

统计语言模型及 汉语音字转换的一些新结果

郭进^[注1]

(清华大学电子工程系)

【摘要】汉语音字转换是一个重要而困难的问题。语料库语言学为我们提供了新思路。作者们通过建立统计语言模型,将基于语料库的方法与传统的基于规则的方法结合,研制了THED新一代音字转换系统。该系统对随机抽取的新华社新闻语料有不低于95%的带调音节和国标汉字的转换正确率。本文侧重报道该系统在汉语音字转换方面及与此相关的汉语切词和词性标注方面的一些实验结果,也简要介绍该系统在语料库应用方面的一些思路。

一、汉语音字转换

汉语音字转换是指由计算机自动地将汉语音节串与方块汉字串正确地互换。对这个问题的重要性近来已有了进一步的认识:正确的音字转换是汉语语音识别与合成的关键环节,也是汉字键盘输入最理想的方式。不仅如此,音字转换所能达到的正确率在工程上决定了汉语拉丁化或拼音化的前景。

汉语音字转换的困难只有一个:音字不一一对应。汉语是一字多音的。例如,《常用多音多义字》一书收“常用”多音字共600多个^[1]。《汉语多音字辩略》则收“实用”多音字1219个^[2]。据作者统计,全部6724个国标方块汉字(不计16区后的39个部首偏旁)^[3]在《新华字典》^[4]中共有7536个读音,平均每个国标汉字有1.12个读音。相比之下,汉语一音多字现象就严重多了。上述6724个汉字在《新华字典》中最多只会1264种不同读音,平均每种读音至少对应5.78个国标汉字。如果认定全部70,000个汉字都可用普通话读出,则平均每种普通话读音要对应近60个汉字!人们从不同角度用不同术语如不确定性、歧义性、混淆性和多重性等来表达这一困难。

解决音字转换问题的办法似乎也只有一条:利用上下文约束。孤立地发一个音节,一般很难猜中它“正确地”对应哪个方块汉字。但若给出对应一个词或词组的几个音节,则正确猜中的可能性就大多了。如果给出一个音节所处的短语、句子或段落篇章,则基本上可以肯定其对应的汉字。一般而言,所给的上下文越多,音字正确转换的可能性就越大。

于是,构造一个计算机自动音字转换系统,首先要决定上下文的范围和利用上下文约束

的策略。现代汉语单字词的使用度约占一半,一般认为仅在词级不可能取得满意的结果。而段落篇章义太大,以目前的技术和条件很难做出一定规模的工作。自然地,目前普遍采用子句级上下文^[5],即在一个句子或短语的范围内利用上下文约束来选择恰当的字或音。

对在汉语音字转换中利用上下文约束的方法的研究,在中国大陆至少已有十年的历史了。比较成型的方法主要有二类。第一类是基于最长词匹配^[6,7,8]。这类方法的优点是易于实现,系统开销小。缺点是对上下文约束利用不充分,很难取得较高的转换正确率。另一类采用自然语言理解的主流技术,即基于规则的句法语义分析^[9,10,11]。这类方法所做的每一步都有理有据,在受限的自然语言理解问题中取得了相当的成功。但实际表明,在当今计算表及计算理论的限制下,对开放的自然语言做到完全规则化,几乎是不可能的^[12]。这类系统一般速度慢,适应面窄。

八十年代中期以来,以信息论为基础的统计建模方法先后在语音识别^[13,14],词性标注^[15,16],拼写校正^[17],机器翻译^[18]等一系列自然语言处理领域中取得成功。这类方法以大型语料库为基础,对非受限的自然语言进行充分细致的调查统计,并在此基础上采用一种统计的语言模型来“理解”自然语言。虽然统计的方法在某种意义上都是靠“猜”,由于大范围的调查收集了大量的不便被语法语义系统利用的细致具体的语言知识,在此基础上的语言模型就有可能更充分地利用上下文提供的约束。实际上,最长词匹配方法和句法语义分析方法都可以看成是统计语言模型方法的特例。因此,只要使用恰当,这种新方法就一定不会比传统的方法差。

