现代汉语复合词内部结构词典的构建[1]

邱立坤[1,2] 张晓巧[1] 毛宁[1]

1 北京城市学院 人工智能研究所，100083；2 北京大学 中文系，北京，100871

E-mail: qiulk@pku.edu.cn

摘要：本文介绍了构建现代汉语复合词内部结构词典的方法和流程。在构建该词典的过程中，我们使用人工分析与自动分析相结合的方法。在自动分析方面，我们共使用了三种方法，包括双向平行类推、成对替换类推两种自动类推方法，以及基于形式与意义同构思想的推导方法。Hownet 54000 双字词中，这三种方法可以给出全部或部分内部结构的词语有 4 万余条。在人工标注和自动标注的基础上，我们通过两种方法来辅助校对。第一种方法是采用启发式规则从人工标注结果中自动发现标注不一致的现象，并从中挑出可能错误的标注结果，收到了良好的效果。第二种方法是将人工标注结果和自动标注结果进行比较，将不一致的提交给校对人员。

关键词：整体语义类，成分语义类，双向平行类推，成对平行互推

Building a Contemporary Chinese Dictionary on Internal Structure of Compounds

Qiu Likun[1,2], Zhang Xiaoqiao[1], Mao Ling[1]

(1 Institute of Artificial Intelligence, Beijing City University, Beijing 100083;
2. Department of Chinese Language and Literature, Peking University, Beijing 100871)

E-mail:qiulk@pku.edu.cn

Abstract: This paper presents a novel approach of building a contemporary Chinese dictionary on internal structure of compounds. Manual tagging and three automatic tagging methods, including bi-direction parallel analogy, paired parallel analogy and inference based on the consistency between form and meaning, are used. More than 40,000 words of all the 54000 bi-syllabic words in Hownet are fully or half semantically tagged by those automatic methods. After manual and automatic tagging, two methods are used for checking. First, several heuristic rules are used on manual tagging results to mine abnormal tagging. Second, the manual and automatic tagging results are comparing together and those inconsistent tagging results are considered as abnormal tagging. All those abnormal tagging results are reserved for further checking.

Keywords: semantic category of compound, semantic categories of constituents, bi-direction parallel analogy, paired parallel analogy

1 引言

现代汉语双音词多数是合成复合词和附加复合词，仅有少部分是单纯词。复合词是双音词的主体，也最为复杂。单纯词没有内部结构，因此双音词的内部结构问题主要就是双音复合词内部结构的问题。

复合词的结构研究在构词法研究中争议较大，存在着两种相对立的观点，一种认为复合词的结构与短语的句法结构基本一致，以陆志韦（1964）、朱德熙（1982、1983）、尹斌庸（1984）、邵

[1]本研究受模式识别国家重点实验室开放课题基金和北京市教委科技发展计划面上项目 KM200800004003 资助。

敬敏（2001）杨同用（2002）、王洪君（2005）等为代表；另一种观点认为词内完全没有句法结构，但有语义结构，以刘叔新（1990a，1990b）、黎良军（1995）为代表。就近年的研究而言，前一种观点占压倒性的优势。

我们认为两种观点本质上并不矛盾，其区别在于着眼点不同。前一种观点侧重于考察词语内部的语法结构，后一种观点侧重于考察词语内部的语义结构。两者是互为补充、相辅相承的关系。正如短语和句子都同时具有语法结构和语义结构一样，复合词也具有其自身的语法结构和语义结构。俞士汶（1999）指出“要解决好计算机系统内未定义词的处理，重要的途径就是注意对合成词构词规律和词间关系的研究。”但是正如傅爱平（2003）所说，“就目前所见到的文献资料而言，在词素构词方式的调查统计中得到的统计规律很少在识别未登录词语的工程实践中得到应用。尤其是那些基于语法属性的构词规律。”基于上述原因，我们希望能够在前人工作的基础上，构建一部面向新词词性猜测和语义类猜测的现代汉语复合词内部结构词典。

本文剩余部分组织如下：第二节介绍相关的工作；第三节以词语和义项的来源为依据对 Hownet 中的词条进行分类；第四节介绍自动标注的方法；第五节介绍计算机辅助人工标注的方案；第六节介绍基于该词典的一些统计结果；最后一节是结论和展望。

2 相关工作

苑春法(1998)构建了汉语语素数据库并基于该库统计了汉语的构词规律，发现名词的构词方式以体素联合和定中偏正为主，其中定中偏正占 80.6%，体素联合占 9.3%；动词以述宾、谓素联合和状中偏正三种构词方式为主，它们各占 39.7%、27.0%、23.3%；形容词以谓素联合为主，占形容词二字词总量的 62.5%。

亢世勇（2004）为 52366 个双音合成词中的每个字标注同义词词林的义类标记和简单释义，之后发现仅有 8.02%的词整体意义与成分意义没有任何关系.

