文章编号:1001 - 9081(2009)09 - 2571 - 04

贝叶斯决策树在英文现在分词词性识别中的应用

徐 哲,刘 循
(四川大学 计算机学院,成都 610064)
(jaioros@163.com)

摘 要:针对英文现在分词词性标注这一特定问题存在的难点分析了隐马尔可夫模型(HMM)的不足,提出了贝叶斯决策树模型。对一个已经标注好的语料库进行统计,运用决策树 C4.5 算法从单边条件和双边条件两个方面对英语现在分词的三种词性进行合理的分类消歧。对于双边条件下仍然存在歧义的情况,用贝叶斯最小风险对决策树改进,用标注好的语料库对模型进行训练。最后,采用一个未经过标注的语料库进行测试,取得了非常好的效果,证明了模型的优越性。

关键词:分类;消歧;贝叶斯决策树;隐马尔可夫模型
中图分类号: TP391.1 **文献标志码**:A

Application of Bayesian decision tree to recognition of English present participle

XU Zhe, LIU Xun
(*College of Computer Science, Sichuan University, Chengdu Sichuan* 610064, *China*)

Abstract: Concerning the difficulties in part-of-speech tagging in English present participle, the authors analyzed the drawbacks of Hidden Markov Models (HMM) and proposed Bayesian decision tree model. Firstly, the tagged corpus was calculated and C4.5 in decision tree was used for proper classification and disambiguation of the three classes of present participle. Then, the decision tree was improved by Bayesian least risk. At last, an untagged corpus was used to test the model and the result is very good, which proves the superiority of the model.

Key words: classification; disambiguation; Bayesian decision tree; Hidden Markov Model (HMM)

0 引言

英文文章中会大量出现一类带有 ing 后缀的分词,这类词叫作现在分词,现在分词在英文文章中出现的频率较高,词性变化较大,位置灵活,对这一类词性的标注比较困难,主要难点如下:

1)稀疏固定搭配。

例句:It is no good hoping that he can make it by himself.
You need to practice writing articles.

2)连续分词。

例句:It's an interesting reading training center.

3)逗号分隔的动名词。

例句:The student must be trained adequately in all 4 skills: understanding, speaking, reading and writing.

根据词性标注的研究现状,可用于解决这类问题的主要方法是决策树分类器[4]、二元语法、三元语法[9]和隐马尔可夫模型(Hidden Markov Model, HMM)[5],还有贝叶斯分类器[6]和人工神经网络等[8]。

上述几类方法中,决策树和贝叶斯分类器各自很难处理现在分词词性复杂的语法规则。人工神经网络运算量大,不利于大数据量文本处理。二元语法、三元语法都是采用最大似然估计的思想,均会受到稀疏数据和逗号分隔问题而降低正确率。HMM 正确率相对较高,但也基于最大似然性的状态表序列,根据有限视野和时间不变性从许多可能的序列中找出一种最可能出现的标记序列作为文本的词性标注。但却很难正确标记连续分词这种稀疏分布的情况,也很难处理逗号分隔的情况,且运算量也很大。

为了避免上述模型的不足,结合模型各自的优点,本文提出了基于贝叶斯的决策树 C4.5 算法模型。对于以上三个难点,决策树 C4.5 算法具有多级分类,分层过滤,运算量小的特点。结合贝叶斯可以降低 HMM 中最大似然估计的较大误差。因此,贝叶斯决策树模型从运行速度还是正确率上,对现在分词词性标注上具有优越性。

决策树 C4.5 分类是根据"属性_值"对,每个实例选择 3 个属性:单边属性 p_-1 表示前一个词的词性,p_+1 为后一个词的词性;双边属性 $p\_\pm1$ 表示前后的词性;w_0 表示需要标注的现在分词。

括号内的符号表示现在分词的正确词性。例句中,hoping, writing 的正确标记是 VBG1—动词现在分词,interesting 的正确标记是 JJ—形容词。Reading, training 的正确标记是 VBG2—动名词。第 3 类难点的例句中四个均为 VBG2—动名词。除此之外,在所有词性的标记是 BEZ—是动词,VB、VBD—动词或动词现在式,NN—名词,IN—介词,RB—副词,TO—不定式 to,PER—标点符号,CC—连词,AT—冠词,JJR—形容词比较级。可以分析 w_0 前面一个词的词性 p_-1,或者后面一个词的词性 p_+1,或者前后相邻两边的词

