CCD的构建及其应用*

咎红英[1, 2]　俞士汶[1]

1 （北京大学计算语言学研究所，北京，100871）

2 （郑州大学信息工程学院，河南郑州，450052）

{zanhy, yusw}@pku.edu.cn

[摘要] 本文介绍与 WordNet 兼容的中文概念词典（Chinese Concept Dictionary,CCD）的结构与构建方法，报告了北京大学计算语言学研究所 CCD 的研究与工程进展状况，并指出 CCD 在中文信息处理领域的应用前景。

[关键字] 中文概念词典;概念;同义词集合;语义关系

The Construction of CCD and its Application

ZAN Hongying[1,2]　YU Shiwen[1]

1(The Institute of Computational Linguistics, PekingUniversity, 100871, Beijing, China)

[2] (The College of Information Engineering, Zhengzhou University, 450052, Zhengzhou,Henan, China)

E-mail: {zanhy, yusw}@pku.edu.cn

Abstract: The paper introduced the structure and module of Chinese Concept Dictionary (CCD) compatible with WordNet. The author reported the research work and project status on CCD by the Institute of Computational Linguistics, Peking University, and presented the application of CCD in the Chinese information processing field.

Key words: CCD, NLP, concept, SynSet, semantic relation

一　引言

随着信息技术（Information Technology, IT）日新月异的迅速发展，计算机的应用越来越深入人们的日常工作和生活，人们急需更加个性化的服务。这不仅仅是计算机外壳更漂亮、各种配件的性能更优良等就可以解决的问题。自从计算机产生的那一天起，人们就把它看作人脑的

[作者简介] 咎红英（1966－），女，河南焦作人，讲师，博士生，主要研究领域为计算语言学,信息提取；俞士汶（1938－），男，安徽宣城人，教授，博士生导师，主要研究领域为计算语言学,自然语言信息处理技术。

*本文的研究工作得到国家自然科学基金项目(69973005)、973 项目（G1998030507-4)、863 项目（2001AA114040） 和北大 985 项目的支持。

延伸，以至于今天计算机更多地被称为电脑。经过五十多年理论和技术的迅猛发展，今天的计算机可以胜任复杂的计算、分析、控制和管理等工作，但是在智能方面还有广阔的发展余地。如何让电脑更接近人脑，让计算机懂得人类的语言，使人与计算机能更好的交流，这不仅是一个自然语言理解（Natural Language Understanding, NLU）和自然语言处理（Natural Language Processing, NLP）的技术问题，也是一个人工智能问题、一个哲学问题。多年来这个问题一直吸引着人们去探讨，在机器翻译、自动文摘、文本分类、语音识别与生成、信息检索、信息提取等诸多方面都取得了一定的进展。国外在这方面起步较早，有公认的评测机构，如由美国国防高级研究计划署（the Defense Advanced Research Projects Agency，DARPA）等资助的 MUC(Message Understanding Conference) 、TREC(Text REtrieval Conference) 等，吸引了国际上众多单位的参加，近年来我国也有部分单位参加了国际机构的评测。早在二十世纪 50 年代我国就对机器翻译有所研究，后来进入低潮。关于中文信息处理（Chinese Information Processing）方面的技术从 80 年代起再一次成为研究热点，特别是因特网（Internet）的普及，网上海量信息潮水般的涌入，人们迫切希望计算机能自动地精选出所需的资料，并且希望得到的网页按照相关度排序。但是目前的网络服务比如搜索引擎等还远没有达到人们的要求，常常是没有语义分析，只是根据所给出的查询词串的逻辑组合机械地给出一系列匹配网页，造成垃圾信息过多。因此，要想使计算机更聪明，使网络信息检索更智能，在自然语言特别是中文（本文中指汉语）的理解和处理方面还需要做大量的基础工作。

