文章编号：1003-0077（2011）00-0000-00

基于知识库的汉语未登录词语义预测*

瞿健菊[1,2]，冯敏萱[2]

（1. 吉首大学师范学院，湖南 吉首 416000；2. 南京师范大学 文学院，江苏 南京 210097）

摘要：未登录词语义预测是自然语言处理研究的难点。该文基于知识库的语素构词知识，采用了分阶段的算法自动预测未登录词的语素构词知识，以此实现对未登录词的语义预测。基本思路是通过语素义组合或语素义类组合的匹配，先预测语义层面的知识，再确定相应语素项，最终获得未登录词多层面的语素构词知识。该算法简单、直观、合理，在首素性类、首素义类、首素义、尾素性类、尾素义类、尾素义、构词方式这七项预测内容全部正确的标准下，实验结果的预测正确率为62.32%，召回率为61.72%。

关键词：未登录词；语义预测；语言知识库；构词；知网

中图分类号：TP391　　**文献标识码：**A

Sense Prediction of Chinese Unknown Words Based on Knowledge Base

QU Jianju[1,2], FENG Minxuan[2]

(1.Normal College of Jishou University, Jishou, Hunan 416000, China;

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

Abstract: Sense prediction of unknown words is a difficult problem in natural language processing. Based on word-formation knowledge in knowledge base, this paper uses the phased algorithm to automatically predict word-formation knowledge of unknown words. Through the combination of morpheme meaning or the combination of morpheme's semantic category, this method first predicts the knowledge of semantic level, then determines the corresponding morphemes, finally gets the knowledge of word-formation of unknown words. The algorithm is simple, intuitive and reasonable. The experimental criteria is seven predictive content are all correct, which are the first morpheme's parts of speech, the first morpheme's semantic category, the first morpheme's meaning, the last morpheme's parts of speech, the last morpheme's semantic category, the last morpheme's meaning, grammatical structure type. The experimental results show that the prediction accuracy is 62.32% and the recall rate is 61.71%.

Key words: Chinese unknown words; Sense Prediction; Language knowledge base; Word-formation; HowNet

1 引言

自然语言的理解应建立在对自然语言中每一个词语语义理解的基础之上。然而，由于存在大量的未登录词，这些词语的语义对于机器而言是未知的，因此对自然语言处理提出了很大的挑战，是自然语言处理领域的一个难题。未登录词的语义预测研究，旨在对未登录词提供预测的语义知识，这对信息检索、机器翻译等自然语言处理的关键研究具有重要的应用价值。汉语未登录词的语义预测难度较大，因此相关的研究较少，且主要集中在未登录词的语义类别预测这个方向[1-8]。

语义类别所能表达的语义知识仍是有限的，近年来有个别学者进行了更细化的汉语未登录词语义预测。张瑞霞[9]以《知网》为语义知识资源、概念图为知识表示方法，进行未登录词的语义分析。然而，概念图这种表示形式较为复杂，不便于计算。吉志薇[10]在定量统计

* **收稿日期：**　　**定稿日期：**

的基础上建立二字词的语义描写体系，预测未登录词语义层面的语素构词知识。然而，该实验仅选取某一特定的语素义类组合为考察对象，缺少整体的结果。田元贺[11]具体探讨了语义构词知识对未登录词理解的应用价值，认为“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”。他所指的语义构词知识是包括语义层面和语法层面的多方面语素构词知识。该研究采用贝叶斯网络的方法，构建汉语未登录词的语义构词分析模型，能较好地预测未登录词的“多层面”词义知识，然而“全层面”语义构词知识的预测正确率较低。因为未登录词的语义预测是研究难点，所以“多层面”语义知识不失为一种可行的方案，然而“全层面”知识具有更大的研究和应用价值，值得进一步深入研究。

2 知识库的构建

汉语的词汇系统总是在不断地发展和变化，因而未登录词的数量是无限的。但是语素作为汉语中词语的构词成分，在数量上是有限的，且其表义功能是相对稳定的。因此在自然语言处理中，可以把语素作为基础资源，获取语素构词的知识，用来识别和理解未登录词。我们构建了《汉语语素构词知识库》（以下简称《知识库》），以《现代汉语词典》（以下简称《现汉》）和《知网》为主要基础资源，结合语文词典和知识系统，以义项为单位描述每个词的语法语义多层面语素构词知识，从而得到人机两用的语言知识资源。《知识库》第一阶段以《现汉》第 5 版、《知网》2009 版和《同义词词林》扩展版所共有的双字词[①]为收录对象，共 39102 个记录。我们制定了详细完整的标注体系，并根据标注规范进行具体的标注。目前已完成全部名词记录和其他词性各 20%抽样记录的标注，共 23500 个记录。下面以《知识库》中“要案”的标注为例来说明词语的描述。其中，方括号“【 】”中的是属性字段名，字段名的右边是该属性字段所对应的属性值。

