利用覆盖歧义检测法和统计语言模型进行汉语自动分词 [1]

王显芳 杜利民

(中国科学院声学研究所语音交互信息技术研究中心 北京 100080)

摘 要 该文探讨了利用覆盖歧义检测法和统计语言模型进行汉语自动分词的问题. 采用了多次迭代的方法来进行汉语词层面统计语言模型的训练. 该方法能够得到更优化的语言模型. 该文详细介绍了统计语言模型的训练过程, 给出了语言模型复杂度随迭代次数增加而减小的实验结果. 还给出了在不同的统计语言模型阶数下切分正确率变化的情况, 分析了切分正确率变化的原因.

关键词 统计语言模型, 覆盖歧义检测法, 自动分词

中图号 TP391.4, TN912.3

1 汉语切分和覆盖歧义检测法

对于汉语而言, 词是具有固定意义的最小单元. 但汉语的书写以字为基本单位, 词与词之间没有区分标识. 将汉语句子切分成词是汉语信息处理的前提 [1,2] . 对于切分歧义字段的处理, 是汉语分词要解决的主要问题之一. 切分歧义一般分为两种: 一种是交叉歧义切分, 另一种是覆盖歧义切分 [2−4] .

交叉歧义切分是指, 一个汉字串包含 A 、 B 、 C 三个子串, 若 AB 、 BC 分别构成词, 则该汉字串有两种切分形式, AB/C 和 A/BC[5,6] . 如汉字串 "着重要" 存在两种切分形式, "着 / 重要" 和 "着重 / 要". 比如说在句子 "在整个世界局势起着重要作用" 中, 应该切为 "着 / 重要". 而在句子 "着重要解决资金问题" 中, 应该切分为 "着重 / 要".

覆盖歧义切分是指包含至少两个汉字的汉字串, 它本身是词, 切开也是词 [2] . 比如说 "马上". 在句子 "他从马上跳下来" 中, 应切分为 "马 / 上". 在句子 "我马上就下楼" 中, 应切分为 "马上".

目前, "长词优先" 准则是解决覆盖歧义的最常用也是最实际有效的一种切分准则 [2] . 所谓 "长词优先" 准则, 就是尽可能地用最长的词来匹配句子中的汉字串. 比如说 "社会主义"、"社会" 和 "主义" 都是词. 当我们在句子中遇到 "社会主义" 这个汉字串时, 就会用 "社会主义" 这个词来匹配它, 使得切出来的词尽可能长. "长词优先" 准则在一定程度上模拟了人工分词的心理过程. 对于绝大多数情况, "长词优先" 准则是适用的.

覆盖歧义检测法采用 "长词优先" 准则, 能够检测所有的交叉歧义, 同时忽略所有的覆盖歧义. 它输出的切分路径集称为最大无覆盖歧义切分路径集. 此集合满足如下条件: 在给定词典的条件下, 一个句子的所有切分路径构成了一个集合 P , 它必然存在一个不包含覆盖歧义的切分路径的子集合 $Q \subseteq P$, 而对任给句子的一种切分路径 $x \in P$, 都能够找到一种切分路径 $y \in Q$, 使得 y 与 x 之间只存在覆盖歧义而不存在交叉歧义.

以句子 "结合成分子时" 为例, 其最大无覆盖歧义切分路径集包含的切分路径数为 4 , 分别为 "结 / 合成 / 分 / 子时"、"结合 / 成 / 分子 / 时"、"结 / 合成 / 分子 / 时"、"结合 / 成分 / 子时".

最大无覆盖歧义切分路径集的意思就是如果向该集合当中加入一种不属于该集合的切分路径, 则此切分路径必然和集合中一种切分路径存在覆盖歧义; 而如果从此集合中删除一种切分路径, 必然会导致句子的一些切分路径无法在该集合中找到与之只存在覆盖歧义而不存在交叉歧义的切分路径.