作为汉语语音识别系统的一部分,我们选定子句为上下文约束范围,系统地将这种新方法用于汉语音字转换,研制了THED新一代汉语音字转换系统。这个系统对随机抽取的新华社新闻语料有不低于95%的带调音节到国标汉字的转换正确率。对比实验表明,这是作者所知目前世界上最好的汉语音字转换结果。

本文第二节简要介绍统计语言模型方法。第三节说明语料库,第四节介绍系统的几个主要模块,第五节是几个实现考虑,第六节给出实验结果。最后是展望。

二、统计语言模型方法

自然语言处理中有很大一类可以直接或间接地看成找对应问题。例如音到字转换就是给音串找对应的字串,字到音转换则是给字串找对应的音串。甚至翻译问题也是对应问题:给一种语言的一个字串(句子),找在另一种语言中对应的字串(句子)。

一般地,说 $X = x_1, x_2, \dots, x_n$, $Y = y_1, y_2, \dots, y_m$, 找对应问题就是给Y求X。如果 $X - Y$ 不是一一对应关系,则一般不存在唯一“正确”的对应。统计的方法首先就是把找“正确”的对应换为找最可能的对应,即找 \hat{X} ,使在Y下的后验条件概率最大,即

$$\hat{X} = \underset{x}{\operatorname{argmax}} P(X|Y) \quad (1)$$

按Bayes公式,有

$$P(X|Y) = \frac{P(X)P(Y|X)}{P(Y)} \quad (2)$$

所以,

本文1992年7月29日收到。

注1 直接参加本文部分工作的还有:王政贤、马国华、孙玉环(中国电子器件公司),蔡奕、王新涛(北京信息工程学院),许惠、王松、李景林、袁岩松、唐武、孙甲松(清华大学)。

$$\begin{aligned} \hat{X} &= \operatorname{argmax}_x \frac{P(X)P(Y|X)}{P(Y)} \\ &= \operatorname{argmax}_x P(X)P(Y|X) \\ &= \operatorname{argmax}_x P(x_1 x_2 \dots x_n) P(y_1 y_2 \dots y_m | x_1 x_2 \dots x_n) \end{aligned} \quad (3)$$

因此,如果能对所有可能的 X 求出 $P(X|X)$ 和 $P(X)$,则只要比较一下这二次乘积的大小即可找出最可能的对应 \hat{X} 。

于是,统计方法有二个基本问题:建立模型以估计概率和构造算法以求对应。作为一种十分简单但非常有效的语言模型,可假设(一阶马尔可夫假设)

$$P(x_1 x_2 \dots x_n) = P(x_1) P(x_2 | x_1) \dots P(x_n | x_{n-1}) \quad (4)$$

和(独立输出假设)

$$\begin{aligned} &P(y_1 y_2 \dots y_m | x_1 x_2 \dots x_n) \\ &P(y_1 \dots y_{m_1} | x_1 \dots x_{n_1}) \dots P(y_{m_{k+1}} \dots y_{m_k} | x_{n_{k-1}} \dots x_{n_{k-n}}) \end{aligned} \quad (5)$$

在这个模型下,如果我们已经有了一个足够大的语料库,一种可行的方法就是用相对频率来估计概率 $P(x_i | x_{i-1})$ 和 $P(Y|X)$ 。基于动态规划的Viterbi算法则是一种找出对应的有效方法。

对这种方法比较详细的介绍请参见[16,19]。

三、语料库

1. 字典

在THED音字转换系统中,收齐二级国标汉字共6724个,及其在新华字典中的所有发音共7536个(其中不同的音节共1264个)。另外收齐国标中所有字母符号。

2. 词典

词典共收词56729条。包括北航词典[20]中除罕用二字地名外所有词共4万余条和二本中型成语词典[21,22]中所有四字成语。选择北航词典是因为它的统计规模最大,且是在14万余词条中选出的常用词。补入1万左右的成语是因为它们在统计上和语法上都很难处理,不如让其“直进直出”。

