基于词典的汉语名词语义信息的自动分析与获取

翟高寿* 张永奎** 杨尔弘** 查建中*

* 北方交通大学机械工程系智能工程研究室，北京 100044

** 山西大学计算机科学系，太原 030006

摘　要： 针对词汇语义研究和计算词典方法的兴起及汉语信息处理中名词语义研究与其所占重要地位很不相称的现状，我们立足于《现代汉语名词机读词典》，主要采用基于规则的语义分析策略和参照性分析方法并综合考虑词频、句法位置等因素，对其释义文本实现了基本语义关系和相应语义核心词的识别和提取，以推动汉语名词机用词典的建造进程。

关键词： 词汇语义，计算词典方法，机读词典，机用词典

Automatic Acquisition of Semantic Information of Chinese Nouns based on Defining Texts of MCNMRD

Zhai Gaoshou*, Zhang Yongkui**, Yang Erhong**, Cha Jianzhong*

* IERL, Dept. of Mechanical Engineering, Northern JiaoTong University, Beijing 100044, P.R.China

** Dept. of Computer Science, ShanXi University, Taiyuan 030006, P.R.China

ABSTRACT: In view of the situation of the researches on semantics of nouns that are very unworthy of their importance in Chinese Information Processing(CIP), we have tried to analyse and acquire semantic knowledge of nouns automatically from Defining Texts of MCNMRD that has been derived from Modern Chinese Dictionary. The methods exploited are mainly based on rules, and semantic characteristics of Chinese nouns,word occurence and syntactic position are also considered.This work can be used to carry out further semantic research and will lay foundation for the construction of Machine Tractable Dictionary for CIP.

KEYWORDS: Lexical Semantics, Computational Lexicography, MRD, MTD

一、 引　　言

自然语言处理各类任务实现自动化的需要和现代语言学理论体系中词汇主义的兴起，使得词汇资源日益受到人们的重视[1]。计算词典方法发展成为词汇信息获取的重要技术之一[1-5]。语法理论的日臻成熟，句法分析的局限性，使自然语言处理技术向语义分析倾斜[6]，词汇语义的研究成为焦点所在。

词典作为基本常识的主要来源，已成为获取词义以建造和组织词汇知识库的基础。目前对机读词典语义信息的提取主要采用启发式的归纳方法，即通过抽取最基本的概念与其间的语义关系，对自然语言定义中发现的共同元素逐级推广，从而把隐含在词典释义文本中的常识形式化、结构化来实现机读词典向机用词典的转化[3]；且语义机用词典的知识表示一般采用“语义分类＋属性描述”的方式[7]。

鉴于汉语中各种词类的句法、语义乃至其它信息各有侧重，各具特色，因而不能一概而论。也就是说，一下子对所有的汉语词类进行句法、语义等语言学信息的分析和研究是不现实和不可行的，应采取“各个针对性击破”的技术路线，有关思想可参图 1：

