

厦门大学第一届中文信息学会句法分析评测系统描述¹

练睿婷 陈毅东 史晓东 蔡科 朱翔 刘智文 刘宁锋
厦门大学智能科学与技术系自然语言处理实验室 厦门 361005
lianlian1022@gmail.com, {ydchen, mandel}@xmu.edu.cn

摘要: 本文描述了厦门大学智能科学与技术系参加第一届中文信息学会句法分析评测的系统。文章分别对参加评测的两项任务的概况以及我们在评测中所使用的系统融合技术做了介绍, 并详细描述了参与测评中的数据配置情况和结果。

关键字: 词性标注 句法分析 系统融合 K-best 融合

The XMU Systems for the 1st China Workshop on Syntactic Parsing

Ruiting Lian, Yidong Chen, Xiaodong Shi, Ke Cai, Xiang Zhu, Zhiwen Liu and Ningfeng Liu
Natural Language Processing Lab, Department of Cognitive Science, Xiamen University, Xiamen 361005
lianlian1022@gmail.com, {ydchen, mandel}@xmu.edu.cn

Abstract: *In this paper, an overview of XMU POS tagging and Syntactic Parsing systems for the 1st China Workshop on Syntactic parsing is given. Brief introduction of two different kinds of POS tagging systems and the system combination technologies are proposed, respectively. It also contains the introduction of two differ models for Parsing system, and K-best combination technologies are proposed. The training data setups and the results are also described as well as some discussions.*

Keyword: *POS tagging, Parsing, System Combination, K-best Combination*

一 引言

本文将对厦门大学智能科学与技术系参加第一届中文信息学会句法分析评测的系统进行描述。本次评测中, 我们参加了汉语词性 (Part-of-speech, POS) 标注处理和句法结构树分析两个项目的评测。其中, 词性标注的评测中, 我们使用了目前效果较好的最大熵模型 (ME) [Adwait Ratnaparkhi, 1998]和条件随机场模型 (CRF) [Lafferty et al., 2001], 并使用系统融合技术融合了基于这两个模型的词性标注结果。句法结构树分析的评测中, 我们也使用了 Kbest 的融合技术[Zhang et al., 2009]对词汇化模型句法分析器 (如 Charniak Parser [Charniak, 2000]和 Stanford Parser[Klein and Manning, 2003]) 的分析结果和非词汇化句法分析器 (如 Berkeley Parser[Petrov and Klein, 2007]) 的分析结果进行了重新打分排序, 从而选择最佳的结果。

文章的第二节将介绍参加评测的词性标注系统情况; 第三节将介绍参加评测的句法结构树分析系统情况; 第四节描述了评测中使用的数据、处理过程、评测结果及讨论; 第五节是相关结论。

¹ 本文的工作得到国家自然科学基金 (批准号: 60573189)、863 高科技计划项目 (批准号: 2006AA01Z139、2006AA010107 及 2006AA010108)、福建省自然科学基金 (批准号: 2006J0043) 和福建省科技重点项目 (批准号: 2006H0038) 的资助。

二 词性标注系统概况

2.1 基于最大熵（ME）模型的词性标注系统

评测中我们采用了最大熵模型进行词性标注，且训练了两个最大熵模型，一个为多标记词模型，一个为未登录词模型。

由训练语料统计出一个字典，字典中的一项为<词语，Label1，Label2，……>，其中的 Label 项列出了该词语在训练语料中的所有出现标记。字典包含训练语料中出现的所有词语。在解码时，当出现的是字典中的单标记词，则直接使用对应的标记。当出现的是字典中的多标记词，则使用多标记词模型估算各个候选标记的概率。当出现的是未登录词，则使用未登录词模型估算各个候选标记的概率，此时的候选标记来自整个标记集。解码过程采用了二元动态规划。

当待标记词是未登录词时，由于候选标记来自整个标记集，造成结果状态数过多，为此我们只保留 M 个最好结果状态，此次测评中，我们选取 M 为 10。

2.1.1 特征选择

多标记词模型选择的特征与未登录词模型选择的特征是不同的。对于多标记词模型，使用如下特征模板进行特征选择：

W _i
W _{i-2}
W _{i-1}
W _{i-2} W _{i-1}
T _{i-2}
T _{i-1}
T _{i-2} T _{i-1}
W _{i+1}
W _{i+2}
W _{i+1} W _{i+2}

