基于复杂特征的VN结构模板获取模型*

赵 军 黄昌宁

(清华大学计算机科学与技术系 北京 100084)
(清华大学智能技术与系统国家重点实验室 北京 100084)

摘要 提出了基于复杂特征的VN结构模板获取模型.首先用统计决策树模型生长动词分类树,然后用最小描述长度原则对动词分类树剪枝,最后由动词分类树推导出VN结构模板.实验证明,在利用结构模板进行VN结构的识别时,这种模型比基于义类和极大似然估计原则的模型具有更高的精确率和召回率.

关键词 自然语言处理,语料库,复杂特征集,统计决策树,最小描述长度原则.

中图法分类号 TP18

在汉语中,动词V和名词N的同现情况有以下3种:偏正结构VN(如"射门方法")、动宾结构VO(如"改进方法")和非法组合IC(如"包括[方法的改进]"中的"包括方法").本文把VO结构和非法组合IC统称为~VN结构.正确地识别VN结构对于句法分析、信息检索、信息抽取等都是至关重要的,其中一种重要的方法是利用词语结构模板来识别VN结构,例如,对于"射门方法"、"改进方法"和"包括方法",如果能够利用某种方法从训练语料中获得如下形式的结构模板:Hh05+"方法"→VN,Ih11+"方法"→VO,Jd05+"方法"→IC(其中Hh05、Ih11和Jd05是《同义词词林》[1](以下简称《词林》)的义类代码),则可以正确地识别它们的结构.

基于词的VN结构模板获取可以形式化地表示如下:设有动词集合$V=\{v_1,v_2,\ldots,v_V\}$,名词集合$N=\{n_1,n_2,\ldots,n_N\}$,给定观察数据$O=\{(v,n)|v\in V,n\in N\}$,求解概率模型$p(v,n)$,使它能够解释观察数据$O$.因为这种概率模型的参数数目众多($|N|\cdot|V|$),所以,在参数估计时存在数据稀疏问题.

建立基于等价类的概率模型是解决数据稀疏问题的重要方法.这种方法可以描述为:在N的划分P_N和V的划分P_V之上,对于$cv\in P_V$和$cn\in P_N$,求解概率模型$p(cv,cn)$,进而$\forall\ v\in P_V,\forall\ n\in P_N,p(v,n)=p(cv,cn)$.集合的等价类划分有两种方法:①自动聚类:从训练语料中自动学习词语的等价类划分,这种方法得到的划分能够客观地反映真实文本.但是聚类中同样存在数据稀疏问题,而且聚类算法复杂,因此实用性较差[2,3];②基于义类词典的划分,方法简单,适用性好,但是义类词典是语义分类体系,而VN结构模板不仅与词语的语义特征有关,还与词语的句法特征有关,因此,基于义类词典的划分对于词语结构模板获取是不充分的.[4,5]

本文提出了基于动词复杂特征的VN结构模板获取模型,该模型在对集合进行划分的同时考虑了动词的语法特征和语义特征,优于单纯基于义类词典的模型;与自动聚类方法相比较,该模型充分利用复杂特征集的多种信息来限制模型的求解空间,实用性更强.

1 基于复杂特征的VN结构模板获取模型

1.1 问题定义

一个动词和名词同现是构成VN结构还是~VN结构,既与动词和名词本身的语法和语义特征有关,也与该同现对的上下文环境的语法和语义特征有关.本文将其本身的特征称为静态特征,将上下文环境的特征称为

---

* 本文研究得到国家自然科学基金资助.作者赵军,1967年生,博士生,主要研究领域为自然语言处理,信息检索,语料库语言学.黄昌宁,1937年生,教授,博士生导师,主要研究领域为自然语言处理,信息检索,语料库语言学.

本文通讯联系人:赵军,北京 100084,清华大学计算机科学与技术系

本文1997-08-14收到原稿,1998-01-23收到修改稿

动态特征.本文主要研究任意动词和特定名词n同现时的结构,在实验中只考虑动词的静态特征,它们有:词性子类SUBV(包括及物动词vt,不及物动词vi等),音节数SYL(包括单音节mon,双音节bi等),义类词典《词林》的大类SENSE1($F\sim J$)、中类SENSE2($a\sim n$)和SENSE3小类(01～67)以及动词的词形WORD等.

