用概率方法处理汉英机器翻译系统中的歧义*

刘颖*　常宝宝**　刘群*　王斌*

*中国科学院计算所二室　100080

**北京大学计算语言所　100871

摘要： 机器翻译中，在词汇、句法、语义等各个层面上，都经常遇到歧义问题，如何有效地把语言学和统计学结合起来处理歧义问题是很重要的。本文就是在一个汉英机器翻译系统基础上（这个系统主要用基于规则的方法实现），运用综合概率打分函数对不确定性方法分析的各个不确定分支进行引导，使机器翻译在词汇、句法层的排歧达到了一定的效果，同时也提高了机器翻译的速度和准确率。

Resolving Ambiguities with Probability method in Chinese-English machine translation

LiuYing*, Chang Bao-bao**, LiuQun*, WangBin*

(*Institute of Computing Technology, Chinese Academy of Sciences)

(**Institute of Computational Linguistics, Peking University)

Abstract: In machine translation, ambiguities are often faced in the lexical level、 syntactic level and semantic level. It is very important how to combine linguistics with statistics effectively to deal with ambiguities . In this paper, a unified probabilistic scoring function and a rule scoring function are used to resolve lexical and syntactic ambiguities in a mainly rule-based Chinese-English machine translation system. Good results have been achieved, and at the same time the speed and accuracy of our machine translation system have also been improved.

一、引言

在自然语言处理中，经常会遇到歧义问题。语言中的歧义既反映在单词上，又反映在单词组成的各种结构上，形成词汇歧义和结构歧义。基于规则方法构造的系统一般使用语言学知识、计算机知识、逻辑知识来组织规则，我们的汉英机器翻译系统就是成功地使用了这些技巧，通过人的内省方式来构造规则的。但由于人对语言的知识很难精确掌握，所以用规则来刻画很难面面俱到。为了保留正确的分析，我们使用不确定性算法，因为如果使用确定性算法，当规则刻画得粗时，就有可能保留了错误的结果而删除了正确的结果。

* 本文研究工作受国家 863 项目（编号 863-306-03-06-2 ）基金资助

为了解决由不确定性算法分析产生的歧义，我们把概率方法结合到我们的系统中。[1]和[2]中提出了解决句法歧义的综合打分函数，这个函数可以把各种不同的知识源用统一的公式结合起来，以前的工作表明这个打分函数对于各种不同的应用能提供很高的识别能力（[3]和[4]）。结合综合打分函数，我们提出了给规则打分的一种方法。本文是在我们的汉英机器翻译系统基础上，这个系统主要以规则方法为主，我们把规则方法分析出的句子或短语用综合的概率打分函数来统计。对于歧义结构或不确定性分支，选择出概率最大结果。

本文通过最大可能估计方法来对歧义结构进行消歧，但如果没有充分的训练数据，最大可能估计方法显然是不可靠的。而且用最大可能估计方法估计未出现事件的概率为零，在许多应用中这个结果是不合适的。为了避免稀疏数据问题，我们使用参数平滑算法[5]。

二、综合的概率打分函数

语言学知识包括词汇、句法和语义对于解决句法歧义都是必不可少的。为了用统一的公式结合各种知识源，[2]中提出了一个综合的概率打分函数。在英汉机器翻译系统（ BehaviorTran[2])和语音处理系统[4]已经成功地使用这个打分函数来解决歧义问题。下面给出综合的概率打分函数。

1 定义

对于输入词串 w={$w_1 w_2 \cdots w_n$}={w_1^n}，其中 w_i 代表这个词串的第 i 个词。Lex_k ($1 \le k \le M$)表示 M 个可能的序列中第 k 个词汇序列，$syn_{j,k}$ ($1 \le j \le N_k$) 为相应于 lex_k 的第 j 个句法结构，N_k 是与 lex_k 相联系的句法结构数量。消歧过程就是对于输入句子{w_1^n}，找到其相应的最可能词汇序列和句法结构的过程，也即发现下标 $(\hat{j},\hat{k})$ 使得 $P(syn_{\hat{j},\hat{k}}, lex_{\hat{k}} | w_1^n)$ 最大，即 $(\hat{j},\hat{k}) = \arg\max_{j,k}\{P(syn_{\hat{j},\hat{k}}, lex_{\hat{k}} | w_1^n)\}$ lex_k ($1 \le k \le m$)表示M个可能的序列中第k个词汇序列，$syn_{j,k}$ ($1 \le j \le N_k$) 为相应于 lex_k 的第j个句法结构，N_k 是与 lex_k 相联系的句法结构数量。消歧过程就是对于输入句子{w_1^n}，找到其相应的最可能词汇序列和句法结构的过程，也即发现下标 $(\hat{j},\hat{k})$ 使得 $P(syn_{\hat{j},\hat{k}}, lex_{\hat{k}} | w_1^n)$ 最大，即 $(\hat{j},\hat{k}) = \arg\max_{j,k}\{P(syn_{\hat{j},\hat{k}}, lex_{\hat{k}} | w_1^n)\}$ 对于句法结构 $syn_{j,k}$，综合的打分函数为：当给定输入词串{w_1^n}时，产生某个词汇序列 lex_k . 分析此词汇序列产生某个句法结构 $syn_{j,k}$ 的概率，也即：

