广东工业大学

硕士学位论文

基于进化策略的仿唐诗自动生成系统

姓名：曹卫华

申请学位级别：硕士

专业：计算机应用技术

指导教师：张灵

20110529

摘　要

诗歌是一种具有独特魅力的文学体裁，是人类文明的象征，用计算机模拟生成诗歌史自然语言生成领域的一大挑战。本文研究机器自动生成仿唐诗，对其可能性和具体实现方法进行详细的研究和讨论。研究内容主要有以下几大方面：

1、总结和分类诗歌生成领域到目前为止的研究成果，对每种方法的主要步骤和代表性诗歌生成系统进行介绍。并简述中国古典诗词的计算语言学研究概况。

2、建立唐诗语料库。将词句按格律细分为子句，统计子句字串，根据频率、共现度等参数抽取结合强度高的字串，结合各种已有的词典资源来建立唐诗词表。根据条件概率对已切分的唐诗进行注音，建立音韵数据库。

3、根据唐诗建立相关的语法规范，用确定性自动机（DFA）进行词句语法合法性判断。计算语义度量：通过潜在语义分析和互信息计算词义相关度；通过词典和语料库统计相结合的方法计算词义相似度；成立专家组对高频词进行风格和情感的分级评判。

4、基于进化策略建立仿唐诗生成模型。根据唐诗特点，编码方式是基于平仄规律的，适应度函数是基于语法和语义加权值的，选择策略是基于精英主义和轮盘赌算法的。各主要操作的实现步骤在文中都有详细介绍。

5、建立基于进化策略的仿唐诗生成系统，包括给出系统框架、主要实现流程和具体的仿唐诗生成实例。并且对实验结果进行了分析和总结。

实验结果表明，本文建立的计算模型和设计的系统基本上可以实现计算机自动生成仿唐诗的目标，为今后进一步的研究提供了理论和实验基础。

关键词：自然语言生成 ；计算诗学；仿唐诗生成；进化策略

ABSTRACT

The computer simulation of the human poetry - a special type of literatureand a typical phenomenon of human creativity - is a great challenge of naturallanguage generation. This paper aims to conduct an elementary research inChinese ancient poetry generation. We analyze the possibility of the computerautomatic poetry generation and discuss the method of its implement in detail. Our research includes the following aspects:

1. Make a summarization of the machine poetry generation research. Weclassify the methods and introduce the representative systems of each sort. Make a summarization of computational linguistics-based Chinese ancientpoetry research.

2. Build the Song Poems corpus and database. According to the rules andforms of Song Poems, sentences are divided into sub-pieces. Closelycombined two-character words are extracted by calculating the frequency andcollocation rate. After completing the segmentation, we get a lexicon. Strategies such as conditional probability are used to implement automaticpinyin-tagging.

3. Establish the grammar criterion of Chinese ancient poetry generation. Propose a Deterministic Finite Automata - based method to judge thegrammatical validity. Combine Latent Semantic Analysis and MutualInformation methods to calculate lexical relevancy; Use corpus statistics andKeenage to do the lexical similarity computation; And retain the expert panel to do the stylistic and emotional measurement of words.

4. Propose the Genetic Algorithms approach to Chinese ancient poetrygeneration. According to the characteristics of Chinese ancient poetry, wedesign the Level and Oblique Tones-based coding method, the grammaticaland semantic weighted function of Fitness, the Elitism and

Roulette combinedselection operator, the partially mapped crossover operator and the heuristicmutation operator. III

5. Construct the Genetic Algorithms — based Chinese ancient poetrygeneration system. Describe its implement, give the flowchart and someinstances of the result, and analyze the result. As is shown by a certain number of tests, the system constructed on thebasis of the computing model designed in this paper is basically capable ofgenerating Chinese ancient poetry, and we hope that this work can serve asthe foundation for further research in the field.

KEYWORDS: Natural language generation; Computational poetics; Imitation Tang poetry generation; Evolution strate

Content

第一章 引言

人类通过自然语言和各种符号语言进行交流和思考。语言是人类智能活动中不可或缺的工具，随着计算机在不同领域逐步替代人类完成各项工作，人们也期待计算机能够真正的理解自然语言，进而实现人与机器的自然交互。

诗歌这种语言形式用语简洁凝练、结构跳跃灵动、富有节奏和韵律、高度集中地反映了人类的思想感情和生活，表现了人类智慧在语言层面的天赋才能。作为诗歌之国，中华民族的诗歌文化源远流长，但一直以来，其研究大多局限于语言和艺术领域，直到 20 世纪 90 年代中期才兴起运用计算语言学手段研究中国古诗词这样一个全新领域。作为人类语言最璀璨的明珠，诗歌创作一直被视为人类，甚至是文人墨客的专利，是人类智慧特有的产物，但是随着当代计算机技术的迅猛发展，人工智能成为一种思潮，让机器产生智慧成为人类最大的野心，越来越多人投入计算机模拟人类思维和创作的研究，并且已经取得了一定的成就，那么计算机能否自动或辅助人类创作古诗词呢？本文以汉语古诗词为研究对象，研究和讨论计算机模拟诗歌生成的可能性和方法。

1.1 研究的技术背景

自然语言生成(Natural Language Generation，NLG)是自然语言处理领域中的重要分支，它以人工智能和计算语言学为基础，研究和模拟人类生成自然语言文本的过程和方法。NLG 研究计算机如何根据信息在机器内部的表达形式来生成一段高质量的自然语言文本。

通常开发运用 NLG 系统主要有两个目的：①从经济角度考虑，NLG 系统可以作为人们生活中的交际工具，它在生产速度、纠错、多语言生成等方面具有很大的优势，可以利用语言知识和领域知识来生成文本、帮助消息、分析报告等；②NLG 系统可以检验特定语言理论，其工作过程与研究自然语言本身有着非常紧密的联系，必然会涉及语言理论诸多方面的内容[1]。自然语言文本生成在早期采用的是罐装文本（canned text）和模版填充（template filling）技术[2]，但这两种方法存在很多缺点，它们缺乏灵活性，难以生成灵活、多

样的文本，在系统维护、修改和扩充等方面都十分困难，所以现在人们通常采用形式化的方法，在语法和文本规划水平上进行语言生成的研究，其主要思想是将用户的输入信息和从信息管理系统的数据库中得到有关信息合并在一起，经过转换得到数据信息的深层语义，由文本规划器进行文本的规划，然后由文本实现器将规划器所生成的中间结果转变成自然语言文本。NLG 的经典结构包括句子规划(也称微观规划) 、内容规划(也称宏观规划)和表层生成三个基本功能模块[3]，其生成过程是系统根据应用目标和用户模式来建立相应的语义表示、语法分析和话语结构。结构如图 1 所示。

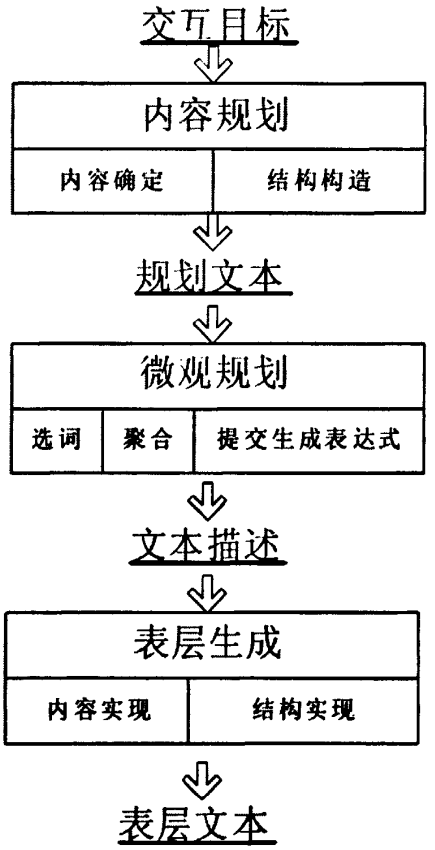

图 1-1.NLG 的经典结构

Figure 1-1.NLG classic structure

内容规划就是对生成的内容进行选择并以将这些信息连贯地组织起来。采用的手段手段一般是基于修辞结构理论(rheto-rical structure theory，RST)的方法和基于 SCHEMA 的规划方法[3]。句子规划就是优化句子的表达方式。近年来，计算语言学的研究有了进一步的发展，在提高生成文本的可读性和描述的

清晰程度上产生了很多新技术。表层生成(surface generator)就是根据预先定义好的语法规则，将前面输出的文本的数据结构(通常是一棵树)进行单词的线性化输出。主要方法和技术包括根据形式语法生成模型、短语结构扩展生成模型、系统功能语言学模型、基于扩展转移网络生成模型和基于合一的生成模型等。经过四十年的发展，自然语言研究的专家们不断提出新的理论和方法，设计出新的生成模型，取得一系列新的进展。目前语言生成的研究侧重于以下几个方面[1]：

①在特定的语法理论框架内更加广泛深入地处理语言现象，如 Fawcett 的 GeneSys 生成系统。

②在同一语法环境下生成多语言，如英国 Stirling 大学的 Nigel 多语种生成系统（包括英语、德语日语、法语、荷兰语、西班牙语），上海交通大学的多语言天气预报发布系统。

③面向实际应用的开发，如英国 Edinburgh 大学 Michaelo ‘Donnell 所设计的在线文件剪接系统、北京交通大学和北京颐和园的导游系统、中国科技大学的机器人足球现场解说系统等。

④对所要表达的信息进行语义和句法方面的聚合。当前语言生成的研究方向主要是在语言表示形式、信息内容规划以及语言生成模型等方面。自然语言生成的研究只有在诸多语言学科、计算机领域和其他学科的通力协作下才能获得新的成果。

1.2 研究的内容与目标

中国的诗歌文化源远流长，优秀的诗歌作品数不胜数，在汉语文化的成长、演变与传播中占有着极重要的地位。诗歌这种体裁语言极度凝练，优美隽永，是对语言的一种高度的应用。用计算机生成诗歌史对自然语言处理领域的一大挑战。国外对这方面的研究起步较早，目前已经取得了一定的经验和成果，有一部分较成型的系统可供使用（具体参见本文第二章的综述）。而我国在这方面的起步较晚，至今不过差不多十年时间，其研究成果较少，并且大都还集中于词汇语义方面，对机器诗歌创作的系统性研究比较欠缺。因此，本文对该课题进行了研究和探讨，研究诗词自动生成的可能性、方法、计算模型和困难分

析，希望通过对诗词自动生成机制的研究，构建较完善的计算模型和系统，以使计算机能够模拟人类思维，生成可以“以假乱真”的诗词作品。由于汉语古典诗词种类繁多，在研究的起步阶段，我们首先尝试从一种诗体入手，建立具有较好推广性的模型。考虑到唐诗具有格律严谨的特点，本文主要针对唐诗建立了生成模型。

