汉语句法成分中心词自动识别方法的研究

任晓娜，王莹莹，周俏丽，蔡东风

知识工程中心 沈阳航空航天大学 辽宁 沈阳 110136

E-mail: rxn_nlp@163.com

摘 要：本文提出一种基于层叠条件随机场的统计和规则相结合的句法成分中心词自动识别的方法。首先将输入的一个标有句法成分的句子分为底层组块和高层短语，分别对这两部分训练两个不同的模型，并逐层进行中心词识别；然后通过规则库和实例库分别进行后处理；最后将这两部分的识别结果进行合并，即得到句子中所有句法成分中心词的最终识别结果。在CIPS-ParsEval-2009评测语料的基础上，本文采用的方法相比于去年本单位采用的方法减少了0.2%错误率。

关键词：自然语言处理，句法成分中心词，句法分析，条件随机场

Automatically Recognize Head Constituents of Chinese Sentences

Ren Xiaona, Wang Yingying, Zhou Qiaoli, Cai Dongfeng

Knowledge Engineering Research Center, Shenyang Aerospace University, Liaoning, Shenyang, 110136

E-mail: rxn_nlp@163.com

Abstract: An effective approach of recognizing head constituents of Chinese sentences is proposed based on the combination of Cascaded Conditional Random Fields (CCRFs) and some other rules. First, a sentence that has been labeled syntactic constituents is divided into two parts: base chunks and high-order phrases, on the basis of which two models are trained. The two models are utilized to recognize the head constituents based on CCRFs. Then we correct the preliminary results on the basis of a rule base and an example base. All head constituents are obtained by merging the results of two parts. Experimental results demonstrate that improvement is obtained over our results on CIPS-ParsEval-2009 (baseline), and the error ratio is 0.2% lower than baseline on CIPS-ParsEval-2009 corpus.

Key words: natural language processing, head constituents, syntactic parsing, Conditional Random Fields.

1. 引言

句法分析一直是自然语言处理领域的一个核心问题之一。随着句法分析新算法的不断提出，以及相关新技术的出现，使得句法分析的效果有了很大的提高。同时，句法成分的中心词识别也成为句法分析的一个重要组成部分。在 CIPS-ParsEval-2009[1]评测中，task5 其中的一个评测标准是“边界+成分标记+中心词识别”，可以看出句法分析器的性能不仅仅依靠句法分析的结果，还包括句法成分中心词识别的效果。很多中心词驱动的统计句法分析模型就是依据句法成分的中心词来驱动句法分析，从而来提高句法分析的性能，但是由于很多汉语树库并没有标注句法成分的中心词，所以中心词驱动的统计句法分析模型一个首要的任务就是要寻找任何父节点对应的中心子节点[1]。另外，由于很多汉语树库多采用短语结构的标注形式，汉语依存树库的建设还存在很

[1] 会议网址：http://www.ncmmsc.org/CIPS-ParsEval-2009/index.asp

多不足，所以，国内外不少研究者都尝试了将短语结构树转化为依存结构树库[2]，转化的第一步就是要标注句法成分的中心词。所以，句法成分中心词识别在句法分析上、树库之间的转化上都有重要的意义。

本文的其他部分组织如下：第二部分介绍句法成分中心词识别的相关研究工作，第三部分详细介绍了 CRFs 模型及特征选择过程，第四部分为实验结果及相关分析，第五部分给出本文工作的总结和下一步展望。

2. 相关研究

通常句法成分中心词识别的做法是人为构造一个中心词映射规则表，然后由算法将规则表强加在树库上。Magerman 提出了核心节点映射表，通过优先序列来确定一个组块中的核心节点[3]。Collins 修改 Magerman 的规则，提出了新的核心节点映射表[4]。Yamada 和 Matsumoto 重新定义了一个核心节点映射表，并且给出了一套转化程序 Penn2Malt[5]，此程序提供了 Penn Chinese Treebank 和 Penn Treebank 的核心节点映射表，现已成为最流行的转化程序，被大量学者在研究过程中采用。Johansson 和 Nugues 对 Yamada 和 Matsumoto 的核心节点映射表进行了修改，对并列短语、介词短语、从句连词、限定代词、名词短语进行了更加详细的区分和处理。

