文章编号: 1003-0077(2008)06-0014-06

# 短语结构树库向依存结构树库转化研究

李正华,车万翔,刘 挺

(哈尔滨工业大学 计算机科学与技术学院 信息检索研究室,黑龙江 哈尔滨 150001)

**摘 要**: 汉语依存树库的建设相对其他语言如英语,在规模和质量上还有一些差距。树库标注需要付出很大的人力物力,并且保证树库质量也比较困难。该文尝试通过规则和统计相结合的方法,将宾州汉语短语树库 Penn Chinese Treebank 转化为哈工大依存树库 HIT-IR-CDT 的体系结构,从而增大现有依存树库的规模。将转化后的树库加入 HIT-IR-CDT,训练和测试依存句法分析器的性能。实验表明,加入少量经转化后的树库后,依存句法分析器的性能有所提高;但加入大量树库后,性能反而下降。经过细致分析,作为一种利用多种树库提高依存句法分析器性能的方法,短语转依存还存在很多需要深入研究的方面。

**关键词**: 计算机应用;中文信息处理;短语结构树库;依存结构树库;依存句法分析

**中图分类号**: TP391 **文献标识码**: A

## A Study on Constituent-to-Dependency Conversion

LI Zheng-hua, CHE Wan-xiang, LIU Ting

(Information Retrieval Laboratory, School of Computer Science and Technology,
Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

**Abstract:** The progress of Chinese dependency treebank construction has fallen behind other languages, such as English, in terms of scale and quality. Building a large scale treebank needs a lot of human and material resources. Meanwhile, it is very difficult to guarantee the quality of the treebank. In this paper, we explore a new method which combines rule-based method and statistical-based method to convert a constituent treebank named Penn Chinese Treebank to a dependency treebank which follows the annatation standard of HIT Chinese Dependency Treebank (HIT-IR-CDT). We increase the size of training data by adding converted treebank into HIT-IR-CDT and re-train the dependency parser. Experiments show that small addition of converted treebank can improve the performance of dependency parser, while large addition will bring it down. Through detailed analysis, we believe that convertion of constituent-to-dependency treebank still needs in-depth research as a method of improving performance of dependency parser by utilizing different treebanks.

**Key words:** computer application; Chinese information processing; constituent-based treebank; dependency treebank; dependency parsing

## 1 引言

对于基于统计的句法分析而言,树库建设是一项非常重要的工作。树库质量和规模直接关系到句法分析的性能。当前最著名的树库是 Penn Treebank。汉语方面,较早的是台北中央研究院标注的 Sinica 树库(繁体)、美国宾夕法尼亚大学 Penn Chinese Treebank (PennCT)树库,以及清华大学的汉语树库。这些树库大多采用了短语结构的标注形式。较著名的依存树库有捷克语的布拉格树库,英语的 PARC 树库,以及俄语、意大利语等语言的树库。汉语依存树库的建设还存在很多不足,无论从规模还是质量上,和其他语言特别是英语有很大差

**收稿日期**: 2008-06-01 **定稿日期**: 2008-09-10

**基金项目**: 自然科学基金资助项目(60675034, 60575042); 国家 863 计划资助项目(2006AA01Z145)

**作者简介**: 李正华(1983—),男,博士生,主要研究方向为自然语言处理;车万翔(1980—),男,博士,讲师,主要研究方向为自然语言处理;刘挺(1972—),男,教授,博导,主要研究方向为信息检索和自然语言处理。

距。这为汉语依存分析的研究工作带来了困难,制约了中文依存分析研究的发展[1]。

标注树库是一件费时费力的工作,需要完善的标注体系和规范的标注流程,从而保证标注的质量。本文尝试利用基于规则与统计相结合的树库转化方法,将 Penn Chinese Treebank (PennCT)转化为 HIT-IR-CDT 的标注体系。称转化后的树库为 Penn Chinese Dependency Treebank (PennCDT)。然后将 PennCDT 作为 HIT-IR-CDT 树库的补充,训练依存句法分析器模型,从而提高依存分析的性能。

本文内容组织如下。第 2 部分介绍树库转化方面的相关工作。第 3 部分介绍 PennCT 和 HIT-IR-CDT 两种树库的标注体系。第 4 部分介绍基于规则与统计相结合的树库转化过程。第 5 部分介绍利用转化后的树库 PennCDT 提高依存分析性能的情况。第 6 部分为结论及进一步工作。