其中 W 表示词语，T 表示标记，i 表示当前位置，例如 W_{i-2}W_{i-1} 表示当前词的前两词的组合。对训练语料中的所有词语运用上述特征模板，得到训练多标记词模型的特征文件。

对于未登录词模型，选用在训练语料中出现次数少于或等于 N 次的词的特征作为其训练特征，在此次评测中，我们选取 N 为 5。所选用的特征模板与上述特征模板的不同之处在于未选用 W_i 模板，而是换成了 First 与 last 模板，分别表示该词语的第一个字与最后一个字。

2.2 基于条件随机场（CRF）的词性标注系统

条件随机场（Conditional Random Fields, CRFs）是由 Lafferty 等人于 2001 年提出的一种无向图模型，可以用来标记和切分序列化数据[Lafferty et al., 2001]。它是一个在给定输入序列的条件下计算输出序列的条件概率的无向图模型。最简单也最常用的一类 CRFs 模型是线性链式 CRFs 模型，很适合用于中文分词的机器学习任务。

2.2.1 特征选择

中文词性标注的特征选择主要是要考虑上下文特征，词长特征，未登录词特征等。利用条件随机场进行词性标注的时候主要是考虑了词的上下文特征，考虑到 task1 任务只有 3MB 左右的训练数据却要训练出 70 个 tag 的模型所以采用较少的特征以避免严重的数据稀疏。以下是采用三字窗口具有 6 个特征的特征模板。

$C_n (n = -1, 0, 1)$

$C_n C_{n+1} (n = -1, 0)$

$C_n C_{n+2} (n = -1)$

其中 C 表示待标注的词，下标 n 表示考虑的词与当前的词外置的相对偏移位置。

2.3 系统融合技术

本次评测中，我们还利用了最大熵模型融合 CRF 标注器标注结果，即在最大熵的特征提取中引入了 CRF 标注结果的特征，这样理论上能弥补基于最大熵模型的词性标注系统中的标注偏置问题，而我们的实验也证明了系统融合后词性结果有了一定的改善。训练时，采用如图 1 所示方法：