二　概念和语义的表示

人们在理解句子或文章时常常是通过分析其中关键词语或短语的概念及其语义关系来得到整体语义的。语义是思维的体现者，是客观事物在人们头脑中的反映，是人们交际过程的中心所在。语义问题涉及到哲学、社会学、心理学、认知科学、人工智能、逻辑学及数学（[11]）。对语言的理解主要是概念和语义的把握。要实现计算机自动分析和理解自然语言，就必须挖掘语言中的知识，形式化地表示语言的概念和语义。目前语义研究的主要成果有 WordNet、FrameNet、Mindnet，中文方面有董振东先生以义元方式描述概念的知网 HowNet（[8]）等。

汉语是基于词汇（包括短语）的意合语言[9]，因此基于词汇的概念和语义的研究很有意义。北京大学计算语言学研究所自 2001 年 4 月开始在国家自然科学基金等项目的支持下，着手中文概念词典（Chinese Concept Dictionary, CCD）的构建。CCD 从关系语义学的观点出发，用同义词集合（set of Synonyms, SynSet）来描述概念（concept），用概念间的关系（relation）来描述语义，方便语义关系的表示和检索，有利于简单地实现语义距离的计算，特别是同义词集合（同义关系）、上下位关系、整体/部分关系等的描述有利于概念的分级扩展，可以直接应用于机器翻译、自动文摘、文本分类、概念检索和信息提取等方面的语义理解。

三　CCD 的结构

著名的语言学家乔姆斯基（Noam Chomsky）认为人类的语言存在相对独立的句法规律，语言的无限性可由产生式规则的递归使用体现出来。这些年计算语言学的主流观点是：句法和语义是不可分割的整体，它们相互构成约束。二十世纪初，一些哲学家开始了语义的形式化的研究，如 Frege、Russell（Russell 在[12]和[13]中分析了人类语义知识、逻辑推理的共性）、Wittgenstein（[14]）、Carnap（[15]）等。进入 80 年代，自然语言的形式语义学和计算语义学开始走入计算

语言学，并逐渐被大多数的计算语言学家接受。这些研究的基础假设是：人们对概念、语义、知识的理解有很大的相似性；基本目标是：构造用于描述自然语言语义的元语言，使之形式化和可计算化。首先要解决的问题是词汇语义及其关系的形式化描述（属于词汇语义学研究范围），并构建一部可用于自然语言处理的语义词典（属于计算词典学的研究范围）。目前，最为成功的案例是 Princeton 大学的 WordNet（[6]）。为了继承已有的研究成果，并与国际标准接轨，北京大学计算语言学研究所构建的 CCD 继承了 WordNet 的主要结构、概念及语义关系，并针对中文特点进行了调整和发展。

Princeton大学认知科学实验室G. A. Miller教授、Fellbaum教授（心理学系）等人于 1983 年就开始了WordNet的设计与开发（真正有效的工作始于 1985 年，美国海军研究室、美国国防高级研究计划署曾为其提供过资金支持），历时近 20 年，现已完成WordNet1.7 的UNIX版本（PC 版本尚未推出），成为事实上的国际标准。在WordNet中，概念就是同义词的集合（SynSet，由可替换性原则确定），上位关系（hypernymy relation）是名词（或动词）概念间的主关系，另外还有一些辅助的关系（例如，名词概念间的对立关系、部分-整体关系等，动词概念间的反向假设关系、致使关系等，详见附录一）。可以简单地认为，WordNet[1]是一部机读的（machine-readable）语义词典，它通过同义词集合表示概念，通过概念间的关系描述英语概念之间复杂的网状语义关系[5，6]。同义词集合SynSet是WordNet词库的基石，也是WordNet构成一个义类词典的根本所在；关系指针代表了一个SynSet中的词与另一个SynSet中的词之间的语义关系。

