现代汉语述语形容词机器词典的研究与实现*

尹一瓴 陈群秀

清华大学计算机科学与技术系 智能技术与系统国家重点实验室，北京 100084

Email: cqx@s1000e.cs.tsinghua.edu.cn

摘 要： 现代汉语语义知识库是自然语言处理过程中汉语语义资源的重要工程之一。目前，已经完成了现代汉语语义知识库的三个组成部分：《现代汉语述语动词机器词典》、《现代汉语名词槽关系机器词典》和《现代汉语语义分类系统》，在此基础上本文描述了现代汉语语义知识库另一重要组成部分《现代汉语述语形容词机器词典》的研究与实现。

关键字： 现代汉语述语形容词机器词典，论旨属性，论元属性，形容词类型，语言工程，计算词典学

The Research & Implementation of the machine tractable dictionary of contemporary Chinese predicate adjectives

Yin Yiling Chen Qunxiu

The state Key Laboratory of Intelligent Technology and System

Department of Computer Science and Technology, Tsinghua University, Beijing 100084

Email: cqx@s1000e.cs.tsinghua.edu.cn

Abstract: The Chinese semantic knowledge base system is a very important part of language knowledge engineering of Chinese semantic resource in natural language processing. Now three parts of The Chinese semantic knowledge base system: the machine tractable dictionary of contemporary Chinese predicate verbs、the system of relations of slots centering noun for contemporary Chinese 、Chinese semantic classification system have been finished, This thesis presents the research and implementation of the machine tractable dictionary of contemporary Chinese predicate adjectives which is another important part of The Chinese semantic knowledge base system.

Key words: the machine tractable dictionary of contemporary Chinese predicate adjectives, thematic properties, argument properties, typology of adjectives, language engineering, computational lexicology

一. 前言

自然语言理解(NLU，Natural Language Understanding)是人工智能学科中一个重要方向，也是知识信息处理中的核心课题。从计算机科学的角度看，自然语言理解的任务是建立一种

*本文承国家社科“九五”重大项目《信息处理用现代汉语词汇研究》（批准号 97@YY001）、国家 973 重点基础研究发展规划项目（项目号 G1998030507-2）和国家 863 高科技项目（项目号 2001AA114040）资助。

计算模型，这种计算模型能够象人那样理解自然语言。为了实现语言信息处理的种种功能，人们研究自然语言的词法分析、句法分析、语义分析、语境分析等技术，并开发诸如电子词典、知识库、语料库等语言数据资源。近年来，随着计算机科学技术的发展以及信息高速公路带来的信息化，自然语言处理更加重视语义的研究及大规模语义知识库的建设。

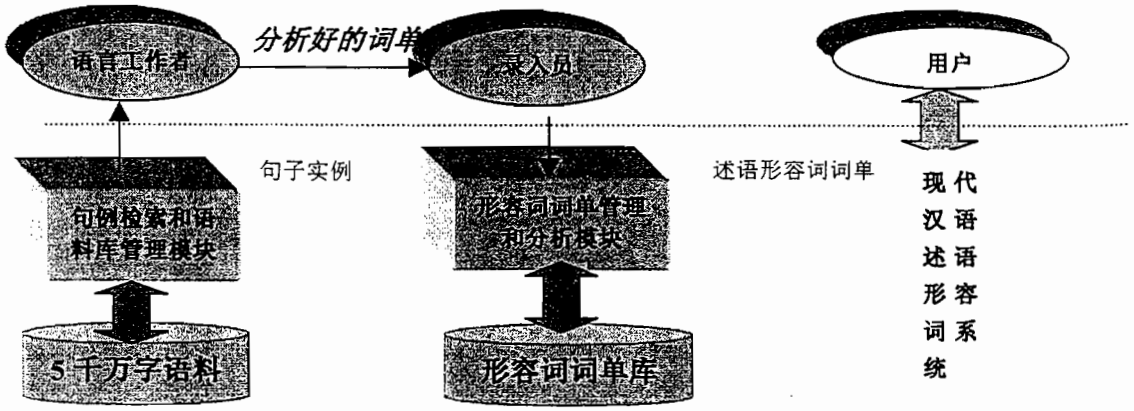
清华大学陈群秀等和中国人民大学林杏光等人研制了现代汉语语义知识库的两个重要组成部分：现代汉语述语动词机器词典和现代汉语名词槽关系词典，现代汉语述语动词机器词典的理论依据是：原则参数语法、格语法、配价语法和研制者自己的一些想法，描写了一批汉语常用动词义项的论旨网格（论元属性、论旨属性，例如基本语义句式及其变换式），描写了每个论旨角色的语义类、语类、句法功能，同时描写了动词的分类、否定式、时态、语义指向动词的动词补语、形容词补语，趋向动词补语等语法、语义、语用信息。现代汉语名词槽关系词典描写了一批以汉语常用名词为中心的槽关系即论旨角色内部的语义组合关系，即多项式定语与名词中心词之间的语义关系。这两个词典已经收录了 3976 个常用的汉语动词的 5199 个义项以及 2466 个常用汉语名词的 3046 个义项有关词形、拼音、词性等众多语义信息。而且由于是采用传统词典学和计算词典学相结合方法在大量客观的语料（5000 万语料）基础上研制的，故而词条描写内容丰富详尽、角度新颖。

