

CCD 语义知识库的构造研究*

刘扬, 俞士汶, 于江生

北京大学计算语言学研究所, 北京, 100871

摘要: CCD (the Chinese Concept Dictionary) 是一个 WordNet 框架下的汉英双语语义知识库 (the Chinese-English WordNet)。在制定语义规范后, 作者提出了构造 CCD 的演化模型 (the model of evolution)。新的构造模型强调双语语义知识库构造中的继承 (inheritance) 和转换 (transformation) 思想, 希望从 WordNet 现有的英语单语语义信息出发, 通过词典编纂者的联机翻译 (online translations) 和可视化操作 (visualized operations), 逐步实现由 WordNet 到 CCD 的计算性转换, 自然地得到一个双语语义知识库, 从而大幅度提高构造此类语义知识库的效率和质量。针对该构造模型, 作者设计并实现了可视化的辅助词典构造软件 VACOL。该软件在北大计算语言所 CCD 项目中得到大规模的应用, 取得了很好的成果。

关键词: 语义分析; 语义知识库; 概念; 双语 WordNet; CCD

A Study on the Construction of CCD

Liu Yang, Yu Shiwen, Yu Jiangsheng

The Institute of Computational Linguistics, Peking University,
Beijing, 100871

Abstract: After the establishment of the specification of the knowledge-base, the author puts forth an evolutive model of the construction of CCD, a bilingual Chinese-English WordNet. A characteristic of this new model is to emphasize the inheritance and transformation of the existing monolingual knowledge-base during the construction of the bilingual one. The bilingual knowledge-base just comes into being through the lexicographer's online translations and visualized operations; hence the efficiency and quality of the construction are highly improved. This approach is a language-independent one and actually brings forward a general solution to the construction of the bilingual WordNets in all nations. For this model, the author designed and eventually realized the auxiliary construction tool VACOL. The ICL (the Institute of Computational Linguistics) has benefited a lot by employing this tool to construct CCD in Peking University.

Keywords: Semantic Analysis; Semantic Knowledge-base; Concept; Bilingual WordNet; CCD

1. 引言

在过去, 计算语言学研究集中在词法分析和句法分析 (morphological & syntactic parsing) 上, 基于规则或基于统计的成熟的语法分析技术率先在各种自然语言处理领域得到广泛运用。目前, 随着 Internet 应用的普及, 多语种内容信息的理解和处理逐渐受到关注。一些语义分析 (semantic analysis) 方面的研究, 如词义消歧 (上下文中的同义词、多义词的理解和处理, 即 Word Sense Disambiguation, 简称 WSD) 和语义归纳、推理等, 开始处于萌芽期并逐渐走向前台, 成为下一阶段计算语言学研究的一个亮点。

自然语言的语义分析和内容信息的理解, 离不开语义知识库 (the Semantic Knowledge-base) 的支持, 它是帮助计算机“了解”人类语言的一个媒介和手段, 也是让计算机逐渐“智能”起来的物质前提。

20 世纪 80 年代中后期, 一些国家开始构造各种面向机器的语义知识库。迄今为止, 语义知识库研究的主要成果有美国 Princeton 大学的 WordNet^[2]、California 大学 Berkeley 分校的 FrameNet 和 Microsoft 公司的 MindNet, 它们描写的对象都是英语语言; 汉语方面则有中国科学院计算机语言信息中心董振东先生的 HowNet (或称知网)^[9, 10]。

在现有的这些语义知识库中, WordNet 以体系简洁、内容详实而著称。由于该语义知识库中刻画的语义内容丰富、完备, 对实际的自然语言处理工作有很大的推动和促进作用, 逐渐成为其中最具有广泛影响力的一部。

该语义知识库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 年正式启动, 经过 10 多年持续不懈的研究与开发, 业已成为自然语言处理领域最重要的公用语义资源和语义知识库事实上的国际标准^[6]。在 1997 年发布的 WordNet 1.6 版本, 已经描写了 4 类实词的近 10^5 个概念节点和近 10^7 个语义关系, 形成了一张庞大的英语概念语义网络。

WordNet 的基本思想是利用关系表示词汇语义

[2]。WordNet使用同义词集 (set of synonyms , 简称 Synset) 代表概念 (concept) , 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。WordNet中概念间的语义关系包括: 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) 。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。

