文章编号：1003-0077（2014）00-0000-00

汉语核心框架语义分析*

石佼[1]，李茹[1,2]，王智强[1]

（1.山西大学 计算机与信息技术学院，山西 太原 030006；

2. 山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006）

摘要：汉语核心框架语义分析是从框架语义角度，通过抽取句子的核心框架，获取汉语句子的核心语义骨架。该文将核心框架语义分析分为核心目标词识别、框架选择和框架元素标注三个子任务，基于各个子任务的不同特点，采取最大熵模型分别对核心目标词识别与框架选择任务进行建模；采用序列标注模型条件随机场对框架元素标注任务进行建模。实验在汉语框架网资源的 10831 条测试语料中显示，核心目标词识别和框架元素标注 F 值分别达到 99.51%和 59.01%，框架选择准确率达到 84.73%。

关键词：汉语框架网；核心框架语义；语义分析

中图分类号：TP391　　**文献标识码：**A

Chinese Core Frame Semantic Analysis

SHI Jiao[1]，LI Ru[1,2]，WANG Zhiqiang[1]

(1.School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China;

2. Key laboratory of Computation Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract：Based on the theory of frame semantics, Chinese core frame semantic analysis is to extract the core frame semantic representation to analyze the semantic content of the sentence. We solve this problem using a three-stage learning model. Taking the tasks' different characteristics into consideration, we choose Maximum Entropy model to take core target in the sentential contexts and predict frame for the core target, choose Conditional Random Field model to label the frame elements defined in Chinese FrameNet. Experimental results on the 10831 exemplified sentences show that the F score of core target identification and frame element identification reached 99.51% and 59.01% respectively, the frame identification with 84.73% accuracy.

Keywords：Chinese FrameNet；Core frame semantic；Semantic analysis

1 引言

汉语核心框架语义分析[1]是以框架语义学[2]为理论基础，基于汉语框架网（Chinese FrameNet，CFN）[3~4]的语义表示与标注资源，通过抽取句子中的核心目标词及其所激起的核心框架语义场景，达到汉语句子级核心语义分析的目的。如对于例句“20年后，他回到了出生时的老家。”，在不区分句子中核心语义的情况下，对其进行框架语义分析的结果如图1所示：

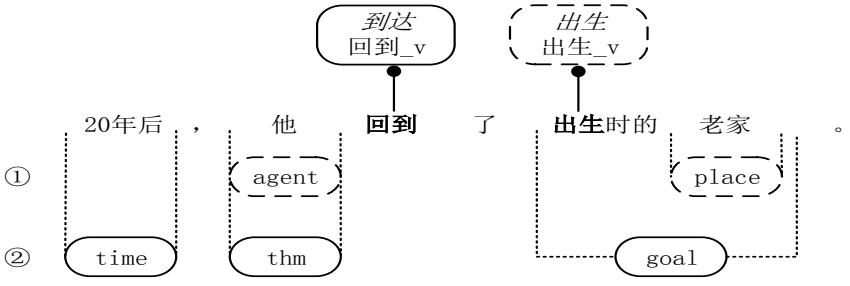


图 1 例句框架语义分析结果

* **收稿日期：** 2014-07-21　　**定稿日期：**

基金项目：国家自然科学基金项目(No. 61373082)；山西省回国留学人员科研资助项目(2013-015)；汉语框架语义网资源及技术平台(2014091004-0103)；国家语委“十二五”科研规划项目(YB125-19)；国家“八六三”高技术研究发展计划基金项目(2006AA0lZ142)

图中①和②是分别以“出生”和“回到”为目标词进行框架语义分析的结果。其中，“回到”和“出生”分别能够激起“到达”和“出生”框架，“他”和“老家”分别填充“出生”框架下的“agent”和“place”框架元素，“20年后”、“他”和“出生时的老家”分别填充“到达”框架下的“time”、“thm”和“goal”框架元素。显然，基于“到达”框架的语义分析结果能够获得句子的核心语义，而基于“出生”框架的语义分析结果只能获取句子部分语义信息，前者即为本文研究的主要内容。核心框架语义分析能够有效获取句子的核心语义，为自然语言处理领域提供了一种新的语义分析方法。

