文章编号：1003-0077（2011）00-0000-00

汉语未登录词的词义知识表示及语义预测*

田元贺[1]，刘扬[2]

（1. 北京大学中国语言文学系，北京 100871；2. 北京大学计算语言学研究所，北京 100871）

摘要：在此前的汉语未登录词语义预测中，构词相关的知识一直被当做预测的手段，而没有被视为一种有价值的知识表示方式，本文在“语素概念”基础上，深入考察汉语的语义构词知识，给出未登录词的“多层面”的词义知识表示方案。针对该方案，本文采用贝叶斯网络方法，构建面向汉语未登录词的自动语义构词分析模型，该模型能有效预测未登录词的“多层面”的词义知识。这种词义知识表示简单、直观、易于拓展，实验表明对汉语未登录词的语义预测具有重要的价值，可以满足不同层次的应用需求。

关键词：未登录词；词义知识表示；语义预测；语义构词

中图分类号：TP391　　**文献标识码：**A

Lexical Knowledge Representation and Sense Guessing of Chinese Unknown Words

Tian Yuanhe[1], Liu Yang[2]

(1. Department of Chinese Language and Literature, Peking University, Beijing 100871;
2. Institute of Computational Linguistics, Peking University, Beijing 100871)

Abstract: In the previous researches in sense guessing of Chinese unknown words, the lexical knowledge related to word-formation has been used as a means to guess, and is not regarded as a valuable form of knowledge representation. This paper, on the basis of the morphemic concepts, with respect to the lexical knowledge of semantic word-formation, provides a multi-level solution to knowledge representation of Chinese unknown words. A model based on Bayesian network is also constructed to analyze semantic word-formation of Chinese unknown words, effectively predicting the multi-level lexical knowledge of Chinese unknown words. This kind of lexical knowledge representation is simple, intuitive and easy to expand. Experimental results show that, this knowledge representation is of important value in sense guessing of Chinese unknown words, and can meet the application needs on different levels.

Key words: Chinese unknown words; lexical knowledge representation; sense guessing; semantic word-formation

1 引言

词义知识的表示和获取是文本理解的基础。在中文信息处理的实践中，汉语未登录词的频繁出现，对机器理解提出了很大的挑战，其语义预测对智能信息检索、机器翻译等典型应用具有重要价值。目前这一领域的研究仍处于起步阶段。

未登录词的语义预测涉及两个方面：预测内容以及预测方法。

在预测内容上，此前的研究 [0000000] 主要是预测未登录词的语义类别，也有预测概念图 [0] 和语义构词知识 [0] 的。语义类别的预测是一种粗线条的预测，只能表示特定语义分类下的大概的词义，而对精细化的词义需求无能为力。例如，对于“选材”一词，语义类别的预测一般将其设定为“获取”这个义类，而“获取”的“具体内容”无法直观得到。相比之下，概念图以图的形式表示构词概念之间的关系，表达的词义信息要多于语义类别，然而，对于未登录词来说，这种表示形式过于复杂，既不直观、也不利于计算。

关于词义，符淮青 [0] 等多位语言学家指出：语素义的组合可以在一定程度上体现词义。因此，将语义构词知识作为词义知识表示并对其进行预测将是一种新的选择。这种词义知识表示具有简单、直观的特点，能够全面、充分地反映构词语素对词义的贡献。例如：在“选材”中，“选”有语素义“挑选”，“材”有语素义“有才能的人”，其“述宾”结构关系及成分意义，准确地反映了“选材”的语义，在精度上高于语义类别，在复杂程度上低于概念图。吉志薇 [0] 是目前唯一尝试预测未登录词语义构词知识，并给出预测方法的人。但遗憾

* **收稿日期：**　　**定稿日期：**

基金项目：国家社科基金一般项目（16BYY137）、国家重点基础研究发展计划资助项目（2014CB340504）、国家社科基金重大项目（12&ZD119）

的是，她只是简单地将未登录词的语义构词知识作为词义知识输出，而没有注意到“多层面”的构词知识在实际应用上的巨大价值，并且，不同的语素及其意义之间无法形成有效的关联，此外，她的实验结果也不太理想。

在预测方法上，目前主要有两类，即基于语料的方法和基于词内部知识的方法。此前，Lu[0]和 Chen[0]尝试了基于语料的方法，并用这种方法预测语义类别，Lu 的 F 值为 37.1%，Chen 的准确率为 34.4%。此外，Lu 还提供了基于词内部知识的方法。