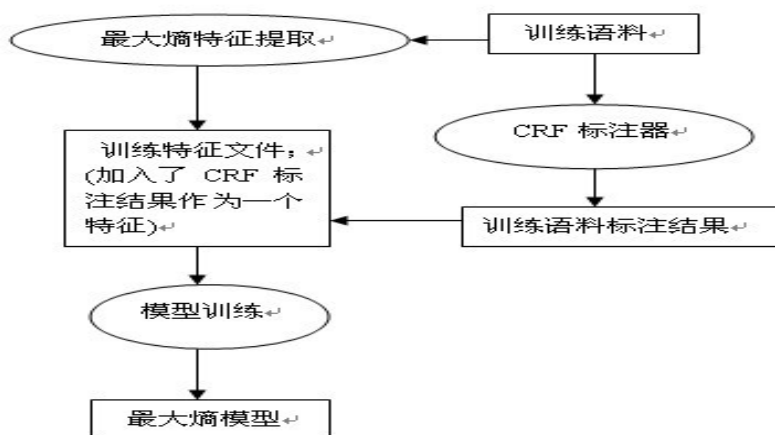


图 1 系统融合的训练

标注时，采用相似的流程，利用了动态规划算法来进行解码，如下图所示：

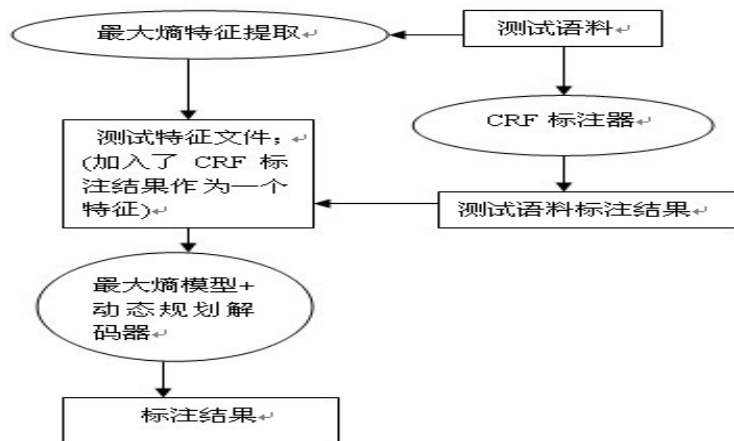


图 2 系统融合的标注

三 句法分析系统概况

统计句法分析在不同语言的结构分析方面已经取得了较好的结果，基于上下文无关文法(PCFG)的句法分析方法一直是该领域研究的主流，但 PCFG 存在的一个重要的问题是语法缺少对词汇的敏感性。因此，对 PCFG 模型的改进可以分为以下两种：引入词汇化信息和扩展非终结符标记。即目前最常见的两大类句法分析模型：词汇化模型[Collins 1997, 1999; Charniak 1997, 2000]和非词汇化模型[Klein and Manning 2003]。词汇化模型中，词汇信息在训练语法规则模型时起主导作用，而非词汇化模型则利用非终结符的潜在信息。两种模型有很好的互补性，因此有必要对他们进行融合。

3.1 词汇化句法分析系统

头驱动模型 (head-driven, 也称为中心词驱动) [Collins 1997, 1999; Charniak 1997, 2000]是最有代表性的一种词汇化模型，即为每个非终结符号都标上一个单词作为其中心词 (也称为 head)。如下图所示。图 3 即为一颗 PCFG 构建出来的分析树，图 4 为通过概率词汇化 CFG 构建出来的分析树。

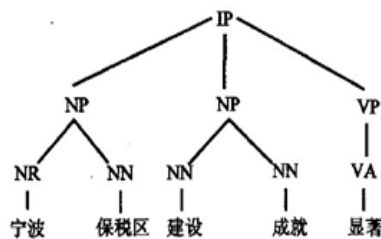


图 3

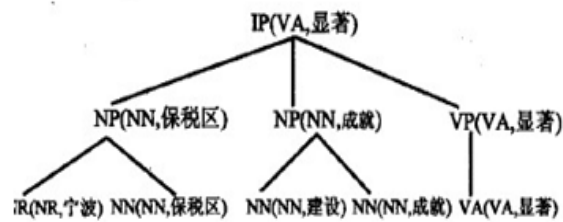


图 4

对文法引入中心词后，每条文法规则的概率计算公式变化为：

$$P(h) \rightarrow L_n(l_n) \dots L_1(l_1) H(h) R_1(r_1) \dots R_m(r_m)$$

其中， H 对应的是中心词短语， h 对应的是中心词。可通过链接规则表达该文法规则的概率计算公式如下：

$$\begin{aligned} & P(L_{n+1}(l_{n+1}) \dots L_1(l_1) H(h) R_1(r_1) \dots R_{m+1}(r_{m+1}) | P(h)) \\ &= P_h(H | P(h)) \times \prod_{i=1, \dots, n+1} P_l(L_i(l_i) | L_1(l_1) \dots L_{i-1}(l_{i-1}), P(h), H) \times \\ & \quad \prod_{j=1, \dots, m+1} P_r(R_j(r_j) | L_1(l_1) \dots L_{n+1}(l_{n+1}), R_1(r_1) \dots R_{j-1}(r_{j-1}), P(h), H) \end{aligned}$$

分析树的概率就是这颗分析树中每条文法规则通过上述公式求得概率的乘积。在本次评测中，我们选择了 Charniak Parser²[Charniak 2000]和 Stanford Parser³[Klein and Manning 2003]为词汇化句法分析系统的代表。其中 Charniak Parser 是目前最好的基于词汇化模型开发出来的句法分析器，该分析器除了引入中心词信息外，还将 pre-terminal 做为一个附加特征引入到文法概率的计算中。Stanford Parser 实现了一个基于 Factored 模型的句法分析器，其主要思想就是把一个词汇化的分析器分解成多个要素 (factor) 句法分析器。在他们的分析器中，即将一个词汇化的模型分解成一个 PCFG 和一个依存模型。