杨梅（2006）探讨汉语合成词的构词模式对合成词语法属性的影响及相关问题，认为合成词的构词方式与词性没有必然的关系，仅是一种倾向性。该文统计表明，90％以上的合成词是向心词，只有不到 10％的合成词是离心词。

在判断词素语法类的方法方面，Packard（2000）提出使用中心原则（Headness Principle），董秀芳（2004）提出参考一个语素在历史上的使用情况来判断这个语素在现代汉语中的语法类别，王洪君（2005）则提出“只要有相当数量的单字不兼类，且是否兼类能够用义场来控制，以单字类控两字组的类、以不兼类字控制兼类字就有可行性。……今后的任务应是进一步扩大考察范围，以弄清汉语到底哪些义场的字不兼类，哪些义场的字兼类，兼类的规律是什么。”

陈保亚（1999）提出用“平行周遍原则”来区分词和短语，之后陈保亚（2005）和陈保亚（2006）又对此理论作了进一步的发展。陈认为所谓的“平行”不仅仅是分布的平行，还可以是语义甚至语音层面的平行，只要可以找到相应的规则。更进一步，陈将“平行不周遍”和“既平行又周遍”两类现象与理解规则和生成规则的区分对应起来，认为前者与理解规则对应，后者与生成规则对应。这一理论极富启发意义，从生成的角度来说，可以认为一部分新词是根据已有的词依据平行周遍原则类推而来；从理解的角度来说，可以依据平行周遍原则推出一部分词语的内部结构。

前人的许多工作分别从某一角度涉及到复合词内部结构的分析，有的是从结构关系类型角度，有的是从成分语义类的角度。与前人的工作相比，我们构建复合词内部结构词典时目的性更强一些，换句话说，我们构建这一词典是为通过内部结构来识别新词的语法类和语义类服务，这一思想贯穿了我们的整体构建过程。

3 汉语复合词的基本构造类型

要分析现代汉语词语的内部结构，首先就要弄清楚这些词语是怎么来的。一般认为，现代汉语复合词产生方式有三种，分别是类推、词汇化和简缩，此外还有一些词语分别来自汉语方言和汉语以外的语言。每个新词语自然会有相应的新词义，除此之外，一些旧词语在使用中也会产生新的意义。因此，我们可以依据义项的来源将语义词典中的词条分别归入三大类九小类(见表格1)。不同类别的词语，具有不同的来源，其内部成分与整体之间的关系也不同。

以上三个大类中，已有词语产生的新义项比较容易处理，因为每个新义项都对应有一个本义，该词语内部结构的分析应参照本义进行。外来词中音译词是单纯词，没有内部结构，方言借词和日语借词数量有限，特殊对待，也不进行内部结构分析。因此，复合词内部结构分析的主要对象是第一大类三个小类的词语。对于第二大类和第三大类的词语，我们仅标注其所属类别，而不进行内部结构分析。

第一大类“来自语言内部的词语”包括三个小类，本文分别称之为缩略词、词汇化词和类推词。其中缩略词的内部结构需要与其原形结合起来分析，缩略词的成分本身并没有语法类和语义类分析的问题。因此，内部结构分析的主要对象是词汇化词和类推词。

表格 1 语义词典义项分类列表

来自语言内部的词语	类推词	来自已有的构词模式（复合、附加）	网民、股民
	缩略词	来自于语言内的短语或跨层结构	科技、南开
	词汇化词	来自于语言内的短语、句法结构或跨层结构	虽然、但是
外来词	方言借词	来自于方言	般配、帮衬
	音译词	来自于其它语言	坦克、吉普
	仿译词	同时来自于其它语言与本语言	啤酒、卡片
	日语借词	来自于日语	空间、组合
已有词语的新义项	修辞构词	来自于语言内的词语	暗礁、暗流
	旧词新义	来自于语言内的词语	小姐、教授