1.3 本文研究的主要贡献

本文对仿唐诗的计算机自动生成进行了初步研究。文章的主要贡献突出表现在以下几个方面：

1、根据大量的中外文献对诗歌生成领域到目前为止所使用的方法进行了总结、分类和分析。概括各种方法的主要实现步骤，并介绍了每种方法的代表性系统及运行实例。

2、对诗歌生成的语法规则和语义度量提出了可供计算机量化操作的方法。在语法合法性检验方面，根据唐诗格律特点，在研究分词模式的基础上，提出了利用 DFA 进行判定和过滤的方法。在语义度量方面，提出了基于词义相关和风格、情感统一性的检验。

3、汉语古典诗词的生成研究尚处于起步阶段，目前没有成型的方法和模型可供借鉴。本文提出的基于进化策略的仿唐诗生成模型是一次大胆的尝试。根据唐诗的特点，在编码方面选择平仄编码方式，适应度函数基于语法和语义加权值，选择策略基于精英主义和轮盘赌算法、部分映射和启发式交叉算子和启发式变异算子。经过实验证明，模型的建立和算法的设计较好地达到了仿唐诗生成的研究目标。

1.4 论文结构

论文主要结构和内容安排如下：

第一章：介绍计算机模拟诗歌生成的研究背景，研究目标，目前开展的工作和本文的主要贡献。

第二章：综述诗歌生成研究背景，介绍自然语言生成领域中与诗歌生成相关的已有研究成果，并分类、总结和分析所用方法，对目前为止最有影响力的

一些代表性系统进行介绍；概述从计算角度出发的汉语古诗词研究现状和进展。

第三章：详细介绍与计算机自动诗歌生成相关的一些基础性研究工作，比如唐诗语料库的建立，唐诗的切分，词汇的自动提取，词典的建立，词性标注和音韵标注等。

第四章：分析仿唐诗生成需要遵循的语法规范，计算语义度量。应用 DFA 检验语法合法性，根据词义相似和词义相关实现语义度量，还提出词汇风格和情感意义的标注办法。

第五章：分析和比较现有研究成果，以唐诗为切入点，结合汉语古诗词的具体情况，提出基于进化策略的研究思路，并详细的介绍编码、初始种群生成、适应度计算、选择、交叉、遗传等主要操作。

第六章：构建仿唐诗生成系统。介绍系统的总框架、算法流程和具体实现情况，给出系统运行的实例，分析实验结果。

第七章：总结和展望已开展的研究工作。指出研究中发现的问题，提出改进和推广建议，为进一步的工作定出方向和目标。

第二章 诗词生成及汉语计算诗学综述

自20世纪70年代以来，国际上对于机器作诗的研究逐渐兴起，目前为止已尝试了许多方法并取得了一定的进展，出现了一些较成熟的方法和一些可供使用的系统。而在中国，20世纪90年代中期才开始汉语古诗词计算语言学方面的研究，至今不过十年时间，而且研究成果目前大都还集中于词汇语义方面。在开展汉语古诗词的机器生成研究之前，我们首先对前人的研究成果作简要的回顾。

2.1 计算机诗歌生成综述

诗歌生成技术历经了几个阶段的发展[4][5][6]，从早期的 Word Salada 发展到现在的较为成熟的基于进化算法和基于实例推理的方法，从简单粗糙发展成复杂完善。在本小节中，我们将回顾诗歌生成的发展，对主要的研究方法进行总结，并简要介绍几个具有代表性的诗歌生成系统。

2.1.1 随机词汇连接（Word Salada）

早期的机器作诗被成为 Word Salada，因为它的生成结果仅是一些词汇的堆砌。这种方法基于简单的计算程序，采用连接随机生成词汇的方法，对诗歌内容、形式和意义的考虑都很少，其作品从严格意义上说并不能称为诗歌，但该方法敲开了诗歌生成这一片崭新的研究领域，是一种很有意义的尝试。Word Salada 代表系统有 Pete Kilgannon 的“LYRIC 3205”[6]，其作品举例如图 2-1 所示下：

judy gotta want upon someone.
wanna sadly will go about.
sammy gotta want the thief him but the
every reason. real distance carry.

图 2-1.基于随机词汇连接的计算机诗作
Figure 2-1. Based on random words connected computer poem

2.1.2 基于模版的诗歌生成系统

基于模版的诗歌生成系统的生成机制相对来说比较复杂。这类系统事先定义好一个模版，其中固定了生成诗歌中的某些词汇或短语片段，其余片段则留出空白用以填充词性、时态等信息，大多为实词，如名词、动词、形容词，偶尔也填充副词。这些词是计算程序从词典中根据条件随机选择的[5]。这类系统的代表有 RACTER 和 PROSE(Hartman，1996)[7]、RETURNER、APPI、Masterman 的俳句生成系统等，以及互联网上的 ELUAR、ALFRED 等实用系统[8]。以下时 RETURNER 原型系统的一个生成模版：

1.IN THE MORNING+noun phrase with a noun as head+WILL+APPEAR/BE/BECOME/SEEM/TURN+adjective phrase

2.Noun phrase with a noun as head+ALSO/NEVER/OFTEN/SOMETIMES+verb in the present tense+AGAIN

3.LAST NIGHT/TODAY/TOMORROW+pronoun+verb phrase(a verb in past/present/future tense)+pronoun+THROUGH THE WILLOWS

基于该模版的输出实例如图 2-2 所示：

In the morning crowbars will be nearly round.

Separate blankets never step again.

Tomorrow I will ring him through the willows.

图 2-2.基于模版的机器诗作示例

Figure 2-2. Poetry example of template-based machine

基于模版的诗歌生成系统虽然有较好的输出，但这类系统也存在一些固有的缺陷：人为参与较多，生成作品的质量很大程度上取决于模版的设计，体现不出机器的自动性。采用减少留白数量，选择充斥诗性词汇的倾向性词库等投机的方法，能很大程度改进输出结果的质量，更多地体现的是人的诗性而不是机器的诗性，虽然在一定程度上满足了合乎语法的要求，但离机器自动作诗的目标还有很大差距。

2.1.3 基于设定模式的诗歌生成系统

基于设定模式的系统与基于模版的方法相同，通常有一个事先设定的模式，不同的是模式的灵活性远大于模版，更加地合乎语法和韵律要求。由于这类系统的模版都不尽相同，以下我们以具体的系统为例逐一进行介绍。

Gerv′ as(2000)的 WASP 系统，事先设定句子数目，每个句子词汇数，形容词与名词的比例，时态等模式信息[5][9]，然后从首个适配位置出发，采用贪婪算法在词库中搜索符合条件的词汇逐一填充所有适配位置。为了保证所选词不重复出现填充过程中还有额外的启发式搜索机制，生成实例如图 2-3：

Mu´erome por llamar Juanilla a Juana,

que son de tierno amor afectos vivos,

y la cruel, con ojos fugitivos,

hace papel de yegua galiciana.

图 2-3.基于设定模式的 WASP 系统诗作示例

Figure 2-3. WASP model system based on poems set an example

Ray Kurzweil 的 Cybernetic Poet(Kurzweil, 2001)系统以人类创作的诗歌为模式，统计大量的已有诗作的词汇、词汇结构及排列顺序、韵律模式、诗歌整体结构等方面，并分析和建模[10]。为保证诗歌主题的连贯性，采用特殊的递归算法。当算法无法找到合适解时，则放宽对特定词的约束，使计算得以继续。系统生成实例如图 2-4 所示：

Scattered sandals

a call back to myself,

so hollow I would echo

Crazy moon child

Hide from your coffin

To spite your doom.

You broke my soul

the juice of eternity,

the spirit of my lips.

图 2-4.基于设定模式的 Cybernetic Poet 系统诗作示例

Figure 2-4. Based on the system setting mode Cybernetic Poet poetry example

Rubaud et al.(2000)ALAMO 小组的 Rimbaudelaires 诗歌生成器通过用空格替换 Rimbaud 十四行诗中的名词、动词和形容词来构造诗句模版，然后从 Baudelaire 的诗中选取相应的词进行填充。选词算法加入了强句法和韵律约束，从而保证了系统输出作品的诗性[11]。

2.1.4 基于进化算法的诗歌生成系统

基于进化算法的诗歌生成模型包括生成模块和评价模块两部分。生成模块根据词法、句法、概念等信息产生备选诗作，再由评价模块依据制定的准则对备选输出给出等级评价。该模型的评价模块是一个两层的评价体系。其中，低层评价器基于主观和客观两种评价机制：用大量已有诗作的人为评判意见训练神经网络，用以模拟人类作出的主观评判；同时，通过语法、韵律等规则进行客观评判。高层评价器则由用户制定低层评价器中各种评价参数的比重[12][13][14]。

基于这一思想 Levy(2001)构造了原型系统 POEVOLVE，虽然没有完整实

现该模型，但它让人们看到了计算机在诗歌生成方面的潜力[12]。POEVOLVE从词库中选词生成初始种群。词库中的每个词均有有强弱音、押韵等标注信息。评价系统以基于受训神经网络的主观评价为主要手段。

Hisar Maruli Manurung（2003）提出诗歌必须满足的三个条件：语义（Meaningfulness），语法（Grammaticality）和诗性（Poeticness）。他的MCGONAGALL系统，将诗歌生成问题看成一个状态空间搜索问题，目标状态是满足三个条件的文本[18]。在语义表示上，采用了词汇化树邻接文法；在评估函数上，采用了编辑距离算法和结构相似度两种度量。MCGONAGALL是迄今为止最为成熟的一个基于进化算法的诗歌生成系统，图2-5给出了它的一个生成实例：

There is a young lady called bright.

She (will) travel much faster than light.

She set out one day relatively.

She is on (a) preceding night.

