基于主题模型和统计机器翻译方法的中文格律诗自动生成

蒋锐滢[1] 崔 磊[2] 何 晶[3] 周 明[4] 潘志庚[1),5)]

[1)](浙江大学 CAD&CG 国家重点实验室 杭州 310058)

[2)](哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

[3)](微软公司(美国) 雷德蒙德 98052)

[4)](微软亚洲研究院自然语言计算组 北京 100080)

[5)](杭州师范大学数字媒体与人机交互研究中心 杭州 310025)

摘 要 文中针对格律诗自动生成进行了研究.首先根据创作者提交的若干关键词,利用主题模型进行扩展得到更多的主题相关词,然后通过语言模型自动生成首句.在此基础上通过统计机器翻译的方法生成后续句.在生成过程中,利用主题模型进行诗词的意境扩展,从而得到更加丰富的句子候选.该研究的主要特点和贡献是:首先提出以统计机器翻译为理论基础,将格律诗的上下句关系映射为统计翻译模型中源语言与目标语言的关系,设计了融入诗词领域知识的统计机器翻译模型.其次主题模型用来在生成过程中进行词汇集扩展,从而加强了诗词的主题及意境.另外文中还论述了基于 BLEU 的诗句生成的自动评测方法,并配合所设计的人工评价标准,形成了比较完备的诗词评价体系.实验结果证实了该方法的有效性.

关键词 律诗生成;主题模型;统计机器翻译;自动评测

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2015.02426

Topic Model and Statistical Machine Translation Based Computer Assisted Poetry Generation

JIANG Rui-Ying[1] CUI Lei[2] HE Jing[3] ZHOU Ming[4] PAN Zhi-Geng[1),5)]

[1)](*State Key Lab of CAD&CG, Zhejiang University, Hangzhou* 310058)

[2)](*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

[3)](*Microsoft, Redmond* 98052, *USA*)

[4)](*Natural Language Computing Group, Microsoft Research Asia, Beijing* 100080)

[5)](*Digital Media & HCI Research Center, Hangzhou Normal University, Hangzhou* 310025)

Abstract This paper focuses on automatic ancient Chinese poetry generation. Topic model is leveraged to find semantic related words with the given keywords or key-phrases, and automatically generate the first sentence of the poetry by language model. Then statistical machine translation (SMT) model is used to give the followings step by step. Topic model expands the artistic conception of the poetry during generation, resulting in richer sentence candidates. The features and contributions of this study are as follows: (1) Based on SMT theory we consider two consecutive sentences in the poetry as the source-side and target-side sentences in SMT, under the rhythm and meter constraints of ancient Chinese poetry, we proposed a SMT model which learns poetry

收稿日期:2014-11-28;在线出版日期:2015-08-25.**蒋锐滢**,男,1987 年生,硕士,主要研究方向为自然语言处理、数据挖掘. E-mail: j.yearn@gmail.com. **崔 磊**,男,1986 年生,博士研究生,主要研究方向为自然语言处理、统计机器翻译. **何 晶**,女,1986 年生,博士,主要研究方向为近似计算、社交网络. **周 明**(通信作者),男,1964 年生,博士,首席研究员,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为自然语言处理、搜索和人工智能. E-mail: mingzhou@microsoft.com. **潘志庚**,男,1965 年生,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为虚拟现实、多媒体、计算机图形学、教育游戏.

creation knowledge from an ancient poetry corpus. (2) Topic model is leveraged to strengthen the artistic conception of the poetry by extending the keywords to a collection of semantic related words. (3) We also discuss automatic evaluation of poetry generation with BLEU metric, cooperating with our human evaluation standards, and having formed a comprehensive evaluation system for poetry generation. The experimental results show our method is quite promising for ancient Chinese poetry generation.

Keywords　poetry generation; topic model; statistical machine translation; automatic evaluation

1　引　言

格律诗是中国古典诗词的一种，特指唐代之后的古诗体. 根据篇章长度不同可分为绝句和律诗，按照单句诗字数的不同又可分为五言诗和七言诗. 格律诗的创作讲究文字优美、格律押韵，要求在规定字数内完成对于主题内容的描述，并且包涵特定意境与抒发情感. 但格律诗的遣词造句以及韵律等约束对于未经特别训练的文学爱好者而言，是学习创作过程中的一大障碍. 我们通过对大规模诗词数据进行统计机器学习，并将创作格律诗的知识融入到统计概率模型中，利用计算机进行格律诗的辅助创作，不仅为广大古典文化爱好者提供了帮助，而且对中国传统文化的传承及宣扬也具有积极意义. 除此之外，作为自然语言生成领域的研究，本文也提供了可参考的新思路.

本文针对格律诗自动生成进行了研究. 首先根据创作者提交的若干主题词，利用主题模型进行扩展得到更多的主题相关词，然后通过语言模型自动生成首句. 在此基础上通过统计机器翻译的方法生成后续句. 在生成过程中，利用主题模型进行诗词的意境扩展，从而得到更加丰富的句子候选. 本研究的主要特点和贡献是：首先提出以统计机器翻译为理论基础，将格律诗的上下句关系映射为统计翻译模型中源语言与目标语言的关系，考虑到诗词的音律押韵等约束，设计了融入诗词领域知识的统计机器翻译模型. 其次主题模型用来在生成过程中进行词汇集扩展，从而加强了诗词的主题及意境. 另外本文还论述了基于 BLEU 的诗句生成的自动评测方法，并配合所设计的人工评价标准，形成了比较完备的诗词评价体系. 实验结果证实了本方法的有效性.

本文第 2 节介绍相关工作；第 3 节介绍格律诗创作的背景以及辅助创作系统的整体构架；在第 4 节中介绍主题模型在本文中的作用和首句生成算法；第 5 节介绍融入诗学领域知识的统计机器翻译模型，包括模型的特征函数以及解码算法；在第 6 节，引入基于 BLEU[1] 的诗句生成自动评测方法以及对数线性模型的特征函数权重训练方法；第 7 节描述模型训练数据并阐述实验的设计评测、结果及分析；最后一节对本研究作总结.

2　相关工作

国外诗词生成研究始于 1959 年 Lutz[2] 用计算机生成的第 1 首德文诗. 诗歌生成的方法可被分类成基于模板的生成方法、生成并测试的方法、基于遗传算法的方法和基于实例推理的方法：

(1) 基于模板的生成方法. 给定一个模板，在满足语法、韵律等约束下进行填词作诗.

