HNC
概念层次网络

基于HNC的现代汉语句子级语义标注语料库的研究和建立

RESEARCH ON SENTENCE CORPUS WITH SEMANTIC ANNOTATION BASED-ON HNC THEORY

刘智颖◎著

自然语言处理资源建设

中国社会科学出版社

HNC
概念层次网络

基于HNC的现代汉语句子级语义标注语料库的研究和建立

RESEARCH ON SENTENCE CORPUS WITH SEMANTIC ANNOTATION BASED-ON HNC THEORY

自然语言处理资源建设

中国社会科学出版社

图书在版编目(CIP)数据

基于HNC的现代汉语句子级语义标注语料库的研究和建立／刘智颖著．
北京：中国社会科学出版社，2015.2
ISBN 978-7-5161-5576-9

Ⅰ.①基… Ⅱ.①刘… Ⅲ.①汉语-语料库-研究 Ⅳ.①H1

中国版本图书馆CIP数据核字(2015)第032779号

出 版 人 赵剑英
责任编辑 任 明
特约编辑 付 钢
责任校对 季 静
责任印制 何 艳

出 版 中国社会科学出版社
社 址 北京鼓楼西大街甲158号（邮编100720）
网 址 http：//www.csspw.cn
中文域名：中国社科网 010-64070619
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印刷装订 北京市兴怀印刷厂
版 次 2015年2月第1版
印 次 2015年2月第1次印刷

开 本 710×1000 1/16
印 张 12
插 页 2
字 数 203千字
定 价 55.00元

凡购买中国社会科学出版社图书，如有质量问题请与本社联系调换
电话：010-84083683

前　言

本书针对自然语言处理中资源建设方面的不足，研究建立句子级语义标注语料库。从语义的角度、句子的层面对真实文本语料进行自上而下的加工标注，最终形成具有一定规模的基于 HNC 理论的现代汉语句子级语义标注语料库。

基于 HNC 的现代汉语句子级语义标注语料库是以 HNC（Hierarchical Network of Concepts，概念层次网络）理论为指导，以句子为标注的基本单位，为连续文本语料标注语义框架信息的语料库。为语料库中的连续文本标注的信息所提供的丰富的语义知识，不仅对于计算机理解语义，而且对于语言学工作者也是一个非常宝贵的资源。

本书在 HNC 理论句类体系的基础上，重点研究了句子级语义标注语料库的标注内容、标注方法和标注难点，确定了 XML（Extensible Markup Language，可扩展标记语言）的标注规范，进行了语料库查询工具的功能设计。研究内容主要有以下几个方面：

（1）确定了语料库的标注内容，在 HNC 理论的指导下，以句子作为标注的基本单位，分别从语言空间和语言概念空间进行结构和语义两方面的标注；

（2）分析了语料库的标注难点，探讨了语义块核心要素部分的缺省问题、包装成分和分离成分的判定问题、的字短语和所字结构的标注问题、与句式有关的句类判定问题等标注难点，给出了自己的标注方案；

（3）建立了语料库的 XML 标注规范。从篇章段落、句子、语义块、块素等层级确定了 XML 标注元素及属性；对元素和属性进行了 Schema 模式定义，保证了文档的有效性和良构性；用 XSL 进行 XML 语义标注语料库的结构转换，使 XML 语义标注语料库呈现出用户所需要的表现形式；

（4）设计了语料库的管理工具，为语料库使用者提供相应的语料库管理及查询功能，方便用户从语料库中检索到自己想要的信息，最大限度

地发挥语料库的使用价值。

句子级语义标注语料库的建设填补了中文信息处理资源建设的一项空白，它不仅对 HNC 理论的学习和 HNC 句类分析系统的完善具有重要意义，而且也可以为广大的语言学工作者进行语言研究提供帮助。该语料库的建设对整个中文信息处理乃至语言教学与研究都具有重要的意义。

本书对语义标注语料库的构建研究是基于作者对 HNC 理论和语料库的理解，由于本人水平有限，难免存在错误或不当之处，敬请读者批评指正。

目　录

图表目录

第一章　绪论

一直以来，中文信息处理主要集中在词法、句法结构的分析和处理上。随着中文信息处理研究的深入，计算机专家和语言学家越来越体会到，语义在中文信息处理中起着极为重要的作用。认识到中文信息处理要想有所突破，必须关注语义问题。

语言单位上，人们对字和词的处理技术已经比较成熟，在汉字输入与输出、自动分词等方面取得了重大突破，中文信息处理开始向更高的阶段迈进，开始关注对语句的处理。可以说，现在中文信息处理的重心已由词处理阶段向句处理阶段过渡。

