基于 TCRF 的核心框架元素标注

王智强 刘海静 李双红 李茹

山西大学 计算机与信息技术学院，山西 太原 030006

E-mail:zhiq.wang@163.com

摘要： 本文基于TCRF（tree structured conditional random field）模型对汉语框架网（CFN）中的核心框架元素进行了自动标注研究。该方法抽取了依存树中父节点层面的特征，使得标注结果在词与词性层面特征的基础上有一定程度的改善。实验选用了CFN中“发明”框架下的句子库，在加入父节点相关特征的最优模板中，核心框架元素自动标注结果的准确率84.3%，召回率62.0%，F值71.5%。

关键词： 汉语框架网络;框架元素;自动标注;TCRF模型

Automatic Labeling of Chinese Core Frame Element based on TCRF model

Wang Zhiqiang Liu Haijing Li Shuanghong Li Ru

E-mail:zhiq.wang@163.com

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

Abstract: This paper investigates the automatic labeling of core frame element in CFN based on TCRF(tree structured conditional random field) model. This method extracts the feature of parent node in dependency Syntax tree and makes a certain degree of improvement to the label results on the feature of word and part of speech of the word. The experiments select the sentences of “Invention” frame in CFN, the results obtain 84.3% precision and 62.0% recall and F score 71.5% on the best features added dependency relation.

Key words: Chinese FrameNet; Frame Element; Automatically Labeling; Tree-structured Conditional Random Fields

1、引言

语义角色标注是自然语言处理中的一项重要任务。它是浅层语义分析的一种实现方式，该方法并不对整个句子进行详细的语义分析，其实质是在句子级别进行浅层的语义分析。具体而言，语义角色标注即对一个句子中谓词所支配的论元进行识别、分类。

英语的语义角色标注研究，最早是Dan Gildea与Dan Jurafsky[1]基于英语FrameNet语料库的工作。伴随着宾夕法尼亚大学英文PropBank语料库的建立，语义角色标注任务越来越受到许多学者的关注。国际上也先后举行了5次评测，分别是Senseval-3[2]、CoNLL(Conference on Computational Linguistics Learning)会议主办的SRL(semantic role labeling) Shared Task 2004[3]、2005[4]、SemEval2007[5]以及CoNLL Shared Task 2008[6]，最好的评测结果F1-值达到84.86%。

汉语方面有Sun[7]、Xue[8]、刘怀军[9]等的研究。其中Xue基于ChinesePropBank语料库，通过使用手工标记的句法树，得到了94.1%的F1-值。但如果采用自动句法分析，却只有71.9%的F1-值。这说明句法分析的准确性，很大程度上限制了SRL的性能。随后有基于依存分析的SRL也开始兴起，该方法获得了和基于短语结构句法分析的SRL相当的性能，并且还具有很大提升空间。目前还未有利用依存句法分析在汉语框架网络（CFN）上进行SRL的研究。本文将尝试基于TCRF模型，通过利用依存句法层面的特征对汉语框架网（CFN, Chinese FrameNet）[9]中核心框

基金项目：国家自然科学基金（60970053）；山西省国际科技合作项目（2010081044）；2007年度山西省高校拔尖人才创新基金；山西省实验室开放基金（2009011059-4）。

作者简介：王智强（1987-），男，硕士，主要研究方向为计算语言学；刘海静（1985-），女，硕士，主要研究方向计算语言学；李双红（1984-），男硕士，主要研究方向计算语言学。

架元素进行自动标注。

汉语框架网（CFN）是山西大学建设的汉语语义知识库，以C. J. Fillmore的框架语义学为理论基础、以加州大学伯克利分校的FrameNet为参照、以真实语料为依据而建立的。它包含有三个子库[10]：框架库、句子库、词元库。框架库是对汉语词汇按照所表示的活动场景的异同分类描述。句子库是给定词元和所属框架，对句中的框架元素所在成分标记框架元素名称、短语类型和句法功能三种信息，词元库则描述了每一个词元的词义，并根据句子标注结果行程标注报告。其中句子库中所标记的框架元素又分为核心框架元素与非核心框架元素，例如：

