# 面向综合语言知识库的知识融合与获取研究

孙 超，张仰森
（北京信息科技大学 智能信息处理研究所，北京 100192）

**摘 要**：针对如何填补语料库和电子词典的数据结构之间的差异，如何将语料库与电子词典融合到综合语言知识库系统中，并进行多语言知识资源之间的交叉参考等问题，提出并实现了一种便捷的语言知识查阅方法。该方法以语料库为基础，利用鼠标左键点击完成从电子词典中获取相关知识的操作。通过在北大语料检索工具上的实验，使得用户在浏览语料的同时即可获取电子词典中相应的词汇知识，实现了语料库和词典间便捷、准确的对应和参照，体现了此方法的优势。

**关键词**：语料库；电子词典；自然语言处理；知识获取方法
**中图分类号**：TP391 **文献标识码**：A **文章编号**：1673－629X(2010)08－0025－04

# Research of Knowledge Integration and Obtaining Oriented Comprehensive Language Knowledge System

SUN Chao, ZHANG Yang-sen
(Institute of Intelligent Information Processing, Beijing Information Science & Technology University, Beijing 100192, China)

**Abstract**: To solve the problem of filling the "gap" of electronic dictionaries and corpora, the problem of integrating corpora and electronic dictionaries into comprehensive language knowledge system, and the problem of cross－reference among the various data resources, a convenient method for obtaining language knowledge is proposed. On the basis of the corpora, it obtains knowledge form electronic dictionaries with the left mouse button clicking. By experiment on the corpora retrieve system, this method supports convenient and accurate cross－reference of different databases and enable the users to obtain rich language knowledge from various dictionaries when they are viewing the corpora, and this method represents the advantage of itself.

**Key words**: corpus; electronic dictionary; natural language processing; method of getting knowledge

## 0 引 言

随着计算机技术的发展，自然语言处理系统迈入了实用化进程，其中电子词典和语料库成为人们开发的焦点，它们的规模和质量很大程度上决定了自然语言处理系统的成败。北京大学计算语言学研究所一直致力于语言资源的开发和建设，积累了一系列语言资源。目前，正在努力把这些语言数据资源集成到一起建成综合型语言知识库系统。首要目标是实现一个用于知识获取的交叉参照模块，支持组成综合型语言知识库的各个成分数据资源之间便捷的、准确的交叉参照，方便用户（包括人和机器）从数据结构各不相同的多种语言资源获取丰富的语言知识。文中描述了这一模块的设计和实现情况，该模块已经嵌入到综合型语言知识库原型系统中，为用户获取不同层面的词汇知识提供了便利。

**收稿日期**：2009－12－21；**修回日期**：2010－03－02
**基金项目**：国家自然科学基金项目(60873013)；北京市自然科学基金B类重点项目(KZ200811232019)；北京市属市管高校人才强教计划项目(PXM2008_014215_055942)
**作者简介**：孙 超(1982－)，男，硕士研究生，研究方向为智能信息处理；张仰森，教授，硕士生导师，研究方向为人工智能、中文信息处理。

## 1 语言知识资源的可融合性分析

综合语言知识库系统包括了很多模块。目前，其原型系统主要涉及以下语言知识资源：

(1)现代汉语语法信息词典(简称《语法信息词典》或GKB)

(2)现代汉语基本标注语料库(简称"基本标注语料库"或STC)

(3)面向汉英机器翻译的现代汉语语义词典(简称"汉语语义词典"或CSD)

以上三个语言知识资源的知识结构存在着"缝隙"，这些"缝隙"成为语言知识库之间进行相互融合的

障碍。可融合性分析的目的是为了找到知识结构之间的“缝隙”和填充“缝隙”的方法，使得语言知识资源能够很好地融合。下面分别对它们的知识结构进行研究和分析。

### 1.1 《语法信息词典》知识结构

《现代汉语语法信息词典》(Grammatical Knowledge Base，GKB)是语言知识库大厦的第一块基石，它是一部面向语言信息处理的大型电子词典，收录 8 万词语，在依据语法功能(优势)分布完成词语的分类的基础上，又按类描述每个词语的详细的语法属性[1]。GKB 中涉及 26 个词类标记，每一类词都分别存贮在数据库的一个表中，此外还有一个总库。其样例如表 1 所示，其中“词语、词类、同形”三个字段构成数据表的关键字。

