文章编号：1003-0077（2011）00-0000-00

蒙古文原始语料统计建模研究*

白双成 [1,2]

（1.内蒙古社会科学院 蒙古语信息技术研发中心，内蒙古 呼和浩特 010020；

（2.内蒙古蒙科立软件股份有限责任公司，内蒙古 呼和浩特 010011）

摘要：本文针对蒙古文纠错语料稀缺、扩建难度大，原始语料存在严重的拼写多样化和字形拼写错误而无法直接利用的现状，在分析总结蒙古文编码特性基础上，通过搜集整理大规模原始语料和标注部分语料，以蒙古文输入法为技术实现手段和试验平台，重点解决了基于原始语料统计建模和模型优化等研究问题。实验结果证明，该方法可有效提高输入效率，开拓了蒙古文原始本文建模利用的新思路，对所有蒙古文音词转换和形词转换研究都有广泛的参考价值。

关键词：蒙古文原始文本；统计建模；读音错误；字形错误；智能输入

中图分类号：TP391　　**文献标识码：**A

Study of Mongolian Raw Text Modeling

Bai-Shuangcheng[1,2]

(1.Inner Mongolia Academy of Social Science,Hohhot, Inner Mongolia 010020, China ;

(2.Inner Mongolia Menksoft Co.,Ltd , Hohhot, Inner Mongolia 010011 ,China)

Abstract: The corrected Mongolian text is scarce and hard to build, raw text can not be directly used. We based on the analyzing the characteristics of Mongolian encoding, through the collection and collation of large-scale raw text corpus and part of the corpus annotation, as for Mongolian input technology application, mainly solved the problem of statistical modeling and model optimization of original text. Experimental results show that the method can effectively improve the input efficiency. We developed a new idea of using the original Mongolian text modeling, the results in this paper can be directly applied to all Phoneme-to-Word Conversion and Grapheme-to-Word Conversion problems.

Key words: Mongolian Raw Text; Spelling Diversity Phenomena; intelligent Input Method; Spelling Error

1 引言

自然语言处理广泛使用统计语言模型（Statistical Language Model SLM），尤其是自然标注大数据（Naturally Annotated Big Data）、（深度）机器学习（Deep Machine Learning）、知识图谱（Knowledge Graph）等众多方法和理论，促使信息检索（Information Retrieve）、机器翻译（Machine Translation）、校对纠错（Spell Check&Correct）、知识问答（Question Answering）等涉及自然语言应用的各领域研究工作获得了较为显著的进展，基于这些研究成果的各类应用投入使用。这些新技术、新方法的共同点是要以大量数据资源为依托。然而，就是由于“可直接利用”的蒙古文数字资源稀缺，方便获取的未纠错原始文本又是无法直接利用资源，蒙古文信息处理对这些新技术、新方法显得格外的反应迟缓。另一方面，下大力气构建的各类词典（Diction）和知识库（Knowledge Base）得不到充分利用，更是非常遗憾的事情。我们急需在这方面进行深入探索和研究。

1.1. 蒙古文纠错语料现状

Unicode[1]核心规范（Core Specification）13.4 节对蒙古文编码原理(Encoding Principles)这样描述：蒙古文编码模式（Encoding Model）有别于 Unicode 中的任何其他文字，也较为

* **收稿日期：**　　**定稿日期：**

基金项目：国家电子发展基金 2010 年度、2011 年度蒙古文专项；国家自然科学基金（61163020）；内蒙古自治区自然科基金（2011MS0918）

作者简介：白双成（1974—），男，研究员，博士，主要研究方向为蒙古文语言工程。

复杂。因其复杂性，在此仅展示蒙古文映射的基本特性。因派生蒙古文的闪米特字母表（Semitic Alphabet）无法表全蒙古语读音，很多字形被赋予不同读音[1]，其正确读音要依赖上下文来判断。在这方面蒙古文的拼写法（Orthography）与英文类似，但拉丁文中 c 不管读作/k/或/s/，始终被认为是同一个字母（Letter）而赋予相同字符（Character）编码。与此不同，蒙古文的相同字形，可能因有不同读音而被赋予不同编码。这主要源于现代蒙古语语法认为字母的读音才是用于区别的显著特征，而不是它的字形。

正如这段 Unicode 规范所述，我们所熟知的汉、英等多数文字都是“按形编码”，字符编码与字形之间是一对一的简单对应关系，而蒙古文是“按音编码”，字符编码与字形之间是多对多的复杂转换关系，这种特殊编码方式是导致蒙古文文本中存在“拼写形式多样化现象”[2]或称之为“同形异码现象”[3]的根本原因。这种特殊的编码方式也使其拼写错误细分为“字形拼写错误”和“读音拼写错误”两个层次，读音拼写错误指虽然词形正确但字符内码不正确，或者说读音不正确。所以，蒙古文信息处理界一般认为只有“字形拼写正确”且“读音拼写正确”的文本才是可直接利用的数据资源，本文称之为“纠错文本”（Corrected Text）或纠错语料，而未经纠错或未经读音纠错的文本为“原始文本”（Raw Text）或原始语料。

