汉语全文检索中的义项标注技术研究

刘 开 瑛
（山西大学计算机科学系，030006）

摘要：本文主要介绍现代汉语文本语义标注中的几个基本问题，包括语义标注的几个约定，从同义词标注入手的一个实验实例。提出了同义词语义分类体系的构造、选词范围和语词属性描述方式等。

Research on Contemporary Chinese Full Texts Retrieval Word sense tagging Techniques

Liu Kaiying
(Shanxi university, 030006)

Abstract: In this article the auther mainly introduces some fundamental problems in sense tagging in contemporary Chinese full texts retrivcal. Including a few agreements of sense tagging, from an experimental example starting with mark of synonym . The auther advances the constraction, the range of selecting words and the way of describing words' attributes of the semantic classified system in synonym.

现代汉语文本的语义标注对自然语言处理、机器翻译、情报检索等领域都有重要意义。研究目标不同，语义标注的深度不尽相同。在研制汉语全文检索软件中，汉 语文本的语义标注更有其现实意义，直接关系到检索软件的检索效用(Effectiveness)的评价，即查全率和查准率两个指标的确定。本文主要谈谈汉语文本中语义标注的几点认识和体会。

一 汉语全文检索中语义标注的几个约定

1. 汉语文本中的分词原则是以语词为单位，包括词和固定词组。这些语词一般以语文语词为单位，即以语义或语法功能为基本单位。概念是构成人类思维的基本单元，语词是概念的载体。同一个概念可以有不同形式的语词来表示，而同一个语词 形式也可以表达不同的概念。所以从全文检索系统的目的来说，应该以概念为检索单位。情报检索是采用后控词表来补足语词查全率检索问题的。如果能对被检索的汉语文本事先进行语义概念标注，将给汉语文本主题概念研究和查全率的提高奠定良好的基础，检索效用会更高的。

2. 研制现代汉语语词的语义分类体系(也称义类体系) 及其符号系统是语义标注首先要解决的课题。语义分类就是语词分类，在这个语义分类体系中，表达每一个概念的语词只能出现在一个义项的词群中，即每个语词按其义项定义出类别，有几个义项就应有几个不同

* 国家自然科学基金赞助项目，批准号 69373041

的义项类；表达同一概念的不同语词符号，一律归纳入同一个义项类中。研制汉语义类体系，没有现成的计算机可用的资料可以依据。传统的“雅学”、英国的《洛氏分类语汇》(Thesaurus)、日本的《分类词汇表》等可资借见参考，语言学界也都不能照搬照用。八十年代以来出版的《同义词词林》《简明汉语义类词典》《现代汉语同义词词典》等都是一部部汉语类义词典，收词丰富，内容充实，词语分类都各自成体系，是极有价值的参考书。然而这类词典都是为人们的写作和翻译而编写的，不能作为机读词典来使用。所以建立汉语义类体系及其符号系统从理论到具体原则，从收词范围到分类层次，都有一系列的技术性问题有待解决。我们希望汉语语义分类体系具有分类界限分明，可操作性强，能严格避免交叉分类的纠缠，交叉现象的出现，能够满足机读标注的要求。这个任务需要逐步解决。

二　从同义词标注入手的一个试验实例

从宏观上说，语词语义分类是语义学科总结的对客体的认识成果，是人类对客体认识的抽象性概括性的总结，因此它应具有抽象性、概括性和普遍性。从结构上说，汉语语词语义分类具有层次、网络性和开放性的特色。当前语义分类的分析没有统一标准，不够系统，不够科学，带有很大的主观随意性。所以在全文检索系统中，建造机读通用的汉语词语义类标注技术也是不现实的，条件还极不成熟。而应首先建造机读同义词词库，对汉语文本进行同义词标注，实践证明这个方法是行之有效的。为此，我们对《同义词词林》进行了改造成机读词典的试验。