第一节　中文信息处理的研究热点

所谓“句处理”，可以理解为：怎么让计算机处理、理解自然语言中一个句子的意义，怎么让计算机生成一个符合自然语言规则的句子。（陆俭明，2003）

句处理的难度远远超过字处理和词处理，因为句处理不仅面临着复杂的句法规则问题，更面临着复杂的语义、知识背景、文化背景等一系列问题。现在，在句处理方面已形成多种处理、理解汉语句子的策略和方案——一是以理性主义为哲学基础的基于规则的处理方法，这种方法或是以一定的形式文法系统来表述自然语言中大小成分间的组合规则，或是“以概念化、层次化、网络化（简称‘HNC’）为基础”来提供概念组合、语义表述的规则；二是以经验主义为哲学基础的基于语料库统计方法，这种方法是以各种统计数据来显示语言成分间的组合可能性。为解决汉语句处理问题，已开发了好几个不同规模的汉语语料库，不同类型、不同规模的现代汉语信息词典。

单纯的以统计方法从语料库中来获得知识，并不能从根本上实现计算

机对语言的理解。如何做到让计算机理解语言呢？黄曾阳先生创立的概念层次网络理论（简称 HNC 理论）是关于自然语言处理的理论，其目标是以概念联想脉络为主线，建立一种模拟大脑语言感知过程的自然语言表述模式和计算机理解处理模式，使计算机获得消解模糊的能力。目前 HNC 理论发展日渐成熟，而且已经在实际应用中被证明确实是处理自然语言的一种行之有效的理论。本书的研究就是基于 HNC 理论所进行的。

当前中文信息处理句处理的现状是：服务于句处理的资源建设刚刚起步，而为句处理所提供的语言知识则很欠缺，尤其是语义知识的提供更是一个空白。语料库的建设及应用研究已经成为当务之急。我们的目标是建立一个以句子为标注的基本单位的语义标注语料库。为语料库中的连续文本标注一定的语义信息，将语义信息赋予语料库。这样，计算机可以通过对语料中语义信息的提取达到对语句的理解。本书以 HNC 理论为指导，以句子作为标注的基本单位，为连续文本语料标注语义框架信息，这些标注信息不仅显示了句子的语义结构框架，而且还显示了句子的语义类型，语义块（句子的下一级语义单位）的构成和分离情况等。本语料库的建成不仅对于计算机理解语义，而且对于语言学工作者也是一个非常宝贵的资源。可以说，句子级语义标注语料库的建设填补了中文信息处理资源建设的一项空白，具有开创性意义。

第二节　基于 HNC 的现代汉语句子级语义标注语料库

HNC 理论（黄曾阳，2004）认为，自然语言理解是一个从自然语言空间到语言概念空间的映射过程，两个空间各有自己的一套符号体系。语言空间符号体系千差万别，但语言概念空间符号体系全人类只有一个。语言空间依托于语音和文字而存在，语言概念空间依托于概念联想脉络而存在。HNC 设计了对自然语言概念体系进行总体表述的语义网络，并以此为基础，建立了自然语言语句的语义表述模式，构造了句子的语义结构表示式，这些表示式是对语句层面概念联想脉络的形式化表达。语句无限而语句的概念类型（句类）有限，自然无限的句子都可以用有限的句类表示式表示出来。句类是从语言深层对自然语言语句的语义类型所做的分类。

基于 HNC 的现代汉语句子级语义标注语料库（以下简称 HNC 语义标

注语料库）正是以 HNC 理论的句类思想为指导而建立起来的，以句子作为标注的基本单位，对语料进行语义信息标注的语料库。语料标注采用自上而下的标注方式，先标注大的语言单位，再标注小的语言单位。对于连续文本语料来说，先标注篇章、段落，再标注句子，然后是语义块，最后标注词语。HNC 语料库选择以句子作为标注的基本单位，在句子级对语料进行语义标注，因其既是语言理解和表达的基本单位，也是计算机处理自然语言的基本单位。这种标注方式与自下而上的语料标注方式相辅相成，满足了语言本体研究和语言信息处理研究的不同需要。

本语料库采用 XML 作为标注形式。XML（eXtensible Markup Language）是可扩展标记语言的简称，它是一组定义语义标记的规则，这些规则将文档划分为多个部分，并且对文档的不同部分作出标记。XML 非常灵活，我们可以自己定义这些标记及语法结构。通过 XML 的元素和属性对语料进行标注。每个语义单位都由一个特定的元素进行标记，语义知识也通过属性值对进行描述，元素内部可再嵌套元素，这就形成了一个树形结构，显示具有层次性的特点。比如，我们定义 s 元素表示句子（sentence），jk 元素表示广义对象语义块，ek 元素表示特征语义块；每个元素有开始标记和结束标记。句子的句类代码和语句格式用属性/属性值对来表示。方便易读，使句子的语义信息一目了然，一般的语言工作者也能很快理解和掌握。