迄今为止, WordNet 研究的国际性学术组织 GWA (the Global WordNet Association) 和国际性学术会议 ICGW(the International Conference on Global WordNet) 都已成立或召开; 同时, 由于存在“人类语言的认知语义结构具有共通性”的假设, WordNet 作为人类语言语义知识库研究和应用标准的地位以后还会加强。

当前, 世界上许多国家在开发WordNet框架下的双语语义知识库, 这成为语言资源建设的一个热点。这些项目由于语种和应用侧重点的不同, 在“构造什么样的”方面存在一些差异, 但由于最终的语义知识库都是大规模的“双子网”的复杂结构, 在“如何构造”方面基本上面临同样的难题[5]。

在这样的大背景下, WordNet框架下的双语语义知识库, 特别是汉英双语语义知识库的构造具有重要的现实意义, 它不仅是中文信息处理的关键的基础资源, 同时也是全球多语种WordNet资源建设的一个重要的组成部分, 必将对相关理论研究和应用产生巨大的推动[12]。

受国家自然科学基金、863 项目和 973 项目等支持, 北大计算语言所自 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WordNet框架下的汉英双语语义知识库, 直接为机器翻译 (MT) 、信息提取 (IE) 等各项应用服务[12]。该研究属于跨学科的前沿性工作, 需要语言学和计算机科学等多领域知识的配合。

2. 构造 CCD 的难点分析

作为WordNet框架下的双语语义知识库, CCD 必须考虑与WordNet兼容的问题。对于WordNet中的

每一个英语概念, 在CCD里必须存在大体对应的汉语概念, 反之亦然。当然, 由于两种语言的知识体系 (Ontology) 不同, 汉语中的概念和概念间的语义关系跟英语中的可能会有细微差异[5]。

在实际应用中, 这样的汉英双语语义知识库能够提供较大的复用性 (reusability) 和开放性 (openness) 。

然而, 构造CCD这样的语义知识库却很困难: 对汉英双语语义知识库而言, 在同一部知识库里, 同时存在两类不同的知识体系, 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映射的大量工作是不可避免的; 同时, 随着时间的推移和社会的发展, 双语语义知识库能否演化以及如何演化也需要加以考虑[3]。

由于涉及两类不同的知识体系及其概念的映射, CCD语义知识库内部结构十分的复杂。它包含大规模的“双子网”的复杂结构, 每个子网各自约 10^5 概念节点和 10^7 个概念关系, 而且因为不同语种的语义结构存在差异, 这两个子网并不同构。

语义知识库中双语概念的对应还可能是一个多对多 (many-to-many) 的关系, 下图显示了双子网的这种复杂对应关系。例如, 概念对“ C3/ E3 ”、“ C4/ E2 ”之间存在一一对应关系, 但概念“ C5 ”却与多个概念“ E1 ”、“ E5 ”、“ E6 ”同时构成对应关系 (即在概念意义上汉语概念“ C5 ”应当与英语概念“ $E1 \cup E5 \cup E6$ ”相当) 。

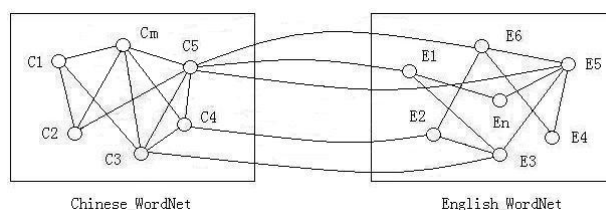


图 1 双子网的复杂对应关系示意图

实际上, 如何得到不同语种的语义子网, “双子网”间的概念如何映射, 双语语义知识库能否演化以及如何演化, 这些问题在构造中都要妥善地加以解决。因此, 构造 WordNet 框架下的双语语义知识库的工程非常庞大, 开发什么样的解决方案至关重要。解决方案的优劣对工程进度 (时间、人工开销以及资金投入等) 和工程结果 (语义知识库的内在质量) 的影响是巨大的。

