文章编号：1003-0077(2011)06-0046-07

汉语框架语义网构建及其应用技术研究

刘开瑛

（山西大学 计算机与信息技术学院，山西 太原 030006）

摘　要：汉语框架语义网(Chinese FrameNet,CFN)是一个以 Fillmore 的框架语义学为理论基础、以英文 FrameNet 为参照、以汉语语料事实为依据的供计算机使用的汉语词汇语义数据库。该文首先介绍了汉语框架语义网的构建基础——框架语义学以及英语的框架语义网工程，然后具体分析了汉语框架语义网的构建技术，并对基于汉语框架网的语义角色自动标注研究进行了介绍，25 个框架的交叉验证的实验结果的准确率、召回率、F1-值分别达到 74.16%，52.70%，61.62%；最后，介绍了几个基于汉语框架语义网的研究课题的进展情况。

关键词：汉语框架网；语义角色标注；框架语义依存图

中图分类号：TP391　　**文献标识码**：A

Research on Chinese FrameNet Construction and Application technologies

LIU Kaiying

(School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract: Chinese FrameNet is an computational lexical semantic database which is based on the frame semantics by Fillmore, referring to the FrameNet by the California University at Berkeley and supported by Chinese real copus. This article introduces the CFN' basic theory: Frame Semantics and the English FrameNet project. Then it analyzes the constructing technology, introduces the ongoing researches on automatic semantic annotation based on CFN. the experimental results on all 25 frames data for the precision, the recall, and F1-value reached 74.16%, 52.70%, 61.62% respectively. At last, it introduces the situation of some researches based on CFN.

Key words: Chinese Framenet; semantic role labeling; frame semantic dependency graph

1　框架语义学和框架语义网

框架语义学(Frame Semantics)是由 Fillmore 提出的研究词语意义和句法结构意义的一种理论方法[1]。该理论试图用经验主义方法，探寻语言和人类经验之间的联系，并研究一种可行的描述方式来表示这种联系，即将词义、句子意义和文本意义统一用"框架"(Frame)进行描述。框架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，是储存在人类经验中的图式化情境，这种范畴系统所描述的既可能是一个实体，也可能是一种行为实践模式，甚至是一些社会制度、习俗等。框架中的各种参与者称为框架元素(Frame Elements)。例如，"煎、焙、煮、烤"等动词在人类经验中激活的是烹饪的场景，该场景涉及做饭的人(Cook)、食物(Food)、盛食物器皿(Container)和热源(Heating-instrument)等要素。因此，描述"煎、焙、煮、烤"等动词的语义性质，就可以将其归入烹饪框架，以做饭的人、食物、器皿和热源等为框架元素进行刻画。框架元素在语义关系中的作用与语义角色或格角色相当，但传统的语义角色或格角色是相对于普遍的词汇而言的，而框架元素仅适用于具有相同框架背景的一小组词语，其类型大大细化。框架元素表达的语义内容更加丰富、更加深入，用其来描述自然语言的语义更为适当且实用性强。

收稿日期：2011-09-15　**定稿日期**：2011-10-10

基金项目：国家自然科学基金资助项目(60873128)；国家 863 高技术研究发展计划资助项目(2006AA01Z142)；国家社会科学基金资助青年项目(07CYY022)

作者简介：刘开瑛(1931—)，男，教授，主要研究方向为人工智能和中文信息处理。

1997 年，美国加州大学伯克利分校即以框架语义学为理论基础，开始构建一个基于真实文本的词汇语义数据库——框架语义网(FrameNet)[2]。该项目在伯克利的国际科学研究所运作，主要是由国家科学基金会的支持，数据可免费下载。它已被世界各地的研究人员下载和使用于各种各样的用途。截至 2011 年 8 月，框架语义网已定义了 1 094 个语义框架，这些框架描述了 12 132 个词语的语义内容，并对其中 65％的词语进行了句子语义标注，每个词语标注 15 到 20 个真实语句。框架语义网为每个框架细致刻画了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；非核心框架元素并不显示框架的个性，可以出现在多数框架中。目前的 FrameNet 共定义了 9 328 个框架元素，平均每个框架约有 10 个。举一个略复杂的例子，例如，attract(牵引)、cast(抛掷)、catapult(弹射)、drag(拖)等 34 个动词所属的使位移(Cause_Motion)有 9 个核心框架元素：施动者(Agent)、区域(Area)、致因(Cause)、目的地(Goal)、始状态(Initial_State)、路径(Path)、结果(Result)、源点(Source)、转移体(Theme)；11 个非核心元素：程度(Degree)、形容(Depictive)、距离(Distance)、解释(Explanation)、手柄(Handle)、工具(Instrument)、方式(Manner)、方法(Means)、空间(Place)、亚区(Subregion)、时间(Time)。此外，该数据库还定义了 8 类框架—框架关系，包括继承关系、整体—部分关系、因果关系等，共建立了 1 589 个关联，几乎将所有的框架连接到了同一个网络图中，因此称为框架语义“网”。