因纠错语料单词拼写正确，我们直接从文本抽取单词构建词库（Lexicon），可以直接进行各类统计分析，可以直接利用现有的各类语言模型解决很多问题。如果语料使用拉丁转写（Latinization[2]）方式，还可暂时忽略蒙古文编码相关问题而直接套用各种模型。目前，我们所了解的蒙古文统计建模的研究基本都是基于拉丁转写的已纠错语料[4][5][6][7][8][9][10]。

目前，蒙古文已纠错语料及其统计研究存在如下几个特点：

（1）现有规模小，新建扩建难度大

只因真正符合读音正确这一苛刻要求的纠错语料建设是个费时、费力、费钱的浩大工程，从目前公开资料来看，只有内蒙古大学蒙古学学院建立的 100 万词级语料[11]（100TUM）是经得起考验的纠错语料库。虽然后来扩充到 500 万词级，近期进一步扩充到了 1000 万词级，但基于这些扩充语料的研究还很少，其可靠性还需时间和实践验证。有研究人员提出通过自动纠错方式构建纠错语料方案，甚至 100TUM 建立初期就已经利用了校对软件。但蒙古文信息处理很多底层基础研究还未解决好，工程化更是薄弱，未能形成完整技术体系现状下，单纯依赖“词典+规则”校对和纠错很难满足纠错语料建设高要求，剩余人工纠正工作仍然不轻松[12]3。

（2）无法满足日趋多样化建模需求

百万词级规模虽然能够满足部分统计建模的科研需求，但是面向实际应用时明显不足，更无法满足需要大数据的统计模型。例如，词向量表示（Word Vector Presentation）是深度机器学习的基础，很显然语料量越大，低维空间（Low Dimensional Space）的词向量越精准。逐步扩容的纠错语料也许能够满足科研和部分实际应用需要，但我们不能单纯等待搜集足够量的纠错语料后才开展相关研究工作。另一方面，仅限于纠错语料的统计思路不利于大数据利用，更无法解决动态流通语料（Dynamic Current Corpus）的实时监控、舆情分析、语情检测等动态语料统计领域需求。新词术语研究，语言动态监测等时效性很强的工作更不可能依赖加工好的语料，必须寻求基于流通语料的利用渠道。由于原始语料量可以很大，即使在筛查特定条件词（如长词）、观察搭配（Collocation）、观察单词长度分布等最基本和最简单的统计必然要比纠错语料表现出更好的统计分布。更何况，直接利用蒙古文原始语料的研究工作是纠错语料建设的必要补充和回旋途径，相铺相成，互为补充。

（3）不易挖掘语言特性

---

[1] 实际上，同形字母使蒙古文具有了“跨方言”特性。

[2] 也称作罗马化（Romanization），就是将蒙古文字母使用拉丁字母转写方式。

[3] 见确精扎布教授此书 694 页“关于蒙古文纠错软件”一文。

如前所述，采用拉丁转写方式的语料便于不受限于蒙古文编码的特殊限制及语言特性带来的麻烦而无障碍利用现有统计模型，甚至早期不支持多字节编码（Multi-Byte Character Set）的统计模型工具都可以直接利用。另一方面，由于拉丁转写方式与标准编码都是“按音编码”，便于转化为标准编码进行进一步研究。但最终，我们必须解决好语言的特性问题并提炼为共性问题，才能利用好通用模型。因为人为纠正掩盖了蕴含在实际应用中真实存在的错误，基于纠错语料的统计方法已失去了接触这些非常有用的“错误”信息的机会了。仅以单词录入不规范而言，我们可以统计观察不同维度的共性，这些共性可能来源于不同录入工具特性，也可能来源于不同区域或方言差异，是纠错中不容错过的重要信息。这一点是本文有别于以往所有蒙古文统计建模的重要特性。

1.2. 原始语料利用面临的困难

既然没有足量可直接利用的纠错语料，又无法满足日趋多样化的统计建模需求，那我们是否可以利用原始语料呢？

虽然不像英文一样有取之不尽的数字资源，也不像汉文一样具有庞大使用人群，但蒙古文也已经积累了足够多的数字资源。除每年各大出版社、报社、杂志社的正规出版物都有电子文档外，近年来蒙古文网站发展迅猛，有日渐增多和繁荣趋势。尤其是“中国蒙古语新闻网”等正规新闻网站内容都是经过多层审核发布，文字相对规范标准，拼写准确度较高。虽说文字内容存在大量读音不正确，甚至不乏字形拼写错误，也存在蒙古文编码不统一，视觉顺序与文本流顺序不统一[13]等一系列问题。但这正是蒙古文目前使用的真实反映和写照，蕴含着很多可利用、可挖掘信息，是我们需要关注和有待解决的问题。我们可以很方便地通过网络爬虫（Web Crawler）爬取这些网络资源，稍作编码转换、行序恢复等预处理即可获得原始文本资源。本文原始文本特指这种只做 HTML 标签和排版格式剔除、行序恢复的未经读音和字形纠正、未经标注和其他额外处理的文本资源。网络文本资源内容丰富且获取便捷。虽然蒙古文网络应用时间不长，但已经显示出了他的生命力，以其可获得性（Accessibility）成为潜力最大的数字资源，虽然还不敢说轻松获得海量大数据，但我们有理由说便利获取较大规模未标注数据时代已经来临。我们已经站在了大数据门口，注重精加工、精处理之外，需要我们适当转变思想，将数据废气（Data Exhaust）化废为宝，从数据里提炼出有价值的信息和知识。