基于复杂特征的VN结构模板获取模型可描述如下:设有特定名词n,动词集合$V(n)=\{v_1,v_2,\ldots,v_V\}$,给定观察数据$S=VN+\sim VN$,其中$VN=\{(v,n)|v$和$n$构成VN结构,$v\in V,n\in N\}$,$\sim VN=\{(v,n)|v$和$n$不构成VN结构,$v\in V,n\in N\}$,其中动词$v$以复杂特征集的形式表示如下:

$$v=\begin{bmatrix} f_1=x_1 \\ f_2=x_2 \\ \cdots \\ f_n=x_n \end{bmatrix},\qquad n>0,$$

其中f_i为特征名,x_i为特征值.

基于动词复杂特征的VN结构模板获取的中心思想是:①识别与结构相关的特征,并依据这些特征对动词集合$V(n)$进行划分;②基于动词的划分,估计每个等价类中的动词与n同现构成VN和$\sim$VN的概率.其中的关键问题是等价类的划分,对于$V(n)$的一个划分P,应满足以下条件:

①$P\subseteq 2^{V(n)}$;　　②$\bigcup_{cv\in P} cv=V(n)$;　　③$\forall\ cv_i,cv_j\in P,cv_i\cap cv_j=\varnothing$;

④$\forall\ cv\in P$,如果$|cv|>1$,则$\bigcup_{v\in cv} v\neq\varnothing$,其中$\cup$表示复杂特征集的合一运算.

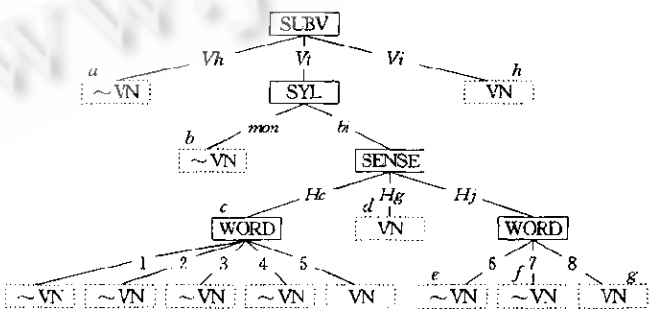
在该划分上建立的概率模型既可以解释例子集,又可以对未观察的动词和名词同现的结构作出精确的判断.本文利用统计决策树SDT(statistical decision tree)模型[6]进行动词等价类的划分,一方面,SDT的表达能力强于N元模型,相当于插值N元模型;另一方面,SDT模型的最大优势在于自动抽取相关特征的能力.

1.2　用SDT模型生长动词分类树

1.2.1　用SDT表示动词分类树

SDT是一个决策机制,它根据一系列特征,赋予每一种可能的选择一个概率值$p(f|h)$,其中h表示一系列特征,f为当前作出的选择,概率值$P(f|h)$由前n个特征提问序列$q_1,q_2,\ldots,q_n$来决定(其中第i个特征提问仅与前$i-1$个特征提问有关).

图1是用SDT表示的一棵动词分类树,它描述了一个具有某些特征的动词与名词"成绩"同现时构成VN结构或$\sim$VN结构的概率.其中,内部结点是提问结点,一个提问结点表示对一个特征的提问,从该结点延伸的树枝代表该特征可能的取值;叶结点是选择结点,表示符合从根结点到该结点的路径上所有"特征—值"的动词与名词"成绩"同现时是构成VN结构还是$\sim$VN结构.所提问的特征有:动词的子类SUBV、动词的音节数SYL、动词的义类SENSE等.例如,结点8表示$SUBV=vt$,$SYL=bi$且$SENSE=Hg$的动词与名词"成绩"同现构成VN结构.



数字1～8分别代表下列词:登记、公布、取消、肯定、考核、发扬、计算、平均.

图1　统计决策树例图

1.2.2 基于极大似然估计 MLE(maximum likelihood estimation)原则的动词分类树的生长算法

动词分类树生长算法的关键是每个提问结点所提问的特征的选择问题,本文利用基于信息增益的特征来选择方法[6]生长动词分类树.设 X 是由任意动词与特定名词 n 的同现对构成的训练集,$X=\{(v,n,c,m)|v\in V,c\in C\}$,其中 $C=\{VN,\sim VN\}$ 是分类集,V 是动词集合,m 是 v 和 n 的同现次数;设动词特征集为 $A=\{A_1,\ldots,A_p\}$,特征 A_k 的取值的集合 $V_k=\{v_{k1},\ldots,v_{kn}\}$,则递归地生长关于特定名词 n 的动词分类树 T 的算法描述如下:

① 建立动词分类树 T 的根结点 *root*,将训练集 X 与 *root* 相关联;