目前伯克利 FrameNet 团队正在对美国国家语料(American National Corpus，ANC)进行句子的标注和全文标注，作为多重标注项目(即 Multiply-Annotated Sub-Corpus 项目)的一部分。另一个项目，正在和国防承包商合作创建军事领域的框架和词元，此国防承包商曾建立了一个士兵战斗报告的自动语义角色标注的系统，用来决定两篇报告什么时候是不同人报告同一件事，什么时候是表述不同的事件。

2　汉语框架语义网的构建

由于框架语义描述是以人的认知经验为基础的，因此，在不同的语种中存在很大的共性。例如，框架购买和出售都包括购买者、出售者、商品、金钱等元素，不论各种语言的具体形式有什么不同，其语义是基本相同的。目前，许多研究人员正在建立与框架语义网平行的词汇语义数据库，包括西班牙语[3]、德语[4]、日语[5]、巴西语[6]和汉语等。汉语的框架语义网(Chinese FrameNet，CFN)是由山西大学从 2004 年开始建立的。该项工程一方面针对汉语词汇，参照英语框架语义网，译建或创建适合汉语语义内容的框架，定义其框架元素以及框架—框架关系；另一方面，以汉语真实语料为支撑，针对各个框架标注了一些例句，显示框架语义在句子表层的表现形式[7]。目前，已对 3 151 个词元(一个义项下的一个词)构建了 309 个框架，标注了 2 万多条句子。除了描述通用领域一些常用词语的框架语义外，对认知语义和法律、旅游等应用领域也进行了系统的语义知识描述。其中，认知领域包括 51 个框架，涉及 512 个词语[8]，法律领域包括 86 个框架。2006 年 10 月 11 日，山西省科技厅组织，聘请国内专家，由倪光南院士主持对该工程的阶段性成果进行了科技成果鉴定，鉴定结论为“该课题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”。

CFN 由框架库、句子库和词元库三部分组成，下面即对各子库的构建技术加以说明[9]。

(1) CFN 框架库构建技术

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。

框架构建的基本原则如下：

a) CFN 框架是可以直接靠 FN 翻译和修改补充完成的。现已有翻译普通词语框架 240 个，自制的只有 19 个。自制框架有：等同、比较、推理、意识、想象、使满足、使呈现等。例如，频率最高的多义词“是 v”属于“属于某类、存现、状态、类似和等同”五个框架，其中“等同”框架为自制。

但是我们对各种不同专业领域的框架包括法律术语、足球、生物医学领域和旅游业就不同。如 CFN 的法律自制框架库工作量很大。美国法律属英美法系，是判例法，判例法的一个重要特征是“遵循先例”。而我国法律是属于大陆法系的，是成文法，人们的权利和义务都以法律的形式确定下来，所

有的人适用同一部法律,法院也不例外,所谓法律面前人人平等。所以在做CFN法律框架库时,应结合我国政治和经济制度,词元选择必须根据我国现行法规术语。在CFN中已有法律翻译34个框架,自制52个框架。

表1 CFN框架库样例

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),属性(Att),初值(Val1),终值(Val2),初状态(Inis),终状态(Finis),变幅(Diff),值区间(Val_ran)		
非核心框架元素	环境条件(cir),倚变因素(Cor),动作时间量(Dur),倚变起点(Cor1),倚变终点(Cor2),修饰(Manr),路径(Path),空间(Place),速度(Speed),时间(Time)		
框架关系	父框架:无	总框架:无	后续过程:无
	子框架:[增殖]	分框架:无	结果状态:[数量]
	参照:无		
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		

b) 框架元素的数量和类型是区分和认定框架的根本标志。例如,"改变"一词有两个义项:事物发生变化;使发生变化。分别出现在以下两种上下文中:

例1 岩石**改变**了形状。

例2 老师**改变**了小毛的想法。

例1中"岩石"自身在某一属性上("形状")发生了变化,而例2则是由一个外部力量"老师"致使某实体发生变化,也就是说,例2比例1增加了一个表示致因的框架元素。因此"改变"作为两个词元归入不同的框架:[经历变化]和[使变化]。

真实汉语语料始终是决定框架元素、概念归类等的主要证据。由于我国同英美国家生活习惯、宗教信仰和人文文化等不同,对于一个词语,如果语料中有某一语义成分出现,则给该词语所承担的框架设立一个框架元素;反之,如果语料中这个词没有和某一语义成分共现,则不作为我们设立框架元素的标准。这样做,可能会因为所考查的语料范围有限,遗漏一些框架元素,但是,随着研究的进展,我们会根据新的语料不断修正前面的构建结果。

c) 对于一个框架,所含词元应该逐一审查,尤其不能直接翻译。因为在英语FN和CFN有的属于同一框架,有的不在同一框架。我们以汉语中表示感知觉活动的两类词语的框架构建情况为例:

<1> 听v,看v,尝v,闻v,嗅v,听听v,闻闻v,看看v,尝尝v,……

<2> 感到v,听到v,听见v,听出v,觉得v,看到v,看出v,看见v,……

两组词语都表示有感知能力的实体利用感觉器官对外界事物进行感知,但是<1>组的动作发出者在活动中是自主的,而在<2>组所表示的活动中,动作发出者是不能控制动作发生的。可见,两组词语的核心框架元素类型不同,前者概括为**自主感知者**,后者为**非自主感知者**。

还有表示同一领域而不同的事件阶段和状态,例如,动词"**知道**"表示一种**认知状态**,即认知者大脑拥有某内容,但"**获知**"表示的则是一种**认知过程**,表示认知者的认知状态由不知道变为知道。这类情况下,词语表示了不同的事件阶段或状态,应该归入不同的框架。

d) 词性与框架没有必然的对应关系。区分框架主要根据框架元素的类型、数量和框架所表示的事件阶段,表示动作行为的框架涉及的词元大部分是动词,但也有少数事件名词(即有配价的名词),表示状态的框架大部分是形容词,但也有动词、名词。

(2) 句子库构建技术

句子库是CFN为每一个词语例句来自真实的汉语语料库,如"北京大学CCL现代汉语语料库"。该语料库是生语料库,需要经过分词F2000加工成熟语料库,每条例句均具有分词、词性标注以及名体信息,并且人工确定每个待表词语取10~20个义项一致的例句,针对每条句子标注了目标词(每个例句只标注了一个目标词)及其框架语义角色。

CFN句子标注,是以框架库为基础,针对一个句子,给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"大型弦乐队的人数增加一倍或一倍以上"的标注结果如下:

<att-np-subj 大型弦乐队的人数> <tgt=[量变] 增加> <diff-np-obj 一倍或一倍 以上>。

其中,tgt表示所标注的目标词"**增加**",该词语属于[量变]框架;att表示框架元素属性(框架元素标记见表1),np表示短语类型是名词性短语,subj表示句法功能是主语,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。

句子库主要是目标词确定，汉语中谓词有：体词谓语句、形容词谓语句和动词谓语句。但当动词或形容词作定语、状语和连用形容词和叠用动词都不是目标词。

(3) 词元库构建技术

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，后者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，见表 2。

表 2　目标词元“看”的汇总表

框架元素	标注数量	句法实现方式
自主感知者	(28)	np. subj (9)；np. ext (11)；ini (4)；np. head (1)；pp[给]. ext (1)；np. dlc (1)；np. subj_s(1)
现象	(24)	np. obj (11)；dj. obj (4)；dni (2)；ini(3)np. ext(2)；np. subj (1)；vp. obj(1)
方向	(3)	sp. obj(2)；pp[往]. adva(1)
修饰	(2)	dj. comp(1)；ap. adva(1)
时间	(6)	tp. adva(4)；tp. subj_s(1)；tp. adva_s(1)
空间	(1)	pp[在]. adva(1)
动作时间量	(1)	np. obj(1)