## 2　相关工作

短语结构和依存结构虽然在表现形式上不同,但是它们都是对句子语法结构的描述,因此在结构上存在一致性。国内外不少研究者都尝试了将短语结构树库转化为依存结构树库。

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[2]。Collins 修改 Magerman 的规则,将这种依存关系作为短语结构句法分析中的中间表示[3]。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[4],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。Nivre 重新实现了 Yamada 和 Matsumoto 的方法,定义了一套启发式规则来确定弧的依存关系类型[5]。其程序 PENN2MALT 提供了 Penn Treebank 和 Penn Chinese Treebank 的核心节点映射表。

Johansson 和 Nugues 为了充分挖掘 Penn Treebank 标注的信息,提出了一套更加完善、细致的转化策略[6]。首先,由于 Penn Treebank 对名词短语采用了扁平的标注方式,他们使用启发式规则分析名词短语的内部结构。然后,他们利用核心节点映射表将短语结构转化为依存结构。他们对 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。

汉语方面,党政法等将清华汉语树库(TCT)转化为依存结构[7]。他们利用核心节点映射表进行结构转化,然后利用规则确定弧的依存关系类型。

上述工作都是直接将短语树库转化为依存结构的树库,其中依存句法关系类型一般都根据短语结构树库中的短语类型获得。本文则尝试将短语树库转化为某一特定体系的依存树库。这样做的好处是可以增大目标树库的规模,因而无需改变句法分析模型学习策略,便可以提高依存分析器的性能。这项工作可以看作是利用多种树库学习句法知识的一种方式。

## 3　树库情况

短语树库我们采用 Penn Chinese Treebank (PennCT)。依存树库采用哈工大信息检索研究室标注的 HIT-IR-CDT。

PennCT 由宾夕法尼亚大学计算机与信息科学学院组织标注。语料规模约 1.9 万句子、50.7 万词语、82.5 万汉字。PennCT 对句子进行了分词、词性标注、短语结构句法树构建[8]。其词性标注集包含 33 种词性。PennCT 的短语结构采用了分层的标注体系,不但包含结构标记,而且还采用了功能标记。如图 1 中,"外商 投资 企业"的短语类型为 NP-SBJ,表示结构类型为 NP,功能类型为 SBJ (主语)。

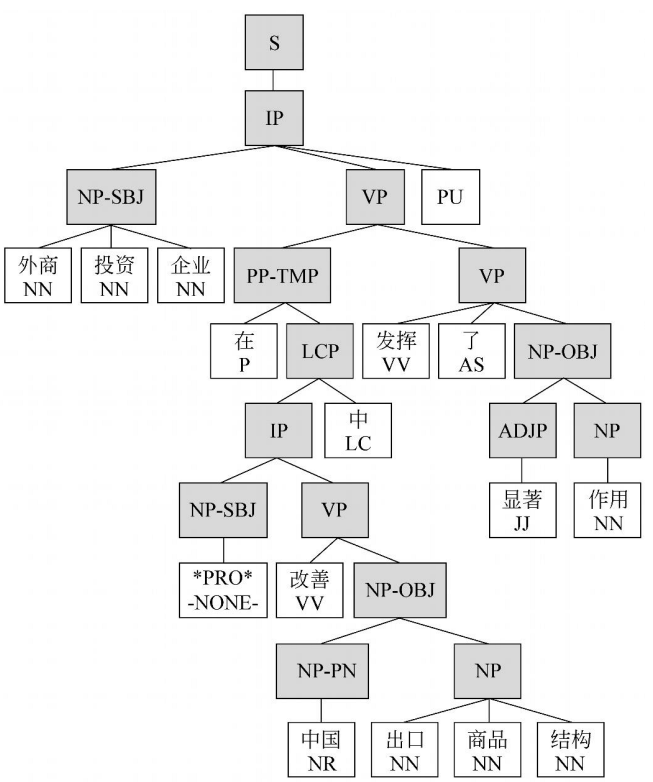


图 1　PennCT 短语结构树示例

HIT-IR-CDT 由哈尔滨工业大学信息检索研究室于 2005 年组织标注[9]。其数据来源于《人民日报》

1998 年上半年语料。包括 1 万个句子,平均句长约为 20 个词语。HIT-IR-CDT 对语料进行了句子切分、词语切分、词性标注、依存结构标注,如图 2 所示。词性标注集采用国家 863 标准,共包含 28 种词性。依存关系类型共有 22 种。图 2 中,< EOS > 为空节点,用来指向句子的核心节点。弧的方向采用标准方式,总是核心(父亲)节点指向依存(儿子)节点。

