

汉语句类依存树库的构建研究

王慧兰

解放军外国语学院二系, 洛阳 471003; E-mail: hlwang9@hotmail.com

摘要 以机器翻译领域为应用目标, 以概念层次网络理论的语义网络和句类分析方法为理论基础, 探讨句类依存树库构建的理论和标注实践等问题, 详细描述了构建树库所需的类别标注集和句类关系标注集, 并给出了句类依存树样例。

关键词 汉语树库; 概念层次网络理论; 句类依存树库

中图分类号 TP391

Construction of Chinese Sentence-Category Dependency Treebank

WANG Huilan

The Second Department, PLA University of Foreign Languages, Luoyang 471003; E-mail: hlwang9@hotmail.com

Abstract Aiming at the area of machine translation applications, this paper conduct research on the construction of Chinese sentence-category dependency treebank (CSCDT) based on the theory of hierarchical network of concepts. Conceptual category tagset and sentence-category relation tagset for the treebank are presented also with the example tree of CSCDT.

Key words Chinese treebank; hierarchical network of concepts; sentence-category dependency treebank

自然语言理解与处理的基本资源包括各类术语词典、多语词典、配价词典(如德语配价词典 BAL-BU)、同义词词典(如 WordNet 和汉语同义词词林)、语义本体(如 FrameNet 和 HowNet)、领域本体、句法树库(如宾州树库)等等。句法树库(treebank)是句法树的集合, 树结构是一种被广泛应用于语言学、音乐和视觉感知、推理、学习和问题解决等领域的知识描述方法, 句法树也成为句子最直观的形式描述模型, 是机器自动学习自然语言句法知识的源泉。句法树的质量直接影响机器自动学习句法知识的效率, 因此句法树以及以句法树为基本元素的树库是自然语言理解与处理的重要资源。

目前国内外的树库基本上可以分两大类: 一类是主要呈现句法信息的树库, 如英国的 Lancaster-Leeds 树库、美国的宾州树库; 还有一类主要呈现词语之间的语义支配关系, 如德国的 Tiger 树库和捷

克的布拉格依存树库 PDT。另外 Fillmore 主导建设的框架网络 FrameNet 中, 每个框架都配有若干经过句法语义分析的例句, 从广义角度讲 FrameNet 也可以看做某种树库。

1 汉语树库建设综述

在汉语的自动理解以及中文信息处理领域(如信息检索、机器翻译)中, 汉语树库也发挥着重要作用^[1]。目前国内外比较有影响的汉语树库主要包括: 宾州大学汉语树库 Penn Chinese Treebank (Penn-CTB)、中国台湾中央研究院汉语树库 Sinica、清华大学树库 Tsinghua Chinese Treebank (TCT)、北京大学汉语树库、哈尔滨工业大学信息检索研究中心汉语依存树库 Chinese Dependency Treebank of HIT-IR (HIT-IR-CDT)以及山西大学汉语框架语义知识库 Chinese FrameNet (CFN)。以上汉语树库均有各自的

国家社会科学基金(10BYY009)资助

收稿日期: 2012-06-10; 修回日期: 2012-08-15; 网络出版时间: 2012-10-26 17:04

网络出版地址: <http://www.cnki.net/kcms/detail/11.2442.N.20121026.1704.003.html>

标注体系以及标注特点, 树库包含的标注信息决定了树库已开发的以及潜在的应用领域。

从句法标注层面来看, PennCTB 以乔姆斯基的短语结构语法为理论基础, 标注了句子的层次关系、短语的结构类型、功能类型以及词语的词类^[2]。采用的词类标注集共包含 33 种词类标记符, 将动词、形容词作为一个大类(共 4 小类: VA, VC, VE, VV), 对于谓词未做详细分类, 对于小品词的分类却非常细致(共 9 类: DEC, DEG, DER, DEV, SP, AS, ETS, SP, MSP)。这种分类体系脱胎于宾州大学英语树库词类标注集, 基本上沿用了基于印欧语系的词类体系。中国台湾中研院语言所和资讯所联合建设的 Sinica 树库以信息为本的格语法为理论基础, 兼顾了句法和语义信息, 不仅标注了句子的层次关系、短语的结构类型以及词类, 还标注了名词短语的语义角色^[3]。Sinica 树库的词类标注集由 46 个词类标记符组成, 与宾州汉语树库相比, 词类划分更为细致, 尤其是对动词的次类进行了细分, 仅动词就有 17 个标记符, 而相对简化了助词的次类划分。

