汉语机用同义词词典的建造技术

杨尔弘　　刘开瑛
（山西大学计算机科学系　太原　030006）

摘要　词汇的语义知识是文本处理的基础。同义、近义和相关是词汇之间语义关系的三个方面，这三种词汇语义关系在文本标注和检索中起着重要作用。本文探讨了建立汉语机用同义词词典的内容和原则。介绍了使用机读资源建立机用词典的方法。词典中描述了词汇的意义，区分了词汇的同义、近义和相关关系，并按同义关系组织了词汇的近义和相关关系。词典采用数据库方式组织。其收词特点决定了该词典的共享性和复用性。可以作为语言信息处理的基础资源。

The Technique In Building Chinese Machine Tractable Synonym Dictionary

Yang Erhong Liu Kaiying
Dept. of Computer Science, Shanxi University, Taiyuan 030006

ABSTRACT: Lexical semantics is the base of text processing. Synonym, near—synonym and relativity are the three aspects of the lexical semantics relations. The three lexical semantics relations play an important part in text tagging and retrivaling. This paper discusses the contents and principles of building Chinese machine tractable synonym dictionary and introduces the method about using machine--readable resource to build a machine tractable dictionary. This dictionary discribes the lexical meaning, distinguishes the relationships among synonym, near—synonym and relativity of lexis and also organises the near—synonym, relativity relations based on the synonym relations. The characteristics it shows in lexis collecting determine it sharable and reusable. The dictionary can be used as a basic resource in language information processing.

1　引　　言

词汇知识的缺乏是真实文本处理的瓶颈问题。在全文检索系统中，查全率、查准率是两个重要指标。当前流行的检索系统在实用中表明，查全率和查准率都不够理想[1]。影响查全率和查准率的直接原因可以有：(1)检索文本上没有标注任何有关词汇意义的信息；(2)没有一个良好的知识库作为支撑，特别是词汇知识的缺乏。经验表明，为全文检索系统建立后控词表是提高系统查全率的一种重要手段[6]，后控词表为检索系统提供了一定的词汇知识。而词表的规模及词汇知识的描述从根本上决定了检索效果的好坏。合理地组织词汇知识便可使得词表不仅可以提高检索的效果，同时可以作为文本的词汇义项标注的基础资源。

同义、近义和相关是汉语词汇意义的三种关系。这三种词汇意义关系在文本处理中起着重要的作用。本文主要介绍利用机读版本的《现代汉语同义词词典》[3](简称《现同》)、《实用汉语用法词典》[4](简称《实典》)、《同义词词林》[5](简称《词林》)建造汉语机用同义词词典

(Machine Tractable Synonym dictionary, MTSD)的技术。在该MTSD中,按词汇的同义关系组织了词汇的近义和相关关系,描述了词汇的同义、近义、相关信息,以期作为汉语文本的义项标注和全文检索的基础性资源。

2 机用同义词词典的建造依据和原则

2.1 MTSD的建造依据

研究词汇意义关系的前提条件是在同一个语言系统中。更确切地说,这意味着:(1)处于某个发展时期的语言;(2)同一种语言;(3)非方言。

为了说明词汇的同义、近义和相关关系,我们定义如下的函数和操作:

设word1、word2表示词形,函数:

mean(word1):表示word1的意义;

object(word1):表示word1的所指对象;

category(word1):表示word1的语义类属。

part－of－speech(word1):表示word1的词性。

操作: mean(word1)－mean(word2)表示词word1与word2的意义的差异。

定义1: 若word1≠word2,而object(word1)=object(word2),且part－of－speech(word1)=part－of－speech(word2),则word1与word2具有同义关系(SR)。

SR是一种等价关系。任何一个语言系统中具有SR的词的数量相对较少。

例如: 在集合{邻居 邻舍 邻人 街坊}上有SR,该集合为一组同义词。

定义2: 若word1≠word2,且object(word1)≠object(word2),而mean(word1)≈mean(word2),则word1与word2具有近义关系(HR)。

HR是一种对称的关系。为了说明近义的程度,我们采取将近义词划分级别的方法。

例如: (禁止 HR 取缔),(禁止 HR 严禁),(禁止 HR 查禁){取缔 严禁 查禁}是"禁止"的近义词集合。

定义3: 若word1≠word2,object(word1)≠object(word2),mean(word1)≠mean(word2),而category(word1)=categroy(word2),则word1与word2具有相关关系(RR)。

词的相关关系根据不同的要求,其相关的意义可以不同,我们这里考虑的相关性其意义就是属于同一个语义分类。

例如: 教师 RR 学生

2.2 MTSD的建造原则

1) 收集现代汉语中通用的同义词组,建造同义词词库(SDB);