与核心框架语义分析相关的研究主要是 SemEval-2007 国际语义评测会议中基于 FrameNet[5]语料库的“框架语义结构抽取”任务[6]。Cosmin Adrian Bejan 等[7]利用支持向量机模型和最大熵模型，实现了一个线性框架语义结构抽取系统，系统整体抽取结果 F 值为 39.25%；Richard Johansson 等[8]提出了一种基于依存句法分析的框架语义结构抽取方法，通过扩充 FrameNet 词元库，来提高系统性能，最终抽取结果 F 值达到 48.8%；Dipanjan Das 等[9]针对目标词识别、框架分配和框架元素标注任务，分别使用半监督学习方法和快速对偶分解算法构建模型，F 值达到 68.45%。随着 FrameNet 语料库资源不断丰富，针对英文语料的框架语义结构分析已取得一定成果。

目前，汉语语义分析的主要手段仍集中在基于“谓词-论元”结构的语义角色标注任务上。使用不同的机器学习方法，针对基本特征及其组合特征，对语义角色标注任务进行了研究。Xue 等[10]使用汉语 PropBank 语料库语料，运用最大熵分类器，在自动的句法分析基础上进行语义角色标注，准确率达到 71.9%；李济洪[11]通过 IOB 策略将语义角色标注问题转化为词序列标注问题，采用条件随机场模型，基于统计学中的正交表，挑选最优特征模板，在给定句子中的目标词以及目标词所属的框架的情况下，F 值达到 61.62%。语义角色标注虽能达到浅层语义分析的目的，但由于其只针对句中给定的谓词标注语义角色，因此并不能对整个句子进行详细的语义分析。核心框架语义分析以“框架语义学”为理论基础，从框架语义角度刻画句子的语义，通过抽取句子的核心框架语义结构，达到分析句子完整语义的目的。

本文将汉语的核心框架语义分析任务拆分为核心目标词识别、框架选择和框架元素标注三个子任务，结合同义词词林资源[12]，使用基于贪心策略的特征选择算法，分别建立不同任务的机器学习模型。第 2 节对汉语核心框架语义分析的相关概念和问题进行描述；第 3 节介绍任务的特征选择算法；第 4 节为实验及结果分析；最后进行结论与展望。

2 核心框架语义分析相关概念与问题描述

2.1 核心框架语义分析相关概念

定义 2.1[13] 汉语框架网.

汉语框架网（Chinese FrameNet，CFN）是以 Charles J. Fillmore 的框架语义学为理论基础，参照加州大学伯克利分校的 FrameNet 工程，构建的以汉语真实语料为依据，可供计算机使用的汉语词汇语义知识库。

汉语框架网由框架库、句子库和词元库三部分组成。框架库以框架为单位，对词语进行分类描述，明确给出框架的定义和这些词语共有的语义角色即框架元素，并描述该框架和其他框架之间的概念关系；句子库包含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；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。为了易于理解框架语义分析任务，以下给出“框架”的概念。

定义 2.2[13] 框架.

框架是指与一些激活性语境相一致的结构化范畴系统，它是储存在人类认知经验中的图式化情境，是理解词语的背景和动因。

以“到达”框架为例，在框架库中的简略描述如表 1 所示：

表 1“到达”框架简表

框架名	到达|Arriving
定义	指转移体到达目标的过程。目标可直接表达出来，或从上下文中得到理解，或者动词本身隐含目标之义。
核心框架元素	目标(goal) ，转移体(thm)
非核心框架元素	伴随者(thm_c) ，形容(depic) ，目标状态(goal_c) ，修饰(manr) ，方法(mns) ，传送模式(mot) ，路径(path) ，源点(src) ，时间(time)
词元	到达 v，来到 v，进入 v，抵达 v，返回 v，走到 v，走进 v，赶到 v，回来 v，归来 v，到 v，回到 v

定义 2.3 核心目标词.