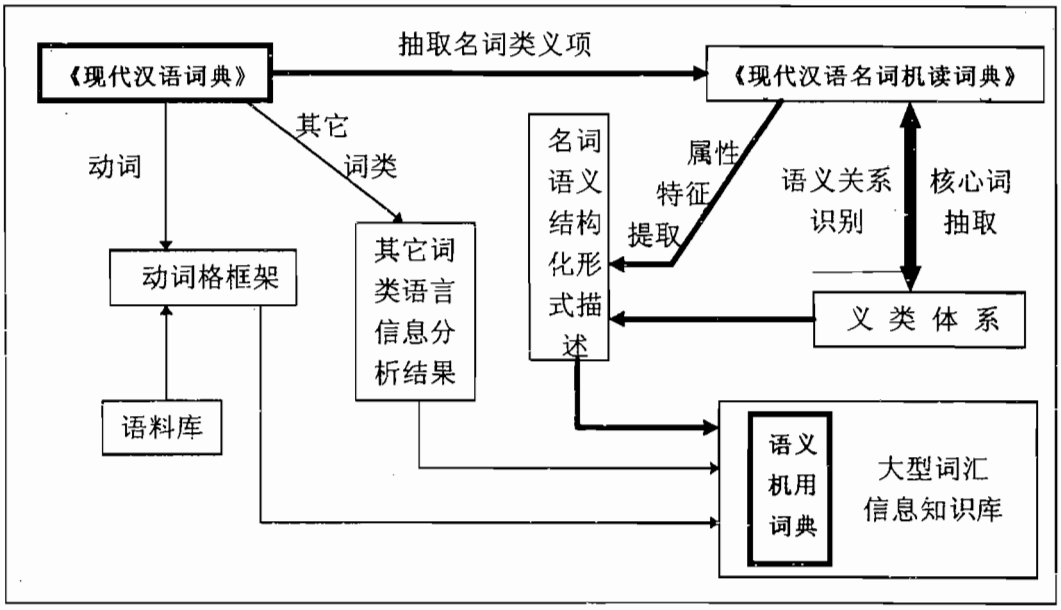

图 1　语义机用词典的建造路线

现代汉语信息处理的关键在于名词和动词两类词的语法、语义研究的深入[11]。目前针对动词的研究很多，规模较大的如《现代汉语述语动词机器词典》[8]；相比之下，专门研究名词的语法、语义特征的人员则不多，并且多数停留在人工分析的层次，这与名词在汉语信息处理中所占的重要地位很不相称；而名词对汉语句子的制导主要通过其语义概念来实现，所以我们拟开展基于词典的汉语名词语义机器分析的尝试性研究，所用资源为从《现代汉语词典》[9]为主体的机读资源抽取所有名词及对应义项而建立的《现代汉语名词机读词典》[11,12]。研究方案具体设想如下：(1) 针对各词的释义文本识别有关语义关系和语义核心词，形成《现代汉语名词机用词典》的初级版本；(2) 对(1)所得的上下位关系类型的语义关系和语义核心词加以整理并完备化，进而建立概念层级结构并辅以统计和人工组织整理，最终初步形成现代汉语名词的语义分类体系；(3) 利用当前流行的文本到数据库转换[10]的思想和技术实现部分语义类中名词的释义文本的形式化结构描述，余者则采用机助方式完成。本文将主要讨论语义关系和语义核心词的识别及抽取技术。

二、 基于词典的汉语名词语义信息获取的基本思想

2.1 语义分析策略的选取

名词的义项释义大多为名词性短语或段落，而后者也多以名词性短语打头。释义中一

般至少包含上下位关系、同义/近义关系和部分整体关系三种基本语义关系之一，且作为标志的对应语义核心词通常位于释义的第一句段或较靠前的句段中——语义核心词随各种关系类型分别指向上位概念、同义/近义概念和整体概念。举例说明如下：

00135 微安

$00 电流单位，一安培的一百万分之一。

00162 口岸

$00 港口

03725 回肠

A00 小肠的一部分，上接空肠，下连盲肠，形状弯曲。

03737 香肠

$00 用猪的小肠，装上碎肉和作料等制成的食品。

“微安”与“电流单位”之间为上下位关系，后者是前者的上位概念并是释义的语义核心词。“口岸”与“港口”之间为同义关系，后者同样是释义的语义核心词。“回肠”与“小肠”之间是部分整体关系，后者是前者的整体概念。“香肠”与“食品”之间也为上下位关系，显然这种情况要稍复杂些，其难点在于“的”字结构的分析。

鉴于词典释义用词的相对集中和句法结构的相对简单，当前从释义中提取语义信息均采用模式匹配为主的分析策略。《现代汉语名词机读词典》的释义同样具有这些特点，故可采用基于规则的分析技术（辅以简单的句法分析主要是“的”字结构的层次分析）决定候选语义核心词集并综合考虑参照性原则和句法位置、词频等因素加以唯一确认的策略。

2.2 基于规则的分析方法

以上对名词释义文本内容和结构的分析说明，依靠规则（即模式匹配的方法）可在一定程度上解决语义核心词的抽取和语义关系的识别问题。在对词典释义进行机助方式分析的基础上，我们总结出十多条规则以实现语义关系的识别和候选语义核心词集的抽取，其中某些词条的义项的语义核心词并得到唯一确认，余者则可根据“参照性”原则和“词频＋位置”的思想加以唯一确认。列举两条语义分析规则如下：

RULE-1 (Tail[Clause]＝"之一")∧(Clause:＝ Clause －"之一")∧ GetNP(Clause,NP_)
∧(Len(NP_)<11)├ NP_∈ SC_SET

RULE-2 (Clause ＝ LMT)∧ Non_DE(Clause)∧(Len(Clause)<7) ├ SYNONYM:＝ Clause

(其中，LMT,Clause 分别表示释义自身及其中的一个句段，Tail[ ]表示指定串的尾部，GetNP(), Len(), Non_DE()为分别用于提取名词、求词长、判断是否含“的”字的函数; NP_∈ SC_SET 表示 NP_仅仅有资格入选语义核心词集并尚待确认,其语义关系为上下位关系; SYNONYM:＝ NP_ 表示语义关系为同义/近义关系,且 NP_ 已被成功确认为对应名词在该义项的同义/近义词)

