一个包含复杂特征的统计英语句法分析模型

孟 遥 黄玉 赵铁军 李 生
（哈尔滨工业大学计算机科学与技术学院 哈尔滨 15001）
E-mail: {meng,hy}@mtlab.hit.edu.cn

摘要 文中设计了一种包含复杂特征的统计句法分析模型，在单纯依靠词性信息的基于概率评价的句法分析模型的基础上，对只依靠词性信息难以消除歧义的规则，适当引入复杂特征集。模型综合考虑了上下文无关规则的结构特性、所处的上下文环境及复杂特征信息。实验结果表明，该方法具有较高的分析精确率和召回率。

关键字：句法分析 统计 复杂特征

A complex-feature-oriented statistic English Parsing Model

Meng Yao, Huang Yu, Zhao Tiejun, Li Sheng
Dept. of Computer Science & Engineering , Harbin Institute of Technology, Harbin, 15001
Email:{meng,hy}@mtlab.hit.edu.cn

ABSTRACT: This paper describes a new statistical parsing approach which is capable of utilizing complex-features. The parser is first trained purely on part of speech information, and then is integrated with related complex features to solve the structure ambiguity. The proposed model possesses the advantage of processing both structure information of CFG rules and localized information provided by complex features. The experiment shows a promising result by precision and recall .

Keyword: Parsing, statistic, complex feature

1 引言

句法分析是自然语言处理的一个核心内容，它的研究和实现具有重要的理论意义和实用价值[C. Manning 1999]。句法分析的任务是生成一棵带有句法功能标记的短语结构树，它能指明短语的构成成分及其之间的相互关系。句法分析的难点主要在于：句法分析中存在着大量的歧义。一个句子，应用上下文无关规则可以生成许多句法分析结果，关键是要从这些句法结果中选择出正确的结果，也就是句法结构的消歧。

目前的研究结果表明，基于规则的方法，存在着知识获取困难，鲁棒性差的问题，建立一个初具规模的句法分析器也大约需要十年的时间[Magerman 1994]。如果在句法分析的过程中只依靠词性信息，许多句法结构歧义难以消除，句法分析的结果也并不理想[G. De Pauw 2000]。因此在基于概率的上下文无关规则中引入词汇信息，成为当前句法分析研究的热点[M. Collins 1996, 1997][Franz Beil 1999]。但词汇信息的引入又带来几个问题：比如，包含词汇信息使规则数激增，为搜索带来难度；大量的数据稀疏问题如何解决等

[Stefan Riezler 2000]。

本文提出了一个新的句法分析模型：它首先依靠统计方法，设计了一个只使用词性信息的句法分析器。用一个组合型的概率评价函数解决句法结构歧义问题，然后基于错误驱动的思想，针对模型产生的错误，用人工或机器自动获取的方式抽取包含有复杂特征的校正规则，在统计语料中获取包含复杂特征的规则的概率，将其并入初始的上下文无关规则集，生成包含复杂特征的规则集，之后将新生成的规则集用于句法分析，针对新出现的错误再次获得校正规则，并入新规则集，形成对原有规则集的进一步扩展，通过对规则库逐渐修正、精化，从而改进只依靠词性信息所生成的句法分析结果。本方法很大程度地解决了词汇规则数量过大的问题，同时也一定程度地解决了数据稀疏问题。

2 包含复杂特征集的英语句法分析

由于词性标注和英语的 BNP 识别精度都很高[荀恩东 1999][Endong Xun 2000]，且这两个部分各自比较独立，所以我们在句法分析时先进行了词性标注和英语的 BNP 识别，本文所描述的句法分析模型是词性标注和 BNP 识别之后所使用的模型，我们所使用的句法分析符号集与 Penn Treebank 一致，而且本文用 Penn Treebank 作为训练语料[Marcus et al. 1994]。

首先我们只保留了 Penn Treebank 中词性符号和短语符号，去掉其中的词汇和短语的成分标志，从中获取句法分析所需要的上下文无关规则，将其作为句法分析的初始规则集。然后我们从训练语料中计算得到每一条规则本身的发射概率和符号之间的三元转移概率。