图2-5.基于进化算法的MCGONAGALL系统诗作示例

Figure 2-5. MCGONAGALL system based on evolutionary algorithms sample poems

2.1.5 基于实例推理的诗歌生成系统

基于经验知识进行推理的人工智能技术的代表是CBR，这种推理方法用案例来表达知识并把问题求解和学习相融合，利用过去积累下来的对于类似情况的处理方案解决新问题，当然在这个过程中要通过适当的修改[16]。采用CBR技术的系统一般包括四个处理步骤[17]：匹配retrieve：以案例的形式向系统表述当前问题的特征变量。在案例库中通过案例的索引与检索寻找与当前问题最为相似的案例。重应用reuse：如果旧案例中存在与当前问题相符的解决方法，则在新案例中重新应用这些方法，直接输出该问题的解决方案。否则，完善修改检索出的案例，生成新案例。修正revise：修改和完善新案例中与旧案例不同的地方，形成一个全部满足新案例的解答。保存retain：评价当前问

题的解，并将新方案增添到实例库中，以备日后求解问题使用。

基于实例推理方法的代表性诗歌生成系统有 ASPERA(Gerv′ as，2001)and COLIBRI(Diaz-Agudo et al，2002)两个系统 [9][15]。ASPERA 系统要求用户描述目标输出，包括给出目标输出的诗歌类型，包括长度、情感倾向等信息。用户输入首先要通过特定的专家系统进行预处理。然后，输入被切分为与生成的诗句相对应语义片段。例如 Gerv′ as(2001)的系统中[9]，若输入信息为‘John loves Lucy they go together to the beach’，系统将其切分为三个片段：

(片段 1 John loves Lucy)

(片段 2 they go together)

(片段 3 to the beach)

在得到切分片段后，CBR 模块开始执行以下四个步骤：

(1)匹配：根据一定的相似性评判机制为每个片段选择适当的诗句模板。相似性判断算法的判断依据主要是片段中出现的关键词和目标诗歌类型。

(2)重应用：通过挑选出的现有诗句模版逐个替换切分片段，最终形成一个包含所有片段的诗歌草样。

(3)修改：在 Gerv′ as 的系统中，自动修改算法还未完全实现，该步骤现阶段是通过用户交互实现的。

(4)保存：将修改后的诗歌作为新的诗句模版保存到实例库。

图 2-6 是基于 CBR 诗歌生成系统的一个实例，(a)为用户输入，(b)为系统匹配的诗句模版，(c)为切分语义片段在模版中适当位置的替换，源于语义片段的词用黑体标出，(d)为修改后的诗作，带*号的词是系统出于韵律和语法作出的修改。

(a)

una boca ardiente pase techo y suelo

(b)

no s'olo en plata o viola truncada
se vuelva mas t'u y ello juntamente
en tierra en humo en polvo en sombra en nada

(c)

no s'olo en **boca y** *viola* **ardiente**
se **pase** *mas t'u y ello juntamente*
en tierra en **techo** *en* **suelo** *en sombra en nada*

(d)

no s'olo **para*** **boca y** *viola* **ardiente**
se **pase** *mas t'u y ello juntamente*
en **t'ia*** *en* **techo** *en* **suelo** *en sombra en* **serpiente***

图 2-6.基于 CBR 的 COLIBRI 系统诗作示例

Figure 2-6. COLIBRI system based on CBR for sample poems

CBR 在知识获取、求解质量、求解效率以及知识积累等方面，有着突出的优势。但对诗歌生成系统而言，自动修改算法的设计是一个难以突破的瓶颈。

2.2 汉语古诗词计算机辅助研究综述

我国学者在20世纪90年代中期已经开始了对古诗词的计算机辅助研究的初步探索，现在已经在语料库建立，创作风格辨析，词汇语义分析，联语应对等方面取得了一定的进展，主要的研究工作与代表性成果包括：在 90 年代后期，台湾元智大学与北京大学计算语言研究所合作开发的“古诗研究的计算机支持环境”模型系统[19]，初步实现了全文检索、超文本阅读、关键词检索、统计以及计算语言学辅助研究等功能。随后以 160 万字的宋代名家诗为研究对象，开发了“宋代名家诗自动注音系统”，将基于统计的语言模型与宋诗自身的音韵特点相结合，采用互信息策、条件概率策略略和规则策略三种多音字自动注音策略，实现了宋诗的自动注音[20][21]。

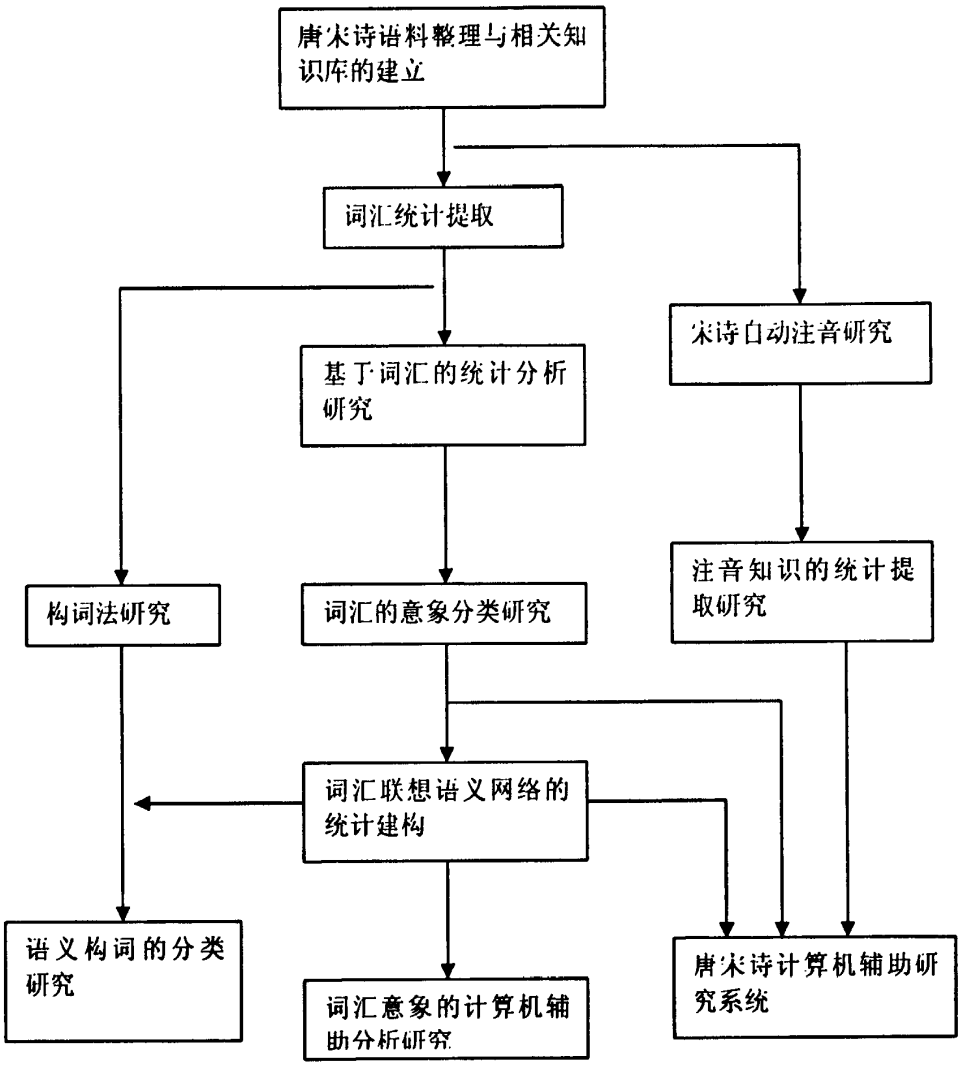

图 2-7.北京大学“唐宋诗计算机辅助研究系统”基本框架图

Figure 2-7. Peking University, "Tang and Song Poems of computer-aided system," the basic framework of the map

北京大学的胡俊峰在他的博士论文“基于词汇语义分析的唐宋诗计算机辅助深层研究”[22]中，根据古诗词语言的特点，改造和应用一些现代计算语言学技术，取得了一些有益的成果。这篇论文涉及的研究包括基于唐宋诗语料库的词汇自动提取、基于统计的词汇语义关系的自动发现、基于词汇的统计知识库的构建、诗词构词规则的提取等，还介绍了基于多条件复合检索技术的唐宋诗计算机辅助研究系统的开发及应用，系统的框架如图 2-7 所示。

重庆大学易勇在他的博士论文“计算机辅助诗词创作中的风格辨析及联语

应对研究”中，着重对诗词风格的机器评判进行了研究[23]。在这篇论文中，他采用向量空间模型来表示诗词，并用 Naive Bayes 等基于机器学习中的方法，首次提出了古典诗词的婉约和豪放风格辨析计算模型，并用遗传算法等方法改进模型，取得较好的诗词风格评判结果。文章在经典诗词语料的机器学习基础上实现了古典诗词的作者辨析计算模型，获得较好的诗词作者评判效果。

重庆大学的李良炎的博士论文“基于词联接的自然语言处理技术及其应用研究”则在诗词语言的理解方面提出了基于词联接的自然语言处理技术，提出了最优句树搜索的初级语言分析算法和词联接最大语义符合度计算，进行了诗词语言初级分析测试、诗词语料标注测试、诗词语言婉约与豪放风格的评价测试，取得了成功，并且深入分析了 NLP 技术背景，从而提出并初步构建了基于词联接的 NLP 技术，并应用到诗词语言处理系统中[24]。

《心脑计算举要》是厦门大学的艺术认知与计算实验室的周昌乐教授的著作，这本书提出了“计算诗学”的概念[25]。周昌乐教授为此成立了相关的研究小组。该小组目前的研究工作涉及汉语隐喻分析与理解、基于情感建模的诗词分析、诗词格律分析、古诗词自动分词及语料库建立、基于风格模拟的计算机辅助诗词生成、诗歌机器翻译系统的开发等内容。

另外，中国科学院自动化研究所的费越潜心研究古诗词自动生成十分相关的联语应对，在他的博士论文“汉语语义的多层次集成研究—及春联艺术系统设计”[26]中，他采用神经网络的方法研究形象思维层次的“语义”，并用春联论域内的词语进行实验。在神经网络的学习过程中，语义的数值表示序列在某种程度上类似于人类学习词语的过程，是一个从无序到有序的动态过程。文章在采取格语法语义表示的基础上，提出了汉语处理的神经网络并行模型，在语义表示和并行模型的基础上，构造了六个汉字以内的计算机春联系统。

重庆大学的易勇也研究过联语，他分析了传统对联的特点，将对联的上下联分别看作两个具有相同长度的语言单位的序列，将联语的应对生成问题抽象为有监督的序列学习问题，采用机器学习方法对其学习。首次提出了不限字数的联语应对生成的计算模型，并分别用隐马尔可夫模型序列学习法、N 元统计语言模型序列学习方法和基于转换的错误驱动序列学习法联语应对生成进行建模分析。在以字为语言单位的春联的应对生成上也取得较好的效果，基于语料库构造了不限字数的计算机联语应对实验系统，取得了较好的实验结果[23]。

微软亚洲研究院自然语言组于2006年完成了微软对联系统和微软对联聊天机器人系统系统能根据用户给出的上联自动提供若干下联供用户选择，如果不满意，用户还可以通过交互手段优选字词来生成满意的下联；当确定一副对联后聊天机器人可以生成若干四字横批供用户参考。目前该系统可处理八字以下的对联，并已在互联网上投入使用[27]。