虽然我们已经有条件获取较大规模原始语料，但正如前所述，蒙古文原始语料还不是可直接利用的数字资源，原始语料的利用困难重重，其障碍主要来源于三个方面。

（1） 字形拼写错误泛滥且隐蔽

由于蒙古文的书写特性，尤其是连写（Cursive Joining[4]）特点注定蒙古文的字形拼写错误发生率远高于其他字母独立式的文字且更具隐蔽性，这一特性也加剧了原始语料的利用难度。据我们初步观察，只出现一次词（Hapaxes）中字形拼写错误词占据了很大一部分。

（2） 拼写形式多样化现象严重

如前所述，即使字形拼写完全正确，其编码可能相互不一致，这种"人机看法不一致"的后果是原本相同的词被区分为多个词，致使无法直接统计建模，甚至基本的查询搜索都难以实现。这种蒙古文原始语料中存在的独特且较为严重的拼写形式多样化现象成为原始语料利用的最大阻碍。

（3） 形态切分歧义多

蒙古文词间有明确的界限（多数为空格），但词由更小的单元——词素(ᠦᠭᠡᠢᠯᠡᠯ Morpheme)组成，词素才是其最小的语法和语义单位。蒙古语属黏着语(Agglutinating Language)，形态变

[4] 这与汉字草书和英文花体等艺术层面的书写方式有本质区别。常见的连写文字还有阿拉伯文（包括国内使用的维吾尔文、哈萨克文和柯尔克孜文）和希伯来文等。

化非常丰富[4][5]，尤其动词具有很强的派生能力,研究发现ᠪᠠᠢᠴᠠᠬᠤ(UILED)这个动词至少有 855 种已确定的变化形式，如果考虑词缀的阴阳性变化的话这个数目还要翻倍。蒙古语语言学一般认为蒙古文以词干(ᠦᠨᠳᠦᠰᠦᠨ stem，而不是词根 ᠦᠷᠭᠡᠯᠵᠢ root)上黏接一个或多个词缀(ᠦᠭᠡᠢᠯᠡᠯ, suffix)来完成构词和构形变化。从词法的角度,可以把蒙古语词分成静词类、动词类和无变化词类三大类。其中静词类和动词类具有各自的一系列形态变化。如果不顾词间这一关系而将各个词作为独立词进行统计学习的话，势必进一步加剧数据稀疏。例如,语料中出现ᠨᠢᠭᠡ ᠲᠥᠷᠥᠯ搭配（Collocation）的概率非常大，但出现ᠨᠢᠭᠡ ᠲᠥᠷᠥᠯᠵᠢᠬᠡᠨ搭配的概率就特别小。有了前者我们就应该给后者一个近似的概率，甚至将两者的概率合并计算才能解决好问题。形态切分是纠错文本和原始文本共同面临的问题，只是原始文本歧义更大，面临的问题更多。

2 与汉文输入法建模的区别

不带统计模型的蒙古文输入法流程如

图 1 所示，从各种词库搜索匹配输入码 k 的候选词依据某种排序算法排序后作为输入法候选词展示给用户进行选择。不管我们搜集多大的词库，都不可能是“全词”词典，不可能彻底解决 OOV（未登录词）问题，所以还需要“音码输入算法”作为补充，让用户录入 OOV，这完全类似于汉字拼音输入法中的全拼输入。除了词组库中词间搭配关系外，各单词独立输入，没有考虑词间影响。