【编号】33519
【词语】要案
【拼音】yao4 an4
【词性】N
【词义类】fact|事情
【现汉词义】重要的案件。
【知网词义】{fact|事情:domain={police|警},modifier={important|重要}}
【首素性类】Ag
【首素义类】PropertyValue|特性值
【现汉首素义】①重要：主～｜紧～｜险～｜～事｜～道。
【知网首素义】{important|重要}
【尾素性类】Ng
【尾素义类】fact|事情
【现汉尾素义】①案件：犯～｜破～｜五卅惨～。
【知网尾素义】{fact|事情:domain={police|警}}
【构词方式】偏正式
【词义和语素义的关系】AB = A + B
【备注】（空）
【说明】（空）

如上所示，每个记录包含 19 个属性字段，能详细描述该词语义项的语素构词知识。这些知识既有语法层面的，如“词性”、“首素性类”、“尾素性类”和“构词方式”等字段；又有语义层面的，如“词义类”、“首素义类”、“尾素义类”、“词义和语素义的关系”等字段。

[①] 从知识库的规模考虑，先以双字词为收录对象。双字词是汉语中最典型的词，对它的研究具有代表性。

这些知识既有基于《现汉》的便于人阅读的，如“现汉词义”、“现汉首素义”、“现汉尾素义”等字段；又有基于《知网》的面向计算机计算的，如“知网词义”、“知网首素义”、“知网尾素义”等字段。每个词语的首尾两个语素既有语义分类知识，又有基于《知网》的完整概念描述，从而更有利于意义的计算。相较于前人的研究[12-14]，《知识库》在语义层面上的颗粒度更小，具有重要的研究意义和应用价值。

3 基本思路

本文是基于《知识库》中已标注的语素构词知识来预测未登录词的语素构词知识。预测的内容包括未登录词的首素性类、首素义类、首素义、尾素性类、尾素义类、尾素义、构词方式。这些内容既有语法层面的，也有语义层面的，而多层面的构词知识可以满足不同的应用需求。在《知识库》中，我们对每个词语记录中的语素标注了基于《现汉》和《知网》的两种语素义。因为《知网》中同一个概念表达式可用来描述同义或同类的不同概念，具有较好的概括性，所以预测所使用的素义指的是《知识库》中基于《知网》的语素义。因为预测的结果是面向计算机的，而《知网》的概念表达式便于意义的计算，所以预测内容中的素义同样也是基于《知网》的语素义。预测的方法是首先将未登录词的两个语素按首尾位置进行语素项[①]组合，然后利用《知识库》中语义层面知识来预测这两个语素所对应的语素项，从而得到未登录词语素构词的多层面知识。预测所使用的语义层面知识主要包括语素义组合和语素义类组合这两方面的知识。

下面使用两个例子来说明预测方法的基本思路。例 1：“背囊”这一名词是由“背”和“囊”这两个语素构成。“背”在《现汉》中对应两个不同的语素。因为计算机只能区分素形（汉字），因此需要查找同一素形下的所有语素项。“背”在《知识库》名词记录中作为首语素参与构词的共有四个语素项，具体如表 1 所示。

表 1　“背”在名词记录中作为首语素的语素项

编号	首素性类	首素义类	首素义
1	Vg	{AlterLocation|变空间位置}	{CarryOnBack|背起}
2	Ng	{part|部件}{AnimalHuman|动物}	{part|部件:PartPosition={body|身},modifier={hind|后},whole={AnimalHuman|动物}}
3	Ng	{part|部件}{inanimate|无生物}	{part|部件:PartPosition={body|身},modifier={hind|后},whole={inanimate|无生物}}
4	Ag	{BeBad|衰变}	{unfortunate|不幸}

“囊”在《知识库》的名词记录中作为尾语素参与构词的共有两个语素项，具体如表 2 所示。

表 2　“囊”在名词记录中作为尾语素的语素项

编号	尾素性类	尾素义类	尾素义
1	Ng	{implement|器具}	{tool|用具:{put|放置:LocationFin={~}}}
2	Ng	{shape|物形}	{shape|物形}

“背”和“囊”分别有四个语素项和两个语素项，因此一共有八种语素项组合的可能，相应素义组合也有八种可能。如“背”的第一个语素项与“囊”的第二个语素项进行组合，其素义组合为“{CarryOnBack|背起}+{shape|物形}”。我们将八种素义组合在《知识库》的