相比之下，基于词内部知识方法的研究较多，结果也更理想。Lu[0]、Chen[0]、Tseng[0]、Chen[0]、邱立坤[0]等基于《同义词词林》（以下简称《词林》）和《知网》，尚芬芬[0]基于《现代汉语语义词典》，均预测未登录词的语义类别。值得一提的是，他们采用的预测模型，例如“重叠字模型”、“字类别关联模型”等，都用了语义构词分析的思路，却没有意识到可以将语义构词知识应用于词义知识表示。在结果方面，Lu 的准确率为 61.6%，Chen[0]对名词的准确率为 81.0%，Tseng 对名词、动词、形容词的准确率分别为 71.4%、52.8%、65.8%，Chen[0]对双音节 V-V 复合词的准确率为 61.6%，邱立坤的 F 值为 64.7%，尚芬芬的准确率为 77.9%。此外，张瑞霞[0]基于《知网》预测概念图的准确率为 79.3%。吉志薇[0]虽然给出了语义构词知识的预测方法，却没能得到一个整体上的结果，其部分结果（准确率为 43.7%）也因为实验样本少而缺乏足够的代表性。

在这些基础上，我们研究未登录词的语义预测，既包括预测内容，也包括预测方法。

首先，我们关注系统的语义构词知识与词义知识表示之间的关联，原则上，这种表示对已登录词、未登录词都是适用的。针对完整给出未登录词词义知识的难点，我们探究“多层面”的词义知识表示在应用需求上的价值；接下来，设计针对汉语未登录词的自动语义构词分析模型，预测未登录词的“多层面”的词义知识，实现对未登录词的词义预测。

需要说明的是，二字词在汉语中占据主体，对它的研究具有代表性，因此，目前的研究以二字未登录词为主。本文中的知识表示和预测方法具有良好的扩展性，可以方便地拓展到三字及以上未登录词的情形。

2 汉语的语义构词知识及“多层面”的知识表示

凡是对词的理解有意义的语义构词知识，在中文信息处理应用中都是有价值的。因此，本文所讲的语义构词知识，涵盖词性、构词结构、语素类、语素义等在内的广义知识。汉语未登陆词的语义预测也将以此为基础给出，以便在广泛的意义层面上来表示词义。

课题组研发多年并计划推出的北京大学《汉语概念词典》（以下简称《概念词典》，英文名称 the Chinese Object-Oriented Lexicon，简称 COOL）在生成词库理论（GLT 理论）[0]、面向对象思想（OO 思想）[0]、WordNet 理论[0]等观点指导下，以《现代汉语词典（第 5 版）》（以下简称《现汉》）刻画的汉语的语素及语素义为依据，采用“同义语素集”来表征“语素概念”并建立“语素概念体系”；在此基础上，详尽描述汉语词的构词结构，并实现构词结构下的构词成分（即语素）对“语素概念体系”中的“语素概念”的严格绑定，以此来诱导和表达汉语词义，并提供多种应用程序接口。

《概念词典》包含的词的这些语义构词知识，构成本文工作的一个数据基础。

2.1 语义构词知识

2.1.1 词性知识

《概念词典》为其收录的词都标注了词性，其中，51454 个二字词的情况如表 1 所示。

表 1 《概念词典》中二字词词性统计表

词性	数量	比率	例词
名词	25720	49.99%	丈夫
动词	18679	36.30%	上升
形容词	5543	10.77%	严峻
副词	905	1.76%	临时
数词	57	0.11%	好多
量词	90	0.17%	公尺
介词	36	0.07%	为了

代词	114	0.22%	咱们
助词	23	0.04%	不得
叹词	10	0.02%	呜呼
拟声词	115	0.22%	乒乓
连词	162	0.31%	不但
合计	51454	100.00%	

2.1.2 构词结构知识

在语言学界有两种主流的构词结构体系，一种注重表达构词语素间的语义关系（如主体、客体等），而另一种体系注重表达构词语素间的语法关系（如主谓、述宾等）。对于第一种构词体系，傅爱平[0]指出：虽然其在表示词义时更具优势，但是其结构体系较为复杂，对计算机来说，识别难度较大。相比之下，第二种构词体系较为简单，结构标准较为统一，且与句法结构有天然的相似性，苑春法[0]的研究表明，基于语法的构词结构与构词语素类和词性之间存在一定的相关性。因此，采用第二种构词体系更有利于计算的开展。实际上，由于后续要求构词成分对“语素概念”严格绑定，我们获得的依然是广义的语义构词知识。

基于以上分析，我们参考杨梅[0]和北京大学中文系郭锐教授对构词结构的研究成果，构建了基于语法的构词体系，并为《概念词典》中所有 52108 个二字词按义项区分标注了构词结构，见表 2。为保证构词结构知识的可靠性，我们请三位专家对同一词项进行标注，两人以上标注结果相同的一致率为 93.46%。

