汉语句法分析建模中基于模型质量的特征选择方法*

孟遥 赵铁军 杨沐昀 李生

哈尔滨工业大学计算机科学与技术学院，哈尔滨 150001

E-mail: tjzhao@mtlab.hit.edu.cn

摘要：在汉语句法分析建模中，有选择地使用结构消歧特征，有利于把握句法分析的规律，解决数据稀疏问题，提高模型的鲁棒性。本文提出了一种量化的特征选择方法，通过由特征构建的句法分析模型对训练样本的对数似然估计，去评价特征在句法分析建模中所起的作用。初步实验表明，本文提出的方法可以大幅度减少模型所需要的特征数量，使用不超过10%的关键特征构建的句法分析模型，在封闭测试时，精确率和召回率可以接近使用全部特征构建的句法分析模型，而对于开放测试，经过特征选择的模型其效果好于未经过特征选择的模型。

关键词：特征选择 对数似然估计 汉语句法分析

The model-based feature selection for Chinese parsing

Meng Yao Zhao Tiejun Yang Muyun Li Sheng

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001

E-mail: tjzhao@mtlab.hit.edu.cn

Abstract: Feature selection in the modeling for Chinese parsing is the key problem. It can reduce the complexity of the statistics model. Feature selection is also a key step to build the more robust parsing model. This paper proposes a log-likelihood-based parsing feature selection algorithm. The effect of features for Chinese parsing modeling is evaluated by the likelihood of training samples using the model with the selected features. The experimental results indicate the feature selection will compress the features used in modeling largely. The Chinese parsing model with The 20% key features can compare with the model with all the features in the close test. And in the open test, the model with selected features outperforms the model with all the features.

Key Words: feature selection log-likelihood Chinese parsing

1 引言

上下文信息有助于消除句法分析中的结构歧义。但相对于庞大的上下文空间，真正体现结构选择的规律性的消歧特征只占较小部分。通过训练语料分析哪些是体现句法规律性的消歧特征，而哪些是由于语言的灵活和多样而偶然呈现出的一种现象是句法分析建模中的一个困难而必须面对的问题，正逐渐引起研究人员的关注。

目前的特征选择方法主要以语言学知识为指导。Magerman[1]使用一套手工选择的消歧特征，构造决策树进行句法分析；俞士汶、詹卫东等给出了一些基于语言学知识的特征选择方法[2][3]。以语言学知识为指导能够针对所处理的语言选择最有代表性的特征，但这种方法与所处理的语言，以及语言专家对语言的认知存在着很大关系。特征的评价很难量化，可移植

*本文研究受到国家863计划资助(项目号2002AA117010-09)。

性和可维护性都较差。对于统计句法分析而言，研究自动、定量的特征评价方法，更为实用和迫切。就目前而言，这方面的研究还相对薄弱。

针对目前研究存在的问题，本文提出了一种基于模型质量的特征评价方法，以特征所构成的模型的质量为标准，评价特征的优劣。本文方法的优点在于，给出了量化的特征评价方法，并将特征选择与建模统一与一个整体。

为保证模型最大可能体现选择特征的作用，而不受未选择特征的影响，本文建模使用最大熵模型。最大熵模型认为未选定的特征对估计样本没有作用，在理论上保证了在特征评价时不受其它未选定的特征的干扰，做到对特征进行公平的评价。另外，最大熵模型对特征的形式没有要求，允许特征间相互独立，也为最终建立句法分析模型提供了方便。

实验结果显示，本文提出的特征选择方法可以大幅度减少模型所需要的特征数量，使用不超过 10%的关键特征构建的模型，在封闭测试时，性能可以接近使用全部特征构建的句法分析模型，而对于开放测试，经过特征选择的模型其效果好于未经过特征选择的模型。