② 设当前结点为 $node_i$,与 $node_i$ 关联的训练集为 X_i,如果对于任意 $(v,n,c,m)\in X_i$,有 $c=VN$ 或 $c=\sim VN$,则确定为叶结点,返回;

③ 对 $A_i\in A$,分别计算

- 熵 $H(X_i)=-\sum_{c\in C}P_c\log p_c$;
- 条件熵 $H(X_i|A_k)=-\sum_{v\in V_k}\sum_{c\in C}P(c|A_k=v)\log p(c\ \ A_k=v)$;

④ 计算对 A_k 提问的信息增益:$IG(A_k,X_i)=\frac{1}{IV(A_k)}(H(X_i)-H(X_i|A_k))$,其中 $IV(A_k)$ 是为了避免选择具有较多取值的特征的倾向所加的系数,表示为 $IV(A_k)=-\sum_{j=1}^{n}\frac{|X_{ij}|}{|X_i|}\log\frac{|X_{ij}|}{|X_i|}$,其中 $|X_i$ 是与 $node_i$ 相关联的训练集 X_i 中的例子数,$|X_{ij}|$ 是训练集 X_i 中符合条件 $A_k=v_j$ 的例子数;

⑤ 确定具有最大信息增益的特征 $A_m=\mathrm{argmax}_{A_k}IG(A_k,X_i)$;

⑥ 依据特征 A_m 的取值的集合 $V_m=\{v_{m1},\ldots,v_{mn}\}$ 生长结点 $node_i$ 的儿子结点 $node_{i1},\ldots,node_{in}$,并将训练集 X_i 划分为 n 个子集 $X_{i1},\ldots,X_{in}$,分别将 $X_{i1},\ldots,X_{in}$ 与 $node_{i1},\ldots,node_{in}$ 相关联;

⑦ 从特征集合 A 中删除特征 A_m;

⑧ 对于结点 $node_{i1},\ldots,node_{in}$,分别执行②~⑦,进行儿子结点的生长和训练集的划分.

这样,一棵基于复杂特征的动词分类树就生成了,所有的叶结点构成动词集合 $V(n)$ 的最优划分,其中每个叶结点所表示的复杂特征集由从树根到该结点的所有结点的复杂特征组成.

1.3 用最小描述长度 MDL(minimum description length)原则搜索动词分类树的最优划分

1.3.1 基于 MLE 原则的语言获取模型和基于 MDL 原则的语言获取模型

以上讨论的动词分类树的划分和基于划分的概率模型的建造是语言获取的问题.基于 MLE 原则的语言获取模型为

$$M=\mathrm{argmax}_M p(O|M).$$

其挑选模型的标准是模型与训练数据的拟合性,即模型 M 要最大可能地解释训练集 O,而通常情况下,提供给学习者的数据只是目标语言的一小部分,于是依据 MLE 原则获取的语法虽然能够很好地解释训练集中的数据,但是对训练集以外的数据的解释能力很弱,这就是语言获取中的过度适合(Overfitting)问题,即对训练数据的不规则性和特异性过分敏感,缺乏归纳能力.而从已知的观察数据归纳出既可以解释已知数据又可以解释未知数据的语法是语言获取中的关键问题.例如,在图 1 所示的基于 MLE 原则的生长的 SDT 中,其最优划分是所有叶结点构成的划分,最优的概率模型是建立在最优划分上的模型,可以看出模型的概括能力很弱,无法判断训练集以外的同现"否定/Hc 成绩"的结构.这种过度适合的问题,使得开放测试中 VN 结构识别的召回率不高.

而贝叶斯[7]的语言获取模型为

$$M=\mathrm{argmax}_M p(M|O)=\mathrm{argmax}_M\{p(O|M)\times p(M)\}.$$

与基于 MLE 原则的语言获取模型相比,贝叶斯语言获取模型除了考虑模型和训练数据的拟合性以外,还考虑了模型 M 的先验概率 $p(M)$.本文依据最小描述长度 MDL 原则[8]来定义模型的评价函数,即对于给定的观察数据的最好的概率模型是具有最短描述长度的模型,其中描述长度由以下两部分组成:①模型描述长度 $l(G)$,即模型的编码长度;②数据描述长度 $l(O|M)$,即将模型作为数据的预测时,数据的编码长度.本文将在 1.3.2 节中具体定义 $l(M)$ 和 $l(O|M)$,这里先定义 $p(M)=2^{-l(M)}$.于是 $P(M)$ 给简单的语法赋予高的概率,这与 Occam's

Razor 的直观意义相符,即简单的语法优于复杂的语法.[7]另一方面,基于 MDL 原则的语言获取模型又超越了 Occam's Razor,即搜索使 $p(o|m)\times p(m)$达到最大的模型 M,其中 $P(M)$倾向于简单的模型,而 $P(O|M)$倾向于与训练数据拟合性好的模型.MDL 原则就是要在数据拟合程度和模型复杂度之间找到一个鞍点.