表 2 《概念词典》二字词构词结构统计表

构词结构	数量	比率	例词
主谓	524	1.01%	年轻
连谓	1709	3.28%	进攻
联合	11414	21.90%	丰满
述宾	8141	15.62%	选材
述补	630	1.21%	提高
定中	19581	37.58%	红旗
状中	4215	8.09%	热爱
介宾	157	0.30%	从小
重叠	310	0.59%	哥哥
名量	78	0.15%	纸张
数量	56	0.11%	一些
方位	189	0.36%	野外
复量	20	0.04%	场次
前附加	698	1.34%	老虎
后附加	2308	4.43%	忘却
单纯词	2078	3.99%	克隆
合计	52108	100.00%	

2.1.3 语素类知识

语言学上的“语素”指的是“最小的音义结合体”，在本文中，为方便起见，汉语语素暂且限定为一个汉字。由于《现汉》只为部分（主要是成词语素，约 48%）语素标注了语素类，我们采用专家人工标注的方式补齐了其余（主要是不成词语素，约 52%）的语素类，《概念词典》全部 20175 个语素的语素类知识见表 3。

表 3 《概念词典》语素类统计表

语素类	数量	比率	例词
名语素	9782	48.49%	仗（打仗）
动语素	6331	31.38%	习（学习）
形语素	2346	11.63%	光（光滑）

副语素	572	2.84%	万（万全）
数语素	80	0.40%	一（一世）
量语素	414	2.05%	年（光年）
介语素	113	0.56%	按（按例）
代语素	129	0.64%	何（何必）
助语素	121	0.60%	者（或者）
叹语素	84	0.42%	嘘（嘘唏）
拟声素	104	0.52%	乒（乒乓）
连语素	70	0.35%	若（倘若）
缀语素	29	0.14%	们（我们）
合计	20175	100.00%	

2.1.4 语素义知识

此前，学界对于语素义系统的研究较少。亢世勇[9]曾构建了《汉字义类信息库》，但他所选取的义类体系源于《词林》，用这种词义体系对字义分类的做法难免偏颇。借鉴 WordNet 理论，课题组成员陆顾婧[9]在其硕士论文中用“语素特征”（现在称其为“语素概念”）来称谓汉语中可计算的最小意义单元，并采用“同义语素集”的形式来加以表示，该集合中的元素为具有相同或基本相同意义（即语素义）的那些语素，其中的每个语素都携有独特的“语素义编码”。例如，语素“选”有多个语素义，其中的一个语素义的“语素义编码”为“选 1_04_01”，这表明：它是该单字在词典中的第 1 次条目出现（即“选 1”），该条目共有 4 个义项（即“选 1_04”），当前为第 1 个义项（即“选 1_04_01”）。

目前，对以上 20175 个语素所表达的语素义，我们按释义计算相似度，形成初步的“同义语素集”，并经反复的人工校对、核对，获得了 5113 个“语素概念”。在这些“语素概念”之间，我们进一步构建了初步的上、下位语义关系，形成了一个树状结构的“语素概念体系”。在后续的知识表示中，如果确定了特定语素的语素义，携有了“语素义编码”，就意味着特定语素在该体系中绑定了一个“语素概念”，并接受该体系的意义表达和约束。

以表达“选择、挑选”意义的动语素“语素概念”X 为例，X={刷 3_01_01，抡 1_01_01，拔 1_08_03，拣 1_01_01，择 1_02_01，择 2_02_01，挑 1_02_01，擢 1_02_02，调 4_02_02，选 1_04_01，遴 1_01_01，铨 1_02_01}，在“语素概念体系”中，其所处的“语素概念”位置如图 1 所示。

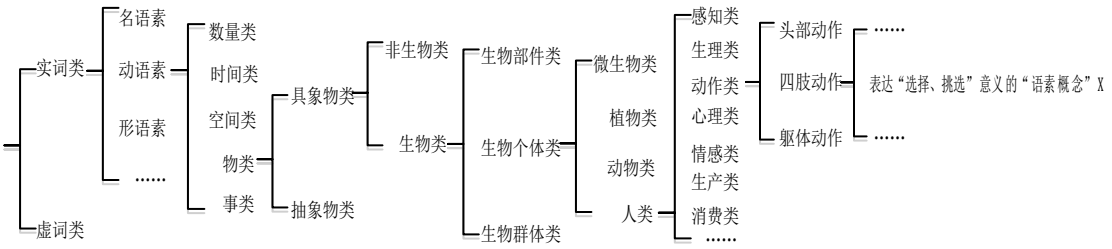

图 1 树状结构的“语素概念体系”示例

在标注二字词的构词结构和前、后语素类后，我们继续把《概念词典》中所有二字词的前、后语素按其语素义与对应的“语素义编码”挂钩。于是，二字词的前、后语素与它们在“语素概念体系”中的“语素概念”就建立了严格绑定关系。这样一来，单一的语素义就拥有了更丰富的、便于计算的意义形式。