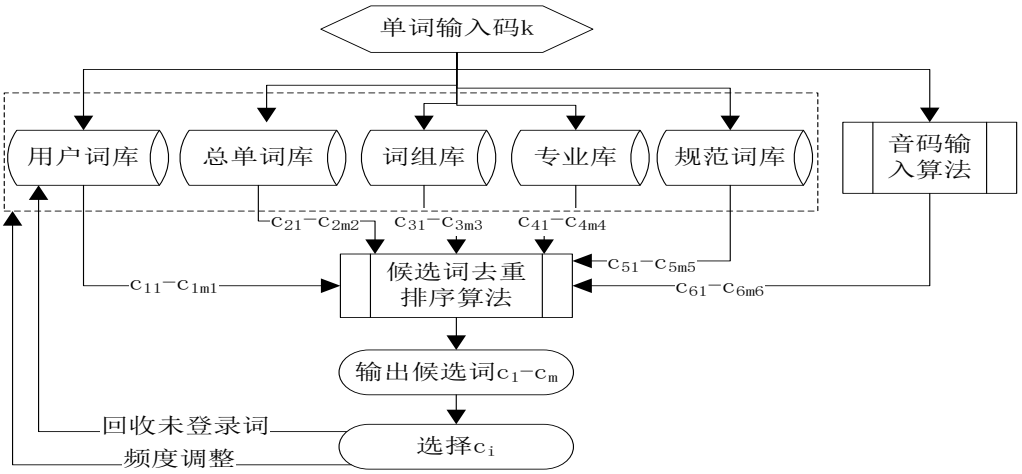


图 1 不带统计模型的蒙古文输入法流程图

基于统计语言模型的汉字输入法概念模型可简化为如图2。用户输入的拼音串经过音节切分器（Syllable Segmentor）切分为音节图（Syllable Graph 可能存在切分歧义，如果考虑简拼形式，切分歧义更多），再经过晶格构建器（Lattice Builder）构造晶格（Lattice），最后由静态统计语言模型（Static Statistical Language Model）获取最佳路径。

如果有了足量蒙古文纠错语料库，我们就可以从语料直接抽取出词典，经过ID化、三元统计、计算回退参数等过程后就可以建立带回退的静态3元模型。因单词间有明显的间隔，词库建立过程也相对简单。如果将蒙古文形态变化问题搁置，将所有单词视作完全独立的单词，那么蒙古文SLM的输入法应用上，除了输入码（相当于汉字的拼音）的切分和词库匹配策略有不同外整体思路与汉文没有什么大区别。

因为汉字是“按形编码”，所以晶格中候选项是没有读音的，原始文本中的词也是没有读音的，两者可直接匹配，词的读音标注是独立于编码的NLP问题。但蒙古文是“按音编码”，晶格中候选项是有读音的词，原始文本拼写形式多样化现象非常严重，恰恰读音不确定，字形拼写错误也比汉字更严重，原始文本中直接匹配晶格词的可能非常非

常小，更不可能形成词间搭配，从而为最佳路径选择提供概率贡献。例如，在本文6000万词级原始语料MGLNews中[illegible]一词总共出现了90485次，其读音完全正确的只有14634次，而同形异码拼写形式共出现了272种，总出现频次高达75851，占据84%，这还不包括[illegible] [illegible]等数量庞大的字形拼写错误形式。也就是说，我们根本无法直接利用原始文本解决蒙古文输入法词间最佳路径选择问题，除非我们先对文本进行有效纠错。

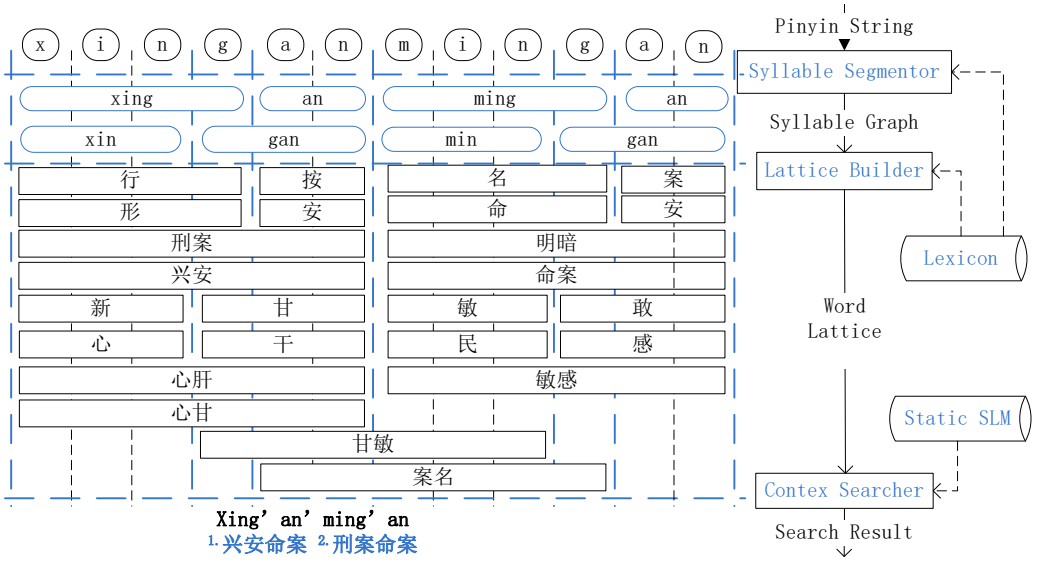


图 2 基于统计语言模型的汉字输入法概念模型图

另一方面，在某一汉字编码标准下，汉字是个封闭集合，可输入字数是确定的[5]，多音字的数量也是确定的，所以对于拼音串中的每个音节来说，其对应候选单字是固定的，这种字典较为容易获得。字层面进行统计静态语言模型即可获得单字组成的晶格中的最佳路径。如果还要支持词（尤其是多字词和词组）级别的连续录入，必须具备带拼音的词库和词组库，并对语料进行与之匹配的分词再进行建模。如果没有分好词的语料，或者是想利用原始文本，首要任务就是分词。采取什么样的分词策略主要依赖于手中资源。例如，sunpinyin 采取的是先使用正向最大匹配方法（FMM，forward maximum matching）分词。“为人民、人民、民办、办实事”等都是可能切分，所以“为人民办实事”六个单词为一个交集型歧义切分块。分词结果中的此类歧义切分部分视为一个特殊单词，以此降低后续 3 元统计受歧义切分干扰。再利用构建好的统计模型进行二次切分，尽力解决上一步遗留的交集型歧义（Overlapping Ambiguity）切分并重新构建模型。除了自己做以外，可利用的第三方切词工具和资源也很多。相对而言，蒙古文词库就是稀缺资源。

