

一种结合汉字信息的 LPCFG 汉语句法分析器

徐文智

北京邮电大学智能科学与技术中心

earl808@gmail.com

王小捷

北京邮电大学智能科学与技术中心

xjwang@bupt.edu.cn

摘要

本文提出了一种新的基于词汇化 PCFG 模型的句法分析模型。该方法能有效利用汉语中的字信息，在一定程度上缓解了词汇化 PCFG 模型中词信息的数据稀疏问题。本文进而探讨了能明显提高句法分析性能的重要特征。同时我们还对原始树库中的标注进行一些必要的修改，以减少标注歧义。基于 TCT 树库所进行的实验结果表明，字信息以及其他特征能带来句法分析性能的提升。

关键词 句法分析 词汇化的 PCFG 模型 最大熵模型 字信息 TCT 树库

1. 介绍

句法分析是自然语言处理中非常重要的基础任务。随着宾州中文树库的发布，很多不同的中文句法分析模型被提出来。(Bikel,2000)采用中心驱动模型对汉语进行句法分析。(Levy,2003)详细分析了中文与英文语法上的差别所带来的困难。(Wang,2006)利用 shift-reduce 决策的方法，极大的提高了句法分析的解码速度。这些研究大部分沿用英语分析的方式，采用基于词的模型。

然而中文与英文存在重要的区别：英文表达的基础单位是词，而在中文中还存在着更小的单位：字。由于汉语词切分的困难，也有很多研究直接探索基于字的模型。(Luo,2003)直接对没有分词的句子进行句法分析，在句法分析中融入了分词的过程，但句法树的内部结构中并没有利用字信息；(Zhao,2009)的依存分析完全舍弃了词的概念，是字与字间的依存树。

我们试图综合词信息与字信息来得到更为有效的句法分析性能。尽管分词的标准很难得以统一，但是不同的分词标准大部分都是在一个短语的内部，对句法树的内部结构不会有太大影响，所以我们还是选择词作为句法分析的基础单位。基于中心标记的句法分析中的词能够极大的改进句法分析的性能(Collins,1999)，但是中心词的依赖数据非常稀疏。而在中文中，我们注意到大量的意义相近的词共用同一个字，比如：科学家，历史学家...这些词在一定层次上，意义是一样的。因此词的信息可以通过综合字的信息来一定程度上缓解词的数据稀疏问题。

本文的句法分析实验所采用的语料是清华大学新开发的 TCT 树库。这个树库主要以二叉树为主，加上少量的多叉树和单叉树，因此实验中，先把所有的子树转化为二叉树。句法树的表示模型沿用(Collins,1999)中的词汇化的 PCFG 模型以利用词与字的信息。和(Charniak,2000)一样，利用最大熵模型来估计句法树的概率。最后的解码算法采用简单的 CYK 算法。

接下来一节，详细介绍利用字信息的句法分析模型和特征选择，以及基于对语料的研究，所做的一些改动。第三节是实验结果及讨论。第四节是本文的结论和展望。

2. 字信息句法分析器

2.1 模型介绍

词汇化的 PCFG 模型是我们工作的基础，我们的句法分析的表示方法和这个模型是一致的，在这做个简单的介绍。

假设 P 是子树的标记， $L_1...L_n$ 是中心节点左边的修饰成分， H 是中心成分， $R_1...R_n$ 是

右边的修饰成分。则词汇化的 PCFG 模型的推导规则可以写成：

$$P(hw, ht) \rightarrow L_n(lw_n, lt_n) \dots L_1(lw_1, lt_1) H(hw, ht) R_n(rw_1, rt_1) \dots R_n(rw_n, rt_n) \quad (1)$$

其中, (hw, ht) 表示中心节点的中心词及词性, $(lw_1, lt_1) \dots (lw_n, lt_n)$ 以及 $(rw_1, rt_1) \dots (rw_n, rt_n)$ 分别表示左右修饰成分的中心词及词性, 而父节点 P 的中心词是中心成分 H 的中心词。为了计算这个推导规则的概率, 把他分成三个部分: 先由 P 生成中心成分 H , 再由 P, H 分别生成左边和右边的修饰成分。这样语法规则的概率为:

$$\Pr_h(H | P, hw, ht) * \prod_i \Pr_l(L_i(lw_i, lt_i) | P, H, hw, ht) * (\prod_j \Pr_r(R_j(hw_j, ht_j) | P, H, hw, ht)) \quad (2)$$