<cog-np-subj 日本 nsy 的 u 研究 v 人员 n > <time-tp-adva 最近 nt > <tgt 发明 v > <null 了 u > <inv-np-obj 一 m 种 q 利用 v 废弃 v 轮胎 n 加固 v 坡面 n 的 u 新 aq 方法 n > 。 w

标记成分<cog-np-subj 日本 nsy 的 u 研究 v 人员 n >与<inv-np-obj 一 m 种 q 利用 v 废弃 v 轮胎 n 加固 v 坡面 n 的 u 新 aq 方法 n >为核心框架元素，它是框架在概念理解上的必有成分；<time-tp-adva 最近 nt>为非核心框架元素，它并不显示框架的个性，表达了时间、空间、环境条件、原因、目的等外围语义成分。

本文选用"发明"框架下的句子库，基于TCRF模型对其核心框架元素进行自动标注的研究。通过加入依存句法层面特征，核心框架元素自动标注结果的准确率84.3%，召回率62.0%，F值71.5%。

2、系统描述

基于TCRF模型的核心框架元素标注与一般的标注过程类似：1）核心框架元素识别，识别出句法层面上哪些成分是语义角色。2）核心框架元素分类，区分出第一步中识别出来的成分属于哪个框架元素。我们把框架元素标注看成是基于依存句法分析的序列标注问题。本文将利用TCRF来进行核心框架元素的识别和分类。具体标注流程如下图：

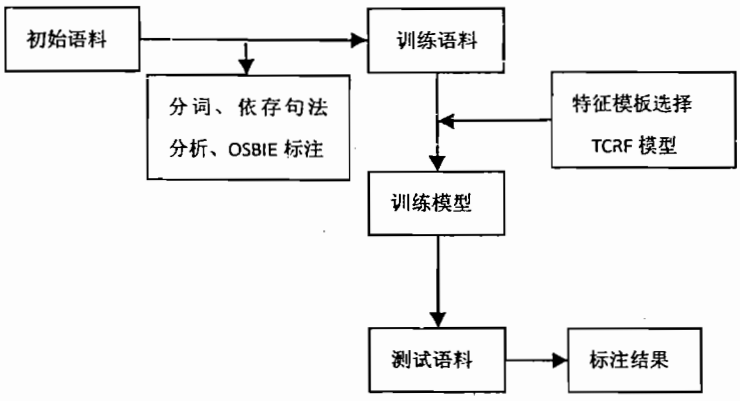

图1 系统流程图

3、基于TCRF模型的核心框架元素识别

3.1 TCRF模型

近年来，条件随机场模型被广泛地应用于自然语言处理序列标注的问题中。条件随机场模型（CRF，Conditional Random Fields）[11]由Lafferty和McCallum等人于2001年提出。它将无

向图中的团函数和最大熵框架有机地融合到一起，得到了一个用来解决序列标注和分割的概率模型。条件随机场模型不仅克服了隐马尔科夫模型[12]的强独立性假设，而且不具有最大熵马尔科夫模型[13]的标注偏执问题。继 CRF 模型之后，Jie Tang 等在 2006 年提出 Tree Structured Conditional Random Fields（TCRF）[14]，它能将随机变量组织成一个树结构，很适用于在句法树上进行框架识别。在该模型中，我们抽取了节点之间的父子关系侧面特征，对于观察值 x，最终的输出标记的概率如下可得：

$$p(y/x)=\frac{1}{Z(x)}\exp\left(\sum_{e\in\{E^{pc},E^{cp},E^{ss}\}}\lambda_j t_j\left(e,y|_v,x\right)+\sum_{v\in V,k}\eta_k s_k\left(v,y|_v,x\right)\right) \tag{1}$$