2 基于模型质量的特征评价方法

2.1 基于模型质量的特征评价方法

在经验风险最小化原则下，模型的评价可以考虑模型与训练样本的分布是否拟合。样本的平均对数似然估计是一种常用的评价模型拟合程度的函数，可以通过平均对数似然估计的变化，评价特征的质量。基于模型对样本的平均对数似然来评价特征可分为三个步骤：1、从特征空间中选择一组特征；2、基于选定特征构造模型；3、计算模型对样本的平均对数似然值（log-likelihood），平均对数似然值越大说明模型的分布与训练样本分布越接近，特征选择效果越好。

由于对特征的评价建立在模型评价基础上，因此采用何种模型是评价是否公平可靠的关键。本文基于最大熵原则，选取满足选择特征要求的最大熵模型为最终确定的统计模型。熵最大模型只与选择的特征相关，而认为未选择的特征对估计样本的分布不起作用。因此选择熵最大的模型可以消除未选特征的干扰，真正体现选定特征对构造模型的贡献，使特征进行公平的比较。

2.2 特征的表示形式和最大熵模型的建模方法

设$\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$是给定的样本，在句法分析中 x 指句法分析的当前状态和它所对应的上下文信息，y 是某一种结构选择。训练样本可以通过训练树库得到。

设 $\tilde{p}(x,y)$ 为(x,y)的经验概率，$\tilde{p}(y\mid x)$ 指 x 出现的情况下，y 的经验概率，$p(y\mid x)$ 指 x 出现的情况下，模型 P 对 y 的概率。

模型 P 对训练样本的平均对数似然函数为：$L(P)=\sum_{x,y}\tilde{p}(x,y)\log P(y\mid x)$。

L(P)越大，模型对样本的分布估计越准确，模型对训练样本拟合程度越高。

x 与 y 之间存在的特征，可以用一个输出为 0 或 1 的特征函数 f(x,y)表示，特征的经验概率为：$\tilde{p}(f)=\sum_{x,y}\tilde{p}(x,y)f(x,y)$。

从训练样本中学得的模型 P 对特征的概率为: $p(f)=\sum_{x,y}\tilde{p}(x)p(y\mid x)f(x,y)$。

如果训练语料中的某个特征能够体现句法分析规律，则可令选定的特征满足约束等式 $p(f)=\tilde{p}(f)$。约束等式可以有多组，约束等式的集合叫约束集。

最大熵模型，是满足约束集条件的所有模型中条件熵最大的模型。最大熵模型的形式为：

$$p_\lambda(y\mid x)=\frac{1}{Z_\lambda(x)}\exp(\sum_i\lambda_i f_i(x,y))$$

其中 $Z_\lambda(x)\equiv\sum_y\exp(\sum\lambda_i f_i(x,y))$，$\lambda_i$ 为特征 f_i 的权值。

2.3 回溯式的特征选择算法

从 D 个候选特征中选择 d 个特征，所有可能的组合数为 C_D^d，最优的选择是把各种可能的组合都算出来加以比较，以选择最优特征组，D 的数目较大时这种穷举的方法无法实现，应采用启发式方法。

D.Pietra[4][5][6]等人在介绍最大熵模型时，提出了一种顺序前进式的特征选择算法，采用贪心算法，每次选择一个特征，直到选定特征达到一定数量。此方法存在的问题是特征一旦选入，不能去除，并且还会对随后的选择产生影响。本文改进了 D.Pietra 的特征选择算法，提出了一种带有回溯机制的特征选择方法，步骤如下：

令候选特征集 F，已选定特征集 S，对应权向量 λ，$P_s(y\mid x)$ 为满足特征约束下的最大熵模型。令 $P_{s.f}^a(y\mid x)=\frac{1}{Z_a(x)}P_s(y\mid x)e^{af(x,y)}$ 为加入 f 后的最大熵模型，其中 a 为 f 的权。

