文章编号：

基于统计翻译框架的蒙古文自动拼写校对方法

苏传捷[1]，侯宏旭[1]，杨萍[1,2]，员华瑞[1]

（1．内蒙古大学计算机学院，内蒙古自治区 呼和浩特 010021；

2．临汾职业技术学院计算机系，山西 临汾市 041000）

摘要：在以国际标准编码存储的传统蒙古文电子文本中，拼写错误十分普遍。人工校对这些错误不仅速度慢而且成本高。本文提出了一种基于统计翻译框架的传统蒙古文自动拼写校对方法，将拼写校对看作是从错误词到正确词的翻译。本文使用改进的基于短语的统计机器翻译模型来构建拼写校对模型，然后对测试文本进行校对。实验结果表明，本文方法可以快速、有效地校对拼写错误，而且不依赖于特定语言的语法知识。使用本文方法对包含 1026 个正确词、1102 个错误词的测试集进行拼写校对，校对后文本中的正确词所占比例最高可达 97.55%。

关键词：蒙古文；拼写检查；拼写校对；机器翻译

中图分类号：TP391　　**文献标识码：**A

A Spelling Correction Method for Traditional Mongolian Based on Statistical Translation Framework

SU Chuanjie[1], HOU Hongxu[1], YANG Ping[1,2], YUN Huarui[1]

(1. College of Computer Science, Inner Mongolia University, Hohhot, Inner Mongolia 010021, China;

2. Linfen Vocational And Technical College, Linfen , Shanxi 041000, China)

Abstract: In traditional Mongolian electronic texts encoded in Unicode, spelling errors are very common. The cost of correcting spelling errors artificially is extremely high. This paper proposed an automatic spelling correction method for traditional Mongolian based on statistical machine translation framework, and we regard spelling correction task as a translation work which translates the wrong words to the correct words. This paper used the improved phrase-based statistical machine translation model to build spelling correction model. We use this model to correct the raw text. We used a test set which contained 1026 correct words and 1102 wrong words to test our method, Experimental results show that our method can correct spelling errors quickly and efficiently without special language knowledge. The percentage of correct words in our proofread text can reach to 97.55.

Key words: Mongolian; spelling check; spelling correction; machine translation

1 引言

随着信息时代的到来，各种各样的文本资料以数字形式被广泛地存储在本地硬盘或者互联网上。我们可以很容易地获得大量的电子文本。但是这些电子文本的质量普遍不高，存在不少拼写错误，需要进行拼写校对。人工校对不仅无法应对飞速增加的文本数据，而且成本极高。所以，许多学者致力于自动文本校对方法的研究。Pollock 等人曾就科学文章的自动校对问题进行了深入的探讨[1]。Kukich 对英语文本中存在的错误进行了系统的分类，并总结

* **收稿日期：**　　**定稿日期：**

基金项目：工信部电子信息产业发展基金课题“蒙古文辅助翻译与电子辞典软件”

作者简介：苏传捷（1987——），男，硕士研究生，主要研究方向：自然语言处理、机器翻译；杨萍（1987——），女，硕士研究生，主要研究方向：自然语言处理；员华瑞（1988——），男，硕士研究生，主要研究方向：自然语言处理。

通讯作者：侯宏旭（1972——），男，教授，博士生导师，主要研究方向：中文信息处理、信息检索。

了相应的校对方法[2]。上世纪九十年代，国内的一些学者也开始了对中文文本自动校对技术的研究。骆卫华、张仰森等很多学者从不同的角度讨论了中文文本中存在的问题、解决方案、以及中英文自动校对的异同[3,4]。

近几年来，国内学者对蒙古文等少数民族文字信息处理的研究越来越多，对文本质量的要求也越来越高，这就对蒙古文电子文本自动校对技术提出了更高的要求。传统蒙古文是拼音文字，其拼写错误类型与英文相近。值得注意的是，传统蒙古文是典型的黏着语，其形态变化丰富而且拼写系统复杂。因此，传统蒙古文文本的自动校对面临着更大的挑战。很多学者们针对传统蒙古文的自动校对问题提出了可行的方案。斯•劳格劳利用基于字形的词法分析器和基于读音的词法分析器，设计了基于不确定有限自动机和有限自动机的自动拼写校对算法[5,6]。敖敏等人对蒙科立编码的传统蒙古文文本中的同形异码问题进行了深入的研究，并依此编制了基于同形字符替换以及符合字符拆分的语言规则来进行自动拼写校对[7]。赵军等人介绍了一种基于音节的统计语言模型的自动校对方法，并重点探讨了模型的设计和可行性[8]。