1.3.2　基于 **MDL** 原则的动词分类树的最优划分算法

给定一棵动词分类树,可以得到动词集合 V 的若干个划分.例如,对于图 2 的动词分类树,可以得到 V 的以下划分:

- 全集 V,
- $\{[SUBV=Vi],[SUBV=Vt]\}$,
- $\left\{[SUBV=Vi],\begin{bmatrix}SUBV=Vt\\SENSE=H\end{bmatrix},\begin{bmatrix}SUBV=Vt\\SENSE=I\end{bmatrix}\right\}$,
- $\left\{\begin{bmatrix}SUBV=Vi\\SENSE=H\end{bmatrix},[SUBV=Vt],\begin{bmatrix}SUBV=Vi\\SENSE=I\end{bmatrix}\right\}$,
- $\left\{\begin{bmatrix}SUBV=Vi\\SENSE=H\end{bmatrix},\begin{bmatrix}SUBV=Vi\\SENSE=I\end{bmatrix},\begin{bmatrix}SUBV-Vt\\SENSE-H\end{bmatrix},\begin{bmatrix}SUBV=Vt\\SENSE=I\end{bmatrix}\right\}$.

因为建立在动词分类树的叶结点组成的划分之上的模型与训练数据的拟合性最好,因此,基于 MLE 的模型认为这种划分是最优划分;而 MDL 原则认为,最优划分的评判目标应该是在模型复杂度和数据拟合度上的整体评分最好.

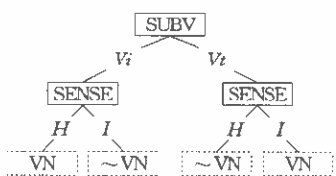


图2　动词分类树例图

对于特定名词 n、动词集合 $V(n)$和观察数据 $S=VN+\sim VN$,设 $V(n)$的候选划分集合 $\Omega=\{P_1,P_2,\ldots,P_S\}$,其中每个 $P_i\in\Omega(1\leqslant i\leqslant s)$都对应一个基于类的概率模型 M_i,目标是在候选概率模型中搜索最优的模型.本文用 MDL 原则构造模型的评价函数,描述如下:

① 对于每个候选划分 $p_i=\{c_1,c_2,\ldots,c_r\}$,构造相应的概率模型 $M_i=\{p(v,n)|v\in V(n)\}$,其中$\forall\ v_i,v_j\in c_k$,$p(v_i,n)=p(v_j,n)=p(c_k,n)$,而 $p_{VN}(c_k,n)=\sum_{v\in c_k}\sum_{(v,n)\in VN}f(v,n)\Big/\sum_{v\in c_k}\sum_{(v,n)\in S}f(v,n)$,其中 $f(v,n)$表示(v,n)同现的频度.

② 计算每个概率模型的数据描述长度:对于任意的动词和名词的同现(v,n),它的结构可由随机变量 $X=\begin{cases}1,(v,n)\in VN\\0,(v,n)\in\sim VN\end{cases}$表示,因此 X 服从两点分布 $P\{X=x\}=\begin{cases}1-p, & x=0\\p, & x=1\end{cases}$.简化上式得 $P\{X=x\}=p^x(1-p)^{1-x},x=0,1$.设 $x_1,x_2,\ldots,x_n$ 是随机变量 X 的容量为 n 的观察样本值,由于样本中 $X_i(i=1,2,\ldots,n)$相互独立,所以观察值 $x_1,x_2,\ldots,x_n$ 出现的概率是

$$L=P\{X_1=x_1,X_2=x_2,\ldots,X_n=x_n\}=\prod_{i=1}^{n}P\{X_i=x_i\}=\prod_{i=1}^{n}(p_i)^{x_i}(1-p_i)^{1-x_i}.$$

定义似然函数为

$$\log L=\log\prod_{i=1}^{n}(p_i)^{x_i}(1-p_i)^{1-x_i}=\sum_{i=1}^{n}x_i\log p_i+\sum_{i=1}^{n}(1-x_i)\log(1-p_i),$$

本文将$-\log L$ 作为数据描述长度.