2.3 参照性方法和"词频＋位置"分析法

经过规则性语义分析后，绝大多数义项的语义关系类型得到识别，且其中部分义项的

语义核心词得到确认；而对于上下位关系类型的义项一般得到的是候选上位词集，尚待进一步处理。汉语具“义合性”特征，且名词和名词性短语多为偏正结构，故而相同尾部的名词的上位概念通常相同，因此可参照已确认者对同尾字的相应词条义项做上位词的唯一选择。此外，就上上位概念词而言，其出现在释义中的机会较多故而入选“候选上位词集”所得的词频较高，并且语义核心词在释义文本前面出现的概率较大，这些也可作为抽取和识别的重要依据。以上所述分别构成参照性方法和"词频＋位置"分析法的基本思想。

三、实现与结果

3.1 词典结构的调整

《现代汉语名词机读词典》的原词条结构用 BNF 范式形式化描述如下：

〈 Entry 〉 ::= 〈 Entry-No 〉 〈 Word 〉 〈 Meaning-Description-Part 〉 〈 End-Mark 〉

〈 Entry-No 〉 ::= 〈 D_1 〉 〈 D_2 〉 〈 D_3 〉 〈 D_4 〉 〈 D_5 〉

〈 D_1 〉 ::= 0 | 1 | 2 | 3 | 4 | 5

〈 D_i 〉 ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 （i = 2, 3, 4, 5）

〈 Meaning-Description-Part 〉 ::= { 〈 Meaning-Item 〉 }$^+$

〈 Meaning-Item 〉 ::= 〈 Meaning-Item-Code 〉 〈 Defining-Text 〉

〈 Meaning-Item-Code 〉 ::= 〈 PinYin-Code 〉 〈 Meaning-Item-No 〉

〈 PinYin-Code 〉 ::=$ | A | B | C | D | E

〈 Meaning-Item-No 〉 ::= 00|01|02|03|04|05|06|07|08|09|10|11|12|13|14|15|16|17|18|19

〈 End-Mark 〉 ::= $ end

语义信息分析和获取的整个过程是逐步完善的，单就语义关系的识别和语义核心词的抽取的工作而言也是如此，所以做好分析过程进展情况的记录相当重要。为此，需对词典的义项代码做以下扩展，以利于信息保存和方便操作：

〈 Meaning-Item-Code 〉 ::= 〈 PinYin-Code 〉 〈 Meaning-Item-No 〉 〈 Semantic-Code 〉

〈 Semantic-Code 〉 ::= 〈 Semantic-Relation-Code 〉 〈 SKs-Num 〉 〈 Result-Type 〉

〈 Semantic-Relation-Code 〉 ::= 0|1|2|3 （分别代表“语义关系未知”、“上下位关系”、“同义／近义关系”和“部分整体关系”）

〈 SKs-Num 〉 ::= 0|1|2|3 （表示候选语义核心词的数目：识别失败则为 0；抽取后得到唯一确认时应为 1；所得候选词不止一个且暂时难以确认，则为 2 或 3）

〈 Result-Type 〉 ::= 0|1|2 （分别代表“未识别”、“识别所得候选词多于一个且无法确认”、“识别得到唯一确认”）

与此同时，一旦识别了语义关系即当〈 Result-Type 〉不为 0 时，词典的语义描述项〈 Meaning-Description-Part 〉中应增加语义信息项，以记录（候选）语义核心词（集）：

〈 Meaning-Description-Part 〉 ::= { 〈 Meaning-Item 〉 [〈 Semantic-Information-Item 〉] }$^+$

〈 Semantic-Information-Item 〉 ::= 〈 Semantic-Relation-Type-Identifier 〉 { 〈 SK 〉 } $^{+}$

（〈 SK 〉是从释义中提取而得的[候选]语义核心词，其个数与〈 SKs-Num 〉项所标明的相同）

〈 Semantic-Relation-Type-Identifier 〉 ::= CLASS | SYNONYM | PARTOF

3.2 实现过程中的关键技术

经过规则性语义分析后，为对上下位关系类型的义项的上位词进行唯一确认，需建立上位词库和候选上位词库。由于词库文件庞大，故词库的压缩即合并相同的上位词或候选上位词并进行分块整理和统计是难点和关键所在。为避免由于词形比较而频繁打开／关闭文件所造成的内、外存切换引起的时间浪费，我们采用设定文件指针来回移动对词库进行二次扫描的算法。同时，因为语义分析过程将频繁地访问这两个词库从而使得检索成为一笔不可忽视的时空开销，所以需把它们调整为按词尾字分块排列结构，以便采用基于分块索引的方法实现较快速的搜索访问。