本文提出了一种基于统计翻译框架的传统蒙古文自动拼写校对方法，将拼写校对工作看成从错误词到正确词的翻译。我们首先利用人工校对前后的对应文本，训练基于改进的短语翻译模型[9]的拼写校对模型。然后应用该校对模型对测试文本进行拼写校对。我们对包含 1026 个正确词、1102 个错误词的测试集进行了校对测试。测试结果表明，使用本文方法校对后的文本中正确词所占比例最高达到了 97.55%。而且，本文提出的方法不依赖于语法知识，很容易移植到其它语言上或者与其它方法结合使用。

本文的第二部分介绍基于统计翻译框架的拼写校对模型。在第三部分，我们将讲解模型的训练与校对方法。第四部分是实验以及相关分析。最后是总结与展望。

2 基于统计翻译框架的拼写校对模型

2.1 自动拼写校对与机器翻译

从本质上看，自动拼写校对与机器翻译的工作是十分相似的，都是将一种形式的文本转化成另一种形式的文本。因此，在本文中我们将错误词 E 校对成正确词 $\hat{C}$ 的过程看作是一个翻译的过程。根据基于噪声信道模型的机器翻译方法[10]，有：

$$\hat{C} = \underset{C}{\operatorname{argmax}} P(C|E) = \underset{C}{\operatorname{argmax}} P(E|C) \cdot P(C) \tag{1}$$

然而，自动拼写校对与机器翻译的目标略有不同。拼写校对更加注重校对后词汇的正确性，也就是上式中等号右端的 $p(C)$。如果想要利用机器翻译框架进行自动拼写校对就需要改进上述公式为：

$$\hat{C} = \underset{C}{\operatorname{argmax}} P(E|C)^{\alpha} \cdot P(C)^{\beta} \tag{2}$$

其中，α 和 β 是权值，用来调解 $P(E|C)$ 和 $P(C)$ 的权重。

2.2 传统蒙古文拼写错误的特点与翻译模型的选择

传统蒙古文的词汇结构复杂且形态变化丰富，字母的书写形式与其所在的位置以及前后字母有着密切的联系。因此，传统蒙古文的国际标准字符编码系统也比较复杂，包括“名义字符”和“变形显现字符”两个部分。“变形显现字符”本身不占码位，只是“名义字符”在当前位置、前后字符或者控制符作用下的显现形式。这种设计导致不同的字符编码的组合可以键入形状完全相同的字符串（在本文中，字符串特指字符编码的序列，而不是屏幕上显示的形状序列）。混用这些字符串是造成传统蒙古文拼写错误的主要原因。

目前常用的统计机器翻译模型有：基于词的统计机器翻译模型[11]、基于短语的统计机器翻译模型、基于层次短语的统计机器翻译模型[12,13]以及基于语言学句法知识的统计机器翻译模型[14,15]。根据传统蒙古文中拼写错误的原因与特点，我们选择基于短语的统计机器翻译模型作为框架来构建拼写校对模型。基于词的统计机器翻译模型很难捕获传统蒙古文中的字符搭配信息。基于层次短语的统计机器翻译方法的优势在于可以进行长距离调序。但传统蒙古文的拼写错误中不存在长距离错序，所以不选择基于层次短语的翻译模型作为校对框架。另外，由于没有传统蒙古文的词法树库，所以基于语言学句法知识的统计机器翻译方法目前无法迁移到传统蒙古文的拼写校对工作中。