句法分析的整个过程是分层进行的，在每一层的输入是其下面一层的输出。每一层得到所有能够匹配句子子串的规则。随后对这些规则进行挑选，挑选时使用一个组合型的概率评价函数，选取概率函数评分最大的规则序列，执行规则序列中的每一条规则，合并生成新结点。以此类推，自底向上完成整个句法分析。

在规则选择时，我们使用一个组合型的概率评分函数将规则本身的出现概率与规则所处的上下文环境一起考虑，以此选取概率最优规则链。

虽然组合型的概率评分函数考虑了规则本身的概率和规则所处的上下文环境，而且我们还使用了三元文法体现规则所处的上下文环境，比二元文法包含更多的上下文信息。但这种完全依赖词性信息的句法分析模型，它所具有的消歧知识是粗粒度的，它忽视了词汇、语义、特殊句型等消歧中的重要信息[C. Manning 1999][D.M. Magerman 1995]。面对自然语言的各种丰富现象，单纯依赖词性信息的句法分析根本不能解决所有的歧义问题。

为此我们提供了一个用于修正上下文无关规则的修正树库，先用基于词性的句法分析器分析修正树库中的每个句子，在每一层句法分析之后，抽取出分析错误的位置，以错误的合并生成前的状态为初始状态，采用人工或机器自动提取的方法，书写一条包含有复杂特征的校正规则，改正错误的合并。修正树库分析后，去除校正规则中的重复规则，将其合并成校正规则集。根据训练语料统计每个校正规则的概率，将含有概率的校正规则并入初始规则集，形成包含有复杂特征的规则集，用新的规则集分析新的修正树库，再次得到校正规则，并入原有规则集，这样，采用循序渐进的方法，逐渐改进句法分析的精度。

该模型与通常的错误驱动的方法有所不同，常用的错误驱动方法是在错误的结果下书写一条规则，改变错误的结果，而文中的校正规则是基于正确的分析结果书写的，它相当于对原有的上下文无关规则的细化，它通过对包含复杂特征的规则赋予比简单的上下文无关规则更高的概率，使这种规则在规则选择中被优先选中，从而达到改正错误的目的。

相对于以往的词汇型的统计句法分析模型，本文提出的方法，有选择地对上下文规则增加复杂特征集，大大减少了规则的数量，并且一定程度地解决了数据稀疏问题。

3 组合型概率评价函数

在规则集中所有能够匹配输入符号链一部分的规则构成候选规则集。候选规则集中的某些规则是不能够同时出现的，我们把一组互不包含，且覆盖句子所有节点的规则，叫做候选规则链。每一条规则链即对应着一种句法结构，也就是一种句法分析结果，输入一个句子，它可以生成的规则链可能很多，句法结构消歧的任务是通过某种选择机制，选取最适合该句子的候选规则链。选取使概率评价函数最大的规则链，作为最优规则链。

给定一条规则 $R^i: W_k, W_{k+1}, \ldots, W_l \rightarrow N_i$，规则的概率评价函数为：

$$Score(R^i) = S_{stru}(R^i) * S_{cont}(R^i \mid R^{i-1}, R^{i-2})$$

其中 $S_{stru}(R^i) = \alpha P(N_i \rightarrow w_k, w_{k+1}, \ldots w_l)$ 为概率上下文无关文法中规则的概率，在这里称为规则本身的概率，它表示规则本身出现的概率，α 为权值，它表示在概率评分函数中规则本身的概率的权重。$S_{cont}(R^i \mid R^{i-1}, R^{i-2}) = \beta P(N_i \mid N_{i-1}, N_{i-2})$ 其中，N_{i-1}和N_{i-2}为在规则链上该规则前两个规则所生成的非终极符号，$P(N_i \mid N_{i-1}, N_{i-2})$ 它表示了规则出现时它的上下文对它的认可程度，β 权值，表示在整个概率评分函数中规则的上下文认可程度的比重。在这里称为规则的上下文概率。实验显示，α：β=6：1 时获得分析结果最为理想[孟遥 2001]。