(2) 生成并测试的方法. 根据形式要求生成随机词序列，用相应约束以及评价准则判断该序列是否符合要求. Manurung 的 Chart 系统[3]、WASP 系统[4] 和 Tra-la-Lyrics[5] 都基于此方法.

(3) 基于遗传算法的方法. 结合遗传算法和评测模块，生成模块根据语法等信息用遗传算法生成备选诗作，评价模块则依据一定准则对备选输出进行评分. 代表系统有 POEVOLVE[6] 以及 McGonnagall[7].

(4) 基于实例推理的方法. 通过检索已有诗句，依据用户所要描述的目标信息对已有诗句作内容上的调整. ASPERA[8] 和 COLIBRI[9] 是此类方法系统的代表.

国内诗词生成研究始于 20 世纪 90 年代中期，迄今为止已积淀了不少前人工作. 如台湾罗凤珠的格律检查和同韵词查找系统①. 周昌乐等人[10] 在宋词生成上的研究，其方法是在给定词牌及韵律模板基础上，用遗传算法来进行宋词的自动生成. 在中文对联生成方面，微软亚洲研究院自然语言计算组研

① http://cls.hs.yzu.edu.tw/tang/PoemTone/index.asp

发的计算机自动对联系统[11]①，将统计机器翻译应用于下联的自动生成. He 等人[12]借鉴了上联生成下联的思想，将其扩展到了格律诗的自动生成上. Genzel 等人[13]也曾用统计机器翻译思想作有韵律约束的诗歌机器翻译.

本文采用基于统计机器翻译的格律诗生成方法，通过将格律诗的上下句关系看作统计机器翻译中源语句和目标句的对应关系，并对统计机器翻译模型进行合理特化，使其符合中文格律诗生成的特点. 通过从训练语料中学习诗词创作知识，从而在已有上句的情形下实现下句的自动生成. He 等人[12]的基于统计机器翻译的格律诗生成方法需借助一个人工语义分类辞典——《诗学含英》②，创作者必须在该辞典里选择关键词进行首句生成，这便造成了创作主题的限制. 而且《诗学含英》写于 17 世纪，许多词汇已不通用，容易生成难以理解的诗句. 另外统计机器翻译过程不考虑生成内容在主题表述上的一致性，所以并未将生成诗句与主题的相关度作为特征纳入考虑范围，生成结果虽然辞藻华丽，但读者往往很难从中体会到较为一致的主题或意境. 最后，因为缺少自动评测方法，统计机器翻译中用于解码的对数线性模型并未作参数调整，而是根据经验设定参数，这种方法未必能够搜索到模型的局部最优解.

本文是对 He 等人[12]工作的改进. 主要特点与贡献在于：(1) 提出以统计机器翻译为理论基础，将格律诗的上下句关系映射为统计翻译模型中源语言与目标语言的关系，设计了融入诗词领域知识的统计机器翻译模型；(2) 引入了主题模型作语义相关词集扩展，创作者能够自由输入关键词，使得创作主题更为自由和广泛. 利用主题模型衡量生成诗句与主题词之间的语义相关性，增强了生成诗句内容在主题表达上的一致性，从而增强其意境；(3) 讨论了基于 BLEU 的诗句生成自动评测方法，并结合最小错误率训练方法(Minimal Error Rate Training, MERT)[14]为对数线性模型作自动权重调整.

3 格律诗创作的背景及框架

一般的格律诗创作过程可大致分为如下几步：

(1) 有清晰的描述主题，这些主题可以通过具体的主题词来表述. 如写一首想念故乡的诗，可由“思乡”或“故乡”等确切的词汇来表述.

(2) 根据主题词联想出更多与其语义相关的词汇，如由“思乡”联想到“慈母”和“家书”等.

(3) 在符合格律诗约束下，通过对这些词的合理组织，斟酌着创作诗句.

(4) 在完成全诗过程中，参考上文与主题逐步创作下句，循环往复直至完成全诗.

相应地我们把格律诗辅助创作系统的具体架构设计如下(图 1)：

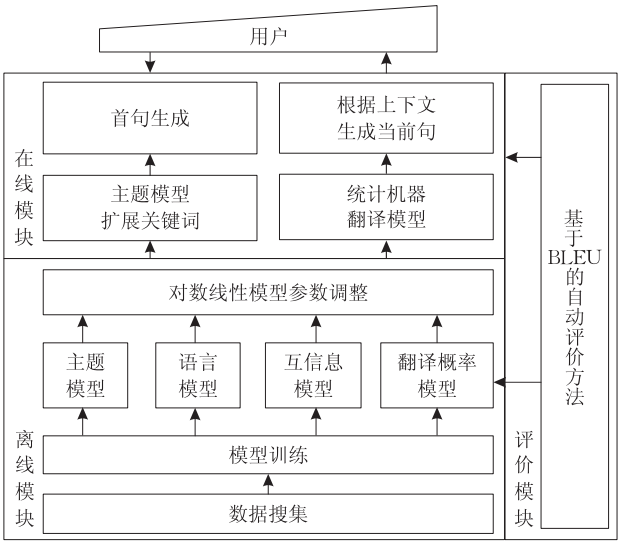

图 1 格律诗辅助创作系统架构

系统在统计机器翻译的基础上，引入主题模型作词汇集的扩展，辅助首句的生成和交互式单句生成，并且在统计机器翻译模型中融入主题模型加强了生成诗句与主题的相关性. 基于统计机器翻译的格律诗生成，将格律诗中上下句看作源语言与目标语言，利用已生成上句通过统计机器翻译模型实现下句的自动生成. 统计机器翻译把翻译过程看作一个最优目标语句的搜索过程，对于一个给定的上句，根据以往的翻译经验，生成多种可能的译文. 翻译经验包括词汇之间的翻译模式和源语言短语序列翻译为目标语言短语序列的模式等. 已有的研究表明基于短语的翻译模型在表现上要优于基于字的翻译模型)[15]，另外通过对格律诗的观察，我们发现上下句之间的对应关系往往建立于短语级别上，所以本文以短语作为翻译基本单位，采用了基于短语的统计机器翻译模型.