③ 计算每个概率模型的模型描述长度:在本问题中模型描述长度由两部分组成,即划分描述长度 L_{par}和概率描述长度 L_{pro}.设动词集合 $V(n)$的候选划分集合为 Ω,则划分描述长度为 $L_{\text{par}}(V(n))=\log|\Omega|$;概率描述长度的计算方法如下:因为 MLE 的标准差为$O(1/\sqrt{|S|})$,每个标准差的编码长度为$O(|\log(1/\sqrt{|S|}|)=O(|-\log(1/\sqrt{|S|}|)=O(\log|S|/2)$,于是每个自由参数的编码长度为$O(C+\log|S|/2)=O(\log|S|/2)$,本文定义概率描述长度为 $L_{\text{par}}=(|\Omega|/2)\times\log|S|$.

④ 用 MDL 原则挑选最优划分和最优模型:$M_{\text{opt}}=\text{argmin}_M\{L_{\text{dat}}(M)+L_{\text{par}}(M)+L_{\text{pro}}(M)\}$,因为$\forall\ M_i$ 和 M_j,$L_{\text{par}}(M_i)=L_{\text{par}}(M_j)$,因此,$M_{\text{opt}}=\text{argmin}_M\{L_{\text{dat}}(M)+L_{\text{pro}}(M)\}$.

给定一棵动词分类树 T，$node_i$ 是其中的一个结点，与结点 $node_i$ 对应的训练集表示为 X_i，将与结点 $node_i$ 对应的复杂特征集表示为 $[node_i]$，$\{[node_i]\}$ 构成 X_i 的一个划分，建立在划分 $\{[node_i]\}$ 上的概率模型为 M_i. 基于 MDL 原则搜索其最优划分 opt 的递归算法如下：

① 将当前结点设为根结点；

② 设当前结点为 $node$，如果 $node$ 是叶结点，则返回 $[node]$；

③ 否则，对于 $node$ 的每个子女结点 $child_i$，递归地搜索与它相关联的动词子集 C_i 的最优划分 P_i，令 $P=\sum_i P_i$，构造基于划分 P 的概率模型为 M，如果模型 M_i 的描述长度小于模型 M 的描述长度，即 $L(M_i)<L(M)$，则返回划分 $\{[node_i]\}$，否则返回划分 P.

利用以上的算法可以在一棵动词分类树上搜索到一个动词集合的最优划分. 例如，图 1 所示动词分类树的最优划分为结点 $a\sim h$ 构成的划分.

1.4 由动词分类树推导 VN 结构模板

在生长动词分类树并搜索它的最优划分以后，由经过剪枝的动词分类树可以容易地推导出 VN 结构模板. 例如，在图 1 中，与最优划分对应的结构模板为：

①$[SUBV=Vh]$+成绩 $\xrightarrow{1.0}\sim VN$，

②$[SUBV-Vi]$+成绩 $\xrightarrow{1.0}VN$，

③$\begin{bmatrix}SUBV=Vt\\SYL=mon\end{bmatrix}$+成绩 $\xrightarrow{1.0}\sim VN$，

④$\begin{bmatrix}SUBV=Vt\\SYL=bi\\SENSE=Hc\end{bmatrix}$+成绩 $\xrightarrow{0.907}\sim VN$，

⑤$\begin{bmatrix}SUBV=Vt\\SYL=bi\\SENSE=Hg\end{bmatrix}$+成绩 $\xrightarrow{1.0}VN$，

⑥$\begin{bmatrix}SUBV=Vt\\SYL=bi\\SENSE=Hj\\WORD=发扬\end{bmatrix}$+成绩 $\xrightarrow{1.0}\sim VN$，

⑦$\begin{bmatrix}SUBV=Vt\\SYL=bi\\SENSE=Hj\\WORD=计算\end{bmatrix}$+成绩 $\xrightarrow{1.0}\sim VN$，

⑧$\begin{bmatrix}SUBV=Vt\\SYL=bi\\SENSE=Hj\\WORD=平均\end{bmatrix}$+成绩 $\xrightarrow{1.0}VN$.

