汉语框架语义网（CFN）构建现状[1]

刘开瑛

山西大学计算机与信息技术学院，山西 太原 030006

E-mail:liukaiying99@vip.163.com

摘要：本文介绍了汉语框架语义网（CFN）构建现状，并详细的论述了 CFN 自动标注方法和构建面向中文问答处理知识库的探索过程。提出了一种基于层叠条件随机场模型的句法语义自动标注方法。该方法利用层叠条件随机场对汉语框架语义知识库（CFN）中的“陈述”“包含”“拥有”框架进行自动标注。实现了其核心框架元素自动标注结果的召回率 66.7%-75.7%，准确率 76.4%—83.9%，F 值 71.8-79.6%。

基于叙词表的本体构建的基础上，《中国分类主题词表》的领域本体构建研究，提出了基于本体的构建方法。该方法对本体中的概念体系的规范化和标准化处理提出了具体的措施和步骤。

关键词：汉语框架语义网 叙词表 层叠条件随机场 框架元素

Status Quo of Project of Constructing Chinese FrameNet

Liu Kaiying

School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China

E-mail:liukaiying99@vip.163.com

Abstract: This paper explains status quo of project of constructing Chinese FrameNet (CFN) and approach and process of CFN automatic tagging and knowledge base of face Chinese question and answer.This paper presents an approach of Syntactic and Semantic automatic tagging based on cascaded conditional random fields model. This approach takes advantage of cascaded conditional random field model to tag the frame of “Statement” “Inclusion” “Possession” in CFN automatically. Finally, core of frame elements can be tagged automatically through our automatic tagging system with its recalling rate 66.7%-75.7% and the precision rate 76.4%—83.9% and the F-measure 71.8-79.6%..

This paper is thesaurus-based ontology construction，the Construction of the Domain Ontology Based on Chinese Classified Thesaurus. The paper describes the concrete measures and steps of how to normalize the conceptual system of domain ontology and how to standardize ontology .

Key words: Chinese FrameNet, thesaurus，Cascaded Conditional Random Fields, Frame Element

1 构建汉语框架语义网（CFN）

框架语义学[1] (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，即试图用经验主义方法，寻找语言和人类经验之间的紧密关系，并研究一种可行的描述方式，表示语言和经验之间的这种关系。框架语义学把词义、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。

[1] 本文受国家 863 计划项目（2006AA01Z142）资助。

Fillmore 对框架语义学的研究，始终与构建语义知识库这一实践紧密结合，框架语义学为语义知识库的构建提供了一个基本思路（但又没有限制框架及框架元素的范围和数量），反过来，在这种实践中不断得到修正，进一步明确一些理论上的问题。

FrameNet 是由 Fillmore 亲自主持的一个基于语料库的计算词典编纂工程，从 1997 年开始于美国加州大学伯克利分校进行构建研究，截至 2008 年 3 月，共收录 10,000 词元，构建了 825 个框架，其中 6,100 个词元完成了例句标注，共标注了 13.5 万例句的框架语义信息，主要用到的语料库是不列颠国家语料库（BNC），FrameNet 数据库已在网络上公布。

就语义知识库的构建来说，FrameNet 工程的结果包括两部分：词典资源（即 FrameNet 数据库）和相关软件工具，其中 FrameNet 数据库又包括：

（1）框架库（FDB）：词语义项的语义框架及框架元素的细致描述。

（2）词汇库（LDB）：数千词和短语的配价表示，包括语义搭配模式和框架元素句法表现形式。

（3）例句库（Annotated Sentences）：一个带框架语义标注信息的语料库，包括框架元素及其句法表现（框架元素所在成分的句法功能、短语类型等）。

由于 FrameNet 描述的是词元背后的认知框架，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。但是，各种语言对同样的情境不一定是用词的形式表示出来的，也就是说，语言之间的词汇化方式有差别，因此，框架及其框架元素需要根据不同语种的个性加以改造。