主题模型的应用体现在词汇集扩展以及主题意境的强化两方面. 本文所采用的主题模型是概率潜在语义分析(Probabilistic Latent Semantic Analysis,

① http://duilian.msra.cn

② 《诗学含英》，清代，刘文蔚著，提供使用者依类别检索，以作创作填词时应用典故之参考.

PLSA)[16]. PLSA 是基于双模式和共现数据分析方法延伸的经典统计学方法，被广泛应用于信息检索、自然语言处理和机器学习等领域. 所谓共现数据是指词和文档的共现关系矩阵，所谓双模式是指在词和文档上同时进行考虑. PLSA 基于混合矩阵分解得到词与文档之间的共现关系，考虑到词和文档之间的共现形式，PLSA 用多项式分布和条件分布的混合对共现矩阵进行建模. 通过将词和文档投影到由 K 个潜在主题决定的潜在语义空间，根据它们在空间上的主题分布向量，可以做语义相关度计算. 例如：给定主题词“春日”，根据它在潜在主题空间中的分布向量，可以找出“玉魄”、“红泥”和“燕”等空间距离比较近的语义相关词. 另外通过将生成诗句与给定主题词间的语义相关度作为特征加入统计机器翻译模型中，加强了生成诗句与主题的相关性，使生成的诗句紧扣主题并具备意境.

4　主题模型对格律诗辅助生成的作用

本节介绍了 PLSA 及其训练方法，并阐述语义相关词集的扩展方法，以及生成诗句与主题词之间的语义相关度衡量方法，同时给出了根据主题词生成首句的算法.

4.1　PLSA 及其训练

PLSA 是基于主题模型统计信息上的文档自动索引方法，能够处理面向领域的同义词或者多义词的词聚类及文档聚类等问题. 不同于直接由训练语料得到文档和词的共现关系，PLSA 借助未知变量潜在主题，通过分析它与文档、词之间的共现关系，间接构建了文档与词的共现关系. 该理论有两个独立性假设：观测数据中文档与词不具备依赖关系(bag-of-words)；给定主题后生成的词与文档是不相关的. 从生成模型的角度，可如下描述 PLSA：

(1) 以概率 $P(d)$ 从文档集中选取文档 d；

(2) 该文档以概率 $P(d|t)$ 描述主题 t；

(3) 描述主题 t 用到词 w 的概率 $P(w|t)$.

那么词与文档的关系最终可通过一个联合概率公式建立，如下：

$$\begin{aligned} P(d,w) &= P(d)P(w|d) \\ &= P(d)\sum_{t\in T}P(w|t)P(t|d) \\ &= \sum_{t\in T}P(w|t)P(d|t)P(t) \end{aligned} \tag{1}$$

用最大似然的思想，可以对 $P(w|t)$、$P(d|t)$ 和 $P(t)$ 作估测. 最大似然估计过程可以由 EM(Expectation Maximization)算法来实现，最终得到训练集上主题与词之间的共现矩阵、主题与文档之间的共现矩阵. 预先初始化 $P(t)=\frac{1}{K}$，$P(w|t)=\frac{1}{N}$ 和 $P(d|t)=\frac{1}{M}$(其中 K 是预定义的潜在主题个数，M 为训练数据中包含文档的数量，N 是词的个数)，随后用 EM 迭代优化估测值. EM 迭代步骤：根据当前预估参数计算观测到词和文档在各个主题上的概率分布；由词和文档在主题上的概率分布进一步优化对于模型参数的估测. EM 迭代公式如下：

E-Step：

$$P(t|d,w)=\frac{(P(t)P(d|t)P(w|t))^{\beta}}{\sum_{t'}(P(t')P(d|t')P(w|t'))^{\beta}} \tag{2}$$

M-Step：

$$P(w|t)=\frac{\sum_{d}n(d,w)P(t|d,w)}{\sum_{d,w'}n(d,w')P(t|d,w')} \tag{3}$$

$$P(d|t)=\frac{\sum_{w}n(d,w)P(t|d,w)}{\sum_{d',w}n(d',w)P(t|d',w)} \tag{4}$$

$$P(t)=\frac{1}{R}\sum_{d,w}n(d,w)P(t|d,w) \tag{5}$$

其中 $R\equiv\sum_{d,w}n(d,w)$，$n(d,w)$ 指词 w 在文档 d 中出现次数. 本文采用 TEM 算法(Tempered EM)，较之于常规 EM 算法，通过对估测式(2)加入系数 β 可有效预防参数过拟合. 训练结束后可得到文档-主题共现矩阵和词-主题共现矩阵. 在共现矩阵中，每个文档或词可以由一个维度为潜在主题个数 K 的主题分布向量所描述. 利用词或文档的对应主题分布向量，结合相似度衡量方法，可在向量空间模型中作词聚类或文档聚类等操作. 例如：有 6 个文档 $d_1\sim d_6$ 和 2 个潜在主题 t_1、t_2，PLSA 训练后的文档在潜在主题空间中的表示如图 2，若以欧氏距离计算语义相关度，则 d_1 与 d_2 的语义相关度大于 d_1 与 d_6 的语义相关度. 由图可推断 d_1、d_2 和 d_3 描述的主题内容比较相关，而它们与 d_4、d_5 和 d_6 描述的主题内容则会有比较大差异.

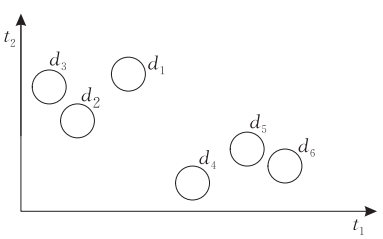

图 2　文档在二维潜在语义空间中的示例

4.2 基于主题模型的词汇扩展

在格律诗创作中,可以由多种多样的词汇来描述某一给定主题.为了得到丰富的目标诗的描绘词集,我们通过选择语义相关的词对给出关键词作扩展(语义相关指词 A 和词 B 同属某个主题的常用词,在人为撰写某类主题文章时经常被混用,且在训练文档库中存在大量共现关系).经实验对比,相较于一般的基于词上下文的词聚类方法,基于 PLSA 的聚类方法更为鲁棒,通过将高维的词之间的关系映射到了低维的词与主题关系上,原本缺少共现信息而未建立起语义相关的词,也可以由 PLSA 的分析表示而建立起语义关系;再者,PLSA 训练过程中预设的潜在主题个数 K 的值等于共现矩阵中词用潜在主题表示后的维度大小,它的取值影响词聚类结果.通过调整潜在主题的个数 K,可以控制词之间语义关联概念强弱,当词的潜在主题分布向量维度越高,那么它们在空间分布上更为分散,对应语义相关性的概念则更严苛,当维度越低则空间上分布更为密集,语义相关的概念就相对粗犷.

