北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis

# 汉语并列关系的识别研究

郑略省　吕学强†　刘坤　林进

北京信息科技大学网络文化与数字传播北京市重点实验室, 北京 100101; † 通信作者, E-mail: lv.xueqiang@trs.com.cn

**摘要**　针对汉语并列关系的标注方式, 提出一种基于条件随机场模型的并列关系自动识别方法。从语料库中自动抽取并列关系的角色信息, 进行角色标注, 在条件随机场模型的基础上实现并列关系的识别。与基于图的依存分析方法比较, 并列关系的召回率和正确率分别提高了 9.1%, 13.8%。

**关键词**　依存句法分析; 条件随机场; 角色标注; 并列关系

**中图分类号**　TP391

# Automatic Identification of Chinese Coordination Relations

ZHENG Lüexing, LÜ Xueqiang†, LIU Kun, LIN Jin

Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101; † Corresponding author, E-mail: lv.xueqiang@trs.com.cn

**Abstract**　The authors presented an approach of Chinese coordination relations recognition based on CRFs. Tokens were tagged with different roles according to their functions in the generation of Chinese coordination relations. Then coordination relations were recognized by CRFs (conditional random fields). Compared with the maximum spanning tree dependency parsing, the experiment shows that recall and precision of coordination relations increase by 9.1%, 13.8%.

**Key words**　dependency parsing; CRFs; role tagging; coordination relations

句法分析一直是自然语言深层处理的核心问题之一。在依存句法中, 词与词之间是直接发生依存关系, 构成一个依存对, 其中一个是核心词, 另一个是从属词。依存关系使用一个有向弧表示, 称为依存弧。每个依存弧上都有一个标记, 称为关系类别, 表示该依存对中的两个词之间的依存关系。如图 1 所示, 例句“马文瑞、汪峰等老同志也出席了茶话会。”中, 每个词都依存于一个其他的词,其中“出席”是句子的根节点, 依存于虚根(Root)。“老”依存于“同学”, 依存关系为定中关系(ATT)。“汪峰”依存

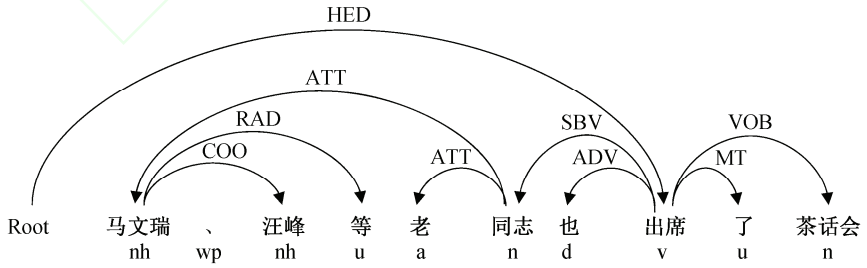


**图 1**　例句
Fig. 1　Example

国家自然科学基金项目(61171159)、北京市教委科技发展计划项目(KM201110772021, KM201211232023)、国家科技支撑计划课题(2011BAH11B03)资助

收稿日期: 2012–06–06; 修回日期: 2012–08–13; 网络出版时间: 2012-10-26 17:49

网络出版地址: http://www.cnki.net/kcms/detail/11.2442.N.20121026.1749.010.html

于“马文瑞”, 依存关系为并列关系(COO)。

近几年, 国内外众多的研究者投入到依存句法分析的工作中来。目前研究的重心放在统一建模上, 很少针对某些特定的语言现象, 识别它们的依存关系, 这制约着句法分析的发展。McDonald 等[1]将依存分析问题归结为在一个有向图中寻找最大生成树(maximum spanning tree)的问题。Nivre 等[2]采用了确定性的分析算法。中国科学院段湘煜等[3]提出基于动作的多阶段算法。哈尔滨工业大学辛霄等[4]提出基于最大熵的依存句法分析。而马金山针对中文特定的语言现象单独进行分析, 提出了一种基于动态局部优化的搜索算法[5−6], 提高了特定结构的识别效果。

