第 21 卷 第 3 期
2002 年 6 月

情 报 学 报

Vol. 21, № 3
June, 2002

基于 EM 算法的汉语自动分词方法[1)]

李家福 (解放军理工大学通信工程学院,南京 210016)

张亚非 (解放军理工大学理学院,南京 210016)

摘要 汉语自动分词是中文信息处理中的基础课题。本文首先对汉语分词的基本概念与应用,以及汉语分词的基本方法进行了概述。接着引出一种根据词的出现概率、基于极大似然原则构建的汉语自动分词的零阶马尔可夫模型,并重点剖析了 EM(Expectation-Maximization)算法,对实验结果进行了分析。最后对算法进行了总结与讨论。

关键词 分词 汉语 EM 算法 语料库 HMM

Segmenting Chinese by EM Algorithm

Li Jiafu

(*Institute of Communication Engineering, PLA University of Science & Technology, Nanjing* 210016)

Zhang Yafei

(*Institute of Science, PLA University of Science & Technology, Nanjing* 210016)

Abstract Word segmentation is a basic task of Chinese information processing. In this paper we present a simple probabilistic model of Chinese text based on the occurrence probability of the words, which can be seen as a zero-th order hidden Markov Model(HMM). Then we investigate how to discover by EM Algorithm the words and their probabilities from a corpus of unsegmented text without using a dictionary. The last part is conclusion and discussion about the algorithm.

Keywords word segmentation, EM Algorithm, corpus, HMM.

1 引 言

自然语言处理是人工智能的重要分支。词是自然语言处理系统中重要的知识载体与基本操作单元。在书面汉语中词与词之间没有明显的切分标志。于是在中文信息处理中汉语自动分词这一研究领域应运而生,并成为中文信息处理中的基础课题。

自动分词系统在汉语分析与理解、机器翻译、中文文献自动标引或全文信息检索、汉字识别、汉语语音识别与合成、中文简繁体自动转换及文本处理(中文文稿自动校对)等领域中得到广泛的应用。

本文分析的算法基于生语料库,不用词典,根据词的出现概率和极大似然原则(Maximum Likelihood Principle)进行分词[1]。

收稿日期:2001 年 5 月 18 日

作者简介:李家福,男,1974 年生,博士研究生。主要研究方向为自然语言处理、数据挖掘、数字图书馆、电信网网管。张亚非,男,1955 年生,复旦大学外文系语言学与话语分析专业毕业,博士。现为解放军理工大学理学院院长、教授、博士研究生导师。研究领域为:军事通信学、自然语言处理。

1) 文章得到国家自然科学基金项目(编号 69975024)、国家自然科学基金重点项目(编号 69931040)资助。

2 汉语分词及其概率模型

汉语中词的抽象定义(即‘词是什么’)与具体判定(即‘什么是词’)问题在语言学界并未完全解决。中国大陆制订了国家标准《信息处理用现代汉语分词规范》,在《规范》中提出用‘分词单位’来替代‘词’的位置,并给出了一套比较系统的‘元规则’。这个《规范》虽对中文信息处理研究产生了一定程度的积极影响,但并未能在‘词’层妥善解决问题。问题的实质在于,除定性信息外,必须引入定量信息,分词处理、分词词表的构造,应该和汉语语料库结合起来考虑[2]。

国内外自 80 年代以来,已陆续开发出一些分词系统,其使用的分词方法可归结为以下 3 类[3]:

(1)基于词典的方法(Lexical Method);

(2)基于统计的方法(Statistical Method);

(3)混合方法(Hybrid Method)。

基于词典的方法如纯粹根据词表机械地作字符串匹配,则可能存在‘交集型’歧义切分字段(Intersecting Ambiguity)和‘包孕型’歧义切分字段(Combined Ambiguity),或是这两种基本类型的变体或组合[2,3]。纯词典方法中还存在如下问题:

● 词典捕捉的是语言特定时期的特定状态。没有哪个词典是完备的,因为语言中常出现新的词语,也没有一个词典能囊括所有领域的词语。

● 词缀:一些词语可以通过词尾变化生成,如复数后缀‘们’可以加到代词和人称名词后。

● 名称:汉字的人名通常包含一个单音节的“姓”和一个双音节的‘名’。

● 译名:不同的汉语方言采用不同的译名。如‘Hollywood’在普通话中译为‘好莱坞’,在粤方言(香港)中译为‘荷里活’。

基于统计的方法基于(两个或多个)汉字同时出现的概率,通过对语料库有监督或无监督的学习,得到描述一种语言的‘语言模型’(常用一阶隐马尔可夫模型(1st HMM))。基于统计的方法有许多优点:生词和名称(包括译名)的影响降低了,只要有足够的训练文本就易于创建和使用。

混合方法主要是将两种方法结合在一起,吸取两种方法的优点。

本文分析的算法是一种基于统计的分词算法,它基于词的出现概率构建汉语分词的概率模型。算法基于以下假定:

● 长度为 1,2,…,K(如 K = 4)词的数量(即使是非常大的)是有限的;

● 每个词都有一个未知的出现概率;

● 词相互独立,即两个词同时出现的概率仅与各自的出现概率有关。

给定词的出现概率,根据极大似然原则 MLP (Maximum Likelihood Principle),一个句子分成词语 $w_1, w_2, \cdots, w_k$,须使 $\Pi p(w_i)$ 最大,其中概率 $p(w_i)$ 是词 w_i 的出现概率。

如下例中,将句子 $C_1\ C_2\ C_3$(其中 C_j 代表一个汉字,以下同)切分为词 $w_1\ w_2 \cdots\ w_m$(其中 w_i 代表一个词),有 4 种可能的切分,其中切分 2 可能性最大,被选定。

表 1 极大似然原则例示

序号	选定	切分	可能性
1		$C_1 \wedge C_2 \wedge C_3$	0.005
2	√	$C_1\ C_2 \wedge C_3$	0.08
3		$C_1 \wedge C_2\ C_3$	0.001
4		$C_1\ C_2\ C_3$	0.03

那么,根据 MLP,如果已知二元组集 $\{w_i, p(w_i)\}$,就可以对文本进行分词处理。本模型可以看作是 HMM(零阶隐 Markov 模型)。EM 算法采取一种特殊的方法训练模型。

3 EM 算法

从以上的分析可以看出,根据 MLP 进行分词处理,未知参数是每个词的概率。如果有‘熟’训练语料库(文本已进行分词处理),就可以对其中的词计数并算出概率。参见图 1,如训练语料库已进行分词处理,就可以得到词集 $\{w_i\}$,并可通过词计数算出词的出现概率 $p(w_i)$。

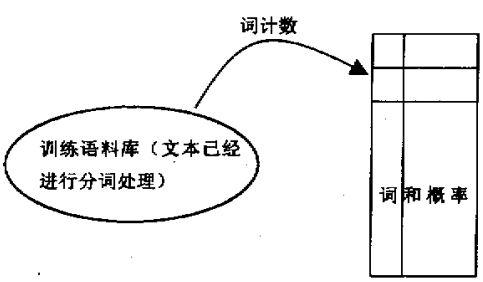

图 1 从‘熟’训练语料库得到词和概率

反之,如果已经知道词和概率,根据 MLP,就可以对句子进行分词处理(图 2)。

这种情况类似‘先有鸡,还是先有蛋’的问题。EM(期望最大值 Expectation Maximization)算法这样来解决这个问题(参见图 3):

● 先拿出一个“蛋”,即随机给出词的概率的初始值;

● 使用当前的词的概率值,对语料库中的句子进行分词处理;