$$score(syn_{j,k}) = P(syn_{j,k}, lex_k | w_1^n) = p(syn_{j,k} | lex_k, w_1^n) \times P(lex_k | w_1^n)$$

$$= s_{syn}(syn_{j,k}) \times s_{lex}(lex_k)$$

其中 $s_{syn}(syn_{j,k}) = p(syn_{j,k} | lex_k, w_1^n)$ 表示句法打分函数，即给定词串 w_1^n 和相应的词汇序列 lex_k，产生句法结构 $syn_{j,k}$ 的概率。$s_{lex}(lex_k) = P(lex_k | w_1^n)$ 表示词汇打分函数，即给定词串 w_1^n，产生词汇序列 lex_k 的概率。

其中$为尾标识，∅为空字符。$L_i$和$r_i$表示要被考虑的左右上下文。上面公式假设每个短语层与其直接的先前短语层相关，而与其它先前的短语层很少相关。为计算的可行性，在公式中仅考虑有限的左右上下文符号。

其中 P(C|{∅},F,G,{B})=

P(F,G 被规约 |输入是{B,F,G})×

P(C←FG| F,G 被规约；输入是{B,F,G})

根据实验上面公式中第一项近似为 1，因此 P(C|{∅},F,G,{B})≈ P(C←FG| F,G 被规约；输入是{B,F,G})

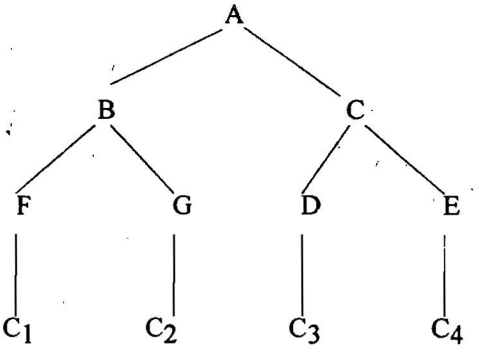


图 ——$Tree_X$

三、计算规则概率的算法

算法 1：假设有一个实例库，库中的每个句子的分析都是正确的，则求规则的概率步骤为：

（1）找到每个句子的分析树。

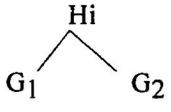
（2）对 H←G_1G_2 的每次使用，相应的计数器加 1：

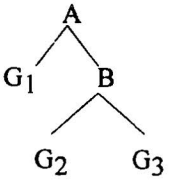
$C(H \leftarrow G_1G_2) = C(H \leftarrow G_1G_2)+1$

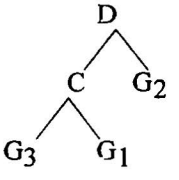
（3）估计概率

$$P(H \leftarrow G_1G_2) = \frac{C(H \leftarrow G_1G_2)}{G_1G_2\text{被规约的次数}}$$

解释：G_1G_2 被规约包括以下几种情况：







算法 2：如果语料库很大，人工则无法确定所有句子的正确分析树。此时在每次应用规则时，用句法结构综合打分函数作为应用这条规则的可信度，则求规则概率的步骤为：

（1）找到每个句子的所有分析树。

（2）对于句法结构 $syn_{j,k}$ 中应用规则 H←G_1G_2，相应的计数器为：

$C(H \leftarrow G_1G_2) = C(H \leftarrow G_1G_2)+S_{syn}(syn_{j,k})$

（3）估计概率

$$P(H \leftarrow G_1G_2) = \frac{C(H \leftarrow G_iG_2)}{G_1G_2\text{被规约的次数}}$$

注：开始统计时，首先挑选出一些分析正确的句子并放进实例库，形成一个小规模的语料库，用它来统计规则概率，根据计算出来的结果，来统计每个句子的综合概率打分函

给定词串 w_1^n，产生词汇序列 lex_k 的概率。

假设生成的句法结构词 $syn_{j,k}$ 主要由词汇序列 lex_k 决定，串 w_1^n 给句法结构提供很少的信息，则 $s_{syn}(syn_{j,k})= p(syn_{j,k} \mid lex_k, w_1^n) \approx P(syn_{j,k} \mid lex_k)$