在一条包含多个目标词的句子中，如果某个目标词激起的框架及其在句中所支配的框架元素依存项相比其他框架更能完整表达句子的核心语义，该目标词即为核心目标词。

如引言图 1 所示的例句中，“回到”和“出生”均为目标词，核心目标词是“回到”。

定义 2.4[13] 框架元素.

框架元素又称框架语义角色，体现一个框架的语义参与者。框架元素包括核心与非核心框架元素，其中核心框架元素显示框架的个性，而非核心框架元素表达框架中的一些通用语义成分，如时间、地点等。

如表 1“到达”框架中的“目标”、“转移体”是“到达”框架的核心框架元素；“伴随者”、“方法”、“路径”、“源点”、“时间”属于非核心框架元素。

2.2 核心框架语义分析问题描述

核心框架语义分析任务是根据句中的核心目标词元及其激起的框架，确定该框架在句中支配的框架元素。

汉语框架核心语义分析任务问题可形式化定义为：

核心目标词识别 给定句子 $S=\{w_1,w_2,...,w_s\}$，对于句中能够激起框架的目标词元集合 $T=\{t_1,t_2,...,t_n\}$，识别出核心目标词 t_i，形式化表示为：

$$\hat{t}=\arg\max_{t_i\in T} R(t_i\mid S) \qquad (1)$$

框架选择 根据当前上下文场景 S，为核心目标词 t_i 分配一个合适的语义框架 f_i。若 t_i 能够激起多个框架，即为歧义词元，设 t_i 激起的框架记为 $f=\{f_1,f_2,...,f_m\}$，则框架选择任务可形式化地表示为：

$$\hat{f}=\arg\max_{f_i\in f} R(f_i\mid S,t_i) \qquad (2)$$

框架元素标注 给定目标词 t_i 及其所属框架 f_i，设 f_i 支配的框架元素集合为 $R_{f_i}=\{r_1,r_2,...,r_j\}$，为语义场景 S 中的连续子集 $X_{f_i}=\{w_{i+1},w_{i+2},...,w_{i+t}\}$ 填充其对应的框架元素 r_R，设 $\Delta_q(r_i)$ 是语义场景 S 中的框架元素的集合，则框架元素标注任务可形式化表示为：

$$\Delta_q(r_i)\leftarrow\arg\max_{X_q\in S} \mathrm{R}(X_q\mid w_q,f_i,r_i,S) \qquad (3)$$

3 核心框架语义分析建模

核心框架语义分析的研究目的是抽取句子的核心语义表示，本文将核心框架语义分析研究拆分为核心目标词识别、框架选择、框架元素标注三个子任务，根据各子任务不同的特点，分别建模。

3.1 模型

3.1.1 核心目标词识别与框架选择模型

核心目标词识别和框架选择任务的标注着眼点是句中的一个词，识别该词是否是核心目标词或该词所属框架。因此，本文将这两项任务看作分类问题来解决，使用最大熵思想构建分类模型。

在本实验中，用向量 X 表示候选词是否为核心目标词或其所属框架，用 y 表示候选目标词是否为核心目标词或者其所属框架， $p(y|X)$ 为预测 X 为 y 的概率，熵定义为：

