文章编号：1003-0077(2008)02-0010-08

中文词法分析与句法分析融合策略研究

米海涛[1,2]，熊德意[1,2]，刘 群[1]

（1. 中国科学院 智能信息处理重点实验室 北京 100080；
2. 中国科学院 研究生院 北京 100039）

摘 要：利用外部资源是提升句法分析性能的一种有效方法。本文利用中文词法分析器这一外部资源，提出了一种通用转换方法将中文词法分析器与句法分析器有机地融合在一起。通过基于转换的错误驱动学习和条件随机场解决不同切词、词性标注标准间的转换问题。在句法分析方面，本文提出了多子模型句法分析器，将中心词驱动模型和结构上下文模型有效结合在一起。融合后的中文句法分析性能在宾州中文树库 1.0 版①测试集上 F1 值达到了 82.5%的最好水平。

关键词：计算机应用；中文信息处理；中文句法分析；中文词法分析；融合策略；基于转换的错误驱动学习；条件随机场
中图分类号：TP391 文献标识码：A

Research on Strategies for Integrating Chinese Lexical Analysis and Parsing

MI Hai-tao[1,2], XIONG De-yi[1,2], LIU Qun[1]

(1. Key Laboratory of Intelligent Information Processing Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China;
2. Graduate University of Chinese Academy of Sciences, Beijing 100039, China)

Abstract: External resources can be used effectively to improve the parsing accuracy. In this paper, we introduce an external Chinese lexical analysis system to parsing and propose a general transformation methd to integrate them. The Transformation-based Error-driven Learning and Conditional Random Fields are used to solve the problem of transformation between two different standards of segmentation and POS tagging. We also propose a parsing model which combines the Head-driven parsing model and Structural Context parsing model effectively. Experimental results show that our new integrated parsing model achieves an F1 score of 82.5% on the Penn Chinese Tree-Bank Version 1.0, higher than the state-of-art parsers.

Key words: computer application; Chinese information processing; Chinese parsing; Chinese lexical analysis; integrating strategy; transformation-based error-driven learning; conditional random fields

1 引言

英语句法分析性能已经达到了 90%的水平，然而中文句法分析还是在 80%左右，如何提高中文句法分析性能成为一个关键问题。当前中文句法分析在很大程度上受到训练数据规模小的限制，宾州中文树库 1.0 版只有四千多句、5.0 版也只有近两万句，如果能够有效利用外部资源，将对中文句法分析有很大帮助。

在宾州中文树库 1.0 版上，Bikel 博士论文[1]使用外部词性标记模块对未登录词进行标记，该模型

收稿日期：2007-05-15 定稿日期：2007-07-13
基金项目：国家自然科学基金资助项目(60603095, 60573188)

作者简介：米海涛(1981—)，男，博士生，主要研究方向为中文句法分析、机器翻译；熊德意(1979—)，男，博士生，主要研究方向为机器翻译、中文句法分析；刘群(1966—)，男，博士，研究员，主要研究方向为机器翻译和自然语言处理。

① 文章中所有实验都按照宾州中文树库 1.0 版标准划分进行。训练集：1—270 篇；开发集：301—325 篇；测试集：271—300 篇。

标记召回率/准确率(LR/LP)为 78.0%/81.2%；Xiong[2] 利用了外部资源的语义类信息，LR/LP 为 78.7%/80.1%。他们都有效利用外部资源提升了句法分析的性能。

