中国诗的宏观分析

词汇发现、主题演变与骈偶建模

钱鹏　汉语言
复旦大学中国语言文学系

1 引言

中国诗歌区别于其他文学形式的最重要的特点，就在于其丰富的形象、多变的主题以及特殊的修辞结构。诗歌主题、体裁的多样性发展经历了文学观念、文化语境等历史因素的变化。诗歌意象的模式与聚合，新意象的创造与突围，也代表了一批诗人在探索诗歌语言形式和诗歌意象道路上，竭力突破已有范式的反常创造。我们通过考察大量诗歌文本，以唐代诗歌作为切入点，纵横数百年的发展历史，试图用计算和定量的研究范式，发现意象聚合的历史事实，勾勒出诗歌主题的历史波动，建模中国诗歌语言注重骈偶的修辞特征，验证文学史研究者提出的题材演化与诗歌变迁的理论。

2 相关工作

利用自然语言技术处理文学文本是计算语言学领域近年来兴起的热门话题。以往的自然语言处理研究关注较为正式的语言文本的处理，而后有很多的研究开始转向文学性语言的计算分析。但是，绝大多数工作都是集中于计量风格学领域的研究，如作者鉴定、文本分类、风格分析等。近年来出现了从计算分析的角度关注文学理论的研究，如 Klein（2010）关注了如何从小说文本中自动抽取出相关的人物社会关系网络。Huges et al（2012）对英语文学 1800 年之后两百多年的历史作了风格学分析，考察了文学作品的影响与时代风格的关系，从英语文学样本中发现较早时间段的作家对邻近时期的作家影响比较大，但较晚阶段的影响较小。Jurafsky（2013）对中国当代诗歌进行了分析，认为内地当代诗人的诗歌呈现出较高的远离传统诗歌的特征，但同时发现台湾诗人的创作中似乎保留了一些古典诗歌的风格。Jurafsky 还对英文诗歌做了分析，试图从计算语言学的角度，找出优秀诗歌区别于一般诗歌的特质。Klingenstein（2014）对伦敦大法院的庭审记录进行主题建模，从不同法律主题的历史变化来考察英国的民主化进程。

诗歌在中国是一个极为重要且历史悠久的文学体裁。纵横千年的时间跨度、数量巨大的诗人群体、卷帙浩繁的诗歌文本都使得在传统的文本阅读方式下，中国诗歌的宏观把握变得极其不易。Jurafsky（2013）虽然也关注了中国现代诗歌与古典诗歌的不同，但我们认为其使用的数据量较小，提供的分析手段与评判指标仍较为粗糙。因此，我们希望利用现有的自然语言处理技术，对具有丰富语言学限制的中国传统诗歌文本进行建模。试图从更一个视角出发，更高效地处理中国文学大数据，尝试诗歌文本生成的任务，以此提供计算语言学视角下对中国诗的刻画、描述与阐释。

3 浅层分析

3.1 词语发现

汉语文本的考察必然要牵涉到字与词的问题。一方面，从形式上能够完全确定的是字的边界，但在语言学意义上，词是真正能够独立运用的最小单位。另一方面，对于类似传统诗歌等古汉语文本，尚且缺乏直接可用的词表。因而，从以字为单位的文本中发现潜在的词语，是一项很重要

的工作。

词语发现主要是基于二元组合点态互信息（Point-wise Mutual Information）的搭配发现算法。只不过，在原始版本的搭配发现算法中，待发现的是词组搭配，构成搭配的是词。而在这里，待发现的是词语，其基本的形式化单位是字。因而，我们通过更改相关变量的定义，直接将此算法迁移至由字组词的过程中，以发现具有特殊含义的诗歌词汇。此外，值得说明的是，汉语具有强烈的双音化倾向，大多数词是双音节形式。因此，在词语发现的过程中，我们暂且只考虑二元模式。

定义两个汉字 C_1、C_2 结合成的二元模式 C_1C_2 能够被当成一个词的条件为：

$$\text{IsWord}(C_1C_2) = \begin{cases} True, & PMI(C_1, C_2) > \theta \\ False, & PMI(C_1, C_2) < \theta \end{cases}$$

其中：

$$\text{PMI}(C_1, C_2) = \log \frac{P(C_1, C_2)}{P(C_1)P(C_2)} \propto \log \frac{\#(C_1, C_2)}{\#(C_1)\#(C_2)}$$

关于阈值 θ 的确定，我们观察了二元模式点互信息量的分布。从图 1 的累积分布图可以大致看出信息量的分布情况。高信息量的词汇较少，而信息量中等的二元组居多。同时，曲线大约在信息量上升到-13 时出现了走势上的剧烈变化。因此，我们选取-13 作为信息量阈值，以作为汉字二元组成词和非成词的判断条件。