$$H(X)=-\sum_{X,y}p(y|X)\log p(y|X) \quad (4)$$

采用拉格朗日乘数法求解最大熵，计算公式为：

$$p(y|X)=\frac{1}{Z(X)}\exp(\sum_{i}^{n}\lambda_i f_i(X,y)) \quad (5)$$

$$Z(X)=\sum_{y}\exp(\sum_{i}^{n}\lambda_i f_i(X,y)) \quad (6)$$

其中， f_i 表示每个特征， n 代表特征总数， λ_i 为特征的权重，每个特征对词性选择的影响大小由特征权重 λ_i 决定。

3.1.2 框架元素标注模型

框架元素标注是给定目标词及其所属语义框架，根据该框架在 CFN 框架库中的定义，对该框架支配的相应框架元素进行标注。本文使用 IOB[14]策略，选择词作为标注单元，把框架元素看作一个词序列的集合，将框架元素标注任务转化为句子层面的词序列标注问题，用{B-X,I-X,O}标记角色标注集合，标注示例如下。

20|B-time 年|I-time 后|I-time ，|O 他|B-thm **回到**|O 了|O 出生|B-goal 时|I-goal 的|I-goal 老家|I-goal 。|O

其中，“回到”是目标词，激起“到达”框架。B-time 代表“到达”框架中角色“time”的开始；I-time 表示“到达”框架中角色“time”的延续；非框架角色元素用 O 标记。

用随机变量 X 表示待标注的数据序列，随机变量 Y 表示相应的标注序列，对 Y 中的每一个 Y_i 在有限状态标记中取值。设输入的序列标记为 $X=x_1x_2...x_n$ ，输出的序列标记为 $Y=y_1y_2...y_n$ ，则在已知输入序列 X 的条件下，输出为 Y 的概率为：

$$P(y|x)=\frac{1}{Z(X)}\left\{\exp\left(\sum_{i}\sum_{k}\lambda_k f_k\left(y_{i-1},y_i,X\right)\right)+\exp\left(\sum_{i}\sum_{k}\mu_k g_k\left(y_i,X\right)\right)\right\} \quad (7)$$

其中， x 代表句中的词， y 代表词 x 填充的框架元素， $f_k(\cdot)$ 表示输出序列 Y 中位置为 i 和 $i-1$ 的转移特征， $g_k(\cdot)$ 为输入序列 X 与输出序列 Y 在 i 位置的特征 y_i 之间的特征， λ_k 和 μ_k 是权重。

3.2 特征选择

特征选择的目的是从候选特征中选出与任务最相关的特征子集。本文针对核心框架语义分析任务，设置基本特征和同义词词林五层编码信息两类特征。特征描述如表 2 所示：

表 2 特征描述

编号		特征名称	特征描述	窗口大小
基本特征	F1	词	当前词词形特征	[-1,1]、[-2,2]、[-3,3]
	F2	词性	当前词词性特征	
	F3	命名体	当前词命名体特征	
	F4	依存关系	当前词依存关系特征	
词林特征	F5	1 级特征	词林编码的大类信息	
	F6	2 级特征	词林编码的大类和中类信息	
	F7	3 级特征	词林编码的大类、中类和小类信息	
	F8	4 级特征	词林编码的大类、中类、小类和词群信息	
	F9	5 级特征	词林编码全部信息	

特征选择的关键是挑选出包含尽可能多的与目标类相关的特征信息。本文采用基于贪心策略的特征选择算法，通过打分策略，选出最优特征模板。主要思想是：在给定的特征候选集中，每次从中选出一个特征加入基本特征模板中，对其预测结果打分，选取最好结果加入基本特征模板中，直到相邻两次打分不再增加。算法如下：