[①] 语素项指的是一个语素的一个义项（本义、引申义或比喻义）。

名词记录中进行匹配，结果只有一种素义组合“{CarryOnBack|背起}+{tool|用具:{put|放置:LocationFin={~}}}”能找到匹配记录。因此，我们预测“背囊”的首素义和尾素义分别为“{CarryOnBack|背起}”和“{tool|用具:{put|放置:LocationFin={~}}}”。根据预测的素义组合，我们可以确定“背囊”的语素项组合为“背”的第一个语素项和“囊”的第一个语素项，从而可以根据语素项预测“背囊”的首素性类和尾素性类分别为“Vg”和“Ng”，首素义类和尾素义类分别为“{AlterLocation|变空间位置}”和“{implement|器具}”。构词方式则根据预测的素义组合在《知识库》中所匹配记录的构词方式判断。因为匹配记录都是偏正式，所以我们预测“背囊”的构词方式也是偏正式。我们预测所依据的素义组合在《知识库》中所匹配的四个词语记录分别为“背包①”、“背包②”、“背篓”和“担架”。其中，“背包①”和“背包②”是“背囊”的相似词语，而另外两个词语从语义上来看，也有一定的相似性。由此可见，利用《知识库》已登录词的素义组合知识来预测未登录词语素构词是合理的。

例 2：“草根”这一名词是由“草”和“根”这两个语素构成。其中“草”在《知识库》的名词记录中作为首语素参与构词的共有三个语素项，具体如表 3 所示。

表 3　“草”在名词记录中作为首语素的语素项

编号	首素性类	首素义类	首素义
1	Ng	{plant|植物}	{FlowerGrass|花草}
2	Ng	{plant|植物}	{FlowerGrass|花草:MaterialOf={material|材料}}
3	Ag	{PropertyValue|特性值}	{careless|粗心}

“根”在《知识库》的名词记录中作为尾语素参与构词的共有四个语素项，具体如表 4 所示。

表 4　“根”在名词记录中作为尾语素的语素项

编号	尾素性类	尾素义类	尾素义
1	Ng	{part|部件}{plant|植物}	{part|部件:PartPosition={base|根},whole={plant|植物}}
2	Ng	{part|部件}{inanimate|无生物}	{part|部件:PartPosition={base|根},whole={inanimate|无生物}}
3	Ng	{part|部件}{event|事件}	{part|部件:PartPosition={base|根},whole={event|事件}}
4	Ng	{cause|原因}	{cause|原因}

“草”和“根”分别有三个语素项和四个语素项，因此一共有 12 种语素项组合的可能，相应素义组合也有 12 种可能。然而，在《知识库》的名词记录中没有匹配到任何一种素义组合，因此我们接着看素类组合的情况。因为“草”的第一个语素项和第二个语素项的语素义类是相同的，所以一共有八种素类组合的可能。这八种素类组合在《知识库》中共出现了四种，按频次从大到小分别为“{plant|植物}+ {part|部件}{plant|植物}”、“{PropertyValue|特性值}+ {part|部件}{plant|植物}”、“{PropertyValue|特性值}+ {part|部件}{event|事件}”和“{PropertyValue|特性值}+ {cause|原因}”。我们把频次最大的素类组合“{plant|植物}+ {part|部件}{plant|植物}”作为“草根”的预测素类组合。根据预测的素类组合，我们可以确定“根”对应的是第一个语素项。但是，“草”因为第一个和第二个语素项的语素义类都是“{plant|植物}”，因而无法确定对应的语素项。在这种情况下，我们可以通过语素项的构词频率将频率更大的语素项作为预测语素项。我们根据频率确定“草”为第一个语素项，从而可以预测

“草根”的首素性类和尾素性类都为“Ng”，首素义和尾素义分别为“{FlowerGrass|花草}”和“{part|部件:PartPosition={base|根},whole={plant|植物}}”。构词方式根据预测素类组合所匹配的词语记录判断。这些词语的构词方式有“偏正式”和“并列式”。我们根据“偏正式”在这些词语中出现频次更大，预测“草根”的构词方式为“偏正式”。我们预测所依据的素类组合在《知识库》中对应的记录有“树干”、“豆荚”、“花茎”等 34 个词语，是一种能产性较强的素类组合。这些词语的共同特征是首语素表示某种植物，而尾语素表示植物的某个部件。由此可见，利用《知识库》已登录词的素类组合知识来预测未登录词语素构词也是可行的。