收稿日期:2009 - 03 - 19;**修回日期**:2009 - 05 - 19。 **基金项目**:国家自然科学基金资助项目(60773169)。
作者简介:徐哲(1984 -),男,河北磁县人,硕士研究生,主要研究方向:计算机智能; 刘循(1963 -),女,四川自贡人,副教授,博士,主要研究方向:计算机智能。

性 $p\_\pm1$，遍历语料库查看在前、后的单边词性条件或者前后双边条件下现在分词可能出现的类型。对 $p\_\pm1$ 不定的情况，结合贝叶斯最小风险处理。

本文结合决策树 C4.5 和贝叶斯，用一个含有 160 000 英文单词的语料库来进行先验统计和训练模型，然后用一个含有 100 000 英文单词的语料库来测试模型分类的正确率。训练和测试语料库均来源于普特英语论坛他人提供。

1 决策树的词性标注 C4.5 算法

1.1 不同标记类别的词分布相对比例

我们统计整个语料库，发现现在分词词性只有 3 种可能：VBG1，VBG2，JJ；因此，使用决策树对所有现在分词进行三种分类，对于三类问题，定义它们的相对比例：

$$p_c(X_i)=\frac{f_c(X_i)\times\prod_{j=1}^{3}f(X_j)}{\sum_{h=1}^{3}[f_c(X_h)\prod_{k=1}^{3}f(X_k)]};\ i\neq j,h\neq k \tag{1}$$

属性值 $X_i(i=1,2,3)$ 分别为 VBG1，VBG2，JJ 三类属性标记，$f(X_i)$ 表示 X_i 在整个语料库中出现的次数，$f_c(X_i)$ 表示在当前上下文，即当前 p_-1，p_+1 中，X_i 出现的次数，$p_c(X_i)$ 表示在整个语料库中，在当前上下文环境 C 中标注为 X_i 的相对比例。

1.2 上下文 C 实例分布的熵

定义三类问题实例分布的熵值：

$$E(S_C)=-\sum_{i=1}^{3}P_C(X_i)\mathrm{lb}\,P_C(X_i) \tag{2}$$

以实例分布的熵值来定量描述当前上下文 C 带来的信息量或者信息纯度，纯度为 0 则最高，信息量最小。信息量包含 $X_i(i=1,2,3)$ 的分布，若 C 可以出现的不同分词词性种类越多，纯度越低，信息量越大。

显然，当某一个 X_i 的相对比例 $p_c(X_i)$ 为 1，$E(S_C)=0$。说明当前 C 下只有一种词性存在，此时信息量最小，纯度最高，最利于决策。

1.3 分类函数——属性信息熵期望值

属性信息熵期望值：

$$T(S,A)=\sum_{C\in Values(A)}\frac{|S_C|}{|S|}E(S_C) \tag{3}$$

$Values(A)$ 表示属性 A 的所有上下文 C（包括 IN，NN 等）的集合。$|S_C|$ 表示当前属性 A（A 包括 p_-1，p_+1，$p\_\pm1$）下，每一类上下文 C（例如 $p\_+1=$ IN）在语料库中出现的次数。$|S|$ 表示语料库中所有分词的总数。

$T(S,A)$ 表示属性 A 的总体上下文 C 信息熵期望值，$T(S,A)$ 越小，越利于分类。

2 决策树模型训练

2.1 单边条件和双边条件分类计算

以 p_+1 为例，对含 160 000 英文单词的语料库进行统计、训练模型，得到三类分词词性分布 $f(\mathrm{VBG1})=2\,352$，$f(\mathrm{VBG2})=478$，$f(\mathrm{JJ})=463$，统计 VBG1，VBG2，JJ 三类在上下文环境 C 中出现的次数 $f_c(X_i)$，得到下列分布矩阵如表 1 所示。