本文利用 PLSA 得到词与主题的共现矩阵,之后根据每个词对应的主题分布向量衡量词之间的语义相关度.表 1 为不同 K 所对应给定词条的语义相关词集,K 为潜在主题个数,从中可以看出随着 K 增大,扩展词汇便越具象地贴近给定主题词.从表达多样性的需求及计算代价均衡考虑,本文取 K 为 50.

表 1 语义相关词表示例

K\词条	秋水	远山	流水
$K=10$	销鬓,歌功,宏声,浪涌,筝音,殿,仕,兆,腔	征鞍,为隐,鸟巢,每向,阻断,衿,烤,店,匿	中州,风水,半酣,上林,入川,眉,舟,怅,乃
$K=50$	惊影,一襟,园亭,独照,深秋,萧,锁,逐,处	群山,展堂,秋景,晚鸦,流烟,枫,浦,浮,碧	曲破,清溪,明月,楼头,分付,系,随,觅,去
$K=100$	凭栏,曾照,玉笛,月照,箫声,悄,冷,碎,瘦	轻云,几片,小径,野水,雨过,笼,噪,朦,蓬	浮萍,逐水,恋恋,紫陌,落处,过,觅,随,似

主题词扩展方法.对给出的每个主题词,从语义相关词库中选取与该主题词语义关联度大于 0.6 的词加入扩展词集(关联度的值域为[0,1]).例如给出 2 个主题词 A 和 B,分别将 A 对应的语义相关词与同 B 语义相关词去重后取并集加入扩展词集.

4.3 生成诗句与主题词间语义相关度衡量方法

为衡量新生成诗句与给定主题词之间的语义相关度,本文参照主题模型应用于信息检索中的原理来计算两者的语义相关度.其方法为:将主题词看作一个特殊短文档,通过 PLSA 模型训练过程中的 EM 迭代过程将其映射到潜在主题空间中,得到对应的主题分布向量并记为 $P(t|q)$.对新生成诗句做同样操作记作 $P(t|d)$,由余弦相似度便可以计算得到主题词与生成诗句之间的语义相关度.

将新文档映射到潜在主题空间的方法为:将 4.1 节 EM 算法中 $P(t)$、$P(w|t)$ 固定,迭代更新式(2)、(4)中 $P(t|d,w)$ 和 $P(d|t)$,直至收敛或迭代次数满.表 2 为根据词"征战"检索得到的语义相关诗句.

表 2 根据主题词检索语义相关诗句示例

主题词	语义相关诗句
征战	转战疆场屡建功,勋章授予大英雄. 倭侵国土燃眉睫,驶抵江阴御日戎.
	玄武门耀明光甲,铁戟双弓战突厥. 将军三箭定天山,战士长歌入汉关.
	战火硝烟漫,倭戎劲敌攻. 团城形势急,戍守怒张弓.

通过语义相似度计算,根据关键词语义检索出的诗句内容紧扣主题且具备对应意境.反向思考,若在生成过程中加强对候选诗句的语义相关度衡量,则能使生成内容更紧扣主题并体现意境.为此,在统计机器翻译模型中本文加入了主题相关度这一特征.

4.4 依照主题词生成首句的算法

首句生成借助主题模型对给定主题词进行扩展,得到语义相关词集,以此词集中词汇作为造句的基本元素,结合语言模型作为评价函数搜索较优候选句.首句生成采用的柱状搜索算法是在宽度优先搜索的基础上加上扩展宽度限制,每次只从扩展节点中选择评分最优的 N 个节点进行扩展,评分函数为二元语言模型.算法 1 描述了首句生成的方法,记候选句集为 $Cand$,词集为 $WordCollection$,当前假设节点为 $hypo$,扩展节点最小堆为 $extendHypoHeap$,假设队列为 $hypoQueue$,扩展宽度为 N,目标诗句长度为 len,所需候选数为 M.几个主要子函数功能以及伪代码描述如表 3 所示.

表 3 子函数功能的描述

子函数	功能描述
$Ex_{TEND}(hypo,w)$	将词 w 连接到假设节点句尾扩展
$S_{ATISFY}R_{ULE}(hypo)$	从格律、长度和唯一性三方面判断假设节点是否能加入候选集
$S_{CORE}(hypo)$	假设节点的语言模型评分

算法 1.　首句生成算法.

输入：主题扩展词集 *WordColletion*，首句候选集 *Cand*（空），句长 *len*，扩展宽度 *N*，所需候选数 *M*

输出：首句候选集 *Cand*

```
1.  hypo=EMPTYHYPO //初始化假设节点
2.  E_NQUEUE(hypoQueue,hypo)
3.  REPEAT
4.    hypo=D_EQUEUE(hypoQueue) //取队首假设节点
5.    IF S_ATISFYR_ULE(hypo) THEN
6.      A_DDC_ANDIDATE(Cand)
7.      CONTINUE
8.    END IF
9.  C_LEAR(extendHypoHeap) //清空扩展假设节点队列
10.   FOR w∈WordColletion DO
11.     IF NOT w∈hypo.Sentence THEN
12.       tHypo=E_XTEND(hypo,w)
13.       IF L_ENGTH(tHypo.Sentence)<=len THEN
14.         I_NSERT H_EAP(extendHypoHeap,tHypo,
                          S_CORE(tHypo))
15.       END IF
16.     END IF
17.   END FOR
18.   FOR i←1,…,N DO //取前 N 最优节点插入假设队列
19.     tHypo=P_OP H_EAP(extendHypoHeap)
20.     E_NQUEUE(hypoQueue,tHypo)
21.   END FOR
22. UNTIL C_OUNT(Cand)>=M or I_SE_MPTY(hypoQueue)
```

第 5～8 行代码将满足条件的候选加入候选集中并且不再对其作扩展，子函数 $S_{ATISFY}R_{ULE}(hypo)$ 包含 3 方面约束：格律约束要求假设节点诗句仅包含一个单字，且单字只能出现在倒数第一或第二的位置上；长度约束要求假设节点诗句的句长应等于 *len*；唯一性约束要求假设节点诗句未在候选集中出现过. 第 10～17 行功能是对假设节点做扩展，枚举 *WordCollection* 中的词 w，若 w 未出现于假设节点诗句内，则用其扩展 *hypo* 并插入扩展节点最小堆 *extendHypoHeap*. 第 18～21 行代码从 *extendHypoHeap* 中选取最优的 N 个扩展节点加入假设队列 *hypoQueue*.