[1] 2002-03-19 收到, 2002-08-09 改回

对于不存在交叉歧义的句子，其最大无覆盖歧义切分路径集中只存在一种切分路径．而对于存在交叉歧义的句子，其最大无覆盖歧义切分路径集中存在多种切分路径．在此情况下，就需要利用其它知识在多种切分路径中选择一种切分路径．统计语言模型可以被用来实现这一功能 [8]．下面简单介绍统计语言模型的一些基本知识．

2 统计语言模型和汉语分词

将人类平常所说的句子、文章看成一个个基本单元 (音节、字、词) 的串性连接，假设这些基本单元是由一个发生器产生的，该发生器一个一个地产生基本单元，从而生成这些句子和文章．引入信息论的观点，将该发生器看成一个 Markov 信源，该信源的基本符号集合为人类语言的基本单元 (音节、字、词)，这些符号构成 Markov 链 [8]．

假设信源符号集合为 $V=\{v_1,v_2,\cdots,v_L\}$，符号个数为 L．

信源发出的符号串为 $W=w_1w_2\cdots w_n$，$w_i\in V$，$i=1,2,\cdots,n$．

将该符号串的产生过程看成一个 Markov 过程，则产生 w_i 的概率仅与 w_i 出现之前的符号有关，w_i 的产生概率为 $P(w_i|w_{i-1,1})=P(w_i|w_1w_2\cdots w_{i-1})$，符号串 W 的概率由概率乘法公式得到 $P(W)=P(w_1w_2\cdots w_n)=P(w_1)\prod_{i=2}^{n}P(w_i|w_1w_2\cdots w_{i-1})$

在建立语言模型时，通常把语言的产生过程简化为 $N-1$ 阶马尔可夫过程．即对于 w_i 的存在概率，只考虑 w_i 之前的 $N-1$ 个语言单元的影响，忽略其他上下文信息的影响．此时，w_i 的存在概率采用 $N-1$ 阶条件概率 $P(w_i|w_{i-1,i-N+1})$ 刻画，因此，上式转化为 $P(W)=\prod_{i=1}^{M}P(w_i|w_{i-1,i-N+1})$．$P(w_i)|w_{i-1,i-N+1})$ 的值可由训练数据得到．设 $N(w_{i,i-N+1})$ 为 $W_{i,i-N+1}$ 在语料中出现的频度，$N(w_{i-1,i-N+1})$ 为 $W_{i-1,i-N+1}$ 在语料中出现的频度，根据大数定理，语料容量越大，$N(w_{i,i-N+1})/N(w_{i-1,i-N+1})$ 就越逼近于 $P(w_i|w_{i-1,i-N+1})$，因此，在语料容量较大时，可以近似认为 $P(w_i|w_{i-1,i-N+1})=N(w_{i,i-N+1})/N(w_{i-1,i-N+1})$．

统计语言模型的评价有一个标准，即看它的复杂度 (perplexity, 又称 PP 值) 的大小．PP 值是从信息熵的角度得到的．语言模型的信息熵定义为 $H(P_M)=\lim_{n\to\infty}-\frac{1}{n}\log P_M(w_1w_2\cdots w_n)$．

语言模型的复杂度可以通过下式来计算，$\mathrm{PP}=2^{-(1/T)\log P(w_1w_2\cdots w_T)}$，其中 $\boldsymbol{T}$ 为语料库的容量．

语言模型的 PP 值越小，表明语言模型的约束能力就越强，语言模型也就越好．

在利用覆盖歧义检测法和统计语言模型进行分词时，设 S 为要切分的句子，其最大无覆盖歧义切分方式集中共有 N_S 条切分路径，将每条路径用一个对应的矢量 $\boldsymbol{T}_i$ 表示．$\boldsymbol{T}_i=(t_{i1},\cdots,t_{i,ni})$，$t_{i1}+\cdots+t_{i,ni}=S$, 则最优的切分路径为

$$\boldsymbol{T}_j=\arg\max_i P(\boldsymbol{T}_i)=\arg\max_i P(t_{i1},\cdots,t_{i,ni})$$

