汉语同音词调查及拼音输入法基线模型研究

丁大斌　黄昌宁

微软亚洲研究院，北京 100190

Email: v-dadi,v-cnh@microsoft.com

摘要： 本文在一个大规模分词语料库的基础上，对现代汉语的同音词现象进行了调查，分析了汉语同音词的特点。调查结果有助于深入了解现代汉语的同音词问题，进而为以词为输入目标的汉语拼音输入法提供了一种基于“高频先见”的基线模型。文中给出了这种输入法模型的 TOP1 和 TOP5 正确率测试结果，并指出未登录词（OOV）是影响该模型正确率的主要因素。

关键词： 同音词 拼音输入法 基线模型 TOP1 正确率 TOP5 正确率

An Investigation on Chinese Homophone & Study on Pinyin IME Baseline Model

Dabin Ding, Changning Huang

Microsoft Research Asia, Beijing 100190

Email: v-dadi,v-cnh@microsoft.com

Abstract: Based on the statistics of one large segmented corpus, this article makes an investigation into Chinese homophone problem, and analyzes the characteristic of homophones in Chinese. This investigation promotes the understanding of Chinese homophone problem. Furthermore, the article proposes a baseline model of Chinese pinyin IME that takes words rather than sentence as input unit and base on the strategy of “High Frequency First Appear”. In the last, this article gives the testing result of the TOP1 and TOP5 accuracies of the model, and point out that the OOV is the main factor which affect the input accuracy.

Keywords: Homophone, pinyin IME, baseline model, TOP1 Accuracy, TOP5 Accuracy

1 引言

每一种自然语言都有同音词。从根本上说，同音词反映了语音的有限性和词语的无限性之间的矛盾。在中文信息处理领域，同音词的大量存在影响了语音识别和以拼音作为汉字输入方法的效率。将中文录入计算机是编辑和打印中文文本、进行网上交流的第一步，是中文信息处理的关键问题。同音词的辨识成为提高拼音输入法效率的关键问题。

自 20 世纪 80 年代以来，对于汉语词汇的专题研究日渐增多，对于同音词问题，也有一些理论上的讨论和基于词典的静态统计结果。尹文刚对《新华词典》[1]收录的汉语同音字进行了统计，提出了“同音率”和“同音度”两个概念作为度量同音字特性的量化指标，得出了汉字语音符合“清晰原则”的结论[2]。然而多数文献都是从语言学角度对汉语同音词问题所作的分析，从信息处理的视角对汉语同音词进行分析的文章并不多见。冯志伟、张普等所著的中文信息处理的相关书籍中曾提到汉语的词频统计、音节总数等数据，然而鲜有对语料中同音词规模的统计[4][5]。本文将从信息处理的角度出发，分析汉语词典和大规模语料库中同音词的分布状况，以期服务于同音词的辨识，提高拼音输入法的效率。

本文讨论的拼音输入法，是以词而不是句作为输入单位，目的是把这种输入法作为将来实现汉语文本词式书写的重要工具。如果一种输入法在人机交互过程中把用户已经确认的每个词的边界记录下来，就不会像当前的字式书写那样在输入后完全“遗忘”掉输入的词边界，造成汉语信息处理中的一种极大的资源浪费。

2 术语定义

为了进行汉语同音词的调查，定义和使用了以下术语：

（1）拼音、音节与音节形式：单字的读音称为音节（syllable），音节是语音的基本单位，无调音节称为音节形式。单音节读音和多音节组成的复合读音统称为拼音（pinyin）。因此，在汉语中每个词（字）都有三个属性可以进入统计：词（字）形、词（字）次和拼音。本文的统计分析未涉及词义或字义。

（2）同音词：具有相同读音的一组词形被称为同音词（homophone）。一个同音词可能有不只一个义项，这些义项之间也可能没有直接关系，本文只把它视作一个词形[1]。对于词典中收录的汉字，本文不区分它究竟是语素字、语素还是词，统称为字或单字。所以文中提及的单音节同音词实指同音字。按照读音是否带调，同音词（字）又分为无调同音词（字）和带调同音词（字）。