McDonald 方法的训练效率和分析性能等方面都表现比较好。其以 HIT-IR-CDT[7]作为训练和测试语料库, 整体的识别效果指标 LAS 为 78.2%, 但并列关系识别率偏低, 召回率和正确率为 54.8%和 64.0%。汉语中平均两个句子就存在一个并列结构, 这较大程度地影响了整体的识别效果。本文将采用分而治之的策略, 利用并列结构在空间上的连续性和平行性特点, 在条件随机场[8]的基础上, 识别并列关系, 改善了并列关系的识别效果。

## 1 并列关系的标注方式

依存语法中并列关系(COO)的标注方式主要由并列词组、核心词和尾词组成。并列词组, 指的是在同一并列结构中发生并列关系的所有并列成分。核心词, 指的是在并列结构中有一个并列成分充当核心节点的作用, 是并列结构同句子的其他结构或词发生依存关系的词, 其他并列成分均以核心词为父亲节点。尾词, 指的是距离核心词最远的并列成分。如图 2 所示, 该例句并列关系的标注方式是遵循左核心原则, 最左边的并列成分为核心词。

1) “贵州南部、江南、华南西部”为一个并列结构。

2) “南部”, “江南”, “西部” 为一个并列词组, 均为并列成分。

3) “南部”为“江南”和“西部”的父亲节点, 为核心词, 最右边的“西部”为尾词。

4) “南部”与“江南”, “南部”与“西部”是两对并列成分。识别成对的并列成分是识别汉语并列关系的主要任务, 即识别非核心词的并列成分和其依附的父亲节点。

## 2 基于条件随机场的并列关系自动识别

### 2.1 并列关系的构成角色

角色表是识别并列关系的基础。根据角色表, 计算机能够理解汉语并列结构。制定角色表则需要对汉语并列关系进行统计分类。

在汉语依存语法语料库中, 并列关系可以分为无标记和有标记并列关系两类。无标记并列关系相对于有标记并列关系而言, 数量比较少, 结构复杂, 不易识别。如“指手画脚, 照本宣科”, “深入细致, 扎实有效”。有标记并列关系长度跨度大, 结构上由并列标记连接, 是并列关系的主要特征, 该类的特点对识别并列关系有很重要的意义。有标记并列关系主要有以下两种[9]。

1) 连词: 主要是中置连词, 在语料库中的词性标记为“c”, 包括“和、与、并、及、或、或者……”, 例如“中国和南非”。

2) 标点符号: 主要是逗号为主, 例如: “一国两制”、“港人治港”、“高度自治”。

汉语依存句法分析中有标记并列关系比较难识别的是嵌套并列关系, 主要困难在于个别并列成分充当多重角色。如图 3 所示。

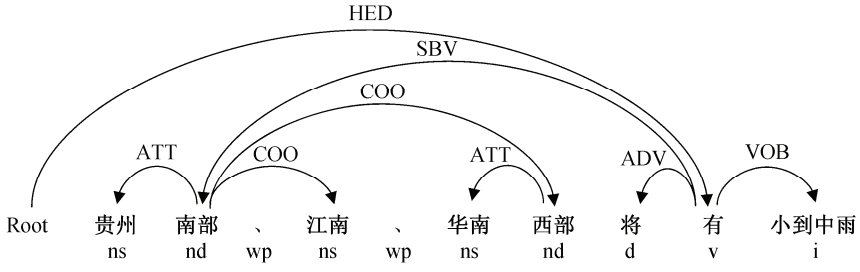


**图 2 例句**

Fig. 2 Example

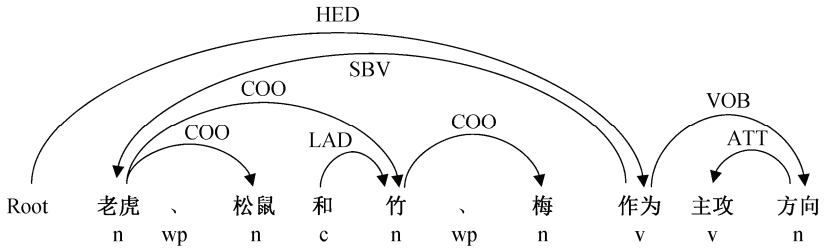


**图 3　例句**

Fig. 3　Example

“老虎、松鼠”和“竹、梅”是两个并列词组, 它们的核心词“老虎”与 “竹”也是一个并列词组。也就是说“竹”在“竹、梅”中充当核心词, 但在“老虎”和“竹”中充当尾词。这种嵌套的并列关系通常会被错误地识别为单一的并列词组, 即“老虎”、“松鼠”、“竹”和“梅”组成一个并列词组。