第三章 唐诗切分语料库及音韵数据库的建立

从真实的语言材料出发研究语言，统计、归纳语言学规律，加以应用和创新，这是研究语言的正确途径。建立针对古代诗歌处理用的语料库是计算机诗词研究系统开发首先要解决的问题，因为语言知识库是自然语言处理系统的基本组成部分。从数量和成就角度而言，唐诗和宋词都是汉语古诗词的代表。在唐宋诗语料研究方面，比较成熟的语料库主要是北京大学和台湾元智大学建立的熟语料库。而厦门大学计算诗学小组则主要构建全唐诗语料库，也是比较成熟的语料库。本文的诗词生成试验主要集中于仿唐诗的机器生成，本章简要介绍语料库建立中与生成试验有关的部分，主要包括唐诗的切分、音韵标注和词性标注。

3.1 全唐诗切分语料库的建立

通过纯统计的方法北大计算语言所将使用稳定、结合强度较强以及带有隐喻义的二字词抽取出来，奠定了建立词表的良好基础[28]；台湾元智大学罗凤珠教授则主要根据诗词格律[29]来切分诗词。结合以上两种方法的优点本文采用一种新的方法来建立唐诗切分语料库。切分步骤如图 3-1 所示。

首先，除了通常考虑通常的词的定义之外，我们还提出了 4 种特殊类型的字串对切分中的“词”进行界定，即：专有名词，领字，典故和具有较明确隐喻意义的高频复合词。以张忠纲等人编著的《全唐诗大辞典》[30]为基础对特殊子串进行处理，结合前人所作的一些归纳，建立专有名词数据库，该数据库共分为天文、时令、人名、地名、人伦、人事、音乐、闺阁、形体、珍宝、建筑、文事、服饰、饮食、草木百花 15 大类；而典故数据库的建立以台湾元智大学罗凤珠老师的诗词典故资料数据库和范之麟 吴庚舜等人编著的《全唐诗典故辞典 》[31]为基础。以上两个数据库共含有词条 6873 条。

其次，根据唐诗切分和格律之间的关系，建立了格律数据库。唐诗格律的句法都有固定的总字数、总句数，每一句的字数也是固定的。通过对部分唐诗进行手工切分，可以发现唐诗切分和格律存在一定关系：词句可以根据对应的

句法来细分成子句。我们以徐青编著的《唐诗格律通论》[32]、张忠纲等人编著的《全唐诗大辞典》[30]和周勋初编著的《唐诗大辞典》[33]为基础，建立句法数据库，该数据库含有不同诗体的句法2415种，标注了各诗体的单字领字位置和句法。我们在数据库中收录了同种诗体的不同句法，并在生语料库中对该类诗体的唐诗所属句法类别进行人工标注。

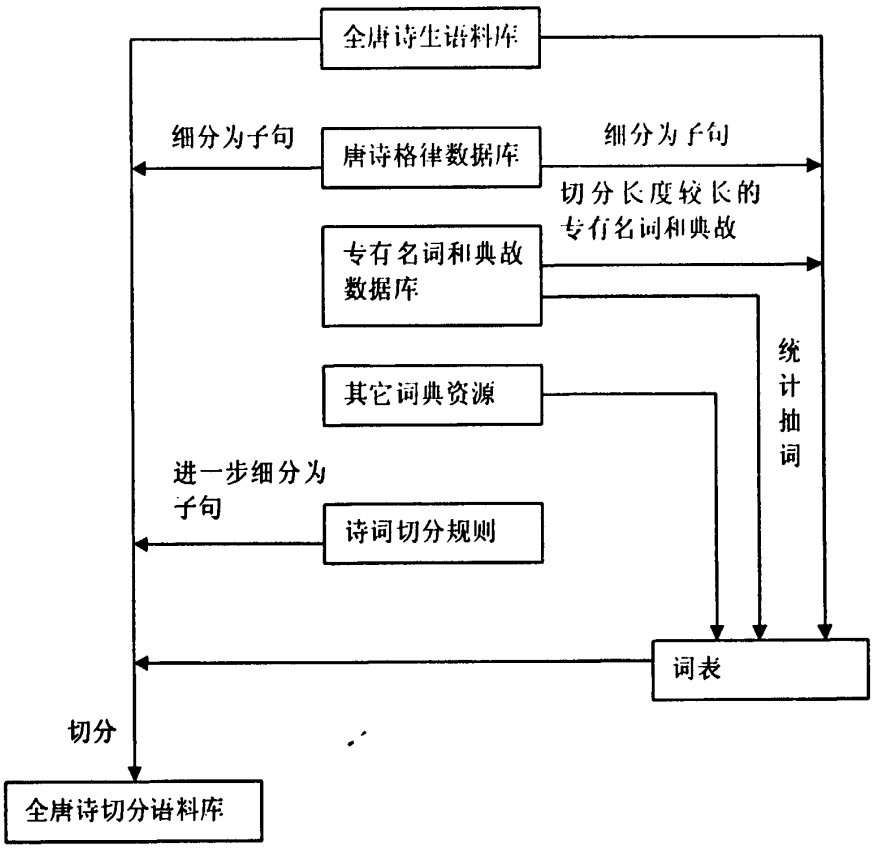

图 3-1.全唐诗切分语料库构建框图

Figure 3-1. Full Tang Segmentation Corpus Construction Diagram

最后，我们根据诗体格律数据库把诗句细分为子句，对子句字串进行统计，从中抽取结合强度较强的二字字串，并结合各种已有的词典资源来建立唐诗词表。在统计抽词过程中，我们主要采用了频率、共现度和互信息三个参数抽取可能成词的字串，并结合《现代汉语词典》，《辞源》等词典来对字串进行人工判断。其中，二字词的自动抽取采用互信息计算法，并结合频率来改进互信

息的提取效果。

$$\mathrm{I}(xy)=\ln\frac{P(xy)}{P_1(x)*P_2(y)}$$

公式中 P(xy)是汉字 x，y 在语料库中的一起同现的概率；P1(x)表示字 x 在语料库所有相邻二元字串中作为前字出现的概率；P2(y)表示字 y 在语料库所有相邻二元字串中作为后字出现的概率。需要说明的是，互信息不是衡量二元字串之间依赖性的好方法，虽然它可以衡量二元字串之间出现信息增加的统计量。由高频率字组成的相邻二元字串的互信息要小于底频率字组成的二元字串，而这会影响互信息的使用效果，为了改进互信息提取的效果，在此使用频率截断的方法，只考虑出现频率大于某个阈值的相邻二元字串。

由于在汉语中，词往往都是由不同字组合而成的。由于在衡量二元字串之间依赖性上互信息表现不足，所以为了衡量相邻二元字串中字的相互依赖性，实验中还采用了共现度[28]作为补充。

$$C(x,y)=R_1(xy,x)+R_2(xy,y)$$

$$R_1(xy,x)=\frac{f(xy)*\ln(f(xy))}{f_1(x)-\sum f(xz_i)}$$

$$R_2(xy,y)=\frac{f(xy)*\ln(f(xy))}{f_2(y)-\sum f(u_jy)}$$

公式中 $f_1(x)$为字 x 在语料库作为前字出现的次数，$f_2(y)$为字 y 在语料库中作为后字出现的次数，f(xy)为相邻二元字串 xy 在语料中的出现次数，$f(xz_i)(i=1，2，\cdots，n1)$为已经抽取的以 x 为前字的二元字串的出现次数，$f(u_jy)(i=1，2，\cdots，n2)$为已经抽取的以 y 为后字的二元字串的出现次数。

通过以上 3 个步骤，我们建立了初步词表，该词表包含专有名词、典故等，共含有词条 43387 条。

对唐诗进行机器自动切分分成四个步骤完成，分别为设置句法切分点、设

定词结构切分点、切分长度大于等于 3 的专有名词和典故、结合词表对字串进一步切分四个步骤。在对以曹寅、彭定求等编撰的《全唐诗》[34]为基础建立的包含 40863 首唐诗的生语料库进行切分的实验中，我们得到的分词正确率为 83.92%。

经过一定的人工校对和修正，我们对最后的切分结果进行了词提取，将切分得到的单字词，以及新出现的长度大于等于 2 的词加入词表，得到一个包含 49462 个词项的词库[35][36]。该词库是计算机模拟唐诗生成的词源。

3.2 词库的词性标注

根据词的语法功能，古汉语的词类可分为以下几种：名词、动词、形容词、介词、助词、副词、代词、区别词、状态词、语气词、数词、量词、连词、象声词、叹词。每个词类还可进行细分，如动词包含及物动词、不及物动词、使动词等，名词包含时间名词、地点名次词、方位名词等。

本文研究中，对生成的词句进行语法合法性的判断是通过词性标注信息来实现的。这一应用主要考虑的是词在句中所处的位置。作为对仿唐诗生成的初步研究，我们暂将研究范围限于出现频繁的词类，因为我们主要着眼于方法的提出。词类简化为以下 6 种：

N：名词，表示人和事物以及时间、地点等名称的词。

VT：及物动词，表示该动词后可以接名词。

VI：不及物动词，与及物动词相对应，这里表示的是相反的概念，即为了有意义，其后不能再接宾语。

EN：使动词，这是一种特殊的动词，由于在古汉语中大量出现，且与普通动词在使用上有所不同，故而单独列出。

ADJ：形容词，表示人和事物的形状、状态和性质的词。

ADV：副词，基本语法作用是用来修饰形容词、动词和其他副词，用作补语或状语，表示行为、动作、状态、性质的程度、时间、范围、肯定与否定、语气等语法意义。

在不同的上下文中某一个词可能具有不同的词性。例如：“月华如水”，此处的“如水”是不及物动词，而“如水碧云”中，却是作为修饰名词的形容

词出现的。针对这个问题，有两个方案可以选择：

(1)每个词只采用一种词性表示，而完善词性搭配的规律本身。如上例中，规律将允许在名词前面出现不及物动词，而不是仅仅允许形容词可以修饰名词。这种方案虽在词性标注上有所简化，但是会导致处理的规律复杂化，而且必须能够区分哪些不及物动词可以放在名词前面，哪些不可以。

(2)第二个方法就是允许出现不同词性的相同词，即允许“如水”作为不及无动词，也允许其作为形容词。这一做法可以使规律得以简化，虽然增添了词性标注的工作量。本文实验中采取第二种方案，因为扩充规则比较困难并且难以完备。多义词现象在实验中较为常见，为了避免在同一首诗歌中出现选择不同词性的同一个词的现象，程序设置了专门的处理机制。

3.3 词库的音韵标注

唐诗具有抑扬顿挫、高低起伏的特点。唐诗的每种体裁都有相应的平仄和押韵要求。平仄是诗词格律的一个术语。古代诗人们把四声分为平、仄两大类，平就是平声，仄包括上、去、入三声。韵是诗词格律的基本要素之一。诗人在诗中用韵，叫押韵。从《诗经》到后代的诗词，差不多没有不押韵的。所谓押韵，就是把同韵的两个或更多的字放在各句的同一位置上，一般总是把韵放在句尾，所以又叫“韵脚”。唐诗创作受严格的格律限制，因此，我们对词库中的每个词项进行了读音、平仄和韵部标注。