输入： 预处理语料，特征集合 $f=\{f_1, f_2, ..., f_n\}$，训练集 $train\_i$，测试集 $test\_i$

输出： 最优特征模板 F

```
初始化 F = ∅，Score[1]=0,Num=0
for j=1 to n                          //将 f 中的特征依次加入特征模板中进行训练
    for i=1 to 5                              //5-fold 交叉验证
    R_j += R_j
    end  for
    R_j = R_j / 5                     //将 F ∪ f_j 作为特征模板，用模型训练得到结果评分 R_j
     if  R_j > Score[1]{              //如果评分 R_j 高于前一轮，则将第 j 个特征加入集合 F
       Score[1] = R_j
       F = F ∪ f_j }
     else  if  R_j = Score[1]{        //如果评分 R_j 与前一轮相同，比较特征 j 和 j−1 的优
                                      先级，确定是否加入特征集合 F 中
       f_max = compare(f_j, f_{j−1})
       F = F ∪ f_max }
     else if  R_j < Score[1]{
          Num++;
           continue;
          if (Num ==2)                //直到连续两次评分不再提高，循环结束
           break;}
    end for
  return  F
```

4 实验设置及分析

4.1 实验语料

实验语料构建源于 CFN 的例句库，共 10831 条标注例句。本文使用哈尔滨工业大学社会计算与信息检索研究中心的语言处理集成平台 LTP[15]对语料进行预处理。实验语料统计结果如表 3 所示：

表 3　实验语料统计结果

	例句数	词元数	框架数
语料集	10831	12200	120

4.2 评价指标

本文使用准确率、召回率和 F 值三个评价指标对实验结果进行评价。设 A 为实验模型预测正确的数据个数，B 为实验预测出的实验数据个数总和，C 为测试集中正确标注的数据个数，则：

$$P=\frac{A}{B} \tag{8}$$

$$R = \frac{A}{C} \tag{9}$$

$$F = \frac{2PR}{P+R} \tag{10}$$

其中，P 表示准确率，R 表示召回率，F 表示准确率和召回率的加权平均值。

4.3 核心目标词识别实验设置及结果分析

针对核心目标词识别任务，设置三组特征模板 T1、T2、T3，分别测试窗口大小和基本特征设置对核心目标词识别模型的影响，实验结果如表 4 所示：

表 4 核心目标词识别窗口大小实验结果

模板编号	特征描述	F 值		
		[−1, 1]	[−2, 2]	[−3, 3]
T1	基本特征（词和词性）	95.99%	97.32%	97.36%
T2	在 T1 基础上添加当前目标词的词形	98.08%	97.85%	98.05%
T3	在 T2 基础上添加当前目标词的词性	98.58%	98.35%	98.16%

由表4可以看出，窗口大小设为1时，在特征模板中加入目标词词形和词性特征对提高核心目标词识别模型性能效果显著，核心目标词识别效果相对较好。因此，以T3为基线模板，构建最优核心目标词识别模型。

选取表2中的基本特征，使用3.2节提出的特征选择算法，构建最优特征模板，实验结果如表5所示。从表中可以看出，特征模板选择最优模型F值为99.50%。在最优特征模板上分别增加同义词词林编码信息，由表5的数据可以看出，F值并没有升高，反而下降。分析其原因，可能是因为核心目标词识别任务主要与目标词元的属性特征相关，在已有最优特征模板中加入同义词词林编码信息后，造成特征信息冗余，导致模型性能下降。

表 5 核心目标词识别实验结果

实验编号	特征模板	F值
1	特征模板选择最优模板	99.50%
2	最优模板+1级特征	99.39%
3	最优模板+2级特征	99.29%
4	最优模板+3级特征	99.19%
5	最优模板+4级特征	99.29%
6	最优模板+5级特征	99.29%

4.4 框架选择实验设置及结果分析

目前，CFN 词元库中共 332 个歧义词元，本文选取 47 个常见歧义词元作为研究对象构建实验例句集。由于测试集中有些句子包含多个歧义目标词，因此只使用准确率而不使用召回率来评价该实验模型。设置开窗口范围内的词形和词性为基本特征，对所有词元构建统一特征模板，测试窗口大小对模型效率的影响。实验结果如表 6 所示：

