统计句法分析建模中基于信息论的特征类型分析

穗志方[*] 赵军[**] 俞士汶[*]

* 北京大学 计算机科学与技术系 计算语言学研究所 北京 100871

** 香港科技大学 计算机科学系 人类语言技术中心 香港

摘要：

统计句法分析利用概率评价模型评价每棵候选句法树存在的可能性，选择概率值最高的候选句法树作为最终的句法分析结果。因此，统计句法分析的核心是一个概率评价模型，而各种概率评价模型的本质区别主要在于它们分别是根据上下文中的哪些特征来赋予句法树概率的。

在统计句法分析研究领域，虽然已经提出了大量的概率评价模型，然而，不同的模型用到了不同类型的特征。如何评价这些特征类型对于句法分析的作用呢？针对以上的问题，本研究为统计句法分析提出了一种特征类型的分析模型，该模型可以从信息论的角度量化地分析不同类型的上下文特征对于句法结构的预测作用。其基本思想是利用信息论中熵与条件熵的度量来显示一个特征类型是否抓住了预测句法结构的主要信息。如果加入某个特征类型之后当前句法结构的不确定性（熵）明显下降，则认为该特征类型抓住了上下文中影响句法结构的某些主要信息。特征类型分析的信息论模型利用预测信息量、预测信息增益、预测信息关联度以及预测信息总量四种度量从不同的侧面量化地分析各种特征类型及特征类型组合对于当前目标的预测作用。实验以 PennTreeBank 为训练集，将上下文中不同的特征类型对于句法分析规则的预测作用进行了系统的量化分析，得出了一系列有关不同特征类型及特征类型组合对句法结构的预测作用的结论。

关键词：统计句法分析 信息论 概率建模 特征类型分析

1 前言

在统计句法分析建模中，不同时期的研究者提出了不同的概率评价模型，这些模型所用的特征各异。早期的句法分析模型[Church, 1988]借鉴语音识别研究中的 N 元模型，利用固定物理距离内的词性标记作为特征来评价句法树的概率，而这些特征是非结构化的特征；概率型上下文无关语法 PCFG [Margerman, 1991][Briscoe, 1993]以句法标记（包括句法成分标记和词性标记）为特征来评价句法树的概率，其中句法成分标记是结构化的特征。后来越来越多的研究表明：（1）N 元模型所利用的非结构化特征对于处理自然语言句法分析中大量的句法结构歧义问题是远远不够的，统计句法分析的概率评价模型应考虑更多类型的特征；（2）仅靠句法成分标记等结构化特征对于自然语言句法分析也是不够的，必须利用词汇信息以及其它更多的上下文信息[Margerman,1991] [Margerman,1992] [Margerman,1995] [Collins 1996] [Charniak,1997] [Black, 1993]。

随后的概率评价模型分别利用了上下文中不同种类的语言信息。[Margerman,1991] [Margerman,1992] 的概率评价模型在 PCFG 基础上考虑了用规则扩展一个成分时的结构和词汇上下文特征，这些上下文特征包括当前句法成分的直接父结点类型以及输入句中以当前规则左端的第一个成分为中心的词性三元组。[Collins,1996]以句法树的中心词之间的依存关系为特征建立统计句法分析的概率评价模型。[Charniak,1997]利用当前句法成分的直接父结点类型及其中心词为特征评价句法树的概率。[Black,1993]将句法树中的词汇、句法、语义以及结构等多种类型的特征结合到一个统一的概率评价模型之中。[Margerman,1995]在更大的上下文环境中利用决策树学习机制从中选择有价值的特征类型。

语言所能提供的信息种类很多，在统计句法分析中，这些不同种类的语言信息的作用是不同的。因此，在概率语言模型中要有选择地加入这些语言信息。那么，如何评价不同类型的特征对于句法分析的作用呢？针对以上问题，本研究从信息论的角度为统计句法分析提出了一种特征类型分析的信息论模型。该模型利用预测信息量、预测信息增益、预测信息关联度以及预测信息总量等度量从不同侧面量化地分析不同的特征类型及特征类型组合对于句法结构的预测作用。本研究不仅将为实际的句法分析建模提供一系列有关上下文中不同特征类型及特征类型组合对句法分析作用的启发性结论，而且还可以从方法论上为句法分析建模提供一种指导——在建立一个实际的句法分析模型之前，可以首先量化地分析这些特征类型对于句法分析的作用，然后从中选择对于句法分析的预测作用较大的特征类型或特征类型组合来建立具体的概率评价模型，从而可以使得特征类型提供尽量多的句法分析信息。

以下，第 2 章描述句法树的概率评价模型；第 3 章提出了一种基于信息论的特征类型分析模型；第 4 章介绍了实验的方法；第 5 章用基于信息论的方法量化地分析了这些特征类型和特征类型的组合，得出了一系列有关不同特征类型及特征类型组合对句法分析作用的结论。

