文章编号: 1003-0077(2007)05-0096-06

汉语框架语义知识库及软件描述体系

郝晓燕[1],刘 伟[2],李 茹[2],刘开瑛[2]

(1. 太原理工大学 计算机与软件学院,山西 太原 030024;
2. 山西大学 计算机与信息技术学院,山西 太原 030006)

摘 要: 汉语框架网络工程是以框架语义学为理论基础的基于语料库的计算词典编纂工程,用于语言学、计算语言学研究及自然语言处理研究。该工程的结果包括两部分: 汉语框架语义知识库(即词典资源)和相关软件。其中,汉语框架网络知识库包括框架库、句子库和词元库三部分,相关软件主要包括汉语框架语义知识库管理系统和基于 Web 的展示系统。本文介绍了汉语框架语义知识库的语义描述体系以及软件描述体系。

关键词: 计算机应用;中文信息处理;汉语框架网络;框架语义;描述体系;软件
中图分类号: TP391 **文献标识码**: A

Description Systems of the Chinese FrameNet Database and Software Tools

HAO Xiao-yan[1], LIU Wei[2], LI Ru[2], LIU Kai-ying[2]

(1. Academe of Computer & Software Engineering, Taiyuan University of Technology, Taiyuan, Shanxi 030024, China; 2. School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract: The Chinese FrameNet project is producing a lexicon of Chinese for both human use and NLP applications, based on the principles of Fillmore's Frame Semantics. It includes two parts. One part is the Chinese FrameNet databse(CFN), which contains frames bank, sentences bank, and lexical unit bank. The other part is a suite of software tools related to the CFN, which includes the database management system and the Web-based demonstration system. The paper will give a brief introduction about the description systems of these two parts.

Key words: computer application; Chinese information processing; Chinese FrameNet; Frame Semantics; description system; software tools

1 引言

汉语框架网络工程是以框架语义学为理论基础的基于语料库的计算词典编纂工程,用于语言学、计算语言学研究及自然语言处理研究。该工程的结果包括两部分: 汉语框架语义知识库(Chinese FrameNet database,简称 CFN)[1](即词典资源)和相关软件。汉语框架语义知识库包括框架库、句子库和词元库三部分。相关软件主要包括汉语框架语义知识库管理系统和基于 Web 的展示系统。

汉语框架语义知识库(CFN)是一个以 Fillmore 的框架语义学[2,3]为理论基础、以加州大学伯克利分校的 FrameNet[4]为参照、以汉语真实语料为依据的供计算机使用的汉语词汇语义知识库,研究内容涉及语义知识库内容的编写、辅助软件的开发和应用研究等。

汉语框架语义知识库(CFN)由框架库、句子库和词元库三部分组成。目前,CFN 课题组已就汉语 1 760 个词元构建了 130 个框架,涉及动词词元 1 428 个、形容词词元 140 个、事件名词(即有配价的名词)词元 192 个,标注了 8 200 条句子;涉及认知领域用词、科普文章常用谓词以及部分中国法律用词。框架库以框架为单位,对词语进行分类描述,明

收稿日期: 2007-04-10 **定稿日期**: 2007-06-26
基金项目: 国家 863 高技术研究发展计划资助项目(2006AA01Z142)
作者简介: 郝晓燕(1970—),女,博士生,主要研究方向为自然语言处理;刘伟(1982—),男,硕士生,研究方向为自然语言处理;李茹(1963—),女,教授,研究方向为智能信息处理。

确给出框架的定义和这些词语共有的语义角色(框架元素),并进而描述该框架和其他框架之间的概念关系;句子库记录带有框架语义标注信息的句子,即按照框架库所提供的框架和框架元素类型,标注句子的框架语义信息和句法信息,它可以作为训练数据供计算机处理语言使用;词元库记录词元的语义搭配模式和框架元素的句法实现方式,它们是从句子库提供的标注结果中自动生成的。

2006 年 10 月 11 日,国内有关专家对"有限汉语框架语义知识库构建技术研究"课题进行了科技成果鉴定。与会专家认为,该课题"运用框架语义分析方法构建汉语框架语义知识库,开创性地研制了汉语框架、框架元素体系以及句子标注体系。……在信息处理用汉语框架语义研究领域中达到了国际领先水平。"课题组目前就有限范围的词语所取得的阶段性成果,为扩展领域的 CFN 构建提供了成熟的技术和样本。本文即对汉语框架语义知识库的三个部分及软件的描述体系分别总结说明,并提出后期工作展望。