在音韵标注方面，根据《古今字音对照手册》、《现代汉语注音字典》和《广韵全字表》进行标注。对于多音字的处理，我们参考了论文《宋代名家诗自动注音研究及系统实现》提出的方法[20]，采用条件概率策略、互信息策略以及规则策略相结合的方法进行读音的判断。在自动标注的基础上，我们还进行了人工校对，最后生成的音韵标注数据库结构如表 1 所示：

其中“注音”表示相应汉字的现代汉语拼音，其中的数字 1、2、3、4 分别代表现代汉语拼音中的阴平、阳平、上声、去声。“韵部”表示相应汉字在广韵中所押的韵。“广韵声调”表示相应汉字在广韵中的声调，分别标注了代表古代汉语中的平、上、去、入四个声调。

表 3-1.全唐诗切分语料库构建框图

Table 3-1. Full Tang Segmentation Corpus Construction Diagram

字	注音	声调	韵部
烟	yan1	平	先
上	shang4	去	阳
处	chu4	去	鱼
雨	yu3	上	虞
梦	meng4	去	东
何	he2	平	歌
重	chong2	平	钟
生	sheng1	平	庚
帘	lian2	平	盐 A
尽	jin4	上	真 A
别	bie2	入	仙 B
断	duan4	去	桓
长	chang2	平	阳

第四章 语法规范的确立和语义度量的计算

怎样利用计算机模拟人类思维进行诗词创作是我们所要解决的问题。所谓模拟，主要两个层次的含义，一是语法，二是语义。在语法方面，诗歌是自然语言的一种文学形式的表达，有着严格的语法要求，这里所说的语法，不仅包括通常汉语所需遵循的语法，还包括诗歌特有的文法规则，或者说诗歌的格律，如平仄，押韵等规则。在语义方面，要求更进一步，它包括了风格的统一，主题的连贯，情感与意境的传达等。语义模拟最关键的问题是如何使句与句之间更有凝聚力，使产生的句子看起来更有意义，而不是毫无关联的词汇或句子的堆砌。

4.1 语法规则的判定

本节需要解决的问题是制定适合的语法规律，实现唐诗语法规律的计算机化。本文研究中将平仄和押韵规则看作诗歌语法要求的一个部分。唐诗的每种体裁都有不同的平仄和押韵要求。由于在数据库建立部分已经完成了词库的音韵标注，故格律要求的满足较容易实现。以下主要讨论基于词句搭配的语法规则。

古典诗词作为一种特殊的文体，其语法与普通文体的自然语言有着很大的区别。就唐诗而言，每种体裁下的句法都有固定的总字数、总句数，每一句的字数也是固定的。

在确定了分词模式后，我们考虑词的选取问题。在中国古典诗词中“诗家语”被广泛使用，成为一大特性。“诗家语” 有别于现代汉语中的常用词汇，有很强的专用性。本实验中构建的词典直接源于全唐诗的分词结果，很有力的保证了在生成诗作中“诗家语”的使用。但由于这些词与现代汉语的词汇在使用上有较大区别，因此不能套用现代汉语的语法，而需要对特定词库的语法规律进行有针对性的研究。

诗词的句与普通自然语言文本的句子相比有很大的不同。在普通自然语言文本中，通常是由严格的句法分析来保证句子的有效性的。而在诗词语言中，

句法成分往往并不完整，因为诗词要求具有高度凝练的特点，这必然会损害到句法的完整性。但通过对大量诗词语句构成的分析，我们发现，组成句子的有效模式的数目是有限的，并且呈现出了层次化的结构。而这种层次化的模式识别，可以用 DFA 或者 NFA 来表示。对应两种不同的自动机，有两种不同的表示策略：

(1)随机的组合词语，先不考虑他们之间的组合是否有效。在产生大量的备选个体后，让他们逐个的通过 DFA 的测试，好的留下，不好的剔除（由进化策略的进化规则保证，具体参见第 5 章）。或者甚至可以在 DFA 判断过程中加上类似属性文法的过程性的句子，即在通过 DFA 的过程中，不仅判断，而且修改，不断进行优化。

(2)在句子产生的初期就运用 NFA，以不同的概率产生不同模式的句子。这恰是 NFA 最擅长的。

该问题的解决取决于对一对矛盾的态度：更大的随机性与严格的诗词规律之间的矛盾。上面的两种策略恰好对应了对矛盾双方不同的强调。如果采用第二种方式，那么要获得一定的广泛性和多样性，必须要有一个足够大的词库，以避免很快在严格的条件限制下把词库里的备选词耗尽了。鉴于此，实验中采用的是第一种方法，可是因为没有比较，所以说不能说明第一种方法优于第二种，这有待于以后实验的验证。

由于篇幅限制，以下仅给出字长为 7，分词模式为“2212”的词句的 DFA 判断图，如图 4-1 所示：

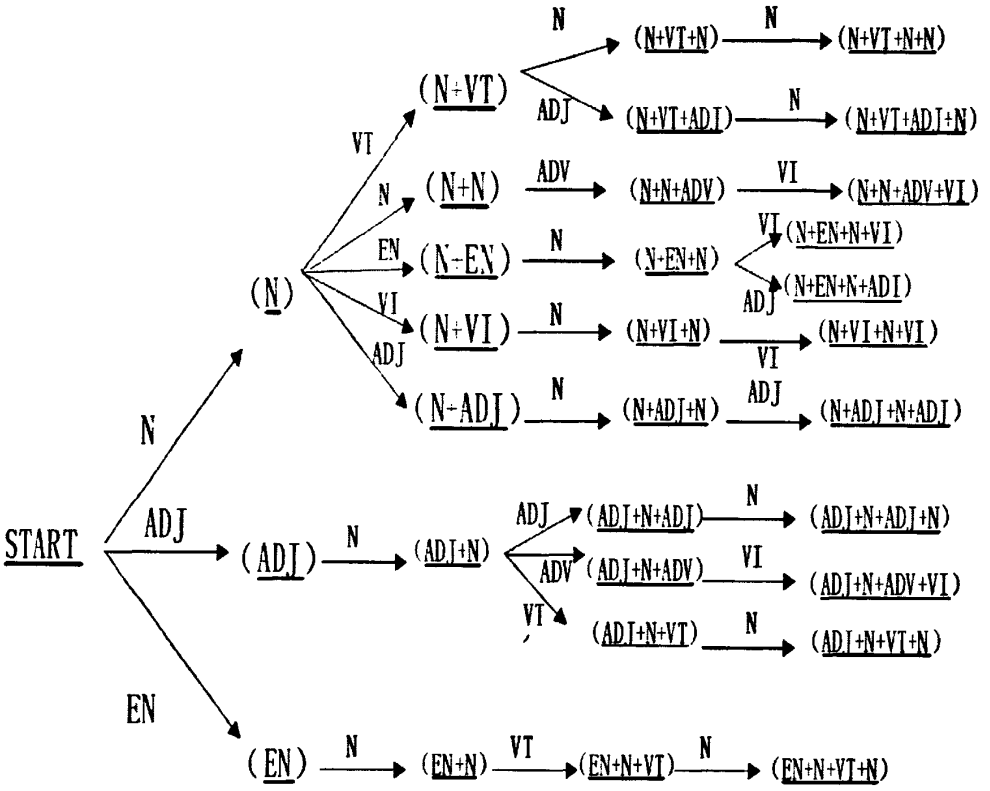

图 4-1.七字词句合法语法模式的 DFA 表示

Figure 4-1. VII explore issues that the legal syntax patterns DFA

4.2 词义相关度的计算

词的搭配不仅是个语法上的限制，也同样是个语义上的限制。计算词义相关，目的是建立词语间的关联，发掘词语共现和搭配的可能，从而保证生成诗词行文和主题上的连贯。本文中，词义相关度的计算主要应用于：

1）对于给定的主题词，计算与主题词相关度高的词

2）对于进化策略执行过程中产生的个体，计算词、句间的相关度与连贯性，作为适应性度量的一部分

3）作为启发式交叉算子和启发式变异算子操作过程中的一个度量

具体的应用方法将在第五章中进行介绍。本节中，我们详细描述采用潜在语义分析和互信息两种方法的基于语料库统计的词义相关度计算。

4.2.1 利用潜在语义分析计算词义相关度

利用潜在语义分析（LSA，Latent Semantic Analysis）计算词义相关度是基于这样的假设：如果给予大规模的文本语料库，词义相关的词语由于有一定的共现规律，一个词可以用一些有共现规律的词来代表它们的语义。

当文献量足够大，文献长度足够小，可以保证两个不同词义的词会出现在不同的文献中时，一个词在各种文献中出现或不出现的统计可以反映该词的词义。如果一个词在各种文献中出现多次，计算另一词在那些文献中每次出现的状况，就可以得到两词词义的相关程度。不同词所表现出的相关程度是不同的，因此，可以用词的搭配分布来表示词义[38]。

4.2.1.1. 准备工作

1、语料库的选择：选用《全唐诗》作为语料库，共包含 20162 首唐诗。

2、待测词的选择：从分词得到的词库中选取了 3000 个高频词，作为待测词。

3、待测文献的选择：为满足文献量足够大，文献长度足够小的要求，我们以一首词作为文献大小的度量，从《全唐诗》语料库中选取待测词出现相对频繁的 6000 首词作为待测文献。

4.2.1.2. 频率矩阵的构造

传统的向量空间模型中向量由词与邻近的同现词构成，在本文中，向量由词与文献构成，我们称这种方法为基于词文献空间模型方法。每个词义向量的分量就是经常和它一起同现的一些文献。一个词可以在多个文献中同现，则一个词义向量就是一个多维向量。参与计算词频的每一文献就构成了词义空间的一维。基于词文献空间模型方法的主要优点体现在由于采用距离方式计算向量之间的相关度，可以有效地避免传统概率统计方法中不可回避的数据稀疏问题，且由于避免了参数估计和特征获取等学习模型计算量过大的问题，词文献空间模型方法在保持高正确率的情况下，还具有简洁和高效的特点[40][41]。在构造频率矩阵时，我们将所有的待测词（t 个）都用在待测文献（d 句）中的出现频率表示出来，形成 t×d 的矩阵。

4.2.1.3. 奇异值分解

奇异值分解（SVD，Singular Value Decomposition）是一种与特征值分解，因子分析紧密相关的矩阵方法。LSA 方法用经过加权后的词－文献矩阵 x 做为输入，计算简化后的词、文献向量，将每一篇文献从基于词的向量空间表示映射到一个级数较低的正交因子张成的空间上.并由这些正交因子的线性组合去近似初始词文献矩阵。奇异值分解的理论是这样的：