2 统计句法分析中的概率评价模型

统计句法分析的任务是：对一个句子，为它的每一个合乎语法的分析结果赋予一个概率；找到最可能的那个分析结果，并将该分析结果作为最终的句法分析结果。用公式表示如下：

$$T_{best} = \arg\max_{T} P(T \mid S)$$

其中，S 表示给定的句子，T 为所有合乎语法的句法树，$P(T|S)$表示当句子为 S 时，句法树为 T 的概率。

句法分析中概率评价模型的任务就是估计 $P(T|S)$。

首先应当确定如何表示结构化的句法树。在使用“ 规则型语法 ”的句法分析模型中，用自顶向下的方法对一个句子进行句法分析的过程可看作是从起始结点 B 开始执行的一个推导 D：

$$B \xrightarrow{r_1} \partial_1 \xrightarrow{r_2} \partial_2 \xrightarrow{r_3} \cdots \xrightarrow{r_m} \partial_m = S$$

其中，r_i 表示第 i 个非终结点的重写规则，∂_i 表示在第 i 步推导后得到的终结符和非终结符串。整个推导 D 对应一个重写规则序列（随机事件序列） $r_1, r_2, \cdots, r_m$，其中 m 表示推导过程为 m 步。

因此，一个句法树的概率可以被定义为在给定输入句 S 后，生成当前句法树的推导过程的概率，于是

$$\begin{aligned} P(T \mid S) &= P(D \mid S) \\ &= P(r_1, r_2, \cdots, r_n \mid S) \\ &= \prod_{i=1}^{n} P(r_i \mid r_1, r_2, \cdots, r_{i-1}, S) \\ &= \prod_{i=1}^{n} P(r_i \mid h_i, S) \end{aligned}$$

其中，h_i 表示由推导规则序列 $r_1, r_2, \cdots, r_{i-1}$ 推导出的部分句法树。

为准确地估计参数，需要通过选择特征类型 $F_1, F_2, \cdots, F_m$ 将规则 r_i 出现的上下文条件 h_i, S 划分为等价类，即：$\Phi : h_i, S \xrightarrow{F_1, F_2, \cdots, F_m} [h_i, S]$，使得：

$$\prod_{i=1}^{n} P(r_i \mid h_i, S) \approx \prod_{i=1}^{n} P(r_i \mid [h_i, S])$$

因此，各种句法分析概率评价模型都可看作一个根据部分句法树以及输入句的等价类预测当前推导规则的概率评价模型，即：

$$P(T \mid S) \approx \prod_{i=1}^{n} P(r_i \mid [h_i, S])$$

在统一了统计句法分析概率评价模型的形式之后，不同的概率评价模型的区别主要在于划

分等价类时所使用的特征类型的不同。不同的特征类型对于句法结构的预测作用是不同的。以下，提出一种基于信息论的模型来分析不同的特征类型或特征类型的组合对于句法结构的预测作用。

3 统计句法分析建模中基于信息论的特征类型分析模型

在预测随机事件时，熵和条件熵的概念可用来评价不同特征类型的预测作用。将某个随机事件作为预测目标，可以计算它在特定语料中的熵，在已知该事件的某个类型的特征之后，可以计算该随机事件的条件熵。如果该条件熵比事件的熵小很多，则可以认为在加入这个类型的特征之后，该随机事件的不确定性显著下降，进而可以认为该特征类型抓住了预测当前随机事件的某些本质信息。则该特征类型对于预测当前目标是很有用的。

以下，利用这种思想来建立特征类型分析的信息论模型。

特征类型分析的信息论模型把句法分析过程中当前非终结点的扩展规则作为预测目标，由预测信息量、预测信息增益、预测信息关联度和预测信息总量四个概念组成。

- **预测信息量 Predictive Information Quantity (PIQ)**

特征类型 F 对于预测目标 O 的预测信息量 $PIQ(F;O)$定义为二者之间的互信息 $I(F;O)$，即预测目标 O 的熵 $H(O)$ 与以特征类型 F 为条件，预测目标 O 的条件熵 $H(O \mid F)$ 之间的差。

$$
\begin{aligned}
& PIQ(F;O) \\
&= I(F;O) \\
&= H(O) - H(O \mid F) \\
&= H(F) - H(F \mid O) \\
&= \sum_{f \in F, o \in O} P(f,o) \log \frac{P(f,o)}{P(f) \cdot P(o)}
\end{aligned}
$$