2　汉语框架语义知识库(CFN)的语义描述体系

2.1　CFN框架库

"框架"(Frame)是跟一些激活性语境(Motivating Context)相一致的一个结构化的范畴系统,是储存在人类认知经验中的图式化情境;框架中的各种参与者称为框架元素(Frame Elementsc),它们在使用中与"语义角色"、"格角色"等概念相当。框架语义学认为,框架是词语理解的背景和动因,因此,可以根据背景框架的不同,对词语(主要是动词、形容词和事件名词)进行分类描述。传统的格语法的"格角色"是相对于所有词汇而言的,而框架元素仅适用于具有共同背景框架的一小组词语,摆脱了格语法难以确定的问题,具有个性特征的框架元素更适合用来描述自然语言语义。例如"波动、增加、提高、减少、降低"等汉语词语有共同的意义基础——都表示实体的某种属性从某个值变成另一个值,因此,汉语框架语义知识库(CFN)的框架库中用一个框架"量变"进行描述,该框架涉及的框架元素包括(括号中的字母是标记符号):ent(实体),att(属性),val1(初值),val2(终值),diff(变幅),val_ran(值区间)等等。

框架元素分为核心框架元素和非核心框架元素。核心框架元素是一个框架在概念理解上的必有成分,它们在不同的框架中类型和数量不同,显示出框架的个性,以上所列举的框架元素都属于"量变"框架的核心框架元素。非核心框架元素并不显示框架的个性,表达时间、空间、环境条件、原因、目的等外围语义成分。

框架库中每个框架都按照以下四方面进行描述:(1)框架的定义;(2)框架元素的基本定义以及部分框架元素的示例;(3)该框架所涉及的词元;(4)框架和框架之间的抽象关系。例如"波动、增加、提高、减少、降低"等汉语词语有共同的意义基础——都表示数量变化,归为一个框架进行描述,表 1 简略地展示了该框架的内容。

表 1　CFN框架库记录样例

框架名	量　　变	
定义	该框架表示实体在某个维度上(即某属性)的相对位置发生变化,其属性值从初值变至终值。	
核心框架元素	实体(Ent)	在某属性上具有一定量值的事物。
	属性(att)	实体的有数量变化的属性。▶西瓜经人工辅助授粉,坐果率可提高到 80%-90%。
	初值(val1)	实体的属性值变化的起点。
	终值(val2)	实体最后达到的量值。▶西瓜经人工辅助授粉,坐果率可提高到 80%-90%。
	初状态(inis)	实体经历属性值变化之前的状态。▶军事课程从原来占总学时的 15%增加到 20%至 25%。
	终状态(finis)	实体经历属性值的变化之后所达到的状态。▶厂里新建分支企业,使产品由一个普钙增加到拥有工业硅、氟硅酸钠、复合肥等 6 个品种。
	变幅(diff)	实体在某维度上变动的幅度。▶坐瓜率可增加约 20%。
	值区间(val_ran)	属性值的变动范围。▶今年头 8 个月,国际市场黄金价格一直在每盎司 380-387 美元之间波动。

续表

框架名	量变	
非核心框架元素	环境条件(cir)	事件发生时所处的事理、物理环境或所需要的条件。▶在遭受同样特大旱灾的情况下,牲畜的死亡率由70年代的10% 下降到1993年的3%以下。
	倚变因素(cor)	属性值往往随着某一因素的变化而增加或减少,这种因素或者表示出某种变化趋势,或者由一系列有序数值组成,还有可能仅仅是一个时间的推移。▶棉纱价格随着棉花调拨价上升而上升。
	动作时间量(dur)	量变持续的时间数量。
	倚变起点(cor1)	与初值相对应的时间、地点、状况等因素。▶从吃饱到玩好我国恩格尔系数下降。
	倚变终点(cor2)	和终值相对应的时间、地点、状况等因素。▶从吃饱到玩好我国恩格尔系数下降。
	修饰(manr)	该元素是一个杂类,它们无法归入以上各类非核心元素,笼统归入该元素。
	路径(path)	表示在起点和终点之间所经的变化方向。▶价格不断向上攀升。
	空间(place)	量变发生的地点。
	速度(speed)	属性值变化的速度。▶产量急剧下降。
	时间(time)	量变发生的时间。
框架关系	父框架	
	子框架	增殖
	总框架	
	分框架	
	后续过程	
	结果状态	数量
	参照	
词元	波动 v,增加 v,增长 v,提高 v,减少 v,降低 v,上升 v,攀升 v,升 v,增 v,下降 v,降 v	

