CCD构造模型及VACOL辅助软件的设计与实现*

刘扬　于江生　俞士汶

（北京大学计算语言学研究所，{liuyang, yujs, yusw}@pku.edu.cn，北京 100871）

[摘要] 作者依据自己在北大计算语言所 CCD（the Chinese Concept Dictionary）项目的工程实践，提出了 CCD 的构造模型。该模型强调双语词典（bilingual lexicon）构造中的继承（Inheritance）和转换（Transformation）的思想，希望从 WordNet 现有的英语单语词汇语义信息出发，通过词典编纂者的翻译和可视化操作（visualized operations），逐步实现由 WordNet 到 CCD 的计算性转换（computational transformation），自然地得到一个汉英对应的双语语义词典，从而大幅度提高此类词典编纂的质量和效率。针对该构造模型，作者设计并实现了一个可视化的辅助词典构造软件 VACOL（the Visualized Auxiliary Construction of Lexicon），该软件目前在计算语言所已得到大规模的应用。作者阐述了 VACOL 软件的设计原理，对其中涉及的一些关键算法和技术，如 WordNet 词汇语义信息抽取、数据敏感的树结构建立及其可视化操作等，也简要做了介绍。

[关键词] WordNet，CCD，双语语义词典，继承，转换

The CCD Construction Model & Its Auxiliary Tool VACOL

LIU Yang, YU Jiangsheng, YU Shiwen

Abstract: In this paper, we would like to put forth a new construction model of bilingual WordNet and to present some of the pivotal algorithms within its realization. A characteristic of this new approach is to emphasize the inheritance and transformation of the existent monolingual lexicon with the aim of a bilingual one. ICL (the Institute of Computational Linguistics) has benefited a lot by employing this model and tool to build CCD (the Chinese Concept Dictionary), a bilingual WordNet-like lexicon, in Peking University.

Keywords: Bilingual Semantic Lexicon, Inheritance, Transformation, WordNet, CCD

一 引言：WordNet 与 CCD

目前，自然语言处理（NLP）的理论和方法日渐关注内容信息的处理。无论做机器翻译（MT）、信息提取（IE）还是词汇语义排歧（WSD），在底层，都需要一部能够表达概念关系（conceptual relations）的语义词典（Semantic Lexicon）的支持，语义词典是所有这些应用的不可或缺的一项基础性资源。

[收稿日期] 2002-xx-xx

[作者简介] 刘扬，男，北京大学博士研究生。于江生，男，北京大学计算语言所副教授。俞士汶，男，北京大学计算语言所教授。

*本项研究得到国家自然科学基金项目 69483003、863 项目 2001AA114040、973 项目 G1998030507-4 以及北大 985 的支持。

从世界范围看，美国 Princeton 大学的 WordNet 无疑是当今最具影响力的一部语义词典。它针对的对象是英语语言词汇语义。该语义词典作为一个语言知识工程，在 G. A. Miller 和 C. Fellbaum 等人的主持下于 1985 年启动，经过 10 多年连续不懈的研究与开展，业已成为语义词典事实上的国际标准[Vossen, 1998]。

WordNet的基本思想是利用关系表示词汇语义[Miller, 1993]。

WordNet使用同义词集合（Synsets）代表概念（Concepts），并且力图在概念间建立不同的关系指针（relational pointers），表达不同的语义关系。这样，抽象的概念第一次被形式化了，变得具体而且可以通过词汇意义加以操作，概念之间还可以建立多种语义关系的联系和推理。这是在传统的义素分析法外，简单而有效地表达词汇语义的另一种新的方式和途径。

WordNet包含名、动、形、副等 4 类开放的实词，对介、连词等表达语法关系的各类虚词不予考虑。在 4 类实词中，又以名、动词为词典中语义刻画的重点，由心理语言学家分别作了深入的义类分类。WordNet词汇概念间的语义关系主要包括：同义（Synonymy）、反义（Antonymy）、上下位（Hypernymy/Hyponymy）、整体部分（Holonymy/Meronymy）、属性（Attribute）、蕴涵（Entailment）、致使 （Cause）等 （实际上，名、动词上具有其中的大部分关系，而形、副词一般仅具有同义、反义关系）。在 1997 年发布的WordNet 1.6 版本，已经描写了 4 类实词的 99,643 个概念节点和超过 5×10^6个语义关系，形成了一张庞大的英语概念语义网络。