4 词典构建方案

构建内部结构词典的主要工作是对Hownet中的所有双字词语，标注其内部成分的词性、语义类以及语法结构关系，具体内容如下：

（1） 成分语法类使用2000年人民日报标注语料库（基于北京大学计算语言学研究所2003版词性标记集，含106种词性）（俞士汶等，2003）。

（2） 成分结构关系类型包括并列、定中、状中、补充、支配、陈述、前缀、后缀、名量、重叠、动介、连动、跨层、同位、简缩等15种。

（3） 成分语义类使用Hownet中定义的语义类。

我们构建此词典的基本方法是以人工标注、自动标注交叉验证，并辅之以启发式规则来发现标注过程中存在的错误和不一致问题。具体流程如下：

首先，标注人员按照规范（将另文叙述）进行手工标注；同时，用三种自动标注方法分别独立进行标注。

之后，用人工标注结果与自动标注结果进行交叉验证，将不一致的地方提交给校对人员核对。同时，利用启发式规则发现人工标注过程中可能错误和不一致的地方，也提交给校对人员。

根据校对人员反馈回来的修改结果进一步扩展，找出自动标注结果没有覆盖到而可能出现同类问题的词语，再提交给校对人员核对。

这一构建流程的主要优势在于可以较好地解决标注一致性问题。限于人力和财力，一般不太可能进行双人交叉验证式标注。常见的做法是在自动标注的基础上辅以人工校对和一些后处理措施。当参与人员较多时，一致性问题比较难以解决。基于这一原因，我们没有采用这种常规做法，而是代之以人工标注、自动标注交叉验证，并辅之以启发式规则来发现标注过程中存在的错误和不一致问题，以提高词典的质量。

5 自动标注方法

在自动标注过程中，我们使用了三种方法。第一种方法主要处理定中式名词，其依据是一部分定中式名词的内部结构比较清晰，词语的意义等于其成分意义之和，因此依据 Hownet 对这些词语的形式定义就可以猜出各成分的意义。另外两种方法是双向平行类推法和成对平行类推法。这两种方法主要是为了处理多义成分歧义的消解。

5.1 方法一：使用 Hownet 中词语的 DEF

方法一基于 Howent 中每个词语的 DEF 来猜测成分的 DEF。Hownet 每个词语都给出一个形式化的语义解释，称之为 DEF。例如，词语“男人”的 DEF 为“{human|人:modifier={male|男}}”。方法一的基本思想是：有一些词语的意义等于其成分的意义之和，因此在已知词语意义的情况下，可以根据词语意义反过来推测其成分的意义。例如，已知词语“男人”的 DEF 为“{human|人:modifier={male|男}}”，“男”的多个义项中有一个是“{male|男}”，“人”的多个义项中有一个是“{human|人}”，因此可以根据“男人”的语义解释猜测出其成分的语义解释。

5.2 方法二：双向平行类推

对于一些通过类推方式产生的词语，可以通过 Bootstrapping 方式猜测类推成分的意义。我们称这类词语为类推词，类推过程中不变的成分为共同成分，变化的成分为替换成分。例如“好人、恶人、歹人、圣人、高人、坏人、完人”为一组类推词，其中“人”为共同成分，“好、恶、歹、圣、高、坏、完”为替换成分。这一组替换成分有很多具有多个义项，但是在这一组类推词中，它们都具有一个类似的义项。

因此，基于类推词所具有的特性，我们可以通过 Bootstrampping 的方式来推测其中的替换成分的意义。虽然有一些替换成分具有多个义项，但通过一组替换成分义项类似的限制，可以找出这一组替换成分在当前这一组类推词中的义项。

有两种 bootstrapping 方式可以用于猜测成分的意义，一种是双向平行类推，一种是成对替换类推。所谓双向平行类推要求一组词语的成分语义类和词语语义类两个方向平行，假定一组语义类相同的词语其替换成分所具有的多个义项中有一个是共同义项，则可以判定此共同义项为替换成分在当前这一组词语中的义项。

双向平行类推法的具体流程如下：

（1） 以首字为共同成分，后字为替换成分，标记后字的语义类。建立词典的首字表及相应的首字索引，每个字下面分别记录以该字为首字的所有词语；

（2） 遍历首字表，通过首字索引查看相应的词语，如后字不在词典中，则忽略该词；如后字在词典中，则保留；

（3） 遍历保留下来的词的所有语义类，将词语语义类分组，每个词语可以归入多组，记录各语义类的频次；

（4） 遍历频次大于 2 的组，看组中替换成分是否有共同语义类，如果只有一个共同语义类，则生成一条规则，其格式为<共同成分，替换成分语义类，整体语义类>，将这些词语中的所有替换成分标记为对应的语义类；若有多个共同语义类或者没有共同语义类，则放弃。