预测信息量与熵和条件熵的关系如图 1 所示。

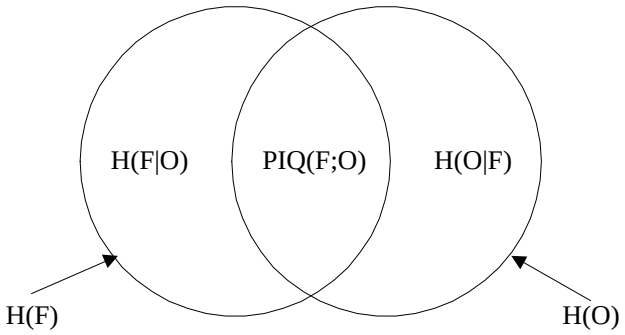

图 1 预测信息量与熵、条件熵的关系

预测信息量在特征类型分析中的意义在于衡量一个单一的特征类型对于预测指定目标的作用。假设存在特征类型F_1和F_2，如果 $PIQ(F_1;O) > PIQ(F_2;O)$，则认为F_1为预测目标O所带来的预测信息量大于F_2的预测信息量。

- **预测信息增益 Predictive Information Gain (PIG)**

在模型已有特征类型序列$F_1,F_2,...,F_i$的情况下，加入特征类型F_x对于预测目标O所带来的预测信息增益$PIG(F_x;O|\ F_1,F_2,...,F_i)$定义为在以$F_1,F_2,...,F_i$为条件时预测目标$O$与特征类型$F_x$的互信息，即在以$F_1,F_2,...,F_i$为条件时预测目标$O$的条件熵与以$F_1,F_2,...F_i,F_x$为条件时预测目标$O$的条件熵之间的差。

$$
\begin{aligned}
&PIG(F_x;O \mid F_1,F_2,\cdots,F_i)\\
&= I(F_x;O \mid F_1,F_2,\cdots,F_i)\\
&= H(O \mid F_1,F_2,\cdots,F_i) - H(O \mid F_1,F_2,\cdots,F_i,F_x)\\
&= \sum_{f_1\in F_1,\, f_2\in F_2,\cdots f_i\in F_i,\, f_x\in F_x,\, o\in O} P(f_1,f_2,\cdots,f_i,f_x,o)\log\frac{P(f_1,f_2,\cdots,f_i,f_x,o)}{P(f_1,f_2,\cdots,f_i,f_x)}\cdot\frac{P(f_1,f_2,\cdots,f_i)}{P(f_1,f_2,\cdots,f_i,o)}
\end{aligned}
$$

预测信息增益的意义在于衡量“ 在给定了一系列特征类型的基础上，再加入一个特征类型将为模型带来的新增加的预测信息 ”。

如果 $PIG(F_x;O \mid F_1,F_2,\cdots,F_i) > PIG(F_y;O \mid F_1,F_2,\cdots,F_i)$，则不论 $PIQ(F_x;O)$ 是否大于 $PIQ(F_y;O)$，都认为在特征F_1,F_2,...,F_i的基础上F_x比F_y为预测目标O带来更多的预测信息。所以，在模型已有特征类型F_1,F_2,...,F_i后，从预测信息的角度来说，下一个应该选择的特征类型为F_x而不是F_y。

- **预测信息关联度 Predictive Information Association Ratio(PIA).**

特征类型 F_x 和特征类型序列 F_1,F_2,...,F_i 在预测目标 O 时的预测信息关联度 $PIA(F_x,\{F_1,F_2,\cdots,F_i\};O)$ 定义为特征类型F_x的预测信息量与在给定特征类型序列F_1,F_2,...,F_i的情况下特征类型F_x对于预测目标O的预测信息增益之间的差。

$$\begin{aligned} &PIA(F_x,\{F_1,F_2,\cdots,F_i\};O) \\ &= PIQ(F_x;O) - PIG(F_x;O \mid F_1,F_2,\cdots,F_i) \end{aligned}$$

预测信息关联度的意义在于衡量特征类型之间对于预测目标的预测信息的关联程度。当预测信息关联度大于 0 时，表示特征类型之间对于预测目标的预测信息存在冗余，这时的预测信息关联度表示的是特征类型之间的预测信息的重叠程度；当预测信息关联度小于 0 时，表示特征类型之间对于预测目标的预测信息存在互补，这时的预测信息关联度表示的是特征类型之间的预测信息的互补程度；当预测信息关联等于 0 时，表示特征类型之间对于预测目标的预测信息是无关的。

- **预测信息总量 Predicting Information Summation (PIS).**

特征类型组合F_1,F_2,...,F_m 对于预测目标O的预测信息总量$PIS(F_1,F_2,...,F_m;O)$表示该特征类型组合为预测目标所提供的所有预测信息。确切的定义如下：

$$PIS(F_1, F_2, \cdots, F_m; O)$$
$$= PIQ(F_1; O) + \sum_{i=2}^{m} PIG(F_i; O \mid F_1, \cdots, F_{i-1})$$

从以上四种预测信息度量的定义可以看出，对于特定的预测目标，特征类型的分析模型定义并描述了单一特征类型的预测信息量，在已有特征类型组合的基础上加入新特征类型后的预测信息增益，特征类型之间的预测信息关联度以及多个特征类型组合的预测信息总量，它们从不同的侧面描述了特征类型以及特征类型组合对于预测目标的预测能力，从而为统计句法分析建模中特征类型的选择提供了一种客观的评价标准和量化的选择依据。