2 基于结构模板的 V+N 型短语的结构识别

基于结构模板的 V+N 型短语的结构识别可以描述为这样一个问题：给定一个同现 (v,n)，其中 v 是一个特定的动词，n 是一个特定的名词，判断它是 VN 结构还是～VN 结构. 可利用的资源有：动词词性词典、动词义类词典、由经过剪枝的动词分类树推导的结构模板. VN 结构识别算法描述如下：

① 对动词 v 标注动词的分类和义类，并建立其复杂特征集的向量表示 $Q_1,\ldots,Q_m$（因为义类歧义没有完全排除，因此可能有多个向量）；

② 分别计算 $Q_1,\ldots,Q_m$ 与 n 同现时构成 VN 结构的概率 $p_{VN}(Q_i,n)=p_{SDT}^{VN}(Q_i,n)+T_V^{VN}\times T_N^{VN}$ 和构成～VN 结构的概率 $p_{\sim VN}(Q_i,n)=p_{SDT}^{\sim VN}(Q_i,n)+T_V^{\sim VN}\times T_N^{\sim VN}$，其中 $p_{SDT}^{VN}(Q_i,n)$ 是以 SDT 结构模板判断的 Q_i 和 n 同现时构成 VN 结构的概率；$p_{SDT}^{\sim VN}(Q_i,n)$ 是以 SDT 结构模板判断的 Q_i 和 n 同现时构成～VN 结构的概率；T_V^{VN} 是动词 v 在 VN 结构中出现的概率；T_n^{VN} 是名词 n 在 VN 结构中出现的概率；$T_V^{\sim VN}$ 是动词 v 在～VN 结构中出现的概率；$T_n^{\sim VN}$ 是名词 n 在～VN 结构中出现的概率；

③ 计算 $k=\text{argmax}_i p_{VN}(Q_i,n)$ 和 $l=\text{argmax}_i p_{\sim VN}(Q_i,n)$；

④ 比较 $p_{VN}(Q_k,n)$ 和 $p_{\sim VN}(Q_l,n)$，

- 如果 $p_{VN}(Q_k,n)>p_{\sim VN}(Q_l,n)$，则同现 (v,n) 的结构为 VN，可信度为 $p_{VN}(Q_k,n)$；
- 如果 $p_{VN}(Q_k,n)<p_{\sim VN}(Q_l,n)$，则同现 (v,n) 的结构为～VN，可信度为 $p_{\sim VN}(Q_l,n)$.

3　模型分析和测试结果

3.1　模型分析

分别从训练集和测试集(训练集和测试集的建造见 3.2 节)中抽出动词与 10 个名词“办法”、“标准”、“产品”、“成绩”、“贷款”、“单位”、“干部”、“公司”、“过程”、“合同”同现的数据作为训练集和测试集，测试各种模型的性能. 测试指标有：①收敛性：比较模型在不同的训练数据时获得的结构模板的变化数(增加的模板数和减少的模板数之和)，变化数越小，收敛性越好；②模板数：比较模型在不同的训练数据时获得的结构模板数，模板数越小，模型越简单；③精确率：$p=\frac{a}{b}\times 100\%$；④召回率：$r=\frac{a}{c}\times 100\%$；其中 a 是能判断 VN/～VN 且判断正确的同现次数，b 是能判断 VN/～VN 的同现次数，c 是总同现次数.

3.1.1　基于 MDL 的模型和基于 MLE 的模型的比较

通过比较基于复杂特征和 MDL 的模型与基于复杂特征和 MLE 的模型的各种指标，比较基于 MDL 的模型和基于 MLE 的模型的性能. 结果见图 3～6.