3. 文本

文本库收集了新华社90年3月至91年3月间每日发往各地的几乎全部的新闻电讯稿共约二千万字。目前选用了其中约70天4百多万字作为基本训练集以统计频率,另选一部分用于调整某些参数用,其余作为测试用语料。

选择新华社新闻语料主要出于其权威性(几乎所有报刊都采用相当数量的新华社新闻稿),实用性(它是我们日常听、看、说甚至写得最多的材料),可比性(该语料可普遍得到,可为类似工作提供一个可比的评估标准)。除此之外,最重要的是其困难性。新闻语料涉及面广,变化快,又有大量词典不可能收录的人名、地名、机构名、动植物名称等等。而

且新闻报道口语化很强、语言往往很不规范严格。所有这些都使它比政论语料难处理得多。后面的实验结果也说明了这点。

四、系统模块

THED音字转换系统以词为单元,以句为单位,采用在等价类上的二词文法统计语言模型与基于上下文无关文法的句法分析相结合的方法。不仅如此,作为一个实用系统,THED还集成了用户词典,自适应调整,用户界面、自学习等一系列功能模块,也采用了一系列旨在提高响应速度和减少空间占用的特殊的数据结构、算法及简化技巧。[如30]事实上,THED音字转换系统首先是作为THED语音识别与理解系统的一部分存在的。后者甚至还包括了语音合成,文书排版等一系列实用功能。无疑这是一个复杂的系统。本文将只介绍几个基本的统计语言模型模块,后面的实验结果也是仅用这几个模块得到的。

1. 切词

因为THED系统以词为单元,所以第一步就是要将文本库变成以词为单位的语流。我们采用自动切词。切词问题可看成对应问题。这只要把(1)式中的 X 看成词串、 Y 看成字串即可。设有一个已切好词的文本库,则可估计

$$P(x_i | x_{i-1}) = \frac{N(x_{i-1} x_i) + \epsilon}{\sum_x (N(x_{i-1} x) + \epsilon)} \quad (6)$$

其中 $N(x_1 x_2)$ 是词对 $x_1 x_2$ 在整个文本库中出现的次数。 ϵ 是为应付统计数据不足而设的一个调整参数。它将根据由此得到的切词系统对另一个独立的文本(即我们专门留下用来调整参数的文本)的切词性能作适当调整。 ϵ 一般取0到1之间的小数。我们取0.5。显然,如果取 $\epsilon = D$,则(6)式就是用相对频率估计概率。

另外,

$$P(Y|X) = \begin{cases} 1, & \text{如果把词串} X \text{看成字串时与字串} Y \text{一样} \\ 0, & \text{其它} \end{cases} \quad (7)$$

它实际上是要求词串 X 确实是字串 Y 的一种切分。

2. 方块汉字串转音节串

给语料库加音节标记是必须的。这是因为测试系统性能要用到,调整参数要用到,甚至估计概率也要用到。我们也采用自动标记的方法。这是因为语料库太大,无法全部人工标记。而实验结果且表明自动标记的精度在一定成度上能满足要求。

我们也将加音节标记问题看成对应问题。这只要在(1)式中把 X 看成音节串, Y 看成词串即可。设有一以词为单位的带音节标记的语料库,则可用类似(6)式的方法估计 $P(x_i | x_{i-1})$ 。因有一音多词和一词多音,求 $P(Y|X)$ 稍复杂些。设拼音词 x_i 对应方块汉字词 y_i ,则可估计

$$P(y_i | x_i) = \frac{\text{在语料库中对应} x_i \text{的} y_i \text{的出现次数}}{\text{在语料库中} x_i \text{的总出现次数}} \quad (8)$$

