汉语框架语义知识库构建工程介绍

山西大学计算机与信息技术学院 刘开瑛
山 西 大 学 管 理 学 院 由丽萍
二零零六年十一月二十日

引 言

- 汉语框架语义知识库(Chinese FrameNet，简称CFN)是一个以框架语义学为理论基础、以真实语料为事实依据的语义词典，其资源将用语义Web标记语言描述，成为一部计算机可读、可理解的语义词典。
- 本文介绍CFN构建的前期考察分析工作、目前的构建结果以及基本的构建方法。

语义知识库是一项基础性资源

- 从国际来看,上世纪80年代以来，自然语言处理从句法学转移到语义学和语用学方面，而语义学是重点，词一级语言单位的语义研究又是重中之重。实践证明，只进行句法规则的描述是不够的，语言描写要落实到词这一级语言单位上来。无论做机器翻译（MT）、信息提取（IE）还是词汇语义排歧（WSD），语义知识库是所有这些应用的不可或缺的一项基础性资源。

现有的几种语义分析方法(1)

- 在句法关系链上添加相应的语义关系标签：
  - 例如Propbank是在Penn TreeBank句法分析的基础上，对与动词有关的语义角色进行标注，包含50多个语义角色类型；汉语方面，以Tesnière的依存语法理论为基础的语义分析方法。如李涓子等的依存语义分析、台湾中研院（1999）的中文句结构树资料库（Sinica Treebank），利用几十个语义角色，在句法关系链上添加语义标签。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。

现有的几种语义分析方法(2)

- Schank(1975)的概念依存理论（Conceptual Dependency, CD）
  - 利用少数几个概念表达丰富的语言意义。该理论对限定领域内的特定应用比较有效，但对常识的描写过于刻板和定式。对于汉语来说，始终是停留在高度抽象的概念表达上，没有落实到具体的语言单位，使得研究者对自然语言的语义表示深度、语义表示标准很难把握。

现有的几种语义分析方法(3)

- 框架语义学
  - 是Fillmore及其同事在格语法基础上进一步提出的语义学理论，其中心思想是词的意义的描述必须跟语义框架相联系。
  - “框架”（Frame）作为一个语言学术语，是指人们理解语言时激活的大脑已有的认知结构，这种认知结构是通过词语反映的。

框架语义学

- 框架语义学的根本特点是经验主义方法，即根据背景框架的不同，对于属于同一个框架的一类词语，明确其具体的框架元素，不同的框架在框架元素的类型和数量上有差别，而传统的格语法的“语义格”是相对于所有词汇而言的。

语义知识库国内外研究现状

- 总体来看，没有词一级的语义知识库，要分析自然语言的意义，是行不通的。
- 目前国内外比较有代表性的语义知识库，有英语的WordNet、MindNet、ILD、FrameNet，汉语的HowNet、《现代汉语述语动词机器词典》、CCD等。
- 这些语义知识库的语义描述涉及多方面内容：词语分类关系，词义组合性质，场景知识，概念与概念之间的多种联系，呈现出“百家争鸣”的局面；有的主要提供了词语之间同义、同类关系，如WordNet,CCD；有的描述了动词与体词性成分之间的组合关系，如HowNet，《现代汉语述语动词机器词典》。

语义知识库构建工程存在问题
（1）理性主义的构建方法

- 理性主义的构建方法使构建结果存在很大的主观性。对语义的认识是“从意义到意义”，以意念为主进行理性思维的结果，而没有客观依据和有效的评价标准。
- 单一地以层级分类组织词义聚合关系，有的事物适合放在层级分类的框架中来认识，有的事物并不适合这样来认识。用什么样的结构去认识各种概念，就目前来讲，应该还是个研究的课题。
- 普遍缺少有针对性的实践检验和评价结果。人们可以提出语义描述的整体体系，可以从各种角度、以各种深度对语义进行表达，但描述是否到位需要靠真实语言现象来检验，描述结果的优劣最终也要靠应用系统的实践来评价，没有实验结果自然不足以服人。