例如：对于“这是民国六年的冬天”这个句子，我们标注如下：

```
<s code = " jD" >
  <jk type = " 1" >这</jk>
  <ek>是</ek>
  <jk type = " 2" >民国六年的冬天</jk>
</s>
```

全句是一个简单句，我们用 s 元素表示；句子的语义类型（句类）为是否判断句，我们用 s 元素内的属性值对 code = " jD" 表示。此句包含两个广义对象语义块：“这”和“民国六年的冬天”，我们用 jk 表示；广义对象语义块在语义表示式中的顺序我们用 type 属性来表示；句子还包含一个特征语义块“是”，我们用 ek 表示。

可以看出，用 XML 标注语料库非常灵活，易标易读，尤其便于计算机对语义信息的提取，是语料标注适合的标注形式。

第三节　本书的研究内容

对自然语言的理解处理最终要靠语义，本书研究的句子级标注语料库以语义为主导，依据 HNC 理论对句子的层次划分方法对语料进行标注。标注的顺序采用自上而下的顺序，从语言层面看，首先标注篇章、段落，再往下一级标注句子，再往下一级是语义块。

一　标注项

以 HNC 理论模式为指导来对句子的语义进行标注，标注项主要有四个：（1）句类，句子的语义类型；（2）语义块，句子的下一级语义构成成分；（3）句蜕，包含在语义块内的子句或其变形；（4）语义成分的共享，包括句与句之间存在的语义上的关联以及句间语义成分的共享等信息。

（一）句类

HNC 定义的句类是指句子的语义类型。HNC 对句类进行了 2-8-57 的划分，首先将句类分为广义作用句和广义效应句两种类型，再进一步下分为 8 种类型，即广义作用句又分为作用句、转移句、关系句和一般判断句四种类型，广义效应句又分为过程句、效应句、状态句和基本判断句四种类型。对 8 种类型又划分了它们的子类，共有 57 种，称为基本句类。这 57 种基本句类是句子语义的基元类型，可以用它们来描述任何句子的语义类型。自然语言中一个句子的语义类型，可能是某一种基本句类，也可能是某两种或多种基本句类的组合——混合句类。例如：

（1）美国要攻打伊拉克。（基本作用句 XJ）

（2）中小型商店难以对抗大型连锁超市。（单向关系句 R310J）

（3）俄罗斯反对美英对伊拉克动武。（单向关系 + 主动反应句 X21R311 * 21J）

（二）语义块

HNC 定义的语义块是指句子的下一级语义构成成分。语义块可以是一个词，一个短语，也可以是一个子句或其变形，还可以是三者的组合。不同的句类需要配置不同的主语义块，例如，反应句需配置三个语义块，分别是反应、反应者、反应引发者及其表现；而信息转移句需配置四个语

义块，分别是信息转移、转移发出者、转移接收者和转移信息，如下面的例句所示（“丨丨”是语义块之间的分隔符，下同）：

（4）国际田联丨丨有条件同意丨丨贾亚辛格参加亚运会。（反应句X20J）

反应者丨丨　反应　丨丨　反应引发者及其表现

（5）一位芬兰同行丨丨兴奋地告诉丨丨记者，丨丨马哈鱼又回到了基米约奇河。（信息转移句T3J）

转移发出者丨丨信息转移丨丨接收者丨丨　转移信息

除了标注主语义块外，我们也标注辅语义块成分，辅语义块有7种：方式、工具、途径、比照、条件、因、果。

（三）句蜕

句蜕是指句子蜕化为语义块或语义块的一部分，也就是语义块中包含的句子。如下面的例句所示：

（6）俄罗斯丨丨反对丨丨｛美国丨攻打丨伊拉克｝。

（7）＜生产丨信息技术产品丨的工厂＞丨丨都转移到了丨丨国外。

（8）＜经济危机丨造成丨的后遗症＞丨丨也减轻了。

（9）这些话丨丨似乎表示了丨丨＜他丨对奴隶的生活境况丨的同情＞。

例句（6）的大括号内的语义块“美国攻打伊拉克”是原型句蜕，蜕化前后句子的基本形式没有变化。例句（7）（8）（9）中，尖括号内的语义块是要素句蜕，蜕化的方式是把句子的某一个语义块作为中心语，其他的语义块作为修饰语。

在语言理解中，不论是原型句蜕，还是要素句蜕，都应该作为句子来处理，要确定它的句类和各个语义块。因此，在我们的语料中，对句蜕都要作为句子来分析，要标明其句类和语义块。

