融合短语统计机器翻译的拼音-英文辅助写作技术的研究 *

尹宝生，陆瑞雪，吴闯，张桂平，叶娜

（沈阳航空航天大学 人机智能研究中心，辽宁 沈阳 110136）

摘要：针对以汉语为母语的用户，区别于传统的多样性的词推荐和句子实例推荐，本文提出了一种基于跨语言输入的英文辅助写作的方法，建立基于短语的拼音-英文统计机器翻译系统，实现了拼音-英文的直接转换，并通过选取词频、语言模型和词性等特征对音英转换的结果进行了消歧。实验证明，本文提出的音英转换的方法能代替用户对写作难点的查找和选择过程，在一定程度上能够提高用户的写作效率。

关键词：英文辅助写作；统计机器翻译；音英转换

中图分类号：TP391　　**文献标识码：**A

Research on Fusion of Pinyin-English Auxiliary Writing Technology and Phrase-based Statistical Machine Translation

YIN Baosheng, LU Ruixue, WU Chuang, ZHANG Guiping, YE Na

(Human-Computer Intelligence Research Center, Shenyang Aerospace University, Shenyang, Liaoning 110136, China)

Abstract: Different from traditional various words and sentence samples recommendation method, this paper proposes an English auxiliary writing method based on inputting method over the language for Chinese. Through building of phrase-based statistical Pinyin-English machine translation system, the paper implements direct conversion from Pinyin to English, and then disambiguate the conversion result through features of word frequency, language model and word tagging. The experiment shows that Pinyin-English conversion method in this paper can replace search and selection process for difficulty words or phrase of a translator, and could enhance writing efficiency to in some degree.

Key words: English auxiliary writing；statistical machine translation；Pinyin-English conversion

1 引言

用户在英文写作过程中，因受语言水平的限制而经常有想不到合适的词语来地道的表达想法，此时往往需要通过查词典或者借助于搜索引擎等方式来搜索写作素材或者范例，这种查询方式查找到的信息量大，工具的切换和人工的筛选等过程也费时费力且很容易打断写作思路，因此针对此情况，市场上出现了很多写作辅助助手。现有的的英文辅助输入分为两种：一种是文章批改形式的助手，即在写作过程中对已输入内容进行错误检查并提示，如 StyleWriter、有道英文助手等，主要包括拼写错误检查、语法错误检查、单词错误检查、句式错误检查等；一种是实时英文输入助手，即在写作交互过程中对用户需要的内容进行个性化推荐，分别在词级别和句子级别对英文写作进行辅助，以减少用户写作时间，提高写作效率。词级别的辅助输入例如谷歌英文写作助手、金山写作助手等，主要是基于英文单词的匹配推荐，即根据用户输入单词的片段对单词或者常用短语搭配进行补全；句子级别的辅助输入主要基于实例的辅助写作系统[1]，即双语例句的匹配检索，根据用户输入的单词或者句子，通过匹配查找和用户输

* **基金项目：**辽宁“百千万人才工程项目”（04021401）；国家自然科学基金（61402299）

作者简介：尹宝生（1975—），男，副教授，主要研究方向为自然语言处理，知识管理，机器翻译；陆瑞雪（1991—），女，硕士研究生，主要研究方向为自然语言处理，机器翻译；吴闯（1985—），男，硕士，主要研究方向为自然语言处理，机器翻译；张桂平（1962—），女，教授，主要研究方向为知识工程管理，机器翻译，人机智能接口；叶娜（1981—），女，讲师，博士，主要研究方向为自然语言处理，机器翻译。

入意思相近或者包含用户输入的例句来给用户借鉴和学习。

有些研究者[2-3]主要针对文章写作中不能明确自己的写作需求对多样性的词推荐或者句子推荐进行了研究。另外一些研究者在非母语的写作辅助输入中考虑到母语和目标语的翻译信息：尹宝生[4]指出了人们在外文输入时经常遇到的困难和难题，提出了开发跨语言输入的方法；张桂平等人[5]设计的多文种输入助理软件实现了日文汉音输入法等，提出以汉语的方式输入日文的方法，为对日文基础少的用户提供了方便；统计机器翻译信息还被考虑融合到汉语拼音输入法中[6-9]，实现了拼音输入法和翻译信息的智能结合，提高了翻译效率。