语义知识库构建工程存在问题
（2）通用性限制了实用性

- 语义知识库多数都追求“大规模”、“覆盖面”，试图描述一种（或多种）语言的全部词语，试图覆盖普遍的语义领域。然而，适用于普遍领域的语义知识库构建还存在着大量的基础问题需要解决，试图将这样的通用词典直接应用于实用系统，似乎还欠成熟。
- 其实，社会对汉语处理的需求不仅仅是通用词典，例如，某些特定领域（游游咨询、股票咨询）的语义信息，将获得的阶段成果解决社会迫切需要解决的实际问题，开拓特别领域应用的市场，是汉语处理研究获得进一步支持的必要条件。

语义知识库构建工程存在问题
(3) 离计算机可读、可理解还有很大距离

- 离计算机可读、可理解还有很大距离
- 现有语义知识库只停留在语义信息的描述上，没有对知识库的形式化表示做专门的研究，尤其是缺少Web环境下机器读取资源的接口，不能满足Web技术飞速发展的要求。

FrameNet明显的优点（1）

- 摆脱了格清单难以确定的问题，具有个性特征的框架元素更适合用来描述千变万化的自然语言语义。提供数量多、类型多的框架元素，并突出框架的个性，适合计算机处理语言的需要。据初步统计，已经公布的800多框架中共有语义元素1000种之多。用这种语义知识库表示句子的语义结构，表示结果更深入、语义信息更丰富。

FrameNet明显的优点（2）

- 具备抽象化概念联系，适合计算机推理的需要。
- 提供语义标注句子库，详尽描绘了词汇的框架语义在真实语料中的实现情况，这就使得该语义知识库可以直接应用于自动语义标注器的研究。

FrameNet明显的优点（3）

- 在资源描述上已向机器可读、可理解迈进了一步。FrameNet工程已经开发了本体自动转换器，FrameNetI版的数据已经转换成了DAML+OIL(The DARPA Agent Markup Language)，这是对XML（可扩展标记语言）和RDF（资源描述框架）的延伸，使FrameNet成为语义web的一个资源。

构建汉语框架语义知识库（Chinese FrameNet, 简称CFN）

- 由于FrameNet描述的是词语背后的认知框架，许多国家的学者通过研究都承认其数据可以跨语言使用，有通用价值，尝试建立与FrameNet并行的词典，包括希伯莱语、德语、日语、西班牙语等。
- 基于以上考查，兼顾现代汉语语义研究不成熟的现状，我们选择了Fillmore的框架语义学作为理论基础，以伯克利FrameNet为参照，构建汉语框架语义知识库（Chinese FrameNet, 简称CFN）。

汉语框架语义知识库内容

- CFN包括三个子库：框架库、句子库和词元库。
- 目前我们定稿的一个以有限词语集合为描述对象的汉语框架语义知识库，共对汉语1760个词元（一个义项下的一个词）构建了130个框架，涉及动词词元1428个、形容词词元140个、事件名词词元192个，标注了8200条句子，为构建大规模汉语框架语义知识库的样本。
- 框架库对汉语词汇按照所表示的活动场景（框架）的异同分类描述，包括该框架的定义、框架元素和框架关系，并带有相应的示例，如[位移]框架：

框架名	位移	
定义	转移体从源点出发，终止于目的地，两个地点之间经过的是路径。	
核心框架元素	转移体	改变地点的实体：我今天去了体育馆。
	方向	移动的方向：向校长走过去。
	目的地	转移体终止的地方：车进了慢车道。
	路径	转移体行进在其上的路线：他绕过爸爸进了客厅。
	源点	转移体在位置改变之前的处所：警察从门口走开了。
	区域	在没有明确路径的情况下，转移体位移的背景：他在屋子里不安地走动。
非核心框架元素	载体	运输转移体的工具。
	形容	描述位移发生时转移体的状态。
	距离	表示位移的长度：小嫩枝在水上漂流了大约100米。
	动作时间量	位移发生的时间数量。
	修饰	位移发生的方式：一艘海军飞艇在暴风雨中疯狂地漂流。
	速度	转移体位移的速度。
	时间	位移发生的时间。
词元	漂流v，漂浮v，飞v，滑行v，走v，移动v，滚动v，滑动v，滑翔v ，漂移v，漂v，进v，走动v，去v	