词性标记(POS tag)准确率对句法分析性能的影响是十分显著的，当词性标记完全使用 Gold-standard 标记(标记准确率为 100%)时，Jiang[3] 作业论文、Wang[4] 的实验结果显示句法分析性能都有显著性提高，F1 值大约都有 5~8 个百分点的提高。要提高词性标记准确率，我们可以采用不同的策略实现：Luo[5] 提出了基于字的最大熵句法分析器，该方法在训练过程中将词级句法分析树转换成字级句法分析树，如图 1 所示；句法分析过程使用最大熵方法对字进行词性标记、句法分析，最后再转换成词级句法分析树。Fung[6] 的模型将切词、词性标记、句法分析整合在一起，与 Luo 不同的是，该模型将切词作为一个单独模块实现不做字级句法分析。从另一角度我们可以认为这两种模型都直接利用宾州中文树库训练语料设计一个词法分析器。词法分析结果可以直接作为句法分析的输入，整个模型具有一致性，但是该方法也有缺点：训练语料规模受到很大的限制！如果直接利用宾州中文树库 1.0 版训练一个词法分析器，由于树库本身规模较小，只有四千句约十万词，使得词法分析器性能较低，另一方面，句法分析性能对词性标记准确率很敏感，所以要利用高性能的词法分析器。我们应当利用外部资源。我们很清楚，当前的一些高效中文词法分析器已经使用大规模切词标注语料库训练，如：北京大学加工的《人民日报》语料库 1 300 万字，约 730 万词。并且词法分析性能已经达到了很高的水平，如果能将外部的词法分析器与句法分析有机地融合在一起，将会是个出色的解决方案。

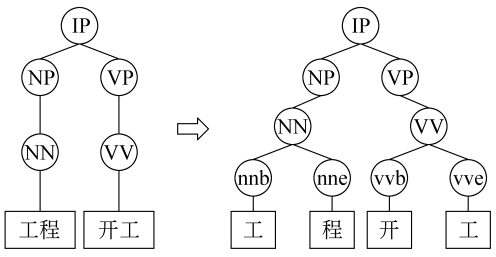


图 1　词级句法树到字级句法树转换

利用外部词法分析器，我们将会碰到很多困难问题。由于词法分析器采用的训练语料库与句法分析器训练中采用的树库完全不同，将导致词法分析结果不能直接作为句法分析的输入，这与 Luo、Fung 的工作是不同的。词法分析与句法分析语料的切词和词性标记标准不尽相同，标准不一致将导致不能直接利用中文词法分析器。不同切词、标记标准之间的转换本来就是个难题，切词方面：在一种标准下可能是一个词，但是在其他标准中可能是两个词；词性标记方面：不同标准下，有很多不一致性，标记集合不同、标记详细程度不同、词性划分也不同等。以上种种难题是我们必须要解决的，文章提出了一种融合策略有效解决了以上难题，将中文词法分析和句法分析有效地融合在一起，句法分析性能在宾州中文树库 1.0 版测试集上 F1 值达到了 82.5%的最好水平。

2　句法分析系统

句法分析系统是多子模型句法分析系统，它将中心词驱动模型[2] 和结构上下文模型[7] 按照对数线性模型融合在一起，并在此基础上加入了标点限制模块。

2.1　中心词驱动模型

该模型类似于 Collins[8] 模型 1，分析树概率为：$P(t)=\prod_{dl\ rl} P(hi)$，相应的每条规则词汇化为：

$$P(h)\rightarrow L_n(l_n)\cdots L_1(l_1)H(h)R_1(r_1)\cdots R_m(r_m) \tag{1}$$

H 是该规则的中心成分；$L_n(l_n)\cdots L_1(l_1)$和 $R_1(r_1)\cdots R_m(r_m)$是中心成分的左右修饰成分。

每条规则概率分解成 3 个部分：

1. 生成中心成分句法标记的概率：

$$P_H(H\mid P, ht(H), hw(H)) \tag{2}$$

$hw(H)$、$ht(H)$分别表示中心成分 H 的中心词和中心词词性标记。

2. 选择修饰成分的句法标记及其对应的中心词词性标记的概率：

$$P_M(M_i, ht(M_i)\mid P, H, ht(H), hw(H), M_{i-1}, ht(M_{i-1}), dir, dis) \tag{3}$$

$M=L$(左)或 R(右)，dir：方向，dis：距离。

3. 选择新生成的修饰成分的中心词的概率：

$$P_{hw(M)}(hw(M_i)\mid P, H, ht(H), hw(H), M_{i-1}, ht(M_{i-1}), dir, dis, M_i, ht(M_i)) \tag{4}$$