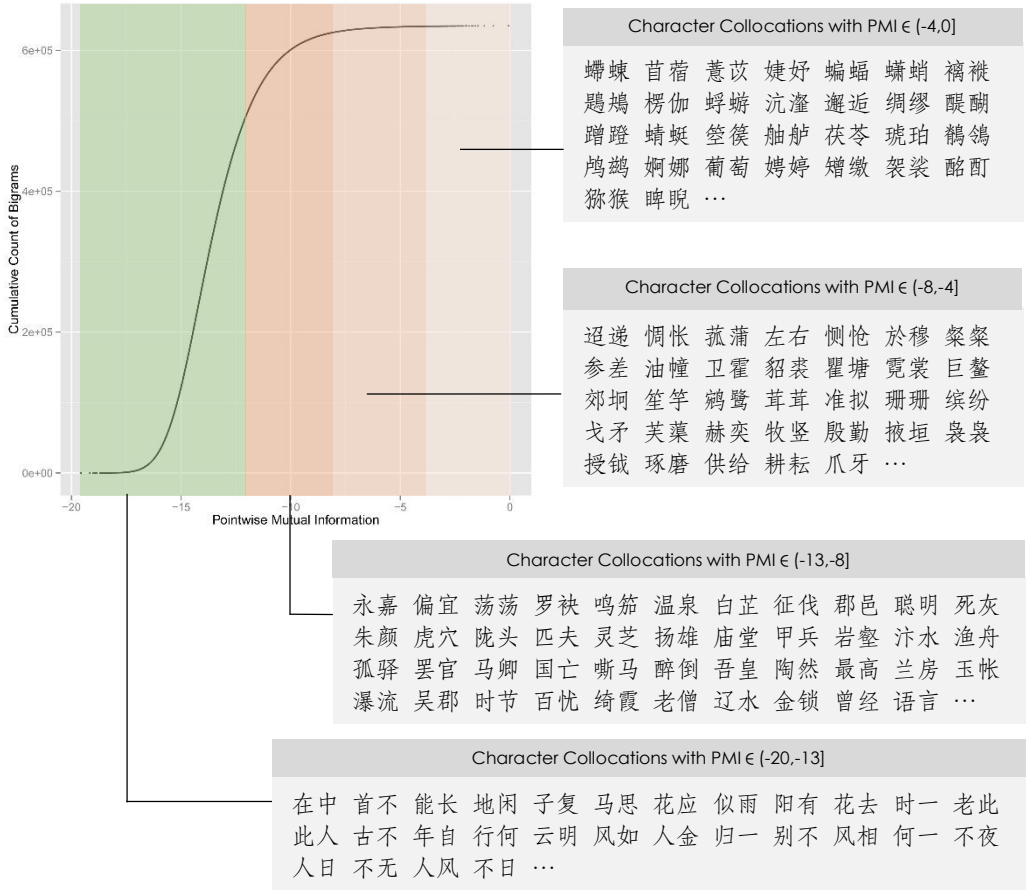

图 1

同时，为了过滤掉无意义的高信息量低频组合，我们也设置了频率限制，要求每一个二元组

至少要出现两次。基于以上条件，我们得到了较为理想的词表。同时，我们也发现，找出的词汇，特别是信息量高的词汇，确实具有特殊的意义。词汇发现不仅在涉及古代汉语文本处理的工程实践中起到辅助作用，对于文学研究者来说亦能够提供帮助，对于希望了解古代文化的人也能提供良有裨益的参考。试举一些 PMI 值较高的词的释义，如下表 1。

词语	解释	词语	解释
蝃蝀	彩虹的别名，一般比喻才气横溢	罘罳	古代的一种屏风
苜蓿	即三叶草	褭褭	羽毛初生的样子
婕妤	为嫔妃的等级之一。	沆瀣	“沆瀣”二字都指夜间的雾气或露水，故有成语“沆瀣一气”
蠨蛸	蜘蛛的一种，脚长。通称蟢子		
胙蚡	散布，传播。	舵艋	形如蚱蜢的小船。《广雅·释水》:“舵艋，舟也。”
旖旎	柔和美丽		
桔槔	古代的一种原始汲水工具。	醍醐	从酥酪中提制出的油。
蟪蛄	蝉，《庄子》有“蟪蛄不知春秋”	薏苡	即薏仁米
彷徨	连绵词，表示来回走，犹豫不决	睢盱	浑朴貌；睁眼仰视的样子
鸑鷟	古代民间传说中的五凤之一，身为黑色或紫色	舳舻	指首尾衔接的船只。舳：指船尾；舻：指船头
葡萄	从西域传来的一种水果	娉婷	形容女子姿态美好

表 1

3.2 探索性分析

3.2.1 基于词语发现的诗歌分词

处理古代汉语文本的常见方式，是将文本全部打散为独立的汉字。这一处理方式立足于一个基本观念，即孤立的汉字也承担了较为充足的语义信息，大多仍是具有较强构词能力的自由语素。但是，我们认为，到了唐代，汉语的双音化已经发展到一定的高度，汉语中产生了一系列使用频繁的双音节词汇，应以词为单位进行切分。

然而，诗歌的特点决定了现有的现代汉语分词技术难以在这一特殊领域获得良好的适应性。因此，基于以上总结的诗歌文本的特点，我们采用一个启发式的分词方法，基于已获得的某些固定古典诗词词汇，结合诗歌文本的格律特征和古代汉语的语素-音节对应关系，采用正向最大匹配法，对诗歌文本作简单的切分。

经过这一分词处理之后，得到的分词结果基本符合我们的预期。由此，得到的诗歌分词结果如下：

篇目	作者	分词后诗歌文本
送杜少府之任蜀川	王勃	城阙 辅 三秦 风 烟 望 五 津 与君 离别 意 同 是 宦游 人 海内 存 知己 天涯 若比 邻 无 为 在 岐路 儿女 共 沾巾
春夜洛城闻笛	李白	谁家 玉笛 暗 飞 声 散 入 春风 满 洛城 此 夜 曲 中 闻 折柳 何人 不 起 故园 情
子夜吴歌	李白	镜湖 三百 里 菡萏 发 荷花 五月 西施 采 人 看 隘 若耶 回舟 不待 月 归去 越王 家
使至塞上	王维	单车 欲问 边 属国 过 居延 征 蓬 出 汉 塞 归 雁 入 胡天 大漠 孤烟 直 长河 落日 圆 萧关 逢 候吏 都护 在 燕然

大林寺桃花	白居易	人间 四 月 芳菲 尽 山寺 桃花 始 盛 开 长恨 春 归 无 觅 处 不知 转 入 此中 来
寄扬州韩绰判官	杜牧	青山 隐隐 水 迢迢 秋 尽 江南 草木 凋 二 十 四 桥 明月 夜 玉 人 何处 教 吹箫
端居	李商隐	远 书 归梦 两 悠悠 只有 空床 敌 素 秋 阶下 青苔 与 红树 雨 中 寥落 月 中 愁

表 2

3.2.2 主题演变

在分词之后的文本上，我们又运用最简单的主题模型（Topic Model）对诗歌进行文学史层面的宏观分析。以往的诗歌分析都基于简单的词频统计，难以有整体把握。本文采用生成式模型，对诗歌中的主题分布可以进行更好的整体建模。