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，构建汉语框架语义网（CFN）[2]。CFN 构建工作工程浩大、语义描述又存在很大的不确定性和模糊性，如果完全从零做起是不可行的。经过 2003 年至 2006 年几年努力，构建了一个以有限词语集合为描述对象的汉语框架语义网，其中，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[3]。其研究结果是构建大规模汉语框架语义网的样本，使 CFN 成为一部计算机可读、可理解的语义词典而努力。

2006 年 10 月 11 日，山西省科技厅组织由侯光南院士主持对《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。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。

2006 年以来我们建立了近 300 个框架，现在谈谈在 CFN 的句子自动标注工作和 CFN 在旅游领域应用开始探索工作。

2 CFN 的句子自动标注工作

目前 CFN 的句子标注工作，是对给定的句子、目标词及其框架，标注目标词的各个直接从属成分所承担元素类型，并标注该短语（或词）的短语类型和句法功能，最终完成三个层面的标注[2]。短语类型标注就是标注框架元素所在的整个短语的句法性质，句法功能的标注时只有作目标词的框架元素成分才标注句法功能。

CFN 语义标注的基本单位是由一个框架承担词和若干框架元素组成的语义结构。框架承担词包括动词、形容词和名词，它们是标注工作的着眼点，因此我们称之为目标词，用tgt（tager）作为标记；框架元素在句法上对应的是目标词直接支配的句法成分，通常是名词性成分。实际上，CFN 语义标注的基本单位就是一个谓词-论元结构。对于只包含一个谓词-论元结构的简单句来说，框架语义结构就是谓词和主语、宾语等所构成的结构单位；但是对于包含多个或多层谓词-论元结构的句子来说，框架语义标注单位即包括核心谓词及其支配成分，也包括小句宾语等内部成分，甚至包括定中结构，可见，CFN 语义表示的基本单位并不是句子。

例句 1：<tot-np-subj 《 w 动画 n 卷 n 》 w >＜tgt=包含 包含 v >＜par-np-obj “ w Animation ws 动画 n 篇 q ” w 和 c “ w Character ws 角色 n 篇 q ” w >。w

在上述例句 1 中，目标词为“包含”，“<tot-np-subj 《 w 动画 n 卷 n 》 w >”中“tot”是整体，“np”表示名词短语，“subj”表示目标词“包含”所支配的句法功能（主语）；“<par-np-obj “ w Animation ws 动画 n 篇 q ” w 和 c “ w Character ws 角色 n 篇 q ” w >”中“par”是目标词“包含”的部分，“np”表示名词短语，“subj”表示目标词所支配成分的句法功能（宾语）。

例句 2：这 r 其中 r 道理 n ， w 浅显 aq 易懂 aq ， w 妇孺皆知 ia ， w 身为 v 农经 jn 博士 n ， w＜freq-dp-adva 经常 d >＜tgt=陈述 自夸 v >＜msg-pp-obj 对 p 历史 n 、 w 哲学 n 有 v 很 d 深 aq 研究 v >＜null 的 u >＜spkr-np-head 李登辉 nh >， w 对 p 此 r 不 d 会 v 不 d 晓 v 。 w

从上述的例句 2 中可以看出 CFN 标注句子并不是对整句进行标注，而是先确定词元，对和词元有支配关系的部分进行标注。在例句 2 中标注的部分“经常自夸对历史哲学有很深研究的李登辉”仅仅是整句中的一个定中结构的短语，“<freq-dp-adva 经常 d >”中“freq ”表示目标词“自夸”陈述的频率，“dp”表示为副词短语类型，“adva”表示为目标词“自夸”的句法功能（状语）。注意，这里的句法功能是指相对于目标词的句法功能。

CFN 自动标注的句子是从北大 CCL 现代汉语语料库中初选（包括当代的报刊、文学、应用文、电视电影，现代的文学、戏剧、网文、口语等方面的文章）。经 F-2000 自动分词和词性标注软件运行后，并手工调整明显错误，再进行人工标注。人工构建语义语料库是一件非常浩大的工程。为了让语义标注语料能够真正的发挥其作用，研究语义角色自动标注的方法是必需的。