北京大学计算语言学研究所的北京大学汉语句法树库^[4]与清华大学 TCT 汉语句法树库^[5]二者一脉相承, 均以汉语传统的层次分析法为理论基础来标注句子层次, 采用了相对较小的词类标注集, 并在词类标注的基础上对直接成分之间的句法关系进行了标注。其中北京大学汉语树库的句法关系标注集包含了 19 个句法关系标记符, 清华大学汉语树库包含了 26 个词类标记符及 27 个句法关系标记符(如 ZW 表示主谓结构), 并且还设计了一个包含 16 个成分标记的短语类型和句子类型的标注集(如 zj 表示整句)。国家语言文字工作委员会的现代汉语语料库也基本上采用了相似的构建思路^[6]。哈尔滨工业大学汉语依存句法树库^[7]以依存语法为理论基础, 采用了国家语言文字工作委员会 863 项目开发的词类标注集(由 23 个词类标记符组成)^[8], 并开发了一个包含 24 种句法依存关系的关系标注集(如主谓关系 SBV)。虽然依存语法本身比较重视语义层面的信息, 但目前的依存树库仍然以句法关系为主, 深层次的语义信息仅仅是树库句法知识的补充。

从语义标注层面来看, 语义标注的类型可以分为语义角色(semantic role)标注和语义框架(semantic frame)标注。Sinica 树库将标注重点放在语义角色(该树库称之为 thematic role)上, 共设计了 63 个语义角色, 如 agent (主事者)、theme (客体)、location

(地点)等^[3]。目前, 随着语言信息处理任务对于语义知识标注的要求不断提高, 语义角色自动标注方法研究也成为热点研究领域^[9]。山西大学 CFN 则由框架库、句子库和词元库 3 个子库构成^[10-11], 其句子库相比 Sinica 树库而言, 句子的语义信息更细化、更丰富, 对于动词标注了动词所属的框架, 对于动词以外的短语则标注了其框架元素类别(如自主感知者 perc_agt)、短语结构类型(如名词性短语 np)以及句法功能(如主语 subj)。

现有汉语树库无论在句法标注还是语义标注层面, 都仍然存在局限, 句法标注层面往往受制于汉语对于西方词类体系的不适应, 语义层面的标注目前主要还停留在语义表层, “由于语义角色类型有限, 忽略了语言表达中的细节, 实用价值受到限制”^[10]。CFN 的标注重点在于语义框架, 对于句子的标注则仅限于短语级别, 没有深入到词一级, 因此在机器翻译领域的应用受限。Boas^[12]探讨了德语和英语的双语 FrameNet 在机器翻译领域的应用, 双语 FrameNet 的利用价值暂时局限于用于消歧的双语词典, 其价值在句子转换层面仍未得到体现。本文拟构建一个面向机器翻译领域的汉语句类依存树库, 重点标注词语级的语义信息和句子级的语义依存信息。

2 汉语句类依存树库的构建理论探讨

本节拟在概念层次网络(hierarchical network of concepts, HNC)理论框架下^[13-15], 研究与构建汉语句类依存树库相关的理论问题。