2002 年 1 月 21 日至 25 日，由全球 WordNet 联盟主办的第一届全球 WordNet 国际会议（the 1st International Conference on Global WordNet）在印度南方的文化古城 Mysore 召开，会议吸引了超过 30 多个国家和地区的 100 余人参加，涉及语种 30 多个。北京大学计算语言学研究所作为中国大陆地区的唯一一家单位参加了次会议，于江生博士和刘扬同学报告了 WordNet 框架的中文概念词典（CCD）的研究和构建工作[1，2，3，4]，并得到了与会人员的广泛认可。

四　CCD 的构建

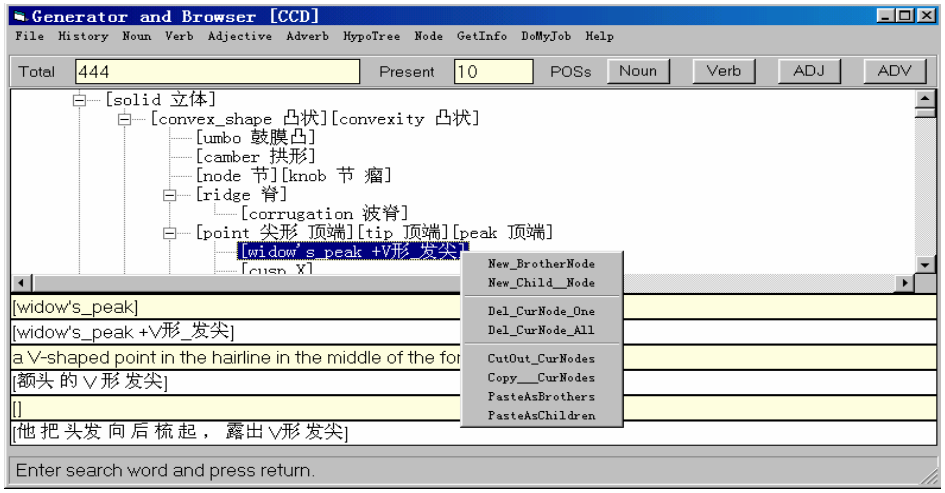
CCD的工作是以 1997 年发布的WordNet 1.6 版本为基础的。WordNet 1.6 版本包含四类实词的 99,642 个概念节点（其中名词概念 66025 个、动词概念 12127 个、形容词概念 17915 个、副词概念 3575 个）和超过 5×10^6个语义关系，形成了一张庞大的英语概念语义网络。其后台数据主要存放在两类ASCII码字符文件中，一类是数据文件（.dat），一类是索引文件.（idx），分不同词性存在相应的文件中。由于文化背景的不同，汉语和英语在本体论（Ontology）和概念上均有一定的差异，因此我们在构建CCD时，一方面从结构上继承了国际标准WordNet的概念及其语义关系和词汇关系，另一方面，CCD又不仅仅是WordNet的简单汉化，而是根据汉语特点和文化习惯进行了调整，并对WordNet自身的部分问题进行了探讨（[10]）。CCD的知识表示反映的是汉语的特点，主要面向的也是中文信息处理领域。

首先我们从名词做起，因为大部分的概念（特别是专业术语）都是名词（占 WordNet 总概念数的三分之二），名词语义的描述可在各种自然语言处理方面得到广泛而直接的应用。但是，直接翻译 WordNet 的概念是不可行的。因为尽管 WordNet 的数据文件和索引文件都是 ASCII

[1] WordNet 1.0 版于 1991 年 7 月公布，并向学术界免费公开。WordNet1.6 可由Princeton大学WordNet 主页免费下载，地址 http://www.cogsci.princeton.edu/~wn/

码文本文件，可以编辑，并且理论上也是可以可扩展的，但由于索引文件和数据文件相关性，实际上直接在原数据文件上编辑几乎是不可能的。因为编辑任何一个数据文件都可能造成错误的偏移字节量，从而造成检索错误。在一个索引文件中，每个记录有一部分是一个或多个字节的偏移量（Offset），每个偏移量指明一个 SynSet 在数据文件中的起始地址。检索 SynSet 或其它信息的第一步一般是在一个或多个索引文件中搜索单词，以获得包含这个单词的所有 SynSet 的数据文件地址。