（3）同音度：一个拼音所对应的同音词形的个数称为同音度（Homophone degree）[2]。

3 语料中汉语同音词的特点

3.1 调查所用的资源

本文调查和分析的主要对象是国家现代汉语平衡语料库[9]，以下简称 NCC。NCC 是由国家语言文字应用委员会主持建立的一个现代汉语平衡语料库，语料由人文与社会科学、自然科学及综合等三个大类约 40 个小类组成，语料抽样能够比较科学地反映现代汉语的全貌。本文使用的语料库共包含 203,395 个词形，150,841 个拼音，17,956,409 词次（不包括标点、数字串和外文字符串），所有语句均完成了词语切分和词性标注，如此大规模的切分语料，在以往的统计分析中是不曾见到的。需要指出的是，NCC 中的词形原本不带拼音。为了能够统计词形及其拼音，本文首先用程序给语料库中的词形标注无调拼音[2]。为了进行对比，本文还分析了《现代汉语词典》（第四版）[8]中同音词的分布。

3.2 基于语料库的同音词分析

从表 1 中可以看出，无调同音字和无调双音节同音词比例较高，三音节及以上的词形中同音词比例显著减少。NCC 词表中，无调同音词所占比例为 37.3%，比《现汉》中比例降低了近 12 个百分点。这是因为语料中包含众多的未登录词，这些未登录词多为无调同音词比例不高的多音节词。然而，同音词在 NCC 语料的词次百分比（出现次数的比例）却很高，达到了 81.2%。比例大幅增加的原因在于，语料中单音节词和双音节词的词次比例很高（93.9%），而根据对词表的统计，它们大多为无调同音词。

	《现汉》(%)	NCC 词表(%)
单字	99.9	99.6
双音节	50.3	69.8
三音节及以上	1.86	4.98
总计	49.0	37.3

表 1 《现汉》和 NCC 中无调同音词对比

[1] 在语言信息处理的初级阶段中，对于词的同一性不加区别。所以算术里的“分数”和考试成绩的“分数”，虽然在词典中属于不同的两个义项，但由于同音同形，本文把它们视为一个词形。

[2] 多音字的读音会有少量的标注错误。

在两种资料的静态统计结果中，无调同音字比例都接近 100%，其主要原因是汉语中单字多而音节形式少，《现汉》中平均每个音节形式的载字量约为 20 个。如果《现汉》中收录的 8465 个单字都在文本中被频繁使用的话，就会给拼音输入法带来极大的困扰。然而周有光指出，汉字中“有两千多个(1/3)是代表‘语词’的‘词字’，它们能独立成词。利用‘以词定字’方法，三分之二的汉字可以避免同音干扰。”[12]。根据对语料中高频词的分析，2286（约为总数的 1/3）个最高频的单音节词累计词次比率为 99%，剩余 4123 个单音节词的累计词次比率仅为 1%。

	数量	词形(%)	词次(%)	无调同音词 (%)	带调同音词(%)
单音节词	2286	35.7	99	99.5	96.3
双音节词	47831	52.0	99	69.6	-

表 2 高频单音节词和双音节词分析

与单音节词类似，双音节词也呈现出少数高频词频繁出现，大量低频词偶尔出现的规律，只是高频词的比例增大。三音节及以上词形在语料中所占比例小于 8%，这里不做深入分析。以上统计表明：汉语词汇中词形很多（NCC 语料中切分出的词语超过 20 万条），但是频繁出现的词形却不多。根据少数高频词被频繁使用的特点，如果能够采取有效的策略，比如“高频先见、用过优先”等，优先解决高频同音词的消岐，就可以大幅度提高拼音输入法的一选正确率。

4 基于“高频先见”的拼音输入法模型

在本节中，为了评价以词为输入目标的拼音输入法，根据“高频先见”策略建立了一个朴素的拼音输入法模型，进而通过一系列的实验，分析了基于该模型的拼音输入法的基线正确率，以及影响正确率的因素。

4.1 测试模型的建立

1) 模型的定义

在使用拼音输入法输入时，将高频的候选词优先显示，就是“高频先见”的思想。本文根据这一思想，创建了简单的拼音输入法模型。介绍模型之前，需要定义以下参数：T 表示测试集，它由词形串组成，即 $T=(w_1,w_2,\ldots,w_i,\ldots,w_m)$。T 中的词形 w_i 可以重复出现多次。P 表示测试集中词形所对应的拼音的集合，$P=\{p_1,p_2,\ldots,p_i,\ldots,p_n\}$。任取 w_i 属于 T，都对应一个或多个 p_i 属于 P。任意 p_i 都有属性 Hd_i，表示该拼音的同音度。拼音输入法模型以拼音作为输入，输出词形，所以拼音输入模型 F 实际上是一个转换函数：