由于数据稀疏对第 3 个概率影响较大，加入语义类信息，第 3 个概率重新定义为：

$$P_{hw(M)new}=\lambda P_{hw(M)}+(1-\lambda)P_{sym} \tag{5}$$

$P_{hw(M)}$为(4)式定义，P_{sym}为语义模型概率，λ通过*EM*算法确定。

该模型在宾州中文树库1.0版测试集上F1值达到了79.35%的水平。

2.2 结构上下文模型

该模型是考虑父节点使用规则和节点位置的条件概率模型(PRORD)。分析树概率为：

$$P_{PRORD}(t)=\prod_{(A\to\alpha)\in R}P(A\to\alpha\mid A,ParentRule(A),Order(A)) \quad (6)$$

其中，$ParentRule(A)$为非终结符A的父规则；$Order(A)$为非终结符A在父规则中的位置。

2.3 多子模型句法分析器

多子模型句法分析系统将中心词驱动模型和结构上下文模型按照对数线性模型融合在一起，形成多子模型句法分析器，分析树的概率重新定义为：

$$P(t)=\frac{\exp\left[\sum_{m=1}^{M}\lambda_m h_m(t)\right]}{\sum_{t'}\exp\left[\sum_{m=1}^{M}\lambda_m h_m(t')\right]} \quad (7)$$

$$\tilde{t}=\arg\max_{t}\left\{\sum_{m=1}^{M}\lambda_m h_m(t)\right\} \quad (8)$$

目标决策函数：

两个子模型按照特征处理。如，结构上下文子模型的特征表示成：

$$h_{m_1}(t)=\log\left(\prod_{(A\to\alpha)\in R}P(A\to\alpha\mid A,ParentRule(A),Order(A))\right) \quad (9)$$

子模型相应的权重我们可以采用最小错误率训练[9]或者单纯形算法[10]进行调节，我们利用开发集对两个子模型的权重进行了调节。该方法有效解决了多个特征的融合问题，引入更多更有效的特征将更有利于句法分析性能的提高。

我们还注意到：汉语中的标点往往带有一定的句子结构信息，这种信息对句法分析是比较有用的。我们利用标点信息制定一些启发式规则对句法分析进行辅助修正。标点符号限制规则主要包括以下两种：

A. 对于对称标点的处理：子树span[①]不能与对称标点符号span交叉；

B. 对于其他标点符号的处理：利用标点符号在子树中的位置信息，制定规则对错误的子树进行限制，例如：子树的第一个词不能是逗号等。

利用标点符号的信息对句法分析进行修正，一方面，它限制了一些边的生成，在一定程度上减少了句法分析的搜索空间；另一方面，使得句法分析的错误率降低，提高了句法分析的性能。

多子模型句法分析器在宾州中文树库1.0版测试集上F1值达到了80.74%的水平，比中心词驱动模型[2]高出了1个多点，显著性测试结果显示该提高在统计上是明显的，证明了我们这种多子模型组合的有效性。

3 词法分析系统(ICTCLAS)

计算所汉语词法分析系统(Institute of Computing Technology, Chinese Lexical Analysis System，简称ICTCLAS)[11,12]，采用层叠隐马模型(Cascaded Hidden Markov Mmodel, CHMM)将切词、词性标记有效地融合在一起。该系统在国家973项目相关主题专家组组织的汉语分词标记评测和国际SigHan2003研讨会组织的汉语分词评测中分别获得多项第一。

ICTCLAS训练语料为北京大学计算语言学研究所加工的《人民日报》语料库，词性标注集合为ICTPOS3.0[13]。ICTPOS3.0总共包括名词(n)、动词(v)等22类标记，每一类还分了子类，如名词分为：人名(nr)、地名(ns)、机构团体名(nt)、其他专名(nz)、名词性惯用语(nl)、名词性语素(ng)等六个子类。与ICTCLAS相比，宾州中文树库的切分和标记标准有些不同。宾州中文树库切词[14]和标记[15]都制定了详细的标准，宾州中文树库POS3.0总共包括专有名词(NR)、一般名词(NN)等33个标记。二者之间的差异是文章要解决的关键问题。

4 融合策略

图2给出了中文句子的句法分析的总过程，如何将词法分析结果转换成宾州中文树库标准结果成为最关键的问题。可以看到两个结果之间有两种不同之处，一是切词不一致，二是词性标记集合不一致。针对不一致性我们提出了一种有效的转换方法，包括两个转换步骤，第一步为切分转换，第二步为标记转换。加入转换方法之后，我们得到完整的