4 实验数据和算法

4.1 实验数据

本实验的测试数据来自《现汉》第 6 版中新增的双字词。因为《知识库》是以《现汉》第 5 版中的双字词为收录对象，所以《现汉》第 6 版新增的双字词就是《知识库》的未登录词。我们使用新增双字词中同时被《知网》和《词林》所收录的部分。因此，我们在分析实验结果时，可以参考和《知识库》标注时相一致的资源，并根据《知识库》的标注标准来有效地评价实验结果的正确性。如第 3 节所述，我们的预测方法主要利用了语素义组合和语素义类组合的知识，而这些语义层面知识的组合分布是比较分散的。因此，根据《知识库》的标注现状，本实验将只涉及名词，但是预测方法本身仍适用于其他词性的词语。我们将《现汉》第 6 版新增双字词中与《知网》和《词林》共有的全部 209 个名词确定为测试数据，并得到测试词表。

根据确定的测试数据，在本实验中我们使用的其他数据有：

（1）名词标注表

从《知识库》标注总表中得到的全部名词记录的标注表。

（2）首语素项表

从名词标注表中统计得到的首语素项表。每一个在名词记录中作为首语素的语素项形成一个记录。每一个记录有“首语素”、“首素性类”、“首素义类”、“首素义”和“首语素项频率”五个字段。其中，“首语素项频率”指的是该首语素项在名词标注表中参与构词的频率。

（3）尾语素项表

从名词标注表中统计得到的尾语素项表。每一个在名词记录中作为尾语素的语素项形成一个记录。每一个记录有“尾语素”、“尾素性类”、“尾素义类”、“尾素义”和“尾语素项频率”五个字段。其中，“尾语素项频率”指的是该尾语素项在名词标注表中参与构词的频率。

（4）首素义频率表

从名词标注表中统计得到的首素义频率表。根据概念表达式，每一个首素义形成一个记录。每一个记录有“首素义”和“首素义频率”两个字段。其中，“首素义频率”指的是该首素义在名词标注表中参与构词的频率。

（5）尾素义频率表

从名词标注表中统计得到的尾素义频率表。根据概念表达式，每一个尾素义形成一个记录。每一个记录有“尾素义”和“尾素义频率”两个字段。其中，“尾素义频率”指的是该尾素义在名词标注表中参与构词的频率。

（6）知网单字语素表

根据《知识库》标注体系对《知网》的单字记录进行调整，且删除姓氏记录得到的知网单字语素表。每一个记录为一个语素项，包括“语素”、“语素性类”、“语素义类”、“语素义”四个字段。

（7）知网语素义频率表

从知网单字语素表中统计得到的语素义频率表。根据概念表达式，每一个语素义形成一个记录。每一个记录有“语素义”和“语素义频率”两个字段。其中，“语素义频率”指的是该语素义在知网单字语素表中形成语素项的频率。

4.2 算法描述

本实验充分利用《知识库》的已标注数据，根据优先使用《知识库》语素项和优先使用素义组合匹配的两个原则，采用分阶段的算法自动预测未登录词的语素构词知识。优先使用《知识库》语素项是指首先使用首语素项表和尾语素项表中的《知识库》语素项进行语素项组合；如果得不到预测结果，再使用知网单字语素表中的知网语素项。优先使用素义组合匹配是指首先使用颗粒度更小的素义组合匹配；如果得不到预测结果，再使用素类组合匹配。结合第 3 节的基本思路，具体算法如下所述，基本流程如图 1 所示。

第一阶段：使用《知识库》语素项以及素义组合匹配

Input：测试词表、名词标注表、首语素项表、尾语素项表

Step1：将测试词表根据首语素和尾语素分别与首语素项表和尾语素项表联接，得到测试词的语素项组合表。

Step2：将语素项组合表根据首素义和尾素义与名词标注表联接，按词统计每种素义组合在名词标注表中的频次，取频次最大的素义组合为该词的预测素义组合，再取该预测素义组合在名词标注表中频次最大的构词方式为该词的预测构词方式。

Step3：根据预测素义组合确定预测语素项组合。如果一个词语有多个记录，取首语素项频率和尾语素项频率乘积最大的语素项组合为预测结果。

Output：第一阶段的预测结果及第一阶段未处理词表

第二阶段：使用《知识库》语素项以及素类组合匹配

Input：第一阶段未处理词表、名词标注表、首语素项表、尾语素项表、首素义频率表、尾素义频率表、语素项组合表

Step1：将第一阶段未处理词表根据首语素和尾语素分别与首语素项表和尾语素项表联接，得到素类组合表。

Step2：将素类组合表根据首素类和尾素类与名词标注表联接，按词统计每种素类组合在名词标注表中的频次，取频次最大的素类组合为该词的预测素类组合，再取该预测素类组合在名词标注表中频次最大的构词方式为该词的预测构词方式，得到素类组合预测表。