主题模型是 Blei（2003）等人提出的一种被广泛应用于文本主题建模的概率图模型。主题模型以词袋假设作为前提条件，引入狄利克雷先验分布，在文档主题聚类等任务上均取得了不错的效果。

本文用主题模型对数万首分词后的诗歌进行建模。我们从互联网上获取到了《全唐诗》文本格式语料。《全唐诗》是清代曹寅、彭定求等人奉敕编纂，共有 900 卷，收录唐代诗人 2873 人的 49403 首诗歌作品及 1555 条全文无考的诗句。经预处理后，得到 42700 首不含错误字形的干净文本。选取主题数 K 为 20，α= 50/ K，$\beta = 0.01$，采用吉布斯采样算法，试图发现每首诗歌中涉猎的主题以及相应的分布。

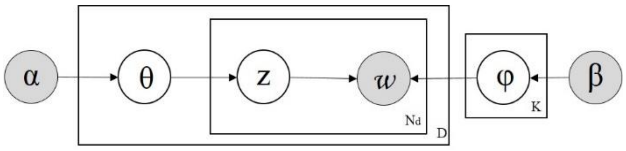

图 2 主题模型

值得说明的是，《全唐诗》编者基本遵照作者生活年代的次序对诗歌进行整理编排。诗歌在全唐诗语料库中的位置也可以在一定程度上反映诗歌所处的时代。因此，我们将训练后得到的每篇诗歌的$\vec{\theta}$值，按位置次序，每 100 篇进行一次区间归并，得到 427 个时间点（Time Tick），并绘制出$\vec{\theta}$的每个主题分量在 Time Tick 上的分布以及对应的平滑曲线。此外，本文认为一首诗歌的某一主题分量如果过低，则实际上并不能准确地体现该主题。所以，在绘图的过程中，只绘制主题分量大于一定阈值的数据点，以消除噪声点，使得主题在时间轴上的变化趋势更为明显。经多次尝试，我们设定阈值为 0.07（接近主题分量的期望值 1/K）。

图 3 选取了 2 幅有代表性的散点图，反映的是某一主题分量在时间轴上的分布。我们发现，主题随时间的变化确实与文学史研究中已有的结论相符合。

主题 9 是颂德主题，从散点图和平滑曲线的走势可以看到，唐初出现了一个极大的峰值。这与明代文学批评家胡震亨《唐音癸签》中的叙述非常一致：“有唐吟业之盛，导源有自。文皇英姿间出，表丽缛于先程；玄宗材艺兼该，通风婉于时格。是用古体再变，律调一新；朝野景从，谣习浸广。**重以德、宣诸主，天藻并工，赓歌时继。**上好下甚，风偃化移，固宜于喁遍于群伦，爽籁袭于异代矣。”当是时，国运昌盛，不少台阁文人参与到诗歌的创作中来，加上君王好写诗，自然会突出“海宇颂皇仁”的主题。

主题 18 是秾艳主题，图中可以看到，平滑曲线在晚唐末期有一个突然的上升。这一结果和文学史研究的结果完全对应。初唐时期，陈子昂引领“风雅兴寄”的运动，诗文相对来说摒弃秾艳之风。而晚唐的思潮有所变化，朝政的暮年之感使得诗人的关注转向个体主体性的方面（章培

恒, 2011）。因而，在创作上更多地表现个体的情感，突出凄艳的风格，孕育着词的产生。而“词为艳科”，表现女性情感的作品在那一时间段更为集中，因而晚唐时期秾艳主题格外突出。

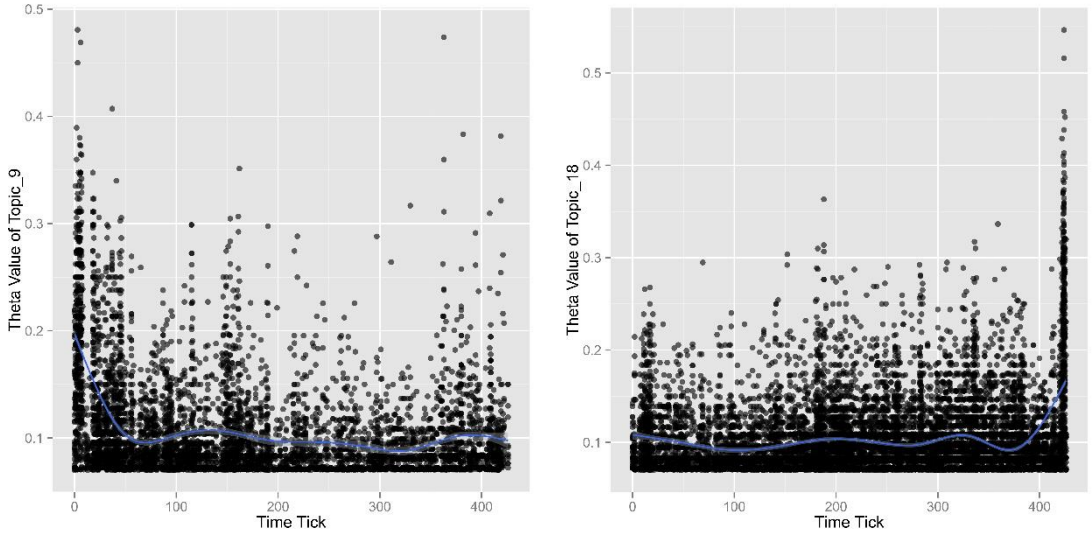

图 3 主题-时间分布图

3.2.3 诗人群体分析

我们试图通过诗歌主题分布来比较不同诗人之间的相似度。对每一个诗人，将其所写过的诗歌的$\vec{\theta}$值相加求算术平均，这样对于每一个诗人都可以计算得到相应的$\overrightarrow{\theta_A}$。定义两个诗人 A_1，A_2 之间的距离为：

$$\text{Distance}(A_1, A_2) = \sqrt{\sum_{k=0}^{K} (\theta_{A_1}^k - \theta_{A_2}^k)^2}$$