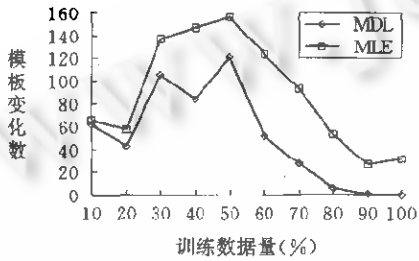


图3　MDL和MLE收敛性比较

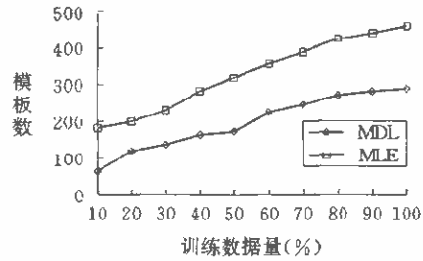


图4　MDL和MLE模板数比较

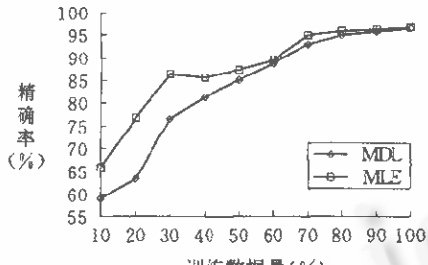


图5　MDL和MLE精确率比较

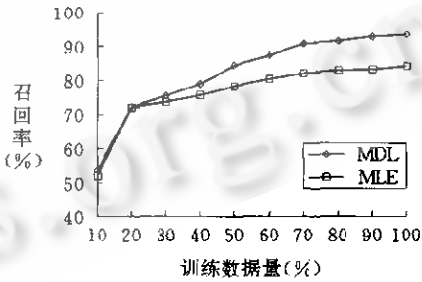


图6　MDL和MLE召回率比较

通过以上实验可以得出如下结论：① 基于 MDL 的模型比基于 MLE 的模型收敛速度快；② 基于 MDL 的模型比基于 MLE 的模型简单；③ 虽然基于 MDL 的模型的精确率比不上基于 MLE 的模型，但随着训练数据的增加，基于 MDL 的模型的精确率逐渐逼近基于 MLE 的模型；④ 基于 MDL 的模型的召回率优于基于 MLE 的模型，主要体现在：由于基于 MDL 的模型的概括能力优于基于 MLE 的模型，使得该模型对于未观察的同现的处理优于基于 MLE 的模型，对于标不上义类的同现的处理优于基于 MLE 的模型.

3.1.2　基于复杂特征的模型和基于义类的模型的比较

通过比较基于复杂特征和 MDL 的模型和基于义类和 MDL 的模型的各种指标，比较基于复杂特征的模型和基于义类的模型的性能. 结果见图 7～10.

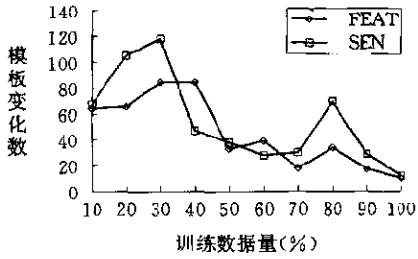


图7　复杂特征模型和义类模型收敛性比较

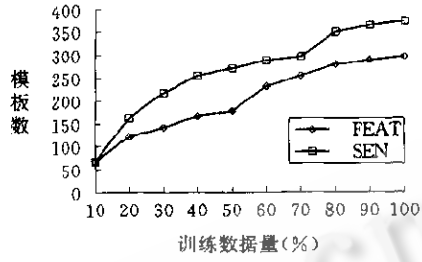


图8　复杂特征模型和义类模型模板数比较

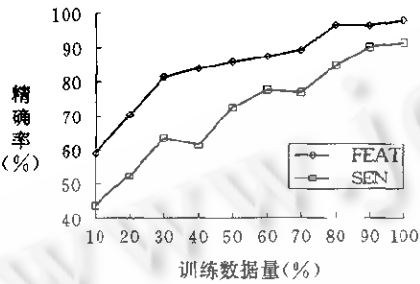


图9　复杂特征模型和义类模型精确率比较

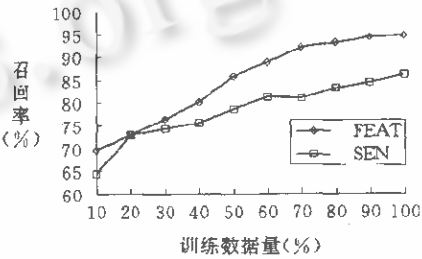


图10　复杂特征模型和义类模型召回率比较

通过以上实验可以得出如下结论：① 基于复杂特征的模型收敛速度比基于义类的模型稍快；② 基于复杂特征的模型比基于义类的模型简单；③ 基于复杂特征的模型比基于义类模型的精确率高；④ 基于复杂特征的模型的召回率优于基于义类的模型.

3.2　测试结果

从70兆字节的新华社语料中抽取(v,n)同现61 324对(238 694次)，涉及5 975个动词和6 568个名词，其中包括VN结构14 081对(38 075次)，涉及2 946个动词和931个名词.对所有的(v,n)同现标注词性标记、义类标记和音节数标记，从而构成训练集.

从训练集中随机抽取3 099对(v,n)同现，构成封闭测试集，涉及1 095个动词和1 468个名词.从训练集外的10万字的语料中抽取(v,n)同现3 165对，标注词性标记、义类标记和音节数标记，构成开放测试集，涉及1 173个动词和1 574个名词.

利用VN结构模板判断封闭测试集和开放测试集中(v,n)同现的结构，测试结果如表1所示.