定义：设 X 为 t×d 矩阵，X ‘X 的特征值是 λ 12， λ 22… λ n2 则 σ 1=| λ 1 |， σ 2=| λ 2 | … σ n=| λ n |，是 X 的奇异值。

定理：任何矩阵，例如一个 t×d 的词频矩阵 X，均可以被分解成三个矩阵的积，

$$\underset{t\times d}{X}=\underset{t\times r}{T_0}\,\underset{r\times d}{S_0}\,\underset{r\times d}{D_0'}$$

上式被称为 X 的奇异值分解。其中，T0’T0=T0’T0=I ，D0’TD0=D0’D0=I’，S0=diag[σ 1,… σ r]，是单值的对角矩阵，r 是 X 的秩， σ 1≥ σ 2≥… σ r>0。

奇异值分解的一个优点就是它允许用一种简单的策略，就是以一个规模较小的矩阵做初始矩阵的最优近似。选取一个合适的 k 值，保留 S0 中的前 k 个最大的奇异值，并保留 T0、D0 中相应的行和列，删去其余的，得到 T、D，则有，

$$\underset{t\times d}{X}\approx\underset{t\times d}{\tilde{X}}=\underset{t\times k}{T}\;\underset{k\times k}{S}\;\underset{k\times d}{D'}$$

是唯一一个秩为 k 的 X 的最小二乘意义上的近似矩阵。可以认为 $\tilde{X}$ 中包含了 X 的主要特征(亦即词与文献的主要语义关系)，并滤除了噪音。我们以奇异值分解推导出的简化模型数据近似初始矩阵中代表词－文献之间的关系的数据，由于其级数 k 远远小于系统中所使用的词数，次要的词与文献就被忽略了[42]。

4.2.1.4. 用 cosine 语义距离计算相关度

由于词 ti 可以表示为 ti = (n1,n2…n3)的向量，而两个词的相关度可以用

cosine 距离表示，cosine 距离又可以利用 $\tilde{X}$ 的两个相应行向量的点积来求得，由于 T、D 是正规矩阵，S 的对角元素大于零，有

$$\tilde{X}\tilde{X}' = TS(TS)' = TS^2T'$$

其中，$\tilde{X}\tilde{X}'$ 的第（i，j）个元素是词 t_it_j 向量的点积。

选择合适的 k 值是一个难点，k 值太小，则无法保留全部的总要结构，无法把握运算的结果；k 值太大，又可能引入噪音，而且会增加计算的时间复杂度。因此我们希望得到一个理想的折衷状态，使 k 足够大可以包括所有现实的结构信息，k 又足够小可以忽略掉取样错误和不重要的细节。实验中，我们取经验值 300。

4.2.2 利用互信息计算词义相关度

基于互信息（MI，mutual information）的算法在语料准备和预处理工作上与 LSA 算法一样，也是构建词文献空间[39]。如果 s 为文献（实验时为句子），w 为候选词语，Fs (w)是候选词 w 出现在句子 s 中的频率，Fi (w)是词语 w 在矩阵 i 列出现的频率，Fs (j)是 s 句子在 j 行出现的频率，

$$N = \sum_i \sum_j F_i(j)$$

是矩阵所有项的统计，θ 是为避免出现分母为 0 的而设的辅助数，暂设为 1。$mi_{w,s}$ 表示 w 和 s 之间的互信息（mutual information）

$$mi_{w,s} = \frac{\frac{F_s(w)}{N}}{\frac{\sum_i F_i(w)}{N} \times \frac{\sum_j F_s(j)}{N} + \theta}$$

为防止互信息在遇到词稀疏时的偏差，还引入纠偏因子（Pantel2002）：

$$\frac{F_s(w)}{F_s(w)+1}\times\frac{\min(\sum_i F_i(w),\sum_j F_s(j))}{\min(\sum_i F_i(w),\sum_j F_s(j))+1}$$

然后计算 cosine 词义相关度：

$$cos\_sim(w_i,w_j)=\frac{\sum_s mi_{w_i,s}\times mi_{w_j,s}}{\sqrt{\sum_s mi_{w_i,s}^2\times\sum_c mi_{w_j,s}^2}+\varepsilon}$$

得到词语间语义关系的度量，ε 是为避免出现分母为 0 的而设的辅助数。

4.2.3 计算结果的综合处理

在将 MI 与 LSA 两种算法的计算结果进行对比时，我们发现两次计算的词义相关度有区别，但计算出的相关词有相当多的重叠。因此，在计算结果的综合上，我们首先选取两种算法的重叠部分，相关度则用两者各占 50%的加权和表示。其次，对于不重叠的部分，我们按相关度从高到低进行排列，并保留相关度大于 10^{-3} 的词。

考虑到所选语料库的局限性，我们需要对生成的词义相关库进行评价。由于评价词语之间的词义相关度需要很庞大的测评体系，因此在条件有限的情况下，我们仅对生成的词义相关库作了适当的人工调整，删除了通过人为判断认为毫不相关的词。

4.3 词义相似度的计算

词语相似度主要用于衡量文本中词语的可替换程度。计算词义相似度，目的在于在保证所选词紧扣主题的前提下，尽量使生成诗词的语言更丰富多变。严格的格律要求，使得诗词创作有别于一般自然语言文本的生成，这一点在词的创作上体现得尤为明显。所谓“作诗”，意味着不仅要找到能准确地表达含义、传达感情的词，还要求所用的词严格遵守平仄和押韵等规则。在这种要求下，词汇量的丰富，同义替换的能力，甚至借代、象征、转喻等修辞手法的使

用，在很大程度上决定了诗词创作的质量。

对于现代汉语的自然语言处理中，词义相似度有两类常见的计算方法，一种是利用大规模的语料库进行统计，一种是根据某种世界知识（Ontology）来计算。

大规模语料统计的方法利用词语的相关性来计算词语的相似度[43]。该方法基于如下假设：凡是语义相近的词，它们的上下文也应该相似。具体做法是事先选择一组特征词，然后计算这一组特征词与每一个词的相关性（一般用这组词在实际的大规模语料中在该词的上下文中出现的频率来度量），于是，对于每一个词都可以得到一个相关性的特征词向量，然后利用这些向量之间的相似度（一般用向量的夹角余弦来计算）作为这两个词的相似度。

改进的模型来描写词汇的上下文语境信息。在给定的语料库Ω和词表δ中，特定词语x在Ω上的语义Sx定义为如下五元组

$$S_x = \{L_x, R_x, C_x, \delta, \Omega\}$$

其中：Lx为x的左同现词汇特征向量， Rx为x的右同现词汇特征向量，Cx为对仗词汇特征向量。特征向量的元素为特征词与特征值组成的二元组

$$V_{xy} = \frac{\log f(xy)}{\log f(x)\log f(y)}$$

其中:f(xy)为y在对应的x的相对位置上出现的频度(同一句的左边、右边或对仗位置上)。x， y∈δ，f(x)、f(y)分别是x、y在语料库Ω中出现的频度。

两个词之间的语义相似度Sim (x ,y)可以通过计算其在三个不同的词汇特征空间(Lx, Rx, Cx)中的距离来得到。距离越小，相似度越大。

$$Sim(x,y) = \frac{1}{k_1 \bullet \Delta L_{xy} + k_2 \bullet \Delta R_{xy} + k_3 \bullet \Delta C_{xy}}$$

其中k_1、k_2、k_3是可以根据语料库实际情况进行调整的加权参数。向量距离的计算公式为:

$$\Delta(x,y)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2}$$

在给定的语料库中，当 Sim (x ,y)超过特定的阈值 R 的时候，就定义这两个词 x， y 在该语料库中具有相似关系，x 的所有相似词组成的集合为 x 的相似词集 Lx。

胡俊峰、俞士汶的论文《唐宋诗中词汇语义相似度的统计分析及应用》[41]中，尝试将该方法用于古诗词的处理，提出用基于上下文的词汇向量空间模型来近似地描述词汇的语义。并将在此基础上定义的词汇相似关系和聚类关系应用于词典编纂、智能搜索引擎的开发等领域。

该方法的优点在于统计结果与所选择的语料库直接相关，因此也适用于古诗词的分析和处理；不足之处在于所得到的词汇相似关系中不仅包含了词汇的同义或近义关系，还包含有反义关系和一些其他相关关系。

根据世界知识（Ontology）计算词语语义距离的方法，一般借助于《同义词词林》和《知网》两个工具[44]。《同义词词林》将所有的词组织在一棵或几棵树状的层次结构中。由于在一棵树形图中，任何两个结点之间有且只有一条路径。于是，这条路径的长度就可以作为这两个概念的语义距离的一种度量。《知网》通过用一系列的义原，利用某种知识描述语言来描述一个概念，再由若干概念来描述一个词语。所有义原通过上下位关系组织成一个树状义原层次体系。计算中，先用最大值优先法将词语相似度转化为概念相似度，再用分类和求加权和法将概念相似度转化为义原相似度，最后通过求义原层次体系中路径长度来计算义原相似度[43]。

该方法的优点在于能得到较精确的同义关系；不足之处在于由于《同义词词林》和《知网》的编纂都是以现代汉语为基础，一些古诗词中常用的同义关系未能被收录。

本文的实验中综合使用了以上介绍的两种方法，考虑到计算的复杂性和词义相似度在应用中较强的针对性，我们将参与计算的词汇规模缩小为词库中收录的高频名词 545 个和形容词 367 个（如一组同义词均出现在高频范围内，仅计算其中频率较高的一个），并对结果进行了人工筛选（主要是删除反义和非

近义相关词）。事实上，经粗略验证，所选高频词及同义词已涵盖出现可能的70%以上。

表 4-1.词义相似计算结果节选

Table 4-1. Results similar to extract meaning

高频词	相似词
船	彩舫画舫小船轻舫舟小舫青舫游舫花舫船舫轻船画船艋船木兰船兰船兰舟扁舟彩舟轻舟渔舟孤舟
蝶	粉蝶蝴蝶双蝶翩翩
笑	欢笑巧笑喧笑盈盈轻笑娇笑笑谈笑语笑靥微笑浅笑笑语谈笑含笑
衣	衣襟衣巾罗衣衣冠乌衣红衣朝衣彩衣绣衣蓑衣锦衣征衣羽衣霞衣衣袖粗衣荷衣单衣舞衣襟袖香襟霞襟袖襟
寒	轻寒晓寒清寒凛凛余寒微寒冷飕飕冻

另外，以上两种方法的一个共同缺陷是只能考察词语的字面含义，而无法发掘通过借代、转喻、用典等修辞手法联系在一起的近义或同义词。而在古典诗词中，这几种修辞的使用又是较为常见的。例如：古诗词中，“高山流水”是一个常用的典故，出自《列子·汤问》：“伯牙鼓琴，志在登高山，钟子期曰：‘善哉，峨峨兮若泰山。‘志在流水，曰：‘善哉，洋洋兮若江河’。“该词常用于比喻知己或知音。也比喻乐曲高妙。《全唐诗》中，牟融的《写意二首》就有应用这个典故：