3 基于层叠条件随机场的 CFN 自动标注方法

条件随机场模型（CRF）作为一种判别型模型，将随机场理论和最大熵原理有效地融合到了序列标注的过程中，并在标注序列上进行全局归一化。因此，CRF 模型可以有效地提高序列标注的精度。但条件随机场直接进行框架元素、短语类型和句法功能的三层标注，消耗的时间和空间特别多，因此需要引入了多个层次的条件随机场模型用于识别这类问题。本文提出了一种基于层叠条件随机场的 CFN 自动标注方法。该方法在低层条件随机场模型中解决了框架元素的识别，将识别结果传递到上层短语类型识别的条件随机场模型，再将前两层识别

结果传递到上层句法功能识别的条件随机场模型，其低层模型为上层模型提供决策支持[4]。

自动标注的基本单元可以是句法成分（Constituent）、短语（Phrase）、词（Word）或者依存关系（Dependency Relation）等等。现在多数的语义标注系统通常都是以句法成分为基本的标注单元的。例如：基于最大熵分类器的语义角色标注[5]中以句法成分为基本单元，使用最大熵分类模型基于Prop Bank语料，研究了语义角色的自动标注问题。这种策略，在句法分析比较成熟的语言（如英文等）上表现较好。然而，在其它语言上，很难自动获得这种深层句法分析的结果，而且现有的句法分析系统，在通用领域表现欠佳。尤其是汉语，很难获得真实语料的句法分析结果，为此以词作为自动标注的基本单元。

首先使用“BIO”策略标注句子库中的句子。特征选取词、词性和相对于目标词的位置，并且在进行后一层自动标注时把前一层的自动标注结果作为特征。

3.1 “陈述”框架自动标注

“陈述”框架下句子库中的句子数为1393句，按5：5的比例分为训练集和测试集。其自动标注的结果如下所示：

“陈述”框架	准确率P	召回率R	F值
第一层框架元素自动标注	75.7%	58.9%	66.3%
第二层短语类型自动标注	66.8%	52%	58.5%
第三层句法功能自动标注	61.6%	47.9%	53.9%

表1：“陈述”框架自动标注结果

从统计结果发现，第一层框架元素自动标注的F值最高，然后逐层递减，主要原因是由于错误的累积。“陈述”框架的框架元素的F值在50%以上的进行统计结果如下：

陈述	训练集	测试集	准确率P	召回率R	F值
媒介*（medium）	653	645	74%	72.20%	73.10%
信息*（msg）	265	280	82.80%	68.60%	75%
说话者*（spkr）	425	412	73.60%	61.70%	67%
听话者（add）	135	130	82.1%	73.8%	77.7%
时间（time）	136	141	80%	56.7%	66%
结果（result）	14	13	81.8%	69.2%	75.5%
合计	1628	1621	76.4%	67.7%	71.8%

表2：“陈述”框架的框架元素标注结果（带*号的为核心框架元素）

“陈述”框架下框架元素的自动标注结果仅列出了F值在50%以上的框架元素，其中的框架元素标注的F值为71.8%，但非核心框架元素的标注效果相差很大，非核心框架元素“听话者（add）”“时间（time）”两个框架元素的自动标注结果的F值在50%以上。但非核心框架元素“ 修饰（manr）”“关涉（top）”“形容（depic）”的标注效果很不理想，其中主要的原因是由于数据稀疏，非核心框架元素“程度（degr）”在训练集和测试集中都没有出现过，但非核心框架元素“致因（cau）”在训练集中出现过两次但在测试集中没有出现过。通用非核心框架元素标注效果最好的为“结果（result）”F值达到75.5%，“结果（result）”在语料库中出现的不多但是它的结构单一，大都是“<tgt 道 v > <result-vp-comp 出 v >”这种结构。其它通用非核心框架元素由于数据稀疏识别效果很不理想。