在 CIPS-CoNLL-2009 评测中，task5 第一名陈晓等人采用 CRF 统计模型把句法成分中心词识别转换为分类问题[6]，根据句法成分的类型和中心词的个数来进行分类，总共分为八类分别进行识别，在句法成分识别 F1 值 88.77%的情况下，句法成分部分中心词识别 F1 值为 86.42%，完全中心词识别 F1 值为 82.53%；而 task5 排名第二的厦门大学练睿婷等人采用基于规则的方法来进行句法成分中心词的确定[7]，其中包括两种句法规则，一种是出现在训练集中的规则，另外一种采用文献[8]提供的决策表来确定每一条句法规则的中心词，在句法成分识别 F1 值 87.37%的情况下，句法成分部分中心词识别 F1 值为 76.02%，完全中心词识别为 70.58%。

本文采用基于层叠条件随机场的统计和规则相结合的方法来进行句法成分中心词的自动识别。根据特征选择和参数估计分别建立基于底层组块和高层短语[2]中心词识别的 CRF 统计模型，并对测试语料分别进行底层组块和高层短语的中心词识别，然后采用基于规则的方法[3]对识别结果进行后处理，最后进行合并得到最终的识别结果，完全识别正确 F1 值达到了 98.96%。

3. 任务描述

句法成分中心词识别就是把一个句子中每个句法成分的中心词识别出来并进行标注。本文所采用的句法成分中心词识别的方法是在句法成分标注完全正确的基础上进行的，与句法成分标注是完全独立的两个过程。举例如下：

输入：[dj　消费者/n　[vp　力求/v　[np　效用/n　最大化/vN　]]]

输出：[dj-1　消费者/n　[vp-0　力求/v　[np-1　效用/n　最大化/vN　]]]

上例中句法成分类型后面的数字代表此句法成分中心词的位置（第一个词的位置为 0），例如，句法成分“[np-1　效用/n　最大化/vN　]”为底层组块，其中心词是位置为 1 的词“最大

[2] 底层组块和高层短语统称为句法成分，其中底层组块用其类型替换后即生成高层短语。

[3] 基于规则的方法即规则库和实例库的后处理阶段。

化/vN”，句法成分“[vp-0 力求/v [np-1 效用/n 最大化/vN]]”中的底层组块用其类型“np”替换后为高层短语“[vp-0 力求/v np/np]”，其中心词是位置为 0 的词“力求/v”，依次类推，得到每一层句法成分的中心词。

上例中的输出分两步进行，同时需要训练两个模型。第一步是底层组块“[np 效用/n 最大化/vN]”的中心词识别，采用底层组块的中心词标注模型（Base-Model）[4]；第二步是高层短语的生成及识别，即底层组块用组块类型来替换后进行中心词识别，如上例中高层短语“[vp 力求/v np/np]”和“[dj 消费者/n vp/vp]”，采用高层短语的中心词标注模型（High-Model）[5]。然后对于打分值低于阈值 0.8 的进行后处理，最后进行合并得到最终的识别结果。其中阈值 0.8 的大小和训练语料的大小有关，需要从实验中得知。本文所采用方法的过程流程图如图 1 所示：

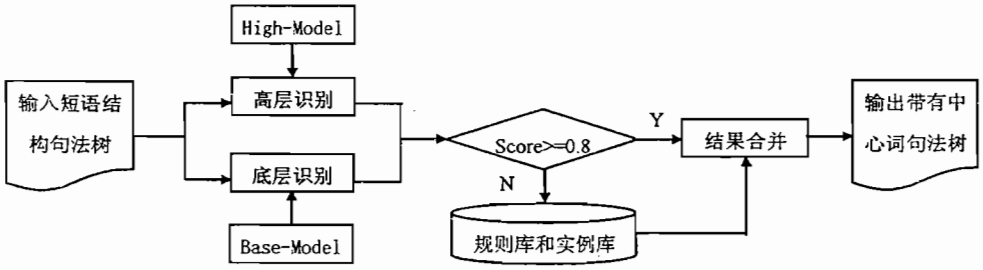

图 1 句法成分中心词识别流程

3.1 CRFs 模型

条件随机场(Conditional Random Fields，以下简称CRFs)是John Lafferty等人于2001年提出的一种基于统计的序列标记和分类模型，也是在给定输入节点条件下计算输出节点的条件概率的无向图模型[9]。