给定一个诗人 A_1，可以通过对 A_1 与其他诗人的距离进行排序，发现最相似的诗人群，构建一个相似诗人群体检索系统。

A_1	A_2	$\text{Distance}(A_1, A_2)$	A_1	A_2	$\text{Distance}(A_1, A_2)$
李白	李颀	0.0678778943473	温庭筠	韩琮	0.0669742430259
	顾况	0.072011592489		刘兼	0.0719581826716
	陈陶	0.0753518831373		成彦雄	0.0725016359302
	欧阳詹	0.0753757591583		薛涛	0.0727166627275
	王昌龄	0.0770582682817		李绅	0.0775427131027
白居易	元稹	0.0669908654741	王之涣	翁绶	0.114263962083
	王建	0.0670158822905		朱琳	0.117622887866
	王绩	0.0683919537222		李约	0.121581528616
	张籍	0.0723872708102		郑锡	0.138013917526
	薛能	0.0771696651509		纪唐夫	0.139967641938

表 3 相似诗人检索结果

表 3 是该系统给出的部分结果，A_1列是待查询的诗人，A_2列是按照距离递增原则排序后的前 5 位最接近的诗人。这一结果与文学批评研究结果也比较接近。例如，与李白最相似的是李颀

与顾况。这种相似不仅有文本上的直觉作为依据，从诗人的交游历史和诗学追求上亦可佐证（章培恒, 2011; Luo, 2011）。李颀与李白都崇奉道教，李颀深受李白诗风的影响。顾况的诗歌则具有很强的多样性。以往的研究过分注重顾况与新乐府的关系，但事实上，顾况的新乐府作品仅占其诗作总量不到十分之一的比例，最新的文学史著作开始注意到他与李白相近的风格特征。这些恰恰与我们的系统结果相符。元稹和白居易在文学史上历来被合称为“元白诗派”。检索结果也显示，与白居易最相似的是元稹。

除此之外，我们发现，系统给出的结果亦能对传统的文学史研究作出很好的补充。例如，对诗人王之涣进行检索，可以看到最相似的诗人是翁绶。盛唐和晚唐诗人之间的相近，或许帮助研究者更好地理解诗风的流变和复归。

4 骈偶结构的建模研究

上文采用的主题模型毕竟太过简单，我们认为它忽略了文档中潜在的重要的修辞结构信息。中国诗歌显著的区别性特征之一就是高频率的骈偶对仗。对仗，或称骈偶，是一种特殊的修辞手法。骈偶关系一般发生在两个相邻的句子之间，上下句字数相等，对应位置的字词性相当，意义相近或相反，以通过字意的某种关联或对立产生出文本的张力与美感。作为诗歌的重要修辞特征，如何在现有的主题模型中嵌入骈偶因素，表达出这一普遍存在的文本结构信息，对于自然语言处理中的诗歌生成任务能够提供有价值的参考意义。而对于当今的中国古代诗歌爱好者、中国文学批评研究者而言，也能够通过骈偶在诗歌作品中的分布，进而揭示出文学形式的内在演化规律。

4.1 骈偶的形式化抽象

试以温庭筠的《更漏子》与李白的《登金陵凤凰台》为例，阐释概率骈偶模型的现实基础与数学抽象的动机。图 4 中，我们将诗歌文本按字拆分，置于语篇格律的坐标系中。字与字之间的连线，表示两字所处的坐标位置决定了它们形成了潜在的骈偶关系。灰色填充意味着，事实上观察到潜在的位置骈偶关系对该字的产生结果并无影响，或者观察到该字根本没有对字。白色填充意味着，事实上已观察到该字的产生过程确实受到了骈偶对字的影响。例如，温庭筠的《更漏子》是晚唐的一首词。由字面观察可以看到，第一句和第二句、第四句和第五句的确构成了典型的对仗，“柳丝长”对“春雨细”的产生有着重要的影响。而在李白的《登金陵凤凰台》中，可以看到第一句和第二句并不对仗。但倘若拆成字对来看，则坐标（1, 1）的字“凤”和坐标（2, 1）的字“凤”、（1, 3）的“台”和（2, 3）的“台”、（1, 7）的“游”和（2, 7）的“流”，似乎仍有一定的骈偶意味。

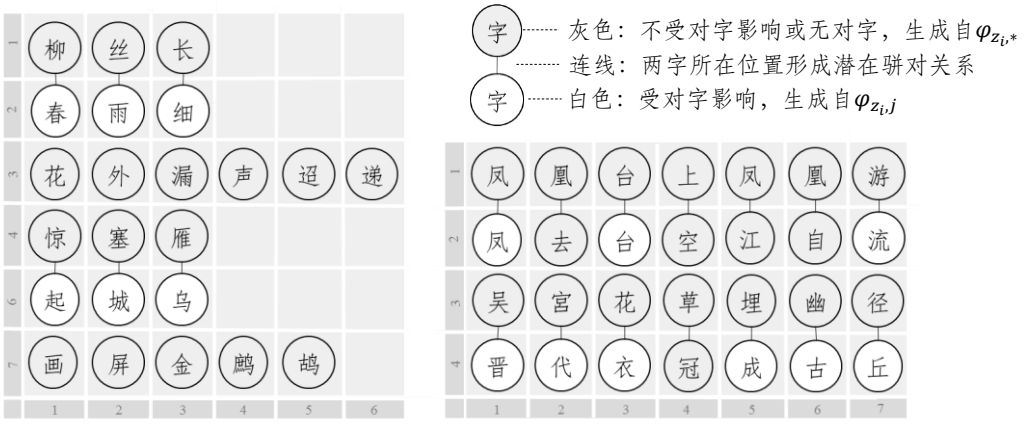

图 4 诗歌骈偶关系示例

由此，我们认为，一首诗歌的文本形式，可以抽象为图 4 所示的语篇格律坐标系。它决定了

其中某些位置上的字会形成骈偶字对。一般来说，每一韵句之内上下相邻的两个小句，若字数相同，就会按对应位置构成潜在的骈偶字对。对于潜在骈偶字对的骈偶下字位置，这些字并不严格与相对位置的上字对仗，其产生过程除了受到全局主题分布的影响之外，还会按一定的概率η满足骈偶上字的对仗要求；而骈偶字对中的上字以及这些字对之外的另一些字则是相对孤立的，其产生过程绝对没有局部骈偶信息的介入，而仅仅受到整首诗歌主题分布的影响。