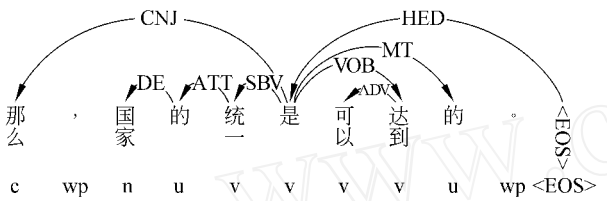


图 2 HIT-IR-CDT 依存结构树示例

## 4 树库转化过程

树库转化系统结构如图 3,主要包括三个方面:

一、词性标注集的转化。目前的依存句法分析器一般都使用词性信息,只使用词性信息会产生严重的数据稀疏问题。因此词性标注集的转化是必要的。

二、短语结构转依存结构。我们借鉴核心节点映射表的思想进行了初步转化,然后使用规则和统计相结合的方法进行后处理。

三、确定每条依存弧的关系类型。我们采用了基于统计的方法来进行依存关系标注。

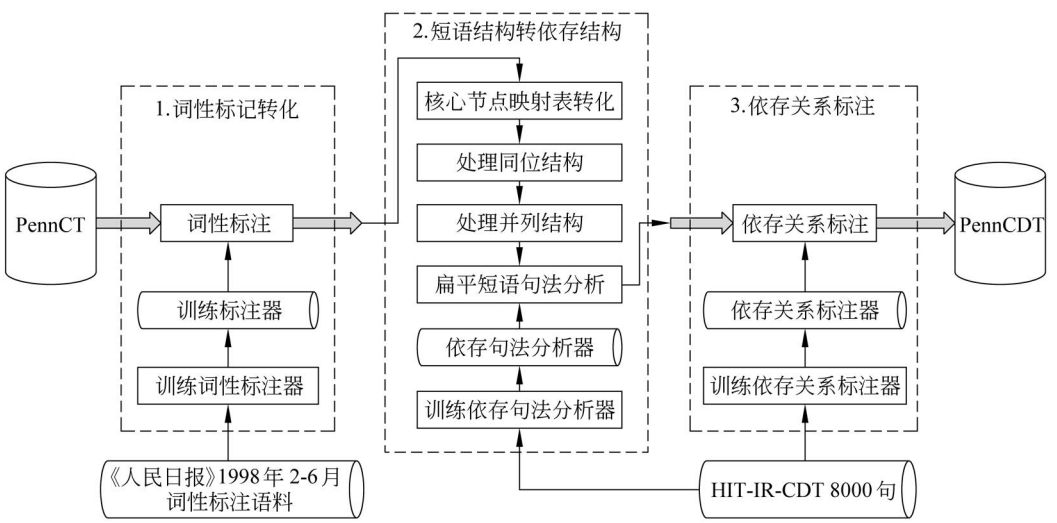


图 3 树库转化系统结构图

表 1 SVMTool 词性标注准确率(%)

| Known | Known Unambiguous | Known Ambiguous | Unknown | Total |
|---|---|---|---|---|
| 98.208 6 | 99.746 1 | 94.046 1 | 78.381 0 | 97.544 4 |

### 4.1 词性标注集转化

SVMTool 是一个开源的基于支持向量机模型的序列标注器,目前在序列标注上取得了较好的效果。我们使用《人民日报》1998 年 2 月至 6 月词性标注语料训练 SVMTool,使用 1 月份语料作为测试集。其准确率性能见表 1。每列含义分别为:Known 表示词典词;Known Unambiguous 表示只有一种词性的词典词;Known Ambiguous 表示有多个词性的词典词;Unknown 表示未登录词;Total 为所有词的准确率。

### 4.2 短语结构转依存结构

#### 4.2.1 利用核心节点映射表进行初步转化

根据 HIT-IR-CDT 的标注体系,我们对 PENN2MALT[①]的核心节点映射表进行了修改,使得转化后的依存弧与 HIT-IR-CDT 标准保持一致。表 2 中列出了我们对映射表修改的部分。