寂寥荒馆闭闲门，苔径阴阴屐少痕。
白发颠狂尘梦断，青毡泠落客心存。
高山流水琴三弄，明月清风酒一樽。
醉后曲肱林下卧，此生荣辱不须论。
萧萧华发满头生，深远蓬门倦送迎。
独喜冥心无外慕，自怜知命不求荣。
闲情欲赋思陶令。卧病何人问马卿。
林下贫居甘困守，尽教城市不知名。

如果能将用典、修辞等信息加以利用，不仅能丰富词汇的使用，更能增添

诗词的文采。第三章中在，我们提到用《全唐诗典故辞典》和罗凤珠老师的诗词典故资料数据库为基础，建立了典故数据库，但该数据库目前的应用还仅局限于分词，典故深层含义的整理和分析工作还在进行中。我们同时还开展了关于隐喻的研究，希望能设计出机器自动发掘隐喻含义的算法。研究进展会在今后的论文中详细阐述。

4.4 词的风格与情感意义标注

诗词作品所展现的风格和所传递的情感，是作者寄托于字里行间，而又超乎于字面含义的表达，是一首词的灵魂所在。对于计算机模拟诗词创作而言，要想达到不仅“形似”，而且“神似”的要求，就必须对风格的模拟和情感的表达进行深入的研究。

陈望道在《修辞学发凡》[46]中把风格分为四组八种：一组—由内容和形式的比例，分为简约，繁丰；二组—由气象的刚强与柔和，分为刚健，柔婉；三组—由于话里辞藻的多少，分为平淡，绚烂；四组—由于检点工夫的多少，分为谨严，疏放。由于该分法具有比较严格的量化定义，可操作性较强，因此适合用于计算诗学的研究。其中第二组的刚健与柔婉，亦即人们通常所说的豪放与婉约，是对唐诗风格最常用的一种评判标准，本文研究也定位于这种标准。

情感意义的赋予是诗词创作的一个重点，也是一个难点。一首好的诗词作品，一定能够准确地传递作者想要表达的情感。关于情感的分类，心理学界和工程界普遍认可所有情感均是在四种原始情感，即快乐、愤怒、恐惧、悲哀的基础上产生的。结合诗词情感表达的特点，本文中将情感模型简化为从快乐到悲哀的一个分级表达。

风格和情感都属于篇章语义，根据语言的组合性原则，应该通过篇以下层次（段、句、词）的有关语义的综合来获得。同样地，段依赖于句，句依赖于词。从语义上看，词是具有完整语义的最小语言单位，因此我们立足于词进行诗词风格和情感的研究[48]。

从词的语义结构分析，一个词包含形式、指称和体验三种语义（如图 4-2）。其中，形式语义包括词的音和形，是语言的物质部分。指称语义即词对应的客观事物，是语言的理性部分。体验语义是人对客观事物产生的体验，是语言的

感性部分，又分为意味体验和情感体验，前者是人对客观事物的形式产生的体验，对应于词的风格，后者则是人对客观事物的内容产生的体验，对应于词的情感[47]。

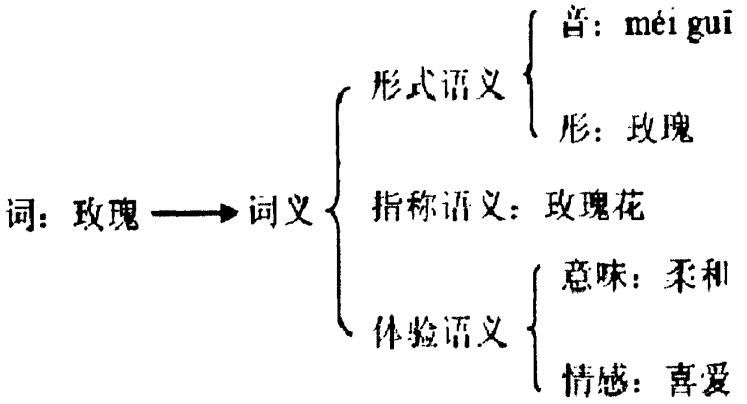

图 4-2.词的三种语义结构示例

Figure 4-2. Words of the three examples of semantic structure

事物的形式包括数量、形体、色彩、声音、味道、重量、力量、节奏、韵律、速度、质感等许多因素，能够给人丰富的意味体验。一般而言，具有数量少、形体小、色彩素、声音柔、味道淡、重量轻、速度慢等形式特征的事物更容易引起人的柔和意味，美学上往往称为“优美”，“指小巧、细腻、柔和的美”；具有数量多、形体大、色彩艳、声音粗、味道浓、重量沉、速度快等形式特征的事物更容易引起人的强烈意味，美学上往往称为“壮美”，“指巨大、粗犷、豪放的美”。作品宏观上的豪放与婉约风格决定于词汇所指事物微观上形式的量。豪放作品中主要包含具有强烈意味的词，婉约作品中主要包含具有柔和意味的词。

将词进行风格上的量化，本文参考李良炎在论文《基于词联接的诗词风格评价技术》的方法[47]，即成立专家组，通过讨论或其它互动方式一致将词集分为柔和、中性、强烈三个子集，然后递归地对各个子集进行相应的操作，递归的层次依精确需求而定。通过这种知识表示方式可以较好地反映人们普遍的审美规律。实验中将词集分为七个不同意味的子集，各词集中词的意味用数字来表示分为-3、-2、-1、0、+1、+2、+3 七种水平，分数越高，代表该词的风格越强烈。另外，由于词的意味是相对于词的指称语义而言的，因此多义词的

多种指称语义都有其对应的意味语义。

为实现度量上的统一，词的情感意义的量化也参照风格量化的标准，从悲哀到快乐分为（－3，3）的七个等级。

第五章 仿唐诗生成的进化策略

这一章中，我们将提出本文研究工作的核心思想，即将计算机自动生成诗歌看作一个状态空间搜索问题，并提出基于进化策略的模型加以解决。

我们以李白著名的七绝《望天门山》为例，将这首诗分词后得到：

天门/中断/楚江/开，碧水/东流/至此/回。

两岸/青山/相对/出，孤帆/一片/日边/来。

在词库里查找这些词的频率，发现每个词的频率都大于 2。这说明，除这首词外其它词的分词结果已经完全涵盖了这首词中使用的每一个词语。也就是说，这首词实质上是词库中某些词的一种排列组合形式。从这个角度出发，我们可以认为，诗词生成问题在本质上是一个解空间中寻求最优化的问题[18]，而解决这类问题正是进化策略的优势所在。

5.1 进化策略简介及适用性分析

20 世纪 60 年代，柏林工业大学的 I.Rechenberg 和 H.P.Schwefel 等在进行风洞实验时，由于设计中描述物体形状的参数难以用传统方法进行优化，因而利用生物变异的思想来随机改变参数值，获得了较好的结果。随后，他们对这种方法进行了深入的研究和发展，形成了一种新的进化计算方法—进化策略(Evolution Strategies,简称 ES)。

5.1.1 基本原理

进化策略中的个体用传统的十进制实型数表示，即：

$X^{t+1} = X^t + N(0, \sigma)$

X^t——第 t 代个体的数值，

$N(0,\sigma)$——服从正态分布的随机数，其均值为零，标准差为 σ。

因此，进化策略中的个体含有两个变量，为二元组<X, σ>。新个体的 X^{t+1} 是在旧个体 X^t 的基础上添加一个独立随机变量 $N(0, \sigma)$。假若新个体的适应度优于旧个体，则用新个体代替旧个体；否则，舍弃性能欠佳的新个体，重

新产生下一代新个体。在进化策略中，个体的这种进化方式称作突变。

5.1.2 两种进化策略

1975 年，H.P.Schwefel 首先提出(μ+λ)—ES，随后又提出(μ,λ)—ES。这两种进化策略都采用含有μ个个体的父代群体，并通过重组和突变产生λ个新个体。它们的差别仅仅在于下一代群体的组成上。(μ+λ)—ES 是在原有μ个个体及新产生的λ个新个体中共（μ+λ）个个体，再择优选择μ个个体作为下一代群体。(μ,λ)—ES 则是只在新产生的λ个新个体中择优选择μ个个体作为下一代群体，这时要求λ>μ。总之，在选择子代新个体时若需要根据父代个体的优劣进行取舍，则使用“+”记号，如(1+1)、(μ+1)及(μ+λ)；否则，改用逗号分隔，如(μ,λ)。近年来，(μ,λ)—ES 得到广泛的应用，这是由于这种进化策略使每个个体的寿命只有一代，更新进化很快，特别适合于目标函数有噪声干扰或优化程度明显受迭代次数影响的课题。本文所采用的进化策略也是(μ,λ)—ES。

5.1.3 进化策略的基本思想

随机产生一个适用于所给问题环境的初始种群，即搜索空间，种群中的每个个体为实数编码，计算每个个体的适应值；依据达尔文的进化原则，选择遗传算子(重组、突变等)对种群不断进行迭代优化，直到在某一代上找到最优解或近似最优解。

5.1.4 进化策略执行过程

（1） 确定问题的表达方式。这种表达式中个体出目标变量 X 和标准差σ两部分组成，每部分又可以有 n 个分量

即

$$(X,\ \sigma) = ((X_1,\ldots,X_n),\ (\sigma_1,\ldots,\sigma_n))$$

X 和σ的关系是：

$$\sigma_i' = \sigma_i \cdot \exp(r' \cdot N(0,1) + r \cdot N_i(0,1)) \qquad (1.1)$$

$$X_i' = X_i + \sigma_i \cdot N_i(0,1) \qquad (1.2)$$

式中：（X_i，σ_i）——父代个体的第 i 个分量；

（X_i'，σ_i'）——子代新个体的第 i 个分量；

$N(0,1)$——服从标准正态分布的随机数；

$N_i(0,1)$——针对第 i 个分量重新产生一次符合标准正态分布的随机数；

r'——全局系数，等于$(\sqrt{2\sqrt{n}})^{-1}$，常取 1；

r——局部系数，等于$(\sqrt{2n})^{-1}$，常取 1；

上式表明，新个体是在旧个体基础上随机变化而来。

（2） 随机生成初始群体，并计算其适应度。进化策略中的初始群体由 μ 个个体组成，每个个体的(X,σ)内又可以包含 n 个 X_i、σ_i 分量。产生初始个体的方法是随机生成。为便于和传统的方法比较，可以从某个初始点(X(0),σ(0))出发，通过多次突变产生 μ 个初始个体，该初始点可从可行域中用随机方法选取。初始个体的标准差 σ(0)=3.0。