而

$$P(Y/X) = \prod_i P(y_i | x_i) \quad (9)$$

我们在(8)式中略去了类似(6)式中的调整参数 ϵ 。后面实验中用到的 $\epsilon = 0.1$ 。

3. 词性标注

“词以类聚”。给词分类是个自然的想法。在此，我们把一个词的分类体系中的一个词的类别称为词类，而把一个具体的词所具有的类别称为词性。显然，词性种类的全体构成一个词的分类体系，而一个具体的词性则决定于词的分类体系。另一方面，词的分类体系则是由分类的目的决定的。

为了便于进行句法分析，需要有一个词的语法分类体系，也需要为每个词标注上语法词性。我们依据《新编汉语多功能词典》[23]将现代汉语词按语法分为13大类，并另加“标点符号”一大类（因为我们把标点符号也看成词）共计14大类。

为了引入语义信息，又需要有一个词的语义分类体系，并为每个词标注上语义词性。我们选用《同义词词标》[24]作为标准。该词典根据语义（也兼顾到部分语法）将汉语词分为12个大类，94个中类，1428个小类。同样，我们另加“标点符号”一大类，再细分为标点、数字、字母、数学符号等数个中类及一些小类。

又为了减少模型参数的数目，特别是减少转移概率矩阵的规模，还需要一个从信息论角度的词的分类体系。这个体系可通过组合上述语法分类体系和语义分类体系，再用一个算法进行系统的分裂合并得到。

选定了词的分类体系，如果再有一个已标记好词性的语料库，则可采用类似方块汉字转音节的方法建立语言模型，并用它来标注更多的语料。

4. 音节串转方块汉字串

这是我们的最终目的。音节串转汉字串也是对应问题。这只要在(1)式中取 X 为方块汉字词串， Y 为音节串即可。类似(6)式可由语料库估计 $P(x_i | x_{i-1})$ ，类似(8)式可估计 $P(Y|X)$ 。

五、实现考虑

1. 启动问题

第一次切词采用北航的切词系统[25]，并屏蔽掉其中所有的“知识”模块。

词性标注采用“滚雪球”的方法，即先人工标记一部分，便让系统建立模型去标记更大的语料，再由人工改正其中的错误，又让机器去做，如此不断扩大规模。

音节标注也是类似的“人助机学，机助人做”的方式：先由机器查词典对无歧义的词（主要是多音节词）进行标注，并根据标注的结果建立音节语言模型以标注其它（有多义）部分。然后人工检查改正部分错误（主要是常用而又使用较平衡的多音字如“行”（xing, hang），等）。据此机器改进音节语言模型，如此往复数次。

2. 零概率问题

训练数据总是不足的。在测试集中总会遇到在训练集中未出现的情况。如何估计该事件的概率会对整体效果产生一定影响。在本文报道的版本中我们一律采用类似在(6)式中引入一“调整参数”的办法。且该调整参数也只是人为地根据系统在“调整语料库”中的性能设定的。采用其它更仔细的方法[如26]的实验正在进行中。

3. 译码算法

基于速度和实现的方便性考虑，我们采用基于动态规划的Viterbi译码算法。

六、实验结果

1. 训练语料的统计特性

总字数	单字词数	双字词数	三字词数	四字词数	五字词数	六字词数
4,190,612	1,507,382	1,212,071	38,407	40,448	3,187	939

单字词与多字词总数基本持平，平均词长约1.5字。多字词也主要是二字词。

此语料仅作计算相对频率用。不包括与此规模相当的用来调整各种参数的“调整语料库”。由于调整语料库也是用来求模型参数的，因此广义的训练语料是指这两部分的总和。以下各组实验结果如不特别指明均由用此训练语料得出的模型求出。