实际上, (Collins, 1999) 认为上式的语法规则再补充上 (P, hw, ht) 的先验概率, 才是这个语法规则的概率。所以这个概率最后的形式为:

$$\Pr_{prior}(P, hw, ht) * \Pr_h(H | P, hw, ht) * \prod_i \Pr_l(L_i(lw_i, lt_i) | P, H, hw, ht) * (\prod_j \Pr_r(R_j(hw_j, ht_j) | P, H, hw, ht)) \quad (3)$$

我们可以看出来, 如果按照修饰成分的马尔可夫假设, 上式概率其实等价于这个语法规则的联合概率, 即:

$$\Pr_{rule}(P(hw, ht) \rightarrow L_n(lw_n, lt_n) \dots L_1(lw_1, lt_1) H(hw, ht) R_n(rw_1, rt_1) \dots R_n(rw_n, rt_n)) \quad (4)$$

这个概率所求的是这个推导规则的联合概率。本文虽然也是采用词汇化的 PCFG 模型, 但是计算推导规则的概率方法是计算条件概率。在二叉树的情况下, 推导规则可以简单的表示为:

$$P(hw, ht) \rightarrow H(hw, ht) R(rw, rt) \quad (5)$$

$$P(hw, ht) \rightarrow L(lw, lt) H(hw, ht) \quad (6)$$

(5)和(6)分别表示中心成分在左边和右边。推导规则的条件概率是基于中心和修饰成分的中心词和词性的概率, 即:

$$\Pr(P, R, H | hw, ht, rw, rt) \quad (7)$$

$$\Pr(P, L, H | hw, ht, lw, lt) \quad (8)$$

为了计算这个推导规则概率, 我们按马尔可夫的思想把(7)、(8)式分别分解为(9)、(10)式中三个概率连乘的形式:

$$\Pr(P, R, H | hw, ht, rw, rt) = \Pr_d(P - DIR | hw, ht, rw, rt) * \Pr_h(H | P, hw, ht) * \Pr_m(R - DIR | P, H, hw, ht) \quad (9)$$

$$\Pr(P, L, H | hw, ht, lw, lt) = \Pr_d(P - DIR | hw, ht, lw, lt) * \Pr_h(H | P, hw, ht) * \Pr_m(L - DIR | P, H, hw, ht) \quad (10)$$

其中 $P-DIR$ 主要是为了区分中心成分的位置, $DIR=LEFT/RIGHT$ 。 $L-DIR$ 和 $R-DIR$ 中的 DIR 同理用于区分修饰成分的位置。

(9)、(10)式的概率的计算过程体现了如下的生成过程。首先由子成分的词生成父节点以及中心成分的位置 (即第一个概率), 我们定义这个概率为词依赖概率 \Pr_d , 如果两个词经常一起出现, 那么这个概率会很大。如果两个词 (以及词里面的字) 没有一起出现, 那他们的概率约等于 $1/|Y|$, $|Y|$ 为词依赖概率的预测目标的数目。然后按照第二个概率 (中心成分概率 \Pr_h) 生成中心成分, 其中我们认为修饰成分的中心词和词性对中心成分的确定没有提供信息, 因此忽略了。第三个概率生成修饰成分, 定义为修饰成分概率 \Pr_m , 衡量的是修饰成分与父节点和中心成分间的依赖性, 因此我们也忽略了修饰成分的中心词和词性的影响。

举个例子，假设有一个子树如图所示：

对于 $vp \rightarrow v \ np$ 这条规则， vp 的中心成分是左边的 v ， np 的中心成分是右边的 n ，所以 vp 的中心词为“组织”， np 的中心词是“专家”。那么词汇化的规则为 $vp(\text{组织},v) \rightarrow v(\text{组织},v) \ np(\text{专家},n)$ 。对于这个推导规则的概率可以表示为：

$$\Pr_d(vp\text{-LEFT}|\text{组织},v,\text{专家},n) * \Pr_h(v | vp,\text{组织},v) * \Pr_m(np\text{-RIGHT} | vp, v, \text{组织},v) \quad (11)$$