4 实验方法

4.1 特征的分类

以下，以图 2 所示的句法树为例说明对预测当前结点的推导规则的特征的分类。

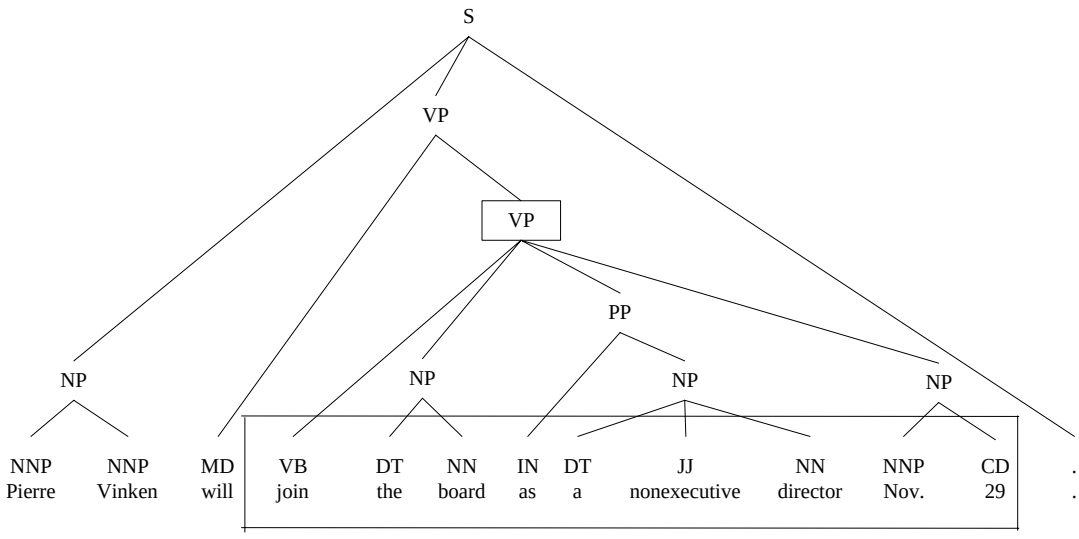

图 2：特征分类示意图

实验的预测对象为当前非终结点（VP）的扩展规则；预测的特征可分为历史特征和目标特征两大类。其中，实线连接的部分为历史特征，是已经扩展的句法树部分，代表当前非终结点所处的结构化的语言环境；矩形框内的部分为目标特征，是包含当前结点的最底层组成词汇的线性的词串，代表从当前结点开始扩展的最终目标。

4.2 实验的语料

4.2.1 PennTreeBank 介绍

Penn Treebank[Marcus, 1993]通过人工为英语语料库中的每个句子标注了语法结构，并且为其中的每个词标注了词性标记(参见附录 1 的 PennTreeBank 词性标记集与附录 2 的 PennTreeBank 短语类型标记集)。标注过的句法树的示例如下：

```
( (S (NP-SBJ (NP (NNP Pierre)
                 (NNP Vinken) )
             (, ,)
             (ADJP (NP (CD 61)
                       (NNS years) )
                   (JJ old) )
             (, ,) )
         (VP (MD will)
             (VP (VB join)
                 (NP (DT the)
```

(NN board))

(PP-CLR (IN as)

(NP (DT a)

(JJ nonexecutive)

(NN director)))

(NP-TMP (NNP Nov.)

(CD 29))))

(. .)))

图 3 PennTreeBank 中的句法分析树示例

4.2.2 树库语法的获取

本实验通过学习 PennTreeBank 中每一个经过句法分析的句子“ read the grammar off the parsed sentences ”[Charniak, 1996]来获取英语语法。

例如，从图 3 的句法分析过的句子中可以得出以下的句法结构规则：

S→NP VP .

NP→NP , ADJP ,

NP→NNP NNP

ADJP→NP JJ

NP→CD NNS

VP→MD VP

VP→VB NP PP NP

NP→DT NN

PP→IN NP

NP→DT JJ NN

NP→NNP CD

用这种方法，实验抽取了 8,126 条上下文无关的英语语法规则，构成实验所需要的英语语法规则集。

4.2.3 中心词的指定

PennTreeBank 语料标注的是英语句子的语法结构及每个词的词性，本实验进一步利用规则驱动的方法为树库语料中的每个非终结点标注了主词。基本方法为首先人工为每一条语法规则指明其中主成分的位置，然后依据这些规则为语料库中的每个非终结点标注主词。

增加了主词信息的语法规则如下：

S(HEAD_POS=”2”)→NP VP .