2. 测试文本

以下各组实验结果除特别指明外均由不包括在上述训练语料和调整语料中的集外测试文本Ⅷ求出。文本Ⅷ是一篇有关包钢的长篇报道的一部分，共7586个汉字（不包括标点符号），其中多处出现包钢的人名地名及只与钢铁有关的专门名词（特别是词典未收的简称），另外还有几处赞美诗句等。

3. 切词

我们对文本Ⅷ分别用THED系统和北航CWSS系统[25]进行切分，人工检查结果如下：

系统	错切次数	正确率(%)
THED	4	99.95
CWSS	30	99.60

在给定切词词典后，切分错误只有覆盖和交叉二类。覆盖错误往往与语义有关，交叉错误往往与语法有关。由于词的定义并不明确，人对切分结果的判断往往不一致。例如“一个”是一个词还是由二个词构成的词组就一直有争论。我们这里采用最宽松的标准：只要语法语

义上说得通就算对。按此标准,“一个”切不切开都算对。

按此标准,THED系统的错误率要比CWSS系统低一个数量级。特别有意思的是:由CWSS启动的THED能纠正CWSS产生的错误。这主要是因为CWSS只利用了词典给出的构词信息和其它个别技巧,而THED还系统地利用了由语料库统计得出的词之间搭配及其概率的信息。

还要指出,采用CWSS系统提供的知识校正对新闻语料往往会带来更多的切分错误。主要原因可能是一方面这些知识规则实际上是挂一漏万,另一方面因未引进概率而过“硬”了。

4. 加音节标记

用THED系统对文本Ⅷ自动加音节标记,然后人工检查,结果如下:

系统	错标字数	正确率(%)
THED	5	99.94

与检查切词错误类似的是:人们对什么是“错”意见并不一致。句中的有些多音字,许多人(包括一些受过良好高等教育的大学生和老师)都很自然地读某个音,但若引经据典仔细推敲,会发现另一个读音才是“对”的。例如“血”,在词语“血浆”中应读xuè,但很多人读成xiě,甚至xuě。还有些字音,甚至权威的字词典之间都不一致。我们这里也只好采用最宽松的标准:检查者读起来顺口、不觉得有明显的错就算对。这个标准主要会放过一些单字多音词的四声错,如“为”(wèi, wéi)等。

按此标准,如果可比的话,THED的结果要比[8]报道的至少好一个数量级。理由很简单:THED系统用到的知识多。

5. 词性标注

这里给出一个独立的关于语法标注的实验结果。如前所述,根据[23]选定了14个大词类,分别请不同的大学生根据[23]人工标注了一本大学政治课本的第一章和第二章。每章的词数见下表。共做了三组实验。第一组用第一章语料做训练集,然后自动标记第一、二章。第二组实验用第二章做训练集,然后分别标记第一、二章。第二组实验用第一、二章全体做训练集,然后也分别自动标记第一、二章。这里定义只要自动标注的标记与人工的不一样就算错。实验结果如下:

训练 \ 测试	第一章 (5847词)		第二章 (6924词)		两章合并 (12771词)	
	错标数	正确率(%)	错标数	正确率(%)	错标数	正确率(%)
第一章	262	95.5	214	96.9	476	96.3
第二章	271	95.4	175	97.5	446	96.5
两章合并	228	96.1	224	96.5	472	96.3

以上结果表明,对只有14个词类的标注问题,只要较小规模的训练量即能达到很好的结

果。例如,只人工标记第一章共5847个词,就可对独立的测试集(第二章)达到96.9%的正确率。这还包括了这二章由不同人工标注所固有的不一致性(尽管大家都依据同一本词典,对一个词在句中的词性的认识远不一致)。而把训练量扩大一倍,即把二章合并,即使是对集内(训练语料与测试语料相同)也无明显改善。事实上,两章合并后的集内正确率(96.3%)比用第一章训练所取得的集外正确率(96.9%)还低些。