$$F(p_i)=\{c_{i1},c_{i2},\ldots,c_{ij},\ldots\}\text{且 } 1\le j\le Hd_i$$

c_{ij} 代表输入拼音 p_i 后，候选窗口中显示的第 j 个候选词形。目标词形 c_{ij} 的序号 j 越小，转换效果就越好。

2) 评价指标

为了评价以词为输入目标的拼音输入法的效率，本文根据目标词形在候选词中的排列位置提出了以下两个评测指标：

i. TOP1 正确率（一选正确率）

TOP1 正确率是指用户的输入过程中，目标词形落在候选词表一选位置上的次数占总的输入次数的比率。TOP1 正确率越高，用户的输入开销就越小，输入体验就越流畅。定义函数 $f_{TOP1}(w_i)=1$，如果 c_{ij} 是 w_i 对应的目标词，并且 j=1，否则 $f_{TOP1}(w_i)=0$。TOP1 正确率定义为：

$$R_{TOP1}=\frac{\sum_{i=0}^{m} f_{TOP1}(w_i)}{m}$$

ii. TOP5 正确率（五选正确率）

与 TOP1 正确率的定义类似，TOP5 正确率是指用户的输入过程中，目标词形落在前五选之内的次数占总的输入次数的比率。定义 TOP5 正确率是因为候选窗口大小为 5 是当前几乎各种输入法的缺省设置，且五选之内的候选词，用户无需翻页直接按数字键就可以选中目标词形，这样的输入开销应在用户可接受的范围内。同样，定义函数 $f_{TOP5}(w_i)=1$，如果 c_{ij} 是 w_i 对应的候选词，并且 j<=5，否则 $f_{TOP5}(w_i)=0$。那么 TOP5 正确率的公式与 TOP1 类似，此处略。在测试集中出现，而在训练集中未出现的词形称为未登录词（OOV）。输入未登录词时，模型给出的候选词形中不包含目标词形或候选词形为 0 个[3]。设测试集 T 中含 k 个未登录词，则未登录词比率定义为 k 与测试集大小 m 的比值。

这里模型得到的 TOP1 和 TOP5 正确率，本文称为基线正确率（Baseline Accuracy）。因为它仅仅利用了训练集上得到的词形的频率信息，没有使用训练集中的上下文信息以及其他外部知识，所以得到的正确率是基于“高频先见”的拼音输入法模型所应该达到的最低指标，可以预见当前主流的拼音输入法在上述两个指标上的正确率都要高于基线正确率。

4.2 实验设计

本文从 NCC 语料中，通过随机抽样来生成 7 个训练集和 1 个测试集，测试集与训练集不交叠。对语料的分割或选择均采用以下两个原则：（1）以句为单位，（2）全文随机抽样。下文中提到的训练集和测试集大小都指二者的相对大小。测试集中包含 179,872 个词次，26,762 个词形。此时得到的测试集是连续文本，为了模拟用户在不同情景下的输入体验，本文从测试集中提取出两个测试序列：词表 T_L 和文本 T_T。以词表 T_L 作为输入目标时，输入的是取自测试集的一个个孤立的词形，其中不含重复的词形（如用户在输入检索词时），得出的 TOP1 和 TOP5 正确率称为词表正确率（Lexicon Accuracy）。以文本 T_T 作为输入目标时，输入的是测试集的句子，词和词之间有上下文关系（如用户在录入文章时），得出的 TOP1 和 TOP5 正确率称为文本正确率（Text Accuracy）。注意这两种输入正确率都是根据同一个测试集得到的，只不过反映了不同输入情景下用户的不同体验。总的来说，从 NCC 平衡语料中抽样得出的测试集规模大，可以较全面地模拟用户的输入目标，保证实验结果的可靠性。

4.3 测试结果

1) 训练集规模变化对 OOV 比率的影响

图 1 中显示了训练集规模逐渐增大时，词表 OOV 比率与文本 OOV 比率的变化曲线（图中使用了两套 Y 轴比例尺）。总的来说随着训练集的增大，词表 OOV 比率和文本 OOV 比率都迅速减小。当训练集较小时，两种 OOV 比率变化较剧烈，随着训练集增大到一定程度后（大于 30 后），两种 OOV 比率的递减就不明显了。

1) 训练集规模逐渐增大时，TOP1 和 TOP5 正确率的变化

图 2 中显示了训练集规模增大时，TOP1 和 TOP5 正确率的变化。随着训练集规模的增大，词表 TOP1 和 TOP5 正确率的增涨幅度较大。对比图 1 中词表 OOV 比率的大幅度递减可以推断出，词表 OOV 比率的递减是造成词表 TOP1 和 TOP5 正确率增加的主要原因。而文本 TOP1 和 TOP5 正确率的变化幅度不像词表输入那么大。相应的，文本 OOV 比率的递减幅度比词表 OOV 比率的递减幅度小得多。总的来说，训练集规模增大时，词表 OOV 和文本 OOV 比率递减、词表和文本 TOP1、TOP5 正确率随之提升。