① span：在中文句子中的覆盖范围。

流程图，如图 3。在下面两个部分中我们会详细介绍这两个转换步骤。

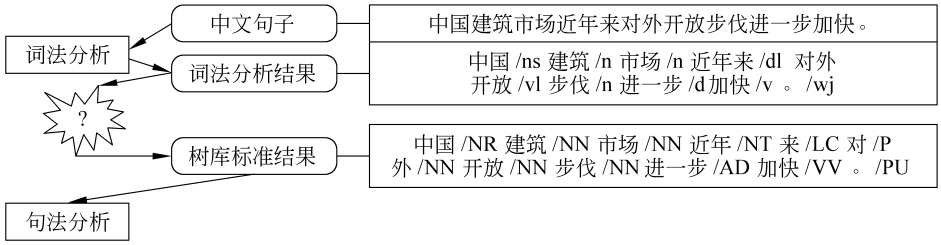


图 2　流程图

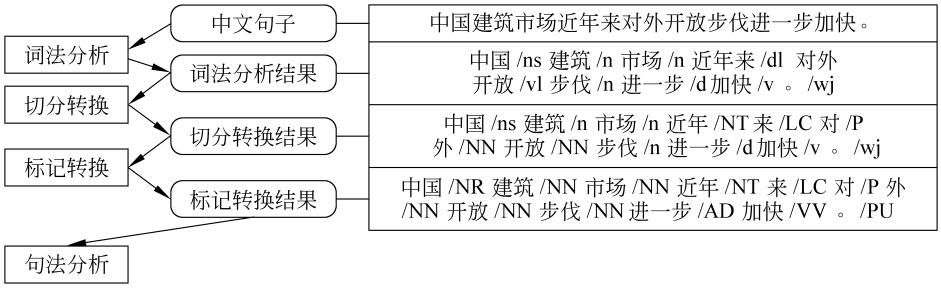


图 3　完整流程图

4.1　切分转换：基于转换的错误驱动学习

由词法分析结果到切分转换结果，我们采用了基于转换的错误驱动学习方法[16]。利用该方法自动学习一个规则集合，再加上人工转换规则（从两种切分标准说明中总结得到），完成切分转换。

基于转换的错误驱动学习方法是一种自动学习规则的方法，这种方法在很多自然语言处理问题中应用，如：词性标记、介词短语附属消歧等等。图 4 给出了基于转换的错误驱动学习方法的流程图。应用于本实例：在训练集上，ICTCLAS 分析结果与宾州中文树库结果进行比较，从差异部分中，算法可以学习到一个转换规则集，这个转换规则集能使 ICTCLAS 切分结果的差异得到最大程度的修正。

每个转换规则包含重写规则和触发环境两部分。

1．重写规则：把前一个词和当前词合并，标记为“CD”，

2．触发环境：前一个词词性为数词 m，当前词为“多”或“成”或“余”，当前词性为数词 m。

该规则可以将“300/m 多/m”转换为“300 多/CD”。

图 4 中从“学习器”到“词法分析结果”的反馈边主要是对一些规则的再判断过程，例如：如果该规则在进行切分转换时引入错误，则会加强该规则的使用条件即触发环境。对转换规则的再判断过程保证了规则的有效性。

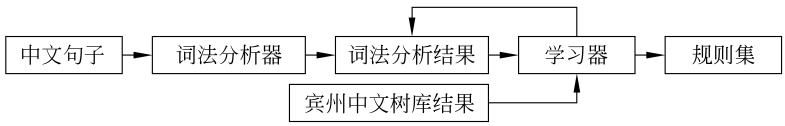


图 4　基于转换的错误驱动学习方法流程图

4.2　标记转换：条件随机场（CRF）

ICTCLAS 词性标记标准为 ICTPOS3.0，而宾州树库词性标记标准为宾州中文树库 POS3.0，我们的任务是将 ICTPOS3.0 转换为宾州中文树库 POS3.0。这个标记转换过程，我们利用序列标记模型条件随机场（CRF）完成。

CRF 是计算具有无向图 G 结构的随机变量集合 Y（标记序列）在给定随机变量集合 X（观察序列）下的条件概率 $P(Y|X)$。无向图 $G=(V, E)$，$Y=(Y_v)$ $v\in V$（每个顶点代表标记序列的一个变量），则当以 X 观察序列为条件，Y 组成的无向图为一个