改进的特征选择算法（Improved Feature-Selection Algorithm）

输入：训练样本$(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$

输出：选定特征集 S

步骤 1、令 S=Φ，K=0，M，N 分别为两个给定常数，N>M

步骤 2、循环当 K<N

步骤 2.1、循环对 F 中每一个特征 f

$$\text{求 }\Delta L(s,f)=\max_a(L(P_{s.f}^a)-L(P_s))$$

步骤 2.2、$\tilde{f}=\arg\max_{f\in F}\Delta L(s,f)$

步骤 2.3、$S = S \cup \{\tilde{f}\}$，$F = F - \{\tilde{f}\}$，K++

步骤 2.4、通过 IIS 算法计算 P_S

步骤 3、K=0，　L=0

步骤 4、循环当 L<M

步骤 4.1、循环对 S 中每一特征 f 令 $\hat{S} = S - \{f\}$，通过 IIS 算法计算 $P_{\hat{S}}$

步骤 4.2、$\hat{f} = \arg\max_{f \in S} L(P_{\hat{S}})$

步骤 4.3、$S = S - \{\hat{f}\}$，$F = F \cup \{\hat{f}\}$，L++

步骤 5、不满足结束条件则返回步骤 2

算法中步骤 2 为特征选入过程，特征选择时优先选择对数似然增加较大的特征。

由于特征间相互关联，所以选入的特征可能并非最优。步骤 4 是一个回溯过程，剔除 M 个最差特征，以弥补步骤 2 特征选择时出现的失误。

3 上下文相关的汉语句法分析模型

汉语的歧义较复杂，消歧需要更多上下文的支持。但汉语语法灵活，并非所有的上下文都对结构消歧有用。因此选择句法分析模型时应优先选择对结构消歧贡献较大的信息。本文句法分析建模时使用第 3 节介绍的基于模型质量评价的特征选择方法。

本文的句法分析过程自顶而下进行，规则的上下文为当前规则父结点和它的左兄弟结点。设 s 为句子，t 为句法分析结果，d1,d2,...,dn 为自顶向下分析所用规则。

最优句法树 $t' = \arg\max(\prod_{i=1...n} p(d_i \mid Father(d_i), Brother(.d_i)))$

规则的父结点和左兄弟结点构成了结构消歧的特征空间。具体特征定义如表 1 所示。

表 1 特征抽取模板

模板	f(x,y)=1 当且仅当
1	y=某规则后项，x 为规则前项， 如 y=vg vg wj ,x 为 S
2	y=某抽象规则后项，x 为某抽象规则前项和某父结点 如：y= vg vg wj ,x 为规则前项 S 和父结点 null
3	y=某抽象规则后项，x 为某抽象规则前项和某左兄弟结点 如：y= vg vg wj , x 为规则前项为 S，左兄弟为 null
4	y=某抽象规则后项，x 为某抽象规则前项和某父结点及某左兄弟结点，如：y= vg vg wj , x 为规则前项 S，父 null，兄 null

特征抽取模板构成了特征选择的候选集，采用特征选择策略选择特征。令 F={从特征集

中选择的特征}。则有约束集 $C=\{p(f_i)=\tilde{p}(f_i)\mid f_i\in F\}$，根据最大熵模型定义，规则的概率为 $p\ (y\mid x)=\frac{1}{Z_\lambda(x)}\exp(\sum_i \lambda_i f_i(x,y))$。其中 x 为句法分析过程中，规则的上下文，y 为当前规则。

设 s 为句子，t 为句法分析结果，则 $p(s,t)=\prod_{i=1\ldots n}\frac{1}{Z_\lambda}\exp(\sum_i \lambda_i f_i(x,y))$。

4 实验与分析

本文所用语料库为哈工大机器翻译实验室建设的哈工大中文树库语料，共 10763 句，以句法层次树形式标注，具体标注规范及 2000 句标注样例可访问 http://mtlab.hit.edu.cn 免费获得。其中训练集为 9763 句，测试集 1000 句。

1、特征选择个数确定

实验样本空间为从语料中抽取的上下文相关句法分析规则，规则按短语类型分组，x 为规则的首部和规则的父结点与左兄弟结点的同现，y 为规则的扩展部分。

最终选择特征由 N 重交叉校验确定。具体实验时，本文将训练样本分为 5 份，保留一份作为校验集，执行特征选择算法时每加入一个特征后，计算新生成的模型对校验集的平均对数似然估计，如平均对数似然估计降低，则终止特征选择过程。选择 5 组实验选择的特征的交集做最终特征集。