因此该文根据并列关系的标注方式, 针对各种并列关系的特点, 将并列关系的内部组成、并列标记和上下文等称为并列关系的构成角色, 如表 1 所示。

## 2.2　并列关系的识别

特征的合理选择是识别并列关系的关键, 特征集是判别某个词或字在并列关系中充当某种角色的主要依据。由于目前汉语依存语料库规模较小, 本文选取词性作为识别并列关系的主要特征。

在同一个并列结构中具有并列关系的成分之间存在一定的规律性。首先通过统计, 语料库的 6815 个并列关系中, 词性相同的就有 5997 个, 比例为 87.9%。

并列结构还有个很重要的特点, 就是结构的平行性[10], 也就是修饰词的共享或相似, 如“各种(/r) X 形(/n)、Y 形(/n)、蝶形(/n)等(/u)”, “当地(/nl)群众(/n)和外地(/nl)游客(/n)”。“各种(/r)”为共享修饰词, “当地(/nl)”, “外地(/nl)”为相似修饰词, “群众”和“游客”是并列成分。但由于汉语是意合语言, 其并列结构还有其它表现形式, 较难识别的是修饰词和被修饰词的词性均为“n”的情况。如“企业(/n)及(/c)投资(/n)机构(/n)”, “企业”与“机构”并列, “投资(/n)”只是机构的修饰词。“政治(/n)和(/c)工资(/n)待遇(/n)”, “政治(/n)”和“工资(/n)”并列, 共同修饰“待遇(/n)”。识别词性均为“n”的并列结构, 重点在于判断哪些词是修饰词。因此本文从训练集中, 将词性为“n”的词根据阈值分为 3 类:

$$P(\mathrm{ATT})=\frac{C_{\mathrm{ATT}}}{C_{\mathrm{ATT}}+C_{\mathrm{VOB}}+C_{\mathrm{SBV}}};$$

$C_{\mathrm{ATT}}$ 表示词性为“n”的依附关系为定中关系; $C_{\mathrm{VOB}}$ 表示词性为“n”的依附关系为动宾关系; $C_{\mathrm{SBV}}$ 表示词性为“n”的依附关系为主谓关系。

1) P(ATT)大于 0.9 为 A, 常为修饰词, 汉语依存语法中一般表现为定中关系, 如 “爱国人士(/n)邵逸夫(/nr)”中的 “爱国人士”。

2) P(ATT)小于 0.1 为 Q, 常为被修饰词, 汉语中一般表现为主谓或动宾关系, 如“大(/a)剧院(/n)芭蕾舞团(/n)演出(/v)”中的“芭蕾舞团”。

**表 1　并列关系角色表**

Table 1　Roles of coordination relations

| 编码 | 意义 | 例子 |
|---|---|---|
| H | 并列词组的核心词 | 经济 、 政治 和 外交 　指手画脚 ， 照本宣科 |
| X | 并列词组的核心词，又是另一个并列词组的非核心词 | 松鼠 、 麻雀 和 竹 、 梅 、 松 、 柏 |
| B | 并列词组的非核心词 | 青草 、鲜花 和 河流 、 湖泊 |
| R | 并列词组内部的并列标记 | 仓库 、 厂房 和 民宅 |
| I | 并列词组内部的非并列标记 | 汇率 和 股票 价格 |
| L | 核心词的上文 | 贵州 南部 、 江南 、 华南 西部 |
| F | 尾词的下文 | 华南 西部 和 北部 有 小到中雨 |
| C | 既是上文又是下文 | 那些 诗句 、 那些 祝辞 ， 喜悦 、 激动 、 欣慰 之 情 |
| O | 以上之外其它的角色 | |

3）M，介于 A 与 Q 之间。

在表 2 特征集中，W 代表词，P 代表词性，D 表示词性为“n”的类别(A, Q, M, U)，U 表示词性非“n”。括号内的数值代表位置信息。注：词性为“n”但未在训练集中出现的词的 D 类别均标为 M。

