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

翟高寿　张永奎　杨尔弘
（山西大学计算机科学系，太原 030006）

摘要：鉴于名词的研究状况与其在现代汉语信息处理中所占重要地位很不相称，本文提出基于语义信息的名词识别方法，用于从以《现代汉语词典》为基础的机器词典中抽取所有名词及其对应义项，进而建造《现代汉语名词机器词典》以服务于汉语名词的语法、语义特征的研究。

Construction of Modern Chinese Noun Machine Dictionary by noun—identifying rules based on Semantic Information

Zhai Gaoshou　Zhang Yongkui　Yang Erhong
(Dept. of Computer Science, ShanXi University, Tai Yuan 030006)

Abstract: In this paper, we describe a method to extract all nouns with their corresponding meaning items from Machine Readable Resource of Modern Chinese Dictionary by noun—identifying rules based on Lexical— Semantic Information; and this method will be used to build Modern Chinese Noun Machine Dictionary.

1 引　言

随着计算机应用领域的拓宽，知识工程的发展，信息量的急剧增长，词汇资源日益引人注目；机器词典的建造已成为自然语言处理系统（Natural Language Processing System，NLPS）的一个瓶颈问题，越来越多的语言学家和计算语言学工作者把其看作是自然语言处理的支柱和语言工程的基石[1,2]。

汉语没有印欧语那样的形态，动词是句子的组织核心，但句法格局面貌却由名词这样的外部形态来确定[3]。现代汉语信息处理的关键问题在于名词和动词两类词的语法、语义研究的深入。这不仅由于二者在汉语语法、语义方面的无可替代的支配地位，更在于二者在汉语词汇量中的压倒优势的份额。目前从事动词研究的计算语言学工作者很多，课题立项初具规模的如《现代汉语述语动词机器词典》的建造工程[4]；相比之下，专门研究名词的语法、语义特征的人员则不多，这与名词在汉语信息处理中所占的重要地位很不相称。因此，开展

针对名词的计算语言学研究工作具有十分现实的意义。由于还没有公开出版的汉语名词词典，所以建造一部现代汉语名词机器词典的工作显得迫切和困难。

2 建造现代汉语名词机器词典的初步设想

当前机器词典的建造方式一般可归为以下三种：(a)在机器辅助下主要依靠人工来生成，其投入大量的人力来描述词条信息，耗资惊人，典型的例子是日本的电子词典开发计划(EDR)；(b)从现有词典印刷版本出发生成机读词典(MRD)，然后抽取各种词汇知识建立机循词典(MTD)；(c)通过对大规模真实文本即语料库的分析获取有关词汇信息来构造机器词典[5]。

我们拟采用建造方式(b)来构建《现代汉语名词机器词典》，然后在此基础上针对名词的语义特征作进一步深入研究 。该名词机器词典构建的基本思想是：从以《现代汉语词典》[6]为主体的机读词典资源(以后为简便起见，一律称作"《机读现汉》")中抽取所有名词及其义项描述信息等，自动生成《现代汉语名词机器词典》。不难看出，其关键在于名词的抽取即识别过程。

《机读现汉》具有两个特点：① 词条的各词汇信息项间有明确的界限标识符；② 词条按词尾字的读音进行排列。这使得我们在名词识别过程中所采纳的一些策略的可行性和可靠性得到保证。

《现代汉语名词机器词典》(以下简称《名词词典》)的词条排列标准仍为词尾字的读音，这意味着以同一字结尾的词条将放置在一起，从而可为后面数量不少的偏正结构名词的语义分析创造有利条件。其词条的框架结构定义如下：

〈词条〉::=〈代码部分〉〈词形〉〈义项部分〉〈结束符〉
〈代码部分〉::=〈代码符〉〈代码值〉
〈义项部分〉::={〈义项描述部分〉[〈义项举例部分〉]}$^{+}$
〈义项描述部分〉::=〈义项符〉〈义项描述〉
〈义项举例部分〉::=〈义项举例符〉〈义项举例〉
〈代码符〉::=№
〈义项符〉::=¥#[(〈义项号〉)]
〈义项号〉::=1|2|3|4|5|…
〈义项举例符〉::=〈义项符°〉〈举例符〉
〈义项符°〉::=〈与对应义项描述部分相同的义项符〉
〈举例符〉::=ex：
〈结束符〉::=¥end

鉴于《名词词典》构建的主要工作是抽取名词，本文将重点讨论名词的识别方法。

3 名词识别与词性标注的区别

显然，名词识别是词性标注的一个重要组成部分。不过，以往的词性标注技术均针对语

料库而设计，就机器词典这一内容和结构与其截然不同的对象而言，有关方法必难以完全类同，这使得我们不可能用已成形的词性自动标注技术来实现名词抽取的目标，而只能在领会有关技术思想实质的基础上来规划我们的名词识别方法。名词识别与词性标注具有如下区别：