表 6 框架选择窗口大小选择结果

窗口大小	特征模板	平均准确率
[-1,1]	词、词性	77.72%
[-2,2]		77.33%
[-3,3]		77.64%

由表6可知，窗口大小设为1时，框架选择准确率最高，为77.72%。选取表2中的基本特征，使用3.2节提出的特征选择算法，得到的最优特征模板准确率为78.44%。在最优特征模板上依次增加同义词词林编码信息，实验结果如表7所示：

表 7 框架选择统一特征模板选择结果

实验编号	特征模板	最优模板准确率
1	自动特征模板选择最优模板	78.44%
2	最优模板+1级词林编码特征	78.11%
3	最优模板+2级词林编码特征	78.44%
4	最优模板+3级词林编码特征	78.43%
5	最优模板+4级词林编码特征	78.57%
6	最优模板+5级词林编码特征	78.51%

加入同义词词林4级和5级编码后，框架选择模型性能都有明显提升，且加入4级特征，模型效率最佳。原因如下：

（1）框架选择任务主要是根据当前歧义词所处的上下文语义场景判断其所属的框架，因此，特征选择中，能否获得有效的上下文特征信息至关重要。同义词词林通过对词语编码，将词语概念抽象，能够表达另一层面的语义场景关系，为框架判别提供词元特征层面更丰富的信息。

（2）加入同义词词林4级编码相比加入其他级的词林编码具有更好的框架选择效果，因为1级、2级、3级信息都是类别特征，对概括上下文信息并没有起到突出表达的作用，4级信息为词群信息，能够很好的抽象当前歧义词开窗口范围内词群特征。

文献[16]提出，针对每个歧义词元分别构建特征模板，使用词、词性、命名体和依存关系能达到更好的排歧效果。本文针对每个词元，在候选特征集中加入同义词词林资源 4 级编码信息，实验结果如表 8 所示：

表 8 各词元单独设置特征模板准确率结果

词元	准确率	词元	准确率	词元	准确率
帮助	97.23%	下降	79.77%	关注	82.85%
包	81.92%	想	94.87%	分为	69.09%
表明	84.55%	修	90.41%	包租	95.42%
表示	87.23%	有	79.10%	回响	96.66%
处理	87.86%	制造	76.40%	坚持	79.27%
倒	88.18%	装载	75.53%	强调	86.77%
合计	96.66%	租借	79.06%	断定	85.13%
回顾	78.04%	租用	81.48%	显示	89.86%
叫	92.42%	坐	79.43%	概括	91.22%
扣	86.77%	联系	93.54%	涉及	76.48%
属于	77.98%	推翻	89.77%	清楚	76.84%
说明	74.50%	归档	73.33%	熟悉	85.42%
探索	66.03%	询问	85.33%	破坏	91.18%
投	90.95%	相信	95.07%	论述	68.39%
推倒	78.57%	供给	78.12%	设想	92.91%
明白	96.56%	让	98.06%	-	-

针对每个歧义词元分别构建特征模板，最终得到的平均准确率为 84.73%，特征平均长度为 2.7。相比文献[16]，平均准确率提高了 2.24%，平均特征长度缩短了 1.2。

4.5 框架元素标注实验设置及结果分析

针对框架元素标注任务，设置 7 组特征模板，开三种窗口大小，分别实验，并计算语料平均 F 值，实验结果如表 9 所示：

表 9 框架元素标注模板实验结果

编号	特征模板	[-1, 1]	[-2, 2]	[-3, 3]
		mF值	mF值	mF值
1	F1、F2	54.23%	50.91%	50.45%
2	F1、F2、F3、F4	58.16%	56.67%	54.65%
3	F5+ F1、F2、F3、F4	59.01%	57.49%	55.88%
4	F6+ F1、F2、F3、F4	58.37%	57.36%	55.35%
5	F7+ F1、F2、F3、F4	58.57%	57.00%	55.22%
6	F8+ F1、F2、F3、F4	58.64%	56.83%	54.74%
7	F9+ F1、F2、F3、F4	58.19%	56.40%	54.62%