2.2 已登录词的“全层面”的词义知识表示

对《概念词典》中的二字词，在以上语义构词分析之后，我们获得了由词性、构词结构、语素类和语素义等四方面知识构成的一个“全层面”的词义知识表示。其中，前三个层面属于语法层，最后一个层面属于语义层。以“选材”一词为例，“选”表示“挑选、选择”的“语素概念”，“材”表示“有才能的人”的“语素概念”。鉴于“语素概念”中的每个语素都携有独特的“语素义编码”，为方便起见，语素对应的“语素概念”只以“语素义编码”的形式标出，“选材”的“全层面”的词义知识表示如表 4 所示。

表 4“选材”的词义知识表示

词	词性	构词结构	前语素类	后语素类	前语素义	后语素义
选材	动词	述宾结构	动语素	名语素	选 1_04_01	材 1_05_04

为了诱导词义的简化的表达形式，我们在构词结构和词义之间搭建意义关联。

亢世勇[0]曾给出包括 A+B=A=B、A+B=A、A+B=B、A+B=C、A+B=A+B、A+B=A+B+D、A+B=A+D、A+B=D+B 等 8 种形式的意义结构体系，其中，A、B 分别表示二字词的前语素义和后语素义，C 代表转义后的意义，D 代表附加意义。这种体系分类详细，但转义和附加义的知识较难于获取，在实际应用中面临较大的挑战。陆顾婧[0]提出了一种简单、方便计算的意义结构体系，见表 5，这也是我们目前采用的方案。需要指出的是，为方便起见，该意义结构体系省略了转义和附加义等附加因素，目前只考虑词的字面意义，即本义。转义和附加义的问题在后续层面单独加以表达和解决，这里不再赘述。例如，“铁窗”有“监狱”的意思，目前只考虑其字面义“铁的窗户”，其转义问题可以在后续阶段加以表示和处理，并不会因此丢失。

表 5 意义结构与构词结构的对应关系

意义结构	语素义和词义的关系	构词结构	例词
00 型	词义与前后语素义相关性均较低	单纯词	克隆、名堂
01 型	词义只与后语素相关性较高	前附加	老虎、仔细
10 型	词义只与前语素相关性较高	后附加、名量	忘却、纸张
11 型	词义与前后语素义相关性均较高	主谓、连谓、联合、述宾、述补、定中、状中、介宾、数量、方位、复量	丰满、红旗

在此基础上，我们给出了词的“意义序列”的输出形式。该序列为构词语素的“语素义编码”的排列，内容和顺序基本由构词结构决定，详情见表 6。以“选材”为例，其“意义序列”一般为“<选 1_04_01,材 1_05_04>”，此外，允许在应用需求中依据约定改变序列顺序，以表达计算的灵活性，如“<材 1_05_04,选 1_04_01>”也是一个合法的“意义序列”。考虑“语素概念体系”的意义表达和约束，词的“意义序列”表达词义的精细程度高于词的语义类别，而复杂程度低于概念图。

表 6 词的“意义序列”示例

例词	意义结构	构词结构	词的“意义序列”
年轻	11 型	主谓	〈轻 1_09_01,年 1_11_04〉
进攻	11 型	连谓	〈进 1_06_01,攻 1_04_01〉
丰满	11 型	联合	〈丰 1_03_01,满 1_07_01〉
选材	11 型	述宾	〈选 1_04_01,材 1_05_04〉
提高	11 型	述补	〈提 2_10_02,高 1_08_05〉
红旗	11 型	定中	〈旗 1_06_01,红 1_06_01〉
热爱	11 型	状中	〈爱 1_05_01,热 1_10_05〉
从小	11 型	介宾	〈从 2_04_01,小 1_09_06〉
哥哥	11 型	重叠	〈哥 1_04_01〉
纸张	10 型	名量	〈纸 1_03_01〉
一些	11 型	数量	〈些 1_02_01,一 1_10_01〉
野外	11 型	方位	〈外 1_08_01,野 1_07_01〉
场次	11 型	复量	〈场 2_09_07,次 1_08_05〉
忘却	10 型	后附加	〈忘 1_01_01〉
老虎	01 型	前附加	〈虎 1_04_01〉
克隆	00 型	单纯词	〈〉

对于三字和多字词，可以采取分层迭代的方法来获取“意义序列”[0]。例如，先将“乱弹琴”输出为“<<弹琴>,乱 1_06_01>”（“乱弹琴”是状中结构），再将“弹琴”输出为“<弹 2_06_04,琴 1_03_02>”（“弹琴”是述宾结构），而完整收集的“意义序列”为“<<弹 2_06_04,琴 1_03_02>, 乱 1_06_01>”。