为方便直观理解，举例如下：有扩展词集 *WordCollection* 为{清溪、石桥、潺潺、绕、过}，扩展宽度 N 为 1，生成候选个数 M 为 1，诗句长度 *len* 为 5. 第 1 次循环，从空串开始枚举词集得到 5 个扩展节点（每个词为一个候选），假定首轮最优扩展节点为“清溪”；进入第 2 次循环，枚举词集得到“清溪-石桥”、“清溪-潺潺”、“清溪-绕”、“清溪-过”4 个扩展节点，假定最优扩展节点为“清溪-潺潺”；进入第 3 次循环，根据长度约束得到两个扩展节点“清溪-潺潺-绕”和“清溪-潺潺-过”，假定最优扩展节点为“清溪-潺潺-过”；进入第 4 次循环，判断“清溪-潺潺-过”满足 3 个约束，将其加入候选集 *Cand*，顺序执行，在外层循环判断语句内 $C_{OUNT}(Cand)$ 大于等于 1，结束生成过程.

扩展步骤每次从假设队列中选取最优的前 K 个候选节点进行扩展，所以搜索空间相当于一棵 K 叉树，算法的时间复杂度为 $O(k^l)$，l 为组句所需要词的平均个数，在格律诗生成中最不理想情况下 l 为 7. 通过对效率与生成候选质量的权衡，本文设定 K 为 15. *WordCollection* 由交互给出的主题词作语义相关词集扩展得到，扩展方法见 4.2 小节.

5　基于统计机器翻译的二、三、四句生成模型

近年来，统计机器翻译的研究发展很快，这得益于训练数据的增加以及不同翻译模型的出现. 总体来讲，目前主流的统计机器翻译方法分为以下几类模型：串到串（string-to-string）、树到串（tree-to-string）、串到树（string-to-tree）以及树到树（tree-to-tree）. 不同的模型各有特点，同时也存在各自的优势与不足. Jiang 等人曾将统计机器翻译的算法应用于中文对联的下联生成[11]，He 等人[12]曾将统计机器翻译应用于格律诗生成. 在本文中，当生成首句之后，采用基于统计机器翻译的方法进行二、三、四句的生成，并加入了新的特征，期望达到诗歌全篇用词意境一致的效果.

基于短语的统计机器翻译技术（Phrase-Based Statistical Machine Translation，PBSMT）是目前一种主流的机器翻译技术，它的优势在于短语翻译结果的选词准确. 由于诗词的生成讲求对仗，不涉及远距离语序调整问题，因此，诗词的生成非常适合采用基于短语的机器翻译算法来解决. 目前，主流的统计机器翻译算法都是基于最大熵框架提出的，即给定源语言句子 f，通过计算和统计不同的翻译特征，利用最大熵模型计算生成目标语言句子 e 的概率，之后按概率排序，并选择概率最大的翻译候选作为结果，形式化描述如下：

$$\hat{e}=\arg\max_{e}\{P(e|f)\}$$
$$=\arg\max_{e}\left\{\frac{\exp\left[\sum_{m=1}^{M}\lambda_m h_m(e,f)\right]}{\sum_{e'}\exp\left[\sum_{m=1}^{M}\lambda_m h_m(e',f)\right]}\right\}$$
$$=\arg\max_{e}\left\{\sum_{m=1}^{M}\lambda_m h_m(e,f)\right\} \quad (6)$$

其中,$h_m(\cdot)$代表最大熵模型中的特征函数,λ_m代表相应的特征权重,M 代表特征函数的数量. 由于只期望获得概率最大的翻译候选,对具体概率值并不关心,因此去掉上面公式中的分母部分,等式依然成立.

在统计机器翻译中,模型的选择固然重要,但是也不能忽略特征函数的选择,因为这直接决定了模型能否生成正确的翻译结果. 一般来讲,基于短语的统计机器翻译系统采用了如下的特征函数:正向短语翻译概率、反向短语翻译概率、正向短语词汇化翻译概率、反向短语词汇化翻译概率、使用短语个数和目标语言模型. 除去上述基本特征外,本文还加入了两个新的特征:互信息特征以及主题模型对于词汇和诗句相关的特征,其中,互信息特征已经在中文对联生成中被使用过,实验结果已经表明其有效性;主题模型特征是第一次被使用在诗词生成领域中,具体如表 4 所示.

表 4 诗歌生成模型的特征函数

特征函数	描述
$h_1(e,f)=\prod_{\forall i,j_i\ \text{s.t.}\ i+1=j} MI(e_i,e_j)$	生成诗句中连续两个词的互信息
$h_2(e,f)=\prod_{i=1}^{I} P_{topic}(e_i\|t)$	生成诗句与给定主题的相关度

新加入的两个特征可以进一步帮助诗歌生成模型进行词汇的优选,期望不同诗句能够达到更佳的主题意境一致性. $P_{topic}(e_i|t)$是词 e_i 与主题 t 之间的语义相关度,t 是将一个或多个主题词作为短文档投射到主题空间后的主题分布向量,其中 l 为句内词的个数.

分词策略参考了文献[17]的方法. 首先随机初始化词表及词频信息,利用词频对格律诗语料进行切分;之后,在切分后语料上优化对词频信息的估测,并过滤掉词频低于设定阈值的词汇(本文取 1.0E−8);重复迭代直至词频收敛,收敛后的词表及词频信息被用作后台分词模型. 出于格律诗单句长度最大为 7 的特点考虑,本文采用基于一元文法概率的分词策略,由分词模型对输入序列的各种切分计算概率值,选择概率最大者作为最终切分序列.