在三个不同窗口大小下，F 值的折线图如图 2 所示：

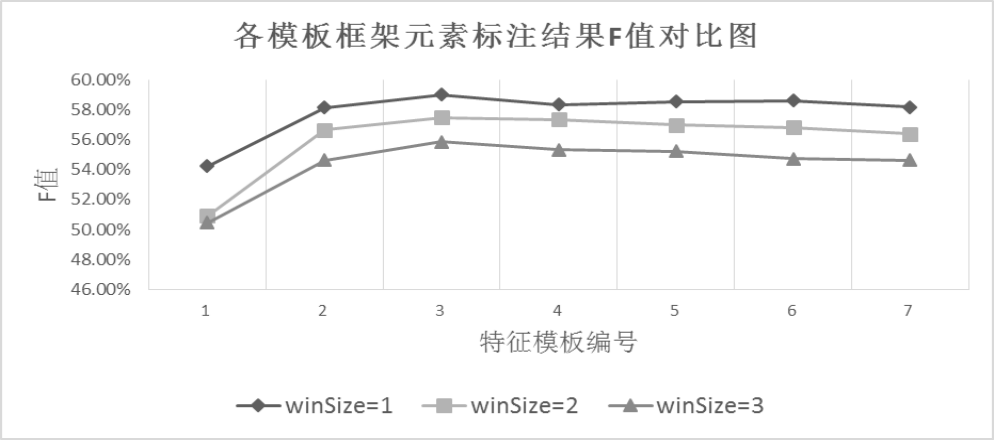


图 2 框架元素标注 F 值比较折线图

根据上图，可以得出以下结论：

（1） 实验 F 值在 58%左右，究其原因可能与实验数据稀疏相关；

（2） 窗口大小设为 1 时，实验效果最佳。因为框架语义分析与其目标词所处的位置及其紧邻上下文信息密切相关，远距离信息可能会导致加入冗余信息，造成模型性能下降；

（3） 2 号实验，加入特征 F3 和 F4 后，模型性能大幅提高。由此可见，词形和词性虽然对框架元素标注有较好的效果，但加入丰富的语义特征能更好地提高模型性能，后续实验可在这方面进行深入研究；

（4） 加入同义词词林信息后，从实验结果对比可以看出，加入 1 级词林信息相比其他实验，效果更突出。因为 1 级同义词词林信息是词语的大类表示，对当前语义场景有更强的概括性。因此更适合做框架元素标注的特征。

4.6 自动抽取实验

为测试本文核心框架语义分析方法的有效性，本文从《人民日报》中构建 300 条测试例句，并人工标注正确答案。分别在人工标注和自动标注下，测试依存结构抽取的性能，实验结果如表 10 所示：

表 10 自动抽取实验结果

	核心目标词识别	框架分配	框架元素标注
人工	-	85.8%	54.74%
自动	89.2%	75.5%	50.26%

经过分析，自动抽取结果偏低有以下原因：

（1） 在开放测试集上的核心目标词识别准确率下降，其主要原因是训练语料中并未涉及所有词元，存在大量未登录目标词，导致模型无法识别；

（2） 框架选择模型训练时，由于实验只对常见的 47 个歧义词元构建了识别模型，使得有些歧义词无法被识别。例如，“地铁口那位靠乞讨为生的老太太已经不在了。”和“他靠在沙发上看着我的作业不说话。”。这两个句子中，词元“靠”分别有“依靠”和“身体姿势”两种框架语义。由于已构建模型的 47 个歧义词元并不包括“靠”，如果待识别例句中含有目标词“靠”，则将导致框架分配不准确，进而影响框架元素标注结果；

（3） 框架元素自动标注结果低，主要原因是实验训练语料较少，模型涵盖范围较小，召回率不高，有些元素无法自动标注；其次，前两个阶段的错误积累，降低了模型的整体性能。