随机场，Y_v 服从马尔可夫性质，$P(Y_v \mid X, Y_w, w \neq v) = P(Y_v \mid X, Y_w, w \sim v)$，$w \sim v$ 表示 w 与 v 相邻，w 与 v 之间有边直接相连。概率公式为：

$$P(Y \mid X) = \frac{1}{Z} \prod_{t=1}^{T} \phi(y_{t,n}, X) \qquad (10)$$

其中，$y_{t,n}$ 表示 y_t 和 y_t 邻节点集合，Z 为归一化因子，$Z = \sum_{\text{所有可能的标记序列 } Y} \prod_{t=1}^{T} \phi(y_{t,n}, X)$；$\phi(y_{t,n}, X)$ 为与 $y_{t,n}$、X 相关的势函数。

在重标记试验中，我们利用一阶 CRF，其中 X 为重切分结果，Y 为相应的宾州标记序列。y_t 表示第 t 个标记，势函数为：

$$\phi(y_{t,n}, X) = \exp\left(\sum_{k,m} (\lambda_k f_k(y_{t-1}, y_t, X) + \mu_m g_m(y_t, X)) \right),$$

$f_k(y_{t-1}, y_t, X)$、$g_m(y_t, X)$为二值特征函数，例如：前一个标记为 NR、当前标记为 NN 的特征函数 $f_k(y_{t-1}, y_t, X)$为：

$$f_k(y_{t-1}, y_t, X) = \begin{cases} 1 & \text{如果 } y_{t-1} = NR \\ & \text{并且 } y_t = NN \\ 0 & \text{否则} \end{cases}$$

两个特征函数代表两类特征。λ_k、μ_m 为相应的特征权重。

在下一部分实验中我们利用了 Phan 等提供的 FlexCRFs 源码①，训练算法采用 L-BFGS 拟牛顿法。CRF 特征选择是个关键点，我们通过定义特征模板来提取特征。例如：

观察序列 X：上海/ns 浦东/ns 开发/vn 与/cc 法制/n 建设/vn 同步/vi

标记序列 Y：NR NR NN CC NN NN VV

词一级特征模板：－2、－1、0、1、2、－1 ∶0、0 ∶1；(0 为当前位置，±N 表示相对 0 的位置)则相应的特征就为：0 ∶上海(表示当前词为“上海”的条件下标记成 NR)；0 ∶1 上海 ∶浦东(表示当前词“上海”且下一个词“浦东”的条件下标记成 NR)等。

4.3 转换实验

在宾州中文树库 1.0 版上我们进行了切分转换和标记转换实验。在切分转换中共生成 4 300 多条转换规则。在标记转换中我们使用了词一级特征(－2、－1、0、1、2、－1 ∶0、0 ∶1)和词性标记一级特征(－2、－1、0、1、2、－2 ∶－1、－1 ∶0、0 ∶1、1 ∶2、－2 ∶－1 ∶0、－1 ∶0 ∶1、0 ∶1 ∶2)，总共生成 98 400 多个特征；训练迭代了 300 次(每次用时 200 秒左右)，选取开发集 F1 值最好(88.8%)的那组特征权重，转换模型在测试集上 F1 值达到 92.0%。

转换实验会引入一些标记错误。如果只给出单一标记结果，错误标记将会对句法分析产生负面影响。因此，我们可以按照以下两种策略减少这种负面影响：1.标记转换结果生成 n-best，每个词都会有多个候选标记结果，按照相应的频数生成候选标记概率，作为句法分析器的输入；2.将标记转换结果(预测标记)作为一个最可能候选标记加入到我们的词性标记模块(Baseline 系统已经使用该模块)标记候选集中。我们的词性标记模块对词性标记概率采用极大似然估计，数据平滑算法为 Good-Truing。相应的词性标记策略为：