2.3 未登录词的“多层面”的词义知识表示

语义构词知识涵盖不同层面，单一层面或多个层面的知识都有助于未登录词的理解,有其独特意义和应用价值。比如，未登录词的词性知识有助于句法分析器性能的提高。再如，未登录词的构词结构知识决定了构词语素对整体词义贡献的差异，对于单纯词类型，获取构词结构知识就够了；对于前附加、后附加、重叠结构、名量结构等类型，还需要获取单一语素义知识；对于其它构词结构类型，在获取构词结构知识的同时，获取单一语素义和全部语素义知识都有价值，这取决于具体的应用需求。例如，对于“红旗”，如果关注对象的属性，那么只需获取前语素义知识，如果关注对象本身，那么只需获取后语素义知识，如果关注整体意义，那么就需要获取所有语素义知识。此外，在某些应用中，甚至语素类都扮演重要角色。例如，如果关注“弹琴”中的独立的实体对象，那么只需分别判别“弹”和“琴”的语素类知识，并据此获取其中的名词性语素的语素义知识即可。

因此，依据应用需求的不同，可以选取不同层面的语义构词知识进行预测并加以组合，以达到对未登录词意义的有效把握，我们称其为“多层面”的词义知识表示。其优点在于，在满足需求的同时，避免了预测“全层面”的词义知识表示的困难，减少了需要预测的知识数目，有助于预测方法性能的提高。

在未登录词的“多层面”的词义知识表示的基础上，其“意义序列”的输出遵循同样的规范，这里不再赘述。

3 基于贝叶斯网络的语义构词分析模型

语义构词知识包括词性、构词结构、语素类和语素义等，苑春法[0]、王淑华[0]等人的研究表明，这些语义构词知识之间具有一定的相关性。因此，可以尝试从二字未登录词的词型出发，以推理的方式获取这些知识。贝叶斯网络正好提供了推理的概率手段，可以用于各种语义构词知识组合性的预测，满足词义知识表示的多层次需求。在本文研究中，我们以贝叶斯最优分类器算法[0]为基础，构建语义构词分析模型。

为表述方便，做如下约定：D表示训练数据，H表示假设空间，$X_{前字}$表示前语素，$X_{后字}$表示后语素，$X_{前类}$表示前语素类，$X_{后类}$表示后语素类，$X_{前义}$表示前语素义，$X_{后义}$表示后语素义，$X_{词性}$表示词性，$X_{结构}$表示构词结构。于是，$X_{前类}$、$X_{后类}$、$X_{前义}$、$X_{后义}$、$X_{词性}$、$X_{结构}$构成了二字未登录词ab（$X_{前字}=a$、$X_{后字}=b$）的语义构词知识，而 V表示依据需求不同而被选入当前词义知识表示的语义构词知识。语义构词分析模型的任务就是预测 V 中最优的语义构词知识组合，即：

$$\underset{v_j\in V}{argmax}\left(P(v_j|D,a,b)\right)=\underset{v_j\in V}{argmax}\left(\sum_{h_i\in H}P(v_j|h_i,a,b)P(h_i|D)\right)\cdots\cdots①$$

进一步，由贝叶斯公式，有：

$$P(h_i|D)=\frac{P(D|h_i)P(h_i)}{P(D)}$$

不妨假设，$P(D)$是常数，且$P(h_i)=P(h_j)$对任何$h_i,h_j\in H$成立。于是，①转化为：

$$\underset{v_j\in V}{argmax}\left(\sum_{h_i\in H}P(v_j|h_i,a,b)P(D|h_i)\right)$$

此外，定义：

$$P(D|h_i):=\frac{h_i在D中预测准确的实例数}{D中全部实例数}$$

接下来，只要给出求解$P(v_j|h_i,a,b)$的方法，就可以预测需要的语义构词知识。为此，需要先构建假设空间 H。

3.1 假设空间的构建

对于贝叶斯网络来说，不同的假设对应于语义构词知识间不同的条件独立性，也对应了不同的网络结构和推理过程。

我们认为，语义构词知识的预测由以下 3 个任务顺序组成：1、语素类知识$X_{前类}$和$X_{后类}$的预测；2、语素义知识$X_{前义}$和$X_{后义}$的预测；3、词性知识$X_{词性}$和结构知识$X_{结构}$的预测。其中，任务 1 有 3 种推理模式：①前字->前类，后字->后类；②后字->后类，前字、后类->前类；③前字->前类，后字、前类->后类；任务 2 有 4 种推理模式：①前字、后类->前义，后字、前义->后义；②后字、前类->后义，前字、后义->前类；③前字、后类->前义，后字、前类

->后义；④前字->前义，后字->后义（该推理模式不使用前类和后类的特征）；任务 3 只有一种推理模式：前义、后义->词性，前义、后义、词性->结构。综上所述，共有$(3\times3+1)\times1=10$种推理模式，分别对应了假设空间中 10 种可能的假设。