90 年代以后，各国开始竞相开发各自的与 WordNet 兼容的（compatible）单语、双语或多语语义词典。迄今为止，P. Vossen 主持了包含欧洲 8 种主要语言（法、德、西、葡等）的 EuroWordNet 的开发工作，并在荷兰成立了“ 全球 WordNet 联盟 ”（the Global WordNet Association，简称 GWA）的国际组织。俄、日、印、韩等国也在开发与 WordNet 兼容的本国语言的语义词典。

2002 年 1 月，由“ 全球 WordNet 联盟 ”主办的“ 第 1 届全球 WordNet 国际会议 ” （the 1st International Conference on Global WordNet）在印度召开，参加会议的国家和地区有 30 多个。会议涉及如下专题：语义关系与词汇语义学（Semantic Relations & Lexical Semantics）、本体论与概念（Ontologies, Concepts & Top Levels）、双语与多语 WordNet 词典构造（Building WordNets）、词义消歧与语义标注（Disambiguation & Semantic Annotation）、WordNet 词典的其它应用（Applications）等。这些情况表明，WordNet 已经成为词汇语义词典的国际标准。在目前阶段，双语与多语 WordNet 词典的构造及其应用是其中最令人关注的一个热点。

北大计算语言学所自 2000 年 9 月启动CCD （the Chinese Concept Dictionary，即中文概念词典）项目，意在构造一个大规模（约 10^5个概念节点）网结构的汉英双语语义词典，直接为MT、IE和WSD等各项应用服务。

构造这样汉英双语语义词典，必须考虑与 WordNet 兼容的问题。具体说，对于 WordNet 中的每一个英语概念，在该双语语义词典里必须存在大体对应的汉语概念，反之亦然。当然，由于两种语言的本体论（Ontology）不同，汉语中的概念和概念间的语义关系跟英语可能会有细微差异。

在实际应用中，这样的双语 WordNet 语义词典能够提供较大的复用性（reusability）和开放性（openness）。它不仅是中文领域自然语言处理的关键基础资源，同时也是全球多语种 WordNet 资源建设的一个重要组成部分，必将对相关理论研究和应用产生巨大的推动。

二 构造双语语义词典的难点

然而，构造这样的词典却很困难。对双语语义词典而言，在同一部词典里，同时存在两类不

同的本体论，一个在汉语语言中，另一个在英语语言中。大体对应的双语概念间做匹配（Mapping）的工作不可避免；同时，随着时间的推移和社会的发展，这样的双语词典能否演化（Evolution）以及如何演化也是一个需要考虑的问题[Yu, 2002]。

考察 WordNet，其根本的组织原则可以概括如下[Beckwith, 1993]：语义词典中基本的组成单位是概念，即同义词集合；基本的概念间语义关系是上下位关系，即聚类分类关系。除上下位主关系将概念构成了树结构外，词典中还存在其它辅助关系，这些关系进一步将库中所有的概念树“编织”成一张巨大的网结构。

对于这么大规模的复杂结构，要实现语义词典中的同义词集合（约10^5个）和语义关系（约10^7个）的合理构成，困难是显而易见的。Princeton大学在构造WordNet过程中，计算语言学家们费了最多精力的事情就是：如何恰当地建立众多的同义词集合和语义关系，以及，如何在修订这些同义词集合和语义关系时仍能确保各种语义关系的一致性（consistency）。不言而喻，设想有一个大规模的网结构维护工具，问题看来就解决了 — 但是，由于维护网结构在时间和空间上的固有的复杂度（complexity），这样的工具根本无法实现。实际上，WordNet网结构的形成仍然靠对后台数据库的一点点加工。

回过头来，构造汉英双语语义词典，理论上讲，可以先按照Princeton大学的做法另外构造一部汉语WordNet，然后在该汉语WordNet和WordNet之间建立匹配映射，这是一个办法。在做CCD项目I期的1,500个概念时，我们就采用了这样的方案，但实际效果不好。在没有任何信息可凭借的情况下，所有的汉语同义词集合和语义关系都要重新构造（关于这一点，也可参考Princeton大学构造WordNet花费的工作量），此外，还要做双语概念的匹配映射，双语词典的演化问题也要解决。词典构造的效率和质量均不能让人满意。