2.1 研究意义与目标

首先, 在理论层面利用句类依存树深化汉语句法语义分析。以往的汉语树库在句法、语义的描写方面各有侧重, 采用的句法分析理论和方法各不相同, 但是其共同点在于, 以往分析都基于现有的词类体系, 该词类体系是从印欧语系引进的舶来品, 对于“词无定类, 离句无品”的现代汉语而言, 由于现代汉语语法体系中词类和句法成分之间不存在一一对应的关系, 导致兼类词和词类歧义偏多, 我们尝试寻找一种面向中文信息处理的更适用于汉语的词分类及句子分析方法, 并在此基础上构建一个更为本土化的汉语树库。

其次, 尝试解决汉语多动词句的句法语义分析难点。汉语多动词句是自然语言理解与处理的难点之一, 汉语动词不带形态标记导致计算机难以准确

判断各动词之间的主从关系,而以往树库对于多动词句的分析并没有突出其特点,以多动词结构“显性轻动词+V”为例,以往树库对于显性轻动词(如“进行”)的词性标注或该结构的句法语义分析,都与语言事实存在较大差距。国家语言文字工作委员会制定的“信息处理用现代汉语词类标记集规范”^[8]将“进行”类动词处理为“形式动词”,区别于一般动词,符合语言学家们的判断,但该标记集并没有在中文信息处理领域获得广泛应用。而宾州树库将“进行”标注为 VV (一般动词), Sinica 树库标为 VC (动作及物动词), 北京大学树库标为 v (一般动词), 均没有将这类在句法和语义上有别于一般动词的显性轻动词做特别标注,这必然导致该类结构的句法语义分析不够准确。句类依存树库在构建过程中将重点关注汉语多动词句的标注规范与技巧,以使之体现和保持汉语本色。

最后,在应用层面深化 HNC 句类理论在机器翻译领域的应用。“目前基于句法的统计翻译方法开始成为研究的热点,特别是在基于语言学句法的翻译模型中,句法知识得到了充分的利用,已经开始超越单纯基于短语的方法或者基于形式化句法的方法”^[16]。但目前基于句法的机器翻译研究仍然依赖于词类体系,而词形和词类都具有较强的语言个性,但是与词形词类相比,概念则具有跨语言共性,可以成为翻译过程中的中介体系。众所周知,语义范畴具有相对性,概念体系的具象化亦非常难,我们仅从语义知识颗粒度相对较粗的概念类别出发,构建基于词的语义类的句类依存树库,为基于句法的汉英机器翻译定义一种新的双语转换模板,探讨该句类依存树库在汉英机器翻译领域的应用价值。

2.2 树库构建的理论背景

HNC 理论是关于人类语言认知机制的理论,也是面向计算机的自然语言理解理论体系,该理论以概念类别为基础,而非词类,根据句子核心动词的语义类别对句子进行分类(即句类),更适合于意合的汉语。

2.2.1 HNC 三大语义网络

HNC 对于概念之间层次性和关联性的描述体现在三大语义网络中:基本概念、基元概念以及逻辑概念语义网络。基本概念语义网络包括时间、空间等基本概念;基元概念语义网络的构建以作用效应链为基础,包含作用、过程、转移、效应、关系及状态 6 个主体基元概念子网络,是 HNC 句类分析的基础;逻辑概念语义网络包括语言逻辑概念(如语义块标识符、句间逻辑说明符等)和基本逻辑概念(如比较、判断等概念)。HNC 语义网络具有概念化、基元化、层次化和网络化等 4 个基本特征,是概念联想脉络的线索^[14]。本节以基元概念语义网络中的关系子网络为例,简要介绍 HNC 语义网络的构成。

关系子网络由 7 个概念节点构成(这 7 个概念节点从哲学角度考虑,未必是完备的),如图 1 所示,每个概念节点都包括对偶的一组概念,如 1 号节点为“结合与分离”,2 号节点为“依存与排斥”;每个概念节点又下辖若干个子概念节点,如 3 号节点下辖 4 个子概念节点:“调和”、“支持”、“反对”和“中立”;每个概念子节点下辖若干个终极节点——词语,如“支持”节点下辖“支持”、“帮助”等 185 个词语,一个概念节点可能下辖上百个词语,一个词语也可能与两个以上的概念节点相关联,从而构成一个有层级的语义网络。