分析表明,词类数目较少并不一定是上述现象的内在原因。事实上,即使只有14个词类,也有词具有7种词性。这也许比在140个词类下有7种词性难办得多。

无独有偶,[27]对英语词类标注也报道了类似的现象。

在理论上,我们曾对所采用的统计语言模型做过扰动分析,发现在一定条件下它是相当稳定的(robust)。也许上述现象可以由此解释。但这是一个专门的问题。我们将另文给出。

6. 音转字

在这个实验中,先用汉字转音节模块将文本转成带四声音节文本,再对该文本做音节串到汉字串转换。然后比较前后二个汉字文本。不一致即算错。结果如下:

文本	总字数	错误字数	正确率(%)
Ⅷ(集外)	7586	357	95.29
G0304.1(集内)	6196	91	98.53

以上结果是在训练集有4,190,612个汉字(不含标点符号)时得出的。对同样的测试文本,当训练集只有1,095,893个汉字时,转换结果是:

文本	总字数	错误字数	正确率(%)
Ⅷ(集外)	7586	658	91.33
G0304.1(集内)	6196	59	99.65

这说明扩大训练量有助于提高系统的稳定性。但即使同样采用一百万字左右的训练集,THED的转换正确率也高于[28]报道的结果(89%)。这说明以词为单位要优于以字为单位。另外,我们还将Ⅷ文本送入音声(lnSun)音字转换系统[5,29]。该系统只取得约76%的转换正确率。

为迎接今年863检查,我们也曾对李鹏总理在七届五次人大上的报告共一万四千多字进行了音字转换。结果正确率竟达98%以上。这既说明了基于统计语言模型的方法的有效,也说明了政论语料要比新闻语料容易处理。

七、结束语

本文简要介绍了THED汉语音字转换系统的一些作法和实验结果。目前该系统正在完善中。我们计划使训练集达到二千万词左右,并平行地针对科技、文艺、政治、综合等再建几套,使系统在相关领域能有更好的表现。另外,词的自动聚类,语法的形式化和基于规则的

系统与基于统计的系统的结合方式, 训练数据不是问题, 自学习自适应功能以及整体训练模式等都是目前正在进行的工作。初步的实验结果是鼓舞人心的, 表明所用方法是有前途的。

致谢: 感谢王作英教授, 本工作作为汉语语音识别与理解系统的一部分, 在他的领导下进行, 并得到国家 863 计划和中国电子器件公司的部分支持。作者感谢陆大铨教授、朱雪龙教授的关心和指导, 黄昌宁教授的鼓励, 以及全体同仁的艰苦劳动。作者也感谢匿名审稿人的衷心细致的建议。