解码算法则参考了 Koehn 等人[15]的方法,与此同时为了满足押韵的约束,在解码第 4 句时,根据第 2 句诗句的最后一个字,删除不符合韵律约束的候选词,另外为保证有足够多的候选词,本文加入了满足韵律约束的词. 解码算法将上述特征引入对数线性模型中,选定性能评价方法之后,利用最小错误率训练算法[14]进行参数训练,之后模型利用训练好的参数对给定诗句进行下句的生成,关于对数线性模型的参数训练具体细节将在下一节介绍. 需要说明的是,本文对一首诗中不同位置的上下句分别构建模型,即分别训练一二、二三和三四句翻译模型.

6 基于 BLEU 的评测方法

自动评测方法在参数预估和系统调整上是非常重要的. 一个自动评测方法往往需要一个标准答案集和一种衡量生成结果同答案集相近程度的指标. 本文中由上句生成下句的方法以统计机器翻译原理为基础,所以用于机器翻译系统自动评测的 BLEU(Papineni 等人,2002)是一种自然的选择. 此外,Jiang 等人[11]在基于统计机器翻译的对联生成研究中即用到了 BLEU 为生成系统作评测,并验证了 BLEU 作为基于统计机器翻译的对联生成系统评测的可靠性. 目前在诗词生成领域尚无公认有效的自动评价方法,而格律诗中诗句间往往两两对仗,可看作是特殊形式的对联,所以本文也试探性地采用 BLEU 作为格律诗自动生成的评测指标.

BLEU 的直观思想是翻译结果越接近参考答案则翻译质量越好. 相应的,本文认为如果根据给定上句生成的下句能够更贴近已有的参考下句则系统的生成质量越好,但由于诗词在内容表现上丰富多样,所以需要搜集拥有多个参考下句的数据样本加入答案集.

6.1 数据集准备

出于诗句内容表现上的多样性,建立答案集时本文选取了拥有较多参考下句的句子作为答案集中的样例. 首先,从几个格律诗论坛抓取了格律诗数据(例如诗词在线、天涯论坛的诗词比兴等),之后根据不同论坛的网页结构特征抽取发帖内容,保留满足格律诗形式的文本. 由于论坛上往往存在对诗句的推敲与讨论,很容易找到拥有多个不同下句的诗句,我们将这些句对提取放入答案集. 最终,数据集经人工筛选,包含了 200 个数据样本,平均每个数据样本

中上句包含对应 24.5 个下句，其中最多包含 40 句参考下句，最少包含 20 句参考下句.

6.2 评价准则

BLEU 通过对翻译候选句与源语句的参考句进行 1 元词到 N 元词的重合度统计，结合式(13)衡量翻译结果的好坏. BLEU 的计算公式如下：

$$p_n=\frac{\sum_{c\in\{candidates\}}\sum_{n\text{-gram}}Count_{clip}(n\text{-gram})}{\sum_{c'\in\{candidates\}}\sum_{n\text{-gram}'}Count(n\text{-gram}')} \tag{7}$$

$$BLEU=BP\cdot\exp\left(\sum_{n=1}^{N}\frac{1}{N}\log p_n\right) \tag{8}$$

其中 p_n 为 n 元词的重合度. 在一般的翻译系统中，BLEU 需要考虑翻译句与候选句的句子长度惩罚 BP 作最终 BLEU 计算的权重调整. 但在格律诗生成中因为参考句和候选句长度相同，所以不需要考虑这一惩罚因子，即 $BP\equiv1$. 另外在一般的统计机器翻译中取 N 为 4，但由于格律诗单句长度较短，一个字便能描述一个对象，所以本文取 N 为 2 计算 BLEU 值.

6.3 BLEU 对比人工评测

BLEU 已经被验证能够可靠地应用于对联生成系统的评测. 但格律诗生成同对联生成仍然存在着一定程度的区别，所以用人工评测对比 BLEU 的评测用以验证 BLEU 评测在格律诗生成系统评测的可靠性. 分别建立了 5 个不同的系统. 系统 S_1 根据 1/50 的格律诗语料库数据训练得到，系统 S_2 根据 1/10 的格律诗语料库数据训练得到，系统 S_3 则包含了所有的训练数据训练得到. S_1、S_2 和 S_3 的格律诗系统的翻译模型的对数线性模型只包含语言模型、正向翻译模型和反向翻译模型. S_4 从整个格律诗语料库训练得到并且融入了词位置信息模型，S_5 则在 S_4 的基础上加入了互信息模型来优化生成候选句.

由于格律诗内容丰富多样，根据同一上句可以生成多种风格及内容的下句. 据此在实验部分，从测试数据集中挑选了参考下句个数在 20～40 的句子. 最后一共得到了 120 个(包括 60 句七言诗句和 60 句五言诗句)测试样例，平均每句对应 28.1 句参考下句. 之后用 5 个系统分别为每句诗句生成最优的下句候选，再用人工评测方法和 BLEU 评测为每个系统进行评分.

人工评测的打分由 2 位评测者以标准为从 1 分(极差)、2 分(一般)到 3 分(优秀)对生成诗句评分，最终取算数平均. 计算得到两种评测方法的评价结果相关性系数为 0.99，证实了用 BLEU 作格律诗生成的评测是切实可行的. 图 3 是对 BLEU 评价结果与人工评价结果作线性回归的图示.

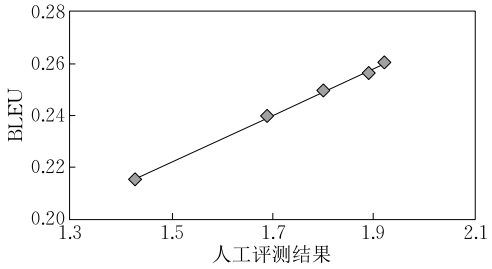

图 3　BLEU 与人工评测结果线性回归图示

6.4 对数线性模型的参数训练

Och[14] 于 2002 年提出了一种新的基于最大熵模型的统计机器翻译训练方法，称为最小错误率训练方法(Minimum Error Rate Training，MERT). 这种方法的最大优势在于，通过任意给定的评价标准，最小错误率训练方法的训练目标是降低给定评价标准下的错误率，使得参数训练向着提高评价标准的方向进行，这样就可以在庞大的搜索空间中进一步提高搜索的准确性. Och 的实验结果表明，基于最小错误率训练方法会显著好于之前最大似然估计的训练方法，同时，也能够显著地提高统计机器翻译系统的性能.