考察WordNet, 其根本的组织原则可以概括为: 语义知识库中基本的构造单元是概念, 即同义词集; 基本的概念间语义关系是上下位关系, 即聚类分类

关系^[2]。除上下位主关系将概念构成了树结构外, 知识库中还存在其它语义关系, 像反义关系、属性关系、整体部分关系、蕴涵关系以及致使关系等, 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, 要实现语义知识库中的约 10^5 个同义词集和约 10^7 个语义关系的合理构成的目标, 工程上的难度是显而易见的。Princeton 大学在构造 WordNet 过程中, 计算语言学家们花费了最多精力的事情就是: 如何恰当地建立起众多的同义词集和语义关系, 以及, 如何在修订这些同义词集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络结构维护工具, 问题看来就解决了——但是, 由于维护网结构的算法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加工。

构造 WordNet 框架下的汉英双语语义知识库, 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 和英语 WordNet 之间建立映射, 这是一个办法 (在做 CCD 项目一期的 1,500 个概念时, 作者就采用了这样的方案), 但实际效果并不理想: 在没有任何信息可凭借的情况下, 所有的汉语同义词集和语义关系都要重新构造 (关于这一点, 可以参考 Princeton 大学构造 WordNet 花费的工作量); 此外, 还要做双语概念的映射, 双语语义知识库的演化问题也要解决。知识库构造的效率和质量均不能让人满意。

总之, 构造双语 WordNet 语义知识库的任务确实十分艰巨。上述问题的存在, 对双语知识库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^[3, 7, 8]。

3. 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 WordNet 框架下的汉英双语语义知识库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^[11]。

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^[5]。

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中, 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) 考虑上下位关系树上的当前英语概念节点, 如果存在与其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, 这要采用“节点修改”的操作。

(II) 如果不存在与其大体对应的汉语概念, 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), 或者在分类上不合适 (inappropriately-classified)。在这种情况下, 应当具体分析属于哪种情况。

(II₁) 如果是“太抽象”, 词典编纂者要对该英语概念创建一些下位汉语概念, 并将所有新建的下位汉语概念同该上位英语概念联系起来, 这要采用“节点添加”的操作。

Get $\{L_{(i+1)1}, L_{(i+1)2}, \dots, L_{(i+1)n'}\}$ from file DAT with their $Pos \equiv \{Pos_{(i+1)1}, Pos_{(i+1)2}, \dots, Pos_{(i+1)n'}\}$ by the encoding rule;

NodeSet = NodeSet $\cup \{L_{(i+1)1}, L_{(i+1)2}, \dots, L_{(i+1)n'}\}$;

END IF

END IF

Gather those members of NodeSet as treenode of the hyponymy tree;

NodeSet' = $\{Node \in NodeSet \mid Node \text{ satisfying conditions of being leaves of the hyponymy tree}\}$;

NodeSet = NodeSet - NodeSet';

LOOP

4.2 实现转换：可视化、数据敏感的语义树及其操作

该过程要实现一个带语义操作的可视化的、数据敏感的上下位关系树。

在功能设计方面，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。

语义树及其操作的工作比较复杂，它同时涉及前台的树操作和后台的数据文件维护。

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为： $\{Pos_i, Ptr_{i1}, Ptr_{i2}, \dots, Ptr_{im}, BasicInfo_i\}$ 。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 $\{Ptr_{i1}, Ptr_{i2}, \dots, Ptr_{im}\}$ ，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 $\{Ptr_{i1}, Ptr_{i2}, \dots, Ptr_{im}\}$ 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COLING'02大会录用^[4]。

如果不拘泥于不同树操作的琐碎细节，树操作算法的伪码可以统一描述如下：

The lexicographers trigger an action on $node_i$;

IF the action is in operations [1, 2, 3, 4]

CASE the action

Operations [1]:

Add a node with its Pos \equiv NewBrother

(Pos_i);

Operations [2]:

Add a node with its Pos \equiv NewChild

(Pos_i);

Operations [3]:

Delete the node with its Pos \equiv Pos_i;

Operations [4]:

Delete all the nodes with their Pos satisfying conditions of being descendants of $node_i$;

END CASE

Recalculate Pos of the rest nodes in the table according to the operation and current Pos_i;

Replace all relevant $Ptr_{j1}, Ptr_{j2}, \dots, Ptr_{jm}$ with new ones according to the operation and current $node_i$;

Refresh the tree;