2) 参照机读词典,为每一组同义词建立近义词和相关词,从而形成近义词词库(HDB)和相关词词库(RDB)。SDB,HDB,RDB构成了MTSD的基础词典(BMTSD),BMTSD的特点是不受任何领域限制,具有共享性和复用性。

3) 根据专用词典或专门语料建立同义词、近义词、相关词子库，从而形成专用子词典。

4) 由于HR关系的非传递性可以导致近义程度的改变，近义词分成若干等级，最多可分为五级，一般为两级，本文以两级为原则，级别表明近义的程度。

5) 属于同一语义类属的词为相关词，相关词具有层级结构。

3 机读词典简介

3.1 《现同》简介

《现同》是一部根据一定的理论原则和方法审定汉语词群的同义关系而编纂的词典，它的主要特点是：1)以现代汉语普通话为收词原则；2)严格区分同义词和近义词；3)同义词具有相同词性；4)给出了同义词的辨析。现代汉语中常见的同义词组，这部词典都收进来了，大体反映了现代汉语词汇上的同义关系面貌。这部词典未收专门学科和专门行业中的同义词，未收同义的固定语。共收录1640组同义词，4600余个词。《现同》符合MTSD的建造原则，因此我们选用该词典的机读版本作为SDB的基础资源。

《现同》的机读版本有两种形式

1) 完全和出版形式的词目音序索引相同，其结构为：

词条::=<拼音><词形><页码>

2) 在每组同义词中，按正文部分选择一个词，给出该组词的词性和释义。其结构为：

词条::=<词形><词性><释义文本>

3.2 《实典》简介

《实典》是一部含有6000余条词的词典。经过统计在该词典中约4800余条词描述了同义词，包括固定语中的同义词，约2000个同义词组。因此选用该词典作为同义词词典的补充。由于该词典没有严格区分同义词与近义词，因而凡是非固定语且与《现同》不交叉的词，我们处理为近义词。《实典》的机读版本形式为：

词条::=<拼音><词形><词性><同义词词组><反义词词组>

3.3 《词林》简介

《词林》是一部类义词典，全书收录词语54000余条。给出了汉语词汇的一种语义分类体系。分类采用层级结构，共有12大类，94中类，1428小类。小类中以同义原则划分词群，共计3925个词群。按照词典建造的原则，我们将3925个词群继续划分为11969个最小同义集。所有收录于《词林》的词均处于一个确定的最小分类中。《词林》的机读版本形式为：

词条::=<词形><义类码>

大写英文字母代表大类，小写英文字母代表中类，每两位阿拉伯数字分别代表了词所处的小类、词群及最小同义集。

4 MTSD的建造方法

MTSD是根据《现同》、《实典》及《词林》的机读版本，自动地和人机交互方式建造的。根据MTSD的建造原则，以《现同》作为SDB的基本资源。对《现同》中未收录的固定语同义词，通过《实典》和《词林》补齐。HDB是通过比较《现同》、《实典》和《词林》产生的。RDB的词汇来源是《词林》。

为了叙述方便，我们设定如下符号表示：

XT表示《现同》中所有同义词组的集合；G1表示XT的一个元素，即一组同义词；SD表示《实典》中的所有同义词组的集合；G2表示SD中的一个元素，即SD中的一组同义词；CL表示《词林》的最小同义集集合，具有相同义类码的词的集合为CL的一个元素；G3表示CL的一个元素，则G3可由一个唯一的义类码标识。

4.1 SDB的建造方法

SDB的形成以《现同》为基础，选用了《现同》的全部1640组同义词，约4600余条词。由于《现同》中不包含同义的固定语，故参照《实典》和《词林》将同义的固定语添加进来。