通过对句法成分中心词识别任务的分析可得，此识别任务可以转化为序列标注问题，同样也可以转化为分类问题，考虑到底层短语词汇化强，而高层短语是在底层短语用其类型替换后生成的，非词汇化强一些，所以采用分层进行处理，即底层和高层分别进行识别的方法。

另外，本文之所以选择条件随机场作为句法成分中心词每层标注识别的统计模型，是因为条件随机场能够综合利用字、词、词性及短语类型等多层次的资源，同时对于长程关联有很好的描述能力，并能避免其他模型中存在的标注偏置问题。本文随机抽取 CIPS-ParsEval-2009 评测语料中的 13000 个事件描述单元作为训练语料，1240 个事件描述单元作为测试语料，采用层叠 CRFs 统计模型且相同的特征模板与其他统计模型进行对比，实验结果如下图所示：

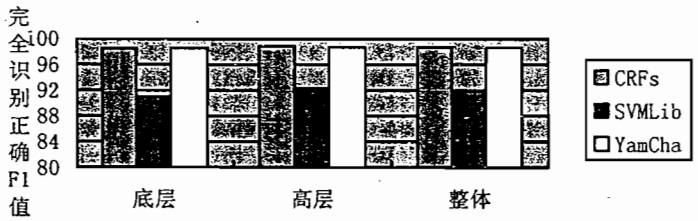

图 2 CRFs 与其他统计模型对比图

[4] Base-Model：只保留最底层短语所生成的模型。

[5] High-Model：除最底层短语之外，上层所有短语所生成的模型。

此对比实验中，CRFs 把此项任务当成序列标注问题来处理，而 SVMLib 当成二分类问题，采用相同语料和模板，完全识别正确 F1 值却相差 6.8%，同样采用机器学习算法 SVMs 的另外一个多分类工具 YamCha，在相同语料和模板的基础上，完全识别正确 F1 值比 CRFs 低于 0.11%。所以，将此项任务转换为序列标注问题，并采用 CRFs 统计模型的识别效果最佳。

3.2 特征选择

在基于 CRF 的句法成分中心词识别问题中，特征的选择通常起着关键性作用。CRF 模型的特点是采用简单的特征表示复杂的语言现象，不做任何独立性假设。根据影响句法成分中心词的各种因素，我们利用词、词性信息、短语类型信息、短语间关系的信息以及它们之间的不同组合信息作为句法成分中心词识别的特征，以下分别介绍 Base-Model 和 High-Model 所采用的特征模板。总的特征空间为：词信息、词性信息、短语类型信息和短语总词数信息。

根据这些特征信息来定义模型中的特征模板，其由原子模板和复合特征模板组成。原子模板可以看作当前上下文的一个特征函数。当特征函数取特定值时，该模板被实例化，得到具体的特征。表 1 为底层识别的原子特征模板 15 种，表 2 为高层识别的原子特征模板 9 种。

表 1. Base-Model 原子模板

原子模板	模板定义
CurWord	当前词
CurPOSTag	当前词的词性
CurPhraseTag	当前词所属短语类型
Word± 1	当前词前或后一个词
Word± 2	当前词前或后两个词
POSTag± 1	当前词前或后一个词的词性
POSTag± 2	当前词前或后两个词的词性
PhraseTag± 2	当前词前或后一个词所属短语的类型
PhraseTag± 1	当前词前或后两个词所属短语的类型