[位移]框架－框架关系

- 父框架：空
- 子框架:[集体位移]，[有向位移]
- 总框架：[位移情境]
- 分框架：空
- 总域：空
- 分域：[到达],[肢体移动],[运送],[伴随],[出发],[扩散],[逃避],[分泌],[液体移动],[光移动],移动声响],[位移情境],[驾驶],[放置],[改道],[移除]，[自动]，[射击]，[旅行]
- 后续过程: 空
- 结果状态: 空
- 参照: 空

汉语框架语义知识库构建方法(1)

- 经验主义的语义描述方法
  - 对于汉语词汇，确定它们属于那个框架，以至对于一个框架，决定它有哪些框架元素时，主要是根据大量的真实语料。首先找跟词元有关的句法成分，看这些成分传递了什么语义信息，然后选择适当的标签去区分这些成分的语义角色。框架元素的基本语义类型应该在各种使用中都一致，如果不一致，就成为不同的框架元素，即使出现在同样的句法位置，也会根据所指不同，而框架元素类型不同。

汉语框架语义知识库构建方法(2)

- “引进消化吸收再创新”的构建思路
  - 由于英汉两种语言差异较大，在我们目前所构建的130个框架中，有80%是从FrameNet已有框架中转化而来的，其余20%中，有些是由于将原英文框架分化为多个，有些则是由于在伯克利FrameNet中找不到对应框架。

汉语框架语义知识库构建方法(3)

- 用语义Web标记语言描述汉语框架语义知识库资源
  - 我们2004年初开始逐步展开对语义Web尤其是本体的研究,将CFN数据的语义Web语言表示技术作为本课题的另一个重要研究内容，并探索OWL自动转换器的实现技术，这是对XML和RDF（资源描述语言）的延伸，使CFN成为一个机器可读、可理解的语义词典，已有OWL描述样例和OWL自动转换器实验软件。

CFN的句子标注

- CFN的句子标注是针对一个句子，出框架元素名称、短语类型和句法功能三种信息。选句主要采用“北京大学CCL现代汉语语料库”，由于题材和体裁丰富，CFN工程所需描述的词元基本上都可以在该语料库中找到实例。

“科学家们正在观察核武器”的标注结果：

- <perc_agt-np-subj 科学家们> <time-dp-adva 正在> <tgt 观察> <phen-pp-adva 核武器>。
- 其中，tgt表示标注的目标词“观察”，该词语属于[自主感知]框架；perc_agt表示框架元素自主感知者，np表示短语类型名词性短语，subj表示主语，其他标记依此类推。

我们取2个Sinica Treebank公布的样例，尝试比较一下标注结果：

- 1a. <location 从水里><Head 看見>了<goal自己>
- 1b. <背景 从水里><tgt 看见><null了><印象 自己>
- 2a. <location 由露天阳台上><Head眺望><goal运河美景>
- 2b. <感知者位置 由露天阳台上><tgt 眺望><现象 运河美景>

这两个句子中跟空间有关的语义成分标注结果不同：a组把“从水里”和“由露天阳台上”都表示为location，b组则予以区分：1b是感知觉活动的背景，即所看到的事物“自己（的影子）”所在的场所，2b是感知者的位置，而不是感知对象的位置。

设想一个问答系统，要回答关于某物/人在哪里的问题，显然b组的标注更有可能输出正确答案。

CFN词元库基本内容

- 词元库描述每一个词元的词义，并根据句子标注结果形成标注报告。下面是词元“看”的语义搭配模式汇总报告示例

标注数量	语义搭配模式			
1total	time(时间)	phen(现象)	tgt（看）	perc_agt(自主感知者)
	np	np		ini
2 total	perc_agt(自主感知者)		tgt （看）	phen(现象)
2	np			np
	subj			obj
1 tota	perc_agt(自主感知者)	supp(支撑词)	tgt（看）	phen(现象)
	np			np
	ext			obj

