用语料库语言学知识指导文本识别研究

常新功　夏莹
北京　清华大学计算机系　100084

摘　要　本文主要讲述了我们近期的研究工作：我们对500万汉语文本进行了字字同现(二元)统计，分别利用BACK－OFF方法和TURING公式法计算得到汉字字字间的二元同现概率矩阵。从结果中分析发现该统计的确能反映汉语语言　的某些规律性。我们把此结果应用到OCR处理中指导文本识别，获得了令人满意的结果。

关键字：二元同现概率　数据稀疏　BACK－OFF法　MARKOV模型

1. 引言

随着国外语料库语言学的发展，近年来我国计算语言学界的一些学者也在这方面做了大量研究。对大规模汉语语料文本进行统计分析，利用互信息、同现信息等结果来进行分词、词性标注、词性排歧等研究，获得了巨大成功。

然而，据目前文献来看，大陆对字字同现的统计工作做的不多，而利用统计方法进行文本处理的应用也很少。目前的文本识别领域中利用语言学识一般限于词一级，即利用词条信息进行前后联想来处理误识、拒识或直接从多候选中选出结果。

台湾在这方面进行了一些研究。工技院电通所开发的“基于统计的中文错字侦错系统”，就用到了汉字间的接续强度等统计知识。中央研究院的“智慧型中文输入方法”，也利用统计得到的一些反映前后文关联的模式(PATTERN)信息。这些应用都获得了巨大成功。

为了更有效地利用上下文知识来提高文本识别率，我们尝试了基于汉字二元同现概率的统计计算方法。首先，我们做了汉字文本的统计工作；然后把统计结果用于指导文字识别。我们认为对汉字的字字同现统计不仅对文本识别、语音识别、汉字输入方法等领域有很大作用，而且对语言学研究也有一定意义。

2. 文本统计

我们在SUN工作站上对500万汉字文本(新华社通讯稿)进行了统计。最后得出一个3763×3763的二元同现矩阵。我们的统计工作采用了3763个标记集，其中一级3755个汉字各为一个标记。下边每组为一个标记：

① 一级之外的所有汉字
② 阿拉伯数字(0－9)
③ 英文字母(A－Z)
④ 句边界类标点(。，！；？)
⑤ 引用类标点左边部(　(　〔〔　〈　【　“等)

本研究由中国国家自然科学基金资助

⑥ 引用类标点右边部( ) 〕 〉 』"等)

⑦ 特殊数字标头符号(3. (2) ⑧ 等)

⑧ 其它符号

这样 3763 个标记集基本能满足真实文本处理需要,而且标记集不致过大。当然在某些应用情况下此标记可能不能满足需要,要加以改进调整。

我们知道,从真实文本统计多元组概率存在数据稀疏(DATA SPARSNESS)问题。即无论统计文本多么大。某些合法的 N 元组可能出现 0 次或很少次。这样,统计同现次数矩阵中可能会有许多 0 值或很小值,所以简单的最大可能方法就使得同现概率不可信。因而我们分别用了 BACK—OFF 和 TURING 公式方法来计算相邻汉字 $W_{i-1}W_i$ 的同现概率 $P(W_i|W_{i-1})$。

2.1 BACK—OFF 法

该方法的思想是:在二元语法不可信时(训练集不足够大,数据稀疏),结合考虑一元语法。计算公式如下:

$$P(W_i|W_{i-1}) = \alpha F(W_i|W_{i-1}) + (1-\alpha)F(W_i) \tag{1}$$

$$F(W_i|W_{i-1}) = N(W_{i-1}W_i)/N(W_i) \tag{2}$$

$$F(W_i) = N(W_i)/NT \tag{3}$$

N( ): ( )出现的次数;

NT: 训练文本规模(总字数);

公式中的 α 可用 MARKOV 插值来求,我们经试验选择了 α=0.8 来计算。

2.2 TURING 公式法

TURING 公式用作计算多元组概率的语言学模型以成功地用于 IBM 语音识别系统。国外许多学者对之进行了各种改进,我们的工作采用的计算公式是基于 KATZ 于 1987 年提出的公式的一种简化,公式如下:

$$P(W_i|W_{i-1}) = \begin{cases} ① N(W_{i-1}|W_i)/N(W_{i-1}) & N(W_{i-1}|W_i) > K \\ ② D_r \times N(W_{i-1}|W_i)/N(w_{i-1}) & r = N(W_{i-1}|W_i) \\ & 1 <= r <= K \\ ③ N_1/NT & N(W_{i-1}|W_i) = 0 \end{cases} \tag{4}$$

$$D_r = \frac{\frac{R^*}{R} - (K+1)\cdot\frac{N_{k+1}}{N_1}}{1-(K+1)\cdot\frac{N_{k+1}}{N1}} \tag{5}$$

$$R^* = \frac{(R+1)\times N_{R+1}}{N_R} \tag{6}$$

N_r: 出现 r 次的二元组个数;