Step3：将素类组合预测表联接第一阶段 Step1 的语素项组合表，得到素类组合对应的语素项组合。如果一个词语有多个记录，首先按首语素项表的首语素项频率和尾语素项表的尾语素项频率乘积最大者为预测语素项组合。如果仍对应多个记录，再取首素义频率表的首素义频率和尾素义频率表的尾素义频率乘积最大者为预测结果。

Output：第二阶段的预测结果及第二阶段未处理词表

第三阶段：使用知网语素项以及素义组合匹配

Input：第二阶段未处理词表、名词标注表、知网单字语素表、知网语素义频率表

Step1：将第二阶段未处理词表根据首语素和尾语素分别与知网单字语素表联接，得到知网语素项组合表。

Step2：将知网语素项组合表根据首素义和尾素义与名词标注表联接，按词统计每种素义组合在名词标注表中的频次，取频次最大的素义组合为该词的预测素义组合，再取该预测素义组合在名词标注表中频次最大的构词方式为该词的预测构词方式。

Step3：根据预测素义组合确定预测语素项组合。如果一个词语有多个记录，取首素义和尾素义在知网语素义频率表中的语素义频率乘积最大的语素项组合为预测结果。

Output：第三阶段的预测结果及第三阶段未处理词表

第四阶段：使用知网语素项以及素类组合匹配

Input：第三阶段未处理词表、名词标注表、知网单字语素表、首素义频率表、尾素义频率表、知网语素义频率表、知网语素项组合表

Step1：将第三阶段未处理词表根据首语素和尾语素分别与知网单字语素表联接，得到素类组合表。

Step2：将素类组合表根据首素类和尾素类与名词标注表联接，按词统计每种素类组合在名词标注表中的频次，取频次最大的素类组合为该词的预测素类组合，再取该预测素类组合在名词标注表中频次最大的构词方式为该词的预测构词方式，得到素类组合预测表。

Step3：将素类组合预测表联接第三阶段 Step1 的知网语素项组合表，得到素类组合对应的语素项组合。如果一个词语有多个记录，首先按首素义频率表的首素义频率和尾素义频率表的尾素义频率乘积最大者为预测语素项组合。如果仍对应多个记录，再取首素义和尾素义在知网语素义频率表中的语素义频率乘积最大者为预测结果。

Output：第四阶段的预测结果及第四阶段未处理词表

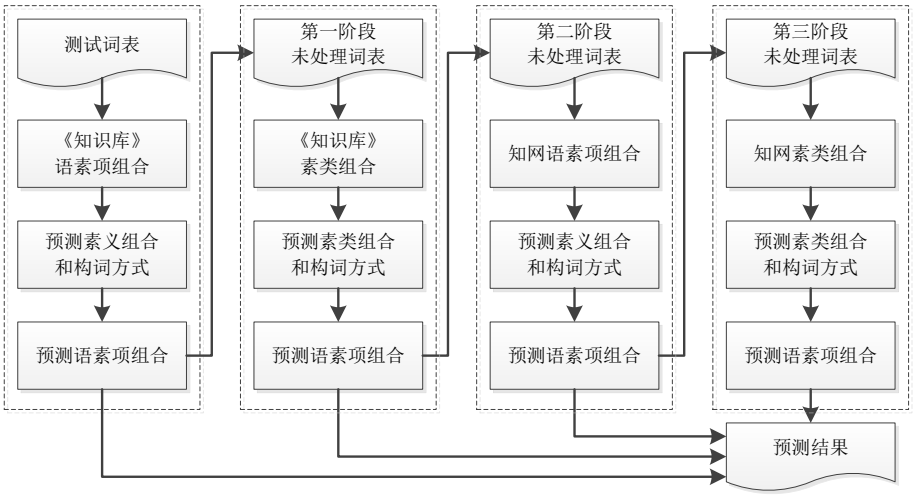

图 1　实验的基本流程

5 实验结果和分析

5.1 实验结果

我们使用《现汉》第 6 版和《知网》等参考资源，根据《知识库》的标注体系，对测试数据的每个词语人工标注了预测内容所包括的首素性类、首素义类、首素义、尾素性类、尾素义类、尾素义、构词方式这七项内容，并以此作为实验结果的评价标准。我们以实验结果的七项预测内容全部正确为预测正确。此外，虽然同一词形下可能有多个词语，而同一词语又可能有多个义项，但是因为测试数据没有上下文语境，所以我们这里规定预测结果只要能符合该词形下词语的任何一个义项即为预测正确。