我们在构建CCD工作中对WordNet的数据文件进行了改良和简化。首先抽取WordNet中的名词初始义类概念的上下位关系信息，形成了 15 个上下位关系树（参见附录三），并将每棵树的信息分别存在独立的Access数据库文件中，便于利用数据库的管理工具对数据进行操作。同时开发了构建CCD的可视化辅助构造软件（Visualized Auxiliary Construction of Lexicon, VACOL），使后续工作可以在上下位树形关系上可视化地进行（[10, 16]）。然后将分解的数据库文件分派给各个领域专家[2]，使他们在VACOL软件上并行工作。领域专家的主要工作是对SynSet中的英语单词、注释、用例进行中文翻译，依据中英文不同的文化背景对WordNet的概念结构进行整理，包括概念节点甚至整棵概念子树的增加、删除或位置调整以及WordNet中原有的多父亲节点、孤立节点、语义指针冲突等问题进行了修正，同时也对WordNet 中缺少用例的概念增加了相应的中文用例。接下来又对加工后的各数据库文件进行合并，对原有概念的语义关系直接继承WordNet中的关系指针，形成一个面向中文的双语概念词典数据库。



集浏览器和生成器为一体的 VACOL 软件界面

为了与 WordNet 格式兼容，我们最后将中英文分离，形成独立的中文概念词典数据库，并把数据库格式的数据转写为 WordNet 格式数据文件和索引文件。在数据转写中，我们先将指向未包含概念的关系指针过滤掉，从而保证了 CCD 阶段成果的封闭性。然后根据实际的中文信息重新计算了每个概念新的偏移地址，同时在数据库中保留 WordNet 原有老的偏移地址，这样既可以保证 CCD 中概念地址的正确，又可以随时得到 CCD 与 WordNet 的对应关系。在数据文

[2] WoldNet概念涉及不同领域的知识，必须靠各个领域的专家来处理。

件的生成过程中，我们还按照 WordNet 规范将各类关系指针规定了顺序，修正了 WordNet 中原来存在的关系指针顺序不一致的问题。最后是根据数据文件生成索引文件，这里重新计算了中文单词的义项个数，并根据其在数据文件中新的偏移地址建立了相应的地址索引列表。索引文件是按照中文词的字符串顺序进行排序的。

截止 2002 年 4 月，项目组按期完成了总计多达 30,000 个与 WordNet 兼容的双语概念的加工，提取出了相对独立的中文概念词典，并且顺利通过了合作单位北佳公司的阶段考核和验收。附录二给出了 CCD 数据文件和索引文件的部分样例。

四　CCD 的应用

CCD 的构建，不仅在概念和语义的表示上靠近了国际标准，而且面向中文信息处理，因此，在我所承担的其它项目中得到了直接的应用。首先，对概念的形式化描述和概念之间语义关系简明的结构使得 CCD 成为词义消歧（Word Sense Disambiguation, WSD）的主要词典资源，并可能在其它语义分析中得到应用。在机器翻译项目中我们所使用的语义词典也正在根据 CCD 的思想进一步扩展。在我们的建立信息科学与技术领域的术语库项目中，术语库的结构设计也借鉴了 CCD 语义关系表示的方法，即用上/下位、部分/整体、同义、同源、相关等关系指针来体现术语间网状的语义关系，使整个术语库的术语之间结构明晰，并方便术语库的维护与扩展。