2.3 基于改进的短语翻译模型的拼写校对模型

从本节开始，我们称 E 为待校对词。对于待校对词 E 和校对后词 C，假设 E 包含 i 个字符（对应于短语翻译模型中的词），记作 E_1^i，C 包含 j 个字符，记作 C_1^j。另设 E_1^i 可以被分割为任意 k 个字符串（对应于短语翻译模型中的短语），记为 $\tilde{E}_1^k$。相应地，校对 E 生成的 C 也包含 k 个字符串，记作 $\tilde{C}_1^k$。那么，公式(2)可以扩展成以下形式：

$$\hat{C} = \underset{C}{\operatorname{argmax}} P(C|E)$$

$$= \underset{C,\tilde{E}_1^k}{\operatorname{argmax}} \sum_{\tilde{E}_1^k} P(\tilde{C}_1^k, \tilde{E}_1^k | E_1^i)$$

$$= \underset{C,\tilde{E}_1^k}{\operatorname{argmax}} \sum_{\tilde{E}_1^k} P(\tilde{E}_1^k | E_1^i) \cdot P(\tilde{C}_1^k | \tilde{E}_1^k, E_1^i)$$

$$= \underset{C,\tilde{E}_1^k}{\operatorname{argmax}} \sum_{\tilde{E}_1^k} P(\tilde{E}_1^k | E_1^i) \cdot P(\tilde{C}_1^k | \tilde{E}_1^k)$$

$$\xrightarrow{improved} \underset{C,\tilde{E}_1^k}{\operatorname{argmax}} \sum_{\tilde{E}_1^k} P(\tilde{E}_1^k | E_1^i) \cdot P\left(\tilde{E}_1^k | \tilde{C}_1^k\right)^{\alpha} \cdot P(\tilde{C}_1^k)^{\beta}$$

$$= \underset{C,\tilde{E}_1^k}{\operatorname{argmax}} \sum_{\tilde{E}_1^k} P(\tilde{E}_1^k | E_1^i) \cdot P\left(\tilde{E}_1^k | \tilde{C}_1^k\right)^{\alpha} \cdot P(C_1^j)^{\beta} \quad (3)$$

使用本文方法进行拼写校对的过程就是为每个待校对词 E 找到相应的 $\hat{C}$。首先，为 E 找到一个合适的分割方式。然后，将分割开的每一个片段分别校对，并按顺序拼接成 C。最后，找出最合适的 $\hat{C}$ 作为校对结果。

3 模型训练与拼写校对器

3.1 训练拼写校对模型

与训练机器翻译模型一样，训练拼写校对模型也需要平行语料。但是，训练拼写校对模型的平行语料与训练翻译模型的平行语料存在一些不同。第一，训练拼写校对模型的平行语料的源端与目标端是同一种语言，其中源端是未进过人工校对的原始文本，而目标端是人工校对后的正确文本。第二，训练拼写校对模型的平行语料中的每一行是一个词。第三，拼写校对模型的基本校对单位是字符，所以平行语料每一行都需要以字符分开。

我们使用 Moses 开源工具包[16]中的短语机器翻译模型训练工具和 SRILM 开源工具包[17]

来训练我们的拼写校对模型。在训练中我们沿用机器翻译的经典特征集，包括：正反向翻译概率、正反向词汇化权重、语言模型概率等。与训练统计机器翻译模型不同，我们在训练参数权重时使用校对后文本中正确词所占比例作为优化目标，与拼写校对的评价标准保持一致。

3.2 拼写校对器

我们使用改进的基于短语的统计机器翻译的 CKY 解码器作为拼写校对解码器。对于每一个待校对词 E，拼写校对解码器使用柱搜索算法搜索最优的校对结果$\hat{C}$。在理想情况下，一个正确的词会被解码成一个新的词，但是该新词与其原有编码完全相同；而存在拼写错误的词将被解码成其应有的编码序列，这个编码序列构成一个正确的词。

4 实验

4.1 实验配置

我们使用包含 3 万词的人工校对前后的两份文本作为平行训练语料。我们称人工校对前后的两个词为一条训练数据，其中人工校对前的词称为源端，人工校对后的词称为目标端。我们将这 3 万条训练数据分为两类：正训练数据和反训练数据。在正训练数据中，源端与目标端相同，且为拼写正确的词；反训练数据中的源端是拼写错误的词，目标端是校对后的正确词。

为了考查正反两类训练数据对拼写校对模型校对能力的影响，我们没有使用全部 3 万训练数据来训练拼写校对模型，而是按不同的比例选取并混合正反两类训练数据，构建五组不同的训练集，分别训练拼写校对模型。在实验中，我们保证每组训练集中的数据数目是相等的。各组训练集中正训练数据与反训练数据的比例见表 1。