N(): ( )出现次数;

NT: 训练文本集大小;

公式中 K 值可经验选择,KATZ 推荐 K=5 左右。

经过上述计算,可得到一个汉字二元同现概率矩阵,该矩阵可服务于各种文本处理。

我们仅对 BACK—OFF 的结果做了进一步分析和应用,对 TURING 计算公式对汉语有效性的验证有待进一步的试验。

3. 统计结果分析

我们对统计所得矩阵(特别是同现在100次以上的5282对二元组)进行了初步分析,发现统计有效地反映了汉语的某些规律,特别是汉字间的依存关系。所包含的以下知识对文本处理非常有用:

3.1 较好地反映了常用词条信息:

对汉语中常出现的词,我们发现对应的二元组都有较高的同现概率。不仅两字词,多字词的相邻字两两同现概率也同样很高。因而可以说统计矩阵包含了常用汉字的前联想、后联想汉字集合。另外,统计最新的文本还反映出一些新词。例如,"暂"字为首的词条,在《新编实用汉语词典》中收入了"暂时、暂且、暂行、暂缓"4个词条,该4个词条在统计中都出现过若干次,而"暂时、暂行、暂停、暂不"具有很高的同现概率。这说明"暂缓"类似的词在现代新闻通讯中用得不多,而"暂停"等未收入词条用得很频繁。因而统计结果的词条知识更具领域倾向性,更能反映词的使用频度,并且能与不断变化的语言保持一致。

3.2 反映出非词的常用搭配知识

在统计结果中同现概率很高的二元组中除词条知识外还包括了许多常用搭配知识。在汉语中有些虚词可与某些字结合出现,没有实词含义,仅起语法作用或表示程度、方式、状态等。例如:"把这"、"得很"、"成了"、"不可"、"们的"等二元组就具有很高的同现概率。诸如"把···","被···","···的"等常用搭配都可由统计结果反映出来。

还有一类字常与数字搭配。例如,"仅···","有···","长···","(几分)之···","。··千","···年","···公(里)"等字。这类搭配关系同样也可由同现概率值大小反映出来。同理,与字母、引用类符号等常用搭配同样具有很高同现概率值。

3.3 常用句头字、句尾字信息

我们的统计结果能反映出汉字充当句首、句尾字的可能性大小。从中发现一级汉字中有1270个字(如峦、虑、肪等)和二级所有汉字不做句首字。而最常出现于句首的字(平均至少每一万字文本中做一次句首字)有85个(如并、不、从、而等)。同样,统计结果也反映了句尾字类似特征。这种汉字充当句首、句尾字可能性大小的知识对于以句子为单位的各类文本处理可能是有用的。

3.4 一些字频信息

我们的统计发现在现代通讯稿中常用字(每4万文本中至少出现一次的字)仅有1818个。所以在某些汉字处理中没有必要追求过大字量。在500万文本中,二级字仅出现6014字次,即一级汉字覆盖文本99.87%,而一级字中有108个字从未出现过。因而我们的统计工作仅包括一级字是明智的。

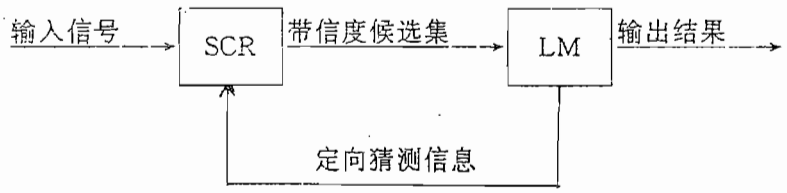
正如我们从BACK—OFF公式看到的,二元同现概率也结合考虑了单字出现的频度信息。

4. 指导文本识别

从上述的分析看,同现概率统计包含了较词条库、前后联想字库更多的信息。

为了在文本识别中更多地利用语言学知识以提高识别正确率,我们尝试了应用该统计结果指导文本识别中,采用方法如下:

这里我们把汉字识别系统分为两个部分：单字识别(SCR)和语言学模型(LM)部分，如下图示。其中 LM 采用汉语 MARKOV 统计模型。



下边我们简单介绍我们的处理过程，我们的处理是以确定边界的字序列(通常为句子)为单位的。

4.1 概率计算公式

一个识别系统要从一输入符号串 $S(S_1\ S_2\cdots S_n)$识别出汉字序列 $W(W_1\ W_2\cdots W_n)$由于输入符号特征的失真，SCR 部分不能确定其对应的汉字而是给出一个带信度值的多个候选。而 LM 部分则考虑 S 可能对应的所有汉字序列，对每一个字串进行概率赋值，最后选择最佳输出。形式化地描述为：