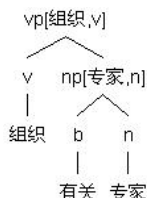


图1 推导规则的句法树

2.2 概率模型以及特征集合

借鉴(Charniak,2000)的方法，我们用最大熵模型来计算概率。最大熵模型通过最大化在一定条件下的概率分布的熵来进行参数估计，这些条件让我们这些参数能体现出训练数据的特点。利用特征函数的方法，最大熵能很灵活的利用各种特征，这些特征往往是提高模型效果的关键。最大熵模型已经成功应用在很多任务中，比如前面提到的基于字的句法分析，分词和 POS 结合起来的最大熵模型等。在我们的实验室中，用的是 Zhang Le 的最大熵工具包，这个工具包利用 L-BFGS 作为参数估计的算法。具体的模型和程序说明请参考(Berger,1996; Zhang,2004)。

我们的特征主要分为四种：基本特征、字特征、上下文特征和上下文与字重叠特征。基本特征主要是传统的词汇化 PCFG 中的特征，由成分中心词、词性及其标记组成；字特征我们只是提取了词的第一个字和最后一个字，对于单字词的第一个字和第二个字就是当前的单字词；上下文特征我们简单的选取了当前推导规则的前面和后面的词的词性，这类特征很好的利用了当前推导规则之外的信息而未增加句法分析解码的复杂度。最后，上下文与字重叠特征，是上下文信息与字信息的重叠特征。具体的特征模板及样例表示如表 1 所示。

注意到在利用单字信息时，单字都是与字所在的词的词性一起作为联合特征。主要是考虑到单字在不同的词性的词中，意义的差别是非常大的。比如，“爱”在动词“爱护”与在名词“爱情”中的意义是不一样的，当然这里所说的意义体现在词的依赖关系上。“爱护”一般会与物品名词相互依赖，而“爱情”则会与形容词或表示人的名字搭配在一起。对于非二叉树的表示方式，在计算左边成分或右边成分的概率时，经常引入一个特征用来表明当前修饰成分和中心成分间是否已经有修饰成分(Collins,1999)。但在二叉树表示中，当前修饰成分与中心成分间是必然没有修饰成分的。由于二叉树的表示方法一般遵循 Chomsky 的 X-bar 理论，因此我们可以修改中心成分的标记来得到这些在二叉树表示中的非局部信息。比如非二叉树推导规则 $vp_1 \rightarrow pp \ d \ vp_2$ ，对应于 $vp_3 \rightarrow pp \ vp_4$ 和 $vp_4 \rightarrow d \ vp_5$ 这两条推导规则，因此在非二叉树表示方式中，计算修饰成分 pp 的概率时， pp 与中心成分 vp_2 之间有修饰成分 d 的。在二叉树表示中， pp 与中心成分 vp_4 间无法体现出这个信息，所以在计算修饰成分概率时，我们把 vp_4 的标记改为 $vp\text{-RIGHT}$ ，表示这个 vp_4 有一个左边的修饰成分，这样就达到了非二叉树中的效果。

	词依赖概率 (vp-LEFT)		中心成分概率 (v)		修饰成分概率 (np-RIGHT)	
基本特征	lw rw	组织 专家	p	vp	p h	vp v
	lw lt rw rt	组织 v 专家 n	p hw	vp 组织	p h hw	vp v 组织
			p hw ht	vp 组织 v	p h hw ht	vp v 组织 v
			p ht	vp v	p h ht	vp v v
字特征	lw frc rt	组织 专 n	p fhc ht	vp 组 v	p h fhc ht	vp v 组 v
	其他词与字组合	...	p lhc ht	vp 织 v	p h lhc ht	vp v 织 v
	frc lt frc rt	组 v 专 n				
	其他字与字组合	...				
上下文特征	lw rw pt1 at1	组织 专家 n v	p pt1	vp n	p h pt1	vp v n
	lw lt rw rt pt1 at1	组织 v 专家 n n v	p pt1 pt2	vp n p	p h pt1 pt2	vp v n p
			p at1	vp v	p h at1	vp v v
			p at1 at2	vp v v	p h at1 at2	vp v v v
			p pt1 at1	vp n v	p h pt1 at1	vp v n v
			p hw pt1 at1	vp 组织 v n	p h hw pt1 at1	vp v 组织 v n
p ht pt1 at1	vp v n v	p h ht pt1 at1	vp v v n v			
上下文与字重叠	lw frc rt pt1 at1	组织 专 n n v	p fhc ht pt1 at1	vp 组 v n v	p h fhc ht pt1 at1	vp v 组 v n v
	其他词与字组合	...	p lhc ht pt1 at1	vp 织 v n v	p h lhc ht pt1 at1	vp v 织 v n v
	frc lt frc rt pt1 at1	组 v 专 n n v				
	其他字与字组合	...				
符号解释						
frc lrc	右边中心词的第一个和最一个字					
pt# at#	前面和后面的词性，数字表示向前话或向后的位置					
frc llc	左边中心词的第一个和最一个字					
fhc lhc	中心词的第一个词和最后一个词					