NP(HEAD_POS=”1”)→NP , ADJP ,

NP(HEAD_POS=”2”)→NNP NNP

ADJP(HEAD_POS=”2”)→NP JJ

NP(HEAD_POS=”2”)→CD NNS

VP(HEAD_POS=”2”)→MD VP

VP(HEAD_POS=”1”)→VB NP PP NP

NP(HEAD_POS=”2”)→DT NN

PP(HEAD_POS=”1”)→IN NP

NP(HEAD_POS=”3”)→DT JJ NN

NP(HEAD_POS=”1”)→NNP CD

4.2.4 实验语料

实验语料为 PennTreebank 树库中 Wall Street Journal 语料库，为其中的每个非终结符标注了主词。用其中的 80%（979,767 词）为训练集，用来计算所需要的各种概率；10%（133,814 词）为测试集，用来计算预测信息量、预测信息增益、预测信息关联度以及预测信息总量，剩余的 10%（133,814 词）为预留集，留作以后使用。基本的语法规则集为从 PennTreeBank 树库中抽取的上下文无关语法规则集，共 8,126 条上下文无关规则。

表 1：实验中语料的规模

语料类型	规模	所占比例
训练集	979,767 词	80%
测试集	133,814 词	10%
预留集	133,814 词	10%

4.3 实验中用到的平滑方法

在特征类型分析中计算各种信息度量时最大的困难是零概率问题。在实验中，我们使用混合方法（blending approach）[Bell, 1992]来解决这个问题。

令L为候选特征类型数，$F_1,F_2,...,F_i$ $(0<i<L)$为到目前为止选定的特征类型序列，模型i为利用$F_1,F_2,...,F_i$ $(0<i<L)$来预测推导规则的概率评价模型，$p(F_1,F_2,...,F_i;O)$为模型i赋予规则O的概率。如果模型i的权值为 ω_i，则混合概率 $\hat{p}(F_1,F_2,...,F_i;O)$ 计算如下：

$$\hat{p}(F_1, F_2, \cdots, F_i; O)$$
$$= \omega_{-1} p_{-1}(O) + \omega_0 p_0(O) + \sum_{j=1}^{i} \omega_j p(F_1, F_2, \cdots, F_j; O)$$

其中，$p_0(O)$表示由一元模型赋予规则O的概率，$p_{-1}(O)$表示规则O的基础概率，即同类（右部相同）规则总数的倒数。权值ω_i用[Bell, 1992]提出的逃逸概率来确定，所有权值的之和为 1。

5 基于信息论的特征类型的量化分析

5.1 词汇类型特征、词性类型特征以及结点成分类型特征的预测信息量分析

- **研究目的**

近年来统计句法分析最大的变化是在概率评价模型中引入了具体词的统计信息。一些统计句法分析的实践已经证明了加入词汇信息之后句法分析系统效率的提高。这里，希望从信息论的角度量化地分析词汇、词性和成分信息为句法分析所带来的预测信息量的差别。

- **特征的选择**

选择历史特征来进行研究，其中历史特征为在虚线所包括的部分句法结构树中，与当前结点的最小结构距离在 2 之内的结点（即：实线内的部分）。在每个结构位置上，选择的特征类型为结点类型、结点的主词以及结点主词的词性（如图 4 所示）。