5 结论与展望

汉语核心框架语义分析是基于 CFN，从框架语义角度，获取句子的核心语义。本文将汉语核心框架语义分析任务分成核心目标词识别、框架选择和框架元素标注三个子任务，针对各个子任务的不同特点，结合同义词词林编码信息，使用基于贪心策略的特征选择算法，分别建立了标注模型，在 CFN 现有数据集中验证了本文方法的有效性。

核心框架语义分析子任务中的未登录目标词识别与框架元素标注仍然是制约最终分析结果的关键环节。而数据稀疏是影响未登录目标词识别与框架元素标注的主要因素，下一步将在不断扩大 CFN 基础标注资源的同时，结合相关词汇语义资源来提高未登录词元识别的准确率，并通过引入半监督机器学习模型提高框架元素标注的召回率，从而改善 CFN 核心框架语义分析的整体结果，为框架语义分析技术进一步应用于自动问答、文摘等领域奠定基础。

参考文献

[1] 李茹.汉语句子框架语义结构分析技术研究[D].山西大学,2012.

[2] Fillmore C. Frame semantics [J]. Linguistics in the morning calm, 1982:111-137.

[3] 刘开瑛.汉语框架语义网(CFN)构建现状[C]//第四届全国学生计算语言学研讨会会议论文集.2008:1-7.

[4] 刘开瑛,由丽萍.汉语框架语义知识库构建工程[C]//中国中文信息学会二十五周年学术会议论文集.北京.2006:64-71.

[5] Baker, Collin F, Charles J. Fillmore,and John B. Lowe. The Berkeley FrameNet project[C].In Proceedings of COLING/ACL. Montreal, Canada. 1998:86–90.

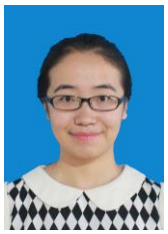
[6] Baker C,Ellsworth M,Erk K.SemEval'07 task 19: frame semantic structure extraction[C]//Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics, 2007:99-104.

[7] Bejan C A, Hathaway C. UTD-SRL: a pipeline architecture for extracting frame semantic structures[C]//Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics, 2007: 460-463.

[8] Johansson R, Nugues P. LTH: semantic structure extraction using nonprojective dependency trees[C]//Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics, 2007:227-230.

[9] Dipanjan Das, Desai Chen, Andr é F. T. Martins,etc.Frame-semantic parsing[J].Computational Linguistics,2013,

40(1):9-56.
[10] Xue N, Palmer M. Automatic semantic role labeling for Chinese verbs[C]//IJCAI.2005, 5: 1160-1165.
[11] 李济洪.汉语框架语义角色的自动标注技术研究[D].山西大学,2010.
[12] 梅家驹,竺一鸣,高蕴琦等编.同义词词林[M].上海:上海辞书出版社,1983.
[13] 郝晓燕,刘伟,李茹,等.汉语框架语义知识库及软件描述体系[J].中文信息报，2007,21(5):96-100.
[14] Ramshaw L A, Marcus M P. Text chunking using transformation-based learning[M]//Natural language processing using very large corpora. Springer Netherlands, 1999: 157-176.
[15] Che W, Li Z, Liu T. Ltp: A chinese language technology platform[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Association for Computational Linguistics, 2010:13-16.
[16] 李国臣,张立凡,李茹,等.基于词元语义特征的汉语框架排歧研究[J].中文信息学报,2013,04:44-51.



石佼（1990—），女，硕士研究生，主要研究领域为中文信息处理。
Email:shijiao0908@126.com；

李茹（1965—），女，博士，教授，博士生导师，主要研究领域为自然语言处理。
Email:liru@sxu.edu.cn；

王智强（1987—），男，博士研究生，主要研究领域为社会媒体数据挖掘、自然语言处理。
Email:zhip.wang@163.com。

通讯作者：李茹　　Email:liru@sxu.edu.cn