《同义词词林》共收录语词七万余条，全部按意义进行编排，它是一部汉语类义词典。该书的分类原则：以词义为主，兼顾词类，并充分注意题材的集中。全书把词语分为大、中、小类三级，共分 12 个大类，94 个中类，1428 个小类，小类下再以同义原则划分词群，每一词群以一标题词立目，共 3925 个标题词。从原则上讲，《同义词词林》所收的每一个词义都可循着由大类到中类到小类这样一条线索找到其应该在的最具体的词群中。与此分类体系相对应的是一个语词分类的编码体系，其编码体系用大写拉丁字母表大类，小写拉丁字母表中类，两位阿拉伯数字表小类，将其编码形式描述如下：

＜编码＞::=＜大类＞＜中类＞＜小类＞

＜大类＞::=＜大写拉丁字母＞

＜中类＞::=＜小写拉丁字母＞

＜小类＞::=＜数字＞＜数字＞

例如：　Bm08 煤 炭 (B 表示大类“物”，m 表示中类“材料”，08 表示小类排号)

煤　煤斤　煤炭　乌金　乌金墨玉
无烟煤　硬煤　红煤　白煤　大砟
焦煤　主焦煤
煤砖　煤饼　蜂窝煤
褐煤　褐炭

} 标题词“煤”的同义词群

* * 原(元)煤　肥煤　瘦煤　烟煤　气煤　乏煤　煤精
煤球　煤砟子　煤核儿

} “煤”的同类词

炭　木炭
草炭　草煤
泥炭　泥煤
焦炭　焦(炼～)　} 标题词“炭”的同义词群

* * 活性炭　骨炭　无定形炭　火炭
煤焦　炭墼　炭精　} “炭”的同类词

这个编码体系比较适合机读语义标记符号体系的建造，为此我们进行了如下加工。在《同义词词林》第三级分类下，仍然隐含着两级分类。首先是以其所列的标题词将第三级类下所收词分成若干词群，在此基础上又按词的修辞色彩和使用范围等方面的差异，分段排列，各个词群之间以空行间隔。

因此，我们在原有的分类体系基础上增加第四级分类，将第三级分类下所列的以标题词开头的各同义词词群分开。对编码的改进是在其原有四位编码的末尾追加两位阿拉伯数字表第四级分类。

经过以上加工后，例如，在“Bm08 煤 炭”下第四级分类为“Bm0801”“Bm0802”分别表示以“煤”“炭”为标题词的同义词词群。然而，在第四级分类下的各段同义词的词义之间仍然具有细微的差别，还不能作为检索用同义词。如在“Bm0801”分类下，“煤”与“无烟煤”“煤砖”“褐煤”之间仍然存在一定的词义差别。因此，在第四级分类的基础上对其进一步分类，即将词群中各段同义词分开，为此增设第五级分类。在其原有的六位编码的末尾追加两位阿拉伯数字表第五类。

加工后的编码形式如下：

<编码>::=<大类><中类><小类><四级类><五级类>

这样，在“Bm0801”下用“Bm080101”表“煤 煤斤 煤炭 乌金 乌金墨玉”这一组同义词；用“Bm080102”表“无烟煤 硬煤 红煤 白煤 大砟”另一组同义词，如此等等。

经过加工的《同义词词林》作为现代汉语文本语义分类标注的机读工具，实用中还存在着多个极待解决的问题。

1. 该词典共收语词七万余条，但大量常用语词没有收入，如：爱护、环保、拼搏、换代、窥测等。据人民日报 1990 年 6 月 5 日、6 日、8 日、10 日四天 32 版全部语料统计，总词数 14800 条，其中常用词共 11730 条，同义词林中包括了 7482 条，未收入的常用词 4248 条，占常用词总数的 36%。大量未收入的词语进行分类定位工作量是很大的。为此我们研制了人机交互软件，进行新词添入的处理。如新词“乐队”查询顺序为抽象事物(D)⇒机构(Dm)⇒文体机构(Dm07)⇒Dm0717，最后得出“乐队”的编码。

2. 该词典没有专有名词的分类标记，大量的人名、地名、机关名、事件名以及标牌名等需要补充。据《人民日报》四天语料统计共有不同的人名 891 条，地名 749 条，机关组织名 512 条，共计 2152 条。占总词数 14800 条的 14.5%。我们建立了人名库、地名库、机构组织名库及事件库。