$S=<S_1\ S_2\cdots S_n>$ 为一个确定边界符号串，

$WS=<W_1\ W_2\cdots W_n>$ 表示一个可能的汉字串，

$P(WS|S)$ 表示输入 S 时，结果为 WS 的概率，

$WS^{\wedge}$ 为识别结果，当 $P(WS^{\wedge}|S)=MAX\ P(WS|S)$，

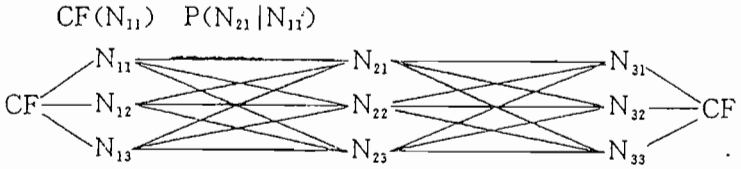
由 BAYES 公式和一阶 MARKOV 模型特性，最后推得公式：

$$WS^{\wedge} = (\prod P(W_i|W_{i-1}) \times CF(W_i) \text{ 最大的 } WS) \tag{7}$$

公式中 W0 和 Wn+1 为确定边界 CB，CF()表示 SCR 部分提供的候选字信度，CF(CB)=1。$P(W_i|W_{i-1})$为所统计同现概率。

4.2 最佳路径选择

对于我们的模型，每一个汉字对应多个候选。这些候选的每一种组合都是一种输出可能，称为一条路径。例如一个 3 字长，每字 3 候选的串的所有路径如下图所示：



因而求最大概率汉字序列即为选择最佳路径。显然，这为一个指数空间求解问题，我们用动态规划法来计算。设若 N 为汉字串长度，M 为平均候选个数，则算法复杂度为 O(M×N)。

这样，按概率赋值计算所得最佳路径即为最终识别结果。

5. 结束语

采用上述方法我们做了手写印刷文本的实验。对手写体我们采用封闭文本，并做了必要的人工干预(反复学习、再识别直到10选率为100%)，这种情况识别从原来的66%，提高到86%。而对印刷体的测试，我们采用了开放文本，未做任何人工干预。这种情况下原首选正确率91。2%(10选率为98。9%)，经处理后达到96。7%的识别率，提高了5个百分点。可见，基于字的二元统计的确反映了汉语语言的一些内在相关性，而这种知识对于类似于文本识别的处理是很有效的。

我们认为：进一步改进汉语文本统计方法和把这种统计结果于多种其它知识结合应用到手写、印刷文本识别、语音识别等语言处理中是非常有意义的研究方向。

参考文献：

【1】Marie Meteer POST：Using probabilities in languange processing IJCAI'91

【2】F. Jelinek,R. L. Meteer,"INTERPOLATED ESTIMATION OF MARKOV SOURCE PARAMETERS FROM SPASE DATA", Pattern Recognition in Practice 1980

【3】Yukiyasu IIDA Knowledge procesing for an OCR REVIEW of theElectrical Communication Laboratories Vol.32,no.5 1984

【4】Steven J. Derose Grammatical category Disambiguation Linguistics by statistical optimization Computational Linguistics Vol.14,no.1 1988

【5】Hung－yan Gu,Chiu－yu Tseng and Lin－shan Lee Markov modelingof Mandarin Chinese for decoding the phonetic sequence into chinesecharacters Computer speech & language (1991) 5 p563

【6】Der－sheng Shy,Larry Wang etc A Statisical Method for Locating TYPO in Chinese Sentence Procedings 1992 International Conference on Chinese Information Processing 1992 Beijing China

【7】SLAVA M. KATZ Estimation of probabilities from Spars Data for the language Model Component of a Speech Recognizer IEEE Trans. VOL ASSP－35 NO. 3 MARCH 1987

【8】白栓虎 "基于统计的汉语语料库词性自动标注方法的研究和实现" 清华大学硕士论文 1992

【9】常新功 夏莹 "利用上下文知识的汉字文本识别系统" 全国智能接口与智能应用专题学术会议'93 1993 黑龙江 镜泊湖

【10】曲洪亚 顾小凤 "手写印刷体汉字识别的一种后处理方法" 第四届全国汉字及汉语语音识别论文集 1992

【11】韩敬体 等《新编实用汉语词典》 社科文献出版社 1989

Text Recognition Using Statistics—based Linguistics Knowledge

Chang Xingong　　Xia Ying
Department of Computer Science, Tsinghua University
Beijing　100084

Abstract

Recently, applying contextual information more than the lexical-level knowledge to text recognition system to improve system performance has intrigued lots of interests. Our recent research in this area was given in this paper.

First, from a chinese text corpus base of size of 5 million characters, we acquired a chinese characters co—occurrence probability matrix. Then, the first—order Markov language mode and dynamic programming strategy were employed to select the optimum characters from several recognition canditates. Here, a bounded sequence of chinese characters (more often, a sentence) is processed as an unit.

Our experimental results are satisficatory and we think this method is promising.

KEY WORDS: bi—gram co—occurrence probability　data sparsness　Back—off method　Markov mode