表 1 三类在 $C=p\_+1$ 下的出现次数

上下文 C	VBG1 $f_C(X_1)$	VBG2 $f_C(X_2)$	JJ $f_C(X_3)$
$p\_+1=$ IN	562	31	0
$p\_+1=$ NN	284	0	362
$p\_+1=$ TO	203	28	0
$p\_+1=$ PN	271	0	0
$p\_+1=$ AT	458	0	0
$p\_+1=$ JJ	133	0	0
$p\_+1=$ PER	139	180	48
$p\_+1=$ CC	58	60	53
$p\_+1=$ BEZ, VB, VBD	10	7	0
$p\_+1=$ RB	244	0	0

根据式(1)，得到三类在当前上下文中的相对概率比例矩阵如表 2。

表 2 三类在 $C=p\_+1$ 下的出现概率比例

上下文 C	VBG1 $p_C(X_1)$	VBG2 $p_C(X_2)$	JJ $p_C(X_3)$
$p\_+1=$ IN	0.7865	0.2134	0
$p\_+1=$ NN	0.1337	0	0.8662
$p\_+1=$ TO	0.5957	0.4042	0
$p\_+1=$ PN	1.0000	0	0
$p\_+1=$ AT	1.0000	0	0
$p\_+1=$ JJ	1.0000	0	0
$p\_+1=$ PER	0.1095	0.6982	0.1922
$p\_+1=$ CC	0.0931	0.4742	0.4325
$p\_+1=$ BEZ, VB, VBD	0	1.0000	0
$p\_+1=$ RB	1.0000	0	0

根据式(2)，实例分布的熵的数据如表 3 所示。

表 3 实例分布的熵

C	$E(S_C)$	C	$E(S_C)$
IN	0.5185	JJ	0
NN	0.3935	PER	0.8101
TO	0.6747	CC	0.9374
PN	0	BEZ, VB, VBD	0
AT	0	RB	0

结果为 0 的上下文 C 最利于决策依据，计算 $|S_C|$。

表 4 熵期望值

上下文	$\lvert S_C\rvert$	上下文	$\lvert S_C\rvert$
BEZ, VB, VBD	1113	\(无单词)	138
IN	460	PER	279
CC	250	PN	17
NN	427	AT	211
RB	114	JJR	126
JJ	158		

又因为，$|S|=f(\mathrm{VBG1})+f(\mathrm{VBG2})+f(\mathrm{JJ})=3293$。

利用式(3)，计算出 $T(S,p\_+1)=0.6777$。

同理，可以计算出：$T(S,p\_-1)=0.8982$，$T(S,p\_\pm1)=2.357$。显然 p_+1 最利于分类，做决策树第一个分类依据。

对语料库学习，生成以下双边条件和单边条件的词性标注规则。

规则 1 双边 p_-1，p_+1 是 IN，AT；IN，JJ；CC，AT；CC，IN；CC，PN；RB，NN；RB，IN；PER，IN；PN，PER，w_0 是

VBG1。

规则 2　双边 p_-1,p_+1 是 NN, \; PN,NN; AT,IN, w_0 是 VBG2。

规则 3　双边 p_-1,p_+1 是 AT,NN; \,IN; RB,PER, w_0 是 JJ。

规则 4　单边 p_+1 是 BEZ,VB, VBD, w_0 是 VBG1。

规则 5　单边 p_-1 是 BEZ, VB, VBD, w_0 是 VBG2。

规则 6　单边 p_-1 是 JJR,AT,w_0 是 JJ。

2.2　贝叶斯最小风险改进决策树

然而一些分词词性不能被单边双边条件确定。例如: p_-1 是 IN, p_+1 是 IN, w_0 可能是 VBG1, VBG2;p_-1 是 CC, p_+1 是 NN, w_0 可能是 VBG1, JJ;若进一步增加判定条件,会使得决策树变得臃肿。若使用最大似然估计法,那么在许多特殊的语境下,会因为估计粗糙而降低正确率。因此,我们使用贝叶斯最小风险估计法。经过对语料库的统计,这种情况基本上只存在两类歧义,于是我们可以采用两类贝叶斯最小风险的方法。

风险系数是对歧义词做决策 α_1 和 α_2 相对于正确的词性造成的损失。损失就是标注词性造成的错误率。于是两种词性和决策构成了一个风险函数空间 $\lambda_{ij}, i,j=1,2$。对于风险的计算,若决策正确,没有任何错误率,所以主对角线上的 $\lambda_{ij}=0$,若决策错误,则对风险的计算方法如下:

统计语料库中特定双边条件下,词性 w_1 出现的次数 $f_c(w_1)$ 和 w_2 出现的次数 $f_c(w_2)$,结合词性 w_1 和 w_2 在语料库中出现的次数 $f(w_1)$,$f(w_2)$,用式(4) 作加权处理,得到的风险系数如表 5。

表 5　风险系数

决策	类别	
	w_1	w_2
α_1	$\lambda_{11}=0$	λ_{12}
α_2	λ_{21}	$\lambda_{22}=0$

风险系数为:

$$\lambda_{12}=\frac{f_c(w_2)\times f(w_2)}{f_c(w_1)\times f(w_1)+f_c(w_2)\times f(w_2)}\times 10$$
$$\lambda_{21}=\frac{f_c(w_1)\times f(w_1)}{f_c(w_1)\times f(w_1)+f_c(w_2)\times f(w_2)}\times 10 \quad (4)$$

两类贝叶斯最小风险,决策 w_1 和决策 w_2 的风险分别为:

$$\frac{p(w_1\mid x)}{p(w_2\mid x)}>\frac{\lambda_{21}-\lambda_{22}}{\lambda_{12}-\lambda_{11}}$$
$$\frac{p(w_1\mid x)}{p(w_2\mid x)}<\frac{\lambda_{21}-\lambda_{22}}{\lambda_{12}-\lambda_{11}} \quad (5)$$

其中:$p(w_1\mid x)=\dfrac{p(x\mid w_1)p(w_1)}{\sum_{i=1}^{2}p(x\mid w_i)p(w_i)}$

$$p(w_2\mid x)=\frac{p(x\mid w_2)p(w_2)}{\sum_{i=1}^{2}p(x\mid w_i)p(w_i)}$$

$$p(w_1)=p(w_1\mid x)$$
$$p(w_2)=p(w_2\mid x)$$

其中:x 表示现在分词,$p(x)$ 表示 x 占词性属于 w_1 或 w_2 的所有现在分词比例;$p(x\mid w_1)$, $p(x\mid w_2)$ 表示在词性 w_1, w_2 的现在分词中 x 的概率比例。

训练决策树模型的过程中,不断地用贝叶斯后验公式调整 w_1,w_2 先验概率,同时更新风险系数 λ_{ij},使得两边比值动态的改变,更加接近现在分词词性实际分布。经过贝叶斯最小风险改进的分类器在整个分类过程中,做出的分类决策始终使得风险值最小,最小化错误率,保证词性标注的正确率。

采用贝叶斯改进的迭代过程如下。

1) 预处理。

$p(w_1)=f(w_1)/|S|$,$p(w_2)=f(w_2)/|S|$;

根据式(4) 计算 λ_{12},λ_{21};

计算归一化因子 $P(x)$;

2) 迭代过程。

FOREACH x IN *Values* (*Bays*)

　更新 $f(w_1)$,$f(w_2)$,$f_c(w_1)$,$f_c(w_2)$;

　更新 λ_{12},λ_{21};

　统计 $p(x\mid w_1)$,$p(w_2\mid x)$;

　计算 $p(w_1\mid x)$,$p(w_2\mid x)$;

END

3) 利用式(5) 做决策。其中 *Values*(*Bays*) 表示需要采用贝叶斯改进的分词集合。

现在分析模型的复杂度,假设 *Values*(*Bays*) 有 k 个分词。显然,更新每一个 FOREACH 中的参数只需要 $O(1)$ 的时间,时间复杂度是 $O(k)$。另外,模型中只用到了有限个额外变量,空间复杂度是 $O(1)$。因此从模型复杂度的角度上分析,模型是高效率的。

3　决策树的生成

首先,$T(S,p\_+1)$ 最小,信息量最纯,因此选取 p_+1 作为决策树分类器的第一个属性,$E(S_{p\_+1=\mathrm{PN}})$,$E(S_{p\_+1=\mathrm{AT}})$,$E(S_{p\_+1=\mathrm{JJ}})$,$E(S_{p\_+1=\mathrm{RB}})$,$E(S_{p\_+1=\mathrm{BEZ,VB,VBD}})$ 均为 0,说明其实例分布最纯,这 5 种情况可以生成叶子节点。