本文方法的测试结果如表 3 所示，“形式”和“程度”被识别为一对并列成分，其中“形式”是核心词。依据并列关系的标注方式，得到并列词组中非核心词的依附关系，即“程度”依存于“形式”，“程度”的父亲节点是词位置为 4 的“形式”，如表 4 所示。

**表 2　特征集**

Table 2　Features

| 原子特征 | 复合特征 |
|---|---|
| W($i$) | W($i$)+ P($i$) |
| W($i$+1) | W($i$+1)+ P($i$+1) |
| W($i$+2) | W($i$+2)+ P($i$+2) |
| W($i$–1) | W($i$–1)+ P($i$–1) |
| W($i$–2) | W($i$–2)+ P($i$–2) |
| P($i$) | D($i$)+ P($i$) |
| P($i$+1) | D($i$+1)+ P($i$+1) |
| P($i$+2) | D($i$+2)+ P($i$+2) |
| P($i$–1) | D($i$–1)+ P($i$–1) |
| P($i$–2) | D($i$–2)+ P($i$–2) |
| D($i$) | P($i$+2)+ P($i$+1)+ P($i$) |
| D($i$+1) | P($i$+1)+ P($i$)+ P($i$–1) |
| D($i$+2) | P($i$)+ P($i$–1)+ P($i$–2) |
| D($i$–1) | P($i$+3)+ P($i$+2)+ P($i$+1)+P($i$) |
| D($i$–2) | P($i$+2)+ P($i$+1)+ P($i$)+ P($i$–1) |
| | P($i$+1)+ P($i$)+ P($i$–1)+P($i$–2) |
| | P($i$)+ P($i$–1)+ P($i$-2)+ P($i$–3) |

**表 3　CRFs 识别结果**

Table 3　Recognition with CRFs

| 分词结果 | 词性 | 词性为“n”的类别 | 角色标记 |
|---|---|---|---|
| 实行 | v | U | O |
| 了 | u | U | O |
| 不同 | a | U | L |
| 形式 | n | Q | H |
| 、 | wp | U | R |
| 不同 | a | U | I |
| 程度 | n | M | B |
| 的 | u | U | F |
| 村务公开 | l | U | O |
| 。 | wp | U | O |

**表 4　并列关系识别结果**

Table 4　Recognition of coordination relations

| 分词结果 | 词性 | 词性为“n”的类别 | 父亲节点 ID | 依存关系 |
|---|---|---|---|---|
| 实行 | v | U | | |
| 了 | u | U | | |
| 不同 | a | U | | |
| 形式 | n | Q | | |
| 、 | wp | U | | |
| 不同 | a | U | | |
| 程度 | n | M | 4 | COO |
| 的 | u | U | | |
| 村务公开 | I | U | | |
| 。 | wp | U | | |

## 3　实验结果分析

本文以 HIT-IR-CDT 前 8000 句作为训练语料，后 1000 句作为测试语料，每个句子的平均长度为 21.3 个词。MSTparser 依存句法分析器是 McDonald 方法[1]的实现，也在同等条件下进行训练和测试。本文方法与 MSTparser 对比的实验结果如表 5 所示，

**表 5　并列关系的识别数据**

Table 5　Results of coordination relations recognition

| 并列关系类别 | 正确数量 | MSTparser | | | 本文方法 | | | $F$ 值变化 |
|---|---|---|---|---|---|---|---|---|
| | | 召回率/% | 准确率/% | $F$ 值 | 召回率/% | 准确率/% | $F$ 值 | |
| ALL | 701 | 54.8 | 64.0 | 0.590 | 63.9 | 77.8 | 0.702 | +0.112 |
| Same_n | 357 | 69.2 | 75.3 | 0.721 | 71.1 | 84.9 | 0.774 | +0.053 |
| Same_v | 122 | 24.6 | 54.5 | 0.339 | 41.8 | 82.3 | 0.554 | +0.215 |
| Diff_n_v | 222 | 67.1 | 68.7 | 0.679 | 70.3 | 72.6 | 0.714 | +0.035 |
| Label | 549 | 66.5 | 82.6 | 0.737 | 76.0 | 81.3 | 0.785 | +0.048 |
| UnLabel | 152 | 40.8 | 39.2 | 0.400 | 33.6 | 81.0 | 0.474 | +0.074 |