用语义Web标记语言描述汉语框架语义知识库资源

- 目前，互联网上的大部分信息都是用HTML语言来表示的，虽然HTML允许我们将网上的信息可视化，但是它提供的描述信息的方式不能用软件方便地进行识别和描述。2000年，国际万维网联盟W3C总裁Tim Berners-Lee提出了下一代万维网——“语义web”的理念,成为人们讨论与研究的热点。XML, RDF和Ontology（本体）是语义Web的关键层，用于表示Web信息的语义。

使用语义Web语言描述CFN资源需要做如下工作：

- （1）用XML标记CFN数据库的文档内容，这种标记语言是灵活的、可扩展的，给使用者提供标记元素自定义功能，让每个人都能创建自己的标签，来对网页或页面的部分文字进行注释。
- （2）用RDF描述CFN词汇语义资源，其基本结构是“主体-谓词-客体”三元组；这些三元组可以用XML语法来表示。用这种结构描述由机器处理的大量数据，是非常自然的方法。
- （3）用本体描述资源之间的联系，解决目前Web 的信息格式的异构性、信息语义的多重性以及信息关系的匮乏和非统一性。

语义Web尤其是本体的调研

- 2006年5月，Tim Berners-Lee宣布，W3C已发布W3C推荐标准80余份，语义Web已经具备了成功所需要的所有标准和技术，包括作为数据语言的RDF、本体语言，以至查询和规则语言，这些国际标准和技术方面的准备为我们的研究提供了可靠的基础。
- 我们2004年初开始逐步展开对语义Web尤其是本体的研究,将CFN数据的语义Web语言表示技术作为本课题的另一个重要研究内容，并探索OWL自动转换器的实现技术，这是对XML和RDF（资源描述语言）的延伸，使CFN成为一个机器可读、可理解的语义词典，已有OWL描述样例和OWL自动转换器实验软件。

《科技查新报告》

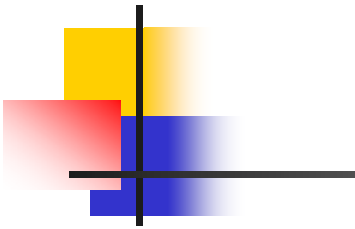
- 2006年7月26日经国内联机信息检索及中文光盘数据库检索提供的《科技查新报告》指出，查到与语义Web相关的文献多篇，查新结论为：“目前国内语义Web研究多数处于学习、消化阶段，没有一个汉语词汇语义知识库作为资源支持。用语义Web标记语言表示汉语框架语义知识，具有一定的新颖性。”

人机交互的技术路线

- 构建CFN是一个庞大的工程，无论是框架和词元内容的编写还是句子标注，都需要高度交互的辅助工具。为了满足构建语义知识库的需要
- 我们自主开发了框架信息编辑、框架信息查询、句子辅助标注、自动生成词元库统计报告等功能的汉语框架语义知识库开发和管理系统。
- 目前正在开发的软件还有CFN本体文件自动生成器、CFN资源文件自动生成器、CFN自动推理器和CFN一致性检测器，等等。

结束语

- 近年来，我们课题组依靠山西大学计算机科学与技术学院，在山西大学网络中心成立“语义Web研究室”和管理学院成立“语言信息处理中心”，并于太原高新技术开发区数码港成立语义Web研发中心，良好的软硬件环境为研究提供了强有力的支持。课题组吸收太原理工大学、上海师范大学、清华大学和北京大学有关单位专家参加或指导，包括计算机、数学、物理、中文、图书情报、信息管理等不同学科的师生，大家聚集在一起，通过交谈、互听报告、互通资料、互改论文，迸发出思想火花，进行前沿探索，通过多学科的前瞻性和战略性思考和研究来激发灵感。



谢谢