选取 p_-1 作为分类器的第二个属性。$E(S_{p\_-1=\mathrm{BEZ,VB,VBD}})$,$E(S_{p\_-1=\mathrm{JJR}})$ 均为 0,说明其实例分布最纯,这 2 种情况可以生成叶子节点。

然后,再选取双边条件的状态空间,做决策树进一步伸展。

最后,采用贝叶斯最小风险分类器,最终生成贝叶斯决策树如图 1 所示。

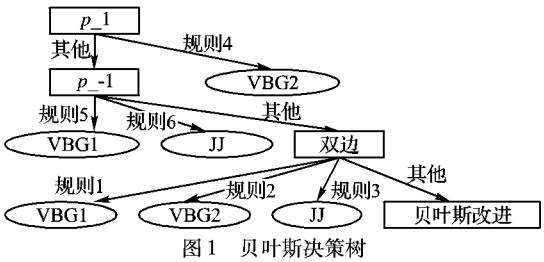

图 1　贝叶斯决策树

贝叶斯决策树的分词词性判定流程如下:

1) 若 p_+1 满足规则 4,分词词性是 VBG2;否则到 p_-1;

2) 若 p_-1 满足规则 5,分词词性是 VBG1;若满足规则 6,分词词性是 JJ;否则到双边;

3)若双边满足规则 1,2,3,分词词性分别是 VBG1, VBG2, JJ;否则对双边不确定词性的情况,用贝叶斯改进。

4　模型测试和实验对比结果

模型训练完毕,测试含有 100000 个未标注的单词的语料库。对比 HMM 和贝叶斯决策树模型在现在分词词性标注上的正确率。其中有 1 471 个词性是 VBG1, 299 个词性是

VBG2,291 个词性是 JJ。

4.1 对比实验1——例句样本测试

对比使用 HMM 和贝叶斯决策树对引言部分的例句样本的现在分词进行标记,结果如表6。

表6 HMM 和贝叶斯决策树例句测试结果

单词	正确标注	HMM	贝叶斯决策树
hoping	VBG1	NN(错误)	VBG1
writing	VBG1	JJ(错误)	VBG1
interesting	JJ	JJ	VBG2(错误)
reading	VBG2	JJ(错误)	VBG2
training	VBG2	JJ(错误)	VBG2
understanding	VBG2	VBG1(错误)	VBG1(错误)
speaking	VBG2	VBG1(错误)	VBG2
reading	VBG2	VBG2	VBG2
writing	VBG2	VBG2	VBG2

可以看出,贝叶斯决策树模型的标记正确数量大于 HMM。

4.2 对比实验2——语料库文本测试

对比使用 HMM 和贝叶斯决策树对含有 100 000 英文单词的语料库所有的现在分词样本进行标记,实验结果数据如表7。

表7 HMM 和贝叶斯决策树语料库测试结果

类别	测试文本	HMM		贝叶斯决策树	
		正确标记数	正确率/%	正确标记数	正确率/%
VBG1	1471	1288	87.6	1463	99.4
VBG2	299	183	61.2	273	91.3
JJ	291	282	96.9	268	92.1

可以看出 HMM 虽然在 JJ 上有很高的正确率,但是在 VBG1 和 VBG2,特别是后者上错误率很高。因此,分词总体正确率只有 85.1%。

贝叶斯决策树模型的正确率在 JJ 上虽然降低,却大大提高了前两类的分词正确率。因此分词总体的正确率达到了 97.6%,仍然大于 HMM。

4.3 对比实验3——英文文本测试

采用贝叶斯决策树模型之前文本正确率是 95.6%,改进之后正确率提高到 98.3%。

5 结语

表6~7 充分说明贝叶斯最小风险决策树模型在标注英文现在分词词性上拥有较大的优势。

当然分词词性不仅仅是受到前后相邻词性的影响,也有一些情况受到整个句子影响,但是 97.6% 的分词标注正确率和 98.3% 的总体文本正确率已经取得了改进。为了体现模型的通用性,应当将模型应用进行扩展,可以应用到对过去分词词性标注,同时可以解决一些英文词性标注诸如定语从句、独立主格内嵌的常见错误。

参考文献:

[1] POLAT K. A novel hybrid intelligent method based on C4.5 decision tree classifier and one against all approach for multiclass classification problems [J]. Expert Systems with Applications, 2007, 36(2): 1587-1592.