2) 影响 TOP1 和 TOP5 正确率的两个因素

[3] 即未登录词的召回率为 0。

根据模型的定义，有两个因素影响模型正确率：1）测试集中的 OOV 比率，2）根据训练集得到的候选词的排序。因为对于 OOV，默认为输入失败；候选词的排列对正确率的影响也显而易见，如果测试集中的高频词形所对应的候选词被排在五选之外，那么正确率就会降低。通过对图 1 和图 2 中曲线的观察，可以预测训练集规模变大时，OOV 比率的单调递减是 TOP1 和 TOP5 正确率按指数规律上升的主要原因，而候选词的排序对正确率的影响不大。原因在于，当训练语料和测试语料足够大时，从训练集上得出的词形的频率顺序是相对稳定的，应与测试集中相同同音词组中词形的频率顺序基本相同，这符合概率中的“大数定律”所揭示的必然性。下面将通过对实验结果的分析，来说明这个规律。

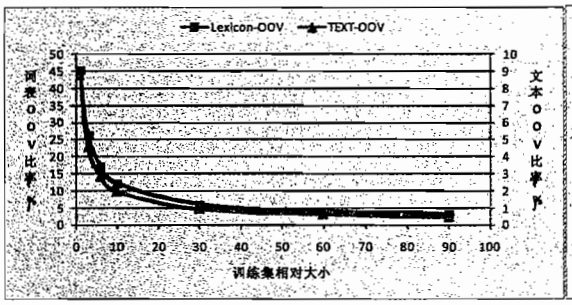

图 1 训练集规模变化对测试集中 OOV 比率的影响

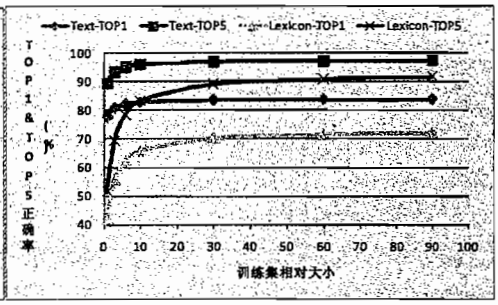

图 2 训练集规模增大时 TOP1 与 TOP5 正确率的变化

由于训练集语料的有限性，测试集 T 中包含两种词形：登录词（IV）和未登录词（OOV）。将未登录词 OOV 剥离出测试集后，观察输入登录词（IV）的 TOP1 和 TOP5 正确率，就可以评价候选词排序对输入正确率的影响。这里将输入登录词时的 TOP1 正确率称为登录词 TOP1 正确率（IV-TOP1 Accuracy），定义如下：

$$R_{IV\text{-}TOP1} = \frac{\sum_{i=0}^{m} f_{TOP1}(w_i)}{m-k}$$

其中 m 为测试集 T 中的词形总次数，k 为未登录词出现的总次数。类似的可以定义登录词 TOP5 正确率（IV-TOP5 Accuracy）。由于未登录词是相对于训练集而言的，当训练集的规模变化时，k 也是变化的。为了保证测试对象是固定的，本文将 7 个训练集上得出的 OOV 集合取交集，将它们从测试对象 T_L 和 T_T 中剥离出去，那么剩下的词形对于 7 个不同的测试集来说都是登录词。这样得到的登录词词表包含 13,318 个词形，用 $T_{L/V}$ 表示，登录词文本包含 162,397 个词次，用 $T_{T/V}$ 表示。将 $T_{L/V}$ 和 $T_{T/V}$ 作为测试序列，得到的词表和文本登录词 TOP1 和 TOP5 正确率如表 3 所示。

大小	1	3	6	10	30	60	90
Lexicon-IV-TOP1(%)	76.1	75.5	75.8	75.9	76.3	76.4	76.4
Lexicon-IV-TOP5(%)	94.1	94.0	94.0	94.1	94.1	94.1	94.1
Text-IV-TOP1(%)	85.2	85.2	85.3	85.4	85.4	85.4	85.4
Text-IV-TOP5(%)	98.0	98.1	98.1	98.2	98.2	98.2	98.2