因汉字音节只有400个左右，相对个数固定，声母、韵母使用较为固定，基本规律清晰，虽然还有歧义问题存在，但总体而言，全拼拼音串自动切分成为可能，简拼形式也有规律可循。但蒙古文输入码连续输入时在音节层面和词层面都可以简拼，所以歧义更大。例如，[illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible]一句的全拼形式为“orgenagvdamhegeretalanamjimhenbaiba.”，按音节简拼时输入码变成“ogagdhgrtlnmjihbb.”，而剩余字母可随意补充而形成多种多样的输入码，如果再考虑按词组简拼，输入码可能性更多。所以，蒙古文输入码几乎任意位置和任何组合都是可能切分，基本需要全概率列举，而且每个切分单元可对应单词量特别大，计算复杂度高。但不管怎么高，始终是个可计算的，且本质上没有区别，所以本文假设输入码已经

---

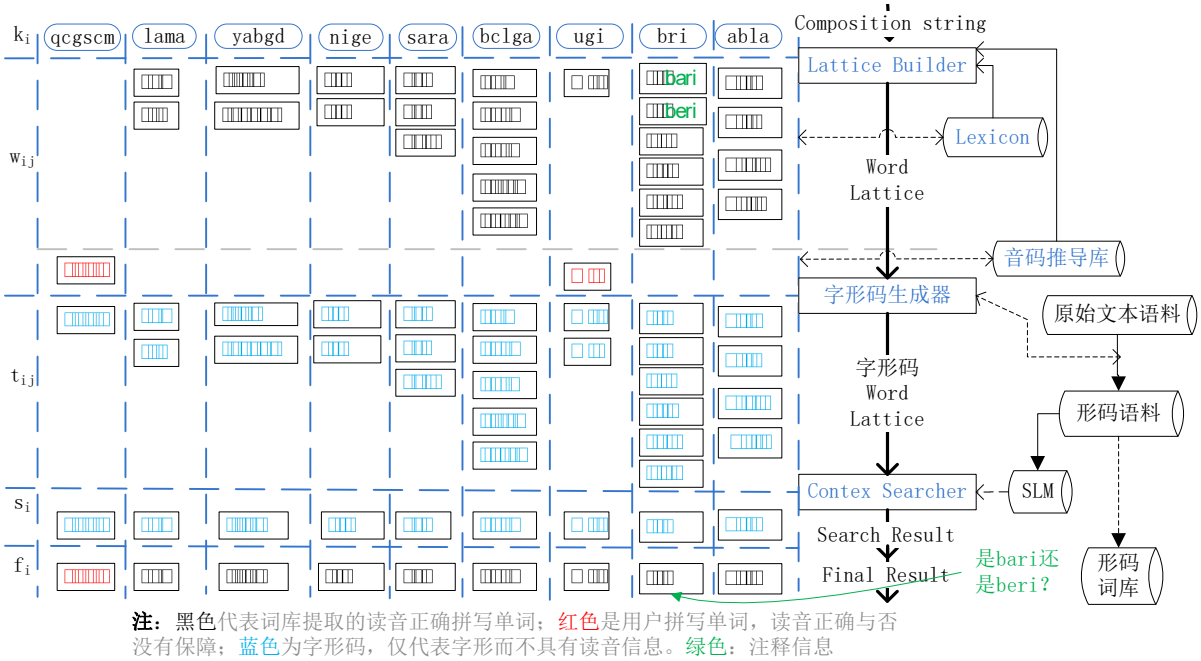
[5] GB2312 收录 6763 个汉字，GB18030 收录 70244 个汉字等，每个标准都有明确的收录汉字数。

按输入单元切分好。

3 原始文本建模

3.1. 模型

基于以上分析，本文提出如图 3 的基于原始语料的蒙古文输入法概念模型。将原始语料映射为形码语料，从形码语料抽取形码词库，构建形码层次的静态语言模型，将原本基于词晶格|w_{ij}|的词间最佳路径搜索问题转换为基于词形码晶格|t_{ij}|的最佳路径搜索问题，利用此模型获得|t_{ij}|的最佳路径|s_i|后，利用|w_{ij}|到|t_{ij}|之间的映射关系获得正确的词序列|f_i|。



