基于最大熵模型的多音字消歧*

刘方舟[1]，施勤[2]，陶建华[1]

(1. 中国科学院自动化研究所，模式识别国家重点实验室，100080；2. IBM 中国研究中心，100083)

文　摘： 字音转换是语音合成系统必不可少的模块，而多音字消歧则是字音转换的核心问题。本文选择了 33 个常见常错的多音字作为研究对象，使用最大熵模型来辨析多音字的读音。在特征选择方面，本文比较了不同领域的多种关键词选择的方法，采用似然比来提取关键词。本文还对比了最大熵模型与决策树算法在多音字消歧上的表现，实验结果表明，最大熵模型的性能要优于决策树算法。

关键词： 字音转换；多音字；最大熵模型；决策树

中图分类号： TP391

1　引言

字音转换是语音合成系统(TTS)必不可少的模块，其正确率直接影响语音合成系统的可懂度。在汉语语音合成系统中，字音转换的任务就是将文字序列转换为对应的拼音序列。大多数情况下，字音转换都是在词典中检索当前词，配以对应的拼音。然而，汉语中有的字对应多个拼音。如“干”字在“干衣服”中读“gan1”，而在“干重活”中读“gan4”。字音转换的关键和难点就是如何解决这种一字多音的问题。汉语中常见的多音字有“为、长、重”等。除去多音字，汉语中还有少量多音词，如“教授 (jiao4shou4 或 jiao1shou4) 、朝阳 (chao2yang2 或 zhao1yang2)”等。本文的研究目标就是根据上下文信息自动的辨析多音字的读音。

一般认为多音字的读音是跟语义和语言习惯相关的，比如“还”表示“归还”时读“huan2”，表示“仍然”时读“hai2”。但按照现在的自然语言处理水平，从语义层面上来解决多音字问题还不太可能。对多音字的读音进行消歧通常有两种主流方法：

1)　基于手工规则的方法：由语言专家总结出多音字消歧的规律，并将这些规律写成计算机可以理解的规则形式，且仅涉及计算机可以获取的信息。计算机发现多音字时就按规则逐条进行条件匹配和消歧处理。

2)　基于统计机器学习的方法：把多音字消歧问题视为机器学习中的分类问题，首先收集包含多音字的语料库并标注多音字的正确读音，然后分别对每个多音字抽取字词、词性等上下文信息，通过机器学习的方法完成多音字消歧。

最初绝大多数语音合成系统都是采取手工规则的方法来进行多音字消歧。然而随着规则数目的增加，某一个多音字的上下文环境可能被多条规则所匹配，这就产生了规则冲突，这是基于规则的方法难以解决的问题之一。随着大语料库在语音合成研究领域的蓬勃发展，很多研究者着手用统计方法来进行多音字消歧。Yarowsky[1]使用似然比选择对多音字读音有辨析作用的上下文特征，然后用统计决策列表对多音字进行消歧，取得了很好的效果。Wang[2]比较了互信息、似然比等多种选择关键词的方法，并采用决策树对多音字的读音进行分类。Zhang[3]采用基于扩展的随机复杂度的随机决策列表来自动提取多音字的读音规则。Zhen[4]将错误驱动的基于转换的规则学习方法(TBL)应用到多音字消歧的问题上，获得了比决策树更高的准确率。

最大熵模型[5]是近年来在自然语言处理中广泛使用的统计分类模型。它在估计概率分布时，除了使之满足约束条件外，不做任何假设，即选取熵最大的概率分布。该模型已经成功的应用于自然语言处理的各个领域，如分词[6]、词性标注[7]、语义消歧[8]等。本文尝试用最大熵模型来解决多音字消歧的问题。

本文下面的章节安排如下：第二节介绍了最大熵模型的基本框架；第三节阐述了多音字消歧中的特征选择；第四节详细的描述了关键词的选择、

*基金项目：国家自然科学基金(No. 60575032)，863(No. 2006AA01Z138)
作者简介：刘方舟（1983），男（汉族），湖南，在读博士。
通讯联系人：陶建华，博士

cutoff 值的选择以及最大熵模型与决策树算法的对比等多组实验，并对实验结果进行了分析；最后第五节总结全文。

2 最大熵模型框架

2.1 特征和约束

自然语言中的许多问题都可以归结为统计分类问题，即估计类y在上下文x中的发生概率 $p(y|x)$。在多音字消歧的问题中，y表示多音字的读音，x表示多音字的上下文环境，包括词性、词长等。