² <ftp://ftp.cs.brown.edu/pub/nlparser/>

³ <http://nlp.stanford.edu/software/lex-parser.shtml>

3.2 非词汇化句法分析系统

通过细化非终结符标记，融入更多的上下文信息来提高句法分析的性能是一种最具代表性的非词汇化方法。Klein [Klein and Manning, 2003]、Matsuzaki [Matsuzaki et al., 2005]、Petrov [Petrov et al., 2006]等人分别通过人工和无监督的方法对非终结符号标记进行了细化 (latent-annotation)，使句法分析器性能得到了很高的提升，接近甚至超过了最好的词汇化方法。简单地说，细化非终结符标记的思想就是将每一个非终结符看成是一个含有一系列潜在变量的非终结符标记。然后通过 EM 算法去自动学习每个潜在变量的概率求解函数，使其在给定的训练数据中概率最大化。以下列二元化的语法规则为例：

$$A \rightarrow B C$$

在非词汇化的模型中，该文法规则被看成一系列如下的规则集合

$$\{A_i \rightarrow B_j C_k | i, j, k \text{ are latent variables}\}$$

在计算树的概率时，首先对树进行二元化（即转化为相应的二叉树），然后用每个非终结符的潜在变量去替换该树中的非终结符，从而可得到一个压缩森林，此时，根节点 (root) 的内部概率 (inside probability) 即为该棵树的概率。

在本次评测中，我们选择了 Berkeley Parser⁴ [Petrov and Klein, 2007] 来作为非词汇化模型的代表进行实验。

3.3 K-best 融合技术

系统融合和分析树的重排序一直是提高句法分析性能的两个重要策略。本次评测中，我们使用了 K-best 融合技术 [Zhang et al., 2009] 将多个系统的分析结果重新计算概率然后重排序，最终输出重排序后的最优结果。

3.3.1 K-best 融合系统框架

K-best 融合技术的主要思想即通过一个线性模型对不同系统的 K-best 分析结果进行重新评分，然后根据新的计算结果进行重排序，从而输出最佳分析树。其流程如图 5 所示。

其中，Charniak Scorer 是指通过 Charniak Parser 中的模型对过滤后的 M 棵树进行重新评分，Berkeley Scorer 是指通过 Berkeley Parser 中的句法分析模型对过滤后的 M 棵树进行重新评分，其评分的计算方法在 3.1 和 3.2 小节中已介绍。Final Scores 是通过 Charniak Scorer 计算出来的概率和 Berkeley Scorer 计算出来的概率加权求和得到的，即

$$\text{Final Scores} = \lambda_1 * \text{Score}_1 + \lambda_2 * \text{Score}_2$$

Score₁ 是指 Berkeley Scorer 计算出来的分析树的概率，Score₂ 是指 Charniak Scorer 计算出来的分析树的概率。 λ_1, λ_2 为相应权值，且 $\lambda_1 + \lambda_2 = 1$ 。