3.2 “包含”框架的自动标注

“包含”框架下句子库中的句子数为1359句，按5：5的比例分为训练和测试集，各个实验的训练集和测试集是相同的。其“包含”框架的自动标注结果如下所示：

“包含”框架	准确率P	召回率R	F值
第一层框架元素自动标注	84%	72.9%	78%
第二层短语类型自动标注	78.9%	68.4%	73.3%
第三层句法功能自动标注	76.6%	66.5%	71.2%

表3：“包含”框架的自动标注结果

对“包含”框架的框架元素的F值在50%以上的进行统计结果如下：

包含	训练集	测试集	准确率P	召回率R	F值
部分*（par）	703	702	90.6%	87.5%	89%
整体*（tot）	668	669	77%	64.1%	70%
角色范围（sco_role）	30	30	71.4%	66.7%	69%
程度（degr）	11	12	58.3%	58.3%	58.3%
合计	1412	1413	83.9%	75.7%	79.6%

表4：“包含”框架的框架元素自动标注结果（带*号的为核心框架元素）

其中核心框架元素的标注效果都很好，非核心框架元素由于数据稀疏标注效果几乎未识别出，通用非核心框架元素“角色范围（sco_role）”和“程度（degr）”两框架元素的自动标注的F值在50%以上，其它通用非核心框架元素的标注效果并不理想。

3.3 “拥有”框架的自动标注

“拥有”框架下句子库中的句子数为697句，按5：5切分，分为训练集和测试集。“拥有”框架的自动标注结果如下所示：

“拥有”框架	准确率P	召回率R	F值
第一层框架元素自动标注	84.5%	57.4%	68.4%
第二层短语类型自动标注	82.8%	56.3%	67%
第三层句法功能自动标注	80.7%	54.7%	65.2%

表5：“拥有”框架的自动标注结果

对“拥有”框架的框架元素的F值在50%以上的框架元素统计结果如下所示：

拥有	训练集	测试集	准确率P	召回率R	F值
拥有物*（pos）	352	353	87.1%	79%	82.9%
所有者*（own）	328	330	71.9%	45.6%	55.9%
方法（mns）	3	2	1	50%	66.7%
合计	683	685	81.2%	62.9%	70.9%

表6：“拥有”框架的框架元素自动标注结果（带*号的为核心框架元素）

其中核心框架元素的标注效果都很好，非核心框架元素由于数据稀疏未识别出，通用非核心框架元素“方法（mns）”的F值为66.7%，其它通用非核心框架元素的标注效果并不理想。

3.4 实验分析

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。以下是“陈述”“包含”“拥有”框架中涉及到的词元。

框架	词元
陈述（102）	介绍 v，评述 v，评议v，评说v，论述 v，讲述 v，讲授v，讲解 v，指出 v，提出 v，提到v，提起v，阐述 v，阐明 v，阐释 v，阐发 v，承认v.....
包含（28）	包含v，包括v，含有v,涉及v,含v,组成v,分为v，涵盖v，列举 v，分成 v，列入 v，分 v，共收 v，共分 v，覆盖 v，融入 v，构成 v ……
拥有（23）	属于v，所有v,拥有v，持有v,占有v,有v，没有v,缺v,缺乏v,缺少v,短缺v,短少v,欠缺v,具有v,只有v，附有 v，收有 v，共有 v ……

我们虽然只解决了三个框架可是解决了153个词元的标注问题。其他框架还在试验中。

4 CFN 在旅游领域应用开始探索

要想建立质量较好的中文旅游问答系统，离不开语义的理解和处理，而语义分析离不开知识；借鉴本体的理论和方法，建立面向旅游问答的中文语义知识库，是现在重要的研究方向。参考国内多家研究成果[7][8]后，我们认为借助《中国分类主题词表》现有的词汇体系结构来进行改造是可行的。在此基础上，我们展开了旅游中文问答系统的旅游景点知识库构建。