在实际写作过程中，因为思维习惯往往以母语的角度进行思考并将母语转换为目标语进行写作，因此可以将以汉语为母语的用户进行英文写作的实际过程看作拼音-英语的转换过程。现有的辅助输入技术在音-英转换过程中更多的考虑到了用户进行选择和修改，没有实现直接的音-英转换，可以在一定程度上帮助用户进行英文的写作，但是对于非英文专业水平的用户来说，英文写作最大的问题是受语言水平的限制，不能表达出地道的英文句子，在选择和修改的过程中耗费的时间也较多，因此在实时写作过程中的辅助效果不高。针对以汉语为母语的用户，本文提出了一种新的辅助输入方法，以提高英文写作效率和推荐正确率为主要目的，结合规则和统计的方法实现了拼音到英文的直接转换，利用短语机器翻译系统生成英文候选词并进行初步排序，并使用语言模型对推荐英文单词进行消歧。

2 音英统计机器翻译系统

传统的英文输入模式，即拼音-中文-英文的输入模式，是基于两级转换的方式生成候选集，首先进行拼音-汉语的转换，再进行汉语-英语的翻译。这种输入模式存在着以下的问题：音字转换和汉英翻译过程中都会产生较大的候选集，信息量很大，以“综合治理”为例，当用户不知道“治理”该如何表达时，根据传统的英文输入方式得到的结果集如表 1 所示；若输入拼音为简拼时，根据谷歌输入法音字转换的结果得到的汉语候选词如图 1 所示，根据汉语候选词会得到更大的单词候选词集，这为选词带来了更多不便性；另外传统的英文输入方式不能很好的对候选进行时态、单复数和动名词等形式的变换。

表 1 拼音-英语转换带来的歧义

拼音	汉语	英语
Zhili	治理	government, administration, administer, bring under control, harness
	致力	endeavor, work for, devote oneself to
	智力	intellectual, intelligence, intellect, intellective
	直立	erect, erective, upstanding
	支离	fragment, broken, disorganized, incoherent, incoherently, trivial and humbled

拼音-英语的跨语言输入模式可以将拼音的输入和查询筛选过程智能的结合起来，屏蔽内部转换过程，实现拼音-英语的直接转换，例如，输入拼音“zgd”，我们将直接得到“Chinese”，“the most expensive”等结果。这种输入方式能更便捷直接的得到用户想要的内容，因此建立一个音英机器翻译系统是有必要的。

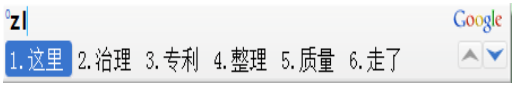

图 1 音字转换结果示例

现有的机器翻译系统是基于汉英的翻译或者其他多语言翻译，是在句子级双语对齐的基础上建立的，可以看作不同的码串之间的翻译。因此我们将拼音-英语的转换看

作不同的码串，建立拼音-英语的短语机器翻译系统。由东北大学开发的开源机器翻译系统 Niutrans[10]支持多个翻译模型，其中包括基于短语的模型，自发布以来，已经被研究者广泛的应用。建立基于短语的音英机器翻译系统可以看作拼音串到英语的翻译，主要步骤如下：

1 对双语平行语料中的汉语语料进行分词和拼音标注，拼音标注使用 MIT 研发的拼音 pypinyin1。

2 对标注好的汉语语料进行多音字替换，如“了解”被标记为 “lejie”、“说服”被标记为“shuofu”等。

4 将汉英词对齐信息映射到音英语料，训练基于短语的音英统计机器翻译系统。

3 语言模型

在拼音输入法中，音字转换是研究的核心内容，主要目的是将汉语拼音转换为若干个相关的汉字单元的候选，主流的解决方法是使用词频信息和语言模型来对候选词进行排序。语言模型的方法可以给出特定语言中某一个句子的先验概率，在统计机器翻译解码过程中，语言模型也用来保证生成的翻译假设符合目标语言语言特性，因此引入目标语语言模型来对候选词集进行排序。

3.1 N 元语言模型

对于拼音音节序列 S = $s_1, s_2, s_3, \cdots, s_n$，音英转换的目的是找到对应的单词序列 W = $w_1, w_2, w_3, \cdots, w_n$，使生成的词序列满足：

$$w^* = \arg\max_{W} \mathrm{P}(W|S,C) \quad (1)$$

其中 C 是用户写作过程中的上下文信息，即历史单词序列 C = c1, c2, c3…cn。由于候选集根据词典得到，对应关系固定，因此可以直接计算在已有的单词序列的条件下当前词的最大概率 $^{\arg\max_w \mathrm{P}(W|C)}$，即通过英语语言模型方法来得到最优的候选词，当前输入没有历史单词序列时，则根据候选词的词频进行排序。语言模型的训练采用美国 SRI International 开发的语言模型工具软件 SRILM。