2 词汇打分函数

$$s_{lex}(lex_k) = P(lex_k \mid w_1^n) = P(c_{k,1}^{k,n} \mid w_1^n) = \frac{P(w_1^n \mid c_{k,1}^{k,n}) \times P(c_{k,1}^{k,n})}{P(w_1^n)} = \frac{s_{lex}^*(lex_k)}{p(w_1^n)}$$

其中 $c_{k,i}$ 表示 w_i 的词类，因为 $P(w_1^n)$对于所有的词汇序列是相同的，所以计算 $s_{lex}(lex_k)$的最大值只要求出 $s_{lex}^*(lex_k)$的最大值就行了。

对于 $s_{lex}^*(lex_k)$中的第一项 $P(w_1^n \mid c_{k,1}^{k,n}) = \prod_{i=1}^{n} P(w_i \mid w_1^{i-1}, c_{k,1}^{k,n})$，假设每个词只与其局部的词类相关，则可近似为：$\approx \prod_{i=1}^{n} P(w_i \mid c_{k,i})$

对于 $s_{lex}^*(lex_k)$中的第一项 $P(c_{k,1}^{k,n}) = \prod_{i=1}^{n} P(ck,i \mid c_{k,1}^{k,i-1}) \approx \prod_{i=1}^{n} P(c_{k,i} \mid c_{k,i-1})$ 二元模型

或 $\approx \prod_{i=1}^{n} P(c_{k,i} \mid c_{k,i-1}, c_{k,i-2})$ 三元模型

因此 $s_{lex}^*(lex_k) \approx \prod_{i=1}^{n} P(c_{k,i} \mid c_{k,1}^{k,i-1}) \times p(w_i \mid c_{k,i})$

$\approx \prod_{i=1}^{n} P(c_{k,i} \mid c_{k,i-1}) \times p(w_i \mid c_{k,i})$ 二元模型

或 $\approx \prod_{i=1}^{n} P(c_{k,i} \mid c_{k,i-1}, c_{k,i-2}) \times p(w_i \mid c_{k,i})$ 三元模型

3 句法打分函数

（1）如果假设一个规则的应用独立于其他规则的应用,并且假设这个规则的应用独立于上下文,也即用概率上下文无关文法来给句法打分,下面用图一的句法结构树作为例子。

$s_{syn}(Tree_X) \approx P(syn_{j,k} \mid lex_k) = P(A|BC) \times P(B|FG) \times P(C|DE) \times P(E|c_4)$ (1)

其中公式（1）中每个概率的求法第三部分给出。

（2）如果假设一个规则的使用不独立于其它规则的使用，并且这个规则的使用也不独立于其上下文，则图一的句法打分函数便为：

$S_{syn}(syn_{j,k}) \approx P(A \mid \{\varnothing\}, B, C, \{\$\}) \times P(C \mid \{\varnothing\}, F, G, \{B\}) \times \cdots \times P(F \mid \{\varnothing\}, c_1, \{c_2, c_3, c_4\}) \approx P(A|L_7, B, C, r_7) \times P(C|L_6, D, E, r_6) \times \cdots \times P(F|L_1, C_1, r_1)$