表 1 各组训练集中正反训练数据的比例

训练集	正训练数据所占比例（%）	反训练数据所占比例（%）
No.1	100	0
No.2	75	25
No.3	50	50
No.4	25	75
No.5	0	100

本文使用没有重复数据的测试集进行测试。测试集包含 2128 个词，其中拼写正确的词的个数是 1026，占总词数的 48.21%；拼写错误的词的个数是 1102，占总词数的 51.79%。

4.2 实验方法与结果

我们首先使用五组训练集分别训练拼写校对模型，并使用五个拼写校对模型分别对测试集进行了开放测试。然后，将测试集及相应的参考集加入五组训练集重新训练拼写校对模型，再进行封闭测试。表 2 中列出了各组实验结果。

表 2 实验结果

编号	开放测试后正确词比例（%）	封闭测试后正确词比例（%）	校对速度（词/秒）
No.1	49.24	97.18	122.44
No.2	90.31	96.33	45.86
No.3	90.78	96.38	33.77
No.4	90.41	97.03	28.64
No.5	93.09	97.55	26.93

从实验结果中我们可以看到，开放测试中，当训练集中的数据全部为正训练数据时，拼写校对效果最差；随着训练集中正训练数据所占的比例减少，反训练数据所占的比例增大，校对效果越来越好；当训练集中的数据全部为反训练数据时，校对效果达到最佳。

通过分析实验结果我们发现，正训练数据与反训练数据的作用是不同的。正训练数据使模型保持正确的词不被校对成错误词，而反训练数据使模型尽量修改错误的词。这是造成表2中实验结果的根本原因。

完全使用正训练数据训练的拼写校对模型中只包含保持正确拼写的信息，不包含将拼写错误改正的信息。因此，该模型几乎无法纠正测试集中拼写错误，反而会将个别正确的词修改成错词。随着训练集中反训练数据所占比例的增大，训练得到的拼写校对模型的纠错能力逐渐增强，校对后文本中正确词所占的比例会逐步增大。当训练集全部由反训练数据组成时，对测试集的校对效果最好，校对后文本中正确词所占比例达到最高。因为全部使用反训练数据进行训练，模型中可以覆盖更多对错误的纠正信息。值得注意的是，在将错词纠正的同时，该模型只会将极少的正确词修改为错误词。这是由于单词中的拼写错误一般发生在局部，因此即使是反训练数据，也会蕴含足够多的正训练数据所包含的信息。

封闭测试的实验结果所反映的趋势与开放测试基本相同。明显的差异在于当训练集中均为正训练数据时，校对的效果仅次于最好的一组实验。为此我们做了一组补充实验：仅使用测试集和相应的参考集训练一个拼写校对模型，再用其对测试集进行校对。校对后正确词的比例高达 97.93%，这个结果是显而易见的。在第一组封闭测试实验中，由于正训练数据不包含任何纠错信息，所以没有为只使用测试集训练的拼写校对模型引入噪声。因此，使用全部正训练数据和测试数据训练的拼写校对模型，可以取得97.18%的好成绩。

从实验结果中我们还可以看到，每一组的封闭测试的成绩都略高于开放测试。这是因为封闭测试的测试集是包含在训练集中的，因此测试集中的错词和正确词都更倾向于被翻译成正确的形式。实际的语料中包含的错误往往是重复出现的。我们对另外一组语料进行了错词统计，45%以上的错词出现不止一次。因此使用本文方法进行实际的拼写校对的效果会介于封闭测试和开放测试之间。

文献5提出的基于自动机的自动校对算法在该文献中表达地比较精炼，难以依照论文实现，因此本文只能列举文献中的校对结果作为对比参考。文献5提出的方法在不同的测试集上的纠错准确率在92.7%到96.1%之间。

另外，本文提出的方法在实际应用中可以根据校对任务的具体需求来选择训练拼写校对模型的语料。如果当前拼写校对任务可以容忍将一些正确的词误改为错误的词，但是需要最大限度地追求错词校正率（纠正为正确词的数目与总错词数的比值），则可以在训练拼写校对模型时使用较多的反训练数据。如果要求尽量保证原有正确的词不被篡改，则可以使训练集中的反训练数据略多于正训练数据，就可以获得理想的校对效果。