如图 2 把"有个小伙子编了句俏皮话"经过依存句法分析后转换成如下结构：

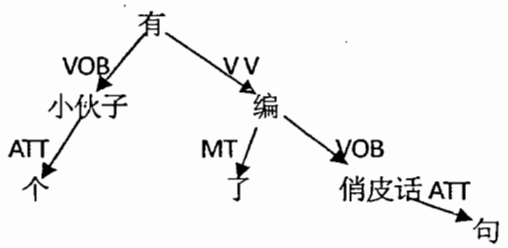

图 2 句法树结构

从依存结构树中可以提取出节点之间的依存信息。本系统抽取了依存树中父节点层面的相关信息，运用 F、G 的特征向量，输出 y 的概率可以表示为：

$$p(y/x)=\frac{1}{Z(x)}\exp\sum_{v\in V}\{F+G\} \tag{2}$$

$$F=\sum_j\lambda_j f_j\left(v,y(v),x\right) \tag{3}$$

$$G=\sum_k\mu_k g_k\left(v,y(v),x,v',y(v')\right) \tag{4}$$

其中，F、G 分别表示当前节点、当前节点的父节点的特征向量。v 表示句中词语对应在树中的节点，v' 表示 v 的父节点。

3.2 特征选择以及模板设置

实验中，我们设置特征模板遵循的主要原则为：在词与词性层面的特征之上逐步加入句法依存特征，即加入我们之前提取的父节点相关特征。通过比较结果的优劣，来选择较好的模板。试验中以词与词性的特征为基础建立特征模板，主要考虑到以下两点：

1.词是组成句子最基本的语言单元，实验前期以词为基本单位对核心框架元素进行了 O-S-B-I-E 划界处理，最重要的是本实验中即将利用的依存句法树中的各个节点也以词为基本单元。

2.以词性作为基本特征是因为对汉语分词的准确率已经达到了一个较高的水平，可以作为框架元素标注的一个信赖的特征。

丁金涛、王红玲等[15]在特征优化组合的研究中表明，选取不多的有用的特征并进行有效优化组合就能取得很好的结果，同时句法分析中重要的谓词和中心词及其相关的词汇特征同样在语义角色标注系统中发挥了重要作用。于是实验之前，根据上述原则初步定义了 5 类特征组

合，共 15 种特征模板如表 1：

表 1 五类特征组合模板

模板	特征组合	窗口大小
T1	词，词性，词与词性的组合	左右分别开一个窗口
T2	词，词性，词与词性的组合	左右分别开两个窗口
T3	词，词性，词与词性的组合	左右分别开三个窗口
T4	词，词性，当前词的父节点	左右分别开一个窗口
T5	词，词性，当前词的父节点	左右分别开两个窗口
T6	词，词性，当前词的父节点	左右分别开三个窗口
T7	词，词性，当前词与父节点的关系	左右分别开一个窗口
T8	词，词性，当前词与父节点的关系	左右分别开两个窗口
T9	词，词性，当前词与父节点的关系	左右分别开三个窗口
T10	词，词性，当前词的父节点及与父节点的关系	左右分别开一个窗口
T11	词，词性，当前词的父节点及与父节点的关系	左右分别开两个窗口
T12	词，词性，当前词的父节点及与父节点的关系	左右分别开三个窗口
T13	词，词性，词与词性的组合，当前词的父节点及与父节点的关系	左右分别开一个窗口
T14	词，词性，词与词性的组合，当前词的父节点及与父节点的关系	左右分别开两个窗口
T15	词，词性，词与词性的组合，当前词的父节点及与父节点的关系	左右分别开三个窗口

我们把 T1-T3 作为基础模板，即只利用词与词性层面的特征。模板 T4-T12 在词与词性特征基础上引入了句法层面的父节点相关特征。模板 T13-T15 集中了实验所设置的所有特征。这样设置模板使得实验具有可比性，利于找出最优模板。

4、实验前期准备与结果分析

4.1 前期语料准备与预处理

现有的 CFN 句子库并没有建立相应的依存句法树库，我们利用哈工大共享的支持中文树库的工具包，将 CFN 语料中“发明”框架下的 198 个句子转换成依存句法树，并对其中明显的错误进行手工校正。经依存句法分析之后，句中的各个词语会通过不同的依存关系体现出句子的句法信息。本系统目前只针对核心框架元素标注进行了研究。我们知道核心框架元素是理解一个框架概念的必有成分，类似于句法功能上的句子主干部分，因此如果能够利用句法层面的特征，理论上将对核心框架元素的标注起到一定的有利影响。另外，框架元素标注跟语义角色的识别类似，我们需要对框架元素进行划界。本文在对句子进行分词、词性标注、句法分析的基础上，使用 O-S-B-I-E 策略对核心框架元素进行块标注，并在块边界识别的同时进行了分类，记标注集合为{S-X,B-X,I-X,E-X,O}(其中，X 为核心框架元素名称)，示例如下：