表1中,框架和框架之间的关系主要包括继承关系(父框架与子框架)、总分关系(总框架与分框架)、因果关系和参照。其中,因果关系分为后续过程与结果状态:有些事件不仅需要动作发出者和动作对象,还需要有一个致使该动作发生的人或事物,这样的事件对应的框架就会导致另一个动作发生,称之为后续过程;动作发生以后,总会引起事物的状态发生变化,因此该动作又导致了某种状态的形成,称之为结果状态。如"提高"有两个义项:比原来高(如"产量提高了80%");使比原来高(如"新技术提高了产量")。这两个义项作为两个词元分别归入框架[量变]和[使量变]。框架[使量变]所表示的动作过程致使数量发生了变化(即[量变]);发生数量改变以后,形成了某一量度状态(即[数量])。"参照"严格地说并不代表现实中各个概念之间的关系,而是为了使用户更加准确地理解框架的含义,提示一些与原框架相似、容易引起混淆的框架。

2.2 CFN句子库

CFN句子库的句子标注语料来源于"121篇阅读理解文本"和"北京大学CCL现代汉语语料库"。我们为词元选取句子时,注意了选择不同的语义搭配模式,保证句子库的多样性。

CFN句子标注,是以框架库为基础,针对一个句子,确定一个词元和该词元所属框架,给框架元素所在的成分标记框架元素、短语类型和句法功能三种信息。例如,句子"军人由于受条件的限制,打电话发传真都不是太方便,同时他们也想通过写信提高自己的文化水平。"的标注结果如下(去掉了句子标号、分词、词性标注标记):

例1. 军人由于受条件的限制,打电话发传真都不是太方便,同时 < ent-np-subj_s 他们 > 也 < supp 想 > < mns-pp-adva 通过写信 > < tgt 提高 > < att-np-obj_s 自己的文化水平 > 。

tgt是目标词标记,目标词"提高"属于[量变]框

架;(ent)实体、(mns)方式、(supp)支撑词等是框架元素标记;np(名词短语)、pp(介词短语)等是短语类型标记;subj(主语)、subj_s(有支撑词支配的主语)、adva(状语)、obj(宾语)等是句法功能标记。其他标记依此类推。

一个框架涉及多个词元,用同一个框架的框架元素集合进行标注;反过来,一个多义词代表多个词元,属于几个不同的框架,即用不同的框架元素进行表示,有了这样的信息,一个应用系统就有可能区分出同一个词形在不同的使用环境中的不同意义。

与基于格语法的语义分析方法相比,CFN 提供的框架元素数量多、类型细化,突出了框架的个性,在语义表示深度上具有明显的优势。

2.3　CFN 词元库

词元库针对每一个词元,记录词元所在义项的具体含义以及该词元的句子标注报告,句子标注报告包括各个框架元素的句法实现情况以及目标词的语义搭配模式,它们是利用软件工具,从标注好的句子中自动汇总出来的。

CFN 的每一个词语都从真实语料中抽象出框架元素的句法实现方式,力求跳出由"意义到意义"描述的局限,因为只有形式特征,才是机器可用的。值得注意的是,CFN 并没有直接刻画动词的语义角色选择限制属性,但它基于真实语料,总结出了哪一个或哪一些语言成分可以充当动词的语义角色,而不是从直觉出发对角色的选择限制进行预测,这可能要比人工描述的结果更具体、更准确,也更有实用价值。

3　汉语框架语义知识库(CFN)的软件描述体系

构建现代汉语框架语义知识库是一个庞大的工程,无论是框架编辑、句子标注,还是实现基于 Web 的信息展示,都需要有一批软件来支撑,以便提高效率,并使得构建结果形式一致,成为结构化的在线数据库[5],同时能够使用户方便的看到我们的成果并得到自己想要的信息。相关软件主要包括汉语框架语义知识库管理系统和基于 Web 的展示系统。

3.1　汉语框架语义知识库管理系统

通过参照美国的框架网络数据库结构[5],并结合汉语框架网络自身的特点对 CFN 框架数据库进行了设计。数据库以语义框架为核心进行信息存储,通过词元与语义框架的联系、词元与标注句子的联系,在逻辑上形成框架库、词元库和标注句子库。在此基础上设计的语义知识库管理系统,实现了对框架信息的编辑(框架信息包括框架基本信息、框架元素信息、词元信息、框架关系信息、框架句子信息)、对框架信息的查询(如框架总体信息展示、框架信息分类查询、图形化关系示意)、句子标注符设置、句子句法语义辅助标注等功能,并将数据信息以不同的视角呈现给不同的用户。系统功能结构如图 1 所示。