首先，我们按实验过程的四个阶段来分析实验结果，如表 5 所示。

表 5　实验结果（分阶段）

阶段	预测词数	百分比（%）	正确词数	正确率（%）	正确词
第一阶段	52	24.88	36	69.23	暗河、背囊、残渣、茶道、车轴…
第二阶段	130	62.20	80	61.34	编委、才貌、参事、餐券、草根…
第三阶段	6	2.87	5	83.33	蚕蛹、跗面、胛骨、栲胶、生姜…
第四阶段	19	9.09	8	42.11	飞蝗、飞瀑、沙浴、食宿、陶俑…

这四个阶段共预测出 207 个词的语素构词知识，占测试词的 99.04%。此外还有 2 个词没有返回语素构词知识，分别是“祁红”和“祁剧”。这两个词是因为在《知识库》和知网单字语素表中都没有“祁”的语素项，从而无法预测相应语素构词知识。四个阶段预测正确的词共 129 个，预测正确率为 62.32%，召回率为 61.72%。

这四个阶段，从处理词占测试词的比例来看，从高到低分别是：第二阶段、第一阶段、第四阶段、第三阶段。从预测正确率来看，从高到低分别是：第三阶段、第一阶段、第二阶段、第四阶段。从所用语素项的来源来看，使用《知识库》语素项的第一阶段和第二阶段，所占比例为 87.08%，预测正确率为 63.74%；使用知网语素项的第三阶段和第四阶段，所占比例为 11.96%，正确率为 52.00%。由此可见，使用《知识库》语素项就能预测绝大多数的测试词，且预测正确率要高于使用知网语素项的方式。从匹配所使用的组合类型来看，使用素义组合匹配的第一阶段和第三阶段，所占比例为 27.75%，预测正确率为 70.69%；使用素类组合匹配的第二阶段和第四阶段，所占比例为 71.29%，正确率为 59.06%。由此可见，大部分的测试词是使用素类组合匹配的方式，但使用素义组合匹配的预测正确率要高于使用素类组合匹配。因此，实验结果体现了优先使用《知识库》语素项和优先使用素义组合匹配的分阶段算法的合理性。

接下来，我们按预测内容来分析实验结果。在表 6 中列出了 7 种单项语素构词知识的预测正确率，在表 7 中列出了 12 种多项语素构词知识的预测正确率。这些不同的语素构词知识可以满足不同的应用需求，如我们可以利用表 7 中项目 6 的“首素义+尾素义+构词方式”进行基于语素的未登录词语义相似度计算。

表 6　单项语素构词知识的预测正确率

预测内容	首素性类	首素义类	首素义	尾素性类	尾素义类	尾素义	构词方式
正确率（%）	90.34	80.19	79.71	96.62	77.78	75.85	91.79

表 7　多项语素构词知识的预测正确率

项目	预测内容	正确率（%）
1	首素性类+尾素性类	87.92
2	首素性类+尾素性类+构词方式	82.13
3	首素义类+尾素义类	68.60
4	首素义类+尾素义类+构词方式	64.73
5	首素义+尾素义	65.70
6	首素义+尾素义+构词方式	62.32
7	首素性类+首素义类+首素义	79.23
8	首素性类+首素义类+首素义+构词方式	74.88
9	尾素性类+尾素义类+尾素义	75.36
10	尾素性类+尾素义类+尾素义+构词方式	70.05
11	首素性类+首素义类+首素义+尾素性类+尾素义类+尾素义	65.70
12	首素性类+首素义类+首素义+尾素性类+尾素义类+尾素义+构词方式	62.32

从表 6 中可以看出，在单项语素构词知识中，预测正确率最高的是尾素性类，达 96.62%；预测正确率最低的是尾素义，为 75.85%。其中，语素性类的预测正确率高于语素义类，而语素义类的预测正确率又高于语素义。由此可见，虽然预测方法是利用语义层面的知识，但是因为语法层面知识的归纳性强，所以预测正确率更高，而语素义是预测内容中最精细的，

所以预测难度也是最高的。因为预测方法是通过匹配的素义组合或素类组合来进一步确定语素项组合，所以如果语素义预测错误，意味着语素项的预测也是错误的，然而语素义类和语素性类的正确率要高于语素义，这是因为虽然语素项预测错误，但是其中部分预测错误的语素项的语素义类或语素性类和正确语素项的是相同的。

从表 7 中可以看出，在语素组合的预测上，素性组合（项目 1）的预测正确率高于素类组合（项目 3），而素类组合的预测正确率又高于素义组合（项目 5）。在语素项的预测上，首语素项（项目 7）的预测正确率高于尾语素项（项目 9）。我们还观察到，素义组合的预测正确率与语素项组合（项目 11）一样，这是因为一旦语素义确定了，相应的语素项也就确定了。当预测内容涉及到两个语素时，如果仅包括语法层面的知识（项目 1、2），有 80%以上的预测正确率；如果包括语义层面的知识（项目 3、4、5、6、11、12），预测正确率在 60%到 70%之间。当预测内容仅涉及一个语素时，表 7 中语素项的预测（项目 7、8、9、10）包括语法和语义两个层面的知识，正确率在 70%到 80%之间。