① http://w3.msi.vxu.se/~nivre/research/Penn2Malt.html

表 2　核心节点映射表

| 短语类型 | 核心节点搜索规则 |
|---|---|
| VCD | l VCD VV VA VC VE;l |
| VP | l VC VP VA VE VV BA LB VCD VSB VRD VNV VCP;l |
| IP | l IP;l VC VP VA VE VV BA LB VCD VSB VRD VNV VCP;l |

4.2.2　后处理

核心节点映射表的表达能力是有限的,它只能解决一个短语中只有一个核心节点的情况。如果一个短语中存在多个核心节点,或者某些结构需要特殊处理的时候,就需要通过后处理来进行。后处理主要包括三个方面: 同位结构、并列结构、扁平短语结构。其中同位结构和并列结构使用规则来处理; 而扁平短语结构需要使用句法分析器,通过统计方法来处理。

· 同位结构的处理

同位结构是一种比较特殊的结构,主要是名词短语。如图 4 中"深圳 等"和"特区"属于同位。一般情况下名词短语中最后面的名词为核心词,我们在核心节点映射表中也是这么规定的。HIT-IR-CDT 规定同位关系的两个短语中的第一个为核心节点,如图 5 所示。对于这种特殊情况,我们需要使用专门的规则来处理。我们的规则是,如果发现 XP = > YP-APP ZP,则把 YP-APP 作为 XP 的核心节点。图 5 是我们采用这种方法得出的一个依存树。可以看到,同位关系可以被正确的处理。

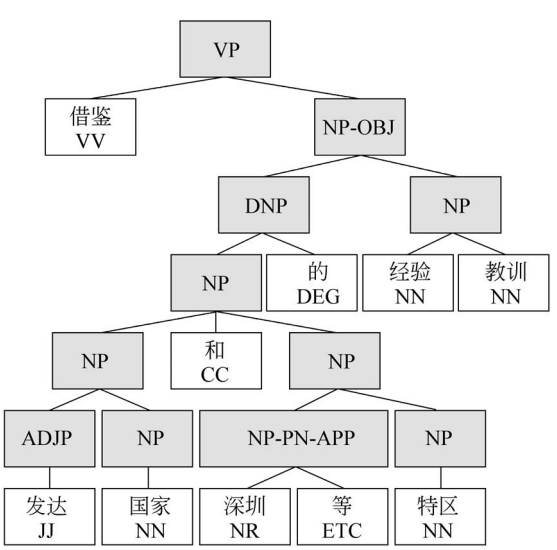


图 4　短语结构(并列、同位示例)

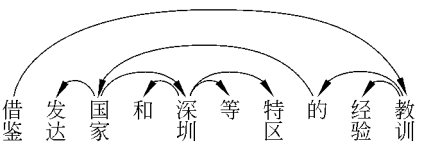


图 5　依存结构(并列、同位示例)

· 并列结构的处理

如图 4 中"发达 国家"和"深圳 等 特区"构成了并列关系。HIT-IR-CDT 对这种情况的处理如图 5 所示。连词"和"依存于后面的短语"深圳 等 特区",前面的短语"发达 国家"为整个 NP 短语的核心节点。这种依存结构也无法在核心节点映射表表达,因此我们使用了规则的方法来处理。首先,我们使用"、"及"CC"将短语切分成片段; 然后,每个片段中确定核心节点,其他节点依存于片段的核心节点; 所有"、"及"CC"依存于其后面片段的核心节点; 最后,所有片段核心节点依存于第一个片段的核心节点。处理后的结果如图 5 所示。

· 扁平短语结构的处理

PennCT 中复合名词短语的标注一般都比较扁平,即不对其内部的结构进行划分。如图 1 中"外商 投资 企业"。为了和 HIT-IR-CDT 标注体系一致,我们需要对名词短语内部的依存结构进行区分。规则的方法很难覆盖所有的情况,为此我们使用了 MSTParser 进行分析。

MSTParser 是 Ryan Mcdonald 等人研究并开发的一个依存句法分析器,将依存分析问题转化为在有向图中寻找最大生成树(Maximum Spanning Tree)问题,并且使用 Online 算法融合多种特征,确定每条弧的概率[10]。

我们使用了 HIT-IR-CDT 前 8 000 句对 MSTParser 进行训练。然后,我们将 PennCT 中的扁平短语抽取出来,利用 MSTParser 处理,确定短语内部依存结构。由于扁平短语的句法结构比较简单,依存分析的效果都比较好。

## 4.3　依存关系标注

我们使用 Online 算法来训练特征向量权值。Online 算法是一种基于距离最大化的学习算法,在文本分类、序列标注、依存分析问题上都表现出了很好的性能。特征选择如表 3 所示。