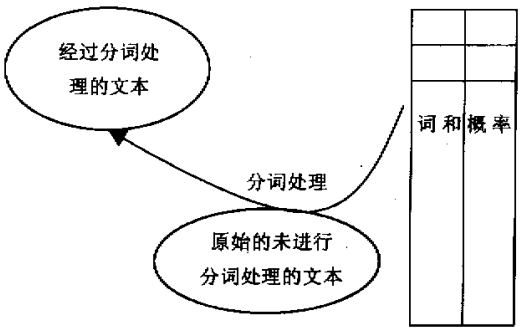

图 2　根据词和概率对文本进行分词处理

● 依据分词结果,重新计算词的概率;

● 重复这个过程,进行多次迭代,直到概率值收敛。

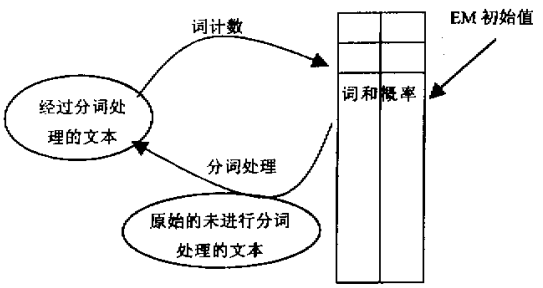

图 3　EM 算法的实现过程

EM 算法具体实现过程如下(见图 4):

(1)对未切分文本中的每个句子:

(a)使用当前词的概率值计算每个可能切分的可能性(likelihood);

(b)对切分可能性进行‘归一化’处理为‘尾数’(fraction),使其和为 1;

(c)对每种切分进行词计数,即将切分的‘尾数’加到词数上。

(2)使用词数更新词的概率。

(3)重复这个过程直到收敛。

从(1)(c)可以看出,EM 算法采用了一种特殊的词计数方法:给定一个长度为 n 的句子,其可能的切分有 2^{n-1} 种,根据对词的概率的估算(尽管可能是有缺陷的),可以计算出每种切分的可能性;在一个可能性为 P_i 的切分中,对每个词增加词数 $\frac{P_i}{\sum_{j=1}^{2^{n-1}} p_j}$。称这种词计数方法叫“软计数”(soft-counting)。

“软计数”实现的动态程序如下:

输入为句子:$C_1\ C_2 \cdots\ C_n$,对句子中的每个词语 $C_{j1} \cdots\ C_{j2}$,它的计数必须增加 $S_{j1}^{left}\ p(C_{j1} \cdots\ C_{j2})\ S_{j2}^{right} / \alpha$,其中:

● S_{j1}^{left} 是在 C_{j1} 左侧子字符串所有可能切分的可能性的和;

● $P(C_{j1} \cdots\ C_{j2})$ 是当前词 $C_{j1} \cdots\ C_{j2}$ 概率值;

● S_{j2}^{right} 是 C_{j2} 右侧子字符串所有可能切分的可能性的和;