SDB按数据库的形式组织，由三个子库组成，分别为同义词词库、固定语同义词词库、释义文本库。

同义词词库与固定语同义词词库的库结构描述为：

元组::=<词形><同义标记><义类码><拼音>

一个元组代表一个词条，一组同义词有相同的同义标记和义类码。

释义文本的词库结构为：

元组::=<同义标记><词性><释义文本>

一个元组描述了一组同义词。

下面分别叙述这三个子库的建造方法。

1. 同义词词库

同义词词库根据《现同》的第一种机读版本，采用自动建立和机助人建的方式来形成。自动建立的过程描述如下：

1) 按页码将词分组；

2) 在同一页码下划分同义词词组；

3) 给同一组同义词确立唯一的同义标记；

4) 在《词林》中确定义类码；

5) 按词库格式写入同义词词库。

在此过程中，划分同义词词组是关键。在此我们采用如下规则划分：

设word1，word2为具有相同页码的词，C(word)表示词word的构成字集合：

rule：IF $C(word1)\cap C(word2)\neq\varphi$ THEN word1，word2属同一词组

rule的确立是根据同义词的构成特点确立的。该规则划分了1368组同义词，剩余同义词组的划分靠人机交互完成。

2. 同义的固定语词库

该词库词的来源是从《实典》和《词林》中挑选的。选择方式如下：

若 (G1∩G2)and(G1∩G3)≠φ,任取 X∈G2−(G1∩G2)或X∈G3−(G1∩G3)，

length(X)≥3

则 X 为固定语,且与 G1 中的词具有相同的同义标记。

3. 释义文本库

释义文本库的建立直接由《现同》第二种机读形式转换而来。同义标记的确定方法为：

若词形 word∈《现同》的第二机读版本,且 word∈唯一的 G1，则同义标记为 G1 的同义标记;否则 word∈G11,且 word∈G12(G11,G12∈XT),则同义标记的确定由人机交互完成。

4.2 HDB 的建造方法

词汇的近义关系是一种非传递性关系,这便决定了词汇近义的程度。根据对机读词典的分析决定 HRD 设计为二级形式：

设 HW1、HW2 分别为词 word 的近义词,且

mean(word)−mean(HW1)=δ_1, mean(word)−mean(HW2)=δ_2

而 $\delta_1<\delta_2$，则将 word 的近义词分为两个等级,HW1 为 word 的一级近义词，HW2 为 word 的二级近义词。

HDB 的结构形如：

元组::= <同义标记><近义等级标记><近义词词串><义类码>

因而,HDB 描述的是对一组同义词的近义关系。如果一组近义词没有相应的同义词，那么就给这一组近义词一个特殊的标记,该标记存放在"同义标记"字段中。

1. 一级近义词的确定

(1)、有相应同义词的近义词词组的 确定

1) 任取 G1∈XT

2) 取出相应的固定语同义词组 SG1

3) 选择 G2∈SD,选择 G3∈CL

4) 若 (G1∩G) AND (G1∩G3)≠φ 则

(G2−(G1∩G2))∪(G3−(G1∩G3)−SG

为 G1 的一级近义词。以 HM1 表式这些近义词词组构成的集合。该集合共含 1108 组近义词,约 4100 条词。

(2)、无同义词的近义词词组的确定

凡是属于以下集合元素均为近义词词组。

(SD−HM1)∪(CL−HM1)

该集合共含 10732 组近义词,约 45000 条词。

2. 二级近义词的确定

我们只考虑某一同义词组二级近义词，二级近义词的来源是《词林》中同一分类下具有相同次小类代码的词组。

任意 G31,G32∈CL,若 G31 的义类码为 $Xyn_1n_2n_{31}$,G32 的义类码为 $Xyn_1n_2n_{32}$,即

G31 与 G32 只是最小类不同，且 G1∩(G31∪G32) ≠φ，

则 (G31∪G32)－G1－ G1 的固定语同义词集 － G1 的一级近义词词集

即为 G1 的二级近义词。该集合共含 991 组近义词，约 3700 条词。

4.3 RDB 的建造方法

RDS 词的来源是《词林》。根据《词林》相关词定义为：若词 word1 的义类码为 $Xyn1_{n21}n_{31}$，word2 的义类码为 $Xyn_1n_{22}n_{32}$，word1 与 word2 相关。则与 G1 相关的词即为与 G1 的义类码中 Xyn 相同，且除去含于固定语和近义词中的词。

RDB 格式为：

元组::=＜同义标记＞＜相关词的义类码＞

凡是没有相应同义词的词群，只要在同一语义分类下，就认为是相关词。

为了使该 MTSD 能和专门学科、专门行业的词汇自然接壤，设计词典时，保留了建造专用词典的接口。同时由于近义词和相关词的开放性，在 HDB 和 RDB 可方便增删。

5 结束语

这样的一部 MTSD 可以用于文本的同义、近义标注，作为检索的知识库。在词典实现时，独立设计了词典的管理系统，系统为汉语全文检索和文本标注提供了接口。结果表明这部 MTSD 对全文检索的查全率的提高起到了较大的作用。作为一部基础性资源，即可服务于汉语文本信息处理，亦可提供给语言学家研究词汇使用。

参 考 文 献

[1] 赖茂生，王延飞，赵丹群，计算机情报检索，北京大学出版社

[2] 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

[3] 刘叔新，现代汉语同义词词典，天津人民出版社

[4] 周行健，实用用法词典，国际文化出版社

[5] 梅家驹等，同义词词林，上海辞书出版社

[6] 穗志芳，汉语全文检索中的义项标注研究，山西大学硕士论文