为此，我们在普通的主题模型（LDA）之外，提出了概率骈偶主题模型（Probabilistic Couplet LDA）。其概率图模型的表示如下：

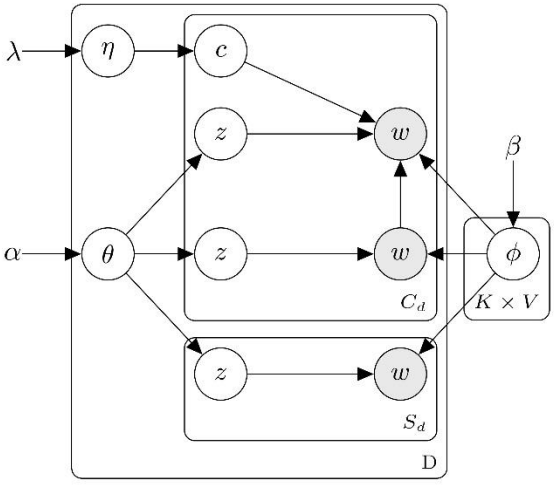

图 5 概率骈偶主题模型

4.2 生成过程与参数学习

因为骈偶讲究字数相当、一一对应，所以在本节的模型中，为简化模型输入，不再以前文发现的诗歌词汇作为基本单位，而是代之以独立的汉字。虽然这一理想化假设与实际情况不尽相同，如双声叠韵词是两字成词相对而不是单字相对，但我们认为对参数的学习来说或无大碍。基于以上的模型假设，概率骈偶主题模型的生成过程可以描绘为：

Draw word distribution (conditioned on $*$ and possible coupling word) $\varphi \sim Dirichlet(\beta)$

For each poem $m \in$ [1:D]

Get the register parameters C_m as the number of potential coupling words pairs, S_m as the number of single words

Draw topic distribution $\theta_m \sim Dirichlet(\alpha)$

Draw $\eta_m \sim$Beta(λ)

For each word in single positions, $i \in$ [1:(S_m+ C_m)] **do**

Draw topic $z_{m,i} \sim Multinomial(\theta_m)$

Draw word $w_{m,i} \sim Multinomial(\varphi_{z_i,\#})$

End for

For each word in coupling positions, $i \in$ [1: C_m] **do**

Draw coupling indicator $c_i \sim Bernoulli(\eta_m)$

If c_i = 0 **then**

Draw topic $z_{m,i} \sim Multinomial(\theta_m)$

Draw word $w_{m,i} \sim Multinomial(\varphi_{z_i,*})$

End if

If c_i = 1 **then**

Draw topic $z_{m,i}$~$Multinomial(\theta_m)$

Get the already generated coupling upper word $j = couple\_map(m,i)$

Draw word $w_{m,i}$~$Multinomial(\varphi_{z_i,j})$

End if

End for

End for

记号	释义
D	诗歌文档总数
η_m	第 m 篇诗歌中产生骈偶字对的概率
c_i	骈偶指示变量，$c_i = 1$表示两个字具有对偶关系，$c_i = 0$表示两个字没有对偶关系
θ_m	第 m 篇诗歌中的主题分布
φ	主题、骈偶上字在字表上的联合分布
C_m	第 m 篇诗歌中骈偶字对的数量
S_m	第 m 篇诗歌中独立汉字的数量
$z_{m,i}$	第 m 篇诗歌中第 i 个字的主题
$couple\_map(m,i)$	第 m 篇诗歌中与第 i 个字对应的骈偶上字

表 4

对于参数的学习，我们采用吉布斯采样算法。通过对一系列的公式推导，对隐变量积分，我们得到如下的采样概率。

如果w_i处在可能的骈偶位置，则同时对主题z_i和骈偶指示变量c_i进行采样。采样中使用的条件概率$\mathrm{p}(z_i = \mathrm{k}, c_i = \mathrm{c} \mid \overrightarrow{z_{\neg i}}, \overrightarrow{c_{\neg i}}, \vec{w}, \vec{j})$：

$$\mathrm{p}(z_i = \mathrm{k}, c_i = \mathrm{c} \mid \overrightarrow{z_{\neg i}}, \overrightarrow{c_{\neg i}}, \vec{w}, \vec{j}) = \frac{\mathrm{p}(\vec{z},\ \vec{c},\ \vec{w})}{\mathrm{p}(\overrightarrow{z_{\neg i}}, \overrightarrow{c_{\neg i}}, \vec{w})} \propto \frac{n_{k,j,\neg i}^{(t)} + \beta_t}{n_{k,j,\neg i}^{(\cdot)} + V \cdot \beta_t} \cdot \frac{n_{m,\neg i}^{(t)} + \alpha_k}{n_{m,\neg i}^{(\cdot)} + K \cdot \alpha_k} \cdot \frac{n_{m,\neg i}^{(c)} + \lambda_c}{n_{m,\neg i}^{(\cdot)} + \lambda_0 + \lambda_1}$$