如 O 前 O 几 O 年 O，O 一 B-cog 名 I-cog 大学生 E-cog 发明 tgt 了 O 一 B-inv 种 I-inv 名为 I-inv "I-inv 虫 I-inv" I-inv 的 I-inv 电脑 I-inv 病毒 E-inv。O

O 表示当前词不是核心框架元素，S 表示当前词单独构成一个核心框架元素，B 表示当前词是一个核心框架元素的开始词，I 表示当前词是一个核心框架元素的中间成分，E 表示当前词是一个核心框架元素的结尾词。

试验所用测试集与训练集语料均来自于山西大学手工构建的 CFN 语料库。选取了"发明"框架下的 198 条例句，首先对其进行还原，然后通过分词、依存句法分析、O-S-B-I-E 核心框架元素划界，并对明显的句法错误进行校正。并把语料拆分为 5 等份，其中 4 份作为训练集，1 份作为测试集。拆分时遵循均匀分配的原则，把"发明"框架中不同词元的例句均匀分配到每一份中。这样可以避免语料规模不大的情况下带来过多的数据稀疏。

4.2 试验结果分析

试验中对 15 个特征模板逐一进行测试，通过最终核心框架元素标注结果的优劣比较，在 5 类不同的特征组合中分别给出了最优模板，分别为 T3、T4、T7、T12、T13 模板，测试结果如表 2：

表 2 最优模板下的标注结果

特征组合	模板编号	准确率	召回率	F 值
词，词性，词与词性的组合	T1	82.5%	55.8%	66.6%
	T2	84.9%	58.4%	69.4%
	T3	85.1%	59.3%	69.9%
词，词性，当前词的父节点	T4	81.7%	53.1%	64.4%
	T5	83.1%	52.2%	64.1%
	T6	82.5%	50.4%	62.6%
词，词性，当前词与父节点的关系	T7	84.3%	62.0%	71.5%
	T8	84.1%	55.8%	67.1%
	T9	85.3%	54.0%	66.1%
词，词性，当前词的父节点及与父节点的关系	T10	84.1%	61.1%	70.8%
	T11	82.4%	60.2%	69.6%
	T12	82.1%	59.3%	68.9%
词，词性，词与词性的组合，当前词的父节点及与父节点的关系	T13	83.6%	54.0%	65.6%
	T14	84.5%	51.3%	63.8%
	T15	82.8%	51.3%	63.4%

从结果中我们可以发现，第二类模板在词，词性特征的基础上加入依存树中父节点的特征时，结果不如完全利用词与词性层面的组合特征，而当加入当前词与父节点关系的特征时，测试结果明显有所提高，这说明当前词与父节点关系的特征对核心框架元素标注具有不错的贡献价值。第四类模板中同时加入了父节点与父节点关系特征，此类模板总体的测试结果相对较好，并且最大优势是随窗口的增加，结果比较稳定。第五类模板组合了实验设置的所有特征，测试结果相对较差。以上模板大多是随着窗口的增加，结果有所下降，主要因为句法分析的不准确会随着特征数目增多而表现的越发突出，另一方面是数据的稀疏性。

由于实验中语料规模较小，为了更客观地评价加入依存句法特征对系统的贡献，我们对"发明"框架下按照词的标注结果做了详细统计。这里给出第一类"词，词性，词与词性的组合"的最

优模板 T3 与第四类"词，词性，当前词的父节点及与父节点的关系"最优模板 T12 的统计结果，如表 3：

表 3 基于词块的统计结果

最优模板编号	框架元素包含总词数	标对词数	正确率	标错词数	错误率
T3	307	212	69.1%	95	30.9%
T12	307	228	74.3%	79	25.7%