近年来，最小错误率训练方法已广泛应用于统计机器翻译中：通过以对翻译系统的自动评价指标为优化目标，有效调整对数线性模型的参数. 式(6)所给出的对数线性模型一共包含了 8 个不同特征，不同特征函数对生成结果质量好坏的影响各异，所以应给各个特征函数赋予不同权重. 本文利用上述 BLEU 评测标准，结合最小错误率参数训练方法[14] 为对数线性模型作自动权重调整.

7 实　验

本节介绍了各统计模型的训练数据，实验设计以及结果分析. 用人工评价方法验证了主题模型生成的语义相关词库替代《诗学含英》的可行性，并对格律诗生成人工评价和自动评价的结果作了分析，最后列举了几首由系统生成的格律诗示例.

7.1 统计模型的训练数据

格律诗训练语料来自互联网，其中包括《唐诗》、《全唐诗》、《全台词》和《毛泽东诗词选》等文献，以及从各大诗词论坛(例如诗词在线、天涯论坛诗词比兴等)抓取并筛选后的格律诗，总计 287 000 多首. 语料库中每首诗的一二句、二三句和三四句分别训练

对应翻译模型,包括正反向的短语以及词汇化的翻译模型.所有单句诗句用于训练语言模型和单句内互信息模型.一首诗中的不同句用于训练句间互信息模型.整个语料库中的格律诗用于训练主题模型.

7.2 实验设计及结果分析

7.2.1 主题模型替代《诗学含英》的可行性

为了验证主题模型可以替代《诗学含英》这一人工分类的辅助创作词典.本文从《诗学含英》中选择了15组关键词,分别由《诗学含英》和主体模型扩展词集,根据同样算法生成首句.由6位评测者对生成结果作评价,如《诗学含英》好于主题模型则得1分;否则为0分.最终取算术平均,评测结果如表5所示.

表5 词汇扩展方法比较

词集扩展方法	《诗学含英》	主题模型
人工评分均值	0.4625	0.5375

由表5可以看出,主题模型替代《诗学含英》作词汇集的扩展是可行的.不但破除了关键词选择上的约束,而且从评测结果看,根据主体模型作词汇扩展的单句生成的结果优于《诗学含英》选词生成的结果.

7.2.2 格律诗生成的人工评测

考虑到目前对于机器艺术作品质量的评测主要通过图灵测验性方式进行,在用BLEU评测的同时,本文也采用专家组人工评价的方法.评价标准主要参考了古代一些学者对于诗词评价准则的见解,参考书籍有司徒空的《二十四诗品》、欧阳修《六一诗话》和姜夔的《白石道人诗说》.虽然这些名家对于诗词好坏的评价标准都各有己见,但总体而言需要满足如下的这些要求:语言流畅、韵律优美、主题鲜明、内容表述一致以及有意境.据此对应地建立了人工评测的5个标准:语言流畅度、韵律符合度、主题相关度、内容表达一致性和意境.

本文将He等人[12]的格律诗生成系统作为基准系统,用此基准系统来对比加入了主体模型后的格律诗生成系统.为了得到客观的评测结果,实验部分所用的格律诗并非由人工交互生成,而采取了自动生成的方法:先由主题生成首句,之后自动下句生成,直至全诗完成.评测数据包括20首根据不同主题词生成的格律诗,五言诗和七言诗各10首,分别从两个系统各自生成的50首五言诗和50首七言诗中随机抽取得到.单句评测是对抽取出的20首诗中的第1句作评价,上下句生成评测则是从这20首诗句随机选取一二、二三或三四句句对作评价.由6位具备诗词领域知识的评价者根据评测标准进行打分,按1分(极差)到5分(优秀)分别对单句生成、上下句生成和整首诗3个不同方面打分.实验结果见表6,表中单行上方数据为基准系统的评价结果,下方数据为新系统的评分.

表6 人工评测结果

评测维度	单句	上下句	整首诗
语言流畅度	4.062 4.275	3.575 4.004	3.525 3.912
韵律符合度	— —	3.796 4.096	3.658 3.992
主题相关度	3.942 4.250	3.663 3.825	3.408 3.896
内容一致性	3.983 4.225	3.600 3.788	3.433 3.667
意境	3.921 3.996	3.563 3.863	3.417 3.883

从单句的生成的评测结果中可以看出,加入了主题相关度后的系统,在主题相关表现上有很大的提高.语言流畅度上也因内容表达一致性的增强较于基准系统有很大的提升.随着上下句之间的信息量的增多,评价数据在各个数据上都比单句生成要有所下降,但从最终的人工评测结果上看,系统的输出是可接受的.整首诗的评价同样由于内容的增加使得在一致性等方面有所降低.总体而言,新系统是优于基准系统的.

7.2.3 格律诗生成的自动评测

在第6节验证BLEU有效性部分,针对格律诗内容上的多样性,本文选择了拥有多于20个参考下句的句子作为测试集的数据,所以数据集规模较小,仅包含120个数据样本.而为了验证加入主题模型后的生成系统确实好于基准系统,本文选择了共1500个数据样本,每句参考句个数在10～40之间,平均每句拥有14.2个参考句子.在更大规模数据集上得到的BLEU值要小于图5中的最高BLEU值,究其因,一方面验证BLEU可靠性实验部分每句的参考句更多,另一方面则是由测试数据集的大小不同所致.

表7展示了两个系统中不同句对翻译模型的BLEU分数.从表7的数据可以看出,无论哪个系统一二句和三四句的BLEU分数都高于二三句,这与诗词中的对应翻译关系是契合的,在格律诗中往往是一二与三四句之间更为对仗.加入了主题模型后的生成系统得到了平均0.0212的BLEU值提升.证实了主题模型的引入是有效的.

表 7　不同句对翻译模型的 BLEU 值

BLEU	一二句	二三句	三四句	平均值
基准系统	0.2177	0.2051	0.2167	0.2132
新系统	0.2385	0.2266	0.2380	0.2344

从新系统的评测结果看，系统本身已经能够得到较好的生成结果，在韵律以及语言流畅度上表现得都比较出色.然而内容表达一致性上仍旧是主要的缺陷，究其因则是由于在生成过程中未考虑谋篇布局，换言之句和句或词和词之间的逻辑关系虽然通过了互信息模型有所加强，但在整体上布局仍旧有欠缺.将两个系统对比地看，新系统较之于基准系统有较大的提升.随着主题模型的加入，一方面针对给定主题词作语义相关词汇集扩展，使得诗本身在内容的描述上能够更好贴近所欲描述的主题，另方面将主题相关度特征引入统计机器翻译翻译模型，使得生成结果在主题表现以及意境上都要好于基础模型.虽然在内容上的连贯一致性和意境方面仍旧需要得到加强，但系统已经能够生成比较优秀的单句以及不错的自动生成的下句，这已经达到了本研究的目的.未来将进一步去克服表现得不够好的方面，优化生成结果.