(1) 实施对象不同，二者分别为语料库和机器词典；

(2) 目标不同，传统的词性标注技术实现语料库的所有的词的词性显式化、把生语料转化为熟语料，名词识别则试图把《机读现汉》中所有的名词抽取出来；

(3) 结果不同，前者是从语义角度出发来进行名词辨别与抽取，所得名词其实不只是"名词"，而且包括名词性短语(这主要是《机读现汉》中的词条包含不少短语的缘故)，后者得到的标注结果则是从语法功能角度出发的真正的词性；同时，前者所得词条的名词或非名词特征是针对其相应义项而言的，仍存在一定的歧义性(词义与词性并非一一对应关系)，后者则是经过兼类词的排歧过程确认该词在特定语言环境中的唯一词性；

(4) 所利用资源及使用技术的出发点不同，前者立足于词典词条的释义文本即语义信息、通过分析其结构与语义信息来确认词条是否为名词，后者则利用后缀表和受限框架或对人工标注的熟语料库统计所得的概率计算模型。

4 基于语义信息的名词识别方法

成熟的语料库词性自动标注技术可分为基于规则的、基于概率统计模型的及混合型三类。鉴于《机读现汉》的最具可用性的良好资源首推释义文本(Lexical Meaning Text，LMT)，而名词的语法、语义特征在很大程度上有赖于词尾部分即词缀，因此我们可首先在计算机辅助下对词形的结构以及释义文本的内容和结构进行详尽的分析，总结出若干规则构成名词抽取技术的操作算子，然后扫描机器词典中词条的释义文本并执行对应操作完成词条"是否为名词"的判别。需指出的是，扫描过程将是反复的，随着未识别词条数量的减少，不断地加入新规则，直到所有的词条被识别完毕。

4.1 LMT 结构规则

通过对词条的释义文本结构的机助分析，我们总结出若干条针对其特点可正确区分名词与非名词的规则，按规则条件对应释义文本的首句或尾句部分可分为首句规则与尾句规则，注意这里所说的首句或尾句部分均指文本中"，"、"。"、"；"分隔的句段，可能是短语，与日常意义下的"句子"不同。为便于说明，约定如此的符号表示：First 、Last 分别表示释义文本的首句和尾句，Word 表示词条的词形，STRING[Head]、STRING[Tail]分别表示字串的首部子串与尾部子串(具体应用时 STRING 以 First、Last、Word 替代)，SINGLE 表示释义只有一个句子或当前分析者为由"；"分隔的若干句段之一；SL、WL 、Len(STRING)分别表示当前所分析的释义句、词条或字串的长度(以汉字计数)；STRING[i]表示字串的第 i 个字。

4.1.1 首句规则

根据词条的释义文本的第一句的结构特点，建立首句规则。举例如下：

R1－1： First[Tail]="之一" ⇒ BelongTo(Word,Noun)

如：阿昌族　我国少数民族之一，…

白露　二十四节气之一，…

雌蕊　花的重要部分之一，…

R1－5：　设 KeyWordSet＝{简称，旧称，尊称，也叫，又称}，则

First[Head]∈KeyWordSet ⇒ BelongTo(Word，Noun)

如：保山　旧称保人或媒人。

令爱　尊称对方的女儿。…

R1－18：　(First[Tail]＝Word[Tail])∧(First[SL－Len(Word[Tail])]＝"的")

⇒ BelongTo(Word，Noun)

如：绝对高度　以平均海水面做标准的高度。

屠刀　宰杀牲畜的刀。

词人(1)　擅长作词的人。

R1－19：　(4≤SL≤9)∧SINGLE∧(First[k]＝"的")∧(2≤k≤SL － 1)

∧(First[Tail]≠"样子") ⇒ BelongTo(Word，Noun)

如：墨迹　墨的痕迹

墨吏　贪污的官吏。

老妪　年老的妇女。

R1－a：　Single∧(First[Tail]＝"的") ⇒ BelongTo(Word，NotNoun)

如：矮　高度小的

白　没有加上什么东西的；…

R1－c：　First[Head]＝"形容" ⇒ BelongTo(Word，NotNoun)；

如：井井有条　形容条理分明

缕缕　形容一条一条，连续不断

R1－d：　设 KeyWordSet＝{不，使，在}，则

Single∧(First[Head]∈KeyWordSet)∧("的"∉First)

⇒BelongTo(Word，NotNoun)

如：不安　不安定；…

妨碍　使事情不能顺利进行；…

立案　在主管机关注册登记；…

4.1.2 **尾句规则**

根据词条的释义文本的最后一句的结构特点，建立尾句规则。举例如下：

R2－1：设 KeyWordSet＝{叫，称，是}，则

(∃ x)((x∈KeyWordSet)∧(Last[SL－WL]＝x)

∧(Last[Tail]＝Word)) ⇒ BelongTo(Word，Noun)

如：茭白　…，做蔬菜吃叫茭白。

反比　…，就是反比。

哀子　旧时死了母亲的儿子称哀子。

R2－2：设 KeyWordSet＝{也叫，又称，也作，通称，泛指，也称作，也称做，也泛指}，则

Last[Head]∈KeyWordSet ⇒ BelongTo(Word，Noun)