数。随着翻译句子的增多，放进语料库中的句子也越来越多，人工要确定出每个句子的正确分析似乎越来越难以胜任，此时便可根据算法 2 来进行统计。

四、平滑算法

令 N 为训练标识数，n_r 为发生 r 次的事件数，则下面的公式成立：

$$N=\sum_r r \times n_r$$

事件 e 发生 r 次的概率的最大可能估计为：$P_{ML}(e)=\frac{r}{N}$

根据[5]中 Turing 的公式，事件 e 发生 r 次的概率为：

$$P_{TU}(e)=\frac{r^*}{N} \quad \text{其中 } r^*=(r+1)\frac{n_{r+1}}{n_r}$$

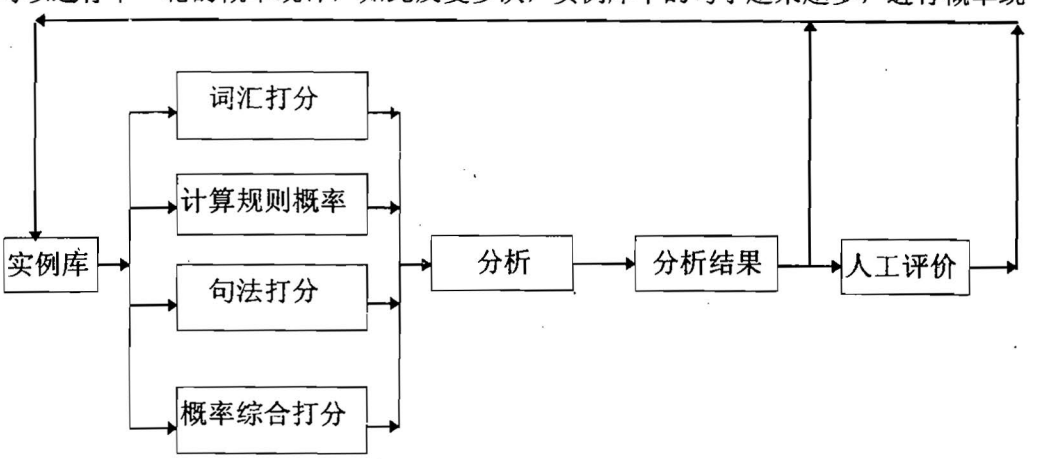
使用 Turing 的公式，真正发生在训练集的所有事件概率的和为：

$$\sum_{e:c(e)>0} P_{TU}(e)=1-\frac{n_1}{N}$$

则从没发生的事件概率和为：$\sum_{e:c(e)=0} P_{TU}(e)=\frac{n_1}{N}$

五、概率处理排歧的框架及实验结果

对实例库（含树库）中的句子分别进行词汇、句法、规则和概率综合打分统计，统计出来的结果用来指导翻译的分析阶段，然后，把规则方法与概率打分函数结合起来分析的结果经过人工校对，便得出正确的分析和错误的分析，最后再把正确的分析放进实例库。这样可以进行下一轮的概率统计，如此反复多次，实例库中的句子越来越多，进行概率统



图二　统计排歧处理的框架

计也越来越客观。统计流程为：分析正确的实例库⟶词汇打分⟶计算规则概率的句法打分⟶概率综合打分⟶概率统计结果来指导分析⟶人工校对分析结果⟶正确的分析结果放进实例库。

如果翻译出的句子特别多，人工校对很难胜任时，那么没有经过人工校对的句子放进放进语料库后，计算规则使用的概率时可以利用上一轮分析时的句法打分函数来累加（第三部分中的算法 2 ）。统计流程为：实例库⟶词汇打分⟶计算规则的概率⟶句法打分⟶概率综合打分⟶概率统计结果来指导分析⟶分析结果放进实例库。

本文使用我们的汉英机器翻译系统翻译出的句子，共挑选出 3500 个已被分析的句子，这些句子基本上覆盖了现代汉语的基本句型，把这些句子分成 3000 个句子的训练集和 500 个句子的测试集，训练集的平均句长为 9.32，测试集的平均句长为 8.66，我们的系统共有短语结构规则 308 个，这些规则是由词类标记 18 个，短语标记 12 个生成。使用二元模型并在句法打分时考虑左边一个上下文信息，对于训练集词类标注准确率为 96.7%，分析树的准确率为 75.3%，对于测试集词类标注准确率为 88.3%,分析树的准确率为 50.4%。

六、结论

没有一种方法可以完全解决排歧问题。我们综合运用了词汇知识、语法知识来统计。在我们的系统中，对于语法约束、语义约束都难以解决的问题，统计排歧可以作为有益的补充，对于运用这些知识和统计结果之后，还存在歧义或分析不正确的句子则需人来校正。

进一步的工作应是利用概率方法来对语义知识进行统计，或者把词汇、语法和语义综合起来进行统计。

参考文献

[1]. Su Keh-Yih, and Chang Jing-Shin, 1988, <<Semantic and syntactic aspects offscore function.>>, In Proceedings,12th International Conference on Computational Linguistics. Budapest, P22-27 ;

[2] Su Keh-Yih, Chiang Tung-Hui, and Lin Yi-Chung, 1991,. << A robustness and discrimination oriented score function for intergrating speech and language processing>> In Proceedings, 2nd European Conference on Speech Communication and Technology. Genova,P207-210;

[3] Su Keh-Yih, Chiang Tung-Hui, and Lin Yi-Chung, 1992, <<A discriminative approach for ambiguity resolution based on a semantic score function>> In Proceedings,1992 International Conference on Spoken Language Processing. Banff,P149-152;

[4] Chen Shu-Chuan, Chang Jing-Shin, Wang Jong-Nae, and Su Keh-Yih, 1991, <<ArchTran: A corpus-based tatistics-oriented English-Chinese machine translation system>> In Proceedings, Machine Translation Summit III. Washington, D.C.,P33-40;

[5] Good I. J., 1953, << The population frequencies of species and the estimation of population parameters>> Biometrika,40,P237-264;