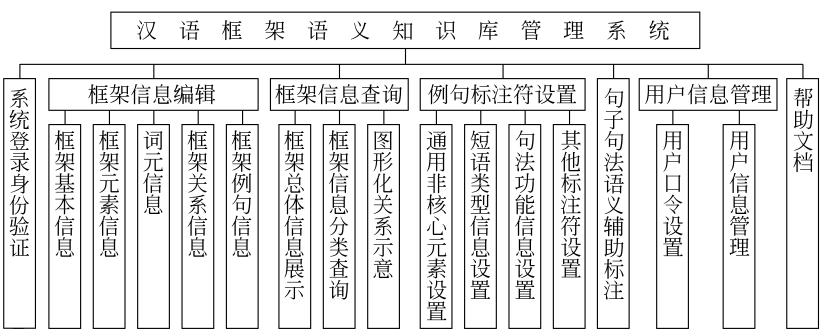

图 1　汉语框架语义知识库管理系统的功能结构图

篇幅所限,以下主要介绍词元的句子标注报告及句子句法语义辅助标注的的描述体系。

3.1.1　句子标注报告

词元库针对每一个词元,记录词元所在义项的具体含义以及该词元的句子标注报告。句子标注报告包括各个框架元素的句法功能抽取以及目标词的语义搭配模式。

〔1〕框架元素的句法功能抽取

框架元素的句法功能抽取是针对已标注好的句子文件,提取出其中的框架元素的句法功能的实现所采用的一种方法。其统计样式如图 2 所示。

图 2 中 0 代表最少出现的个数是零个,n 代表

最多出现 n 次,n 的个数没有限制;内部的大括弧为一个单一个体,其分为两个部分,第一个部分是短语类型,第二个部分是句法功能。一个框架元素可以搭配多个单一个体,即一个框架元素可以有多个短语类型和句法功能;外部的大括弧表示所统计内容由多个框架元素组成,最少出现的单一个体为一个,最多为 n 个。

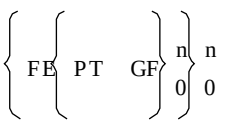

图 2 句法功能抽取样式图

多个句子与一个句子抽取出的句法功能模式类似,但不同的是如果短语类型和句法功能一致则标注数量与出现次数都加 1,如果不一致则只将标注数量加 1;句法功能模式的提取仅限于框架中定义的元素和标注过程中用到的通用非核心元素。

句法功能记录的是词元所支配的框架元素的句法特点,句法功能的抽取是统计每个框架在已标文本中出现的次数,以及此框架元素作为某种短语类型及句法功能所出现的次数,这样统计出的框架元素与某一特定词元搭配时所充当的短语类型及句法功能,为以后可能用到的这一特定词元所具有的特定信息做准备。具体的实现方式是:从数据库中读入一个框架元素,然后从文本中读入一个句子,查找此框架元素是否在这个句子中出现,如果出现则抽取出其相应的短语类型和句法功能,抽取完后再检查这个句子中是否还有此框架元素出现,如果出现则按上面的步骤继续抽取,如果没有则读入下一个句子,直至文本文件全部处理完;然后再从数据库中读入下一个框架元素进行同样的处理,直至所有框架元素全部处理完成。需要强调的是,这里处理的不仅仅是框架中所定义的框架元素,还包括框架中没有定义的但在实际标注过程中用到的通用非核心元素。

〔2〕框架的语义搭配模式抽取

配价模式抽取针对的是已标注好的句子文件,对某一已给定的词元进行其配价模式的抽取,格式仿照英文 FrameNet[3] 中所给出的样式。其统计样式如下:

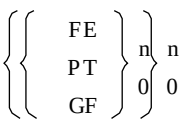

图 3 配价模式统计样式图

图 3 中的 0 代表最少出现的个数是零个,n 代表最多出现 n 次,n 的个数没有限制;内部的大括弧为一个单一个体,其分为三行,自上而下分别是框架元素、短语类型和句法功能。外部的大括弧表示所统计内容由多个单一个体组成,最少出现的单一个体为一个,最多为 n 个。

多个句子与一个句子抽取出的配价模式类似,但不同的是如果配价模式一致则只写入短语类型与句法功能,并将 Total 后的数字加 1;抽取的配价模式中的元素包括框架中定义的核心元素与非核心元素,还包括通用非核心元素,也包括只有框架元素而没有短语类型和句法功能的元素,例如“tgt”,同时还包括零形式(有定零形式和无定零形式)。