表1　(%)

	精确率	召回率
封闭测试	97.8	96.3
开放测试	94.1	90.3

3.3　对实验中几个问题的说明

3.3.1　《词林》收词不足的问题

在识别阶段，存在着由于《词林》收词不足引起的某些动词没有义类代码的问题.在缺少义类代码的情况下，基于复杂特征的模型可能利用词性信息和音节数信息.例如：动词"造林"的词性标记为vi，音节数为2，但在《词林》中没有义类代码，在识别"造林成绩"的结构时，由2.1节中模板②判断它为VN结构.实验表明，开放测试中，对于动词没有义类的(v,n)同现，基于复杂特征的模型的识别精确率为85.3%，召回率为66.7%.

3.3.2　义类兼类问题

对训练集和测试集中(v,n)同现数据中的义类兼类动词，虽然利用词性标记进行了排歧，但仍然存在义类兼

类问题.本文采取的办法是保留歧义.实验表明,这些义类歧义对于训练和识别的影响都不大.原因是:基于复杂特征和 MDL 原则的模型得到的结构模板都具有共同的特征.例如,"研究(Gb01/Hg14)成绩"中义类 Hg 是成组出现的,即"教育(Hg)、训练(Hg)、学习(Hg)、创作(Hg)",而义类 Gb 是个别出现的.在开放测试中,对于动词义类兼类的(v,n)同现,基于复杂特征和 MDL 原则的模型的识别精确率为 89.5%,召回率为 95.3%;而基于义类和 MLE 原则的模型的识别精确率为 82.5%,召回率为 88.4%.

4　结束语

本文提出了基于复杂特征和 MDL 原则的 VN 结构模板获取模型.实验表明,在利用结构模板进行 VN 结构的识别时,这种模型比基于义类和极大似然估计原则的模型具有更高的精确率和召回率.但还有以下两方面的不足:①对于 VN 结构的识别,除了应该考虑其组成成分的句法和语义特征外,还应该考虑它出现的上下文环境.例如,在句子"这个分析系统性能可靠"中(v,n)同现"分析系统"构成 VN 结构,而在句子"这个实验用于分析系统的性能"中"分析系统"构成～VN 结构.虽然本文考虑的主要是它的内部组成成分,但是基于复杂特征和 MDL 原则的 VN 结构模板获取模型是通用的.如果将上下文环境特征加入复杂特征集中,这种模型则可以同时考虑 VN 结构的上下文环境;②虽然基于复杂特征和 MDL 原则的模型的鲁棒性优于基于义类和 MLE 原则的模型,《词林》收词不足的问题仍然是影响 VN 结构的识别精确率和召回率的主要因素.我们正在尝试用分布相似的方法解决这个问题.

参考文献

1　梅家驹等.同义词词林.上海:上海辞书出版社,1983
(Mei Jia-ju *et al*. Tongyici Cilin. Shanghai: Shanghai Dictionary Press, 1983)

2　Brown P F *et al*. Class-based n-gram Models of natural language. Computational Linguistics, 1992,18(4):467～479

3　Dagan I *et al*. Contextual word similarity and estimation from sparse data. Computer Speech and Language, 1995,9(2): 123～152

4　Li Hang *et al*. Clustering words with the MDL principle. In: Proceedings of the 16th International Conference on Computational Linguistics. Copenhagen, Denmark: the Association for International Computational Linguistics. 1996

5　Pereira F *et al*. Distributional clustering of English words. In: Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics. Ohio, USA: Association for Computational Linguistics, 1993. 183～190

6　Magerman S F. Natural language parsing as statistical pattern recognition [Ph. D. Dissertation]. Stanford: Stanford University, 1994

7　Stanley F Chen. Building probabilistic models for natural language [Ph. D. Dissertation]. Cambridge, Massachusetts: Harvard University, 1996

8　Quinlan J R. Inferring decision trees using the minimum description length principle. Information and Computation, 1989,80(2):227～148

The Complex-feature-based Model for Acquisition of VN-construction Structure Templates

ZHAO Jun　HUANG Chang-ning

(*Department of Computer Science and Technology　Tsinghua University　Beijing*　100084)
(*State Key Laboratory of Intelligent Technology and Systems　Tsinghua University　Beijing*　100084)

Abstract　In this paper, a complex-feature- and MDL-based model for acquisition of VN-construction structure templates is put forward. First, a verb classification tree is created using statistical decision tree model. Then, the tree is pruned based on MDL (minimum description length) principle. Finally, structure templates are derived based on the verb classification tree. The experiments show that using the structure templates acquired with the model to recognizing VN-structure, the system has its advantages over the model based on the sense and the MLE (maximum likelihood estimation) principle in precision and recall.

Key words　Natural language processing, corpus, complex feature, statistical decision tree, minimum description length principle.