表1 特征模板以及符号解释。示例的推导规则为 vp(组织,v) -> v(组织,v) np(专家,n)的示例，完整的句子是：
委员会/n 由/p 农业部/n 组织/v 有关/b 专家/n 组建/v 完成/v

2.3 对标注系统的修改以及中心成分的处理

(Klein,2003)详细讨论了对标注系统里面的词性进行分割，取得了明显的效果。本文没有对词性进行分割，而是对子树的标记进行了分割。进行标记的分割主要原因是因为标注系统没有充分地考虑到不同情况，而把这些不同的情况都对待为同一种标记，造成了歧义。由于本次实验所采用的语料是基于二叉树的，二叉树虽然让规则的数目基本变成了一个封闭的集

合,但也带来了更强的独立假设,因此进行标记分割能够让子树的节点表达更多的内部节点信息(Jonson,1998)。和词汇化的 PCFG 模型的出发点一样,这也是一种把非局部的特征变成局部特征的方法。我们主要考虑的三种情况下的修改,依次如下:

首先,是对动词短语 vp 的分割。动词短语可以分为两种情况,动词前面有修饰成分(比如副词)、已经有宾语以及并列的动词或动词加助词之类的形式。前两种情况的动词短语不能在后面跟宾语(有些双宾动词可以继续加宾语,我们的语料把这部分动词与别的动词分别开来了),而有些动词短语可以跟宾语(也许实际上没有宾语)。如果没有对这些情况加以区分,很容易造成句法分析测试的时候把单个的宾语分割成双宾语,如图所示。我们把第三种情况的动词短语的标记改为 vb,表示后面可以接宾语,这样可以很好的把这两种情况加以区分。我们举个简单的例子说明修改的意义。假设把宾语限制为名词短语 np,语料中 vp->vp np 这样的情况出现了 5284 次,进行 vp 分割后,这样的情况只有 166 次。

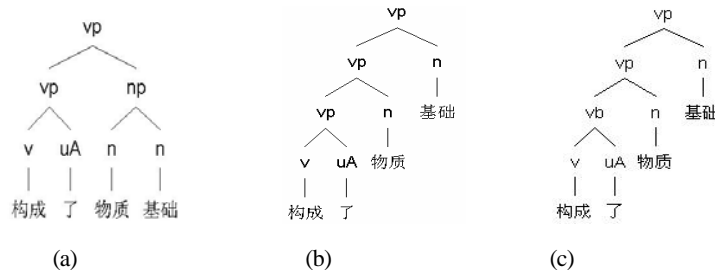


图2 图中的名词或名词短语我们称为宾语, (a)是正确的句法树 (b)一种错误的情况, vp 后接宾语后继续后根宾语 (c) 当进行 vb 修改后, vp 后面很少跟宾语, 这种错误情况概率很小

其次,是对名词短语 np 的修改。我们注意到,名词短语如果由一个非名词修饰成分加名词或名词短语构成,一般这都是最终的名词短语,很少作为另一个非并列名词短语的子成分。因此我们把由非名词修饰成分构成的名词短语修改为 nm。从语料的统计中,我们可以发现, np->np n 出现了 4502 次,而 np->nm n 只有 826 次。

最后我们是对介词短语的中心词的位置进行修改。语料中介词短语的中心词是介词后面的部分,而我们发现介词短语的语法行为和介词本身是非常相关的。以介词“以”和“对”为例,两个介词分别出现了 755 次和 1300 次。介词“以”所构成的介词短语有 745 次修饰动词短语,占 98.7%;而介词“对”只有 744 次,只占 57.2%,有 452 次用于修饰名词短语。

在语料中,并列成分都可以作为短语的中心语,为了处理的方便我们把这类情况的中心成分都处理成最后一个并列成分。比如 np->n cC n,中间的 cC 是并列连词,我们认为这个子树的中心成分为最后的那个成分 n。

3. 实验结果及分析

我们的实验是基于 CIPS-ParsEval-2009 评测会议所提供的 TCT 语料库中的标准训练测试数据进行的。训练和测试时,忽略了其中句子长度为 1 的句子(原始为 16210 个句子,删除后句子数为 7939)。测试时所输入的句子都是附带词性标注的。