配价模式是词元库中最重要的一个组成部分,它将为计算机在今后理解语义方面发挥不可估量的作用。配价模式记录的是词元与框架元素的组合方式,框架元素出现时的短语类型与句法功能在这里不作为主要划分依据,即同一词元的两个句子中出现的框架元素相同并且出现次序也相同,那么不论框架元素所充当的是什么短语类型与句法功能,我们均认为这两个句子同属于一个配价模式;并且在这里词元与框架元素的组合方式是有先后顺序的,即同一词元的两个句子中如果出现的框架元素相同,但是出现顺序不同,则我们认为这两个句子的配价模式不同。

3.1.2 句子句法语义辅助标注软件

句子句法语义辅助标注软件主要实现对已经过分词软件加工后的句子文件标注。软件设计为用右键弹出三级子菜单进行标注,即针对一个句子,给定一个词元和该词元所属框架,根据预定的标记集合,交互地对句子进行标注,标注目标词元所支配的成分的框架元素类型、短语类型和句法功能三种信息。

句子句法语义辅助标注软件目的是为了减少句子标注人员在标注过程中出现的格式问题和拼写错误问题,同时也为词元库的自动生成保证标注质量。

3.2 基于 Web 的展示系统

基于 Web 的展示系统是一个基于浏览器/服务器架构的系统,内嵌 CFN 网站。用户通过浏览器可以进行注册,用于商业用途和研究用途的用户(机构)将获得不同级别的权限,同时下载资料的范围也有所不同。

系统可以从数据库(框架库和句子库)中归纳出各种用途的报告,如框架信息报告、词元信息报告、

(下转第 138 页)

基于统计方法的阿拉伯语到英语的机器翻译系统,其他一些语言的翻译和其他方面的应用也取得了很大的进展。计算语言学一个好的趋势是和语音研究团体的联系更加紧密了。但也有一个令人担心的现象,就是我们和语言学家变得疏远了。他表示,希望计算语言学能够像物理学一样赢得人们的尊敬。

物理学之所以重要的原因是:

Everything is made of particles, so Physics is very important.

现在我们同样可以说:

The World-Wide-Web is made of Language, so Computational Linguistics is very important.

致谢:

感谢孙茂松、王海峰、赵军、车万翔给本文提出的修改意见。感谢王海峰给作者转发了 ACL Newsletter No. 6,使作者得以修正了原文中的一些数字。感谢《中文信息学报》编辑部将本文初稿转发给上述老师修改补充,使本文更加完美。

(上接第 100 页)

每个词元的标注句子链接(这些句子就是各种语义结构模式的具体实现,都标注了语义信息)、框架元素的句法实现方式汇总表、词元的语义搭配模式汇总表、框架 - 框架关系图示报告,也可以进行框架元素、短语类型和句法功能的分别提取以及演示等。

4 结语

CFN 构建工程量大,难度很高,目前课题组只能在有限领域下开展工作,但是,已经探索了一条可行的技术路线,取得了阶段性成果,为实现语义 Web 中的语义知识共享以及智能化、个性化的 Web 服务提供了基础资源。近期我们计划扩展到其他领域,继续构建一批框架,并将其应用到阅读理解式问答系统和智能搜索系统等应用研究中,以检验 CFN 的实践效果。研究小组近期正在进行基于 CFN 的句法语义角色自动标注软件的设计与开发,目的是开发高性能的汉语句法语义分析器,为进行大规模真实文本的语义信息标注提供有力支持。

参考文献:

[1] 刘开瑛,由丽萍. 汉语框架语义知识库构建工程[A]. 中文信息处理前沿进展,中国中文信息学会成立二十五周年学术会议论文集[C]. 2006, 11: 64-71.

[2] Charles J. Fillmore. Frame semantics and the nature of language [A]. In: Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech[C]. 1976, 280: 20-32.

[3] Charles J. Fillmore, Charles Wooters, and Collin F. Baker. 2001. Building a large lexical data bank which provides deep semantics[A]. In: Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation[C]. Hong Kong: 2001, 3-26.

[4] Charles J. Fillmore, Collin F. Baker et al. The Berkeley FrameNet project [A]. In: Proceedings of COLING/ACL[C], Montreal, Canada: 1998. 86-90.

[5] Collin F. Baker, Charles J. Fillmore and Beau Cronin. The Structure of the Framenet Database [J]. International Journal of Lexicography, 2003, 16(3): 281-296.

[6] Charles J. Fillmore, Collin F. Baker and H. Sato. The FrameNet Database and Software Tools[A]. In: Proceedings of the Third International Conference on Lan- guage Resources and Evaluation [C]. Las Palmas, Spain: 2002.