注：准确率=正确识别的数目/识别出的数目 × 100%；召回率=正确识别的数目/实际正确数目 × 100%；$F$ 值=准确率 × 召回率 × 2/(准确率+召回率)。

ALL 表示在测试语料中所有的并列关系。为更好地评价识别效果, 将每对并列成分划分为以下 5 个类: 1) 两词性均为“n”(Same_n), 表一般名词并列; 2) 两词性均为“v”(Same_v), 表动词并列; 3) 两词性非 1 和 2 的情况(Diff_n_v); 4) 含有并列标记(Label); 5) 不含有并列标记(UnLabel)。

实验结果证明了本文的方法有效地提高了并列关系识别的效果。统一建模的分析器学习某些特定结构的能力较差, 采用分而治之的方法, 可以弥补此不足。从表 3 可以发现并列关系的整体识别效果有较大提升, 正确率和召回率分别提高了 9.1%, 13.8%。

由于依存语料库规模较小, 词性成为识别并列关系主要依据, 本文的方法有效地利用词性的信息, 提高了识别同类型的并列成分的效果。同时对比重较大的 Same_n 和 Label 并列关系的识别, 也优于统一建模的方法。

分析主要的识别错误, 可以分为以下两类。

1) “全市(/n) 党政(/n) 机关(/n)、(/wp)事业(/n) 单位(/n) 公款(/n)。”该例句识别为“机关”与“公款”并列。主要原因在于汉语是意合言语, 目前还很难利用语义的信息进行句法分析。这样造成多名词的并列结构识别比较困难。

2) “校园网(/n)和(/c)外面(/nd)的(/u)世界/n。”该例句识别为“校园网”与“外面”并列。主要原因在于汉语语料库规模较小, 对大部分词或字的学习不够充分。由于“的”字常充当并列关系的下文角色, 使“世界”无法成为相应的并列成分。

## 4 结论

本文采用分而治之的策略, 研究了汉语依存语法中并列关系的标注方式和内外环境信息, 利用并列结构的汉语特点, 改善了并列结构的识别效果。实验证明该方法是行之有效的。下一步的研究工作是将并列关系的识别效果进一步提高, 总结汉语其它语言现象, 改善汉语依存句法分析的效果。

致谢 感谢哈尔滨工业大学信息检索研究中心语言技术平台提供的依存树库(HIT-IR-CDT)。

### 参考文献

[1] McDonald R, Lerman K, Pereira F. Multilingual dependency analysis with a two-stage discriminative parser // Proceedings of the 10th Conference on Computational Natural Language Learning. New York , 2006: 54–62

[2] Nivre J, Hall J, Nilsson J, et al. Labeled Pseudo-Projective dependency parsing with support vector machines // Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL). New York, 2006: 221–225

[3] 段湘煜, 赵军, 徐波. 基于动作建模的中文依存句法分析. 中文信息学报. 2007, 21(5): 25–30

[4] 辛霄, 范士喜, 王轩, 等. 基于最大熵的依存句法分析. 中文信息学报, 2009, 23(2): 18–22

[5] Ma Jinshan, Zhang Yu, Liu Ting, et al. A statistical dependency parser of Chinese under small training data // Workshop: Beyond shallow analyses-Formalisms and statistical modeling for deep analyses, IJCNLP-04. Sanya, China, 2004: 43–51

[6] Liu Ting, Ma Jinshan, Zhu Huijia, et al. Dependency parsing based on dynamic local optimization // Proceedings of Tenth Conference on Computational Natural Language Learning, CoNLL-shared task. New York, 2006: 65–73

[7] Che Wanxian, Li Zhenghua, Liu Ting. LTP: a chinese language technology platform // Proceedings of the Coling 2010: Demonstrations. Beijing, 2010: 13–16

[8] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data // Proceedings of the 18th International Conference on Machine learning. San Francisco, 2001: 282–289

[9] 吴云芳, 石静, 金彭. 基于图的同义词集自动获取方法. 计算机研究与发展, 2011(4): 610–616

[10] 苗艳军. 汉语并列结构的自动识别[D]. 苏州大学, 2009