本次评测会议我们所提供的数据结果是基于另一个模型的(由于篇幅原因我在此只做简单说明,此论文模型提交结果后才想到)。老模型与新模型的区别在于对于推导规则概

率的衡量。老模型的概率模型为：

$$Pr_{rule} = \Pr(P - DIR | L(lw, lt), R(rw, rt)) \quad (12)$$

老模型之所以这样构造是借鉴线性的马尔可夫模型。但是这个在句法分析中会失败，因为线性的马尔可夫的结构是确定的，但是句法分析中，句法树的树状结构是不定的，需要我们能学习出来。

我们使用国际上通用的评测标准。实验结果如表 2 所示，第一行结果是老模型的结果，后面的 8 行是新模型不同的特征的结果。表中评测标准的意义分别为：LR、LP 和 F 分别是标记的准确率、召回率和前两者的综合衡量。CB 表示每个句子平均交叉的子树（用括号表示），0CB 和 2CB 分别表示没有交叉子树以及少于 2 个交叉子树的句子百分比。同时注意到老模型 79.76 结果比网上所公布的结果高了一点 0.03%，主要是因为我所删除的句子长度为 1 的句子有很少一部分进行了标记。

从结果可以看出，在基础模型上加上字特征带来了 F 值 1.75 的提升，在加入上下文特征的基础上引入字特征（包括和上下文特征的重叠特征），提升了 0.8。这证明字信息能极大提升我们的模型在基础特征上的效果。而在引入上下文特征后（表中五行与第七行对比）字信息仍然能提升我们的句法分析性能，说明字信息能部分解决引入上下文特征后所不能解决的句法分析的歧义。

模型及特征	LR	LP	F	CB	0CB	2CB
老模型	80.03	79.50	79.76	1.30	56.12	81.76
新模型基础特征	80.19	79.61	79.90	1.20	56.10	83.49
新模型+字	81.91	81.38	81.65	1.10	58.34	84.95
新模型+上下文	85.53	85.34	85.44	0.83	65.62	88.86
新模型+字+上下文	86.17	85.94	86.06	0.80	66.61	89.62
新模型+字+上下文+重叠	86.34	86.13	86.24	0.79	66.65	89.81
新模型+字+上下文+重叠+并列词依赖	86.47	86.26	86.37	0.78	66.73	89.87
新模型+字+上下文+重叠+并列词依赖+语料修改	87.03	86.77	86.90	0.75	67.06	90.36
新模型+字+上下文+重叠+并列词依赖+语料修改中心位置	87.20	86.94	87.07	0.74	67.43	90.40

表 2 不同特征的句法分析结果，没有限制句子长度

字信息之所以有用，是因为字信息起到了缓解数据稀疏问题的作用。字信息缓解数据稀疏问题的能力可以从(Kang, 2005)的工作看出。Kang 对 5 万的双音节的词进行统计，对于构词的意义变化主要分成了 8 类：

- (1) A+B=A=B (2) A+B=A (3) A+B=B (4) A+B=C
 (5) A+B=A+B (6) A+B=A+B+D (7) A+B=A+D (8) A+B=D+B

其中 A、B 代表构成词的前后两个字的字义，C 表示一个完全新的词义，D 表示一个附加的意义。等号的后面表示所构成词的词义，其“+”表示意义的混合。比如 A+D 表示新的词的词义保留了 A 的意思，同时附加了新的意义 D。各类的分布情况如下表 3 所示。从表中我们可以看出，与构词的字的意义完全无关的情况 4 只占了 8.02%。数据表明字义与词义有密切的关系。但实验中字信息并没有带来十分显著的作用。这也印证了表 3 的数据：有些词是前面的字可以代表词义，而有些是后面的字；有些是两个字的混合意思，有些和

字义完全无关。

类型	1	2	3	4	5	6	7	8
合成词数量	4035	1031	297	4201	14455	23562	2780	1886
比例(%)	7.71	1.97	0.57	8.02	27.60	44.99	5.31	3.60

表3 词义与字义的关系统计表

在本方法中，上下文对句法分析的性能很有帮助。与(Rantnaparkhi,1999) 和(Wang,2006)所采用的线性方法，或者(Collins,2000)的重排序方法不同，都能很好地利用子树的前后子树标记。CYK 解码算法限制我们利用子树标记，但是我们可以很方便利用子树前后的词性标注来提高效果而不增加解码的复杂度。