（5） 以后字为共同成分，前字为替换成分，重复（1）-（4），标记前字的语义类。

根据上述流程标记成分语义类，前后字都标出语义类的共 5237 条词语；仅前字标出语义类的共 14684 条词语，仅后字标出语义类的共 11557 条词语。

5.3 方法三：成对平行类推

另外一种 Bootstrapping 方式为成对替换类推。所谓成对替换类推要求一对单字具有共同的语义类，并且它们与多个其它字组合成的词对分别具有共同的语义类。其基本思想是：假设字 A、B 分别与字 C_1、C_2、……C_n组成词对{AC_1、BC_1}、{AC_2、BC_2}、……{AC_n、BC_n}，每个词对中的两个词均具有共同的语义类，我们称之为替换词对；如果字 A 与字 B 具有相同语义类，则称之为词对中 A 与 B 的共同语义类。

成对平行类推算法具体流程如下：

（1） 建立同义字对表，将语义类相同的两个字归为一个同义字对，存入表中；如果两个字具有两个或两个以上的共同语义类，则放弃；

（2） 以首字为共同成分，后字为替换成分，标记后字的语义类。建立词典的首字表及相应的首字索引，每个字下面分别记录以该字为首字的所有词语；

（3） 遍历同义字对表，通过首字索引查看相应的词语；如果一个同义字对存在两个以上的替换词对，则该词对中的替换成分为语义类为同义字对的共同语义类；

（4） 以后字为共同成分，前字为替换成分，重复（2）-（3），标记前字的语义类。

根据上述流程标记成分语义类，前后字都标出语义类的共 4285 条词语；仅前字标出语义类的共 8801 条词语，仅后字标出语义类的共 8749 条词语。

5.4 自动分析结果汇总

上述三种方法彼此独立，可以分开使用。其处理的结果中也有许多是重合的。汇总结果如表格 2 所示，共有 16072 个词语前后字均获得结果，25782 个词语前字或后字获得结果，总计 41854 个词语得到了全部或部分标注结果。

表格 2 自动分析方法标注结果汇总

方法	有前后字结果	仅有前字结果	仅有后字结果
1	10272	/	/
2	5237	14684	11557
3	4285	8801	8749
汇总	16072	14503	11279

6 计算机辅助人工标注

6.1 基于词性兼容规则的异常标注结果发现

Hownet 中每个成分的义项都标注有相应的词性，但是 Hownet 从本质上讲是一个语义词典，而且在设计之初较多地考虑了机器翻译的需要，比较注重与英语词语语义的对应关系，因此其词性体系受英语词性的影响比较大，与主流的词性体系差别较大。基于这一原因，在进行内部结构分析与标注时，在成分语法类这一方面我们选择了北大 2003 版词性标注集。由于 Hownet 中的每个义项在 Hownet 的体系中都对应着一个词性，因此在人工标注之后，我们就可以得到两套词性标注结果。

假设两种词性体系之间存在着一定的兼容规则，那么不符合这些兼容规则的就可能是异常的标注结果。基于这一思路，我们总结出如下词性兼容规则[2]：

（1） Hownet 中的词性与北大版中的对应语素语法类兼容，即 H(N)与 P(Ng)、H(ADJ)与 P(Ag)、H(V)与 P(Vg)分别兼容。

（2） Hownet 中的大类与北大版中对应大类的小类兼容，即 H(N)与 P(nt)、P(nz)、P(nr)、P(nrf)、P(nrg)，H(V)与 P(vt)、P(vi)、P(vu)分别兼容。

（3） Hownet 中的形容词与北大版中的名词兼容，即 H(ADJ)与 P(n)、P(nt)、P(nz)、P(nr)、P(nrf)、P(nrg)兼容。

不符合兼容规则的标注结果如果是错误的标注，可能是由以下几种原因引起的：

（1） 词性标注错误。虽然语义类标注正确，但词性标注错误。

（2） 语义类标注错误。虽然词性标注正确，但语义类标注错误。

不符合兼容规则的标注结果均被提交给校对人员做进一步的校对。

6.2 方法 1、2、3 结果中词性的自动补充与人工标注结果中词性的自动校正

上述三种自动分析方法主要用于猜测成分的语义，我们还需要用其它的方法为上述三种方法的猜测结果补充相应的词性。

我们采用的方法是根据人工标注结果用投票的方法决定自动分析结果中成分的词性，这一方法同时也可以用于标正人工标注结果中的词性。其基本思路是：从理论上讲，每个字的每个义项只应属于一个词性；人工标注结果中，标注的不一致会导致同一个字同一个义项被标记成多个词性的情况出现。其具体方法是：