参 考 文 献

- [1] 朱盛科, 常用多音多义字, 四川人民出版社, 1979
- [2] 林然, 汉语多音字辨略, 上海外语教育出版社, 1986
- [3] 汉字信息字典, 科学出版社, 1988
- [4] 新华字典, 商务印书馆, 1991
- [5] 王晓龙, 音字流切分及相互转换的理论研究与系统实现, 哈尔滨工业大学博士论文, 1989
- [6] 李慧勤, 普及型拼音—汉字变换系统设计, Proc. CCPCOL'90, pp. 383—387, 1990
- [7] 仲兴国, 多词组一次性拼音, 汉字变换, 中文信息学报, Vol. 4, No. 2, pp. 55—64, 1990
- [8] 苟大举等, 汉语语音合成中多音字的处理, 中文信息, 91(1), 33—36, 1991
- [9] 杨长生, 何声钧, 汉语同音词汇的辨析, 计算机研究与发展, Vol. 24, No. 1, pp. 46—50, 1987
- [10] 俞士汶, 中文输入中语法分析技术的应用, 中文信息学报, Vol. 2, No. 3, pp. 20—25, 1988
- [11] 唐武、杨行峻、郭进, 用于语音识别的拼音汉字转换系统SW-1, 中文信息, 91(2), 25—27, 1991
- [12] 黄昌宁, 语料库语言学, 中国计算机用户, 1990.11
- [13] Bahl, L. R., Jelinek, F. and Mercer, R. L., A Maximum Likelihood Approach to Continuous Speech Recognition, IEEE Trans. on PAMI, PAMI-5(2), 179—190, 1983
- [14] Jelinek, F., The Development of an Experimental Discrete Dictation Recognizer, Proc. IEEE. Vol. 73, no. 11, pp. 1618—1624, 1985
- [15] DeRose, J. S., Grammatical Category Disambiguation by Statistical Optimization, Computational Linguistics, Vol. 14, No.1, 1988
- [16] Garsjide, R. G., Leech, G. N. and Sampson, G. R. The Computational Analysis of English: a Corpus-Based Approach, Longman, 1987
- [17] Mays, E., Dameran, F. J. and Mercer, R. L., Context Based Spelling Correction, Proc. of IBM Natural Language ITL, 1990
- [18] Brown, P. F. et. al, A Statistical Approach to Machine Translation, Computational Linguistics, Vol. 16, pp. 79—85, 1990
- [19] Jelinek, F., et. al., Principles of Lexical Language Modeling for Speech Recognition, in Advances in Speech Signal Processing, Furui, S and Sondhi, m. M(eds.)1992
- [20] 刘原、梁南元等, 现代汉语常用词同义词典(音序部分), 宇航出版社, 1989
- [21] 汉语成语考释词典
- [22] 现代汉语成语词典
- [23] 冯志纯、周行健主编, 新编汉语多功能词典, 国际文化出版公司, 1989
- [24] 梅家驹、竺一鸣、高蕴琦, 殷鸿翔, 同义词词林, 上海辞书出版社, 1983
- [25] 梁南元, 书面汉语自动分词系统—CDWS, 中文信息学报, 1988.2
- [26] Katz, S. M., Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer, IEEE Trans on ASSP, Vol. ASSP-35, pp. 400—401, 1987
- [27] Merialdo, B. Tagging text with a probabilistic model. Proc ICASSP, pp.809—812, 1991

- [28] Lee, L. S. et. al, System Description of the Golden Mandarin(1) Voice Input System for Unlimited Chinese Characters, Proc. ICCPOL'91, pp. 45—50, 1991
- [29] InSun 3.01系统说明书, 哈尔滨工业大学, 1991
- [30] 唐武、杨行峻、郭进, 汉语音字转换中同音字(词)的概率后处理, 中文信息学报, Vol. 6, No. 2, pp. 52—56, 1992

Statistical Language Modeling and Some Experimental Results on Chinese Syllables-to-Words Transcription

Guo Jin*

Dept. of Electronics Engineering, Tsinghua University
Beijing, 100084, People's Republic of China

Abstract

The problem of transcribing a sentence of Chinese syllables into a sentence of Chinese graphic words is an important and yet difficult one. A new technology is highlighted by the recently booming Corpus Linguistics. By combining corpus-based approaches with traditional rule-based ones, the author has developed a new generation Chinese syllables-to-words transcription system which is capable of transcribing tonal syllable sentence into standard simplified Chinese graphic word sentence with a character accuracy of no less than 95% for randomly sampled news from the Xinhua News Agency of China. The goal of this paper is to report some experimental results the system achieved on Chinese syllables-to-words transcription as well as on related Chinese sentence word boundary detection, words-to-syllables transcription and word category tagging. Some general considerations on ways of constructing and utilizing language corpora for the developing of this system are also briefly discussed. Key words: Chinese syllables-to-words transcription, Statistical Language Modeling, Chinese sentence word boundary detection, Word Category Tagging, Chinese words-to-syllables transcription.

*Current Address, Institute of Systems Science, National University of Singapore, Kent Ridge, Singapore 0511, Tel: (65) 7726381, Email: guojineiss, nus. sg