举例来说，选取任务 1 中的推理模式③、任务 2 中的推理模式①和任务 3 中的推理模式，它们组成的一种假设的贝叶斯网络如图 2 所示。

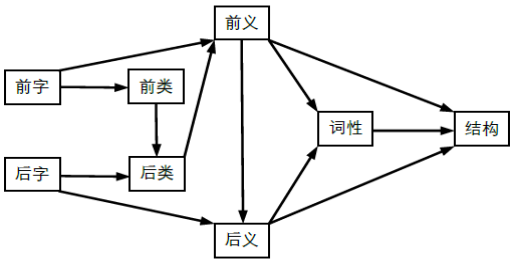

图 2　一种假设的贝叶斯网络

3.2 $P(v_j|h_i,a,b)$的计算方法

V 中不同的语义构词知识组合对应了不同的计算方法，设$V=\{(X_{m_1},X_{m_2},\ldots\ldots,X_{m_n})\}$，其中$X_{m_i}$是 V 中的语义构词知识。

于是，有：

$$P(v_j|h_i,a,b)=P(x_{m_1},x_{m_2},\ldots\ldots,x_{m_n}|h_i,a,b)$$

进一步，由全概公式，有：

$$P(v_j|h_i,a,b)=\sum_{X_m\in V}P\left(x_{前义},x_{后义},x_{词性},x_{结构},x_{前类},x_{后类}\middle|h_i,a,b\right)$$

依据假设h_i下的条件独立性，可以给出$P(v_j|a,b)$的计算方法。

特别地，当h_i为图 2 所示的假设时，有：

$$P(v_j|h_i,a,b)=\sum_{X_m\in V}P\left(x_{前类}\middle|a\right)P\left(x_{后类}\middle|b,x_{前类}\right)P\left(x_{前义}\middle|a,x_{后类}\right)P\left(x_{后义}\middle|b,x_{前义}\right)P\left(x_{词性}\middle|x_{前义},x_{后义}\right)P\left(x_{结构}\middle|x_{前义},x_{后义},x_{词性}\right)$$

当$V=\left\{\left(X_{前义},X_{后义},X_{词性},X_{结构}\right)\right\}$时，有：

$$P(v_j|h_i,a,b)=\sum_{x_{前类},x_{后类},}P\left(x_{前类}\middle|a\right)P\left(x_{后类}\middle|b,x_{前类}\right)P\left(x_{前义}\middle|a,x_{后类}\right)P\left(x_{后义}\middle|b,x_{前义}\right)P\left(x_{词性}\middle|x_{前义},x_{后义}\right)P\left(x_{结构}\middle|x_{前义},x_{后义},x_{词性}\right)$$

其它假设和语义构词知识组合的计算方法与此类似。

3.3 数据稀疏问题的应对方法

对于数据稀疏问题，有两种应对方法：

方法 1 是使用结构简单的假设推理。在假设空间的 10 个假设中，既有贝叶斯网络结构十分复杂的假设，如图 2，也有十分简单的假设，如图 3。理论上，这种假设可以覆盖《概念词典》中全部二字词，增强了模型的适用性。

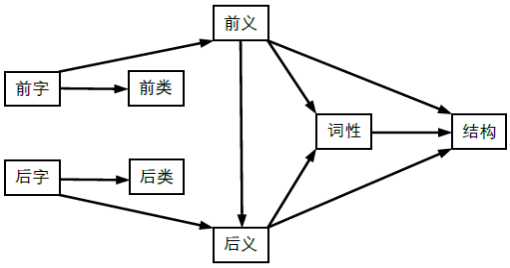

图 3 一种结构弱化的贝叶斯网络

方法 2 是在推理中使用“语素概念体系”中的上层“语素概念”节点。由于在全体“语素概念”间构建起了树状结构，当使用上层节点的语义知识进行推理时，发生数据稀疏问题的可能性大大降低。

4 实验结果与数据分析

4.1 实验数据说明

如前文所述，我们请多位专家对《概念词典》中所有的二字词标注了构词结构、语素类和语素义等语义构词知识。对以上标注结果，按如下原则计算人工标注的准确率：①如果三

人标注一样，则三人均正确；②如果两人标注一样，则标注一样的两人正确，另一人错误；③如果三人标注均不相同，则三人均错误。人工标注的准确率见表 7，由于《概念词典》中已有词性知识，不需要人工标注，所以没有给出相关数据。

表 7 语义构词知识人工标注情况

比较项目	准确率
构词结构	78.86%
前语素类	91.70%
后语素类	92.39%
前语素类＋后语素类	86.51%
前语素义	84.14%
后语素义	85.21%
前语素义＋后语素义	72.99%
词性＋构词结构＋前语素义	68.98%
词性＋构词结构＋后语素义	70.36%
词性＋构词结构＋前语素义＋后语素义	61.87%