总之，构造双语WordNet语义词典的任务确实十分艰巨。上述问题的存在，对双语词典的构造模型和开发工具都提出了特殊的要求。任何行之有效的解决方案，对此类问题的复杂性都必须有充分的认识和把握。

三 CCD 构造模型及其特征

通过以上探索和分析，作者认识到，要成功构造双语 WordNet 就必须找到这样一种模型：一，它能复用 WordNet 中原有的英语语言知识，并直接将这些知识视为汉语待用的词汇语义基础；二，这种复用不只是简单的英汉翻译，还应依据当前双语的实际情况对语义关系具备一定的调整手段。如果做到了这些，汉语 WordNet 的构造和双语 WordNet 的匹配显然都会受益。

实际上，在这种模型下，构造和匹配是在复用和调整中自动得到实现的，而且由于本来就提供了对双语调整的手段，双语词典的演化也不再成为问题。换句话说，新模型的特点是强调对现有的单语词典的继承（Inheritance）和转换（Transformation），由此得到新的双语词典，不试图另起炉灶。可以预期，这种模型会有很高的效率[Liu, 2002]。

与之对应，新模型需要两个过程来完成相应的功能：一，抽取 WordNet 中原有的英语语言知识作为汉语待用的词汇语义基础，其中，同义词集合及其上下位关系无疑是需要最先考虑的；二，开发可视化的（visualized）、数据敏感的（data-sensitive）工具来显示抽取出来的语言知识，词典编纂者（lexicographers）直接在上面操作来表达双语语义和语义关系的异同。

前一个过程只是为了得到一个可用的数据基，实际上，词典编纂者的工作总是集中在后一个过程上，双语 WordNet（包括双语词典构造、概念匹配及词典演化）就是在后一个过程中逐步、

自然地形成的。

在实际的语言知识工程中，新模型涉及计算语言学家和词典编纂者等两类人，他们需要相应的分工与合作。计算语言学家首先抽取出 WordNet 中的初始义类概念的上下位关系信息，并将此信息通过树结构组织成上下位关系树。树中的每个概念节点，即同义词集合，同时携带了其在 WordNet 中的所有其它辅助关系信息。接下来，词典编纂者在上下位关系树上交互操作来表达双语语义的异同。词典编纂者基本的操作情形如下：

I. 对树中的每一个英语概念节点，如果存在大体对应的汉语概念，词典编纂者只需将此英语概念翻译为汉语概念。

II. 如果没有对应物，情况可能是这样：该英语概念在上下位关系上相对于汉语习惯而言，或者太抽象（general），或者太具体（specific）。

II_1. 对于前者，词典编纂者要对该英语概念创建一些下位汉语概念，并将所有新创建的下位汉语概念同该上位英语概念联系起来。

II_2. 对于后者，词典编纂者要以一种特殊的方式“ 删除 ”该英语概念，亦即，该英语概念在汉语中没有对应物，只需将此英语概念同其上位汉语概念联系起来。

上述的语义“ 动作 ”就体现在对树定义的各种可视化操作（visualized operations）中。

实际上，在使用这些操作调整上下位关系时，记录在概念节点中的所有其它辅助关系也要通过程序系统化的合理的计算（computing）来加以调整。词典编纂者本身只是简单地在前台树上选择合适的操作、表达他所期望的语义关系，根本无需关心为了体现该操作，后台数据库究竟应当如何进行修改。该模型通过对前台树结构的可视化操作，完全实现了对后台网结构信息的计算性修改（computational modifications）。

可以看到，上述构造 CCD 词典的模型，实际上并不依赖汉语本身的特点。在不同的国家和地区，构造本国语和英语对应的双语 WordNet，都能使用该模型，它是一个通用的双语 WordNet 解决方案，对此类词典的构造具有一般的方法论意义。

四 VACOL 辅助软件的设计与实现

在项目组成员的大力协助下，作者经过近两年的研发，用于构造 CCD 词典的可视化的辅助软件 VACOL（the Visualized Auxiliary Construction of Lexicon）目前已具备雏形。构造模型中涉及的两个核心问题都得到了很好的解决，现将有关算法简要介绍如下。

首先是抽取 WordNet 中原有的上下位关系信息。

实际上，WordNet 潜在的上下位关系树是十分不均衡的（unbalanced），在一个网结构的海量空间里，常规的搜索算法通常难以奏效。比如，在 WordNet 1.6 中输入“ entity ”，查找其所有下位概念，系统除了“ Search too large. Narrow search and try again. ”的提示外，什么有用的信息都得不到。