表 3 依存关系标注使用特征( + dir_dist)

| Unigram Features | Between Features |
|---|---|
| word_h pos_h<br>word_c pos_c | pos_h pos_k pos_c<br>pos_h pos_c bet_pos_h_same_num<br>pos_h pos_c bet_pos_c_same_num |
| Bigram Features | Surrounding Features |
| word_h pos_h word_c pos_c | pos_h-1 pos_h pos_h + 1 pos_c-1 pos_c pos_c + 1 |

表 3 中,h 表示核心节点;c 表示依存节点(或称儿子节点);h - 1 表示核心节点左边的节点;Between Features 中,k 表示节点 h 和节点 c 中间的一个节点;bet_pos_h_same_num 表示节点 h 和节点 c 中间的节点中,与节点 h 词性相同的数目。+ dir_dist 表示每个特征加入距离、弧方向信息之后构成新的特征。对每一类特征,我们都使用了回退的方法,将具体的特征进行泛化。如 Unigram Features 中的特征"word_h pos_h",我们又使用了"word_h"和"pos_h"两个特征。

我们使用 HIT-IR-CDT 前 8 000 句作为训练语料,后 1 000 句作为测试语料,对关系标注器进行训练和测试,标注器的准确率 93.08 %。通过观察我们发现,某些关系如 DC、IC、IS、VV、APP 等的准确率和召回率都比较低。这种错误将会影响转化后树库的质量。

## 5 转化后树库的利用

为了验证 PennCDT 的质量,发现转化过程中存在的问题,以及利用 PennCDT 提高现有句法分析器的性能,我们将 PennCDT 加入到 HIT-IR-CDT 8 000 句中,重新训练并测试 MSTParser,实验结果如图 6。图 6 中 LAS 表示依存关系准确率;UAS 表示依存弧准确率,见公式(1)、(2)。

$$LAS = \frac{\text{弧正确并且依存关系正确的词数(不含标点)}}{\text{所有词数(不含标点)}} \times 100\% \quad (1)$$

$$UAS = \frac{\text{弧正确的词数(不含标点)}}{\text{所有词数(不含标点)}} \times 100\% \quad (2)$$

我们发现,当加入少量 PennCDT 时,依存分析的性能有所提高。继续加入的时候,依存分析的性能不再提高;当加入大量 PennCDT 后,依存分析的性能下降。这说明 PennCDT 的转化过程还过于粗糙,没有完全切合 HIT-IR-CDT 的标注体系。当加入少量 PennCDT 数据时,补充了一些正确的 HIT-IR-CDT 训练语料中没有涵盖的特征,而 PennCDT 引入的错误特征被 HIT-IR-CDT 训练语料的正确信息覆盖并且修正。因此依存分析的性能得到提高。但是当加入大量 PennCDT 数据时,HIT-IR-CDT 的正确信息被 PennCDT 带来的错误特征覆盖,反而导致性能下降。

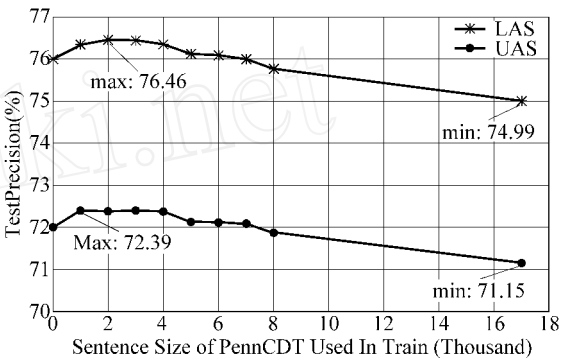


图 6 依存句法分析器在不同规模树库上训练后的测试结果

我们将转化过程中的错误分为三种类型:

一是 PennCT 和 HIT-IR-CDT 在分词标准上的差异。

二是短语结构转化为依存结构时产生的错误,我们的规则还过于简单,对于核心节点映射表的修改不够精细,没有完全了解 PennCT 和 HIT-IR-CDT 之间的差异。这个工作还需要继续研究。短语转依存这一步骤是整个转化过程的基础,直接影响后续的依存关系标注、依存分析训练的效果。

三是依存关系标注过程中带来的错误,有些关系类型的识别效果太差,直接影响 PennCDT 中各种关系的分布平衡。我们需要仔细分析原因,同时研究 HIT-IR-CDT 的标注标准。