3.3 《现代汉语名词机用词典》初级版的生成

根据语义分析所得结果，我们生成了《现代汉语名词机用词典》初级版本；同语义分析时相比，其词条结构主要在〈 Meaning-Item-Code 〉和〈 Semantic-Information-Item 〉两项发生了变化，用 BNF 范式描述如下：

〈 Meaning-Item-Code 〉::=〈 PinYin-Code 〉〈 Meaning-Item-No 〉〈 Semantic-Relation-Code 〉

〈 Semantic-Information-Item 〉::=〈 Semantic-Relation-Type-Identifier 〉〈 SK 〉

（〈 SK 〉对应为释义中抽取并确认的上位词、同义／近义词或整体概念词，余者定义同前）

四、系统结果评测和结论

4.1 系统结果评测

通过实现上述语义分析过程，占 92.90 %的名词义项(32936/35452) 的语义关系和语义核心词得到识别、抽取和确认，而未识别的义项数为 2516, 仅占 7.10 %，处理结果如表 1 所示。另外，大致按每隔 30 条义项抽取一个已被成功提取语义信息的义项作为样本的方式共抽取了 1000 个词的 1000 条义项的识别结果来进行抽样测试，所得测试数据如表 2 所示。在语义分析结果出错的 158 个义项中，因名词识别错误（即本来不是名词而被误认为是名词）引起的有 44 个，占到 24.85 %；而由于原机读词典录入错误、释义文本自身或分析规则的不合理等原因导致出错的共 144 个义项，占 72.15 %。这说明前面的一些工作特

表 1 现代汉语名词机读词典释义的语义关系分析与识别结果

语义关系类型	上下位关系	同义关系	部分整体关系	未识别
义项数	**27151**	**5748**	**37**	**2516**
所占比例	**76.59%**	**16.21%**	**0.10%**	**7.10%**

表 2　词典释义语义信息分析测试结果

错误类型	语义关系错	语义核心词错	两者都错	两者都对
义项数	52	53	53	842
所占比例	5.2%	5.3%	5.3%	84.2%

别是名词识别和语义分析规则的提取等还有待于改进。

4.2 结论及启示

实验结果表明：基于机读词典释义文本提取汉语名词语义信息是可行和有效的；同时，分开词类来进行语义研究的技术路线具有针对性强、简化处理、提高可操作性等优点，特别对于词类不同而区别明显的分析对象如词典释义更是如此。

我们还应看到，虽然语义关系及其核心词的识别与抽取仅主要采用基于规则的方法便可奏效，但其分析焦点大多集中在模式简单、易于处理的释义文本首部，若要进一步提取深层次的语义信息，一定程度的句法、语义分析将必不可少。

本项课题得到国家自然科学研究基金的资助。

参　考　文　献

[1] B.Boguraev&TedBriscoe(eds)., Computational Lexicography for Natural Language Processing(Longman, London), 1989

[2] Huang Chang-Ning, Derivation of Definition Primitives from a Monolingual Dictionary, in: Natural Language Processing Pacific Rim Symposium'95(Seoul)

[3] N.Calzolari,A.Zampolli, Methods and Tools for Lexical Acquisition, Advanced School in Artificial Intelligence(2nd:1990:Guarda),Natural Language Processing:EAIA'90:Proceedinds,pp.4-24

[4] H.Alshawi, Processing Dictionary Definitions with Phrasal Pattern Hierarchies, Computational Linguistics, 1987,13(3-4),pp.195-202

[5] Y.Wilks, D.Fass, Cheng-Ming Guo, J.E.McDonald,T.Plate,B.M.Slator, Providing Machine Tractable Dictionary Tools,Machine Translation,1990,(5),pp.99-154

[6] 俞士汶,自然语言的歧义与机器翻译的对策,《中文信息学报》,1989,3(2)

[7] 姬东鸿,黄昌宁,赵　军,汉语名词的内在知识及其表示,《计算语言学进展与应用》, 清华大学版社, 1995,pp.37-42

[8] 陈群秀,黄昌宁,程　红,现代汉语述语动词机器词典研究初探,《计算语言学研究与应用》, 北京语言学院出版社,1993,pp.231-236

[9] 中国社会科学院语言研究所词典编辑室编,《现代汉语词典》,商务印书馆,1994

[10] 张永奎,从文本中提取信息,情报学报,1994,13(2),pp.102-107

[11] 翟高寿,张永奎,杨尔弘,利用基于语义信息的名词识别方法来建造现代汉语名词机器词典,《计算语言学进展与应用》,清华大学出版社,1995,pp.195-200

[12] 翟高寿,基于词典的汉语名词语义信息的自动分析与获取,山西大学硕士学位论文,1996