最大熵模型的特征 f 定义为描述事件(x, y)是否发生的二值函数，即：

$$f(x,y)=\begin{cases}1 & \text{如果x与y共现}\\ 0 & \text{其他}\end{cases}$$

事件(x, y)在训练样本中的期望：

$$E_{\tilde{p}}f=\sum_{x,y}\tilde{p}(x,y)f(x,y) \qquad (1)$$

其中，$\tilde{p}(x,y)$ 为该事件在训练样本中的经验分布。

该事件在模型中的期望：

$$E_{p}f=\sum_{x,y}\tilde{p}(x)p(y|x)f(x,y) \quad (2)$$

其中，$\tilde{p}(x)$ 为上下文 x 在训练样本中的经验分布，$p(y|x)$ 为模型中的条件概率分布。

事件(x, y)的样本期望值与模型期望值应该一致，即：

$$E_{p}f=E_{\tilde{p}}f \quad (3)$$

该式称为特征 $f(x,y)$ 的约束，它限制概率模型 $p(y|x)$ 从统计意义上接近训练样本的分布。

2.2 最大熵原则

假设存在 k 个特征，满足这 k 个特征的约束的所有概率分布构成一个集合：

$$P=\{p \mid E_{p}f_i=E_{\tilde{p}}f_i, 1\le i\le k\} \quad (4)$$

最大熵模型为满足以下条件的模型：

$$p^{*}=\arg\max_{p\in P}H(p) \qquad (5)$$

其中，$H(p)=-\sum_{x,y}\tilde{p}(x)p(y|x)\log p(y|x)$。

即从满足所有约束的概率分布中选取条件熵最大的概率分布作为最大熵模型。

2.3 指数形式

用拉格朗日乘子法求解(5)式，可得最大熵模型具有如下形式：

$$p^{*}(y|x)=\frac{1}{Z(x)}\prod_{i=1}^{k}\alpha_i^{f_i(x,y)} \qquad (6)$$

其中，α_i 为特征 f_i 的权重，可用 GIS 迭代算法从训练样本中求得，$Z(x)=\sum_{y}\prod_{i=1}^{k}\alpha_i^{f_i(x,y)}$ 为归一化因子。

3 特征选择

最大熵模型的关键在于选取合适的特征模板。由于最大熵模型不对特征作独立性假设，所以可以任意的选择和组合特征。

本文参考并改进了 Zhen [4]所选特征，选取了多音字前后两个字或词范围内(关键词例外，下文会单独解释)的 8 类上下文信息作为基本特征，如表 1 所示。

表 1 基本特征模板

基本特征	意义
LC-2, LC-1, LC1, LC2	多音字前后的字
LW-2, LW-1, LW0, LW1, LW2	多音字前后的词
POS-2, POS-1, POS0, POS1, POS2	多音字前后的词的词性
LEN-2, LEN-1, LEN0, LEN1, LEN2	多音字前后的词的词长
KWB	多音字前的关键词
KWA	多音字后的关键词
KWBPOS	多音字前的关键词的词性
KWAPOS	多音字后的关键词的词性
TONE-1, TONE1	多音字前后的字的声调
POSINWORD	多音字在词中的相对位置 (词首、词中、词尾、单字词)
POSINSEN	多音字在句中的相对位置 (句首、句中、句尾、单字句)

其中关键词指上下文中能对多音字的读音起辨析作用的词。以多音字“为”字为例，“称”位于“为”字前面时，“为”通常作动词，读作

“wei2”，如：

1) 维吾尔族农民**称**他**为**种棉大王

2) 他**称**这一举动**为**希望马拉松

显然“称”是“为”的一个关键词。Wang[2]仅用了前后一个词作为特征，Zhen[4]所用的词特征也局限在前后两个词的窗宽内，如：“LW-1_称”表示多音字的前一个词为“称”。然而，从上面两个例句来看，关键词到多音字的距离可远可近，句 1“称”和“为”仅相隔一个词，句 2“称”和“为”则相隔了 3 个词。因此，本文在选择关键词时既不限制窗宽，也不使用词到多音字的距离，即整个句子的词都可以作为多音字的关键词。