## 6 结论

我们尝试使用基于规则与统计相结合的方法,将 PennCT 转化为 HIT-IR-CDT 的体系结构。我们首先利用词性标注器对词性标注集进行了转化;进而利用核心节点映射表及各种启发式规则将短语结构转化为依存结构。对于 PennCT 中存在的扁平短语结构,我们采用依存分析器进行分析,得到短语内部的依存关系;最后我们利用 Online 算法训练出一个依存关系标注器,标注每条依存弧的依存关系。我们将转化后的树库 PennCDT 加入到 HIT-IR-CDT,训练并测试依存句法分析器。通过对实验结果仔细分析,我们认为这种基于规则与统计相结合

的树库转化方法是有效的。但是由于 PennCT 和 HIT-IR-CDT 在标注体系上存在较大的差异,转化方法还需要继续深入研究。

下一步工作主要是研究 PennCT 和 HIT-IR-CDT 在标注体系上的差异性,细化短语结构转依存的过程。并借此将 PennCT 融入到 HIT-IR-CDT 中,对 HIT-IR-CDT 的标注体系进行改善。

致谢　王丽杰同学为本文的实验提供了汉语词性标注器,胡禹轩同学为本文提供了短语树的可视化程序,曹浩同学为本文提供了依存结构的可视化程序,在此一并表示感谢。

## 参考文献:

[1] 马金山. 基于统计方法的汉语依存句法分析研究[D]. 博士毕业论文, 哈尔滨工业大学,2007.

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

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

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

[5] Joakim Nivre, Mario Scholz. Deterministic Dependency Parsing of English Text [C]// Proceedings of COLING. 2004: 64-70.

[6] Richard Johansson, Pierre Nugues. Extended constituent-to-dependency conversion for English [C]// Proceedings of NODALIDA,2007: 105-112.

[7] 党政法,周强. 短语树到依存树的自动转换研究[J]. 中文信息学报, 2005(3) : 21-27.

[8] Nianwen Xue, Fei Xia. The Bracketing Guidelines for the Penn Chinese Treebank (3.0)[R]. 2000.

[9] Ting Liu, Jinshan Ma, Sheng Li. Building a Dependency Treebank for Improving Chinese Parser [J]. Journal of Chinese Language and Computing, 2006, 16(4): 207-224.

[10] Ryan McDonald. Discriminative learning and spanning tree algorithms for dependency parseing [D]. Ph. D. thesis, University of Pennsylvania, Philadelphia. 2006.

(上接第 7 页)

封闭测试的效果与上面两个实验没有什么差别,都取得了非常理想的效果。开放测试的调和平均值虽然没有含有嵌套的联合结构特征模板二的识别效果 86.91 %高,但考虑到模板三所用特征的数量,其结果还算合理。最长联合结构识别的精确率和召回率比较接近,这从另一个侧面说明了如果语料规模足够大、数据稀疏问题解决得好,从整体上可以提高联合结构识别的精确率和召回率,而不像表 3 一样一个偏高一个偏低。

## 6　结论

本文使用 CRF 统计模型,分别用基于复杂特征的特征模板和增加语言学特征的特征模板在含有嵌套的联合结构、无嵌套联合结构和最长联合结构语料上进行了实验,取得了相对满意的结果。我们下一步的工作主要为: 在提高计算机硬件如内存的前提下,扩大训练语料库的规模,更深入地检验 CRF 的性能;从已标注的联合结构语料中挖掘新的语言学知识,在 CRF 模型中添加新的语言学特征,进而观察 CRF 模型的性能。

## 参考文献:

[1] 周强. 汉语语料库的短语自动划分和标注研究[D]. 北京: 北京大学, 2002.

[2] 孙宏林. 现代汉语非受限文本的实语块分析[D]. 北京: 北京大学, 2001.

[3] Fei Sha, Fernando Pereira. Shallow Parsing with Conditional Random Fields [C]// Proceedings of Human Language Technology Conference and North AmericanChapter of the Association for Computational Linguistics (HLT-NAACL), 2003,135-136.

[4] 周俊生,陈家骏. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报,2006,(5): 805.

[5] 李双龙,刘群. 基于条件随机场的汉语分词系统[J]. 软件天地,2006,(10): 178-179.

[6] 吴云芳. 面向中文信息处理的现代汉语并列结构研究[D]. 北京: 北京大学,2003.