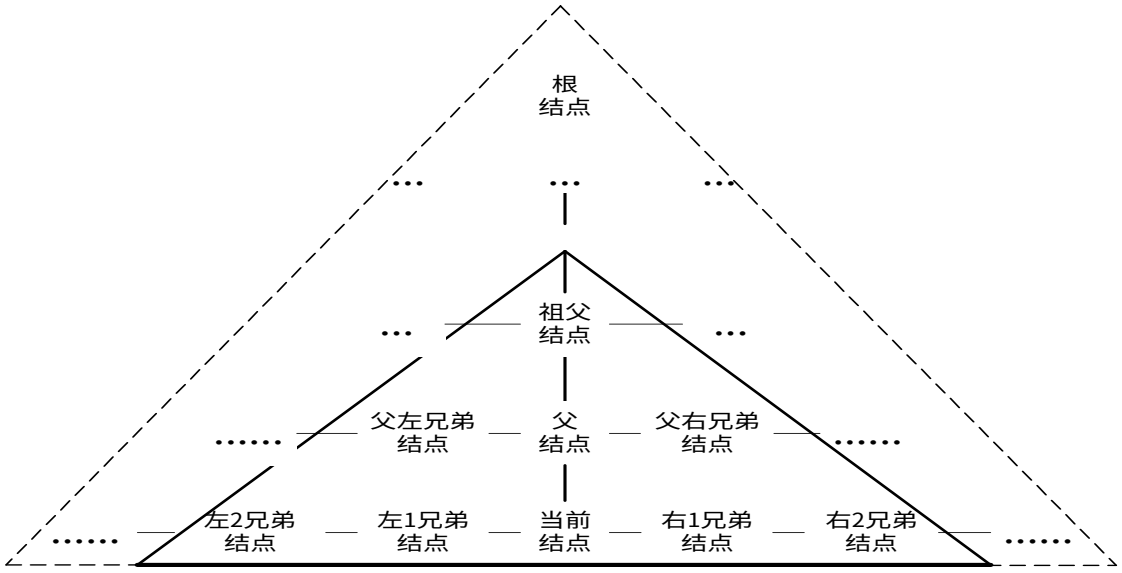

图 4: 历史特征候选类型

- **实验数据**

实验的预测目标为非终结点的扩展规则，用 R 表示。

表 2：历史特征类型的预测信息量

PIQ(当前结点类型;R)=2.3609	PIQ(当前结点主词;R)= 3.7333	PIQ(当前结点主词词性; R)= 2.7708
PIQ(祖父结点类; R)= 0.6483	PIQ(祖父结点主词;R)=1.6808	PIQ(祖父结点主词词性; R)= 0.6612
PIQ(父结点类型;R)= 1.1598	PIQ(父结点主词;R)=2.3253	PIQ(父结点主词词性;R)= 1.1784
PIQ(父右兄弟结点类型;R)=0.1068	PIQ(父右兄弟结点主词;R)=0.3717	PIQ(父右兄弟结点主词词性;R)=0.2133
PIQ(父左兄弟结点类型;R)=0.2505	PIQ(父左兄弟结点主词;R)=1.5603	PIQ(父左兄弟结点主词词性;R)=0.6145
PIQ(右 1 兄弟结点类型;R)=0.4730	PIQ(右 1 兄弟结点主词;R)=1.1525	PIQ(右 1 兄弟结点主词词性;R)=0.7502
PIQ(右 2 兄弟结点类型;R)=0.1066	PIQ(右 2 兄弟结点主词;R)=0.5044	PIQ(右 2 兄弟结点主词词性;R)=0.2525
PIQ(左 1 兄弟结点类型;R)=0.5832	PIQ(左 1 兄弟结点主词;R)=2.1511	PIQ(左 1 兄弟结点主词词性;R)=1.2186
PIQ(左 2 兄弟结点类型;R)=0.0949	PIQ(左 2 兄弟结点主词;R)=0.6171	PIQ(左 2 兄弟结点主词词性;R)=0.2697

- **结论**

在句法树的同一个结构位置上的特征类型中，词汇特征类型的预测信息量大于词性特征类

型的预测信息量，而词性特征类型的预测信息量大于结点成分特征类型的预测信息量。

5.2 结构关系和结构距离对于历史特征类型预测信息量的影响

- **研究目的**

如图 4 所示，历史特征可以有很多候选，它们对于预测当前非终结点的扩展规则的预测信息量是不同的。一个特征类型的预测信息量是否与该特征类型所属的结点与当前结点在句法树中的结构关系和结构距离有关？

- **实验数据**

表 3：结构关系和结构距离对于历史特征类型预测信息量的影响

结构关系 / 结构距离	父子关系	兄弟关系	父子与兄弟关系的混合
1	PIQ(父结点类型;R)=1.1598	PIQ(左 1 兄弟结点类型;R)=0.5832	PIQ(父左兄弟结点类型;R)=0.2505;
		PIQ(右 1 兄弟结点类型;R)=0.4730	
2	PIQ(祖父结点类型;R)=0.6483	PIQ(左 2 兄弟结点类型;R)=0.0949	PIQ(父右兄弟结点类型;R)=0.1068
		PIQ(右 2 兄弟结点类型;R)=0.1066	

- **结论**

在历史特征类型中，当结构关系相同时（同是父子关系，或同是兄弟关系），与当前结点的结构距离越近预测信息量越大；当结构距离相同时，与当前结点具有父子关系的特征的预测信息量比与当前结点具有兄弟关系以及父子、兄弟关系的混合的特征（例如：父结点的兄弟结点特征类型）的预测信息量大。

5.3 历史特征类型和目标特征类型的预测信息量的比较

- **研究目的**

现在选择特征类型时，有的很多概率评价模型往往倾向于选择历史特征类型，而不选择目标特征类型。本实验选择一些历史特征类型和目标特征类型，量化地比较它们各自的预测信息量，从而发现对于预测当前结点的扩展规则来说，哪种特征类型的预测信息量更大？

- **实验数据**

选择历史特征候选中预测信息量最大的特征，即：父结点主词；随机选择目标特征，即：目标词串中的第 1 个词和目标词串中的第 2 个词。

表 4：历史特征类型和目标特征类型的预测信息量的比较

历史特征预测信息量	目标特征预测信息量
PIQ(父结点主词;R)=2.3253	PIQ(目标词串中的第 1 个词;R)=3.2398
	PIQ(目标词串中的第 1 个词;R)=3.0071