一个词是否是关键词还与它是出现在多音字之前还是出现在多音字之后有关，如：

3) 以毛泽东**为代表**的中国共产党人

4) 参加会议的**代表为**了各自国家的利益

句 3 中“…为代表”的结构在语料中经常出现，这时的“为”通常作动词，读作“wei2”，因此“代表”位于“为”字之后时应该是关键词。而句 4 中的“代表”位于“为”字之前，作介词，读作“wei4”。如果对前后不加区分的话，所选关键词可能会混淆多音字的读音。因此关键词要分多音字之前和多音字之后分别选择。

除单独使用基本特征外，本文还将基本特征组合成复合特征来描述更复杂的上下文环境。去掉在实验中表现不好的复合特征后，本文保留了如表 2 所示的复合特征。

表 2 复合特征模板

复合特征	意义
LW1LW2, LW-1LW-2	多音字前后词的组合
LC-1LC1	多音字前后字的组合
POS-1POS1, POS-2POS1, POS-1POS2, POS1POS2, POS-2POS-1, POS-1POS0, POS0POS1, POS-1POS0POS1	多音字前后的词性组合
LW-1POS1, LW1POS-1, LW1POS2, LW-1POS-2	多音字前后词与词性的组合
LC-1POS1, LC1POS-1, LC-1POS0, LC1POS0	多音字前后字与词性的组合

根据表 2 和表 3 中的特征模板，以句 5 为例提取特征实例。

5) 直径(n) 大约(d) **为**(v) 六(m) 英寸(q)

LC-1 表示多音字的前一个字，实例化后得到特征 LC-1_约；POS0 表示多音字本身的词性，实例化后得到特征 POS0_v；LC-1POS0 表示多音字的前一个字和多音字本身的词性的组合，实例化后得到特征 LC-1_约_POS0_v。

将特征模板实例化后，还需要对特征实例进行筛选，删除干扰噪声，保留重要特征。常用的自动选择特征的方法有门限裁剪法(CCFS)和似然值增益法(IFS)[9]。门限裁剪法认为出现频率过小的特征不可靠，将出现次数低于某一阈值(cutoff 值)的特征删除。似然值增益法是以特征对模型似然值的贡献为依据来判断特征优劣的迭代算法，每次迭代都选出对模型的似然值贡献最大的特征加入特征集。Ratnaparkhi[9]指出，似然值增益法计算复杂度高，训练时间长，而效果并不一定比门限裁剪法好。因此本文采用了简单有效的门限裁剪法。

4 实验及结果分析

4.1. 语料

汉语的多音字数目众多，《现代汉语词典》共收录了 1036 个多音字[3]，其中很多只存在于没有发音歧义的多字词中，包含这些多音字的常用词条可以收录在电子词典中供 TTS 系统查询，本文所要处理的主要是那些能单独成词的多音字和发音有歧义的多字词。据 Zhang[3]的统计，在 250 万字的《人民日报》语料中，有 688 个多音字可以单独成词，另有 170 个多音词出现；多音字的使用频率相差甚远，前 180 个高频多音字的使用频率占到全部多音字使用频率的 95%，并且大部分多音字都有一个占主导地位的读音(以下称为高频音)，在前 180 个高频多音字中只有 41 个字的高频音的使用频率低于 95%。 本文选择了其中常见常错的 33 个多音字(如“为、长、重”等)和 24 个多音词(如“背着、教授”等)作为主要的研究对象。

本文实验的语料来自 1982 年至 2001 年间的《人民日报》。首先用语音合成系统的前端对原始文本进行自动分词、词性标注和拼音标注。由于语料中出现了很多重复的上下文信息，比如在多音字“朝”的语料中，“中朝友谊”大量出现，不加限制的话会产生许多冗余语料，因此在筛选语料时，规定多音字前后的字词重复出现的次数不能超过 10 次。经过多人反复的手工校对多音字的拼音，本文构建了一个平均每个多音字 5000 个句子的多音字语料库，按照 8:1:1 的比例划分为训练集、开发集和测试集。