● α 是归一化常数,是句子所有可能切分的可能性的和,等于 S_{n+1}^{left}。

S_{j1}^{left}、S_{j2}^{right} 使用动态程序计算,如,S_{j1}^{left} 的递归

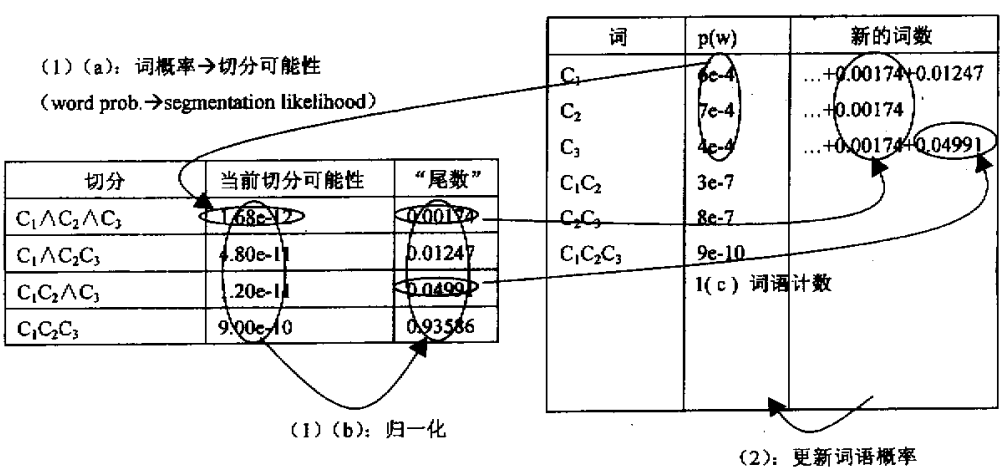

图 4　EM 实现过程示例

函数为：

$$S_{j1}^{left} = \begin{cases} 1 & i = 1; \\ p(C_1) & i = 2; \\ \sum_{j=1}^{i-1} p(C_j \cdots C_{i-1}) S_j^{left} & i = 3; \end{cases}$$

第一遍从左到右扫描计算 S_i^{left}（i = 1，2，…，n + 1），得到 $\alpha = S_{n+1}^{left}$。然后从右向左扫描计算 S_i^{right}（i = n，n - 1，…，3，2，1），同时得出每个词的数量。

本算法的复杂性的 O（*KIN*），其中 k 是单词的最大长度，I 是迭代的次数（通常为 5～10），N 是语料库的大小。

4 实验结果分析

分词系统的性能常用回调率（recall）和精度（precision）来衡量。如系统输出 n_1 个词，正确切分应得出的词数为 n_2，设 c 为两者共有的词数。则，回调率 = c/n_2，精度 = c/n_1。

从表 2 可以看出，同基于词典的算法相比，本算法尽管没有使用字典，并在生语料库上进行训练，分词效果也很好。通过在一个 100Mbyte 的生语料库上训练模型，算法的回调率与精度分别达到 65.65% 和 71.91%。研究发现，大多数分词错误来自 20 个单字符的助词（如‘的’等）。经过一个简单的预处理，将少量的这类词与别的词分开，分词处理的回调率和精度都得到了提高，分别达到 97.72% 和 91.05%[1]。

表 2 分词算法的性能

分词系统	回调率（%）	精度（%）
软计数	65.65	71.91
（经过预处理）	97.72	91.05
基于词典	93.63	95.87

5 总结与讨论

本文在对汉语自动分词简单分析的基础上，重点剖析了一种中文分词的零阶马尔可夫模型和一种在生语料库中训练模型的有效算法。

从本文分析中可以看出，本算法在研究中主要依赖纯粹的统计方法，除利用汉语词长一般为 1，2，…，4 外，没有使用别的汉语知识。因而在未来的研究中应充分考虑以下两个方面的问题：

- 分词知识 知识是分词系统的源泉，分词算法靠分词知识驱动。无论采用基于统计的方法，还是基于词典的方法，或是采用混合方法，都必须靠丰富的分词知识的积累；
- 注重混合方法的开发，即研究本算法如何与基于词典的方法、语法知识、语义规则如词长分布、语法限制等知识结合使用，实现多种分词知识与策略的集成。

参 考 文 献

1 Xianping Ge，Wanda Pratt，Padhraic Smyth. Discovering Chinese words from unsegmented text. SIGIR '99（Proceedings on the 22nd Annual International ACMSIGIR Conference on Research and Development in Information Retrieval，August 15-19 1999，Berkeley CA USA）：217～272

2 Thomas EMERSON. Segmenting Chinese in Unicode. 16th International Unicode Conference，Amsterdam，The Netherlands，March 2000

3 Jay M. Ponte，W. Bruce Croft. Useg：A retargetable word segmentation procedure for information retrieval. Document Analysis and Information Retrieval 96（SDAIR），1996 Sympoium

（责任编辑 许增棋）