另外特别值得一提的是，CCD 还用于我们的信息提取工作。作为国家自然科学基金项目，我们的信息提取项目主要完成面向新闻的出访、会议等信息提取的原型设计，其中命名实体的识别技术和自然语言的浅层分析技术都得到了进一步发展。在模板构建中，我们参考同义词词林，直接借用 CCD 中有关概念的同义词集合及其上下位概念的词汇集合，提高了提取结果的准确率和召回率。目前我们的信息提取工作侧重于基于网络的中文信息提取，正在与北京大学计算机系网络实验室合作，着手“ 网上名人踪迹 ”系统的研究与开发。万维网（World Wide Wed, WWW）自 1994 年开始登陆中国，短短几年内得到了迅猛的发展。北京大学计算机系网络实验室在李晓明老师的领导下，1996 年推出了 “ 天网 ”，在国内中文搜索引擎方面享有很高的知名度，特别是天网 FTP 检索在国内首屈一指。2001 年网络实验室又推出 “ 燕穹 ” 中国 Web 信息博物馆，为国人了解中文网站的历史提供了可能。据“天网” 搜集的网页估计，目前中文（简体）网页数已经超过 5000 万，网上海量信息的涌现迫使人们越来越依赖于搜索引擎，而目前中文搜索引擎的服务还远远不能满足用户的需要，检索结果中无关或无用的网页过多，大约有一半的结果是无关的，80%用户仅对前 2 页的查询结果感兴趣。因此，提高网上信息检索的智能性，按照用户关心焦点的相关度排序检索结果，提供个性化检索服务已势在必行。以上功能的实现均需要中文信息理解和处理的技术支持。目前正在进行的基础工作是根据不同类别的用户特征，参考 CCD 结构，构建特征概念分级扩展的用户信息表，然后利用分级扩展的特征概念对搜来的网页做相关度评价，最后给用户提供评价后的排序列表。此工作正在进行中，结果有待进一步实验。

五　结语

CCD 经历了近两年的艰苦工作，目前已初见成效，建成了含有近 30000 个概念封闭的、经过人工校对了的名词类双语语义辞典，探索出一条构建中文概念词典可行的模式，并且在我们的词义消歧、机器翻译、术语库建设、信息提取、概念检索等工作中的得到了应用。

但是要进一步完善 CCD 还有大量的工作，概念的扩充和调整是下一步的基本任务，主要包括新概念及其关系的自动获取，以及应用驱动的词典演化。WordNet 是从 80 年代中期开始创建，其收录的概念目前看来不足以科技最新发展的需要；在概念的分布上也不甚平衡，比如动物、医学等类过细，其它类又略粗；还有就是某些概念的分类与中文的习惯也有差异。要想使 CCD 真正成为一部实用的面向中信息处理的语义词典，还需要做大量细致的调整工作。

语义词典是自然语言处理的基础，也是让计算机逐渐智能起来的前提。CCD 的构建，面向中文信息处理，并同时与国际标准接轨，在中文语义词典及双语语义词典的构建方面进行了有益的探索。

参考文献：

[1] Yu, J. S., Yu, S. W., Liu, Y. and Zhang, H. R. 2001. Introduction to CCD. Proceedings of ICCC2001, Singapore.

[2] Liu, Y., Yu, J. S. and Yu, S. W. 2002. A Tree-Structure Solution for the Development of Chinese WordNet. Proceedings of GWC2002, Mysore, India, 2002, pp51-56

[3] Liu, Y., Yu, S. W. and Yu, J. S. 2002. Building a Bilingual WordNet-Like Lexicon: the New Approach and Algorithms. Accepted by COLING2002, Taipei, China.

[4] Yu, J. S., Liu, Y. and Yu, S. W. 2002. Construction of WordNet-Like Lexicon. Proceedings of WCLS2002, Taiwan, China.

[5] Miller,G.A. et al(1993). Introduction to WordNet : An On-line Lexical Database. Specification of WordNet.

[6] Fellbaum,C.(ed) (1999). WordNet : An Electronic Database, The MIT Press.