4.2. 关键词选择

Yarowsky[1]在英语的多音字消歧中使用似然比来选择关键词，似然比定义为：

$$\left|Log(\frac{P(P_1|W)}{P(P_2|W)}\right| \qquad (7)$$

其中 P_i 为多音字的第 i 种读音，W 为词，似然比越大，词对多音字读音的区分能力就越强。

Yang[10]针对文本分类问题，比较了互信息、信息增益等五种关键词的选择方法。Yang 使用的互信息的公式为：

$$\sum_i P(C_i)\log\frac{P(W\mid C_i)}{P(W)} \qquad (8)$$

信息增益的公式为：

$$P(W)\sum_i P(C_i\mid W)\log\frac{P(C_i\mid W)}{P(C_i)}+P(\overline{W})\sum_i P(C_i\mid \overline{W})\log\frac{P(C_i\mid \overline{W})}{P(C_i)} \qquad (9)$$

其中 C_i 为第 i 类文本，$\overline{W}$ 表示词 W 不出现，结果显示互信息表现较差，信息增益效果较好，Yang 指出互信息具有偏爱低频词和对概率估计错误敏感的缺点。

Mladenic[11]针对层级文本分类，比较了交叉熵、优势率等关键词选择的方法。Mladenic 使用的交叉熵的公式为：

$$P(W)\sum_i P(C_i\mid W)\log\frac{P(C_i\mid W)}{P(C_i)} \qquad (10)$$

优势率的公式为：

$$\log\frac{P(W\mid pos)(1-P(W\mid neg))}{(1-P(W\mid pos))P(W\mid neg)} \qquad (11)$$

其中 pos 表示正类，neg 表示负类，实验结果表明，对于贝叶斯分类器而言，优势率表现最好，交叉熵的效果优于信息增益，Mladenic 分析指出，由于 $P(\overline{W})\gg P(W)$，所以信息增益公式的后半部分 $P(\overline{W})\sum_i P(C_i\mid\overline{W})\log\frac{P(C_i\mid\overline{W})}{P(C_i)}$ 所占比重很大，即在信息增益中词 W 不出现的信息量很大，而贝叶斯分类器只能使用关键词出现的信息，所以使用交叉熵(即信息增益的前半部分)选择关键词效果更好。

Yarowsky 的似然比适用于只有两种读音的多音字，为了将其应用到两种以上读音的多音字，本文将似然比的公式改造为：

$$P(W)\sum_i P(P_i)\left|Log(\frac{P(P_i\mid W)}{P(\overline{P_i}\mid W)})\right| \qquad (12)$$

其中 $\overline{P_i}$ 表示读音不为 P_i，乘以 P(W)是为了降低低频词的似然比得分。

Mladenic 的优势率同样只适用于两类的情况，本文也将其改造为：

$$P(W)\sum_i P(P_i)\left|\log\frac{P(W\mid P_i)(1-P(W\mid\overline{P_i}))}{(1-P(W\mid P_i))P(W\mid\overline{P_i})}\right| \qquad (13)$$

针对多音字消歧的问题，本文重新对比了似然比、互信息、信息增益、交叉熵和优势率这五种选择关键词的方法在最大熵模型中的效果。实验从训练语料中选择得分最高的前关键词和后关键词各 100 个，仅使用关键词特征模板(KWB、KWA)，集外测试结果如表 3 所示，其中平均准确率定义为测试集中所有多音字的正确样本总和比上测试集中所有多音字的样本总和，测试集的缺省平均准确率(即高频音所占比例)为 80.66%。

表 3　选择关键词的方法比较

方法	平均准确率
似然比	**85.50%**
互信息	81.66%
信息增益	84.30%
交叉熵	84.84%
优势率	85.27%

本文的实验结果与 Yang 和 Mladenic 在文本分类问题中的实验结果基本吻合。用互信息来选择关键词的效果明显差于其它方法，这主要是由于互信息有偏爱低频词的缺点，当某个词的出现次数很少时，它很可能只是偶然的出现在多音字的某一种读音中，导致它与该种读音的互信息非常大，因此大量的低频词被选为关键词，然而这些低频词实际上并不具备统计意义。交叉熵的表现比信息增益要好，原因主要是，在信息增益中关键词不出现的信息量所占比重很大，而最大熵模型难以使用关键词不出现的特征模板，所以用交叉熵(即只使用关键词出现的信息)来选择关键词效果更好。似然比的方法最简单也最有效，优势率比似然比略逊一筹。因此本文使用似然比来选择关键词，平均每个多音字 400 个关键词，多音字之前与多音字之后各 200 个。