对全部二字词整理之后，共得到 41472 个不同词型的语义构词知识，这些将作为我们的实验数据。未登录词通常从语料中筛选并使用模型对其进行语义预测，但是，这样的未登录词缺乏作为标准的正确的语义构词知识，无法给出模型的预测准确率，无法评价模型效果。基于这些考虑，本文实验的训练数据和测试数据均选自《概念词典》，我们将以上词型随机十等分，采用十折交叉验证的方法来检验模型效果，即轮流将其中九份作为训练数据，剩下一份作为测试数据。这样一来，对模型而言，每轮测试数据中的词即未登录词。

4.2 实验结果和分析

首先，在未对实验数据做筛选的情况下，语义构词分析模型可以处理所有二字未登录词，不同语义构词知识及其组合的预测准确率，见表 8、表 9、表 10。从这些结果不难看出，随着预测语义构词知识种类的增多和叠加，其准确率也随之下降。结合前文分析，这也表明，使用自动方法获取“全层面”的语义构词知识是有难度的，在当前，“多层面”的词义知识表示更具有现实意义。

表 8 语法层的语义构词知识预测准确率

试验项目	词性	构词结构	词性＋构词结构	前语素类	后语素类	前语素类＋后语素类
十折平均	78.41%	67.18%	60.88%	84.17%	87.17%	75.45%

表 9 语义层的语义构词知识预测准确率

试验编号	前语素义	后语素义	前语素义＋后语素义
十折平均	63.29%	66.23%	43.24%

表 10“语法＋语义”层的语义构词知识预测准确率

试验项目	词性+构词结构＋前语素义	词性+构词结构＋后语素义	词性+构词结构＋单一语素义	词性+构词结构＋前语素义＋后语素义
十折平均	41.60%	44.22%	53.96%	30.26%

接下来，将人工标注的准确率和自动方法进行比较，见表 11、表 12、表 13。由于无需人工标注词性，所以表中没有“词性”和“词性＋构词结构”的比较项目。可以发现，人工标注的准确率在一些项目上并不高，比如，人工在“词性＋构词结构＋前语素义＋后语素义”项目的准确率为 61.87%，而这一结果是建立在标注专家已知词性和词义的基础上的。这意味着，如果让人和机器处于同样的条件下——只知词型而不知词义和词性，那么人工标注的准确率应该比目前的更低。这恰好表明，使用自动方法准确获取“全层面”的语义构词知识在目前充满挑战，即使预测模型能够改善，人工标注的准确率便是可供参考的上限。相反，预测部分的语义构词知识，即“多层面”的语义构词知识，由于其准确率较高，更应成为自动方法关注的焦点。

表 11 语法层的人工与模型准确率比较

比较项目	构词结构	前语素类	后语素类	前语素类＋后语素类
人工标注	78.86%	91.70%	92.39%	86.51%
模型预测	67.18%	84.17%	87.17%	75.45%

表 12 语义层的人工与模型准确率比较

比较项目	前语素义	后语素义	前语素义＋后语素义
人工标注	84.14%	85.21%	72.99%
模型预测	63.29%	66.23%	43.24%

表 13“语法＋语义”层的人工与模型准确率比较

比较项目	词性＋构词结构＋前语素义	词性＋构词结构＋后语素义	词性＋构词结构＋前语素义＋后语素义
人工标注	68.98%	70.36%	61.87%
模型预测	41.60%	44.22%	30.26%

进一步，结合各个构词结构的统计数据（见表 2），我们发现“多层面”的词义知识表示的价值更加突现。例如，如果只获取后语素义知识，那么对 3.28%（连谓）＋21.90%（联合）＋37.58%（定中）＋8.09%（状中）＋0.59％（重叠）＋0.11%（数量）＋0.36％（方位）＋1.34%（前附加）＝73.25%的二字词有较准确的意义把握。

最后，将实验结果与前人研究进行比较：1、假定二字词的后语素义基本决定了它的语义类别，那么我们对语义类别的预测准确率达到 66.23%，这一结果和现有的研究 [000000] 基本相当，区别在于，我们给出了完全精准的语素义，其背后有“语素概念体系”的表达和约束，而此前给出的是单一的语义类别；2、在此前的实验中，预测语义类别以及预测概念图的研究 [0]，都是将语料中出现的未登录词作为测试数据——实际上，“能产性构词”类型的未登录词在语料中占了很大的比例，其语义预测更加有规律可循。相比之下，本文实验的测试数据是在《概念词典》中随机抽取，其中属于“能产性构词”类型的词并不多。在测试数据的预测难度和适用范围上，本文研究优于此前的研究；3、同样预测语义构词知识的研究 [0] 给出了预测方法，但该方法建立在 71 个专门挑选的未登录词上，不具有代表性，也没能给出完整的实验结果。与该方法的部分实验结果（其准确率为 43.7%）相比，我们在“语法＋语义”层的预测结果与之大致相当，此外，我们在“语素概念”基础上建立不同语素及其意义之间的广泛关联，语义构词知识的广度和深度都有新的提升。