规则链 L_i 的概率评分函数 $Score(L_i) = \prod_{R^i \in L_i} Score(R^i)$，其中$R^i$为$L_i$中规则。

最优规则链 $= \arg\max\{Score(L_i)\}$，其中L_i为任意一条规则链。

规则链选择时，使用了动态规划方法。

关于评价函数中权值的选择及动态规划的详细过程参见[孟遥 2001]。

4 包含复杂特征的句法分析规则

尽管在基于统计的句法分析过程中，采用了综合的概率评分机制，全面考虑了规则本身的概率和它的上下文信息，而且在表达上下文信息时我们使用了三元文法，比通常使用的二元文法更多地表达了结点的上下文信息。但错误的结果显示，许多歧义，使用词性信息和简单的上下文关系难以消除，而引入语义信息、词本身的信息或较深层次的结构关系且较易消除这些错误。因此我们采用了错误驱动的思想，在每一层基于统计的句法分析结束后，通过机器自动抽取与人工参与的方法针对错误的分析结果获得校正规则，校正规则使用包含复杂特征的上下文相关规则。

校正规则的获取过程如下：

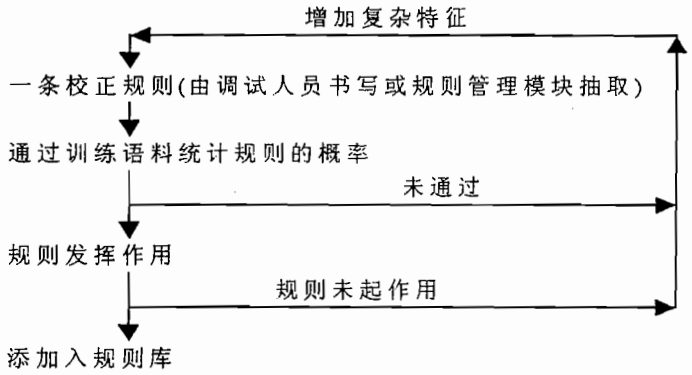

图 3：获取复杂特征规则过程

在每一层句法分析之后，根据错误的位置，书写一条含有复杂特征的规则，从训练语料中获得该条规则的概率，将其并入初始规则库。重新分析错误的句子，如果错误得到改正，则将此规则保留在规则库中，如果在规则选择的过程中，该条校正规则仍未被选中，则修改规则，增加规则所包含的复杂特征，从而提高规则在当前情况下出现的概率。

任意一条规则 L 的格式为：

$$P_1[F_1=v_1],\ \cdots\cdots,\ P_i[Fi=v_i],\ P_{i+1}[F_{i+1}=v_{i+1}],\ \cdots\cdots,\ P_j[F_j=v_j],\ \cdots\cdots,\ P_n[F_n=v_n]$$
$$\rightarrow C(\ i\ ,\ j\ ,\ P_k)$$

其中 P_i 为短语或词性符号，F_i 为一个复杂特征框架，每一结点可以有 0 个或多个复杂特征框架。v_i 为复杂特征值，i，j 为规则合并的开始结点和结束结点。

规则 L 的概率为：

$$P(L) = P(P_i,\ P_{i+1},\ \cdots\cdots,\ P_j/\ P_k,\ ,\ P(F_1=v_1),\ \cdots\cdots,\ P(F_{i-1}=v_{i-1}),\ P(F_{j-1}=v_{j-1}),\ \cdots\cdots,\ P(F_n=v_n),\ F_{i+1}=v_{i+1},\ \cdots\cdots,\ F_j=v_j)$$

从规则的概率计算公式可以看出，包含有复杂特征的规则其概率大于只含有词性信息的上下文无关规则，根据组合型概率评价函数的定义，在规则选择过程中含有复杂特征的规则将被优先选择，而且复杂特征包含的越多，优先选择的可能性越大。通过这种方法，校正了只依靠词性信息产生的错误，从而提高了句法分析的精度。