5.2 错误分析

预测错误的原因主要可分为三种：第一种是虽然正确的素义组合或素类组合在名词标注表中有匹配记录，但该正确组合并不是频次或频率最大的组合，因此预测成错误组合。例如：“冰场”在第二阶段得到预测结果，使用的是素类组合匹配，其正确的素类组合是“{ice|冰}+ {location|位置}”，但是该组合不是名词标注表中频次最大的组合，预测为频次最大的组合“{AlterAttribute|变属性}+ {location|位置}”，最终“冰场”一词的首素义类和首素义都预测错误。

第二种是语素项的所有组合可能中包含正确的素义组合或素类组合，但该正确组合在名词标注表中没有匹配记录，因此预测成错误的组合。例如：“课间”在第一阶段得到预测结果，在最初语素项组合时存在正确的素义组合“{fact|事情:CoEvent={study|学习}{teach|教},domain={education|教育}}+{location|位置:modifier={InBetween|之间},restrictive={?}}”，但是该组合在名词标注表中没有匹配记录，预测为频次最大的“{fact|事情:CoEvent={study|学习}{teach|教},domain={education|教育}}+{room|房间}”，最终“课间”一词的尾素义和尾素义类都预测错误。

第三种是虽然返回了预测结果，但因为在语素项表中并没有收录正确的语素项，所以预测的是错误的语素项，得到的是错误的预测结果。例如：“简牍”在第二阶段得到预测结果，在素类组合时只有一种可能“{PropertyValue|特性值}+{readings|读物}”，这表示首语素项表尚未标注出“简”的语素义类为“{readings|读物}”的相应语素项，最终“简牍”一词的首素义类、首素义、构词方式都预测错误。由此可见，如果《知识库》进一步扩大收词规模，应该可以在一定程度上减少错误，从而提高预测的正确率。

5.3 实验比较

我们将实验结果与前人的研究进行比较。目前只有吉志薇[10]和田元贺[11]的研究是涉及到语义层面的构词预测。这两个研究与本研究的具体测试数据和预测内容都存在差异，因此只能就相似的预测内容进行大致的比较。其中，吉志薇[10]仅选取了该研究中频率最高的素类组合的 71 个词作为考察对象，并将素类组合作为语素意义和该素类组合中频率最高的词化意义以释义模式作为预测内容。该实验结果的正确率为 43.67%，低于本研究在表 7 中所示的与该研究预测内容比较一致的项目 3 和项目 4。在预测方法上，该研究仅使用素类组合的频率，而本研究是先使用素义组合然后再使用素类组合，因为素义组合匹配的正确率比素类组合高，所以整个实验结果的正确率更高。

从测试数据来看，田元贺[11]的研究是将该研究中知识库的标注数据采用交叉验证的方法，比本研究的规模大。本研究未采用交叉验证是因为预测方法是从素义组合入手，而素义组合的分布较分散，因此需确保所基于知识库的覆盖面。从预测内容来看，该研究比本研究

多了词性预测，而本研究是在已知词性的条件下进行，然而目前词性预测的技术已经比较成熟[15]，所以基于词性已知的方法是可行的。该研究预测内容中的语素义是基于《现汉》以“语素义编码”的形式表示的，并建立了一个树状结构的“语素概念体系”。但是，“语素义编码”代表的仅是该语素义在《现汉》中对应的条目和义项，是不能直接用于计算的，而所绑定的“语素概念”体现的只是上下位的语义关系。本研究中的语素义是《知网》形式的概念表达式，是面向计算机可直接计算的属性描述，体现的是多层次的网状关系。从预测方法来看，该研究采用了贝叶斯网络的方法，以推理的方式获取知识，推理过程是先从字预测语素性类，然后预测语素义，最后预测构词方式。从实验结果来看，该方法随着构词知识种类的增多和叠加，正确率也随之下降。比如，“首素性类+尾素性类”的预测正确率为 75.45%，“首素义+尾素义”正确率为 43.24%，而“词性+构词方式+首素义+尾素义”的正确率为 30.26%。由此可见，该研究所使用的语义构词分析模型在涉及语义层面尤其是“全层面”知识的预测上，正确率是比较低的。本研究采用分阶段的算法，通过预测语义层面知识再确定语素项，进而获得多层面的知识，如表 7 中项目 12 的预测正确率为 62.32%，可见本研究在“全层面”语素构词知识的预测上也取得了比较好的结果。