表 1 《语法信息词典》数据库样例

| 词语 | 词类 | 同形 | 拼音 | 注 | … |
|---|---|---|---|---|---|
| 空 | a | A | kong1 | 没有东西，没有内容 | |
| 空 | a | B | Kong4 | 空缺的；没有被利用的 | |
| 抱 | v | A1 | Bao4 | 用手臂围住 | |
| 抱 | v | A3 | Bai4 | 心里存在着想法，意见 | |
| 报告 | n | | bao4gao4 | 书面文件 | |
| 报告 | v | | bao4gao4 | 发表讲话 | |

### 1.2 《现代汉语语义词典》知识结构

《现代汉语语义词典》(Chinese Semantic Dictionary，CSD)是一个面向机器翻译的大规模汉语语义知识库，收录 6.6 万余实词，详细描述每个词语的义项、语义类以及基于配价理论的语义搭配限制，可为计算机语义自动分析、词义消歧等任务提供强有力的支持[2]。其样例如表 2 所示，“词语、词类、同形、义项”四个字段构成了关键字，“主体、配价数、客体”三字段限制了动词对语义的选择。

表 2 《现代汉语语义词典》样例

| 词语 | 词类 | 同形 | 拼音 | 义项 | 备注 | 主体 | 配价数 | 客体 |
|---|---|---|---|---|---|---|---|---|
| 走 | v | 1 | zou3 | 走路 | 屋里～出一个小孩 | 人类\|动物 | 1 | |
| 走 | v | 1 | zou3 | 移动 | ～社会主义的道路 | 人类 | 2 | “*道路”\|“*路线” |
| 走 | v | 1 | zou3 | | 钟不～了 | “*钟”\|“*表” | 1 | |
| 走 | v | 2 | zou3 | 离开 | 客人～了 | 人类 | 1 | |

CSD 是为中文信息处理而建立的最基本的语义知识库，目的是在语法分析的基础上，为计算机自动分析汉语句子和生成英语句子提供更深入的语义信息。

### 1.3 《现代汉语基本标注语料库》知识结构

《现代汉语基本标注语料库》(word - Sense Tagging Corpus，STC)是对《人民日报》语料进行切分和词性标注的成果。目前，《人民日报》的加工规范有两套，分别是《规范 2001》[3]和《规范 2003》[4]①，前者是粗粒度标注的标准，有 40 个标记集，后者是细粒度标注的标准，扩充到 105 个标记集，下面分别是两套标记集的加工语料样例。

《规范 2001》语料例子如下：

19980101 - 01 - 002 - 002/m 我们/r 即将/d 以/p 丰收/v 的/u 喜悦/an 送/v 走/v 牛年/t ，/w 以/p 昂扬/a 的/u 斗志/n 迎来/v 虎年/t 。/w

《规范 2003》语料例子如下：

20001101 - 05 - 014 - 004/m 中国/ns 男队/n 与/p 乌兹别克斯坦队/nt 的/ud 比赛/vn 是/vl 一/m 场{chang3}/qv 硬仗/n 。/wj

### 1.4 知识结构的缝隙与融合

通过分析和研究，发现知识结构的“缝隙”主要表现在两个方面：切分单位的不同、词性标记不同和信息表达方式不同[5]。第一，基本标注语料库的词性标记中有 13 个与《语法信息词典》的词类标记存在多对一的关系。为了实现上述知识资源之间便捷的、准确的交叉参照，那么必须填补资源之间的“缝隙”。针对词性标记的差异，设计了对照表，实现语料库标记到词典词类的映射，以《规范 2001》[3]为例，如表 3 所示。

第二，词典的信息是显性表示的，是确定的，语料库的信息是隐性表示的，是不确定的。词典中用词语、词性、同形以及义项编码来做关键字，然而语料中只有词语和词性标记。因此，要为语料库中的词语增加同形和义项信息。以“除/v”为例来说明。

“除/v”在 GKB 中有两个同形，同形 1 表示“除法”，同形 2 表示“去掉”。在语料中标注了同形，计算机就可以准确地将其映射到词典中的一条记录。

定义：(切分单位集合) $S\{s1,s2,\cdots,sn\}$；(语法知识集合) $G\{g1,g2,\cdots,gn\}$；(语义知识集合) $C\{c1,c2,\cdots,cn\}$；(同形集合) $T\{t1,t2,\cdots,tn\}$；(义项集合) $Y\{y1,y2,\cdots,yn\}$；(词语集合) $W(w1,w2,\cdots,wn)$；其中 $n >= 1$；(综合知识) Knowledge。