4.3. Cutoff 值的选择

为了确定最优的 cutoff 值，本文在平均每个多音字 4000 句语料的训练集上进行训练，在平均每个多音字 500 句语料的开发集上测试不同 cutoff 值的平均准确率，结果如表 4 所示，最佳的 cutoff 值为 2，即出现次数少于或等于 2 次的特征均被删除。

表 4　cutoff 值的选择

cutoff值	平均准确率
0	92.24%
1	92.17%
2	**92.45%**
3	92.24%
4	92.34%
5	92.20%

观察发现，大部分被舍弃的低频特征都是具体的字词特征(LW、LC、KW)。这些特征虽然具有较强的区分能力，但它们在统计意义上不够稳定，容易导致模型过度拟和，因此删除它们有助于提高特征集的整体质量。但如果 cutoff 值取得过高的话，又会丢失许多有用的信息，使得模型性能下降,因此选择一个合适的中间值是必要的。

4.4. 算法对比

决策树算法也是当前自然语言处理中常用的统计分类算法。决策树自顶向下，选择能最大的减少不确定性(如信息增益最大)的属性作为分枝属性，在各个树节点进行属性值的比较，并根据不同的属性值，选择不同的树分枝，直到树的叶子节点，得到分类结论。该算法已成功的应用到众多领域，如句法分析[12]、韵律节奏预测[13]等。本文用决策树算法作为参照来对比最大熵模型的多音字消歧效果。

决策树算法所用语料与最大熵模型的完全相同。在平均每个多音字 4000 句语料的训练集上进行训练，以平均每个多音字 500 句语料作为开发集对决策树进行剪枝，然后在平均每个多音字 500 句语料的测试集上进行测试，其中测试集的缺省平均准确率为 80.66%。当特征集包含具体的字词特征时，决策树的平均准确率为 85.30%；当不使用具体的字词特征时，决策树的平均准确率为 87.83%。可见加入具体的字词特征会大大降低决策树算法的性能。

最大熵模型的平均准确率为 91.38%。图 1 对比了 13 个常见多音字在两种算法下的测试结果，其中决策树算法没有使用具体的字词特征。

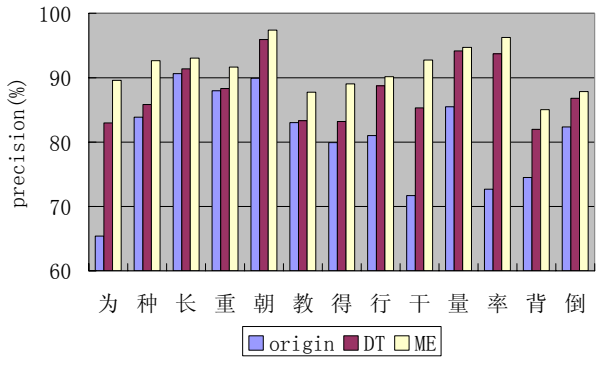

图 1 最大熵与决策树的对比

实验结果表明，对于多音字消歧而言，最大熵模型的效果明显好于决策树算法。与决策树算法相比，最大熵模型主要有以下两个优点：

1) 决策树算法存在严重的碎片问题。使用稀疏的分枝属性(如具体的字词特征)反复划分样本集会产生样本数目过小的子样本集，由于这些分枝不具备统计意义，所以容易导致决策树过度拟和。这也是使用具体的字词特征时，决策树性能下降的原因。而最大熵模型在参数估计时，不需要划分样本集，因此不会产生碎片问题。

2) 最大熵模型具有良好的量化描述能力，它可以通过权重系数准确的描述各个特征对分类结果的贡献，从而将所有特征有效的融合在同一个框架下，而不是孤立等价的使用其中的某几个特征。如果将最大熵模型的特征看作规则，特征的权重视为规则的权重的话，最大熵模型可以被看作一个带权重的规则系统。