语言模型对语料的规模限制依赖性强，尤其是 n-gram 语言模型存在着数据稀疏问题。通过平滑的方法可以解决训练语料没有出现当前 n 元组合而无法处理的情况，以扩大语言模型的覆盖范围。我们采用的平滑方法是基于回退的方法，如公式（2）所示。其中 M 是三元语言模型，P_M 是根据 M 估计的的概率值，abc 是一个三元组，BO_M 是当前单元的回退值，$\mathrm{P}_M(\mathrm{unk})$ 是语料中未登录词的词频信息。

$$\begin{aligned} P(\mathrm{c}|ab) = & \\ & if \quad \mathrm{abc} \in \mathrm{M}: \mathrm{P}_M(\mathrm{c}|ab) \\ & \mathrm{elseif\ bc} \in \mathrm{M}: \mathrm{BO}_M(\mathrm{ab})\,\mathrm{P}_M(\mathrm{c}|b) \\ & \mathrm{elseif} \quad \mathrm{c} \in \mathrm{M}: \mathrm{BO}_M(\mathrm{ab})\mathrm{BO}_M(\mathrm{b})\,\mathrm{P}_M(\mathrm{c}) \\ & \mathrm{else} \qquad : \mathrm{BO}_M(\mathrm{ab})\mathrm{BO}_M(\mathrm{b})\,\mathrm{P}_M(\mathrm{unk}) \end{aligned} \quad (2)$$

3.2 词性语言模型

李鑫鑫[11]在音字转换中在词 N 元语言模型的基础上加入了词性 N 元语言模型，并取得了很好的音字转换效果。考虑到上下文信息，当前词的词性能够影响当前词的选择概率，例如名词+名词，形容词+名词等出现的概率大于形容词+形容词、名词+形容词等，因此考虑词性信息作为除目标与语言模型外的辅助消歧特征。与语言模型类似，我们建立词性的语言模型。采用词性赋码工具 TreeTagger 进行词性标注，定义词性序列 T = $t_1, t_2, t_3, \cdots, t_n$，根据词性语言模型估计的概率值为 P_{MT}。

3.3 线性加权模型

对目标语语言模型估计概率值和词性语言模型估计概率值进行对数线性加权得到当前候选词的可选概率如公式（3）所示，其中 w1 = 1- w2。

$$P_R = w_1 * \mathrm{P}_M + w_2 * P_{MT} \quad (3)$$

表 2 对数线性模型取值

W2 取值	可选候选词出现的位置			
	首位	前 2	前 3	前 5
0	0.478	0.543	0.696	0.826
1	0.413	0.522	0.674	0.804
0.1	0.5	0.565	0.676	0.804
0.2	0.522	0.609	0.717	0.870
0.3	0.493	0.565	0.674	0.826
0.4	0.52	0.587	0.717	0.848
0.5	0.494	0.58	0.705	0.826

表 3 测试语料示例

历史单词序列	拼音输入（预期输入词）	参考译文
comprehensive	zhili（治理）	management
Only through participation and	hezuo（合作）	cooperation
We should speed up	Zhanlve(战略)	strategy adjustment

本文在 Niutrans 中英测试语料中随机选取 500 句，并对英文句子进行随机截取，将截取位置对应的英文单词作为参考译文，将截取位置所对应的汉语词进行拼音标注，根据汉英词典获取所有的英文单词作为候选集，根据公式(3)获得的结果如表 2 所示。

4 候选词集的生成和排序

4．1 候选词生成

根据用户输入的拼音音节序，候选词集合的生成顺序如下：

（1） 选取谷歌拼音的前 5 个音字转换的结果，得到汉语候选集

（2）对音字转换的结果通过 pypinyin 进行拼音标注，获得输入拼音集。

（3） 根据建立好的音英短语机器翻译系统，获得输入拼音集的 1best 翻译结果加入到候选词集。当前拼音不存在翻译结果时，将当前拼音对应的汉语在词典中对应的英语词加入到候选词集。