通常情况下，子树的中心词只有一个。但是我们注意到，对于并列的子树 A，其孩子成分的中心词都是等价的，因此把并列的词语都作为当前子树的中心词。当 A 作为另一个子树 B 的成分是，所有的并列中心词都可以与另一个成分 C 的中心词依赖。当这个并列的成分 A 仍然作为 B 的中心成分，B 的中心词也继承了 A 的并列中心词。这样我们可以提取出更多的词依赖关系。比如，A 有中心词“建立”和“完善”，C 的中心词为“制度”，那么我们认为“建立”、“完善”都与“制度”依赖。同时，A 又作为父节点 B 的中心成分，那么 B 的中心词也为“建立”和“完善”。从实验结果可以看出这项措施带来了 F 值 0.17 的提升。

语料的修改明显地带来了句法分析性能的提高（评测结果时，这些修改我们会还原回来），不过对于语料的修改更需要对标注系统和语言学的理解，这方面的工作我们只是做了一些小的尝试。相信随着语料的成熟，必能减少句法分析的歧义，带来更好的性能。

4. 结论以及未来工作

本文提出了一种新的词汇化的 PCFG 模型，能综合利用中文的词与字信息进行消歧。实验结果证明了字信息能提高句法分析的性能。同时这个模型利用一些上下文和非局部特征，极大的提高了句法分析的准确率。

在未来的工作中，我们的改进工作主要有以下几个方面。首先，CYK 算法的复杂度过高，我们应寻找更加快速的解码算法。其次，在提高改进解码方法的基础上，引入更多的非局部特征。最后，我们需要对中文进行深入研究，特别是中文的词义，构词法的探索，以提高中文句法分析的性能。

参考文献

- [1] A.L. Beger, S. A. D Pietra. And V.J.D Pietra. A maximum entropy approach to natural language processing. Computational Linguistics 22 1(1996), 39-71
- [2] Daniel M. Bikel and David Chiang. 2000. Two statistical parsing models applied to the Chinese Treebank. In Proceedings of the Second Chinese Language Processing Workshop, In Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2000.
- [3] E. Charniak. 2000. A maximum-entropy-inspired parser. Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics. Seattle, WA.

- [4] M. Collins. 1999. Head-Driven Statistical Models for Natural Language Parsing. University of Pennsylvania, Ph.D. Dissertation, 1999.
- [5] M. Collins. 2000. Discriminative reranking for natural language parsing. In Proceedings of ICML, pages 175–182.
- [6] Shiyong Kang, Xiaoxing Xu, Maosong Sun. 2005. The Research on the Modern Chinese Semantic Word-Formation. Journal of Chinese Language and Computing 15 (2): (103-112)
- [7] D. Klein and C. Manning. 2003. Accurate unlexicalized parsing. In Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2003, 423–430.
- [8] Roger Levy and Christopher D. Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2003.
- [9] X Q Luo. 2003. A maximum entropy Chinese character-based parser. In Proceeding of the Conference on Empirical Methods in Natural Language Processing.2003, 192-199
- [10] M. Johnson. 1998. PCFG models of linguistic tree representations. Computational Linguistics, 24:613-632.
- [11] Adwait Ratnaparkhi. 1999. Learning to parse natural language with maximum entropy models. Machine Learning, 34(1-3):151–175.
- [12] M.W. Wang, K. Sagae, and T. Mitamura. 2006. A Fast, Accurate Deterministic Parser for Chinese. In Proceedings of COLING/ACL.
- [13] L. Zhang, 2004. Maximum Entropy Modeling Toolkit for Python and C++. Reference Manual.
- [14] H Zhao. 2009. Character-Level Dependencies in Chinese: Usefulness and Learning. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.2009, 879-887

A Chinese LPCFG Parser with Combining Character Information

Wenzhi Xu

Center for Intelligence Science and Technology Research
Beijing University of Posts and Telecommunications
earl808@gmail.com

Xiaojie Wang

Center for Intelligence Science and Technology Research
Beijing University of Posts and Telecommunications
xjwang@bupt.edu.cn

Abstract

In this paper, we propose a new probabilistic model based on the lexical PCFG model. This model can easily utilize the Chinese character information to solve the lexical information sparse problem in lexical PCFG model. Moreover, we discuss some important features which can improve the performance of parsing. Meanwhile we modify the original corpus from the view of statistics to reduce the ambiguity of label system. Final experiment demonstrate the character information and our modification can bring the improvement of parsing performance.

KeyWords: Parse Lexical PCFG Maxent Entropy Model Character Information TCT tree bank