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　汉语框架语义角色的自动标注技术研究

汉语的语义角色自动标注的较早研究是文献[10]，文献[11]基于中文 PropBank 的自动标注研究工作。文献[12]基于中文 PropBank 语料库，在使用手工标记好的句法分析上，得到了 94.1% 的 F1-值。但若采用自动的句法分析，却只有 71.9% 的 F1-值。PropBank 只对每个句子的核心动词进行了标注，语义角色的定义总共有 50 多个，所有动词的主要角色最多有 6 个，均以 Arg0-Arg5 为标记。正如文献[13]指出：“中文 PropBank 中，论元标记 Arg2 至 Arg5 对于语义角色的严重超载，使用这种标注语料来训练角色自动标注系统，其性能势必会受到影响”。事实上，PropBank 中语义角色类型忽略了语言表达中的细节，词汇义项的描述显得不够精细。

近年来，许多学者已使用条件随机场模型进行语义角色标注的尝试，文献[14]条件随机场模型直接使用到 PropBank 的句子的完全句法分析树上，建立标注模型，实验结果表明条件随机场的标注性能要显著好于最大熵模型。文献[15]以 PropBank 为实验语料，将句法分析树“压平”，并考虑句法树中水平层次上的角色标签之间的马尔科夫依赖关系，以线性链 CRF 为标注模型进行了语义角色标注实验，提高了模型的精度。文献[16]使用条件随机场模型研究了英文 PropBank 的语义角色标注问题，他将完全句法分析树转换成浅层短语块序列，并使用浅层短语块和命名实体块作为标注单位，也取得不错的结果。所有这些结果表明条件随机场模型在语义角色标注中有不俗的表现。然而，文献中使用条件随机场进行汉语框架语义角色自动标注研究的很少。文献[17]基于规则的方法研究了“自主感知”和“非自主感知”两个框架的语义角色的标注，但每个框架构建规则工作量大，适应性差。文献[18]中使用层叠条件随机场的 CFN 对“包含”“陈述”“拥有”“属于某类”“研究”“提供”“适宜性”等十多个汉语框架元素自动标注结果的准确率为80.1%，召回率为 69.9%。不过，至今文献中未看到汉语框架语义角色的自动标注的系统研究。

文献[19-21]使用条件随机场模型(CRF)进行汉语框架语义角色标注。其标注任务为：对于一个汉语句子，在给定目标词及其所属框架的前提下，将其框架元素的自动标注问题通过 IOB 策略转化为整个句子上的词序列标注问题，使用条件随机场模型(CRF)，采用统计学中的正交表实验方案，自动识别出目标词所支配的框架语义角色的边界，并标注出该目标词所支配的语义角色名称，既包括核心框架元素，也包括非核心框架元素。这个任务与 Senseval-3 中针对英文 FrameNet 的语义角色标注

任务是相同的。实验分基于词层面特征和基于基本块特征两大部分,并在每个实验中分别考查将语义角色边界识别和角色分类同时进行和分两步标注两种情况。实验所用 CRF ++ 工具包来自于 Sourceforge(http://crfpp.sourceforge.net/),使用其中的 CRFL2 算法,并选取 C=1 进行参数平滑。所有 CRF 模型中都用一阶转移特征。

实验选取 25 个框架的例句库,将其拆分为 4 份;考虑到语料规模偏小,采用了 2-fold 交叉验证方法,即,任取两份作为训练集,其他两份作为测试集。这样共可以做 3 组 2-fold 交叉验证。最终以 3 组交叉验证实验的 F1-值的平均值来评价标注模型的性能,见表 3。

表 3 CRF 模型实验结果汇总表

		mP	mR	mF	std(mF)
基于词特征	边界识别和分类同时进行	72.60%	49.50%	58.86%	0.001 9
	先识别边界,再分类	62.87%	56.44%	59.48%	0.005 0
基于基本块特征	边界识别和分类同时进行	73.99%	50.73%	60.19%	0.000 8
	先识别边界,再分类	63.97%	57.25%	60.42%	0.005 0
分批正交表的实验	只识别边界	75.69%	68.08%	71.68%	0.005 4