注：黑色代表词库提取的读音正确拼写单词；红色是用户拼写单词，读音正确与否没有保障；蓝色为字形码，仅代表字形而不具有读音信息。绿色：注释信息

图 3 基于原始语料的蒙古文输入法概念模型图

很显然此模型利用了蒙古文输入法中多数重码词不是同形词的事实。

3.2. 评价指标

鉴于蒙古文输入法一直没有可参考的评价方法，本文选定输入码比作为评价指标。输入码比指使用一种输入法录入评价语料库内容时能够正确录入所有单词的最短输入码的长度之和与另一种输入法录入相同内容时能够正确录入所有单词的最短输入码的长度之和的百分比。计算公式如下：

$$R = \sum_{i=1}^{n}\sum_{j=1}^{m_i} len(k1_{ij}) \div \sum_{i=1}^{n}\sum_{j=1}^{m_i} len(k2_{ij}) * 100\%$$

n 是评价语料库句子数，m_i 为第 i 个评价句单词数，k1 和 k2 分别是预比较两种输入法输入码。考虑到目前我们所能比较输入法有限，本文又主要考察引入统计模型后对输入法的贡献度，所以实际选用了最短输入码[6]/生成码[7]比（I1 指标），这个指标体现的是输入法最短输入码相对于生成码的百分比，或者简单认为评价输入法与一般读音输入法输入码长度的百分比，体现该输入法较全拼式一般读音输入法的优化程度。因生成码等价于原文单词，所以

[6]输入上下文环境 $w_1w_2\cdots\cdots w_n$（一般为一个句子或一个短语）下能准确录入这个上下文所有单词的一系列输入码组合中的最短的那个组合 $k_1k_2\cdots\cdots k_n$ 称之为最短输入码序列。最短输入码序列中的 k_i 称之为 w_i 的最短输入码。

[7]如果一个蒙古文读音输入法较为严谨完整，任意一个单词 w 对应唯一一个全拼式读音输入码 k，这个 k 也只能推导出唯一一个 w，我们将 w 的这种没有二义性的输入码 k，称之为 w 的生成码。在键盘映射输入法状态下的蒙古文编码标准名义字符或对应的拉丁字母就可以看成是生成码。

此项指标实际上就是平均码长。比例越小，说明输入法可以用越简短的输入码输入。计算公式中的 $k1_{ij}$ 是第 i 句第 j 词的最短输入码，$k2_{ij}$ 是第 i 句第 j 词的生成码。

3.3. 原始语料

本文原始语料采集自中国蒙古语新闻网 MNN、中央人民广播电台民族网蒙古语网 CNR 和央视网蒙语网 CNTV 的部分新闻稿件，称其为 MGLNews，详情如表 1。

表 1 生语料数据汇总表

网站	新闻稿件数	句数	Tokens	Types
MNN	86189	4008244	49243265	312902
CNR	43857	885970	11493090	84458
CNTV	8195	180570	2039115	57216

3.4. 评价语料

本文使用两种评价语料库，两个评价语料库各有侧重点，形成互补。我们在两种评价集上分别进行评价。

(1). 零散短句评价语料库 C1

评价语料库由各自独立的 500 句人工精选的输入正确(必须保证读音正确)的短句构成。我们在一位语言学家的统一领导和指引下，组织 10 位蒙古语文专业的本科生，让每位搜集并录入校对 60 句，再由一个语言学家进行统一筛选，保留 500 句。构建原则：

① 选择复杂度适中的可连续输入词序列，避免嵌套、引用和跨标点句。虽说是句子，但不必苛求整句，每句原则不超过 10 个词（输入单元）。

② 语句符合语法规则和常规，避免过于生僻或古怪说法。

③ 避免具有明显学术争议的拼写方式和还未完全普及的外来词。

④ 尽量考虑蒙古文形态变化特性对输入的影响。例如 ᠰᠤᠷᠭᠠᠭᠤᠯᠢᠨ ᠤᠨ ᠰᠤᠷᠭᠠᠭᠴᠢᠨ ᠰᠤᠷᠭᠠᠭᠤᠯᠢᠴᠢᠳ ᠰᠤᠷᠭᠠᠭᠤᠯᠤ ᠰᠤᠷᠭᠠᠭᠤ ᠬᠠᠷᠠᠨ ᠪᠣᠯᠤᠨ᠎ᠠ᠃ 就可以考察同一词根/词干的多个派生词的录入情况。

⑤ 考虑同音词或重码词识别能力。例如 ᠡᠷᠡ ᠨᠡᠷ᠎ᠡ ᠨᠡᠷᠡ ᠲᠠᠢ ᠬᠡᠷᠡᠭ ᠮᠡᠳᠡᠭᠰᠡᠨ᠃和 9 ᠰᠠᠷ᠎ᠠ ᠶᠢᠨ ᠪ ᠤᠳᠤᠷ ᠢᠶᠠᠷ ᠬᠤᠷᠢᠶᠠᠨ ᠡᠷ᠎ᠡ ᠲᠠᠢ ᠪᠠᠭ᠎ᠠ中的ᠡᠷ᠎ᠡ ᠡᠷᠡ是同音词，而且两个词都是常用词，需要上下文才能判断出来。为此也可以构造一些原本不存的人造句：