ⅰ. 在训练集中出现过的词，按照可能的标记作为候选标记；

ⅱ. 如果不满足 ⅰ。进行有明显特征的特殊类识别，如：数词、时间词等。

ⅲ. 如果不满足 ⅱ。则取该词的第一个字来代表整个词(字的标记概率在训练时候统计得到)。

ⅳ. 如果不满足 ⅲ。则定义成未登录词，按照所有可能的词性标记作为候选标记。

在第二种策略中，我们以词性标记模块中最高概率的二倍作为预测标记的概率值加入到词性标记候选集中，作为句法分析的输入。

按照词性标记模块标记算法，词性标记主要依赖训练集中的词性标记(历史标记)，第二种策略不但考虑了历史标记也考虑了预测标记，有效增加了句法分析的搜索空间，句法分析就更有可能搜索到更好的句法分析树。所以我们在以下的实验中都按照第二种策略进行。

5 实验与分析

对句法分析进行对比实验，测试集结果如表 1 所示。

Baseline 系统是第 2 节我们讲到的句法分析系统，其输入为宾州树库已切分好的句子，这里我们假定切分标准完全和宾州树库标准一致，通过 4.3 节

① 我们对源码进行了部分改进。

所提到的词性标注模块进行标注，多标注候选结果作为句法分析器的输入。融合模型则按照第 4 节所述进行。

我们可以看到：在测试集上，通过转换模块 F1 值为 92.0%，但在句法分析后 F1 值达到了94.1%。如第 4 节所说，将转换模块的预测标记加入到词性标记候选集合，不但考虑了历史标记也考虑了预测标记，有效增加了句法分析搜索空间，使得句法分析搜索到更好的分析树。句法分析过程也是对词性标记的重选择过程，所以会有效提高词性标记准确率。

表 **1**　句法分析对比实验

模　型	Len≤40 words				Len≤100 words			
	LR	LP	F1	POS	LR	LP	F1	POS
Baseline	80.1%	81.4%	**80.7%**	93.4%	77.8%	79.1%	**78.5%**	93.3%
融合模型	81.7%	83.3%	**82.5%**	94.3%	78.8%	80.8%	**79.8%**	94.1%

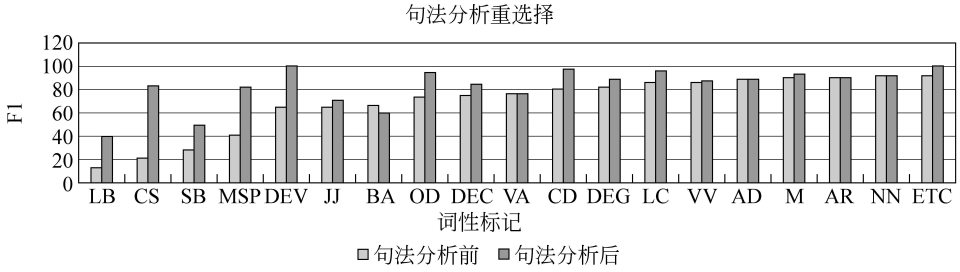
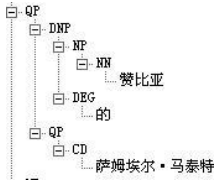


图 5　开发集上句法分析前后词性标记 F1 值提升

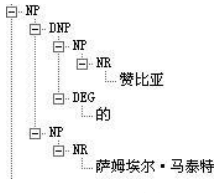
从图 5 我们可以看到，句法分析对词性标记的重选择提高了词性标记准确率。这主要表现在：对转换标记中明显错误修正，例如，将“刘/DER 副主席/NN”修正为“刘/NR 副主席/NN”；对名词、动词句法结构修正，例如，将“专访/VV ：/PU 教练/NN 谈/VV 计划/NN”修正为“专访/NN ：/PU 教练/ NN 谈/VV 计划/NN”。转换模块和句法分析重选择的互补性大大提升了词性标记准确率。