与此同时，在汉语中，述语形容词具有与述语动词同等的重要地位，两者共同构成了句子的骨架，是整个句子的核心，别的成分都受它制约，被它吸收，句子的表层结构可以看作是由述语动词、形容词的论旨网格投射而成。因此，对现代汉语述语形容词的系统研究是非常必要也是非常重要的。

本文就在上述的基础之上，研究了现代汉语述语形容词机器词典的描述内容、描述形式和填写规范，设计、实现了述语形容词词典系统的软件支撑环境，用于建立述语形容词机器词典和分析述语形容词的具体特性。我们的研究填补了现代汉语述语形容词机器词典的空白，以大规模客观语料（5000 万语料库）为研究基础，使得我们的述语形容词机器词典信息丰富、角度新颖，并且包含词法、句法、语义信息和部分语用信息，采用了工程性的描述，是一个规模较大的语言工程，在国内外还是首次。

二. 述语形容词系统的研究方法

以前语言学家研究述语形容词的方法是手工方法，对其关系的描述是语法性(语类)的。而我们的研究方法不是就述语形容词本身研究形容词，而是研究组合框架中的述语形容词和有关名词性成分的语义相互制约关系。并且我们的研究方法也大不相同，是“语料库方法+联想”，即把主要从机贮语料库中获取的大量例证为依据的计算词典学方法同主要以语言工作者的经验、语感为依据的传统词典学方法结合起来，以使述语形容词的研究真正建立在丰



富客观的语言事实基础上，同时发挥语言工作者的经验、语感进行联想和归纳，即我们的研究方法、技术路线和实验方案是基于大规模语料的计算词典学编辑方法与基于语言工作者的语感的传统词典学编辑方法相结合，使得述语形容词的研究真正建立在丰富和客观的语言事实基础上。具体做法是：先从 5000 万的语料库中抽取大量的包含述语形容词的句子实例，再由语言工作者人工分析和研究这些句子实例，填写包含具体述语形容词信息的工作单，然后再由录入员将词单录入计算机进行统一管理，并且依靠计算机来分析和统计述语形容词的各种信息。上图为整个形容词系统的研究框架：

三. 述语形容词词典工作单设计

工作单的设计遵循下列几条原则：

- 工作单设计以述语形容词为轴心，以每个形容词为中心组织一张工作单的信息；
- 研究方法是“语料库方法+联想”，即把主要从机贮语料库中获取的大量例证为依据的计算词典学方法同主要以语言工作者的经验、语感为依据的传统词典学方法结合起来，以使述语形容词研究真正建立在丰富的客观的语言事实基础上，同时发挥语言工作者的经验、语感进行联想和归纳。
- “形容词类型”的设计原则是将“从语义角度分类”、“可操作性”和“计算机处理用”几个原则综合考虑，将形容词分为感情形容词、感觉形容词、属性形容词和其他形容词。
- 欲研究和描写的术语形容词的义项选择原则为综合考虑形容词的分布性、形容词的常用频度和典型性。