5 总结与展望

针对传统蒙古文国际标准编码文本普遍存在的拼写错误，本文提出了一种基于统计机器翻译框架的自动拼写校对方法，并将基于短语的统计机器翻译模型改进成适用于传统蒙古文的自动拼写校对模型。实验结果表明，本文提出的方法可以有效地对传统蒙古文文本进行自动校对，且校对后文本中正确词的比例最高可达97.55%。本文方法不需要特定的语法知识，可以很好地移植到其它语言的自动校对工作中。我们希望将来能够引入传统蒙古文的语言特征，继续提高本文方法的校对能力，同时尽量避免将正确词修改成错误词的情况。我们也将尝试将本文方法与现有的传统的基于规则和词典的方法进行融合，提高现有自动校对系统的

校对水平。我们还希望可以扩展本文提出的方法，使其可以校对上下文相关的真词错误，实现更加实用的自动校对系统。

参考文献

[1] Joseph J. Pollock. Automatic Spelling Correction in Scientific and Scholarly Text[J]. Communication of the ACM, 1984 , (4) : 358-368.

[2] K. Kukich. Techniques for Automatically Correcting Words in Text[J]. ACM Computing Surveys, 1992, 24(4) : 377 - 438.

[3] 骆卫华，罗振声，宫小瑾．中文文本自动校对技术的研究[J]．计算机研究与发展，2004,1(1):244-249.

[4] 张仰森，俞士汶．文本自动校对技术研究综述[J]．计算机应用研究，2006,23(6):8-12.

[5] 斯·劳格劳．基于 DFA 的蒙古文自动校对算法[C]．第二届少数民族青年自然语言处理研讨会，2010．

[6] 斯·劳格劳．基于不确定有限自动机的蒙古文校对算法[J]．中文信息学报，2009,23(6):110-115．

[7] 敖敏,熊子瑜,呼和．基于蒙科立输入法的同形异码词研究[C]．第十一届全国人机语音通讯学术会议，2011,10．

[8] 赵军，敖其尔，吉仁尼格，巩政等．基于统计语言模型蒙古文词汇分析校正器的设计与实现[C]．第十一届全国民族语言文字信息学术研讨会，2007,2．

[9] Philipp Koehn, Franz Josef Och, Daniel Marcu. Statistical Phrase-Based Translation[C]. In Proceedings of HLT-NAACL, 2003, pages 127–133.

[10] Peter F. Brown, John Cocke, Stephen A. Della Pietra. A Statistical Approach to Machine Translation[J]. Computational Linguistics, 1990, 16(6): 79-85.

[11] Peter F. Brown, V. J. D. Pietra, Stephen A. Della Pietra. The Mathematics of Statistical Machine Translation: Parameter Estimation[J]. Computational Linguistics., Vol. 19, No. 2. (June 1993), pp. 263-311

[12] D. Chiang. A Hierarchical Phrase-Based Model for SMT[C]. In: Proc. of the ACL. Ann Arbor: Association for Computational Linguistics, 2005. 263-270.

[13] D. Chiang. Hierarchical Phrase-Based Translation[J]. Computational Linguistics, 2007, 33(2): 201-228.

[14] Yang Liu, Qun Liu, Shouxun Lin. Tree-to-string Alignment Template for Statistical Machine Translation[C]. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 2006, pp. 609-616.

[15] Yang Liu, Yun Huang, Qun Liu, Shouxun Lin. Forest-to-String Statistical Translation Rules[C]. Annual meeting of the Association for Computational Linguistics, 2009. pp. 704-711.

[16] Philipp Koehn, Hieu Hoang, Alexandra Birch. Moses: Open Source Toolkit for Statistical Machine Translation[C]. Proceedings of the ACL demonstration session, 2007.177-180.

[17] Andreas Stolcke. SRILM : an Extensible Language Modeling Toolkit[C]. In Proceedings of the International Conference on Spoken Language Processing, volume 2, pages 901–904.

作者： 苏传捷 地址：内蒙古自治区呼和浩特市赛罕区大学西路235号 内蒙古大学计算机学院 邮编：010021 电话：13848918791 电子邮箱：suchuanjie@foxmail.com