7.2.4　生成格律诗示例

表 8 列举了两首人工根据关键词交互生成的诗，要求仅从返回的前十句中选择诗句，并且不对诗句内容作修改.

表 8　生成结果示例

主题词	基准系统	新系统
松 、远山	浩浩丛林远山衔 悠悠一水秋风入 楼阁半空碧千寻 草堂人日红万急	群山远宿苍松绕 一水秋风白月来 千门晓日青天去 万里飞花落地开
春径 、石桥、流水	一曲阳关空 千山雨有无 万水风何在 白云流水句	清溪潺潺过 高山流水行 一地落花去 三江春月尽

8　总　结

目前国内外的诗歌生成研究较多采用的是直接基于模板的生成方法.而本文独创性地提出了运用结合主题模型与统计机器翻译模型的方法来进行格律诗的自动生成，将格律诗中上下句关系映射为统计机器翻译中的双语关系，并深入讨论了这一模型下的相关理论及技术难题，实现了这一原理下的格律诗自动生成.最后，制定了严格的标准对单句生成、依上文生成当前句和自动生成全诗进行了人工评测.从评测结果看，本研究取得了较好的成果，对格律诗自动生成和自然语言生成都具有一定的参考价值.未来工作需要更多地考虑如何来谋篇布局，使得一首诗的内容在连贯性上和一致性等逻辑层面上能够有更好的表现.另外，诗词的风格分类以及风格化的诗词生成也是值得研究的课题，即如何根据不同作者的写作风格，用机器模拟他们的思维进行诗词创作.

参　考　文　献

[1] Papineni K, Roukos S, Ward T, Zhu Wei-Jing. BLEU: A method for automatic evaluation of machine translation// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia, USA, 2002: 311-318

[2] Lutz T. Stochastische texte. Augenblick, 1959, 4(1): 3-9

[3] Manurung H. Chart generator for rhythm patterned text// Proceedings of the 1st International Workshop on Literature in Cognition and Computer. Tokyo, Japan, 1999: 15-19

[4] Gervás P. WASP: Evaluation of different strategies for the automatic generation of spanish verse//Proceedings of the AISB00 Symposium on Creative & Cultural Aspects and Applications of AI & Cognitive Science. Birmingham, UK, 2000: 93-100

[5] Oliveira H G, Cardoso F A, Pereira F C. Tra-la-Lyrics: An approach to generate text based on rhythm//Proceedings of the 4th International Joint Workshop on Computational Creativity. London, UK, 2007: 47-55

[6] Kempe V, Levy R, Graci C. Neural networks as fitness evaluators in genetic algorithms: Simulating human creativity //Proceedings of the 23rd Annual Conference of the Cognitive Science Society. Edinburgh, Scotland, 2001: 1221

[7] Manurung H. An Evolutionary Algorithm Approach to Poetry Generation[Ph. D. dissertation]. University of Edinburgh, Edinburgh, UK, 2003

[8] Gervás P. An expert system for the composition of formal Spanish poetry. Journal of Knowledge-Based Systems, 2001, 14(3-4): 181-188

[9] Díaz-Agudo B, Gervás P, González-Calero P A. Poetry generation in colibri//Proceedings of the 6th European Conference on Advances in Case-Based Reasoning. London, UK, 2002: 73-102

[10] Zhou Chang-Le, You Wei, Ding Xiao-Jun. Genetic algorithm and its implementation of automatic generation of Chinese SONGCI. Journal of Software, 2010, 21(3): 427-437 (in Chinese)
(周昌乐，游维，丁晓君. 一种宋词自动生成的遗传算法及其机器实现. 软件学报，2010，21(3)：427-437)

[11] Jiang Long, Zhou Ming. Generating Chinese couplets using a statistical MT approach//Proceedings of the 22nd International Conference on Computational Linguistics. Manchester, England, 2008: 377-384

[12] He Jing, Zhou Ming, Jiang Long. Generating Chinese classical poems with statistical machine translation models//Proceedings of the 26th AAAI Conference on Artificial Intelligence. Toronto, Canada, 2012: 1650-1656

[13] Genzel D, Uszkoreit J, Och F. "Poetic" statistical machine translation: Rhyme and meter//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Massachusetts, USA, 2010: 158-166

[14] Och F J. Minimum error rate training in statistical machine translation//Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003). Sapporo, Japan, 2003: 160-167

[15] Koehn P, Och F J, Marcu D. Statistical phrase-based translation//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT/NAACL). Edmonton, Canada, 2003: 48-54

[16] Hofmann T. Probabilistic latent semantic indexing//Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Boston, USA, 1999: 50-57

[17] Ge X, Pratt W, Smyth P. Discovering Chinese words from unsegmented text (poster abstract)//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley, USA, 1999: 271-272

JIANG Rui-Ying, born in 1987, M. S. His main research interests include natural language processing, data mining.

CUI Lei, born in 1986, Ph. D. candidate. His main research interests include natural language processing, statistical machine translation.

HE Jing, born in 1986, Ph. D. Her main research interests include approximate calculation, social networks.

ZHOU Ming, born in 1964, Ph. D., principal researcher, Ph. D. supervisor. His main research interests include natural language processing, search and artificial intelligence.

PAN Zhi-Geng, born in 1965, professor, Ph. D. supervisor. His main research interests include virtual reality, multimedia, computer graphics and entertainment education.

Background

Computer assisted poetry generation not only helps fans of Chinese ancient poetry in writing their own poetries, but also helps in carrying forward traditional Chinese culture. Currently, poetry generation approaches include template-based generation, case-based reasoning, generate and test approach, and genetic algorithm based generation. This paper provides a novel statistical way. This research is a part of couplet and poetry generation project led by Dr. Ming Zhou, who is the manager of Natural Language Computing group in MSRA. By learning knowledge from poetries and couplets, statistical machine learning methods like SMT and text generation technology are used to automatically generate poetries. In addition, we also discuss the automatic evaluation metric of poetry generation.