表 2 High-Model 原子模板

原子模板	模板定义
CurPhraseTag	当前短语所属类型
LCh_Word	当前短语最左边孩子
RCh_Word	当前短语中最右边孩子
LCh_Pos	当前短语最左边孩子词性
MCh_Pos	当前短语中间孩子词性
RCh_Pos	当前短语最右边孩子词性
NumCh	当前短语中包含的孩子数目
CurPhraseTag± 1	当前短语前后短语的所属类型

仅使用原子特征不足以标识上下文中的一些现象。通过对上表中原子特征模板进行组合，构成一些复合特征模板来表示更为复杂的上下文环境。当特征函数取特定值时，复合特征模板中的各个原子模板被实例化，从而产生具体的特征。

Base-Model 的复合模板共 7 种分别如下：

CurPosTag/CurPhraseTag，Word-1/PosTag-1，Word+1/PosTag+1，PhraseTag+1/PhraseTag+2，

CurPhraseTag/PhraseTag+1/PhraseTag+2，PosTag-1/CurPosTag/PosTag+1，Word+1/PosTag+1/PhraseTag+1

High-Model 的复合模板共 16 种分别如下：

CurPhraseTag/ NumCh，CurPhraseTag/ LCh_Word，CurPhraseTag/ LCh_Pos，CurPhraseTag/LCh_Pos/RCh_Pos，

CurPhraseTag/NumCh/LCh_Pos/ RCh_Pos，

CurPhraseTag/NumCh/LCh_Word/LCh_Pos/MCh_Pos/RCh_Word/RCh_Pos，LCh_Word/LCh_Pos，

CurPhraseTag/MCh_Pos，NumCh/LCh_Pos/ MCh_Pos/ RCh_Pos，CurPhraseTag/ NumCh/ MCh_Pos，

CurPhraseTag/ LCh_Word/LCh_Pos/MCh_Pos/RCh_Word/RCh_Pos，LCh_Word/ LCh_Pos，LCh_Pos/ MCh_Pos，

CurPhraseTag/NumCh，RCh_Word/RCh_Pos，NumCh/LCh_Word/LCh_Pos/MCh_Pos/RCh_Word/RCh_Pos

原子特征模板和各种复合特征模板共同构成了模型的所有特征模板。所以，Base-Model 共有 22 种模板类型；High-Model 共有 25 种模板类型。

3.3 后处理阶段

通过分析错误实例，发现部分 CRF 的识别结果与实际情况明显不符，后处理阶段可以针对一些识别错误的情况进行补充识别和纠正，从而来提高句法成分中心词识别的正确率。其中后处理阶段采用规则库（Rule Base）和实例库(Example Base)分别进行处理。

通过总结分析错误实例，针对一些特殊的错误实例获得一些启发式规则，这些规则组合起来即形成一个规则库。规则库的形成流程如下图：

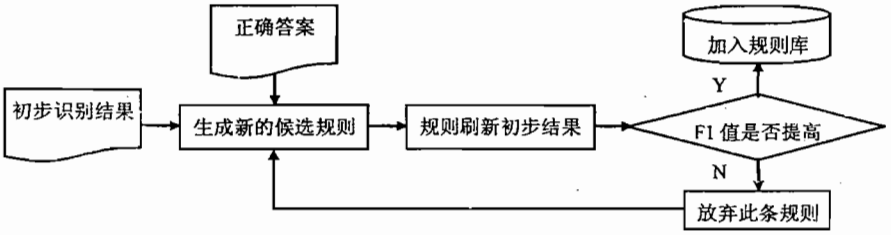

图 3 规则库形成的流程图

通过两种错误类型实例来说明候选规则的生成：

（1）[np-0-1-2-3 诗/n 词/n 歌/n 赋/n] [np-0-1-2-3 晋/nS 冀/nS 鲁/nS 豫/nS]

（2）[vp-2 走/v 一/d 走/v] [vp-2 抓/v 一/m 抓/v] [vp-2 看/v 了/uA 看/v]