5 结论

本文尝试将最大熵模型用于多音字消歧的问题，取得了令人满意的实验结果。本文还将文本分类中关键词选择的方法用于多音字的关键词选择，验证了文本分类中互信息偏爱低频词和交叉熵优于信息增益的结论。算法的对比实验显示，对于多音字消歧而言，最大熵模型明显优于决策树算法，其原因主要在于决策树算法存在严重的碎片问题和最大熵模型具有良好的量化描述能力。鉴于最大熵模型在多音字消歧中的出色表现，下一步工作考虑将它运用到其它类似的文本消歧的问题中去，比如数字、符号读法的消歧。

6 致谢

本文的部分工作是第一作者在 IBM 中国研究中心做实习学生期间完成的，非常感谢语音组各位老师提供的建议和帮助。

参 考 文 献

[1] David Yarowsky. “Homograph disambiguation in speech synthesis.” In J. van Santen, R. Sproat, J. Olive and J. Hirschberg (eds.), Progess in Speech Synthesis, Springer-Verlag, 1997, pp. 159–175.

[2] Wern-Jun Wang, Shaw-Hwa Hwang, Sin-Horng Chen. “The broad study of homograph disambiguity for mandarin speech synthesis”, ICSLP96, pp. 1389-1392.

[3] Zi-Rong Zhang, Min Chu. "An efficient way to learn rules for grapheme-to-phoneme conversion in Chinese", ISCSLP2002, pp. 59.

[4] Min Zheng, Qin Shi et al. “Grapheme-to-phoneme conversion based on TBL algorithm in Mandarin TTS system”, INTERSPEECH2005, pp. 1897-1900.

[5] Adam L. Berger, Stephen A. Della Pietra, Vincent J. Della Pietra. “A maximum entropy approach to natural language processing”, Computational Linguistics, 1996, 22(1). 39-71.

[6] Jin Kiat Low, Hwee Tou Ng, Wenyuan Guo. “A Maximum Entropy Approach to Chinese Word Segmentation”, the Fourth SIGHAN Workshop on Chinese Language Processing, 2005, pp. 161-164.

[7] Adwait Ratnaparkhi. “A Maximum Entropy Model for Part-of-speech Tagging”, EMNLP96, pp. 133–142.

[8] Hoa Trang Dang, Ching-yi Chia et al. “Simple features for Chinese word sense disambiguation”, COLING2002, pp. 88-94.

[9] Adwait Ratnaparkhi. “Maximum Entropy Models for Natural Language Ambiguity Resolution”, Ph.D. Dissertation. University of Pennsylvania, 1998.

[10] Yiming Yang, Jan O. Pedersen. “A comparative study on feature selection in text categorization”, ICML97, pp. 412-420.

[11] Dunja Mladenic, Marko Grobelnik. “Feature selection for unbalanced class distribution and Naive Bayes”, ICML99, pp. 258-267.

[12] David M. Magerman. “Statistical decision-tree models for parsing”, the 33rd Annual Meeting of the ACL, 1995, pp. 276-283.

[13] Michelle Q. Wang, Julia Hirschberg, “Automatic classification of intonational phrase boundaries”, Computer Speech and Language, 1992, 6(2). 175-196.

Maximum Entropy Based Homograph Disambiguation

Fangzhou Liu[1]， Qin Shi[2]， Jianhua Tao[1]

(1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, 100080, Beijing, China； 2. IBM China Research Lab, 100083, Beijing, China)

Abstract：Grapheme-to-phoneme conversion is an essential component in Text-to-Speech system, and homograph disambiguation is the core issue of grapheme-to-phoneme conversion. This paper selects 33 key polyphones which are frequently used and often read wrong to study, and presents a maximum entropy model for homograph disambiguation. In feature selection, this paper evaluates various keyword selection methods in different domains, and adopts the log-likelihood ratio to extract keywords. This paper also gives a comparison of maximum entropy model with decision tree algorithm on the performance of homograph disambiguation, the experimental results showed that maximum entropy model surpass decision tree algorithm obviously.

Key words：Grapheme-to-phoneme; Polyphone; Maximum entropy model; Decision tree