在形容词的研究中包括词形、拼音、释义、论元数目、义项数目、义项序号等词性和论元属性，也包括形容词类型的语义分类属性。我们的研究把形容词分为感情形容词、感觉形容词、属性形容词、其他形容词四种语义类型。感情形容词是表示感情、心理活动和感情色彩比较强的形容词，例如高兴、愉快、可爱、痛苦、绝望等词。感觉形容词是指身体、皮肉、骨头、五脏六腑、眼、鼻、嘴、舌、耳等的感官感觉和对周围环境的整体感觉（除去明显的属性之外），例如渴、甜、苦、香、痛等词。属性形容词是指根据陈群秀拟定的《信息处理用现代汉语语义分类系统》中在“3.2.5”中表示物体和事物的各种属性的词。例如大、小、宽、窄、深、浅等物理属性和幼稚、成熟等生理属性。其他形容词是除感情、感觉、属性之外的形容词，例如凑巧、过分、平等等词。

在述语形容词的研究中，论旨属性信息是研究的重点。它包括基本式 1、变换式 1、基本式 2、变换式 2 等四个论旨模式和各自的句例、论旨角色的名称、句法功能、语义分类、语义特征、论旨实例等的具体信息。“论旨模式”，指该述语形容词该义项下各论旨角色名称与形容词的排列顺序表达式。“基本式 1”，表示该动词该义项下最基本的论旨模式，“变换式 1”，表示“基本式 1”的变换形式。“基本式 2”指区别于基本式 1 的第二种基本论旨模式，“变换式 2”表示“基本式 2”的变换形式。“句例”，指每个模式的一个或几个句例。“论旨名称”，指该动词该义项下基本式中的必要论旨角色的称呼。论旨名称的选择共有 22 个，包括：“施事”、“当事”、“领事”、“系事”、“受事”、“客事”、“分事”、“与事”、“同事”、“结果”、“基准”、“数量”、“范围”、“工具”、“材料”、“方式”、“依据”、“原因”、“目的”、“时间”、“处所”、“方向”。 “语类”，指论旨角色的句法范畴，规范使用的语类有 18 个：（1）N：名词；（2）V：动词；（3）A：形容词；（4）R：人称代词；（5）D：指示代词；（6）W：疑问代词；（7）ML：数量词；（8）NP：名词词组；（9）VP：动词词组；（10）AP：形容词词组；（11）S：小句；（12）DE：“的”字结构；（13）N（时间）：时间名词；（14）N（处所）：处所名词；（15）N（方位）：方位名词；（16）N（专名）：专有名词；（17）N（普名）：除 N（时

间）、N（处所）、N（方位）、N（专名）之外的其他名词；（18）X：表示任一语类，X=（1）+（2）+（3）+（4）+（5）+（6）+（7）+（8）+（9）+（10）+（11）+（12）。

“句法功能”，指该论旨角色在句子中充当的句法成分，采用传统的6种句法成分“主、谓、宾、定、状、补”，由于系统研究的是述语形容词，而此处又不研究形容词作为定语的情况，所以，实际上在本词典中用得上的是“主、宾、状、补”。“语义分类”指该论旨角色在语义分类体系中的上位义类。本规范使用的语义分类体系，为陈群秀拟定的“信息处理用现代汉语语义分类系统”。语义类用“{}”将其括起，若论旨角色有几种语义分类时用“|”分隔，表示“或”的关系。“语义特征”，指作为限制论旨角色的语义分类的辅助手段，例如“液体”、“毒性”、“片状”等。“论旨标记”，指论旨角色所带的前置或后置的介词或方位词，例如：“在”，“对”。“论旨实例”，指论旨角色的实例，用以对论旨角色的“语类”、“语义分类”、作形象化的或补充说明。