5.4 进一步讨论

从预测内容来看，除了本实验的七项内容以外，我们可以使用和预测构词方式一样的方法来预测词义和语素义的关系。我们还可以根据预测构词方式判断未登录词的核心语素，并将该语素的预测语素义类作为未登录词的预测语义类别。我们将实验结果的 207 个词根据如上所述的方法得出预测语义类别，再将这些词在《知网》中的概念根据《知识库》的语义分类体系得到的语义类作为评价标准，最后语义类别预测正确率为 67.63%，这一结果和前人的研究[1-8]基本相当。

从测试数据来看，本实验根据《知识库》的标注情况，为了得到有效的评价结果，所以只包含了名词，但是我们的预测方法本身同样适用于其他词性的词语。我们将《现汉》第 6 版新增双字词中与《知网》和《词林》共有的其他词性词语在相同的预测方法下进行了测试，结果显示算法是同样可行的，只是因为这些词性已登录词的数量不够多，所以四个阶段处理词的比例与名词相比有较明显差异。比如，我们利用 20%抽样的动词标注数据对共有的 154 个动词进行实验，结果四个阶段所处理词的比例分别为：3.90%、29.87%、3.90%、56.49%。由此可见，已标注数据的数量对实验结果的影响较大。

此外，本实验根据《知识库》的标注情况，为了得到有效的频次或频率信息，只包含了名词标注数据，因而加上了已知词性的条件。但是，我们通过名词标注数据和其他词性抽样数据进行了统计分析，发现不同词性的素义组合和素类组合分布情况存在明显差异。因此，在《知识库》第一阶段标注完成后，我们可以尝试将素义组合或素类组合在整个《知识库》中进行匹配，并根据匹配词的词性来预测词性。因为《知识库》是以汉语中最典型的双字词为收录对象，所以实验测试数据使用的也是双字词，但是本文的预测方法是从语义层面入手，具有良好的扩展性，在具有已登录词标注数据的基础上，可以进一步运用到三字及以上的未登录词上。

6 结语

本文基于《知识库》已登录词的语素构词知识，采用了分阶段的算法自动预测未登录词的语素构词知识。实验结果显示，在首素性类、首素义类、首素义、尾素性类、尾素义类、尾素义、构词方式这七项预测内容全部正确的标准下，预测正确率为 62.32%，召回率为 61.72%。与前人研究相比，本文方法在“全层面”语素构词知识上也取得了较好的结果，且预测的语素义是《知网》可直接计算的概念表达式，在自然语言处理领域中具有重要的应用价值。虽然本实验的测试数据较小，但实验结果体现了利用语义层面知识，且优先利用颗粒度小的语素义组合知识，首先预测语义层面知识，再确定相应语素项，进而获得多层面的

语素构词知识，这种研究思路的合理性和有效性。下一步我们将继续扩大《知识库》标注数据的数量，利用更完善的《知识库》，优化现有的未登录词语素构词预测算法，并增加预测内容，如词性等。我们将扩大测试数据规模，进行更多的相关实验，并尝试将其运用于实际的应用系统中。

参考文献

[1] Lua K T. Prediction of meaning of bi-syllabic Chinese compound words using back propagation neural network[J]. Computational Processing of Oriental Languages, 1997, 11(2): 133-144.

[2] 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.

[3] 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.

[4] 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.

[5] Lu X. Hybrid models for Chinese unknown word resolution[D]. Ohio: The Ohio State University, 2006.

[6] Lu X. Hybrid models for semantic classification of Chinese unknown words[C]//Proceeding of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2007: 188-195.

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

[8] 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.

[9] 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.

[10] 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.

[11] 田元贺,刘扬.汉语未登录词的词义知识表示及语义预测[J].中文信息学报,2016(06):26-34.

[12] 苑春法,黄昌宁.汉语语素数据库的建造与应用[J].Communication of COLIPS,1997,7(1):1-4.

[13] 俞士汶,朱学锋,李峰.现代汉语语素库的开发及应用[J].世界汉语教学,1999(02):39-46.

[14] 亢世勇.面向信息处理的现代汉语语法研究[M].上海:上海辞书出版社,2004:26-61.

[15] 刘慧敏.中文词性标注及未登录词词性预测研究[D].南京:南京师范大学,2015.

作者联系方式：瞿健菊，地址：湖南省湘西自治州吉首市北一环 188 号吉首大学师范学院，邮编：416000，电话：13739048657，电子邮箱：simplequ@163.com

瞿健菊（1982—），通讯作者，讲师，主要研究领域为自然语言处理。
Email：simplequ@163.com

冯敏萱（1978—），副教授，主要研究领域为计算语言学。
Email：fengminxuan@njnu.edu.cn