⁴ <http://nlp.cs.berkeley.edu/Main.html>

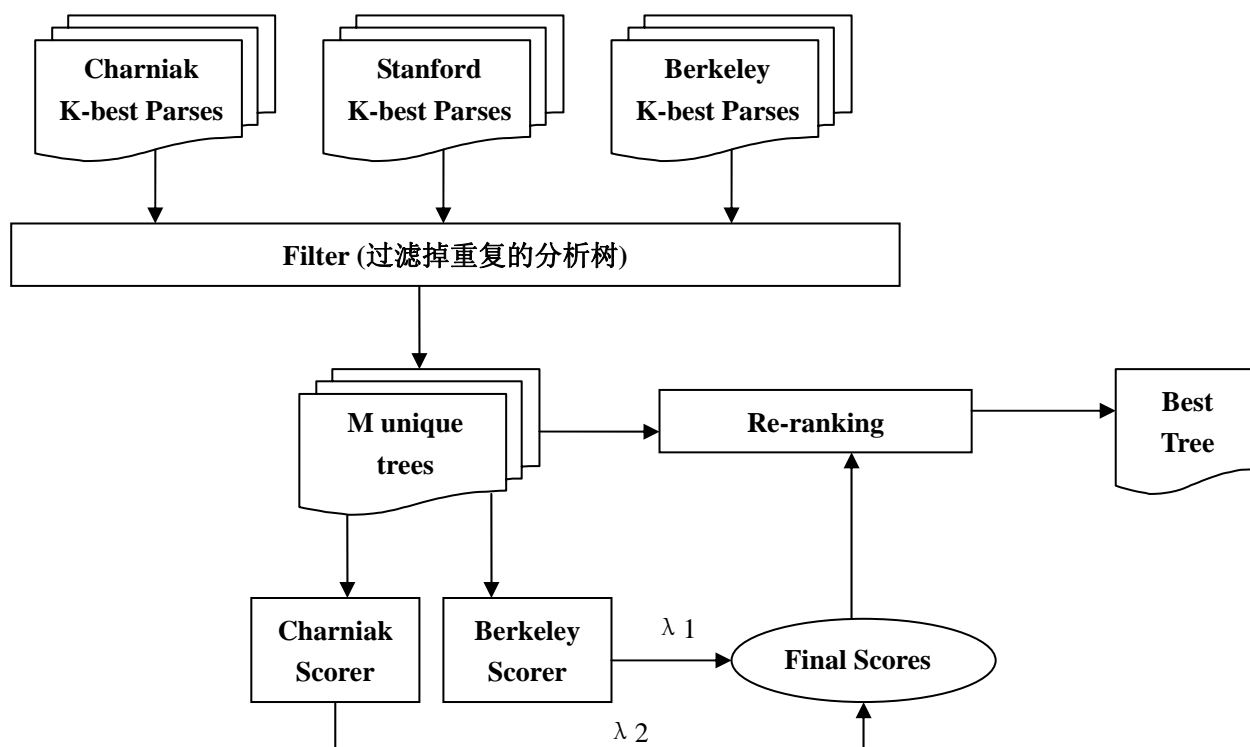


图 5 K-best 融合系统框架

3.3.2 参数估计

我们所使用的技术和[Zhang et al., 2009]所使用的技术类似。采用了最小错误概率准则来调整参数的值（在实验中我们采用最大化 F1 的得分），在开发集上调整最佳参数，同样采用了模拟退火法来进行参数估计。

由于模拟退火法容易陷入局部最优，我们的实验只需要估计一个 0 到 1 之间的参数 λ_1 ($\lambda_2=1-\lambda_1$)。因此我们还使用了最简单有效的穷举法来进行参数估计，本次评测中， λ_1 ， λ_2 的最优值分别为：0.72 和 0.28

3.4 中心成分的确

对于中心成分的确，我们采用了基于规则的方法，将要确定中心词成分的句法规则分成以下两种情况：

- 1、该句法规则出现在训练集中：
如果在训练集中出现该句法规则，则按照该句法规则标记中心词。
- 2、该句法规则没有出现在训练集中：

这种情况下，我们采用文献[Fei Xia, 1999]提供的决策表来确定每一条句法规则的中心成分。决策表中的每个条目都是由一个 3 元组构成的，形式如下：

<Parent Non-terminal: Direction, Priority List>

其中，Parent Non-terminal 表示句法规则左手侧的短语符号；Direction 为方向，取值为 Left 时指明是从左往右搜索，取值为 Right 时指明是从右往左搜索；Priority List 为一个由一系列非终结符号组成的优先表。我们的具体做法如下：

1) 根据训练集抽取每个非终结符的中心词优先队列和左优先还是右优先规则。

2) 对某个规则如: $np \rightarrow tp\ uJDE\ vN$, 有 np 优先队列: $n\ np\ vN\ nP\ vp\ nR$, 并且为右优先, 则按优先队列从左到右求中心词, $tp\ uJDE\ vN$ 从右到左。在此例中先对 n 从右到左搜索规则, 没有搜索到; 接着搜索 np , 也没有; 再搜索 vN 出现在 $tp\ uJDE\ vN$ 中, 则 vN 为其中心词; 以此类推。

四 评测数据与结果

4.1 词性标注评测数据与结果

我们在本次词性标注评测中采用的训练数据为评测小组提供的清华树库的数据, 未使用其他数据。实验中, 我们采用了张乐博士的最大熵工具包⁵ 和CRF++工具包⁶, 其中最大熵模型的训练采用L-BFGS算法迭代100次。本次实验使用的标记总共为70个, 训练数据为大约3MB的已标注数据, 测试数据为大小为495KB的3751个语句。