应用这种方法，对于封闭语料几乎可以校正句法分析中出现的所有错误，而对于开放语料，可以保证其精确率和召回率高于不包含复杂特征规则时句法分析精确率和召回率。

5 实验结果

实验包括封闭测试和开放测试两部分，封闭测试使用 Penn Treebank 中随机抽取的 500 个句子作为测试文本，开放测试文本选自某机器翻译比赛所用英语测试文本（平均每句 15 个单词），从中随机抽取 160 个句子，由我校外语系教师，参考 PennTreeBank 的标注标准进行手工标注。

测试时使用了三个测试指标。

其中：精确率：$\frac{\text{正确的短语个数}}{\text{识别的总的短语个数}}$，召回率：$\frac{\text{正确的短语个数}}{\text{测试句中总的短语个数}}$，

F-估计(measure)：$\frac{2*(\text{精确率}\ *\text{召回率}\ )}{\text{精确率}\ +\text{召回率}}$。

具体实验结果如下：

句法分析评测		出现短语数	识别短语数	正确识别数	召回率	精确率	F 估计
封闭	未含复杂特征	7161	6822	5388	0.7524	0.7897	0.7706
	含复杂特征	7161	7046	5832	0.8144	0.8277	0.8210
开放	未含复杂特征	1134	1102	907	0.7998	0.8230	0.8112
	含复杂特征	1134	1104	966	0.8518	0.8750	0.8632

测试结果中开放测试高于封闭测试，主要原因是开放测试所选句子相对封闭测试所选句子简单，另外在整个句法分析测试中开放测试的召回率高于封闭也有我们的标注的结果不如 PennTreeBank 细致的问题。

整个的实验结果显示，在基于统计的句法分析过程中，采用包含复杂特征的规则可以提高英语句法分析器的性能。

6 结论

本文提出了一种新颖的包含复杂特征的句法分析方法，有效地消除了在句法分析过程中出现的大量句法结构歧义。组合型的概率评价函数，不仅可以考虑规则本身的出现概率和规则所处的上下文环境，还可以调整规则本身的概率和规则的上下文概率在评价中的比重。通过实验选择最优的比例。基于错误驱动的思想选取复杂特征规则，不仅可以充分应

用词汇、语义、复杂的上下文结构关系的消歧作用，而且规则库中规则的数量也远远小于对任意规则都包含词汇信息的基于词汇的统计句法分析模型，一定程度地解决了词汇级句法分析模型数据稀疏模型。实验表明，采用包含复杂特征的统计句法分析模型可以提高句法分析的精度。

参考文献

[C. Manning 1999] Christopher Manning, Hinrich Schutze. Foundations of Statistical Natural Language Processing. The MIT Press, 1999

[Collins, 1997] Michael John Collins. A New Statistical Parser Based on Bigram Lexical Dependencies. cmp-lg/9605012

[Collins, 1997] Michael John Collins. Three Generative, Lexicalised Models for Statistical Parsing ACL1997

[Magerman, 1994] David M.Magerman. Natural Language Parsing as Statistical Pattern Recognition. Doctoral Dissertation. Stanford University, 1994

[Franz Beil 1999] Franz Beil, Glenn Carroll. Inside-Outside Estimation of a Lexicalized PCFG for German. CL/9905009

[G. De Pauw 2000] Guy De Pauw. Aspects of Pattern-matching in Data-Oriented Parsing. Coling 2000

[Stefan Riezler 2000] Stefan Riezler, Detlef Prescher. Lexicallized Stochastic Modeling of Constraint-Based Grammars using Log-Linear Meaaures and EM Training. ACL2000

[荀恩东,1999] 荀恩东 1999 统计与学习并举的渐进式英语句法分析 哈尔滨工业大学博士论文 1999

[Endong Xun 2000] Endong Xun, Changning Huang and Ming Zhou. A Unified Statistical Model for the Identification of English BaseNP ACL 2000.

[孟遥,2000] 孟遥 赵铁军 李生 姚建民. 一个渐进式的统计与机器学习结合的英语句法分析器 863 计划智能计算机主题学术会议. 2000