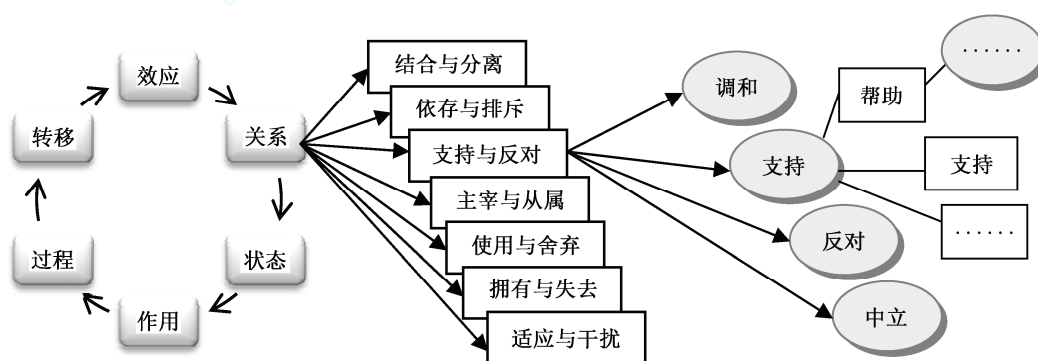


图 1 HNC 基元概念语义网络

Fig.1 HNC Semantic Network of Primitive Concepts

2.2.2 HNC 句类分析体系

句类由语义块构成,是句子的语义类型,HNC 句类划分的标准是“作用效应链+判断”,它们表述事物的基本侧面及句子的基本语义信息,一共有作用句、过程句、转移句、效应句、关系句、状态句和判断句,共 7 大句类、57 组基本句类,基本句类彼此之间又可以形成混合句类,依靠 57 组基本句类及其混合句类可以穷尽表达自然语言中的所有句子,且所有句类都可以利用句类表达式进行形式化表达^[14]。

以“中国今天公布了打击走私的巨大成果。”为例,句子的特征语义块(主要动词)为“公布”,是整个句子的支配者,该句的句类为 T3(信息转移)与 Y30(效应中的显隐)的混合句:

T3J=转移者 TA+信息转移 T3+
转移对象 TB+转移内容 T3C,
Y30J=显隐的对象 YB+效应中的
显隐 Y30+效应内容 YC。

“中国”为转移者 TA;“今天”为 j1(时间概念),充当条件辅语义块 Cn;“了”为 hv(特征语义块的后附加成分,也称为特征语义块核心的下装);语义块“打击走私的巨大成果”为效应内容 YC。如下所示:

中国 今天 公布了 打击走私的巨大成果 。
TA Cn T3Y30 YC

T3J 与 Y30J 两个基本句类复合之后,形成该例句的句类表达式为 T3Y30*21J=TA+T3Y30+YC。该句类表达式可以预测句子的主要语义类别以及包含的主语义块数目和类型等相关句类知识,“*21”表示复合句类从基本句类 T3J 和 Y30J 中选取的语义块数量和顺序。

2.3 句类依存树库与现有树库的关系

国内外不少研究者都尝试了将短语结构树库转化为依存结构树库,并且取得不错的成果。目前,汉语的短语结构树库相对规模较大,也较为成熟,而汉语依存树库的规模相对较小,李正华等^[17]尝试利用规则和统计相结合的方法,经过 3 个步骤:词类标注集转化(准确率 98%)、短语结构转依存结构、确定依存弧的关系类型(准确率 93%),将 PennCTB 短语结构树库转化为哈尔滨工业大学依存树库的体系结构,以扩大汉语依存树库的规模并用于训练和测试依存句法分析器的性能。从以往研究转化的过程和结果来看,在具有一定规模句类依存树及相应