融合过程为：

$S + T -> G, S + T + Y -> C, (S) =$ new $S$，(新切分单位集合)$(S)\{(s1),(s2),\cdots(sn)\}$，

Knowledge $= (S) + G + C$

通过以上分析，设计语料库与词典的词性映射表、对语料库中的词语增加同形和义项信息可以减少资源“缝隙”，增加语言知识资源之间的融合度，逐渐把三个

①《规范 2001》是指发表在《中文信息学报》上的“北京大学现代汉语语料库基本加工规范”；《规范 2003》是指发表在《汉语语言与计算学报》上的“北大语料库加工规范：切分·词性标注·注音”。

资源融合在一起。

表 3　语料库和词典的词性对照表

| 词性 | 词典标记 | 语料库标记 |
|---|---|---|
| 名词 | n | n ns nt nz |
| 形容词 | a | a an ad |
| 方位词 | f | f |
| 代词 | r | r |
| 状态词 | z | z |
| 介词 | p | p |
| 语气词 | y | y |
| 前接成分 | h | h |
| 非语素字 | x | x |
| 简称略语 | j | j |
| 时间词 | t | t |
| 数词 | m | m |
| 动词 | v | v vn vd |
| 区别词 | b | b |
| 连词 | c | c |
| 拟声词 | o | o |
| 后接成分 | k | k |
| 成语 | i | i |
| 处所词 | s | s |
| 量词 | q | q |
| 副词 | d | d |
| 助词 | u | u |
| 叹词 | e | e |
| 语素 | g | Vg Ag Ng Dg Tg |
| 习用语 | l | l |
| 标点符号 | w | w |

## 2　知识获取与系统模块设计

对知识资源实施融合技术之后,就可以从它们中获取相关知识。获取知识的切入点是标注语料库,通过语料库的切分单位(标注词语)来获取词语的语法知识和语义知识,以实现语言知识资源之间的交叉参考。

获取第 $k$ 个切分单位的融合知识的过程可以这样描述:

$(wk, tk, gk) = \mathrm{Get}((sk))$,

$\mathrm{Knowledge} = \mathrm{Get}((wk, tk, gk), G, C)$。

按照上述描述,知识获取流程如图 1 所示。

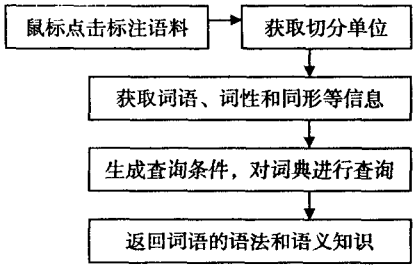


图 1　知识获取流程图

根据知识获取流程图设计系统功能模块。系统功能模块设计主要分为三部分:(一)获取切分单位;(二)将切分单位信息映射到词典信息;(三)进行查询。

第一、获取切分单位。

系统采用鼠标取词的方法,无需用户输入词条。以语料库切分单位的分隔符为标识,从鼠标光标的左右开始截取字符串,到分隔符为止。

第二、将切分单位信息映射到词典信息。

根据词语标注的映射关系,本模块对获取到的切分单位进行如下分析:(1)取词语;(2)取词性标记;(3)取同形标记;(4)取义项标记。

以"活动/vn! 2-1"为例进行说明:

(1)取词语,在"/"之前,得到"活动"。

(2)取词性标记,在"/"和"!"之间的,得到 vn。根据词性对照表,映射到动词标记 v。

(3)取同形标记,在"!"和"-"之间的,得到 2。

(4)取义项标记,在"-"到空格之间,得到 1。

第三、查询和获取知识。

根据获取的切分单位信息生成查询条件,去数据库中查询,获取融合知识。

在查询 CSD 和 GKB 的时候,由于两个数据库的数据结构的不同,对其查询的过程也是不同的。CSD 库是没有层级结构的,所以只需根据查询条件到相应的数据表中进行查询,即可返回结果。GKB 是有层级结构的,它的层级包括总库和一级分库和二级分库,其中动词和代词分别有二级分库。对于二级分库的查询需要用到一级库的某些属性条件。以"做/v! 2"为例子说明如何查询分库,流程如图 2 所示。