表 2 人造测试句示意表

绑定	名	有	汉子	酒	酿
ᠡᠷᠡ	ᠨᠡᠷ᠎ᠡ	ᠲᠠᠢ	ᠡᠷ᠎ᠡ	ᠠᠷᠢᠬᠢ	ᠨᠡᠷᠡ
ere	nere	tai	ere	arihi	nere

⑥ 构造类似于ᠮᠣᠩᠭᠤᠯ ᠬᠦᠨ ᠪᠠᠢᠭ᠎ᠠ ᠪᠠᠢᠳᠠᠯ和ᠮᠣᠩᠭᠤᠯ ᠬᠦᠨ ᠪᠠᠢᠭ᠎ᠠ ᠪᠡᠷ ᠪᠠᠢᠳᠠᠯ的对比句。

(2). 整篇整句评价语料库 C2

虽然第一种评价语料库是针对输入法构造的专业评价语料库，但我们认为还应该建立一个由整篇整句的更接近实际应用的语段构成的稍微大一点的数据作为一种评价语料库更为合适。由于我们暂时无条件独立构建这样的测试集，所以我们从 100TUM 中选择了高中语文部分（排除古文部分），总共 48 篇文章，自动断句为 490 句，5000 词形（token）。选择高中语文课是考虑其涵盖短篇小说、长篇小说节选、诗歌、新闻报道等多种题材，语句类型丰富，与实际应用接近，具有较强的代表性。虽然语料库收纳内容时代较早、部分文章内容具有明显时代烙印、部分单词拼写方式经过文字规范已有少许变化，但我们认为这些问题还是可以忽略的。

3.5. 实验结果

我们以没有统计模型的蒙科立输入法 2012 版(NOLM)为基准输入法，对基于 MGLNews 的统计模型输入法（RAW）、RAW 基础上进行基于字形归并的模型优化输入法（RAW^{+}）

进行了对比评价，指标如表 3。

表 3 性能评价指标统计表

指标	零散短句评价语料库 C1				整篇整句评价语料库 C2			
	NOLM	统计模型输入法		改进	NOLM	统计模型输入法		改进
I1	69%	RAW	60%	9%	71%	RAW	62%	9%
		RAW $^{+}$	57%	12%		RAW $^{+}$	60%	11%

实验结果表明，本文提出的生语料建模方法有效提高了指标，明显缩短了输入码长度，且基于字形归并的模型优化后进一步获得提高。

C2库上的RAW指标和RAW^{+}指标普遍要比C1库指标略低。句子普遍长于C1语料、大篇幅文章切句、切词可能存在歧义、量大等都有可能是原因。但这些对本文说明目标没有明显影响，所以暂时没有进行进一步细致分析。

4 总结

这一模型可以很好地解决针对输入法应用的原始语料建模问题，开启了针对原始语料利用和建模的研究，对蒙古文所有音词转换（语音输入、读音输入等）和形词转换（校对、OCR识别等）应用都有很好的参考和借鉴作用，也是纠错语料建设的必要辅助途径。

本文提出的模型存在如下几个特点。

4.1. 需要大词库支持

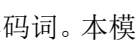
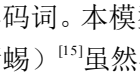
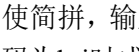
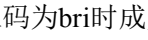
本模型中我们要利用的是原始语料，单词读音是不可靠的，甚至字形都有可能是错误的，所以不能再像基于纠错语料的模型一样从语料中直接抽取单词构建词库。独立于语料构建的词库必然无法保证一一映射语料单词，语料已无法用词库ID化，所以还需要一个额外的字形码的词库。大词库词表和语料库形码词表是有交集的两个独立集合。对于输入码$|k_i|$，对应的$|w_{ij}|$只取决于大词库，与语料无关，所以大词库必须大到能够满足日常一般应用为宜。因为大词库中单词都是读音字形拼写正确的单词，所以可以使用多种简拼方式录入这些单词，并形成候选单词的晶格。显然，如果大词库不够大，实验难以获得理想结果。

4.2. 用音码输入 OOV

词库再大也不可能避免OOV问题，所以本模型依然采用音码输入方式录入OOV。对于语料中存在，大词库中未能涵盖的OOV，本模型同样可以利用SLM选择包含这个OOV的最佳路径。对于语料库和词库中都未能出现的OOV，我们不做额外处理。因对OOV暂未能采取任何措施，其正确性完全依赖于用户录入。如图3中的  一词不一定是qcgscm，也有可能是qvgsvm或者其他什么。

工程实现上可以保存OOV，构建单独的用户词典来允许简拼方式录入已录入OOV。如果能够构建一个较好的OOV回收策略，从语料库众多拼写形式中猜测出正确拼写形式并扩充词库是新词回收的重要途径。但本文暂且搁置此问题而认为用户输入是正确的，本文实验数据只考虑OOV用音码全拼方式录入情况。