如果w_i处在没有骈偶对应关系的位置，则只对主题z_i进行采样。采样中使用的条件概率$\mathrm{p}(z_i = \mathrm{k} \mid \overrightarrow{z_{\neg i}}, \vec{c}, \vec{w}, \vec{j})$如下：

$$\mathrm{p}(z_i = \mathrm{k} \mid \overrightarrow{z_{\neg i}}, \vec{c}, \vec{w}, \vec{j}) = \frac{\mathrm{p}(\vec{z},\ \vec{c},\ \vec{w})}{\mathrm{p}(\overrightarrow{z_{\neg i}}, \vec{c}, \vec{w})} \propto \frac{n_{k,j,\neg i}^{(t)} + \beta_t}{n_{k,j,\neg i}^{(\cdot)} + V \cdot \beta_t} \cdot \frac{n_{m,\neg i}^{(t)} + \alpha_k}{n_{m,\neg i}^{(\cdot)} + K \cdot \alpha_k}$$

由此我们可以进一步得到词表$\Phi(\varphi_{k,j,t})$、文档主题分布$\Theta(\theta_{m,k})$以及文档骈偶度$\mathrm{H}(\eta_m)$等参数的更新法则，如下：

$$\varphi_{k,j,t} = \frac{n_{k,j}^{(t)} + \beta_t}{n_{k,j}^{(\cdot)} + V \cdot \beta_t}$$

$$\theta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{n_m^{(\cdot)} + K \cdot \alpha_k}$$

$$\eta_m = \frac{n_m^{(1)} + \lambda_1}{n_m^{(\cdot)} + \lambda_0 + \lambda_1}$$

4.3 实验结果

我们同样采用了《全唐诗》作为训练语料。采用吉布斯采样算法对诗歌文本的主题隐变量进行学习，设定主题数 K 为 20，待采样过程基本稳定收敛后，输出得到的主题。首先，我们可以观察一下基本主题，也就是没有骈偶上字的主题分布。表 5 按顺序罗列出主题及按概率递减排列的主题词，并人工标注出主题名称：

主题序号及对应的主题词			主题名称
topic 0	couple * ：	不一生有天大时如百来中公骨人得为死行士当	酬志抒愤
topic 1	couple * ：	别路去归远行山离故客望乡里处思何尽愁云国	羁旅送别
topic 2	couple * ：	月夜秋风寒明声空落清寂霜思梦云夕影暗孤惊	秋寒孤寂
topic 3	couple * ：	天仙云龙三神气九灵海丹道真中金帝千化阳玄	游仙方术
topic 4	couple * ：	春风花处柳日雪草何树色尽水芳上年怅杨落新	春景
topic 5	couple * ：	里一千年君日长万三阳人去东西十上来朝南前	叙事
topic 6	couple * ：	不我有何为之子者所君人言兮可与吾此以心如	古体
topic 7	couple * ：	应无得诗事是闲知多后曾到吟来莫只经难身名	闲适闲愁
topic 8	couple * ：	江水南波山流楚舟岸海客云湖远浪吴鱼雨浦烟	山水
topic 9	couple * ：	德圣方乐明帝王命礼大文皇臣至惟贤国神代既	颂德
topic 10	couple * ：	马将城河汉边北天塞山西关黄军下骑出剑国胡	边塞
topic 11	couple * ：	雨叶树风竹色露阴晚清林池野烟绿静景疏秋草	郊野
topic 12	couple * ：	朝文恩重从书同主分传才官阙方当节府入笔御	政治讽谏
topic 13	couple * ：	家年不为人老无少生子自多问儿事一田身苦未	田园民生
topic 14	couple * ：	山石云松泉林峰寺溪幽下隐寻僧高鹤鸟入岩野	僧寺归隐
topic 15	couple * ：	酒日相客来醉白逢时今还长后下未年发饮对地	饮酒寻欢
topic 16	couple * ：	不人见如此知时君一有何心无来上相中在可今	对话
topic 17	couple * ：	玉金歌宫曲飞女翠光双舞珠香王碧凤楼轻转殿	宫闱艳词
topic 18	couple * ：	花春红不香似小桃语愁深情相看欲画恨怜新垂	爱情春思
topic 19	couple * ：	无已在高事犹多自旧时知久心怀方终期虽游岁	登高怀古

表 5

可以看出，这些主题确实较为不错地聚合了意义相关的汉字，也基本符合我们对中国诗歌的认知。此外，我们还考察了概率骈偶模型找出的受骈偶上字影响的字词分布。如表 6 所示。

	topic 2	couple 天	：	日月烟春江佳别芳风云鲜是残故归野又离解泪
，	topic 10	couple 汉	：	胡秦周云梁天尧牛晋一汉未河荆不鲁乘今王蕃
	topic 0	couple 白	：	红青黄绿朱碧翠金紫香锦银丝拂楼罗飞绮狂芙
	topic 0	couple 雪	：	丝珠香花烟金风罗锦楼红春冰银琼碧霞粉霓露
	topic 0	couple 轻	：	细薄红嫩暖重小艳浓掩软纤明断澹腻含新飘双

表 6

此外我们还发现，概率骈偶主题模型不仅找某一骈偶上字最可能的骈偶下字，其找出来的骈

偶下字的分布也与对应的主题相关，这也是一般的频率计数方法和普通的主题模型所不能取得的效果。以“殿”这个字为例，在不同的主题下，筛选出的概率较大的字基本与对应的无骈偶上字主题的语义非常相关。例如，主题三的关键字大多是关于游仙诗一类。相应地，在主题二下发现的、最有可能与“殿”形成骈偶字对的字，也大多是神仙色彩的字词，如“宫”、“龙”、“莱”（应当是“蓬莱”的后字）等等。主题十是边塞战事一类。同样，在主题十下发现的、最有可能与“殿”形成骈偶字对的字，也偏重于塞外事物相关的地点处所类字词，如“城”、“营”、“河”等。具体可见表 7。