从两类最优模板统计结果来看，没有加入依存句法层面特征的 T3 模板相比加入依存特征的词的标注错误率要高 5 个百分点。在本实验的语料规模下，5 个百分点大约是 16 个词，16 个词则相当于两个例句的核心框架元素。说明依存句法层面特征的加入对系统性能的提高是非常可观的。

虽然依存句法层面特征的加入对于 CFN 中核心框架元素标注有一定的改善，但是本系统的性能还是偏低。分析原因主要有以下几点:

（1）训练语料不足，导致数据稀疏，以至于识别错误。

（2）缺乏较好的句法分析器的支持。

本系统建立在自动句法分析基础上，由于目前对汉语的据法分析效果很不理想，这就不可避免的在系统中输入了一定错误的句法信息，导致整个系统的性能下降。句法分析器的准确率也在一定程度上限制了本系统的性能，它将是今后影响我们进行大规模语料处理的一项重要因素。

5、结束语

本文使用 TCRF 模型，将依存句法分析中父节点层面特征有效地融入 CFN 核心框架元素自动标注。实验表明，虽然句法分析的准确率不够高，但依存句法层面的信息仍可以一定程度的改善 CFN 中完全基于词层面特征的核心框架元素标注。另外本系统中由于语料规模的限制、句法分析的不准确、句法中子节点的特征未加入等因素，导致系统性能偏低。因此下一步我们将主要从以下几个方面着手提高系统框架元素标注的性能：一是抽取依存树中的子节点和子节点的关系等特征，来提高系统的性能。同时不断扩充 CFN 中每个框架下例句的数目，克服数据稀疏的影响。另外如何提高句法分析的准确率也是下一步重点考虑的问题。

参考文献

[1] Gildea D, Jurafsky D. Automatic Labeling of Semantic Roles, Computational Linguistics, 2002, 28(3): 245-288.

[2] Litkowski KC. Senseval-3 task Automatic Labeling of Semantic Roles. Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, 2004. 9-12.

[3] Carreras X, Marques L. Introduction to the CoNLL-2004 Shared Task: Semantic role labeling. Proceedings of the CoNLL 2004, 2004. 89-97.

[4] Carreras X, Marques L. Introduction to the CoNLL-2005 Shared Task: Semantic role labeling. Proceedings of the CoNLL 2005, 2005. 152-164.

[5] Baker CF, Ellsworth M, Erk K. SemEval'07 Task 19: Frame Semantic Structure Extraction. Proceedings of the 4th International Workshop on Semantic Evaluations, 2007, 99-104.

[6] Surdeanu M, Johansson R, Meyers A, Marquez L, Nivre J. The CoNLL 2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. Proceedings of CoNLL-2008, 2008.

[7] Sun HL, Jurafsky D. Shallow Semantic Parsing of Chinese. Proceedings of NAACL-HLT 2004, 2004.

[8] Xue NW, Palmer M. Automatic semantic role labeling for Chinese verbs. Proceedings of the Nineteenth International

Joint Conference on Artificial Intelligence, 2005.

[9] Liu K, You L. The project of building Chinese FrameNet knowledge base. The Progress and Forefront of Chinese Information Processing. Beijing: Tsinghua University Press, 2006.

[10] 刘开瑛,由丽萍. 汉语框架语义知识库构建工程. 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, 2006, 11:64-71.

[11] Lafferty J, Mccallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data In Proceedings of the 18th International Conf. on Machine Learning, 2001, 282-289.

[12] Rabiner LR. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 1989, 77(2):257-286.

[13] Mccallum A, Freitag D, Pereira F. Maximum Entropy Markov Models for Information Extraction and Segmentation. Proceedings of ICML, 2000, 591-598.

[14] Jie Tang, Mingcai Hong, Juanzi Li, and Bangyong Liang. 2006. Tree-structured Conditional Random Fields for Semantic Annotation. In Proceedings of 5th International Conference of Semantic Web (ISWC'2006), Athens, GA, USA.

[15] 丁金涛,王红玲,周国栋,朱巧明,钱培德. 语义角色标注中特征优化组合研究. 计算机应用与软件, 2009, 26(5):17-21.