4.3. 同形重码的影响

所谓同形重码是指字形相同的重码词，就是一个输入码k的候选词列表中的c_i和c_j是同形异码词。本模型可较好解决多数同形词问题，例如,  (horbel 到那里)和  （gurbel 蜥蜴）[15]虽然在SLM中被混为一个字形，但两个同形词的读音不同，必输首字母h，g不同，即使简拼，输入码也不会重合，不受影响。但是  (beri 媳妇)和  (bari 抓这个)在输入码为bri时成为同形重码（如图3中的输入码bri一列）。这是这一模型有待进一步研究问题。有两种解方案：a、需要结合纠错语料来解决。当遇到同形重码时用纠错语料的SLM进行选择；b、遇到同形重码时提示补充输入码或从候选中手工选择。

输入法没有必要区分᠋ᠠ᠌ᠮᠠᠭᠠᠨ᠋（去家里）和ᠮᠠᠭᠠᠨ᠋ ᠲᠤᠨᠭᠠᠯᠠᠭ（迸火苗）、ᠠᠪᠤᠯ（家庭）和ᠠᠭᠤᠷᠠᠮ ᠠᠪᠤᠯ（搅拌酸奶）等多义词（同音同形异议）。蒙古文作为音素文字，如果始终用全拼方式录入，候选词必然只有同音异形词，不存在同形重码。

4.4. 词形拼写错误和形态变化影响

原始语料库中词形拼写错误和形态变化是导致数据稀疏的两大重要因素，而这两项对原始语料的利用也非常重要，每一项都是值得专项研究的重要课题。为此本文只是对比了基于字形归并的模型优化结构，有待进一步细化研究。

4.5. 自动纠错

本文模型的核心思想是在对原始语料不进行纠错情况下的利用和建模问题，但词形归并和形态分析实际上又部分回归自动纠错问题上，只是暂时不进行武断纠错后再利用，而是利用中逐步对文本进行排歧和纠错。汉字拼音输入法中也有文章引入了纠错功能[15][17]，这本身对蒙古文来说具有很好的参考价值，值得我们进一步深入研究。

正文在 8000 字左右为宜。

参考文献

[1]. The Unicode Consortium[EB]. http://www.Unicode.org .

[2]. 白双成，张劲松，苏雅拉图. 蒙古文拼写形式多样化问题研究[C]//CCL2015 论文集.广州.2015.

[3]. 张小衡. 中文的同形异码字问题[J]. 中文信息学报, 2015, 29(4):144-150.

[4]. 那顺乌日图,雪艳,叶嘉明. 现代蒙古语语料库加工技术的新进展--新一代蒙古语词语自动切分与标注系统(Darhan Tagging System) [C]//第十届全国少数民族语言文字信息处理学术研讨会论文集. 青海. 2005.

[5]. 那顺乌日图. 蒙古文词根、词干、词尾的自动切分系统[J]. 内蒙古大学学报(人文社会科学版). 1997.

[6]. 侯宏旭,刘群,刘志文. Skip2N 蒙古文统计语言模型[J]. 内蒙古大学学报, 2008, 39 (2).

[7]. 赵伟,侯宏旭,从伟等.基于条件随机场的蒙古语词切分研究[J]. 中文信息学报. 2010, 24(5).

[8]. 应玉龙，李淼，乌达巴拉等. 基于条件随机场的蒙古语词性标注方法[J].计算机应用 2010.8 月.

[9]. 姜文斌,吴金星,乌日力嘎等. 蒙古语有向图形态分析器的判别式词干词缀切分[J]. 中文信息学报. 2011(04).

[10]. 苏传捷,侯宏旭,杨萍等. 基于统计翻译框架的蒙古文自动拼写校对方法[J]. 中文信息学报. 2013(06) .

[11]. 确精扎布.关于现代蒙古语文语料库.内蒙古大学学报（蒙文版）.1992.第一期.

[12]. 确精扎布. 确精扎布蒙古文信息处理专辑[M] . 呼和浩特: 内蒙古教育出版社, 2014.

[13]. 白双成.蒙古文网站内容管理系统研究[J].第十二届全国少数民族语言文字信息处理学术研讨会. 拉萨. 2009.

[14]. 白双成，张劲松，呼斯勒. 蒙古文输入法输入码方案研究[Ｊ]. 中文信息学报 2013(06):169-174.

[15]. 淑琴. 蒙古文同形词知识库的构建[D]. 内蒙古大学. 2010.

[16]. Chen Zheng, Lee K F. A new statistical approach to Chinese Pinyin input[C]// Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2000:241--247.

[17]. Zheng Yanbin, Li Chen, Sun Maosung. CHIME: An Efficient Error-Tolerant Chinese Pinyin Input Method.[C]//Twenty-second International Joint Conference on Artificial Intelligence-volumethree. 2011:2551-2556.

 白双成（1974—），男，博士，研究员，主要研究领域为蒙古文信息处理。Email:331869327@qq.com

作者联系方式：
白双成
呼和浩特市大学东街 129 号 内蒙古社会科学院 MIT 中心
010020
13947141379
bsc@menksoft.com