主题	无潜在骈偶上字	主题	骈偶上字为“殿”的词表（按概率大小取字排序）
2	月夜秋风寒明声空落清寂霜思梦云夕影暗孤惊	2	庭楼灯光空窗声明规秋寒前阶裴深晖蝉云凉
3	天仙云龙三神气九灵海丹道真中金帝千化阳玄	3	龙宫天坛莱台明云虚壶微元芝仑甍氤精扶廓
8	江水南波山流楚舟岸海客云湖远浪吴鱼雨浦烟	8	亭船湖楼烟堤飘岛洲波路西荆蒲游汀通临陵
10	马将城河汉边北天塞山西关黄军下骑出剑国胡	10	城营河关旗旌衣边王侯黄蛮榆氛军都原车尘
11	雨叶树风竹色露阴晚清林池野烟绿静景疏秋草	11	池园庭窗渠亭阁微秋新莲台尘森砌高枝斜筠
12	朝文恩重从书同主分传才官阙方当节府入笔御	12	门台闱臣墀旒銮舆朝元阁轩书禁官时京侍恩
14	山石云松泉林峰寺溪幽下隐寻僧高鹤鸟入岩野	14	山松房廊云龛空经杉磴堂蹊峰林潭竹禅室幡
17	玉金歌宫曲飞女翠光双舞珠香王碧凤楼轻转殿	17	宫楼台筵王房鸾池歌弦光廊衣妃绿盘昭阶差
18	花春红不香似小桃语愁深情相看欲画恨怜新垂	18	楼花房帘人深墙眉炉香窗来愁欲春红桃渠浓

表 7

“殿”的骈偶下字频数统计							
（宫 48）	（楼 33）	（门 15）	（台 14）	（山 13）	（人 10）	（城 10）	（房 9）
（池 9）	（云 8）	（天 8）	（庭 8）	（王 8）	（光 7）	（窗 7）	（阳 7）
（龙 7）	（亭 6）	（园 6）	（廊 6）	（堂 5）	（微 5）	（朝 5）	（条 5）
（松 5）	（枝 5）	（烟 5）	（舆 5）	（营 5）	（阁 5）	（坛 4）	（声 4）
（壶 4）	（尘 4）	（明 4）	（旗 4）	（来 4）	（河 4）	（深 4）	（渠 4）
（空 4）	（筵 4）	（经 4）	（臣 4）	（花 4）	（衣 4）	（车 4）	（金 4）
（闱 4）	（阶 4）	（陵 4）	（风 4）	（黄 4）	（侯 3）	（元 3）	（关 3）

表 8

我们同时还直接统计语料中“殿”的所有骈偶下字的频率，如表 8 所示。对比表 7 和表 8 的数据，我们可以直观地感受到，概率骈偶主题模型比简单的条件频率统计更具有模型优势：不

仅发现了汉字之间的骈偶关系，还能够对这些骈偶字对的主题进行区分。这可以进一步为自动写诗提供更有表达能力（Expressive）、主题驱动（Topic-driven）、风格区分（Style-discriminative）的数学模型。

4.4 基于联句任务的定性对比

为了能够对我们提出的概率骈偶主题模型进行更好的评价，我们引入古代诗歌创作中的“联句”任务，以此来对比不同模型的性能差异。

“联句”是指给定一句诗，续写下一句。若给定前人的诗句，则续接的必须是全新的句子。联句可以是完全对仗的形式，也可较为自由。本文中，我们在该任务上测试了四个模型：

- 平均随机模型（Random Model）是指从字表中随机选字组成下句。
- 主题模型（LDA）是给定隐变量主题，在该主题下按概率选字，组成下句诗。这里，为优化生成过程，下句所有字暂且共享同一个给定的主题。此外，下字的概率分布表仅选择在语料中出现过的、概率较高的部分汉字。
- 条件频率模型（Conditional Frequency Distribution Model）是指在已有的语料中，以给定的上字为条件，统计所有出现过的下字的频率。在生成的过程中，根据给定诗句中的汉字，按照其对应的下字条件频率分布，产生下句中相对应的位置上的字。为优化生成结果，下字的频率分布表中剔除了出现频率过低的汉字。
- 概率骈偶主题模型（Probabilistic Couple LDA）是指对于上句中的每一个字，在给定的主题下，按下字频率分布随机选择汉字，并由这些字组成诗句。同样，为了优化词袋模型对于短文本的生成能力，待生成文本的主题已事先给定，下字的概率分布表也仅选择在语料中出现过的、概率较高的部分汉字。

给定诗句：白日依山尽（[唐]王之涣《登鹳雀楼》）				
模型	平均随机模型 (Random Baseline)	主题模型 (LDA)	条件频率模型 (ConditionalFreqDist)	概率骈偶主题模型 (ProbCoupletLDA)
生成结果	仆瞥钬递瘉 凄谕照醺雏 弩驿欧巑缆 谟宝桩蝘钧 航薜怨鲤灭	琴指惊流响 鸿洒槐井阳 火兵营朔帐 映岫杳槛亭 云乡故到风	青云逐日新 青云向山来 人君带人生 绿春去人皆 碧心归夜迟	青山远客遥 沧蘋远江波 东潮棹峡船 孤流越波帆 还涯乡雁空

表 9

我们给定唐人王之涣的名句“白日依山尽”，按照上述方式利用四个模型进行“联句”。不同模型的生成水平参差不齐，这里暂且选登了各模型中较好的诗句做定性比对，结果如表 9 中所示。可以明显看到，平均随机模型的生成结果，可理解性非常低。相比随机模型，主题模型的生成能力有明显提升，但在上下句的意义对应方面有所不足。条件频率模型的生成结果可以说是非常不错的。尤其是，条件频率模型强烈地表达了骈偶对应关系，如“白”和“青”、“碧”形成了非常典型的对仗，但有时这种强烈的对应也会表现得过度充分。例如，“绿春去人皆”中的“皆”与“白日依山尽”中的“尽”相对，从孤立的字层面来说非常好，但整体看来，“尽”实际上在句子中做动词而非副词，所以用表示副词义的“皆”相对似有不妥。概率骈偶主题模型似乎是能够综合主题敏感和骈偶对应两方面的特征，在所选下句主题与上句意境较为契合的情况下，甚至能够有一定概率生成很精彩的句子，如“沧蘋远江波”。