句蜕如果与其他词语或短语再连接，就成为包装句蜕。如：

（10）我们的眼睛成为\　｛交流感情｝的工具/。

（11）能耗限制了\　＜计算机的运行＞速度/。

例句（10）中“交流感情”是原型句蜕，与词语“工具”连接，形成原型包装句蜕；例句（11）中“计算机的运行”是要素句蜕，与词语“速度”连接，形成要素包装句蜕。

（四）语义成分的共享

共享现象不仅出现在复句中，单句与单句之间，单句与句蜕（或块扩）之间都可以发生语义块的共享现象。语义块的共享是复句、单句、句蜕（或块扩）共同具有的语义特征。

要描述这一语义信息，我们需对如下几方面做出回答：哪个句子存在语义块的共享现象？它共享了另外句子的哪个语义块？如何表示共享语义块和被共享语义块之间的对应关系？这些都是本书要解决的问题。

二　标注方式

HNC 语义标注语料库在标注方式上从 2005 年开始进行了全面的革新。由原来的纯文本的线性标注更改为采用 XML（可扩展标记语言）格式进行标注。本书主体部分将全面介绍 XML 标注及检查规范。

采用 XML 语言进行标注，不仅可以更加清楚细致地表明语义成分及各部分的关系，而且 XML 其数据和显示形式相分离的特点，为数据的共享提供了可能。使 HNC 语义标注语料库可以为更多的人服务。

下面给出纯文本和 XML 标注样例，以突出 XML 标注的优越性。

(12a)！07T31Y3＊211J 由一位班主任｜｜教授｜｜［@ 葡语］，算术、体育和社会课。

(12b) <s code = " T31Y3＊211" form = "！07" >
<jk type = " 1" >由一位班主任 </jk >
<ek > 教授 </ek >
<jk type = " 2" > < word type = " 1" > 葡语 </word > ，算术、体育和社会课 </jk > 。
</s >

这两例中，例（12a）是 HNC 原有的语料标注形式，例（12b）是用 XML 进行标注的形式。采用 XML 的形式标注语料，使语料具有易读性、易检性、层次性、扩展性等特点和优势。

而且，利用 XML 的 Schema 来定义文档的模式，可以验证文档所包含的内容是否是形式规范的，以此来提醒标注者进行修改，达到自我检验的目的，提高了语料标注的质量。再有，通过 XSL 可扩展样式语言来定义文档的外观，可以使文档内容按照不同的需要呈现给读者。其结构及内容与显示形式分离的特点，便于阅读和信息共享与交换，使之成为实现语料

库标注和共享的理想工具。

三　管理工具功能设计

语料管理系统的建设一般包括数据维护（语料更新、存储、修改、删除及语料描述信息项目管理）、语料自动加工（分词、标注、文本分割、合并、标记处理等）、用户功能（查询、检索、统计、打印等）等几个方面。

本书重点设计系统的查询功能。根据所标注的语义信息，设计查询请求，力求详尽及最大限度地利用标注语料库，使之为语言研究和处理服务。如查询句类信息，语义块信息，句蜕信息等。

系统可以按照使用者不同的要求来查找语料，输出的是符合使用者要求的句子。查询主要分三个层面进行。第一层面，从句子一级设定查询条件。我们可以设定要查询句子的句类，语句格式等语义信息及句子所拥有的语义块数量等；第二层面，从语义块一级设定查询条件。我们可以设定语义块的构成类型，语义块中是否包含子句（包括句蜕和块扩两种），子句内是否还嵌套子句，语义块是否存在分离等。第三层面，从词语一级设定查询条件。我们可以查询到动态新词、伪词等。

第四节　已有的研究

语料库通常指为语言研究收集的、用电子形式保存的语言材料，由自然出现的书面语或口语的样本汇集而成，用来代表特定的语言或语言变体。经过科学选材和标注，具有适当规模的语料库能够反映和记录语言的实际使用情况。人们通过语料库观察和把握语言事实，分析和研究语言规律。语料库已经成为语言学理论研究、应用研究和语言工程不可缺少的基础资源。（傅爱平，2003）

语料库的应用方面。经过科学选材、具有适当规模的语料库能够反映和记录语言的实际使用情况，为语言学研究和应用提供统计数据和各种语言材料。对于计算语言学基于统计的研究方法来说，语料库的建设更是不可缺少的基础。目前我国已有多个百万字以上容量的汉语语料库和双语语料库，用于语言信息处理的各种研究和应用目的：汉字识别、智能汉字输入、文本自动分类、汉语自动分词、汉语人名地名自动识别、汉语关联词