在这方面，作者目前已经实现了用于获取所有下位概念信息的一个优化搜索算法[Liu, 2002]。

大致说来，新算法包含多回合的二路扫描（the Two-Way Scanning Process）和收滤编码（the Gathering/Sieving and Encoding Process）的过程，而每一回合都试图得到上下位关系树中处于当前层上的所有节点的信息，回合的次数等于特定概念的上下位关系树的深度。

应用该算法，实际的搜索空间和时间复杂度都大为降低，从而使完整地抽取 WordNet 中原有的上下位关系信息成为可能。

其次是带语义操作的可视化的、数据敏感的上下位关系树的设计与实现。这方面的工作比较复杂，因为它同时涉及前台的树操作和后台的数据文件维护。

树结构的实现可以采用 Microsoft Visual Studio 6.0 开发套件提供的 Treeview 控件。

用来编辑树结构的可视化操作，作者提供了“节点添加”、 “节点修改” 、 “节点删除”和“子树移动”等 4 类。词典编纂者在树上选定了概念节点，就可以从所有这些类操作中选择合适的一个来加以运用。这些操作都是经过精心考虑后决定采用的，基本的指导原则是，要保证选取的操作在含义和功能上足够简洁（concise）、足够有效（capable）。易于证明，任何形状的树结构，都可以通过反复运用这些操作来达到。

考虑数据结构，一个具有 n 个节点的上下位关系树的后台数据文件可以描述如下：

Pos 1	Ptr 11	Ptr 12	…	Ptr 1m	BasicInfo 1
Pos 2	Ptr 21	Ptr 22	…	Ptr 2m	BasicInfo 2
…	…	…	…	…	…
Pos n	Ptr n1	Ptr n2	…	Ptr nm	BasicInfo n

表中每条记录实际上包含了 3 部分的信息：树结构信息{Pos_i}，网结构语义关系信息{Ptr_{i1} , Ptr_{i2}, … , Ptr_{im}}，以及，仅与当前概念有关的概念自身信息{$BasicInfo_i$}。

对前台的每一种树操作，都需要恰当地处理后台数据文件中的这些信息。首先，至关重要的是维护两类信息的一致性：其一是树结构信息{Pos_i}的一致性，其二是网结构语义关系信息{Ptr_{i1}, Ptr_{i2}, … , Ptr_{im}}的一致性，这些都要靠程序系统化的合理的计算来加以调整。此外，概念自身信息{$BasicInfo_i$}因为局限于各记录内部，只涉及英汉翻译的问题，与结构调整没有任何关系。

上述思想实现了前台树操作和后台数据文件的维护，相应算法已被 COLING2002 大会录用。感兴趣的读者可以在参考文献[06]中找到较详细的算法描述。

五 VACOL 在 CCD 项目中的应用与实践

按照该模型的工作流程，在 CCD 项目 II 期（暂且只考虑名词类），作者首先抽取 WordNet 中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，并将每个树的信息分别存在单独的 Access 数据库文件中。之后，项目组将这些数据库文件分派给不同的词典编纂者，供其在 VACOL 软件上独立操作处理，不同词典编纂者的工作可以并行。

在 4 个月的时间内，项目组按期回收了加工过的总计多达 30,000 个的双语概念，并且顺利通过了合作单位北佳公司的阶段考核和验收。与 I 期方案相比，新模型无论在效率上还是在质量上，优越性都十分明显。

作为该模型及工程实践的一部分，在研发过程中，作者还首次发现了 WordNet 1.6 版本在语义表达方面一些严重缺陷和错误，它们在语义关系方面的逻辑是不严格的，需要做进一步的改进（作者曾就这些问题向 C. Fellbaum 本人和台湾中研院黄居仁教授请教，得到确认）：

1. 在名词中存在“ 处于同一语义范畴（Semantic Category）上的多个上位（Hypernym）”现象（总共 772 个，如{radish}）和“ 处于不同语义范畴上的单个上位 ”现象（总共 2172 个，如{prayer_wheel}）；