ELSE IF

The lexicographers translate Current BasicInfo_i from English to Chinese;

END IF

5. 构造 CCD 的工程实践

按照该模型的工作流程，在 CCD 项目二期和三期，作者首先抽取 WordNet 中初始义类概念的上下位关系信息，形成了 25 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。

在近两年的时间内，项目组按期回收了加工过的总计多达 60,000 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。与一期方案相比，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 Princeton 大学 C. Fellbaum 本人和台湾中研院黄居仁教授请教，均得到确认）：

(1) 在名词中存在“处于同一语义范畴上的多个上位概念”现象和“处于不同语义范畴上的单个上位概念”现象。前者总共出现了 772 例，如 {radish}，后者总共出现了 2,172 例，如 {prayer_wheel}；

(2) 在名词中存在“整体概念即为自身”现象和“部分概念即为自身”现象。前者总共出现了 3 例，如 {science, scientific_discipline}，后者总共出现了 3 例，如 {science, scientific_discipline}；

(3) 在名词中存在“整体概念与上位概念交叉”现象和“部分概念与下位概念交叉”现象。前者总共出现了 8 例，如 {car_seat}，后者总共出现了 8 例，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“上位概念即为自身”现象。这种现象总共出现了 1 例，如 {reserve, hold, book}；

(5) 另外，在各个词类中均存在“数据文件中语义关系指针不遵循位置约定规范”现象。比如，数据文件中的上位关系指针“@”和下位关系指针“~”不能保证各自的连续性，有交叉出现的情况。

这些问题的存在，说明 Princeton 大学在“知识库语义规范”和“知识库结构检查”方面的工作做得不够，Princeton 大学的 Grinder 软件对大规模复杂结构的数据处理还存在某些漏洞。

6. 结语

截止目前，在 CCD 二期、三期项目的实施中，CCD 的构造工作是在 VACOL 辅助软件的支持下完成的，概念（同义词集）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。在以后，作者希望尝试用程序实现概念的自动翻译、抽取反义词对等工作，进一步探索自动构造 CCD 的模式。目前，作者已经有了一些初步算法。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^[1]。

未来的构造模型的基本设想是：在初步得到自动翻译的版本后，才考虑在 VACOL 辅助软件上由词典编纂者进行下一步的人工校对和语义结构的修改。如果自动翻译的办法可行，再配以 VACOL 辅助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

北大计算语言所自 2000 年 9 月启动 CCD 项目，本文即是近三年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、黄居仁教授（台湾）、詹卫东副教授、张化瑞老师、宋春燕同学、咎红英同学、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。

参考文献

- [1] Chang, J. S., You, G. N. et al. Building a Bilingual WordNet and Semantic Concordance from Corpus and MRD. WCLS'02, Taipei, China. 2002.
- [2] Fellbaum, C. WordNet: an Electronic Lexical Database. Cambridge, Mass.: MIT Press. 1999.
- [3] Liu, Y., Yu, J. S. and Yu, S. W. 2002. A Tree-Structure Solution for the Development of Chinese WordNet. ICGW'02, India. 2002.
- [4] Liu, Y., Yu, S. W. and Yu, J. S. 2002. Building a Bilingual WordNet-Like Lexicon: the New Approach and Algorithms. COLING'02, Taipei, China. 2002.
- [5] Pianta, P., Pala, K. VisDic -- a New Tool for WordNet Editing. ICGW'02, India. 2002.
- [6] Vossen, P. EuroWordNet: a Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer. 1998.
- [7] Yu, J. S. Evolution of WordNet-Like Lexicon. ICGW'02, India. 2002.
- [8] Yu, J. S., Liu, Y. and Yu, S. W. Construction of WordNet-Like Lexicon. WCLS'02, Taipei, China. 2002.
- [9] 董振东，董强，知网简介，网上资料，<http://www.keenage.com/>，1999 年。
- [10] 董振东，语义关系的表达和知识系统的建造，《语言文字应用》，1998 年第 3 期。
- [11] 刘扬，于江生，俞士汶，CCD 构造模型及 VACOL 辅助软件的设计与实现，《语言文字应用》，2003 年第 1 期，P83-88。
- [12] 于江生、俞士汶，CCD 的结构与设计思想，《中文信息学报》，2002 年第 4 期。