我们提交了两个结果(ME和ME+CRF), 其中ME为基于最大熵模型的词性标注系统, ME+CRF指的是通过2.3小节介绍的融合技术融合了基于最大熵模型的词性标注系统和基于条件随机场的词性标注系统。

	overall F1	small-class average F1
ME (16_a)	92.72	78.45
ME+CRF (16_b)	92.30	78.07

表1 词性标注评测的实验结果

从上述实验结果中可以看出, 系统融合后的结果在性能上还是有一定的提高。这也可以说明引入CRF的词性标注结果的特征, 能在一定程度上弥补最大熵模型造成的标记偏置问题。

4.2 句法结构树分析评测数据与结果

我们在本次词性标注评测中采用的数据为评测小组提供的清华树库的数据, 未使用其他数据进行训练。其中训练集包含60245个片段, 开发集包括1674个片段, 测试集包括16210个片段。

我们在此项任务中提交了4个结果, 分别为Charniak Parser、Berkeley Parser、Combination 1(使用穷举法来进行参数估计)和Combination 2(使用模拟退火来进行参数估计), 其中系统融合均使用了Charniak Parser、Berkeley Parser和Stanford Parser三种分析器的结果。我们分别对K-best中k=1和k=5做了实验, 发现k=1的效果反而稍好, 因此采用了k=1。

	without-head match F1	partial-head match F1	complete-head match F1
Berkeley Parser (16_a)	87.37	76.02	70.58
Charniak Parser (16_b)	83.87		
Combination1 (16_c)	87.65		
Combination 2 (16_d)	87.61		

表2 句法分析评测的实验结果

⁵ http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

⁶ <http://crfpp.sourceforge.net/>

说明：在提交数据的时候我们弄错了 charniak parser 的数据，该表中列出的实验结果为答案公布后重新实验所得。

从上述实验结果可看出，K-Best 融合后的结果在性能上有一定的提高，但不是很高，这可能是因为在很小的开发集上估计出来的参数不一定适合较大的测试集，可能还和要融合的系统之间性能差异大有关；另外，我们还对 K=5 进行了实验，发现 without-head match F1=85.84，这可能因为噪声大引起的，我们将在下一步工作中找出 K 的最佳值，并引入新的特征进行重新评分。

五 结论

本文对厦门大学智能科学与技术系参加第一届中文信息学会句法分析评测的系统进行描述。本次评测中，我们参加了包括词性标注和句法结构树分析等两个项目的评测。词性标注方面，我们将进一步研究系统的融合技术，并希望将句法分析引入到词性标注过程中以提高其性能。句法分析方面，我们将进一步改善 K-best 融合各种句法分析器的实验，还将在重新评分的线性模型中添加新的特征，并进一步对中心词成分的确定进行研究。

参考文献：

- [1] Adwait Ratnaparkhi, 1998, *Maximum Entropy Models for Natural Language Ambiguity Resolution*, Phd Thesis, University of Pennsylvania
- [2] Michael Collins.1997.Three generative, lexicalised models for statistical parsing . ACL-97, pages 16-23.
- [3] Michael Collins.1999. Head-driven statistical models for natural language parsing . Doctoral Dissertation, Dept. of Computer and Information Science, University of Pennsylvania, Philadelphia 1999.
- [4] Eugene Charniak .1997. Statistical parsing with a context-free grammar statistics. AAAI-97, pages 598-603.
- [5] Eugene Charniak .2000. A maximum-entropy-inspired parser. NAACL-2000
- [6] John Lafferty, Andrew McCallum, Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the 18th International Conference on Machine Learning [C] San Francisco, pages 282—289.
- [7] Fei Xia, 1999, Automatic Grammar Generation from Two Different Perspectives, PhD thesis, University of Pennsylvania
- [8] S. Petrov et al., 2006, Learning Accurate, Compact, and Interpretable Tree Annotation, ACL'06
- [9] T. Matsuzaki, Y. Miyao, and Tsujii, 2005, Probabilistic CFG with latent annotations, ACL'05, pp.75-82
- [10] D. Klein, C. Manning, 2003, Accurate Unlexicalized Parsing, ACL'03, pp 439-446
- [11] Hui Zhang, Min Zhang, Chew Lim Tan, Haizhou Li, 2009, K-Best Combination of Syntactic Parsers, In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1552–1560