[7]俞士汶，朱学锋等（1998），现代汉语语法信息词典，清华大学出版社

[8]董振东，董强 “知网” http://www.keenage.com/

[9]鲁川（2001），汉语语法的意合网络，商务印书馆

[10]刘扬（2002），CCD 构造模型及 VACOL 辅助软件的设计与实现，第一届学生计算语言学会议论文集

[11] Partee, B.H. et al (1990). Mathematical Methods in Linguistics, Kluwer Academic Publishers.

[12] Russell, B. (1948). Human Knowledge --- Its Scope and Limits, Simon and Schuster.

[13] Russell, B. (1989). Logic and Knowledge, Unwin Hyman Ltd.

[14] Wittgenstein, L. (1953). Philosophical Investigations, Basil Blackwell Ltd.

[15] Carnap, R. (1966). Der Logische Aufbau Der Welt. Felix Meiner Verlag, Hamburg.

[16] Yu, J.S. (2001). Evolution of WordNet-like Lexicon, The First WordNet Conference, Mysore, India, 2002, pp134-142

[17] Yu, J.S. and Yu, S.W. (2002). Word Sense Disambiguation based on Integrated Language Knowledge Base, in Proceedings of International Conference on East-Asian Language Processing and Internet Information Technology (EALPIIT'2002), Hanoi, Vietnam, Jan. 8-11, 2002, pp411-417

[18] Yu, J.S., WSD and Closed Semantic Constraint, accepted by The First SigHAN Workshop on Chinese Language Processing

附录一：Wordnet中的关系指针符号及其含义说明 (WordNet Relational Pointers)

名词		动词		形容词		副词	
反义关系 Antonym	!	反义关系 Antonym	!	反义关系 Antonym	!	反义关系 Antonym	!
下位关系 Hyponym	~	下位关系 Troponym	~	近义关系 Similar	&	导出形式 Derived from	\
上位关系 Hypernym	@	上位关系 Hypernym	@	关系型形容词 Relational Adj.	\		
部分关系 Meronym	#	蕴涵关系 Entailment	*	又见 Also See	^		
整体关系 Holonym	%	致使关系 Cause	>	属性 Attribute	=		
属性 Attribute	=	又见 Also See	^				

附录二：CCD结果部分样例

1、 数据文件（NOUN.dat）

00669606 11 n 02 开端 2 开始 2 007 @ 00653451 n 0000 ~ 00669813 n 0000 ~ 00669903 n 0000 ~ 00669997 n 0000 ~ 00671206 n 0000 ~ 00671298 n 0000 ~ 00671394 n 0000 | 任何事物的开始; "良好的开端是成功的一半"

00669813 11 n 01 初潮 0 001 @ 00669606 n 0000 | 女性的初次经期; "初潮标志青春期的到来"

00669903 11 n 03 开始 3 起事 0 来临 0 001 @ 00669606 n 0000 | 开始或早期阶段; "冬季的来临"

00669997 11 n 02 开端 3 初现 0 001 @ 00669606 n 0000 | 最初的时期; "文明的开端"

00670080 11 n 01 原因 0 005 @ 00669089 n 0000 ~ 00670254 n 0000 ~ 00670475 n 0000 ~ 00670572 n 0000 ~ 00670979 n 0000 | 为事物的开始提供原动力的事件; "寻找飞机失事的原因"

00670254 11 n 01 前因 0 001 @ 00670080 n 0000 | 在一个事件之前发生的事件或原因; "战争的前因和后果"

00670356 11 n 02 开端 4 序幕 0 001 @ 00669089 n 0000 | 作为前导事件，引导或介绍以后的事物; "一首歌揭开了演出的序幕"

2、索引文件（NOUN.idx）

开头词 n 1 1 @ 1 0 00614881

开始 n 4 3 @ ! ~ 4 0 00473204 00669089 00669606 00669903

开尔文 n 1 1 @ 1 0 02170844

开战借口 n 1 1 @ 1 0 00653684

开明 n 2 2 @ ! 2 0 00591588 00594970

开普勒天体运动定律 n 1 2 @ ~ 1 0 00477260

附录三 CCD 名词概念分类