表中,mP——表示多组交叉验证的平均准确率;mR——表示多组交叉验证的平均召回率;mF——表示多组交叉验证的平均 F1-值;mA——表示多组交叉验证的平均精确率;std(mF)——为 mF 的标准差的估计,也即 std(mF)=$\sqrt{Var(mF)}$。

基于 CRF 的汉语框架语义角色自动标注实验可得以下基本结论:(1)边界识别与角色分类同时进行的情况下,标注效果较好;(2)基本块特征对角色分类有显著作用,但对边界识别作用不显著;(3)采用分批正交表实验(方案三)比使用全部特征一起建模的结果好。

CRF 与其他标注模型结果比较。主要阐述了使用支撑向量机(SVM)和最大熵模型(ME)两种模型进行语义角色标注的实验结果,并将它们与给出的基于条件随机场(CRF)模型的实验结果进行了对比。所有方法实验结果汇总表将所有结果汇总,见表 4。

表 4 所有模型实验结果汇总表

		mP	mR	mF	std(mF)
基于词特征					
条件随机场 CRF	先识别边界,再分类	63.04%	60.20%	61.58%	0.004 9
支撑向量机 SVM	先识别边界,再分类	71.67%	55.27%	62.41%	0.002 8
最大熵模型 ME	先识别边界,再分类	61.36%	54.59%	57.78%	0.003 6
基于基本块特征					
条件随机场 CRF	先识别边界,再分类	64.28%	61.03%	62.61%	0.004 4
支撑向量机 SVM	先识别边界,再分类	72.53%	54.84%	62.45%	0.002 5
最大熵模型 ME	先识别边界,再分类	60.43%	58.07%	59.23%	0.004 2
基于完全句法分析树					
用 Stanford 大学句法分析器,再使用 ME 分类器	仅边界识别使用文献[12]的特征	71.70%	45.46%	55.64%	0.004 0

5 结论

模型以词为基本标注单元,将标注步骤分为 1)边界识别;2)角色分类;3)后处理三个步骤。全部实验是在选出的 25 个框架的 6 692 个例句的语料上进行。将语料均匀分为 4 份,分 3 组作 2-fold 交叉验证,以 3 组交叉验证的平均 F1-值作为最后评价指标。基于条件随机场标注模型(CRF)与基于支持向量机(SVM)模型的标注结果没有显著差异,但 CRF 显著好于基于最大熵(ME)模型的标注结果。在全部 25 个框架的所有实验中,语义角色边界识别最好的结果(mF)为 71.68%;在给定语义角色边界下角色分类的最好结果(mA)为 84.08%;在给定句子中的目标词以及目标词所属的框架情况下,最好结果(mF)达到 63.26%。

4 其他基于 CFN 的自然语言处理研究进展

4.1 汉语框架自动识别中的歧义消解研究

框架自动识别即给定一个句子及其目标动词

(或事件名词)，自动识别出其所属框架。由于框架语义网中定义的框架元素是从属于特定框架的，因此，框架的识别实际上是实现句子语义分析的前提。该任务的难点在于，有些动词，尤其是一些常用词，有不只一个义项，分属于多个框架，如动词“有”在CFN中属于拥有、存现和形成三个框架，这就需要根据具体的上下文消解歧义。例如，当“有”出现的句子**“全书的观点有创意”**中时，应该标注为拥有框架，而不是存现或形成框架。框架自动识别的歧义消解可分解为三个子任务：(1)词元检测(Lexical Unit Detection)，(2)未知框架检测(Unknown Frame Detection)，(3)框架消歧(Frame Disambiguation)。文献[22]经过词元检测、未知框架检测后，确定有88个词语对应两个以上框架，涉及框架14个，相应的例句2 077条。研究将框架消歧任务看做典型的单点分类问题，使用最大熵对其进行建模，选用词、词性、基本块、依存句法树上的若干特征，并且借助于开窗口技术和边界识别策略，采用3-fold交叉验证方式进行了实验。初步实验结果表明，框架消歧的精确率达到69.28%。文献[23]基于依存句法分析，并借助T-CRF模型，针对7个可激起多个框架的词元进行了框架消歧的研究，最终在940句的训练集与128句的测试集中获得了81.46%的准确率。