词性标记准确率的提高进一步提高了句法分析的性能。提高点主要表现在：

1. 对命名实体的准确预测，ICTCLAS 可对命名实体识别，对未登录词的命名实体标记比较准确。



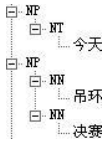
Baseline：错误的标记



融合模型：正确的标记

2. 对动词、名词的正确标记影响到整个句法分析树。

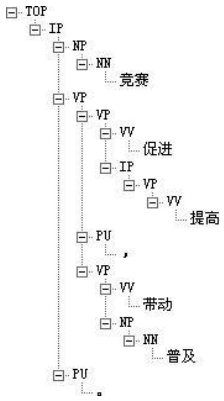
Baseline：错误的分析树



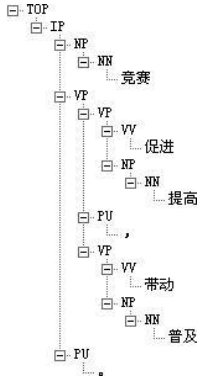
融合模型：正确的分析树

这个例子出错原因是：按照 Baseline 词性标记模块，“吊环”为未登录词，按照第一个字选择候选，“吊”标记成动词 VV 的概率最大，所以会出现错误。转换模块进行标记时在多数情况下可以标记正确。

3. 对并列结构调整。



Baseline：错误句法分析



融合模型：正确句法分析

在测试集上，与 Baseline 系统相比，词性标记 F1 值提高了近 1 个点(93.4%到 94.3%)，句法分析 F1 值也直接提高了 1.8 个点(80.7%到 82.5%)。测试结果表明这种提高在统计上是明显的。

6 比较

表 2 列出了相关中文句法分析器在宾州中文树库 1.0 版测试集上的结果。Bikel and Chiang[17] 使用了两个模型，一个是基于 Collins[8] model 2 的词汇化 PCFG 模型，另一个是统计树粘接语法(Tree-adjoning Grammar)模型，该模型 LR/LP 达到 76.8%/77.8%；Levy and Manning[18] 使用乘子模型将 PCFG 模型和依存模型结合在一起，该模型 LR/LP 达到 79.2%/78.4%；Fung[6] 的整合模型 F1 达到 79.6%；Bikel 博士论文[1] 模拟 Collins 模型，使用了修改过的中心词规则映射表，并且构造了一个词性标记模块对没有出现过的词进行词性标记，该模型 LR/LP 达到 78.0%/81.2%；Chiang and Bikel[19] 用 EM 算法对中心词规则映射表进行优化，该模型 LR/LP 达到 78.8%/81.1%；Wang[4] 采用了基于分类的决策模型，LR/LP 达到 79.2%/81.1%；Jiang 作业论文[3] 将 Collins 模型移植成中文句法分析器，LR/LP 达到 80.1%/82.0%的好成绩；Luo[5] 基于字的最大熵句法分析模型 F1 值达到 81.4%，不过他们使用的训练数据为宾州中文树库 2.0 版本，大约是 1.0 版本的 2.5 倍，所以不具备可比性。与以前的工作相比，我们的融合模型主要是有效利用到了外部资源，包括：语义类信息(知网、词林)、ICTCLAS(中文词法分析器)，词性标记准确率达到了 94.3%，LR/LP 达到 81.7%/83.3%的最好成绩。

表 3 进一步表明了词性标记准确率对句法分析性能的影响。可以看出，当词性标记完全使用 Gold-standard 标记(标记准确率为 100%)时，所有句法分析性能指标 LR/LP 都有显著性提高。我们可以充分利用外部资源提高词性标记的准确率，从而进一步提高句法分析性能。

表 2 句法分析性能比较

模型	Len≤40 words				Len≤100 words			
	LR	LP	F1	POS	LR	LP	F1	POS
Bikel and Chiang[17]	76.8%	77.8%	77.3%	—	73.3%	74.6%	74.0%	—
Levy and Manning[18]	79.2%	78.4%	78.8%	—	—	—	—	—
Fung[6]	80.9%	78.3%	79.6%	—	—	—	—	—
Bikel 博士论文[1]	78.0%	81.2%	79.6%	—	74.4%	78.5%	76.4%	—
Chiang and Bikel[19]	78.8%	81.1%	79.9%	—	75.2%	78.0%	76.6%	—
Wang[4]	79.2%	81.1%	80.1%	92.5%	76.7%	78.4%	77.5%	92.2%
Jiang 作业论文[3]	80.1%	82.0%	81.1%	92.4%	—	—	—	—
我们的系统	81.7%	83.3%	82.5%	94.3%	78.8%	80.8%	79.8%	94.0%