5 结语

综上所述，本文研究的贡献体现在如下两个方面：

一、在预测内容上，此前的汉语未登录词语义预测，构词相关的知识一直被当做预测的手段，而没有被视为一种有价值的知识表示方式，我们在“语素概念”基础上，深入考察汉语的语义构词知识，给出未登录词的“多层面”的词义知识表示方案。这种“多层面”的词义知识表示,针对未登录词的完全语义预测的困难,可以依据不同的任务性质和指标要求，给出不同的语义构词知识及组合，表现出高度的灵活度和可裁剪性，预测结果简单、直观、易于应用。

二、在预测方法上，针对“多层面”的词义知识表示的需求，我们采用贝叶斯网络方法预测未登录词的多样化的语义构词知识。该模型实现简单，可以依据任务需求的变化快速给出相应结果，可以预测任何汉语二字词，表现出良好的适用性。与同样预测语义构词知识的此前方法想比，本文方法首次给出了整体实验结果，与此前部分实验结果的预测准确率相当。此外，该方法能够预测精准的语素义，其背后也有“语素概念体系”的表达和约束，而此前给出的多是单一的语义类别。

总体上看，未登录词的语义预测仍旧是研究上的难点，“多层面”的词义知识表示不失为一种有效的应对方案，它通过对预测内容的选取和组合，可以满足不同应用对不同层面词义知识的灵活需求。但是，也应看到，我们对未登录词的词义知识表示和语义构词分析进行了初步的探索，所使用的语义资源和分析技术仍有较大的提高和改善的空间，这也是未来

需要继续展开的工作。此外，目前只探讨了汉语二字词的情形，多字词资源仍在加紧开发中，将研究成果由二字词拓展到多字词，也是我们下一步需要展开的工作。

参考文献

[1]Lu X. Hybrid Models for Semantic Classification of Chinese Unknown Words[C]//HLT-NAACL. 2007: 188-195.

[2]Chen H H, Lin C C. Sense-tagging Chinese corpus[C]//Proceedings of the second workshop on Chinese language processing: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 12. Association for Computational Linguistics, 2000: 7-14.

[3]Chen K J, Chen C. Automatic semantic classification for Chinese unknown compound nouns[C]//Proceedings of the 18th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2000: 173-179.

[4]Tseng H. Semantic classification of Chinese unknown words[C]//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 2. Association for Computational Linguistics, 2003: 72-79.

[5]Chen C J. Character-sense association and compounding template similarity: Automatic semantic classification of Chinese compounds[C]//Proceedings of the 3rd SIGHAN Workshop on Chinese Language Processing. 2004: 33-40.

[6]邱立坤.现代汉语未登录词词类和语义类标注研究[D].北京大学,2010.

[7]尚芬芬，顾彦慧，戴茹冰，等.基于《现代汉语语义词典》的未登录词语义预测研究[J].北京大学学报(自然科学版)，2016，01：10-16.

[8]张瑞霞，杨国增，闫新庆.基于知网的汉语普通未登录词语义分析模型[J].计算机应用与软件，2012，08：126-130.

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

[10]符淮青.词义和构成词的语素义的关系[J].辞书研究，1981，01:98-110.

[11]Pustejovsky, J. The Generative Lexicon[M]. Mass: MIT Press, 1994.

[12]Grady Booch, Robert A. Maksimchuk, Michael W. Engle, etc. Object-Oriented Analysis and Design with Applications, 3rd Edition[M]. Addison-Wesley Professional, 2007.

[13]Fellbaum C. WordNet: An Electronic Lexical Database [M]. Mass: MIT Press, 1998.

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

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

[16]杨梅.现代汉语合成词构词研究[D].南京师范大学，2006.

[17]亢世勇，李毅，孙道功，等.汉语系统语料库的建设与词典编纂[C]//上海辞书学会.2004 年辞书与数字化研讨会论文集.上海辞书学会：2004：7.

[18]陆顾婧.汉语构词分析与词义知识表示研究[D].北京大学,2013.

[19]王淑华.双字组合理解模式探索[J].上海大学学报(社会科学版)，2007，03:43-47.

[20]Tom M. Mitchell 著，曾华军，张银奎译.机器学习[M].北京：机械工业出版社，2014：125-126.

作者简介：

田元贺（1994-），男，本科，主要研究领域为应用语言学、语言知识工程、中文信息处理。

Email: tianyh94@sina.com

刘扬（1971-），男，博士，副教授，主要研究领域为语言知识工程、中文信息处理。

Email: liuyang@pku.edu.cn

联系方式：

田元贺 北京大学中国语言文学系 北京 100871 13520891191 tianyh94@sina.com

刘扬 北京大学计算语言学研究所 北京 100871 13021117630 liuyang@pku.edu.cn