3. 该词典收入了部分成语、俗语、方言词、古语词、冷僻词等，这些词语只能适用于某些汉语文章，在一般的文章中出现次数甚少，这些词语的引入造成了同义词群的混乱。如古语词以单字词为主，几乎是字字多义。我们在加工《同义词词林》时，单另建立了古语词库和偏

僻词词库，收集了这部分词语。如“Bm08”中我们将“大砟 Zha”“炭墼 Ji”等归入偏僻词库。在“太原市地方志大事记”全文检索系统中，由于收录了从先秦（公元前 514 年）到 1990 年长达 2504 年的历史，原文中 1911 年以前的记载，全部采用古文汉语叙述，从朝代名、官职名、姓氏名直至古文字的使用都需要古词语词典进行标记处理。据初步统计，占这部分材料总词数的 20%。

4. 同一个词群中，将古语词、方言词、冷僻词抽出后，有的同义词、近义词夹杂，有的词义之间的特殊意味、色彩、词义轻重以及语词搭配关系等仍有明显关系。为此我们需进再分类，并进行第“五级类”标记重新排序。同理，对同类词（划 * * 号者）亦一一列出，进行同样处理。例如“Ai0501 模范”重新标记后结果如下：

Ai050101　模范 标兵 表章 榜样
Ai050102　师表（为人～）
Ai050103　劳模 劳动模范 劳动英雄
Ai050104　先进工作者 先进生产者
Ai050105　冠军
Ai050106　亚军
Ai050107　季军
Ai050108　殿军（后四个语词是由同类词列出的）

三　研制现代汉语同义词标记体系的设想

1. 机读现代汉语同义词词典的收词范围和原则。

① 该词典收集（从“五四”前后直到现在）汉语现代时期和普通话范围内的语词，不包括古语词和方言词。

② 同义词是指意义上相同的等义词和意义基本一致指同样事物对象而材料构造上却不相同的词。排除词义相近但并不指同一事物对象的近义词，即意义在外延和内涵上都不相同的语词。如“繁荣”和“繁华”，“观察”和“视察”都是近义词。

③ 语词与语词之间只能在某一意义上有同义关系，所以同一同义词组只是就一个意义而构成，同一个多义词会出现在不同的同义词组里。

④语词的词性对词义理解有一定的作用，词典中在同一同义词组的语词应属于同一词类。

⑤ 该词典只收常用或较常用的对检索有价值的同义词组。专有名词（如人名、地名、机关组织名等）、专门学科和行业的同义词组一概不收，另行建库。

《现代汉语同义词词典》基本上具备了以上条件，在审定同义词词群中同义关系上，已形成一定的理论原则和操作方法，我们认为它可以作为建造通用的同义词词典选词的基础。特别是在同一个同义词组中，列出其共性的同时，分别揭示了每个词的“个性”或特点（包括特殊意味、色彩、词义轻重和搭配关系等特点）。给建造每个语词的属性描述，奠定了良好基础。

2. 当前主要有语义分类树和属性描述的方法来建造语义词典，这两种方法我们都进行了比较和应用。我们认为，采用这两种方法相结合的办法更为有效。

结构主义语义学者用语义场来研究词与词之间的同义关系、下义关系（hyponymy）、反

义关系。其中下义关系反映了词汇的层级结构(hierarchical structure),这种层级结构可以用树结构来表示。因而由词汇的分类而构成的层级结构可以称为语义分类树。这种语义分类树由于它有属性的继承性、关系的可传递性和结点的唯一性等性质,所以使其具有各结点的属性描述简洁,避免了冗余信息,并可揭示概念之间的蕴含关系,便于知识推理等优点。但是由于概念体系本身的复杂性和词语的共性和个性至使语义分类树建造过程中遇到了不可避免的困难,建立一个总的汉语语义分类树是不可能的。而对于某部分词语群,层次结构清楚,语词词汇是一个封闭性的聚合结构,可以组成一个独立的语义分类树。如植物分类体系,动物分类体系等。