类型（1）属于并列结构且词性相同的名词短语，其中每个词属于同一类别并且词数大于 2，则此类型短语的中心词为其中所有的词；类型（2）属于动词重叠式的动词短语，其中词数为 3，第一个词和第三个词是重叠的两个动词，第二个词是表示短时、惯常（如“一”、“了”）等的体标记词，则此类型短语的中心词为最后一个词。

实例库是通过对训练语料进行分析及处理后构建的一个基于短语的实例库。此实例库加入了训练语料中所有最底层的组块和高层的短语，高层的短语即通过用中心词替换底层组块后的高层短语，总共由 168655 个短语组成。实例库生成过程如下：

例：[dj-1 [np-1 温病/n 学说/n] [vp-0 达到/v [tp-1 成熟/a 阶段/nT]]]

首先把最底层的组块[np-1 温病/n 学说/n]和[tp-1 成熟/a 阶段/nT]加入到实例库中，然后最底层的组块用其中心词来替换，生成新的短语[vp-0 达到/v 阶段/nT]加入到实例库，依此类推，可得到一个基于短语的实例库。此句子生成 4 个实例，即[np-1 温病/n 学说/n]、[tp-1 成熟/a 阶段/nT]、[vp-0 达到/v 阶段/nT]、[dj-1 学说/n 达到/v]。每个句子都采用同样的方法，即构成了一个基于短语的实例库。

规则后处理流程如下所示：

背景知识：规则库 Rule Base 和实例库 Example Base

输入：CRF 识别句法成分中心词后打分值低于 0.8 的短语

基本操作：

① 判断是否满足规则库中的规则，若满足，则进行修正；否则执行②；

② 判断实例库中是否存在此短语，若存在，则按照实例库进行修正；否则执行③；

③ 判断是否为最高层的短语，若满足，则执行④；否则用识别出的中心词来替换此短语，返回①执行高一层的短语；

④ 合并 CRF 识别结果和规则后处理结果；

4. 实验结果及分析

4.1 数据及评价方法

本文采用清华大学信息技术研究院语音和语言技术中心开发的清华汉语句法树库（Tsinghua Chinese Treebank ver1.0, TCT Ver10），即 CIPS-ParsEval-2009 评测提供的实验数据。该树库中每个句子都人工进行了分词、词性标注及句法成分和句法成分中心词的标注。我们采用此次评测的训练集共 33693 个事件描述单元，发放的测试集共 8241 个事件描述单元，并且在句法成分完全正确的情况下，进行句法成分中心词自动识别的实验。

评测指标为句法成分中心词部分识别 F1 值（partial-head match F1，简称 PM）和完全识别 F1 值(complete-head match F1，简称 CM)，其中，句法成分中部分中心词识别正确的个数为 PNum，所有中心词都识别正确的个数为 CNum，句法成分的总个数为 SNum，则计算公式如下：

$$PM = \frac{PNum}{SNum} * 100\% \qquad (1)$$

$$CM = \frac{CNum}{SNum} * 100\% \qquad (2)$$

4.2 实验结果

本文对训练语料中句法成分中心词的分布进行了详细统计，统计结果见表 3。

表 3 句法成分中心词分布情况

句法成分中心词的数量	所占比例
没有中心词	1.3177%
一个中心词	92.1733%
多个中心词（两个或两个以上）	6.5090%

从表中可见，汉语句法成分中大部分都只有一个中心词，而含有一个中心词的句法成分中，中心词位于左边第一个词所占的比例为 22.4839%，位于其它位置的所占比例为 77.5161%。

本文所采用的句法成分中心词识别的方法是在句法成分标注完全正确的情况下进行的，并且与本单位 CIPS-ParsEval-2009 评测结果（Baseline）以及 2009 年评测第一名实验结果（Chen's result）进行对比。句法成分中心词识别结果见表 4，对比实验结果见表 5。

表 4 句法成分中心词 PM 和 CM 实验结果

方法	PM	CM
CCRFs	99.2523%	98.8272%
CCRFs + 规则库	99.2684%	98.8593%
CCRFs + 规则库 + 实例库	**99.3101%**	**98.9555%**