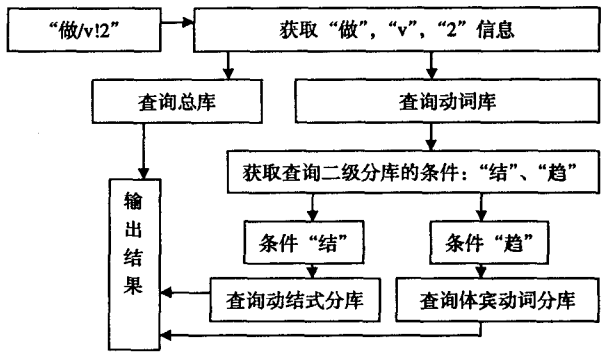


图 2　查询二级分库示例流程图

## 3　实验结果及结束语

将用于知识获取的交叉查询系统嵌入到北京大学标注语料检索工具中,运行如图 3 所示,演示了鼠标点击切分单位"帮助/v",在右侧显示从 GKB 库和 CSD 库返回的结果,分别显示在现代汉语语法信息词典属性信息栏和现代汉语语义词典属性信息栏。

文中通过对自然语言处理中的基础资源的分析,设计了交叉查询参考系统,它可以从各种角度获取不同层次的词汇信息,为语言学家研究语言本体提供便


图 3 语料库与词典的交叉参照图

利的支持。

切分单位在查询中是孤立的,今后努力的主要方向是在当前切分单位的基础上获取其相关的搭配单位。例如"苹果/n",希望能得到其左边的"很/d"、"大/a"和"的/ud"等修饰信息[6~11]。此外,考虑到切分单位不等同于词典词,因此如果获取不到相应切分单位的属性信息,可以对切分单位进行处理,例如"企业经营者/n",可以分析为"企业/n"、"经营/n"、"者/k",从词典获取各个部分的属性返回。这样可以使得本系统模块得到更大的改进,更好地服务于汉语语言学本体研究和教学研究。

**参考文献:**

[1] 俞士汶,朱学锋,王 惠,等. 现代汉语语法信息词典详解[M]. 第 2 版. 北京:清华大学出版社, 2003.

[2] 王 惠,詹卫东,俞士汶. 现代汉语语义词典规格说明书[J]. 汉语语言与计算学报, 2003,13 (2) :159 - 176.

[3] 俞士汶,段慧明,朱学锋,等. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报, 2002,16(5):49 - 64、(6):58 - 65.

[4] 俞士汶,段慧明,朱学锋,等. 北大语料库加工规范:切分·词性标注·注音[J]. 汉语语言与计算学报, 2003, 13 (2) 121 - 158.

[5] 俞士汶,段慧明,朱学锋,等. 综合型语言知识库的建设与利用[J]. 中文信息学报,2004,18(5):1 - 10.

[6] 陈素萍,谢丽聪. 一种文本特征选择方法的研究[J]. 计算机技术与发展, 2009,19(2):112 - 115.

[7] Suarez A, Palomar M. A Maximum Entropy - based WSD System[J]. COLING,2002,2:960 - 966.

[8] 张燕平,徐庆鹏,苏守宝,等. 一种基于贪婪覆盖的文本分类方法[J]. 计算机技术与发展, 2009,19(1):74 - 76.

[9] 刘 琼,李宝敏. 一种果品领域本体库的构建方法[J]. 计算机技术与发展, 2009,19(1):197 - 199.

[10] Nivre J. MaltParser: A Language - independent System for Data - driven Dependency Parsing[J]. Natural Language Engineering, 2007,13(2): 95 - 135.

[11] Otero P. Learning Bilingual Lexicons from Comparable English and Spanish Corpora[C] // Proceedings of MT Summit XI. [s.l.]:[s.n.],2007:191 - 198.

(上接第 24 页)

[7] 尚冬娟,郝克刚,葛 玮,等. 软件测试中的测试用例及复用研究[J]. 计算机技术与发展,2006,16(1):69 - 72.

[8] 马瑞芳. 计算机软件测试方法的研究[J]. 小型微型计算机系统,2001,24(12):2211 - 2213.

[9] RTCA/DO - 178B. Software Consideration in Airborne Systems and Equipment Certification[M]. USA: Radio Technical Commission for Aeronautics, Inc, 1992.

[10] 张克东,庄燕滨. 软件工程与软件测试自动化教程[M]. 北京:电子工业出版社,2002.

[11] 周章慧,王同洋. 智能卡操作系统自动测试中的脚本技术[J]. 计算机工程与设计,2008,29(8):2068 - 2071.

[12] 蒋 云,赵佳宝. 自动化测试脚本自动生成技术的研究[J]. 计算机技术与发展,2007,17(7):4 - 7.