表 3　词性标记准确率对句法分析性能的影响

模　型	LR	LP	F1	POS
Jiang 作业论文[3]	80.1%	82.0%	81.8%	92.4%
	84.5%	88.0%	86.2%	100%
Wang[4]	79.2%	81.1%	79.6%	92.5%
	88.3%	88.1%	88.2%	100%
我们的系统	81.7%	83.3%	82.5%	94.3%
	88.3%	89.5%	88.9%	100%

7　结论

在这篇文章中，我们提出的转换策略有效地将不同切分、标注标准的词法分析和句法分析融合在一起。使中文句法分析器性能有了显著性的提高（F1 值从 80.7%提高到 82.5%）。通过两步转换步骤，将一种词法分析结果转换成另一种词法分析结果。该方法具有很好的扩展性，可以应用到其他语种。此外，文章中提出的多子模型句法分析器，能将现有的子模型结合在一起，提高了句法分析的性能。文章中提出的方法不失为在现有条件下提高句法分析性能的有效途径。

参考文献：

[1] Daniel M. Bikel. On the Parameter Space of Generative Lexicalized Statistical Parsing Models [D]. Ph. D. thesis, 2004, University of Pennsylvania.

[2] Deyi Xiong, Shuanglong Li, Qun Liu, Shouxun Lin, and Yueliang Qian. Parsing the Penn Chinese treebank with semantic knowledge [A]. In: Proceedings of IJCNLP 2005 [C]. 2005. 70-81.

[3] Zhengping Jiang. Statistical Chinese parsing [Z]. Honours thesis, 2004, National University of Singapore, .

[4] Mengqiu Wang, Kenji Sagae, Teruko Mitamura. A Fast, Accurate Deterministic Parser for Chinese[A]. In: proceedings of ACL 2006 [C].

[5] Xiaoqiang Luo. A maximum entropy Chinese character-based parser [A]. In: proceedings of EMNLP [C]. 2003.

[6] Pascale Fung, Grace Ngai, Yongsheng Yang, and Benfeng Chen. A maximum-entropy Chinese parser augmented by transformation-based learning [J]. ACM Transactions on Asian Language Information Processing, 2004, 3(2): 159-168.

[7] 张浩，刘群，白硕等. 结构上下文相关的概率句法分析[A]. 第一届学生计算语言学研讨会论文集[C]. 北京大学，2002. 46-51.

[8] Michael Conllins. Head-Driven Statistical Models for Natural Language Parsing [D]. University of Pennsylvania, 1999.

[9] Franz Josef Och. Minimum error rate training in statistical machine translation [A]. In: Proceedings of the 41st Annual Meeting of the ACL [C]. 2003. 160-167.

[10] 付雷，刘群. 单纯形算法在统计机器翻译 Re-ranking 中的应用[J]. 中文信息学报，2007，21(3)：28-33.

[11] Huaping Zhang, Hongkui Yu, Deyi Xiong and Qun Liu. HHMM-based Chinese Lexical Analyzer ICTCLAS[A]. In: the Second SIGHAN workshop affiliated with 41th ACL [C]. Sapporo Japan: 2003.

[12] 刘群，张华平，俞鸿魁，等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展，2004，41(8)：1421-1429.

[13] 刘群，张华平，张浩. 计算所汉语词性标记集 Version 3.0[Z]. 2004.

[14] Fei Xia. The Segmentation Guidelines for the Penn Chinese Treebank(3.0)[Z]. 2000.

[15] Fei Xia. The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0)[Z]. 2000.

[16] Brill, Eric. A simple rule-based part of speech tagger [A]. In: Proceedings of the Third Conference on Applied Natural Language Processing. ACL [C]. Trento, Italy: 1992.

[17] Daniel M. Bikel and David Chiang. Two statistical parsing models applied to the Chinese Tree-bank [A]. In: proceedings of the Second Chinese Language Processing Workshop, ACL [C]. 2000.

[18] Roger Levy and Christopher D. Manning. Is it harder to parse Chinese, or the Chinese Treebank?[A]. In: proceedings of the 41st Annual Meeting of the Association for Computational Linguistics[C]. 2003.

[19] David Chiang and Daniel M. Bikel. Recovering latent information in treebanks [A]. In: Proceedings of the 19th International Conference on Computational Linguistics [C]. 2002.