2. 在名词中存在“ 整体（Holonym）即为自身 ”现象（总共 3 个，如{science, scientific_discipline}）

和“部分（Meronym）即为自身”现象（总共 3 个，如{science, scientific_discipline}）；

3. 在名词中存在“整体与上位互有交叉”现象（总共 8 个，如{car_seat}）和“部分与下位（Hyponym）互有交叉”现象 （总共 8 个，如{seat}）；

4. 动词与名词有类似的情况出现。非常特别的是，在动词中还存在“上位即为自身”现象（总共 1 个，如{reserve, hold, book}）；

5. 另外，在各个词类中均存在“DAT 文件中语义关系指针不遵循位置约定规范”现象（如上位关系指针“@”和下位关系指针“~”各自都不连续）。

这些问题的存在，说明 Princeton 大学在“词典语义规范”和“词典结构检查”方面的工作做得不够。

六 结语：后续工作与打算

目前，在 CCD 项目 II 期的实施中，从英语到汉语的翻译采用的是纯手工方式，由作为词典编纂者的语言学家和领域专家填写完成。在即将开始的 CCD 项目 III 期，于江生博士提出了概念的自动翻译的想法和初步算法。同时，作者也希望借鉴台湾清华大学张俊盛教授在该方面的一些成功做法[Chang, 2002]。

基本的想法是，在初步得到自动翻译的版本后，才考虑采用本模型在 VACOL 软件上由词典编纂者进行进一步的人工校对和语义结构的改进。

如果自动翻译的办法可行，再配以 VACOL 软件已经提供的人工调整的便利手段，则该模型无论是在理论的完整性方面，还是在工程实践的效率方面，都有更大的指导意义。

致谢

北大计算语言学研究所自 2000 年 9 月启动 CCD 项目，本文即是两年多来研究工作的一个总结。作者衷心感谢王逢鑫教授、鲁川教授、周锡令教授、张化瑞老师、宋春燕同学、昝红英同学、温珍珊同学、李佐文博士、亢士勇教授、刘云博士以及北佳公司刘东先生等人对本文工作的支持和贡献。

本文的一部分成果已经在“全球 WordNet 联盟”主办的“第 1 届全球 WordNet 国际会议”上做了报告，作者也感谢与会者深刻的评价和各种好的建议。

[参考文献]

[01] Beckwith, R., Miller, G. A. and Tengi, R. 1993. Design and Implementation of the WordNet Lexical Database and Searching Software. Specification of WordNet.

[02] Chang, J. S, You, G. N. et al. 2002. Building a Bilingual WordNet and Semantic Concordance from Corpus and MRD. WCLS2002, Taipei, China.

[03] Fellbaum, C. 1999. WordNet: an Electronic Lexical Database. Cambridge, Mass.: MIT Press.

[04] Kamps, J. 2002. Visualizing WordNet Structure. ICGW2002, India.

[05] Huang, C. R., Tseng, I. J. E. and Tsai, D. B. S. Translating Lexical Semantic Relations: the First Step towards Multilingual WordNets. SEMANET2002, Taipei, China.

[06] Liu, Y., Yu, J. S. and Yu, S. W. 2002. A Tree-Structure Solution for the Development of Chinese WordNet. ICGW2002, India.

[07] Liu, Y., Yu, S. W. and Yu, J. S. 2002. Building a Bilingual WordNet-Like Lexicon: the New Approach and Algorithms. COLING2002, Taipei, China.
[08] Miller, G. A. 1993. Noun in WordNet: a Lexical Inheritance System. Specification of WordNet.
[09] Miller, G. A. et al. 1993. Introduction to WordNet: An On-line Lexical Database. Specification of WordNet.
[10] Pavelek, P., Pala, K. 2002. VisDic — a New Tool for WordNet Editing. ICGW2002, India.
[11] Vossen, P. 1998. EuroWordNet: a Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer.
[12] Wilson, R. A., and Keil, F. C. 1999. The MIT Encyclopedia of the Conginitive Sciences. London: MIT Press, Cambridge, Massachusetts.
[13] Yu, J. S. 2002. Evolution of WordNet-Like Lexicon. ICGW2002, India.
[14] Yu, J. S., Liu, Y. and Yu, S. W. 2002. Construction of WordNet-Like Lexicon. WCLS2002, Taipei, China.
[15] Yu, J. S., Yu, S. W., Liu, Y. and Zhang, H. R. 2001. Introduction to CCD. ICCC2001, Singapore.