此外，还具体描述了形容词其他一些相关的语义句法信息：同义/近义词、反义词、可否作定语（及句例）；可否作状语（及句例）；可否作补语（及句例）；重叠方式；否定方式；可以用以限制的程度副词；可以使用的时态等。其中重叠方式包括 AABB、ABAB、AA、AAB、ABB、其他，例如红红火火(AABB)、高高兴兴（AABB）、雪白雪白(ABAB)、贼亮贼亮(ABAB)、厚厚（AA）、高高(AA)、甜甜蜜（AAB）、阴沉沉（ABB）、马里马虎(其他)、糊里糊涂（其他）、慌里慌张（其他）等。形容词的重叠在语用上一般表示感觉程度的加重（例如：红红的花、贼亮贼亮的眼睛、亮亮堂堂的屋子）、感情色彩的加强（例如：高高兴兴、红红火火的日子）和属性值的加大（例如：新新鲜鲜的菜、高高的个子、大大的眼睛）。形容词信息举例：

例 1：

<词形>：含糊　　<拼音>：han2hu5　　<形容词类型>：属性
<论元数目>：1　　<义项数目>：3　　<义项序号>：1
<释义>：不明确；不清晰。
<同义词/近义词>：含混|模糊|暧昧　　<反义词>：清晰|清楚|分明|明确
<基本式 1>：当事+含糊　　<句例>：态度含糊。他的话很含糊。
<变换式 1>：当事+是+含糊+的　　<句例>：有的表述是含糊的。
<论旨名称>：当事
<语类>：{N（普）|NP}
<句法功能>：主语
<语义分类>：{抽象物}
<论旨实例>：态度；意思；言语；表述
<可否作定语>：可　<句例>：含糊的色名应予去除。
<可否作状语>：可　<句例>：他含糊地跟我打招呼。　他含糊地表态。
<可否作补语>：可　<句例>：他回答得很含糊。
<可否重叠>：AABB
<否定式>：不 A
<加程度副词>：很 A；A 极了；最 A

例 2：

<词形>：含糊　　<拼音>：han2hu5　　<形容词类型>：感情
<论元数目>：1　　<义项数目>：3　　<义项序号>：2
<释义>：不认真；马虎。
<同义词/近义词>：马虎　　<反义词>：认真

<基本式 1>：当事+含糊　　<句例>：这事一点儿也不能含糊。 芸芸管俱乐部从不含糊。
<论旨名称>：当事
<语类>：{NP|S}
<句法功能>：主语
<语义分类>：{事件}
<论旨实例>：这事；芸芸管俱乐部
<可否作定语>：可　<句例>：聪明人谁做含糊事。
<可否作状语>：可　<句例>：长春百货都毫不含糊地给退还差价款。
<可否作补语>：可　<句例>：你可别做得那样含糊。
<可否重叠>：AABB
<否定式>：不 A
<加程度副词>：很 A；A 极了；最 A

例 3：
<词形>：邋遢　　<拼音>：la1ta1　　<形容词类型>：感觉
<论元数目>：1　　<义项数目>：1　　<义项序号>：1
<释义>：不整洁；不利落。
<同义词/近义词>：　　<反义词>：整洁|洁净|干净
<基本式 1>：当事+含糊　　<句例>：衣着邋遢。办事真邋遢。 她很邋遢。
<论旨名称>：当事
<语类>：{N（普）|N（专）|NP|R|VP}
<句法功能>：主语
<语义分类>：{人类|动作行为|衣着|具体空间}
<论旨实例>：她；办事；仔裤；房间
<可否作定语>：可　<句例>：该女性不是本来就缺乏警戒心的邋遢女子。
<可否作状语>：否
<可否作补语>：否
<可否重叠>：AABB　　<其他>：邋里邋遢
<否定式>：不 A
<加程度副词>：很 A；A 极了；最 A；A 透了