表 5 对比实验结果

方法	CM
Baseline	98.7598%
Chen's result	98.93%
本文采用的方法	**98.9555%**

本文与 Baseline 的不同之处就是底层和高层分别进行识别时，高层短语中心词识别采用另外一种标注体系，并且在特征模板选择上加入了更多的信息，不仅考虑了短语内部的词和词性的信息，同时也把短语间的关系考虑在内，从而使得中心词识别提高了 0.2%。

Collins 采用一个核心节点映射表进行中心词识别，是针对宾州树库开发的一种完全基于规则的方法，常用在短语结构树向依存结构树转换方面。为了与此方法进行对比，本文采用基于非词汇化的 Berkeley Parser 对清华树库进行短语结构句法分析，然后将分析结果中句法成分边界识别错误的进行纠正，这样就形成一个宾州树库格式的完全正确的短语结构句法树，在此基础上采用 Collins 的核心节点映射表进行中心词识别，并与本文中所采用的方法进行对比，随机抽取其

中只含有一个中心词的2000句进行测试，Collins核心节点映射表识别句法成分中心词的CM值为92.76%，而本文采用基于层叠条件随机场的方法识别句法成分中心词的CM值达到了97.37%。

4.3 错误实例分析

通过分析错误实例发现，相同短语在不同语境下的中心词标注不一致。如下句：

训练语料：[dj-1 [np-2 J・M・凯恩斯/nP 的/uJDE [np-1 政策/n 主张/n]] [vp-1 [pp-1 被/p [np-1 各/rB [np-1 主要/b [np-1 西方/nS 国家/n]]]] 采纳/v]]

测试语料：[np-0-2 货币学派/n 及其/cC [np-0-1 政策/n 主张/n]]

这两个句子中都存在[np 政策/n 主张/n]这个名词短语，但是在测试语料中其中心词是两个词，而在训练语料中其中心词却是一个词，并且此短语总共出现5次。诸如这种情况的句法成分中心词都将识别错误，这将对整体性能带来很大的影响。

5. 结束语

句法成分中心词识别目前已经成为句法分析的一部分，通过中心词识别可以来提高基于中心词驱动的句法分析性能，还有利于分析句子成分的整体框架。本文采用统计和规则相结合的方法来进行句法成分中心词的识别，并与CIPS-ParsEval-2009测评结果进行对比，性能有所提高。

虽然目前句法成分中心词识别正确率很高，但是仍存在一些中心词无法正确识别的句法成分，这可能跟相应的语境语义有很大关系，所以下一步将考虑把语义信息引入进来，并且通过改善特征来进一步提高句法成分中心词的识别。

参 考 文 献

[1] 熊德意. 基于括号转录语法和依存语法的统计机器翻译研究. 博士毕业论文，中国科学院计算技术研究所. 2007: 37～38.

[2] 李正华，车万翔，刘挺. 短语结构树库向依存结构树库转化研究. 中文信息学报，2008. 22(6):14～19.

[3] David M. Magerman. Natural language parsing as statistical pattern recognition. Ph. D. thesis, Stanford University. 1994.

[4] Michael J. Collins. Head-driven statistical models for natural language parsing. Ph. D. thesis, University of Pennsylvania, Philadelphia. 1999.

[5] Hiroyasu Yamada, Yuji Matsumoto. Statistical dependency analysis with support vector machines. In Proceedings of 8th International Workshop on Parsing Technologies. 2003: 195～206.

[6] Xiao Chen, Changning Huang, Mu Li and Chunyu Kit. Better Parser Combination. In Proceedings of CIPS-ParsEval-2009. 2009: 81～90.

[7] 练睿婷，陈毅东，史晓东，蔡科，朱翔，刘智文，刘宁锋. 厦门大学第一届中文信息学会句法分析评测系统描述. 第一届汉语句法分析评测学术研讨会. 2009:125～132.

[8] Fei Xia. Automatic Grammar Generation from Two Different Perspectives, PhD thesis, University of Pennsylvania. 1999.

[9] John Lafferty, Andrew McCallum, Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proc of ICML, 2001: 282～289.