[2] SHUKLA S K, TIWARI M K. Soft decision trees: A genetically optimized cluster oriented approach [J]. Systems with Applications, 2009, 36(1): 551-563.

[3] QUINLANN J R. Induction of decision trees [J]. Machine Learning, 1986, 1(1): 81-106.

[4] PULKKINEN P. Fuzzy classifier identification using decision tree and multiobjective evolutionary algorithms [J]. International Journal of Approximate Reasoning, 2008, 36(2): 526-543.

[5] KUPIEC J. Robust part of speech tagging using a hidden Markov model [J]. Computer Speech and Language, 1992, 6(3): 225-242.

[6] CHEN JING-NIAN, HUANG HOU-KUAN, TIAN SHENG-FENG, *et al*. Feature selection for text classification with Naive Bayes [J]. Expert Systems with Applications: An International Journal, 2009, 36(3): 5432-5435.

[7] LI REN-PU. Mining classification rules using rough sets and neural networks [J]. European Journal of Operational Research, 2004, 157(2): 439-448.

[8] GARSIDE R, LEECH G, SAMPSON G. The computational analysis of English [M]. London: Longman, 1987.

(上接第2570页)

[3] JIA Y H. Fusion of landsat TM and SAR images based on principal component analysis [J]. Remote Sensing Technology and Application, 1998, 13(1): 46-49.

[4] 敬忠良,肖刚,李振华. 图像融合——理论与应用[M]. 北京: 高等教育出版社, 2007.

[5] 程英蕾. 多源遥感图像融合方法研究[M]. 西安: 西北工业大学, 2006.

[6] 李振华,敬忠良,孙韶媛,等. 基于方向金字塔框架变换的遥感图像融合算法[J]. 光学学报, 2005, 25(5): 23-27.

[7] 纪启国. 利用小波变换对图像进行像素级融合[J]. 安徽职业技术学院学报, 2008, 7(1): 25-29.

[8] TAO G Q, LI D P, LU G H. On image fusion based on different fusion rules of wavelet transform [J]. Acta Photonica Sinica, 2004, 33(2): 221-224.

[9] 郑心武,周开利. 小波变换在图像融合技术中的应用及探考[J]. 琼州学院学报, 2008, 15(5): 40-42, 46.

[10] 徐胜祥,胡超,徐运清. 利用 Matlab 实现基于小波变换的遥感图像融合[J]. 微计算机信息, 2008, 24(1/3): 302-303.

[11] 曾梅兰,金升平. 基于小波框架的多传感器图像融合[J]. 计算机工程与应用, 2004, 40(25): 56-61.

[12] 张大明,胡茂林,张长耀. 基于方向可调滤波器和小波分析的图像融合[J]. 微机发展, 2005, 15(7): 28-30, 34.

[13] PIELLA G. A general framework for multiresolution image fusion: From pixels to regions [J]. Information Fusion, 2003, 4(4): 259-280.

[14] PARK J H, KIM K K, YANG Y K. Image fusion using multiresolution analysis [C]// IEEE 2001 International Geoscience and Remote Sensing Symposium. [S.l.]: IEEE, 2001, 2: 864-866.

[15] LI S T, JAMES T K, WANG Y. Using the discrete wavelet frame transform to merge Landsat TM and SPOT panchromatic images[J]. Information Fusion, 2002, 3(1): 17-23.

[16] 高继镇,刘以安. 基于小波系数模糊积分的图像融合算法研究[J]. 计算机应用, 2008, 28(7): 121-123.

[17] 赵文吉,段福州,刘晓萌,等. ENVI 遥感影像处理专题与实践[M]. 北京: 中国环境科学出版社, 2007.

[18] TSENG D C, CHEN Y L, LIU M S C. Wavelet-based multispectral image fusion[C]// Geoscience and Remote Sensing Symposium. [S.l.]: IEEE, 2001, 4: 1956-1958.