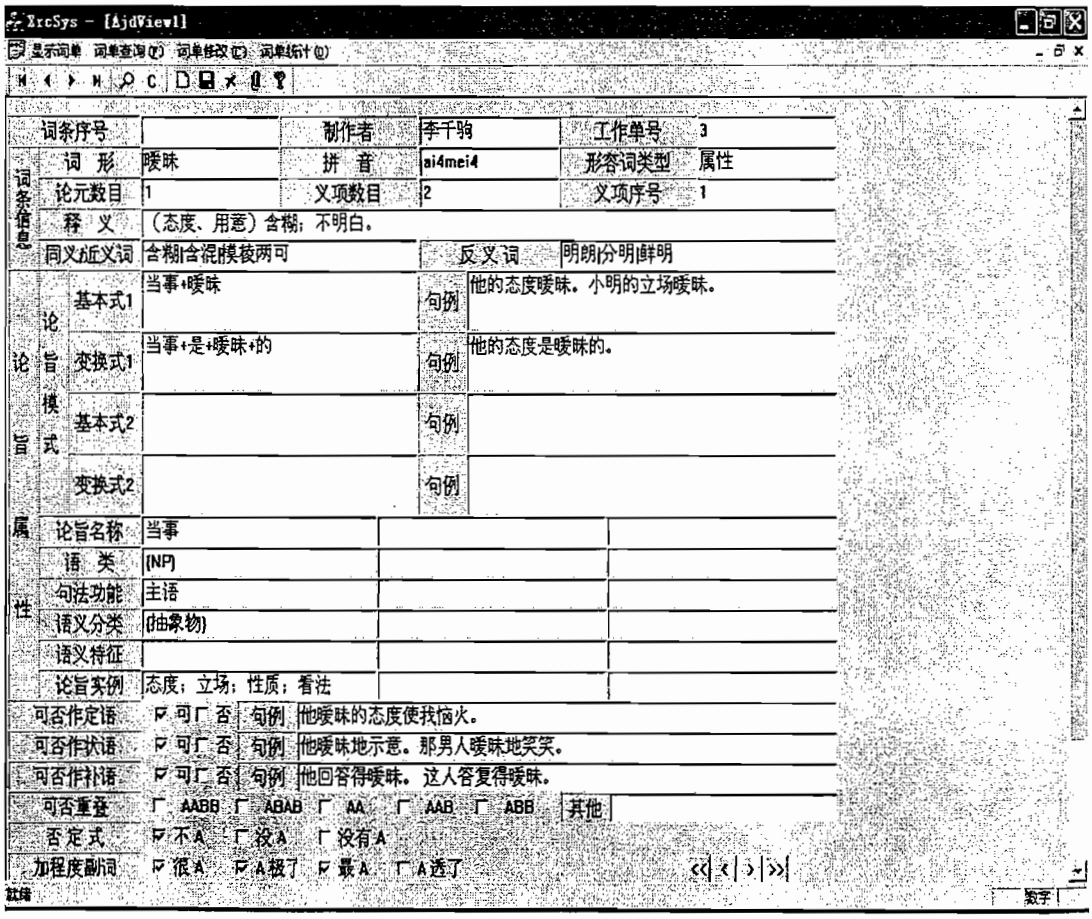
例 4：
<词形>：琐碎　　<拼音>：suo3sui4　　<形容词类型>：其他
<论元数目>：1　　<义项数目>：1　　<义项序号>：1
<释义>：细小而繁多，多用于书面语。
<同义词/近义词>：零碎|繁琐　　<反义词>：单纯|重要
<基本式 1>：当事+琐碎　　<句例>：办公室的事情很琐碎。 家务事琐碎得很。
<论旨名称>：当事
<语类>：{N（普）|NP}
<句法功能>：主语
<语义分类>：{事情}
<论旨实例>：事情；家务事
<可否作定语>：可　<句例>：琐碎的家务事把人弄得婆婆妈妈的。
<可否作状语>：否
<可否作补语>：否
<可否重叠>：AABB
<否定式>：不 A
<加程度副词>：很 A；A 极了；最 A；A 透了