标注规范基础上,将现有的短语结构树或句法依存树自动转化为句类依存树也成为可能,这将为大规模句类依存树库的构建奠定基础。

3 汉语句类依存树库的构建实践

标注树库需要完善的标注体系和规范的标注流程,从而保证标注语料的高质量 and 一致性。

3.1 概念类别标注集

用于标注句类依存树库的共有两个标注集,其一为概念类别标注集,其二为句类关系标注集。概念类别标注集基于 HNC 语义概念体系而建立,用于表示词语的主要概念类别意义,如“普京”标注为 fp(人名)等。我们构建句类依存树库时主要利用了已建成的 HNC 词语知识库资源,该知识库中一共标注了 53000 多个词形,每个词形可能具有多个语义类别,如“后院”具有两个概念类别:wj01(具体空间)和 pj01(社会空间)。

HNC 词语知识库中每个词形的各语义类别都分别标注了该词语的概念类别、HNC 符号、句类代码、语句格式和句类知识等,目前在 HNC 句类自动分析系统中仅仅利用了词语的概念类别知识,还无法对 HNC 符号进行自动解读。我们在构建句类依存树库时主要利用了词语知识库中的概念类别、句类代码等知识,并对其中的概念类别进行了规范统一处理。

概念类别标注集分为以下 11 大类。

v 类概念(动态概念): v; Bv; vB; Cv; vC; vv; vk; hv; qv
g 类概念(静态概念): g; xg
u 类概念(属性概念): u; vu; uv; gu; ug; uu; jx; px; jlu
z 类概念(值概念): z; zz; zzv
r 类概念(效应概念): r
p 类概念(人): p(-); pr; pe; lp; fgp; ff3; fpe; fp
w 类概念(物): w; w9; pw(-); pw5(-); gw; rw; rvw; frw; jgw(-)
jw 类概念(基本物): jw0; jw-; jw1; jw2; jw3; jw4; jw5; jw60(-); jw61(-); jw62(-); jw63-
j 类概念(基本概念): j0; j1; wj1; pj1; j2; wj2; j01; wj01; pj01; pj2(-); pj5; j3; j4; j5; j6
l 类概念(语言逻辑概念): l0; l1; l2; l3; l4; l5; l6; l7; l8; l9; la; lb
f 类概念(语习概念): f1; f2; f3; f4; f5; f6; f7; f8; f9; fa; fb; s; J; pun

例 1 中国/pj2 今天/j1 公布/v 了/hv 打击/v 走私/v 的/l42 巨大/u 成果/r 。/pun

例 2 此外/lb , /pun 委员会/pe 还/uv 相继/uv 派/v 团/pe 赴/v 台/pj2- 访问/v , /pun 与/l02 台湾/pj2- 工商界/pj01 进行/vv 了/hv 广泛/u 的 /l42 接触/v 和/l41 交流/v 。/pun

3.2 句类关系标注集

句类关系标注集用于标注词语之间的语义依存关系，与以往树库关系类标注集的最大区别在于我们的句类关系标注集不是封闭集合，是以 HNC 句类分析理论为支撑的开放集。仍以“中国今天公布了打击走私的巨大成果。”为例(图 2)，具体说明句类关系标注集类别。

句类关系标注集分为以下 2 大类 6 小类。

1) 句类核心成分。与支配者的句类关系利用句类表达式中的符号表示。包括以下 2 小类。

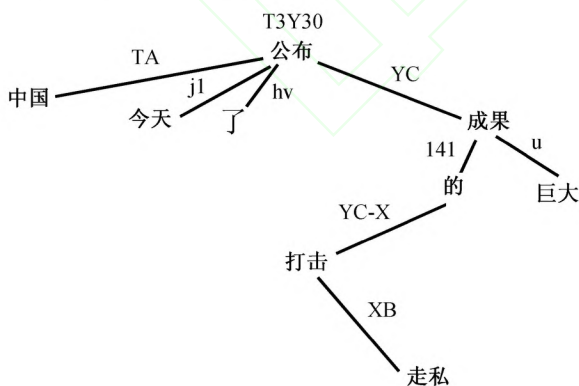
① v 类概念自身的语义类。包括充当全局特征语义块的 v 概念以及充当局部特征语义块(块扩、句蜕的核心成分)的 v 概念，如图 2 中全局特征语义块“公布”，其语义类别为 T3Y30(信息转移与效应)，局部特征语义块“打击”的语义类别为 X(基本作用)。