根据本体的理论，建立中文问答系统所需要的知识库可分为以下这么几个层次：顶层本体，领域本体，任务本体和应用本体[9]。《中国分类主题词表》进过改造后充当了旅游问答系统中的顶层本体，经过与旅游知识的结合形成了我们的领域本体。

叙词表建模的模式有面向概念和面向术语两种。中科院情报中心的毛军在叙词表的RDF中[10]，提出将叙词用概念和词汇两个层次的资源来描述，从某种程度上来看便是利用了面向概念的模式。深圳大学曾新红的文献[7]中分析到，存在于人类思想领域的某个抽象概念构成一组可能的标签，换句话说就是术语代表概念。在面向概念模式中，叙词表的等级和相关关系在概念之间声明。概念是这个概括层次结构中的节点。在一个多语种叙词表中，等同关系在概念之间声明。相对于传统的面向术语模式，面向概念模式将概念和术语分离，更容易维护和更新，因为对术语的修改不会干扰概括层次结构本身。因此，我们在构建旅游景点知识库根据我们的需要和实际情况也采用了面向概念模式。《中国分类主题词表》形式上是面向术语的，但实质上每一个主题词都可以当成一个概念，这样在改造层本体的过程中具有很大的优势和便利。

构建本体的方法很多，因为我们是以《中国分类主题词表》为纲建立领域本体，所以，我们也提出了针对旅游本体构建的一套方法：

A. 确定本专业的领域和范畴。旅游是一个包含很广泛的领域，我们建立的旅游本体主要是针对景点介绍。从中总结出一条本体构建的方法，以方便以后工作的开展。

B. 定义类和类的等级体系。我们最大程度的保持了《中国分类主题词表》的体系结构，从中筛选了 12 个大类，自上而下的建立了体系的总体布局，结合旅游景点所涉及到的文学、艺术、建筑、民俗等各个方面的具体情况，定义了旅游本体的类及类的等级体系，并加上了一定的语义关系。

C. 定义类的属性。我们的知识库在定义类的属性的具体方法上我们要考虑结合 CFN 的优势。根据具体的框架选取来定义类的属性。

D. 定义类的分面。这一步同上一步（定义类的分面）具有同样的特点，是根据 CFN 框架选择的具体情况来定义。

E. 创建实例。创建实例，其实也是一个添加具体文本的过程，在我们的本体中每一个类相对应着一个概念，而类下的实例（即对象）是对应着一个实体（即文本）。创建实例其实穿插在整个本体构建的过程中，也体现了我们自底向上和自顶向下相结合的构建思想。

F. 对本体进行形式化编码。这一步，我们采用了Protégé工具进行本体的形式化编码。

另一步工作，就是知识库中应用到的语料库的建立。现在我们已经从网络上搜集了，不同景点、景区的文档 833 篇，并对其中 23 个景点的 103 篇语料进行统计，选取框架。对框架进行分析，定义类的属性和分面。主要是在 23 个景点的基础上建立一个小规模的语料库，对其进行预处理。然后进行框架元素，句法功能的标注，完成知识库样本库的建立。

参考文献

[1] 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.

[2] 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

[3] 郝晓燕，刘伟，李茹，刘开瑛．汉语框架语义知识库及软件描述体系．中文信息学报，2007.5，21（5）

[4] 周俊生，戴新宇，尹存燕，陈家骏．基于层叠条件随机场模型的中文机构名自动识别．电子学报，2006 年 34 卷 5 期

[5] 刘 挺，车万翔，李 生．基于最大熵分类器的语义角色标注．软件学报，2005

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

[7] 曾新红．《中国分类主题词表》的 OWL 表示及其语义深层揭示研究．情报学报，2005，4

[8] 薛云，叶东毅，张文德．基于《中国分类主题词表》的领域本体构建研究．情报杂志，2007，3

[9] 张亮，黄河燕，胡春玲．中文问答系统模型研究．情报学报，2006，4

[10] 毛军．基于 RDF 的叙词表研究．情报学报，2003，4