四. 述语形容词系统的实现

述语形容词系统存储、管理着所有的形容词信息，而对信息的存储、管理是通过对形容词工作单的存储、管理实现的。在具体的实现中，使用数据库进行存储。使用 Microsoft Visual FoxPro 6.0 建立相应的数据表格，这样，不但可以手动进行操作，也可以用 FoxPro 语言进行管理，通过 ODBC 数据库操作，也可以用 Visual C++等各种通用编程语言进行引用和更改，具有相当大的通用性。而且对于数据库的数据，不需要全部读进内存，而是由数据库系统进行对磁盘的索引、访问。述语形容词系统实现的功能主要有：

1. 词单登陆：完成对新词单的完整性和合法性检查，登录到词单数据库中；
2. 词单查询：完成对库中词单的查询请求，并将查询到的结果反馈给用户；
3. 词单删除：查找到需要删除的词单后，经用户确认，从词单库中删除该词单；
4. 词单修改：查询到需要修改的词单后，由用户对词单进行需要修改取代旧的词单；
5. 词单浏览：提供给用户浏览词单库中词单的功能；
6. 词单统计：根据用户指定的内容，统计词单库中的信息并以文本形式输出；

由于软件实现函数数量较大，这里就不具体描述函数模块的分布和实现，下图为述语形容词系统的操作界面：



五. 结语

现代汉语语义知识库的重要组成部分：现代汉语述语形容词机器词典的完成与实现，使得整个语义知识库的构建工作又向前迈进了一步，将语义只是库中已有的三个部分：现代汉语述语动词词典、现代汉语名词槽关系系统、现代汉语语义分类系统与新完成的述语形容词词典结合起来；可以在框架语义学的理论基础上，进行各种汉语语义、句法有力分析，可以应用在机器学习、机器翻译、句法分析、汉语教学、人机交互接口等广泛的领域。目前我们已经完成了现代汉语常用形容词800个义项的全面描述，正在扩展描述义项，估计2000~2500个义项能够满足中文信息处理的需要。我们也正在研究句子的情态语义表示（汉语虚词的义类），并正在着手将现代汉语述语动词机器词典、现代汉语名词槽关系系统、现代汉语语义分类词典与刚研究建立的现代汉语述语形容词机器词典整合起来，加上机器学习的功能，将来再与现代汉语情态语义系统进行整合，形成一个现代汉语语义知识库，可为中文信息处理提供丰富、全面、可靠的语义知识支持，可为现代汉语语言学、语义学研究、对外汉语教学、中小学语文教学提供有力的工具和资源。

参考文献

[1] 陈群秀：“ 信息处理用现代汉语语义分类体系的设计思想” 《计算机时代的汉语和汉字研究》，清华大学出版社，1996年11月。

[2] 陈群秀等.：“现代汉语述语动词机器词典的设计与实现”， 新加坡1996年中文电脑国际会议（ICCC96）论文集, 1996年6月5日.

[3] Baker. C. F Fillmore, C. J Lowe, J. B: "The Berkeley FrameNet Project".

[4] "The FrameNet Project", International Computer Science Institute UCB Department of Linguistics.

[5] Edward A. Feigenbaum: "The Art of Artificial Intelligence: I. Themes and case studies of knowledge engineering"

[6] 朱育佳，陈群秀：“《现代汉语名词槽关系系统》研究实现与汉语语义知识库系统的初步建立”，2000’ ICMIP，乌鲁木齐，2000年8月。

[7] 陈群秀：“现代汉语名词槽关系系统初步研究”，《计算语言学文集》，清华大学出版社，黄昌宁、董振东主编，1999年10月。

[8]陈群秀：“现代汉语名词槽关系系统槽类型的研究和设计”，JSCL-2001会议论文集，2001年。

[9] 林杏光，张庆旭.：“现代汉语槽关系研究”， 陈力为，袁琦主编,《语言工程》. 清华大学出版社，1997。

[10]董振东，董强：“知网”，《计算语言学文集》，清华大学出版社，黄昌宁、董振东主编，1999年10月。

[10] 俞士汶，朱学锋等著：《现代汉语语法信息词典详解》，清华大学出版社，1998年4月。

[11] 黄曾阳：《HNC（概念层次网络）理论—计算机理解语言的新思路》，清华大学出版社，1998年11月。