其中 $P(t_{i1},\cdots,t_{i,ni})$ 可利用统计语言模型计算得到．

统计语言模型利用基本语言单位 (词、字、音) 共同出现的概率关系来表示语言基本单位之间的依赖关系．它可以用明确的数学公式来表达，算法上也容易实现．在连续语音识别、中文输入、信息检索、机器翻译等领域中，统计语言模型都起到了非常重要的作用．

3 统计语言模型的建立过程

统计语言模型的训练需要首先得到已切分成词序列的汉语语料．覆盖歧义检测法不能输出唯一的结果，因此只能利用其他切分方法得到初始已切分语料．后向最大匹配法 (Reverse Maximum

Matching, RMM) 切分方法简单，切分正确率高 [2]，是得到初始已切分语料的首选切分算法。选用后向最大匹配法作为得到初始已切分语料的首选切分算法的另一原因是 RMM 的切分准则也是长词优先准则，这样就和后续处理中的覆盖歧义检测法所使用的切分准则是一致的。

我们训练语言模型的过程如下 (图 1)：首先，使用 RMM 切分汉语语料，统计初始已切分语料词的出现频率，得到初始的 UGram 模型；然后利用覆盖歧义检测法和初始的 UGram 模型，再对语料进行切分，得到新的 UGram 模型。随着迭代次数的增加，UGram 模型的复杂度将逐渐降低。如此反复几次后，当语言模型复杂度的变化值小于某一阈值时，循环终止。

BiGram 模型的训练方法和 UGram 的方法类似。首先，利用覆盖歧义检测法和已经得到的 UGram 模型，切分汉语语料，统计出已切分语料中词的共现概率，得到初始的 BiGram 模型。然后利用覆盖歧义检测法和初始的 BiGram 模型，再对语料进行切分，得到新的 BiGram 模型。如此反复数次后，当语言模型复杂度的变化值小于某一阈值时，循环终止。TriGram 模型的训练方法也类似。

训练统计语言模型的另一个问题就是数据稀疏问题 [7]。根据大数定理，$N(w_{i,i-N+1})/N(w_{i-1,i-N+1})$ 逼近于 $P(w_i|w_{i-1,i-N+1})$ 的条件是训练数据有足够的容量。但在实际语料库中语料的容量毕竟是有限的，往往达不到这个要求。随着阶数 N 的增大，NGram 模型参量估计的可靠性不断降低。因此需要采用平滑技术来平滑统计数据。我们采用的是删除插值方法 [8]。

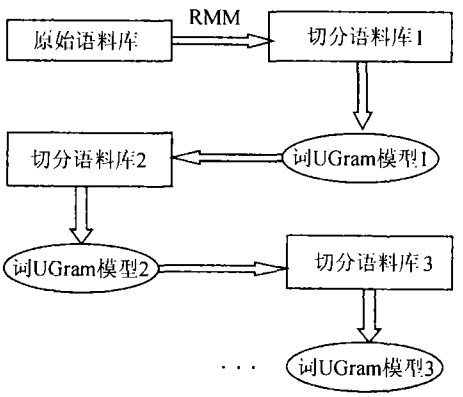

图 1 UGram 模型的训练过程

4 实验结果

我们的语料库为 1990–2000 共 11 年的《人民日报》和 1994 年《市场报》以及 1994 年《中国百家报刊精选》。语料容量共约 520Mbyte，总共有 17494391 句。所使用的词典为北航的词典，词条数为 67480 个，最大词长为 7，其中包含了所有的汉字单字。

根据覆盖歧义检测法的切分结果，这些语料的最大无覆盖歧义切分所集中的切分路径数为 22459830, 平均 1.284 种 / 句. 切分路径个数不为 1 的句子是包含交叉歧义字段的, 这样的句子总数为 3549053，约占句子总数的 20%。这些句子的最大无覆盖歧义切分路径总数为 8514492，平均 2.34 种 / 句。

4.1 统计语言模型的复杂度

下面我们给出训练 UGram，BiGram 和 TriGram 的语言模型复杂度随迭代次数增加而变化的情况，如图 2, 图 3, 图 4 所示。我们发现，随着迭代次数的增加，统计语言模型的复杂度将逐渐减少。