表 3 词表和文本登录词 TOP1 和 TOP5 正确率

表 3 中列出了不同训练集上得出的登录词词表和文本的 TOP1 和 TOP5 正确率（第 2 行到第 5 行）。总体上来说，词表和文本 IV-TOP1 和 IV-TOP5 正确率的变化在 1‰到 1%之间。在训练集小于 10 时，正确率指标略有变化，其后随着训练集的增大，则基本不变。同时候选词排序是否合理，对词表 IV-TOP1 和 IV-TOP5 是基本没有影响的，因为以词表作为测试输入时，每个词形仅输入一次，只要前 5 个候选词在词表中出现过，那么正确率就是不变的。文本 IV-TOP1 和 IV-TOP5 正确率是由候选词的排序决定的，当训练集规模增大时，二者基本

不变的事实说明：根据不同规模训练集得出的、同一同音词组中词形的频率顺序是相对稳定的，对训练集统计结果的观察已经验证了这一点。

通过上面的分析我们知道，要提高拼音输入法模型的正确率，必须减小 OOV 比率。较大的训练集有助于减少 OOV 比率，进而提高 TOP1 和 TOP5 正确率。然而一味的增大训练集，并不是始终有效的。如在训练集规模为 90：1 时，其规模已经超过了 1600 万词次，包含 19.5 万个词形，此时测试集词表中仍有 3.1%的词形为 OOV，经笔者观察多数为人名，其次为机构名和地名。因此可以预见，不管如何扩大训练集的规模，只要测试集与训练集不交叉，那么测试集中仍然会有一定比例的 OOV。如果期望词表和文本 TOP5 正确率都达到 99%，那么必要条件就是词表 OOV 比率小于 1%。为了达到这个目标，一个可能的方法是扩大训练语料的同时，建立人名、地名和机构名的词库或知识库。

5 结论

本文在词典和语料库两种资源的基础上，对现代汉语的同音词现象进行了调查。根据 NCC 大规模平衡语料库的统计，揭示了汉语同音词在真实文本中的分布状况，对前人基于词典的同音词统计做出了有意义的补充。这一调查有助于深入了解现代汉语的同音词问题，服务于同音词的辨识，从而提高拼音输入法的效率。

拼音输入法对中文计算机用户的重要性不言而喻，然而却少有对输入法模型的性能进行深入探讨的文章。本文为以词为输入目标的汉语拼音输入法，提供了一种基于“高频先见”策略的基线模型，并对输入法的评测方法进行了有益的探讨。在根据 NCC 语料构建的训练集和测试集上，本文通过评测揭示了 TOP1、TOP5 正确率和 OOV 比率的变化规律，指出影响模型正确率的诸因素中，OOV 是影响拼音输入法正确率的主要因素。尽管本文只给出了拼音输入法的一种基线模型，但把“高频先见”这一特征单独加以研究，把模型的 TOP1、TOP5 正确率与汉语词汇中同音度（Hd）的分布状况直接联系起来观察，却是同类文献中少见的。本文定义了 TOP1 和 TOP5 两个指标来评价输入法的性能，评测对象分为词表和文本，这样有利于搞清楚不同特征对拼音输入法性能带来的影响，从而使输入法的研究建立在更加扎实的理论基础上。本文工作的下一个目标是在基线模型的基础上，利用上下文信息和用户词表等手段来进一步提高拼音输入法模型的正确率。

参考文献

[1] 中国社会科学院语言研究所词典编辑室. 新华字典[M]. 北京: 商务印书馆, 2003.

[2] 尹文刚. 汉字同音率、同音度及同音字音节个数随同音度增加而递减的规律[J]. 语言科学, 2003, 2(4):3-6.

[3] 代建桃.现代汉语同音词研究[D].中国知网: http://www.cnki.net, 2008.

[4] 冯志伟. 中文信息处理与汉语研究[M]. 北京: 商务印书馆, 1992.

[5] 张普. 汉语信息处理研究[M]. 北京: 北京语言学院出版社, 1992.

[6] 卢偓. 现代汉语音节的数量与构成分布[J]. 语言教学与研究, 2001, (6).

[7] 刘延新, 许皓光. 汉语双音节同音词词典[M]. 沈阳: 辽海出版社, 1999.

[8] 中国社会科学院语言研究所词典编辑室. 现汉汉语词典[M]. 北京: 商务印书馆, 2002.

[9] 教育部语言文字应用研究所计算语言学研究室. 语料库检索系统[R]. 检索来源: 中国语言文字网: http://www.china-language.gov.cn/, 2005.

[10] 黄昌宁, 赵海. 中文分词十年回顾[J]. 中文信息学报, 2007, Vol. 21(3): 8-20.

[11] 尹斌庸. 汉字输入的发展方向是什么[N]. 科技日报, 1996.

[12] 周有光. 中文输入法的两大规律-汉语规律和汉字规律[N]. 计算机世界报, 1994 .