- **结论**

目标词串中第 1 个词和第 2 个词的预测信息量均大于历史特征中预测信息量最大的父结点主词的预测信息量。

5.4 物理位置信息、启发性的主词和修饰词信息以及确切的主词信息对于目标特征类型的预测信息量的影响

- **研究目的**

从 4.1 特征的分类可以看出，不同于结构化的历史特征类型，目标特征类型是线性的。一

般地，候选的目标特征类型是按物理位置来选取的。但从语言学的角度看，物理位置很难抓住语言结构之间的联系。因此，除物理位置之外，这里尝试分别用确切的主词信息和启发式的主词与修饰词信息来帮助选择目标特征类型。通过这个实验，我们希望发现确切的主词信息、启发式的主词与修饰词信息和物理位置信息分别对于依照它们所选定的特征类型的预测信息量的影响。

- **实验数据**

表 5：目标特征类型的预测信息量

用来选择目标特征的信息	**特征类型的预测信息量**
物理位置信息	PIQ(目标词串中的第 1 个词)=3.2398
启发示信息 1：根据词性，判断当前词是否可以做当前结点的主词	PIQ(目标词串中的第 1 个可以做当前结点主词的词)=3.1401
启发示信息 2：根据词性，判断当前词是否可以做当前结点的修饰词	PIQ(目标词串中的第 1 个可以做当前结点修饰词的词)=3.1374
启发示信息 3：在主词已知的情况下，根据词，判断当前词是否可以修饰当前结点的主词，并与当前结点的主词构成当前结点的类型	PIQ(目标词串中的第 1 个可以修饰当前结点的主词，并构成当前结点的类型的词)=2.8757
确切的主词信息--结点的主词	PIQ(当前结点的主词)=3.7333

- **结论**

当前结点主词的预测信息量大于按启发式的主词和修饰词信息选择的特征类型和按物理位置选择的特征类型的预测信息量；而按启发式的主词和修饰词信息选择的特征类型的预测信息量小于按物理位置选择的特征类型的预测信息量。

❖ 主词信息在实际句法分析中的应用

从以上的实验分析可知，统计句法分析概率评价模型中主词信息的引入使对于扩展规则的

估计更准确。然而，在实际句法分析中，当前结点的主词一般是不能事先知道的。在这种情况下如何利用主词的信息呢？[Charniak,1997]提供了一种在统计句法分析模型中使用主词信息的方法。其主要思想是：把每个结点的出现概率分为该结点主词的出现概率和该结点的扩展规则的出现概率两部分，首先利用分析的历史估计当前结点主词的概率，然后估计在给定结点主词的条件下结点的扩展规则的概率，最后递归地计算当前结点的儿子结点的概率。详细计算请参看[Charniak,1997]原文。

6 结束语

本文提出了一种基于信息论的特征类型分析模型，不仅为实际的句法分析建模提供了一系列启发示的建议，而且从方法论上为句法分析建模提供了一种指导，即在建立一个实际的句法分析模型之前，可以首先用这种方法去分析一下所选择的特征，选择其中预测信息量或预测信息总量大的特征类型或特征类型组合，再将它们用于实际的句法分析模型中，从而使句法分析建模少走弯路。

参考文献：

[Bell, 1992] Bell, T.C., Cleary, J.G., Witten,I.H. Text compression. Englewood Cliffs, New Jersey 07632: PRENTICE HALL, 1992

[Black, 1993] Black, E., Jelinek, F., Lafferty, J., Magerman, D.M., Mercer, R. and Roukos, S. Towards history-based grammars: using richer models of context in probabilistic parsing. In: Proceedings of the 31st Annual Meeting of the ACL, Columbus, Ohio, 1993, 31-37

[Briscoe, 1993] Briscoe, T., Carroll,J. Generalized LR parsing of natural language (corpora) with

unification-based grammars. Computational Linguistics, 1993, 19(1): 25-60

[Charniak, 1997] Charniak, E. Statistical parsing with a context-free grammar and word statics. In: Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI Press/MIT Press, Menlo Park, CA, 1997, 598-603.

[Church, 1988] Kenneth Ward Church, A stochastic parts program and noun phrase parser for unrestricted text. In: Proceedings of the Second Conference on Applied Natural Language Processing, ACL, Austin, Texas, s1988, 136-143.

[Collins, 1996] Collins, M. J. A new statistical parser based on bigram lexical dependencies. In: Proceedings of the 34th Annual Meeting of the ACL, Santa Cruz, CA, 1996, 184-191.

[Collins, 1997] Collins, M.J. Three generative lexicalised models for statistical parsing. In: Proceedings of the 35th Annual Meeting of the ACL, Madrid, Spain, 1997, 16-23.

[Eisner, 1996] Eisner, J. Three new probabilistic models for dependency parsing: An exploration. In: Proceedings of COLING-96, Copenhagen, Denmark, 1996, 340-345

[Magerman, 1991] Magerman, D.M. and Marcus, M.P. Pearl: a probabilistic chart parser. In: Proceedings of the European ACL Conference, Berlin, Germany, 1991, http://www-cs-students.stanford.edu/~magerman/pubs.html.