如：八哥　…。也叫鸲鹆。
苦味酸　…。通称黄色炸药。
公案　…，泛指有纠纷的或离奇的事情。
管带　…，又称海军的舰长。

4.2 词形规则

包括词形结构规则与词缀规则(分为前缀规则和后缀规则)。前者主要用于确认非名词，因为结构是 ABB | AABB | ABAB 的词一般为非名词。由于词语的词性和词义受词缀的影响较大，故我们在机器帮助下先生成名词后缀集(NounSufSet)和非名词前/后缀集(NotNounPreSet/NotNounSufSet)，然后进行判断。上述部分形式化规则分别为(W[i] 表示 Word 的第 i 个字)：

R3－1：(WL＝6)∧(W[2]＝W[3])∧(W[1]≠W[2])
⇒ BelongTo(Word,NotNoun)

R3－4：Word[Head]∈NotNounPreSet ⇒ BelongTo(Word,NotNoun)

R3－6：(Word[Tail]∈NounSufSet)∧(Len(Word[Tail])＜WL)
⇒ BelongTo(Word,Noun)

4.3 义类信息的收集、整理与运用

鉴于在词典中，名词词条的释义文本的第一句往往包含着该词的义类等核心语义信息，如下面一些名词的释义中，划线部分可视为义类信息：

鼻咽癌　鼻咽部粘膜的恶性肿瘤。…
艾(1)　多年生草本植物，…
氯化铵　无机化合物，…
火把　夜间行路时照明的东西。
圆盘耙　碎土、平地的农具，…

所以在 LMT 结构规则和词形规则对《机读现汉》处理的基础上，我们可设计程序自动获取所有未处理词的“义类”信息，收集标准包括词语长度、出现频率等，然后机助半自动挑选那些真正是名词义类的短语构成集合 NSTS(NounSemTypeSet)，并由非名词的高频释义句组成集合 NotNSTS(NotNounSemTypeSet)，做为进一步判别的依据。判断规则分别为：

R4－1：(First[Tail]∈NSTS)∧CONSTRAINTS ⇒ BelongTo(Word,Noun)
(其中 CONSTRAINTS 表示对 First 的句型结构及义类以外剩余部分长度等因素的考虑所加的限定性条件。)

R4－2：(First∈NotNSTS)∧Single ⇒ BelongTo(Word,NotNoun)

4.4 名词识别的具体步骤

上述名词识别方法通过以下步骤来实现：

第一步 从《机读现汉》中依次读入词典各词条的代码与释义信息，对照待处理临时文

件的有关记录，检查当前词条“有无未处理义项”。若无，则读入下一词条进行处理，否则利用 LMT 结构规则和词形规则对当前词条的未处理义项逐一进行分析和识别：当应用规则成功时，把词条代码及义项号相应记入名词临时文件或非名词临时文件；否则将词条代码及未处理的义项号记入待处理临时文件中，并把其词形及“义类”信息（即未处理的词条义项的首句）记入伪义类文件，以备语义信息文件的生成和进一步研究对策使用。

第二步 整理伪义类文件生成名词义类文件和非名词伪义类文件，利用语义信息判别规则进行类似于第一步的识别过程，对应更新有关文件记录。

第三步 检查待处理临时文件是否为空，如不为空则继续分析对策进行规则的收集、整理，重复以上步骤直到待处理临时文件为空。

最后按名词临时文件抽取《机读现汉》的对应词条的词形及义项等有关信息生成《现代汉语名词机器词典》。

5 结束语

借用词语的释义即语义信息实现词性识别的思想，对于从语义角度入手来研究词汇的语法特征是一次初步的尝试，并可为汉语信息处理中语法、语义的并行分析提供借鉴。

本文提出的名词抽取算法力求避开释义文本的分词、词性标注等预处理过程直接获取有关语义信息特征作为判断标准，因而对于汉语词典的词条的词性标注研究并为汉语词典中词条词性的标识提供经验。该算法的难点在于词条的释义文本自身特别是其中名词短语结构的复杂性。由于语义信息的介入，该算法一定程度上消除了部分词的兼类现象。

《现代汉语名词机器词典》的生成将可为计算语言学和语言学的同仁们研究汉语名词的有关特性提供一个机读资源。

最后向给予我们课题研究热忱帮助并提供《现代汉语词典》机读版的清华大学智能技术与系统实验室语言信息处理组致以诚挚的谢意！

参考文献

[1] Bran Boguraev、Ted Briscoe，Introduction，Bran Boguraev & Ted Briscoe(eds).，Computational Lexicography for NLP，1989

[2] 朱学锋，计算机辅助编制机器辞典，《中文信息学报》，1989 年第 3 卷第 4 期

[3] 华萍，现代汉语语法问题的两个“三角”的研究，《80 年代与 90 年代中国现代汉语语法研究》，北京语言学院出版社

[4] 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

[5] 杨尔弘、黄昌宁、张 津，利用机读资源建造机用词典，ICCC'94 国际会议论文集

[6] 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