② v 与相关广义对象语义块之间的关系。如图 2 中“中国”与其支配者“公布”的关系类别为 TA(转移者)，“成果”与其支配者“公布”的关系类别为 YC(效应的内容)，“走私”与其支配者“打击”的关系类别为 XB(作用的对象)。

2) 句类非核心成分。与支配者的句类关系利用其本身的概念类别符号表示。有以下 4 小类。

① 特征语义块 Ek 的复合构成。如图 2 中结构助词“了”为特征语义块的后附加成分，也称为 Ek 的下装，关系类别用“了”本身的概念类别 hv 表示。

② 时间、地点、方式、工具、参照等辅语义块。如图 2 中的“今天”概念类别为时间概念，其与支配者的关系类别用概念类别 j1(时间)表示。



T3Y30: 信息转移与效应; TA: 转移者; j1: 时间概念; hv: 特征语义块的后附加成分; YC: 效应的内容; 141: 语义块内部的偏正组合概念; u: 属性; YC-X: 效应的内容中的基本作用概念; XB: 作用的对象

图 2 句类依存关系

Fig 2 Dependent Relationship based on Sentence-Category

③ 语义块核心成分的修饰性成分。包括传统的定语及补语，如图 2 中的属性概念“巨大”，其与支配者“成果”的句类关系利用概念类别 u(属性)表示。

④ 句子的附加成分。主要包括连接词、插入语等成分，如图 2 中的结构助词“的”，其与支配者“成果”的句类关系利用概念类别 141(语义概念 14 中的一类，语义块内部的偏正组合)表示。

3.3 可视化句类依存树

本文在现有研究的基础上，利用 xml 语言构建了基于 HNC 句类理论和 HNC 词语知识库的汉语句类依存树库。依存树的可视化方式基本上沿用哈尔滨工业大学依存树库的可视化方法^[7]，样例如图 3 所示。

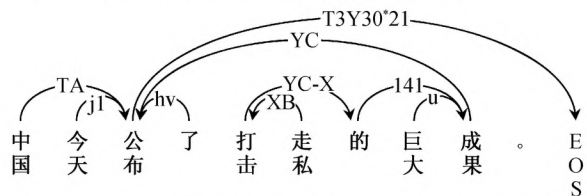
句类依存树上标注的句法和语义信息包括：句子的句类(句子的主要语义类别)与语句格式(语义块排列的表层顺序)，句子中每个词语的概念类别，每一个词语的父节点(支配者)，词语之间的有向弧表示了相关词语之间的语义支配关系，有向弧上的标记表示词语之间的句类关系。在可视化句类依存树上仅显示了语义支配关系及句类关系。

3.4 标注方法与树库应用

本研究开发了基于 HNC 词语知识库和同义词词林的句类依存树辅助标注工具，在构建树库过程中主要采用半自动化标注方法，从 HNC 词语知识库和同义词词林中自动查找词语的概念类别。当有多个语义类别时人工指定其中一个合适的语义类别，并人工标注词语之间的语义支配关系和句类关系。本树库目前的标注规模为试验性的 2000 个汉语复杂句子，语料均为政论语体。

各机器翻译系统采用的双语表征方式不同，由此得到的翻译模板也各有千秋，以往无论是基于实例的还是基于统计的机器翻译方法，主要使用基于非结构化句法的，基于短语结构树的，基于句法依