对于不能构造语义分类树的词语群,可以采用"属性描述"的方法。属性描述就是采用一组"属性——值"对来描述概念的内涵。如"金融"是一个概念,它的内涵包括以下功能:(1)指货币的发行、流通和回笼;(2)货款的发放和收回;(3)存款的存入和提取;(4) 汇兑的往来等经济活动。(现代汉语词典 P578)。它的外延各级各类银行、储蓄所、信用社等办事机构。用类 Lisp 形式:

```
(金融 (读音 (jinrong))
      (词性(名词))
      (义项 1 (指货币的发行、流通和回笼))
      (义项 2 (货款的发放和收回))
      (义项 3 (存款的存入和提取))
      (义项 4 (汇兑的往来等经济活动)))
```

又如"保险"是兼类词,其表达形式如下:

```
(保险 (读音 (baoxian))
      ((词性 (名词))
      (义项 1 (受损后得到赔偿的办法))
      (义项 2 (机械的安全装置))
      (同义词 (保险)))
      ((词性 (动词))
      (同义词 (保证, 担保)))
      ((词性 (副词))
      (同义词 (必定, 一定)))
      ((词性 (形容词))
      (同义词 (可靠, 牢靠, 把稳))))
```

和语义分类树相比,属性描述的方法有如下优点:

(1) 如对各个入构项分别描写,可操作性强,属性描述的方法更便于工程实施。

(2) 可以使语义信息更加详备,可以避免分类的纠缠,集中精力对每个概念进行刻划,有什么信息就记录什么信息,而不必考虑其他概念。

(3) 描写的深度可深可浅,比较灵活。不同类型的系统对语义理解的深度是不同的。有的系统为了效率上的考虑可能要求语义信息少一些,这样只要少一些属性就可以了。

3. 同义词词语的分类体系及其标记符号,由以下几部分组成:

(1) 语法研究中划分词类的目的是把语法性质相同或者相近的语词归在一类。根据语

词的语法功能来确其词类。由于汉语词类和句法成分间有错综复杂的关系，为了语义分类的方便，所以在语义标记符号串的首位增加一位表示词性的记号。其词性分类和标记(大写拉丁字母)为：

名　词　N　　动　词　V　　形容词　A　　数量词　M

时间词　T　　处所词　S　　代　词　P　　虚　词　F

(2)语义分类体系基本上可采用改建后的《同义词词林》的系统及其标记符号。具体的归类方式简单的描述如下：

名词类(N)：　人(A)　物(B)　抽象事物(D)

动词类(V)：　动作(F)　心理活动(G)　活动(H)　现象与状态(I)　关联(J)

形容词(A)：　特征(E)

时间词(T)：　时间词群(Ca)

处所词(S)：　空间词群(Cb)

代　词(P)：　人(Aa)　物(Ba10)　空间(Cb30)　特征(Ed61)

数量词(M)：　数量(Dn)

虚　词(F)：　副词、介词、连词、助词和叹词

例：满意⇒动词(V)⇒心理活动(G)⇒心理状态(Ga)⇒VGa060101

福气⇒名词(N)⇒抽象事物(D)⇒情况(Da)⇒NDa090101

繁荣⇒形容词(A)⇒特征(E)⇒境况(Ef)⇒AEf020101

储蓄 {⇒名词(N)⇒抽象事物(D)⇒经济(Dj)⇒NDj080213
{⇒动词(V)⇒活动(H)⇒生活(Hj)⇒VHj40030!

(3)专业词库、行业词库、人名库、地名库、机关组织名库等需根据汉语文本语料的用词范围来建造。这类词库亦有同义词组审定问题。我们在政府办公文档库、新闻语料库和太原地方志大事记库等都建立了相应的专门词库。实践证明，效果较好。

以上设想都正在实验中，有待进一步改进。

主要参考文献

1. 梅家驹等，《同义词词林》，上海辞书出版社，1983 年

2. 林杏光等，《简明汉语义类词典》，商务印书馆，1987 年

3. 刘叔新主编，《现代汉语同义词词典》，天津人民出版社，1993 年

4. 董振东，机器翻译中词典和文法的关系，《中文信息学报》，Vol. 2，No. 3

5. 孙宏林，语义分类与属性描述，1991 年计算语言学会议论文

6. 孙维张，论语义范畴系统的建构，1989 年全国自然语言理解专题讨论会论文