表 2 为 7 种主要规则最终使用的特征个数。表 3 对选择特征的有效性进行了测试。分别构建了三个模型，其中随机模型为认为规则的出现满足随机分布的模型，最大似然模型为使用所有特征采用最大似然估计构建的模型，最大熵模型为使用选择特征构建的熵最大的模型。表 3 测试了三种模型对样本的平均对数似然估计变化情况。

表 2 主要规则特征选择个数

规则类型	候选特征	选择特征
AP	1067	38
DP	95	9
MP	447	25
NP	2155	39
PP	368	5
S	2262	46
VP	2748	35

表 3 样本对数似然估计变化表

规则类型	随机估计	最大似然	最大熵
AP	-2.325341	-1.116920	-1.146800
DP	-2.132206	-1.451336	-1.458534
MP	-2.555013	-1.001244	-1.021618
NP	-3.287349	-2.266255	-2.349816
PP	-2.519675	-1.407524	-1.507341
S	-0.757190	-0.385735	-0.390911
VP	-3.159474	-2.062316	-2.196551

由表 3 可知，最终确定的特征数目只占候选特征总数的较小一部分，但表 4 显示，在只保证关键特征的分布为最大似然估计时对于训练样本的对数似然估计也接近最大似然估计。

表 2 与表 3 说明与样本分布相关的特征只占特征空间很小一部分，在模型估计时只使用

这些主要特征可以比较准确地估计样本的分布。实验显示在统计建模时引入特征选择是是可行的。

2、句法分析结果测试

针对第 3 节提出的句法分析模型，实验对比了不进行特征选择，采用最大似然估计规则的概率时句法分析的性能，及采用特征选择，规则的概率由选定特征确定的最大熵模型估计获得时句法分析的性能。在分词和词性标注完全正确的情况下，实验结果如表 4 所示：

表 4 句法分析结果比较

模型	封闭		开放	
	精确率(%)	召回率(%)	精确率(%)	召回率(%)
未特征选择	90.2	88.5	77.6	75.8
特征选择	89.5	87.2	80.1	77.2

实验说明，针对本文语料，采用不超过 10%的关键特征对于封闭测试即可以达到与使用所有特征类似的分析性能，而对于开放测试使用关键特征可以取得更好的效果。

5 结论

本文从特征对构建句法分析建模的贡献出发，提出了一种基于模型对样本的平均对数似然估计的特征选择方法，量化地评价特征在句法分析统计建模中所起的作用。针对本文数据的初步实验表明，对于汉语句法分析，关键特征不超过全部特征集的 10%，使用关键特征构建句法分析模型对于已知数据可以达到与使用所有特征构建的模型相似的效果，而对于未知数据，使用关键特征构建的模型其精确率和召回率都有较大提高。在汉语句法分析建模中引入本文提出的特征选择方法可以提高句法分析的鲁棒性。

另外本文的特征选择方法是独立于句法分析模型的方法，具有一定的通用性。

参 考 文 献

[1] David M. Magerman. Natural Language Parsing as Statistical Pattern Recognition. Ph.D. thesis, Stanford University. 1994

[2] 俞士汶、朱学锋、王惠、张芸芸：《现代汉语语法信息词典详解》，清华大学出版社，1998。

[3] 詹卫东著：《面向中文信息处理的现代汉语短语结构规则研究》，清华大学出版社，2000。

[4] A. Berger The improved iterative scaling algorithm: A gentle introduction http://www.cs.cmu.edu/afs/cs/user/aberger/www/ps/scaling.ps 1997

[5] S.D.Pietra, V.D.Pietra and J.Lafferty Inducing features of random fields IEEE Transactions on Pattern Analysis and Machine Intelligence 1997,V19(4): 380-393

[6] A.Berger S.D.Pietra V.D.Pietra A maximum entropy approach to natural language processing Computational linguistics 1996,V22(1):39-71