4.2　汉语的框架语义依存图抽取研究

汉语框架语义依存图是句子语义的一种形式化表示，汉语框架语义依存图抽取是在句子层面进行的一种深层语义分析。

框架核心语义依存图是句子核心语义依存关系的图形化表示，它由目标词、依存于目标词的框架元素的语义核心成分组成。目前，文献[24-25]已经在汉语句子的核心语义依存图抽取研究方面取得了一定的进展。提出了基于多词块标注、条件随机场模型、最大熵模型以及支持向量机模型的核心词块提取方法。经对比试验发现，基于条件随机场模型的框架元素核心词块提取获得了较好的识别性能，达到了93.17%的准确率。

另外，研究了多框架语义依存图的形式化表示技术和表示规范；收集了汉语框架网中用于抽取语义依存图的句子标注语料库。汉语框架语义依存图的抽取技术是与框架元素自动标注、框架排歧技术密不可分的，在进行汉语框架语义依存图抽取的同时，也在不断深入研究如何提高框架元素自动标注及框架消歧的准确率。

4.3　中文阅读理解问答技术研究

中文阅读理解问答技术研究[26]是基于框架语义分析，对中文阅读理解问答技术进行了研究。阅读理解问答系统的研究目的是测试计算机对一篇短文的理解能力。阅读理解的任务是给定任意一篇自然语言文章和一组给定的问题，计算机自动找到相关问题的答案。该研究构建了中文阅读理解语料库，包含121篇完整的文章，3.2万词次，1 633句，平均每篇13.5句。语料库共用232个框架对语料中有关词语进行了框架语义标注(框架名和框架元素)，由于汉语框架数据库规模有限，其中有50多个框架是从英文框架语义网直接翻译使用的。该研究构建了词层面以及句法层面共计35个特征，基于最大熵模型对中文阅读理解问题回答进行了建模，选取35个特征。考虑到特征取值之间的相关性对权重估计的影响，先对35个特征观测值矩阵进行主成分降维，选择适当的主成分个数重构特征，然后再使用最大熵模型进行建模，在测试集上的HumSent准确率达到80.18%。实验结果表明，在阅读理解问答系统中，采用特征的主成分降维方法，能有效融合全部特征信息，回避了最大熵模型中特征筛选的过程，并且提高了阅读理解系统的准确率。

4.4　汉语句子的语义相似度计算研究

汉语框架语义依存图是句子语义的一种形式化表示，计算汉语框架依存图间的相似度是解决汉语句子相似度计算的一种有效途径。

文献[27]设计并实现了基于汉语框架依存图的句子相似度计算模型。主要通过计算汉语框架依存图相似度和外围成分相似度，最终以它们的凸组合作为两个句子的相似度。句子相似度计算模型的整体流程图如图1。

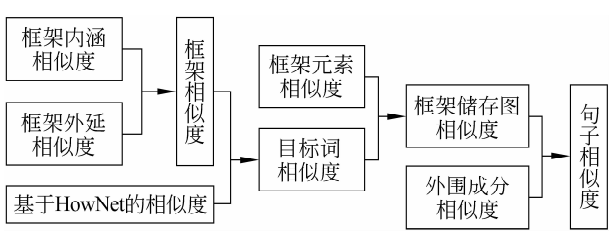

图1　句子相似度计算模型的整体流程图

具体包含：框架相似度计算、目标词相似度计算、词集合相似度计算、框架依存图相似度计算、外

围成分相似度及句子相似度计算。框架网络中两个框架之间的相似度是指它们之间的语义距离。在计算两个框架的相似度时,采用了基于最短路径关系权重乘积的外延相似度和基于框架元素组合的内涵相似度线性组合的策略。其中包括:核心框架元素相似度、框架内涵相似度及框架外延相似度。