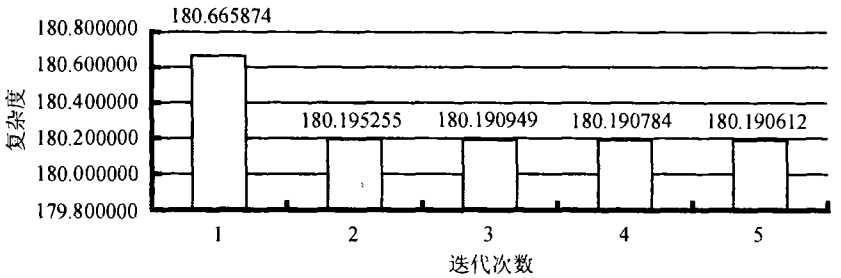

图 2　UGram 模型复杂度随迭代次数增加而减少

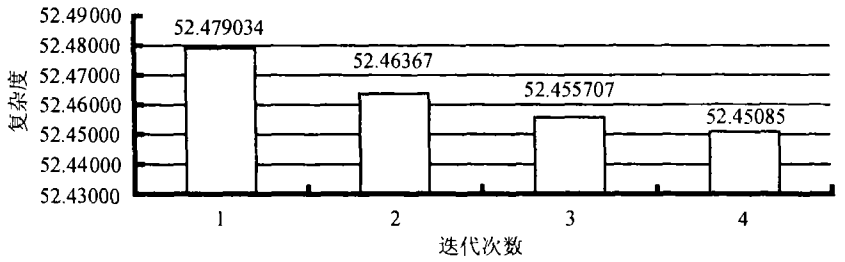

图 3　BiGram 模型复杂度随迭代次数增加而减少

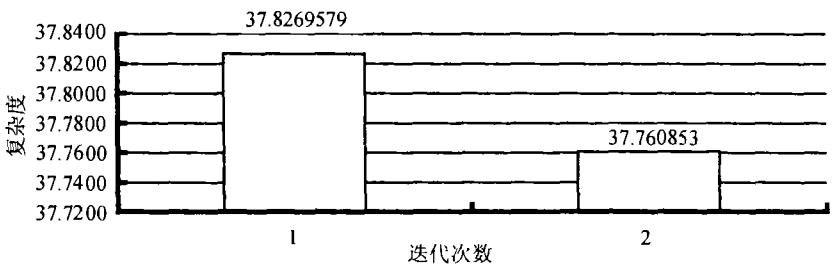

图 4　TriGram 模型复杂度随迭代次数增加而减少

4.2 统计语言模型的阶数不同时的切分正确率

为了测试不同的统计语言模型阶数对切分正确率的影响，我们手工标注了 1994 年人民日报的一段语料，测试语料共 13866 行，174310 字．其中有交叉歧义的句子共 2357 行，39931 字．下面给出后向最大匹配法、覆盖歧义检测法 +UGram 、覆盖歧义检测法 +BiGram 、覆盖歧义检测法 +TriGram 时的句子正确率和词正确率，如图 5, 图 6 所示．

	RMM	UGram	BiGram	TriGram
词正确率(%)	93.42	98.96	99.11	98.97
句子正确率(%)	65.76	94.44	95.25	94.48

图 5 包含交叉歧义句子的句子切分正确率和词切分正确率

	RMM	UGram	BiGram	TriGram
词正确率(%)	98.55	99.76	99.8	99.77
句子正确率(%)	94.18	99.06	99.19	99.06

图 6 全部测试语料的句子切分正确率和词切分正确率

5 总结和讨论

覆盖歧义检测法是基于“长词优先”原则的. 与其它基于“长词优先”原则的切分算法相比，它在忽略覆盖歧义的同时保留了所有的交叉歧义，从而提供了一种能够对覆盖歧义和交叉歧义分开处理的方法. 覆盖歧义检测法并没有直接输出句子的唯一切分结果，还需要利用其他知识进行切分排歧. 统计语言模型利用词共同出现的概率关系来表示词之间的依赖关系，反映了语言的一些统计特性，恰好可以用来在多种候选切分结果之间选择一种切分结果.