（1） 根据投票的方法，记录同一个字同一个义项被标记成各词性的频次，频次高者胜出。

（2） 根据投票结果决定所有单字各义项的词性，并据此为自动标注结果补充词性。

在使用上述方法辅助人工标注时，我们不是直接对人工标注的结果进行修改，而是给出异常提示，由校对人员选择是否接受。

上述流程可以多次运行，一直到不再产生的新的异常结果为止。

6.3 比较人工标注结果和自动标注结果

总计 54000 双音词中，自动标注涉及约 41000 词，其中有一万词因为正确率极高没有提交人工标注。因此人工标注共有约 44000 词，其中有三万条与自动标注重合。将这三万条词语的人工标注结果和自动标注结果相比较，将两者不一致的作为存疑的标注结果，提交给校对人员。

---

[2] H(A)表示 Hownet 中的词性 A，P(A)表示北大 2003 版词性标记集中的词性 A。

7 结论与展望

依据上述人工与自动相结合的方法，我们初步完成了 Hownet 中 54000 双音词内部结构分析的工作。今的我们还计划从以下三个方面进一步发展这部词典。第一，从双字词扩展到三字词和四字词，争取覆盖到 Hownet 中的所有名词、动词、形容词。第二，从 Hownet 扩展到其它的语义词典，建立基于其它词典的内部结构信息词典。第三，从时间的角度区分出词典中所有词语的时间层次，尽可能地找出各词语的词源和出处。

在该词典的基础上还可以进一步开展以下几个方面的工作。首先，可以该词典为基础进行复合词构词规律的挖掘。基于复合词的内部结构信息，可以统计成分词性序列与词语词性之间的关系，成分语义类序列与词语语义类之间的关系。其次，可以使用该词典训练一个双音复合词成分语法类标注工具。词语成分的语法类标注与词语的词性标注有比较大的区别，基于词内部结构信息训练的词语成分语法类标注工具应该比直接使用词语词性标注工具来进行词语内部成分的语法类标注更效。此外，还可以依据内部结构信息进行新词词性和语义类自动标注的工作。

参 考 文 献

[1] 陈保亚，2005，再论平行周遍原则和不规则字组的判定，汉语学习，第 1 期。

[2] 陈保亚，2006，论平行周遍原则与规则语素组的判定，中国语文，第 2 期。

[3] 董秀芳，2004，汉语的词库与词法，北京：北京大学出版社。

[4] 董振东、董强，2006，Hownet And the Computation of Meaning. World Scientific Publishing Co., Inc. River Edge, NJ, USA.

[5] 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》，2003(04)。

[6] 亢世勇，2004，基于数据库的现代汉语语义构词初探，《第五届中文词汇语义学学术会议论文集》，新加坡国立大学出版，还发表于《汉语语言与计算学报》2005 年 2 期。

[7] 黎良军，1995，汉语词汇语义学论稿，广西：广西师范大学出版社。

[8] 刘叔新，1990a，汉语描写词汇学，北京：商务印书馆。

[9] 刘叔新，1990b，复合词结构的词汇属性-兼论语法学、词汇学同构词法的关系，《中国语文》1990 年第 4 期。

[10] 吕叔湘，1962，关于"语言单位的同一性"等等，《中国语文》第 11 期。

[11] 王洪君，1999，《"逆序定中"辨析》，《汉语学习》，1999 年第 2 期。

[12] 王洪君，2005，动物、身体两义场单字组构两字的结构模式，《语言研究》2005 年第 1 期 1-11 页。

[13] 杨梅，2006，现代汉语合成词构词研究[D]，博士论文，南京师范大学。

[14] 尹斌庸，1984，汉语语素的定量研究，《中国语文》1984 年第 5 期。

[15] 俞士汶，1999，自然语言理解与语法研究，马庆株编《语法研究入门》240-251，北京：商务印书馆。

[16] 俞士汶、慧明、朱学峰、孙斌、常宝宝，2003，北大语料库加工规范:切分・词性标注・注音，《汉语语言与计算学报》(新加坡)，第 13 卷 2 期。

[17] 苑春法、黄昌宁，1998，基于汉语语素数据库的汉语语素及构词研究，《语言文字应用》1998 年第 1 期。