— 中国 今天 公布 了 打击 走私 的 巨大 成果 。



T3Y30*21: 信息转移与效应复合句类(21 表示句类的语义块选择方式); EOS: 句末标记。其余符号同图 2

图 3 可视化句类依存树

Fig 3 Visualization of Sentence-Category Dependency Tree

存树的,以及基于语义框架或语义特征的翻译模板几大类。本树库面向的应用领域主要为汉英机器翻译系统,主要研究目标为基于句类依存树库定义面向汉英机器翻译的“句类依存树到串”的语义转换模板,并尝试从经标注的汉英双语平行语料(汉语语料标注为句类依存树)中提取基于句类依存子树的双语语义转换模板库,尤其是针对汉语多动词句的转换模板库,期望为汉英机器翻译提供一种包含更多语义信息的模板。

4 结语

本文分析了现有汉语树库在句法和语义标注方面的特点,以机器翻译领域为应用目标,以概念层次网络理论的语义网络和句类分析方法为理论基础,探讨了句类依存树库构建的理论和标注实践等问题,详细描述了概念类别标注集和句类关系标注集。句类依存树库的构建并不是标新立异,我们希望这种新的树库模式能够为中文信息处理提供另一种视野和思路。当然,树库的构建是一项非常复杂的工程,树库标注一方面需要标注者具备专业的理论知识以及对于现代汉语的分析能力,另一方面,标注的准确性、一致性问题都需要重点关注。

参考文献

- [1] 王跃龙,姬东鸿. 汉语树库综述. 当代语言学, 2009, 11(1): 47-55
- [2] Xue N, Xia F, Chiou F D, et al. The Penn Chinese Treebank: phrase structure annotation of a large corpus. *Natural Language Engineering*, 2004, 10 (4): 1-30
- [3] 陈凤仪,蔡碧芳,陈克健,等. 中文句结构树资料库(Sinica Treebank)的构建. *Computational Linguistics and Chinese Language Processing*, 1999, 4(2): 87-104
- [4] 周强,张伟,俞士汶. 汉语树库的构建. 中文信息学报, 1997, 11(4): 42-51
- [5] 周强. 汉语句法树库标注体系. 中文信息学报, 2004, 18(4): 1-8
- [6] 靳光瑾,肖航,富丽,等. 现代汉语语料库建设及深加工. 语言文字应用, 200(2): 111-120
- [7] Liu Ting, Ma Jinshan, Li Sheng. Building a dependency treebank for improving Chinese parser. *Journal of Chinese Language and Computing*, 2006, 16(4): 207-224
- [8] 国家语言文字工作委员会语言文字应用研究所计算语言学研究室. 信息处理用现代汉语词类标记集规范. 语言文字应用, 2001(3): 16-20
- [9] 李业刚,孙福振,李鉴柏,等. 语义角色标注研究综述. 山东理工大学学报: 自然科学版, 2011, 25(6): 19-24
- [10] 刘开瑛,由丽萍. 汉语框架语义知识库构建工程. 中文信息处理前沿进展 // 中国中文信息学会二十五周年学术会议论文集. 北京, 2006: 64-71
- [11] 刘开瑛. 汉语框架语义网构建及其应用技术研究. 中文信息学报, 2011, 25(6): 46-52
- [12] Boas H C. Bilingual FrameNet dictionaries for machine translation // *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*. Las Palmas de Gran Canaria, Spain, 2002: 1364-1371
- [13] 黄曾阳. HNC(概念层次网络)理论: 计算机理解自然语言的新思路. 北京: 清华大学出版社, 1998
- [14] 苗传江. HNC 理论导论. 北京: 清华大学出版社, 2005: 300-315
- [15] 张克亮,黄曾阳. HNC 作用效应句的汉英句类转换. 中文信息学报, 2003, 17(5): 19-26
- [16] 刘群. 机器翻译研究新进展. 当代语言学, 2009, 11(2): 147-158
- [17] 李正华,车万翔,刘挺. 短语结构树库向依存结构树库转化研究. 中文信息学报, 2008, 22(6): 14-19