020. 现代汉语NN型并列式复合词的生成机制探究. 西北大学硕士论文, 西安.

刘叔新. 1990. 汉语描写词汇学. 商务印书馆, 北京.

刘扬, 林子, 康司辰. 2018. 汉语的语素概念提取与语义构词分析. 中文信息学报, 32(02):12-21.

鲁川. 2010. 知识工程语言学. 清华大学出版社, 北京.

陆志韦. 1964. 汉语的构词法. 科学出版社, 北京.

宋衡, 曹存根, 王亚, 王石. 2023. 一种改进的汉语语义角色分类体系与标注实践. 中文信息学报, 37(01):16-32.

宋作艳, 赵青青, 亢世勇. 2015. 汉语复合名词语义信息标注词库：基于生成词库理论. 中文信息学报, 29(03):27-33+43.

宋作艳. 2022. 基于构式理论与物性结构的动名定中复合词研究——从动词视角到名词视角. 世界汉语教学, 36(01):33-48.

徐枢. 1990. 语素. 人民教育出版社, 北京.

徐通锵. 1997. 语言论：语义型语言的结构原理和研究方法. 东北师范大学出版社, 长春.

王晗. 2013. 现代汉语名+名联合式双音复合词研究. 山东大学博士论文, 济南.

王洪君. 2000. 汉语语法的基本单位与研究策略. 语言教学与研究, 02:10-18.

王笑. 2017. 物性结构与论元结构视域下汉语语义构词研究——以a+b=c类双音合成词为例. 鲁东大学硕士论文, 烟台.

魏雪, 袁毓林. 2013. 基于语义类和物性角色建构名名组合的释义模板. 世界汉语教学, 27(02):172-181.

尹斌庸. 1984. 汉语语素的定量研究. 中国语文, (05):338-347.

袁毓林. 2002. 论元角色的层级关系和语义特征. 世界汉语教学, 03:10-22+2.

袁毓林. 2013. 基于生成词库论和论元结构理论的语义知识体系研究. 中文信息学报, 27(06):23-30.

苑春法, 黄昌宁. 1998. 基于语素数据库的汉语语素及构词研究. 世界汉语教学, (02):8-13.

张念歆, 宋作艳. 2015. 汉语形名复合词的语义建构：基于物性结构与概念整合理论. 中文信息学报, 29(06):38-45.

赵元任. 1980. 中国话的文法. 香港中文大学出版社, 香港.

郑画, 刘扬, 殷雅琦, 王悦, 代达劢. 2022. 基于词信息嵌入的汉语构词结构识别研究. 中文信息学报, 36(05):31-40+66.

周荐. 2003. 论词的构成、结构和地位. 中国语文, (02):148-155+192.

朱德熙. 1982. 语法讲义. 商务印书馆, 北京.

朱彦. 2003. 汉语复合词语义构词法研究. 华东师范大学博士论文, 上海.