[Magerman, 1992] Magerman, D.M. and Weir, C. Probabilistic prediction and Picky chart parsing. In: Proceedings of DARPA Speech and Natural Language Workshop, Arden House, NY, 1992, http://www-cs-students.stanford.edu/~magerman/pubs.html.

[Magerman, 1995] Magerman, D.M. Statistical decision-tree models for parsing. In: Proceedings of the 33th Annual Meeting of the ACL, Cambridge, MA, 1995, 276-283.

[Marcus, 1993] Marcus,M. P., Santorini, B., Marcinkiewicz, M. A. Building a large annotated corpus of English: the Penn treebank. Computational Linguistics, 1993, 19(2): 313-330

[穗，2000]穗志方，语言建模中基于信息论的特征类型分析及其应用，北京大学计算机科学与技术系，中国，北京，博士后出站报告，2000 年 6 月

[Sui, 2000] Sui Zhifang, The information theory based feature type analysis and its applications, Department of Computer Science and Technology, Peking University, Beijing, China, Postdoctoral Report, June, 2000

附录 1 PennTreeBank 的词性标记集

CC 并列连词

CD 基数

DT 限定词

EX WORD"there"

FW 外文

IN 介词或从属连词

JJ 形容词

JJR 形容词比较级

JJS 形容词最高级

LS 列项符

MD 情态词

NN 名词（单数或不可数）

NNS 名词（复数）

NNP 专有名词（单数）

NNPS 专有名词（复数）

PDT 前限定词

POS 所有关系的结束

PRP 人称代词

PRP$ 物主代词

RB 副词

RBR 副词比较级

RBS 副词最高级

RP 小品词

SYM 符号

TO “ to ”

UH 感叹词，语气词

VB 动词，基本型

VBD 动词，过去时

VBG 动词，现在分词或进行时

VBN 动词，过去分词

VBP 非第三人称单数现在时

VBZ 第三人称单数现在时

WDT WH 限定词，如："which"

WP WH 代词，如：" what, who, whom"

WP$ WH 所属代词，如：" whose"

WRB WH 副词，如："how, where, why"

附录 2 PennTreeBank 的短语类型标记集

ADJP 形容词短语

ADVP 副词短语

NP 名词短语

PP 介词短语

S 简单陈述句

SBAR 从句

SBARQ 疑问从句

SINV 倒装句

SQ 疑问句

VP 动词短语

WHADVP 疑问副词短语

WHNP 疑问名词短语

WHPP 疑问介词短语

X 未知或不确定的成分

The Information-Theory-Based Feature Type Analysis in the Modeling for Probabilistic Parsing

SUI Zhifang* ZHAO Jun** YU Shiwen*

* The Institute of Computational Linguistics, Department of Computer Science & Technology

Peking University, Beijing, China

** Human Language Technology Center, Computer Science Department

The Hong Kong University of Science & Technology, Hong Kong

Abstract:

In statistical parsing, the probabilistic models are used to evaluate the possibility of each candidate parse tree, where the parse tree with the largest probability is deemed to be the final result of the parsing. Therefore, the core of statistical parsing is a probabilistic evaluation model. The main difference among the various probabilistic evaluation models lies in which types of features in the context are used to assign the probabilities to the parse trees.

Various probabilistic evaluation models have been proposed in the field of statistical parsing, where different models use different feature types. How to evaluate a feature type's predictive power for the parsing tree? The paper proposes an information-theory-based feature type analysis method. Using the method, we can quantitatively analyze the power of different feature types for syntactic structure

prediction from the viewpoint of information theory. The basic idea is that we use entropy and conditional entropy to measure whether a feature type grasps some of the information for syntactic structure prediction. If the average uncertainty of the syntactic structures declines apparently, the feature type is deemed to have grasped some intrinsic linguistic information in the context that has close relation to the syntactic structures. Using Penn-Treebank as training and testing set, our experiment quantitatively analyze the different feature types' predictive power for syntactic structure prediction in a systematic way and draws a series of conclusions which reflect the predictive power of different feature types and feature type combination for syntactic parsing.

Keyword: Statistic Parsing; Information Theory; Probabilistic Modeling; Feature Type Analysis

作者简历：

穗志方，女， 1970 年 6 月出生， 博士后，研究方向为：计算语言学，机器翻译

赵军，男，1966 年 12 月出生， 博士，研究方向为：计算语言学

俞士汶， 男，1938 年 12 月出生，教授，博士生导师，研究方向为：计算语言学，机器翻译

联系电话：

穗志方：852-92393169（香港）；

赵军：852-92393169（香港）；

俞士汶：86-10-62751892（北京）

Email:

穗志方：suizf@hotmail.com ; suizf@isilk.com

赵军：zhaojun123@hotmail.com

俞士汶：yusw@pku.edu.cn