定性地来评价，对于“联句”任务而言，条件概率模型与概率骈偶主题模型在生成结果的诗性、骈偶等诸多方面都足以分庭抗礼，普通的主题模型效果次之。概率骈偶主题模型时而能产生

意境深远的联句，如“物随秋渐老（[明]边贡），临垂晚风愁”、“人闲桂花落（[唐]王维），鸟静寒叶疏”。当然，生成这样优秀的联句也是概率事件，在联句实验中也常出现“人闲桂花落 ，菊澹荷气起”、“人闲桂花落 ，菊密寒鸟萧”等大为逊色的结果。

不过，值得一提的是，条件概率模型的一个重要缺陷是模型中没有任何主题相关的信息，所以在给定相同的诗句下营造不同意境的能力有所欠缺。而这一点恰恰是概率骈偶主题模型的优势。如表 10 中所展现，一句“鸟惊千树雪”，利用概率骈偶主题模型可以续接不同主题的联句，创造出与原诗“鸟惊千树雪，人语数峰烟”相似或迥异的意境。这一点也与以往研究的诗歌生成模型（He, 2012; Zhang, 2014）有着显著不同。

给定诗句：鸟惊千树雪（[明]高启）				
条件频率模型	无主题			
	帆厌万枝阳 钟怕数林丝	名下四烟冰 帆厌半泉阳	烟照半江阳 烟怕半河梅	名厌数门阳 名入万春春
概率骈偶主题模型	主题 1(羁旅)	主题 2(秋寒)	主题 3(游仙)	4（春景）
	情断故云远 乡归望情寄 旅随未思鸿	卧落雨楼霜 彻滴一明庭 云清尽愁侵	云天万云霞 星骇飞地阴 钟天造氲阴	烟啭杏园枝 桃满送水川 光乱送堤鹋

表 10

本文提出的概率骈偶主题模型能够整合主题和骈偶的信息，主题控制着“怎么写”，骈偶控制着“写什么”，在联句任务中表现出了一定的优越性。但由于时间所限，我们对于上述四个模型，未能做出定量的评价。在未来的进一步研究中，会考虑引入心理学的测试评估方法，对模型生成的诗句进行各维度评分，以期在定性的感受之外，给出更有解释力的定量评价指标。

5 后续工作

在后续工作中，我们希望进一步完善模型，使其能够具有更好的表现力和概括能力，并能够对读者进行诗歌评鉴的直觉加以建模。本文提出的概率骈偶主题模型，假定了骈偶上字直接对骈偶下字产生影响。而当前正在进行的工作则尝试着泛化这一条件，建模骈偶上字的主题对骈偶下字的主题的耦合关系，同时引入汉字二元组（Bigram）的限制条件，以期能够在诗歌研究以及更为复杂、要求更高的生成任务（如次韵、依韵[1]）中取得令人惊叹的表现。

[1]均为旧体诗写作的方式，统称和韵，但规则和难度有所不同。依韵，要求韵脚与原诗韵在同一韵部而不必用其原字，难度中等。次韵，或称步韵，要求后作诗的韵脚必须严格使用前诗的原韵原字，而且用字先后次序也必须相同，难度最大。因此，古人也认为：“步韵最困人，如相欧而自縶手足也。盖心思为韵所束，于命意布局，最难照顾。今人不及古人，大率以此。”（吴乔《答万季野诗问》）

例如苏轼和章质夫著名的次韵词作：

《水龙吟》章质夫

燕忙莺懒芳残，正堤上、柳花飘**坠**。
轻飞乱舞，点画青林，全无才**思**。
闲趁游丝，静临深院，日长门**闭**。
傍珠帘散漫，垂垂欲下，依前被、风扶**起**。
兰帐玉人睡觉，怪春衣、雪沾琼**缀**。
绣床旋满，香球无数，才圆却**碎**。
时见蜂儿，仰粘轻粉，鱼吞池**水**。
望章台路杳，金鞍游荡，有盈盈**泪**。

《水龙吟・次韵章质夫杨花词》苏轼

似花还似非花，也无人惜、从教**坠**。
抛家傍路，思量却是，无情有**思**
萦损柔肠，困酣娇眼，欲开还**闭**。
梦随风万里，寻郎去处，又还被莺呼**起**。
不恨此花飞尽，恨西园落红难**缀**。
晓来雨过，遗踪何在？一池萍**碎**。
春色三分，二分尘土，一分流**水**。
细看来，不是杨花，点点是离人**泪**。

参考文献：

[明]胡震亨. 唐音癸签

章培恒, 骆玉明. 2011. 中国文学史新著. 复旦大学出版社.

Blei, D., A. Ng, and M. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Blei, D. and J. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*.

Hall, David, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the History of Ideas Using Topic Models. In *Proceedings of EMNLP*, 2008:363-371.

He, Jing, Ming Zhou, Long Jiang. 2012. Generating Chinese classical poems with statistical machine translation models. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Hughes, James M., Nicholas J. Foti, David C. Krakauer, and Daniel N. Rockmore. 2012. Quantitative patterns of stylistic influence in the evolution of literature. *PNAS*.

Kao, Justine and Dan Jurafsky. 2012. A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry. *NAACL Workshop on Computational Linguistics for Literature*.

Klingenstein, Sara, Tim Hitchcock, and Simon DeDeo. 2014. The civilizing process in London's Old Bailey. *PNAS*.

Luo, Yuming. 2011. *A concise history of Chinese literature*. Koninklijke Brill NV, Leiden: Netherlands.

McFarland, Daniel A., Christopher D. Manning, Daniel Ramage, Jason Chuang, Jeffrey Heer, and Dan Jurafsky. 2013. Differentiating Language Usage Through Topic Models. *Poetics*, 41 (6): 607-625.

Voigt, Rob and Dan Jurafsky. 2013. Tradition and Modernity in 20th Century Chinese Poetry. *NAACL Second Workshop on Computational Linguistics for Literature*.

Zhang, Xingxing, Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In In *Proceedings of EMNLP*, 2014:670-680.