使用统计语言模型需要对语言模型进行训练，因此需要一种切分方法将训练语料切分成词序列. 后向最大匹配法切分方法简单，所采用的切分准则和覆盖歧义检测法所使用的切分准则是一样的，是得到初始切分语料的首选切分算法. 在得到初始语言模型后，可以通过迭代方法减少语言模型的复杂度，得到更优的语言模型. 由结果所知，模型复杂度的减少并不非常明显，但这在语音识别等应用中还是有一定作用的.

通过对切分正确率的比较，可以发现，后向最大匹配法切分正确率不高. 覆盖歧义检测法 +UGram 切分正确率与之相比有了明显提高，覆盖歧义检测法 +BiGram 的切分正确率又有所提高. 但覆盖歧义检测法 +TriGram 却低于覆盖歧义检测法 +BiGram . 这个现象可以这样解释：一方面，随着统计语言模型阶数的增加，模型精确率逐渐提高. 模型精确度的提高带来切分正确率的提高. 另一方面，因为初始训练语料是由后向最大匹配法切分所得的，而后向最大匹配法的切分正确率不高，也就是初始训练语料中包含了一些切分错误. 随着统计语言模型阶数的增加，切分结果就越有可能再现初始训练语料中的切分错误. 在采用 BiGram 模型时，模型精确度提高因素的影响超过了初始切分错误再现因素的影响，切分正确率达到最高，而在 TriGram

模型时，第一个因素的影响被第二个因素的影响过分抵消，正确率稍有下降。如果训练语料是全部正确切分的语料，则 TriGram 切分正确率将是最高的。

需要指出的是，覆盖歧义检测法并没有考虑未录入词问题，上述实验中仅仅是考虑了交叉歧义，结果当中的切分正确率可以看成是交叉歧义排歧正确率。

参 考 文 献

[1] 刘开瑛，中文文本自动分词和标注，上海，商务印书馆，2000, 30–41.

[2] 陈小荷，现代汉语自动分析，北京，北京语言文化大学出版社，1999, 60–62.

[3] 马晏，基于评价的汉语自动分词系统的研究及实现，语言处理专论，北京，清华大学出版社，1996, 80–105.

[4] 侯敏，孙建军，陈肇雄，汉语自动分词的歧义问题，计算语言学进展与应用，北京，清华大学出版社，1995, 40–43.

[5] 沈达阳，孙茂松，基于统计的汉语分词模型及其实现方法，BYTE China，重庆，1998, 2(2), 38–40.

[6] 孙茂松等，高频最大交集型歧义切分字段在汉语自动分词中的运用，中文信息学报，1999, 13(1), 60–62.

[7] 王雪松，汉语语言的多层面优化统计语言模型研究，[硕士论文]，中科院声学所，1997, 13–15.

[8] 张瑞强，用于汉语连续语音识别中的语言模型的研究，[博士论文]，清华大学，1997, 20–27.

AUTOMATIC SEGMENTATION OF CHINESE USING OVERLAYING AMBIGUITY EXAMINING METHOD AND STATISTICS LANGUAGE MODEL

Wang Xianfang　　Du Limin

(*Center for Speech Interactive Information Technology,*

Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080, *China*)

Abstract　In this paper, the question of Chinese automatic segmentation is discussed using overlaying ambiguity examining method and statistics language model. The multi-time iterative method is applied to train language model, which can produce a better model. The process of training language model is described in detail. The result shows that the perplexity of language model is reduced. The accuracy of segmentation changes with different language model and the reason is analyzed.

Key words　Statistics language model, Overlaying ambiguity examining method, Automatic segmentation

王显芳：　男，1975 年生，博士生，主要研究领域为语音识别、对话系统.

杜利民：　男，1957 年生，博士、研究员、博士生导师、IEEE 高级会员、中国电子学会理事、《电子学报》编委、《电子科技导报》编委. 研究方向为语音识别、自然语言理解、语音交互信息技术.