4．2 候选词排序

利用词频、语言模型、词性语言模型特征对当前的候选词集进行排序。

5 实验

5.1 实验数据及配置

本文实验使用的数据为 Niutrans 翻译系统提供的 100,000 句汉英平行句对，并对汉语语料进行拼音标注，训练了基于短语的拼音-英语的统计机器翻译系统。测试语料为 Niutrans 系统提供的测试语料，随机选取 500 句英文句子并随机进行截取，截取位置的单词或者短语搭配作为音英转换的参考译文，测试语料的情况如表 3 所示。

5.2 实验结果与分析

用户在写作过程中往往会先通过查询词典得到候选词集，然后进行单词筛选，可以将这一过程看做音字转换，汉英翻译两级转换再进行消歧的过程。因此，本文将两级转换得到的候选词集通过语言模型的方法获得的候选集进行排序所获得的结果和融入基于短语的音-英统计机器翻译系统的方法获得的结果进行对比，将推荐正确率作为评价标准，推荐正确率为参考译文出现在候选词集中的位置，对比结果如表 4 所示。

表 4 实验结果对比

方法	两级转换	音英统计机器翻译系统
平均候选词集大小	7.64	4.7
首位推荐正确率	0.41	0.56
前 3 推荐正确率	0,67	0.72
前 5 推荐正确率	0.74	0.804

通过上表可知，经过音字，汉英翻译的两级转换带来的选择歧义可以通过统计的方法进行消歧。另外，建立音英统计机器翻译系统并进行消歧可以在提高推荐准确率的同时给用户减少选择数量，能够在用户的实时写作过程中带来帮助。

6 结论与展望

本文在分析用户在写作过程中往往采用母语思维习惯的基础上，面向以汉语为母语的用户提出了一种基于跨语言输入的英文辅助写作技术，融合了基于短语的统计机器翻译系统，实现了写作过程中遇到的难点词汇或短语的音英转换过程。实验结果表明，本文的方法在减少候选词集大小的基础上能够提高推荐准确率，在一定程度上能够减少用户的工具切换和单词的筛选，可以提高写作效率。因此，跨语言输入的方法来辅助用户进行非母语的写作简便快捷，有实用性。

从理论上讲本文的实验比较简单，对写作的辅助是按从前往后的顺序，后面的选择不能指导用户之前已做出的选择，另外该对候选词集的消歧采用的特征比较少。因此下一步考虑加入用户模型，即利用用户后续选择来指导已作出的选择，全局调整写作的质量。除了词频和语言模型特征外，我们还会进一步考虑句法或者其他消歧特征来提高推荐正确率。

参考文献

[1] 谭咏梅，王枞. 基于实例的机器辅助写作翻译系统[J]. Journal of Beijing University of Posts and Telecommunications，2006，29(Sup)：202-206..

[2] 占飞，刘挺. 面向英文辅助写作的词语相似度应用研究[J]. 智能计算机与应用，2011，1(1)：51-58.

[3] 代贤俊. 面向写作的辅助中文智能输入法系统[D]. 哈尔滨工业大学，2014.

[4] 尹宝生，蔡东风，张桂平. 面向中国人的英文输入助理软件的设计与实现[J]. 中文信息学报，1997，11(1)：23-29.

[5] 张桂平，尹宝生，刘伟松，等. 新一代 Windows 多文种处理平台—“环球使者”的设计与实现[J]. 中文信息学报，1998，12(2)：17-24.

[6] Dong Li. A pinyin input method editor with English-Chinese aided translation function[C]// In 2012 international Conference on Computer Science and Service System. Nanjing : IEEE Computer Society, 2012:446-449.

[7] 方瑞玉. 基于输入法的辅助翻译工具研究和实现[D]. 厦门大学,2013.

[8] Guoping Huang, Jiajun Zhang, Yu Zhou, et al. A New Input Method for Human 9 Translators: Integrating Machine Translation Effectively and Imperceptibly[C]// In Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI-15). Argentina:IJCAI:1163-1169.

[9] Maoxi Li, Jiajun Zhou, Yu Zong, et al. The CASIA Statistical Machine Translation System for IWSLT 2009[C]// International Workshop on Spoken Language Translation (IWSLT). Tokyo: 2009. IWSLT-2009, 83-90.

[10] Tong Xiao, Jingbo Zhu, Hao Zhang, et al. NiuTrans : An Open Source Toolkit for Phrase-based and Syntax-based Machine Translation[C]// In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, Korea. Pages 19-24.

[11] 李鑫鑫, 王轩. 基于级联重排序的汉语音字转换[J]. 自动化学报，2014，,40（4）：624-634.