4.5 旅游自动问答实验系统

文献[28]初步完成了面向山西旅游景点的基于本体的旅游自动问答实验系统。该系统借助汉语框架知识库在语义表达方面的独特优势,对问句进行语义角色标注,提取结构化语义信息。同时,探索了面向特定领域的本体构建方法。依据山西旅游景点网站,系统针对收集的1 566条旅游常问问句,同时用本体语言OwL进行了描述。问句包括六个方面:特色小吃、住宿、娱乐、景点、购物、交通工具。在eclipse3.3平台上进行了实验,并使用了Jena2.3工具包。目前本系统主要针对特指疑问句中的地点、时间、方法及是非疑问句部分问句进行了测试。采用召回率对系统进行评价,实验结果CFN标注后的召回率:特指疑问句(LOC)为72%,特指疑问句(TIME)为73%,特指疑问句(MEANS)为62%。

结束语

汉语框架语义网构建与应用技术研究正在处于发展时期,并得到国家863计划、自然科学基金以及社会科学基金的支持,许多课题正在研究中。目前,汉语框架语义网构建框架和句子库规模偏小,使得其自动标注系统及相关应用技术研究实验结果偏低,而且汉语框架语义网远远没有跟上英语FrameNet进展,直接影响汉语应用技术研究和推广。

致谢 本文撰写过程,及时得到美国加州大学伯克利分校国际计算机科学研究中心FrameNet项目经理Collin F. Baker给予提供的当前FrameNet资料。

参考文献

[1] Fillmore. Frame semantics[C]//Linguistics in the Morning Calm. 1982: 37-111.

[2] Baker CF, Fillmore CJ, Lowe JB. The Berkeley FrameNet project[C]//Proceedings of the CO LING-ACL'98. Montreal: ACL Press, 1998: 86-90.

[3] Subirats, Carlos. Spanish Framenet: A frame-semantic analysis of the Spanish lexicon[C]//Hans Boas, ed. Multilingual FrameNets in Computational Lexicography. Methods and Applications. Berlin/New York: Mouton de Gruyter, 2009: 135-162.

[4] Boas, HansC. BilingualFrameNetDictionariesfor Machine ranslation[C]//Proceedings of the Third International Conference on Language Resources and Evaluation. Eds. Gonz lez M. Rodr guez, and Paz Su rez C. Araujo. Vol. IV. 2002: 1364-1371.

[5] Ohara, Kyoko Hirose, Seiko Fujii, et al. The Japanese FrameNet Project: A Preliminary Report[C]//Proceedings of Pacific Association for Computational Linguistics (PACLING'03), 2003: 249-254.

[6] Salom o, Maria M M. FrameNet Brasil: Um trabalho em progresso[J]. Calidosc pio, 2009, 7:3.

[7] You L P, Liu K Y. Building Chinese FrameNet Database[C]//Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE), 2005: 301-306.

[8] 由丽萍. 构建现代汉语框架语义知识库技术研究[D]. 上海师范大学博士学位论文,2006.

[9] 刘开瑛. 汉语框架语义网(CFN)构建现状[R]. 计算语言学2008年青年学生会议大会邀请报告.

[10] Sun HL, Jurafsky D. Shallow semantic parsing of Chinese[C]//Hirschberg JB ed. Proceedings of NAACL-HLT 2004. Boston: ACL, 2004: 249-256.

[11] Xue NW, Palmer M. Automatic semantic role labeling for Chinese verbs[C]//Bramer M ed. Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence. Edinburgh: IJCAI, 2005: 1161-1165.

[12] Xue NW. Labeling Chinese predicates with semantic roles[J]. Computational Linguistics, 2008, 34(2): 225-255.

[13] 袁毓林. 语义资源建设的最新趋势和长远目标[J]. 中文信息学报, 2008, 22(3): 3-15.

[14] Cohn T, Blunsom P. Semantic role labeling with tree conditional random fields[C]//Knight K, Ng HT, Oflazer K, eds. Proceedings of CoNLL 2005. Ann Arbor: ACL, 2005: 169-172.

[15] 董静,孙乐,吕元华,等. 基于线性链条件随机场模型的语义角色标注[C]//中国中文信息学会二十五周年学术会议, 2006: 32-37.

[16] Yu JD, Fan X, Pang W, et al. Semantic role labeling based on conditional random fields[J]. Journal of Southeast University (English Edition), 2007, 23(3): 361-364.

[17] 刘鸣洋,由丽萍. 汉语感知词语的语义角色标注规则初探[C]//内容计算的研究与应用前沿,2007: 320-325.

(下转第62页)