（3） 计算初始个体的适应度，如若满足条件，终止；否则，往下进行。

（4） 根据进化策略，用下述操作产生新群体：

4.1)重组：将两个父代个体交换目标变量和标准差，产生新个体。一般目标变量采用离散重组，标准差采用中值重组。离散重组：对父代中两个个体实行随机交叉组合。中值重组：从 μ 个父代个体中用随机的方法任选两个个体，然后将父代个体各分量的平均值作为子代新个体的分量，构成的新个体。

4.2)突变：对重组后的个体添加随机量，按照式（1.1），式（1.2）产生新个体。

4.3)计算新个体适应度。

4.4)选择：按照(μ,λ)选择策略，挑选优良个体组成下一代群体。

（5） 反复执行第 4 步，直到达到终止条件，选择最佳个体作为进化策略的结果。将经典进化策略模型直接应用于诗词创作可能存在以下几个问题：

1）进化策略算法能够寻求最优解是建立在如下假设上：编码方案应该保证解空间表达的完备性，即个体的基因编码形式可以全面并且不重复的代表所有的解空间。这在实验中是不可能实现的，因为我们没有办法表述出诗歌句子的所有解。

2）适应度函数能够应用于计算（可计算性）。在大多数应用中，作为概率选择依据的适应度函数应该保证非负且可评估。而诗词作平好坏的评价是一个较为主观的过程，难以量化。

5.2 编码方案

对于诗歌生成问题，编码的选择是一个难点。一种最直接的方法是采用汉字本身的形式编码诗歌的句子，但对诗歌而言，这种编码方式会使得个体的编码串十分冗长，而且，适应度函数是与汉字编码形式密切相关的，也就是说，即使找到了一种新的编码方案，要把原有就很复杂的诗歌评定函数（等价于适应度函数）映射到新的编码空间，也是一个相当困难的操作。

考虑到唐诗的特点，我们提出了将“平、仄”与“0、1”编码相对应的编码方案。以卢纶《塞下曲》为例。

《塞下曲》　卢纶

林暗草惊风，将军夜引弓。

平明寻白羽，没在石棱中。

平仄规则如下：

⊙仄仄平平（韵），

平平仄仄平（韵）。

⊙平平仄仄，

⊙仄仄平平（韵）。

其中⊙表示可平可仄。我们用 0 表示平，用 1 表示仄，用通配符*表示⊙，得到如下编码串：

*1100（韵），

00110（韵）。

*0011，

*1100（韵）。

在实际操作中，为缩小问题的解空间，我们将分词模式固定为出现概率最

大一种

模式：

*1/1/00（韵），

00/1/10（韵）。

*0/0/11，

*1/10/0（韵）。

相应地，我们对词库中的单字词和双字词进行分类：单字词分为平、仄两类，对应编码 0、1；双字词分为平平、平仄、仄平、仄仄 4 类，对应编码 00、01、10、11。由于诗词创作不存在最优解，我们可以将全局最优问题转化为一个求某个局部空间域中的相对最优解问题。因此，不要求基因编码形式能覆盖所有的解空间，只要覆盖范围内存在相对较好的解，甚至是满足一定条件的可行解即可。至于覆盖范围内可能不存在可行解的问题，可以通过变异操作来解决。变异能不断的更新所覆盖的解空间，也就等价于不断的把所搜索的局部解空间进行推移。

5.3 初始种群的生成

考虑到唐诗严格的格律要求，在求解该优化问题过程中，我们始终将格律要求作为必须满足的约束条件。种群初始化的操作主要有一下几个步骤：

1）随机生成满足唐诗要求的韵部。如《塞下曲》要求压平韵，则随机生成一个平声韵部。

2）根据给定的主题词，从词库中挑选和主题词相关度大于 k1 的词，构成一级候选词空间。再从一级候选词中挑选相关度高的一部分词，再分别查找与这些词相关度高的词，构成二级候选词空间。重复类似操作，至候选词空间的词数量大于 n1。

3）从候选词空间随机选择满足押韵要求的词，首先填充每个需要押韵的位置，然后在满足平仄要求的基础上，随机选词填充剩余的位置。同此操作，生成含 N 个个体的的初始种群。

5.4 适应值函数

适应度函数的设计和进化策略中的选择操作直接相关，此外它还影响进化算法的迭代停止条件。适应度函数与问题约束条件。进化策略由于仅靠适应度来评估和引导搜索，所以求解问题所固有的约束条件不能明确表示出来。

遗传算法在优化搜索中基本上不用外部信息，仅用适应度函数为寻优依据。

适应度函数的设计应该尽量满足以下条件：

（1）单值、连续、非负、最大化

（2）合理、一致性。

（3）计算量小

（4）通用性强

本实验中个体适应性的评判主要依据以下 4 个指标：

（1）语法合法性 G：通过 DFA 检验的得分为 1，否则为 0。

（2）主题相关性 R：等于每个词与主题词的相关度之和。

（3）词句搭配的适当性 P：等于每两个连续词的相关度之和。

（4）风格和情感统一性 S：追求高的风格和情感统一性，就是要求同一首词中出现词汇的风格和情感得分都趋于一致。因此，S 等于所有词情感得分的方差与风格得分的方差之和。和越小，说明统一性越好。

考虑到 G、R、P 三个量要求取大，而 S 要求取小，故将 S 取其倒数，记为 S'。适应值函数 F 设计为以上 4 个量的加权和，在进行加权求和之前，还需将 4 个量进行规一化处理，即：

$$F=\lambda_1 G+\lambda_2 R+\lambda_3 P+\lambda_4 S'$$

5.5 选择操作

选择操作也叫复制操作，从群体中按个体的适应度函数值选择出较适应环境的个体。一般地说，选择将使适应度高的个体繁殖下一代的数目较多，而适应度较小的个体，繁殖下一代的数目较少，甚至被淘汰。考虑到诗词作品的优化是一个主观性较强的问题，目前尚无固定的、可量化的标准可以借鉴，我们

在选择这一操作中采用精英主义和轮盘赌算法相结合的模型。

所谓精英主义，是考虑到仅仅从产生的子代中选择基因去构造新的种群可能会丢失掉上一代种群中的很多信息。也就是说当利用交叉和变异产生新的一代时，我们有很大的可能把在某个中间步骤中得到的最优解丢失。精英主义方法在每一次产生新的一代时，首先把当前最优解原封不动的复制到新的一代中，其他选择步骤不变。这样任何时刻产生的一个最优解都可以存活到进化策略结束。

在保留了当前最优解后，我们采用轮盘赌算法完成对剩余个体的选择。令 Σfi 表示群体的适应度值之总和，fi 表示种群中第 i 个染色体的适应度值，它被选择的概率正好为其适应度值所占份额 fi / Σfi。表 3 给出了一个大小为 6 的种群中，每个个体的轮盘选择概率。该种群所有个体的适应值总和 Σfi=6650，则适应度为 2200 的个体被选择的可能为 fi / Σfi=2200/6650=0.331.

5.6 重组算子

进化策略中的重组(Recombination)算子相当于遗传算法的交叉，它们都是以两个父代个体为基础进行信息交换。进化策略中，重组方式主要有三种：

(1) 离散重组。先随机选择两个父代个体

$(X^1, \sigma^1)=((x_1^1,x_2^1,\cdots,x_n^1),(\sigma_1^1, \sigma_2^1,\cdots, \sigma_n^1))$

$(X^2, \sigma^2)=((x_1^2,x_2^2,\cdots,x_n^2),(\sigma_1^2, \sigma_2^2,\cdots, \sigma_n^2))$

然后将其分量进行随机交换，构成子代新个体的各个分量，从而得出如下新个体：

$(X, \sigma)=((x_1^{q1},x_2^{q2},\cdots,x_n^{qn}),(\sigma_1^{q1}, \sigma_2^{q2},\cdots, \sigma_n^{qn}))$

(2) 中值重组。这种重组方式也是先随机选择两个父代个体，然后将父代个体各分量的平均值作为子代新个体的分量，构成的新个体为：

$(X, \sigma)=((x_1^1+x_1^2)/2, (x_2^1+x_2^2)/2,\cdots, (x_n^1+x_n^2)/2),((\sigma_1^1+\sigma_1^2)/2, (\sigma_2^1+\sigma_2^2)/2,\cdots, (\sigma_n^1+\sigma_n^2)/2))$

这时，新个体的各个分量兼容两个父代个体信息，而在离散重组中则只含有某一个父代个体的因子。

(3)混杂(Panmictic)重组。这种重组方式的特点在于父代个体的选择上。混

杂重组时先随机选择一个固定的父代个体，然后针对子代个体每个分量再从父代群体中随机选择第二个父代个体。也就是说，第二个父代个体是经常变化的。至于父代两个个体的组合方式，既可以采用离散方式，也可以来用中值方式，甚至可以把中值重组中的 1/2 改为[0,1]之间的任一权值。

研究表明，进化策略采用重组后，明显增加算法的收敛速度。

Schwefel 建议，对于目标变量 X 宜用离散重组，对于策略因子 σ 及 α 宜用中值重组或混杂中值重组。

5.7 变异算子

变异操作是根据生物遗传中基因变异的原理，按一定概率，对个体编码串上的某个或某些基因位的值进行改变值。本文是对句子中的某个分词用从词库中随机选择的符合格式的分词加以代替。通过随机数确定相关度的大小，再从符合该相关度的词中随机选一个。所以需要两个随机数，前一个随机数是正态分布，后一个是平均分布。变异操作保证算法过程不会产生无法进化的单一群体。因为在所有的个体一样时，重组是无法产生新的个体的，这时只能靠变异产生新的个体。也就是说，变异增加了全局优化的特质。变异增加了进化策略找到接近最优解的能力。

本文研究中采用了启发式变异。操作步骤如下：

步骤 1：对于要进行变异的个体，比较每句的适应度，选出适应度值最小的句子。

步骤 2：若所选句不符合语法规范，找出与原句语法组合最接近的一种合法组合，利用词义相关，替换原句某个或某些词。例如：原句语法组合为 ADV＋N＋VT+N，判断该组合不合法，查找最接近的合法组合为 ADJ＋N＋VT+N，则从词库中选出与第一个 N 相似度最高，且在该个体中没有出现过的 ADJ 替换掉原来的 ADV。

步骤 3：否则，随机选取句中一个词 Wn，获取其邻位词 Wn-1 的词性 P，查找与 Wn 相关度最大且词性为 P 的词，替换 Wn-1（若 n＝1，改对 Wn+1 进行操作）

第六章 仿唐诗生成的系统实现与实验结果分析

我们的设计目标是构建一个仿唐诗计算机自动生成系统，它能按用户输入的关键词和词牌名自动生成仿唐诗。这一章中，我们将介绍系统的总体框架、生成模块的主要算法流程，实验中各相关参数的确定和检验以及系统的实现和运行情况。

6.1 系统框架

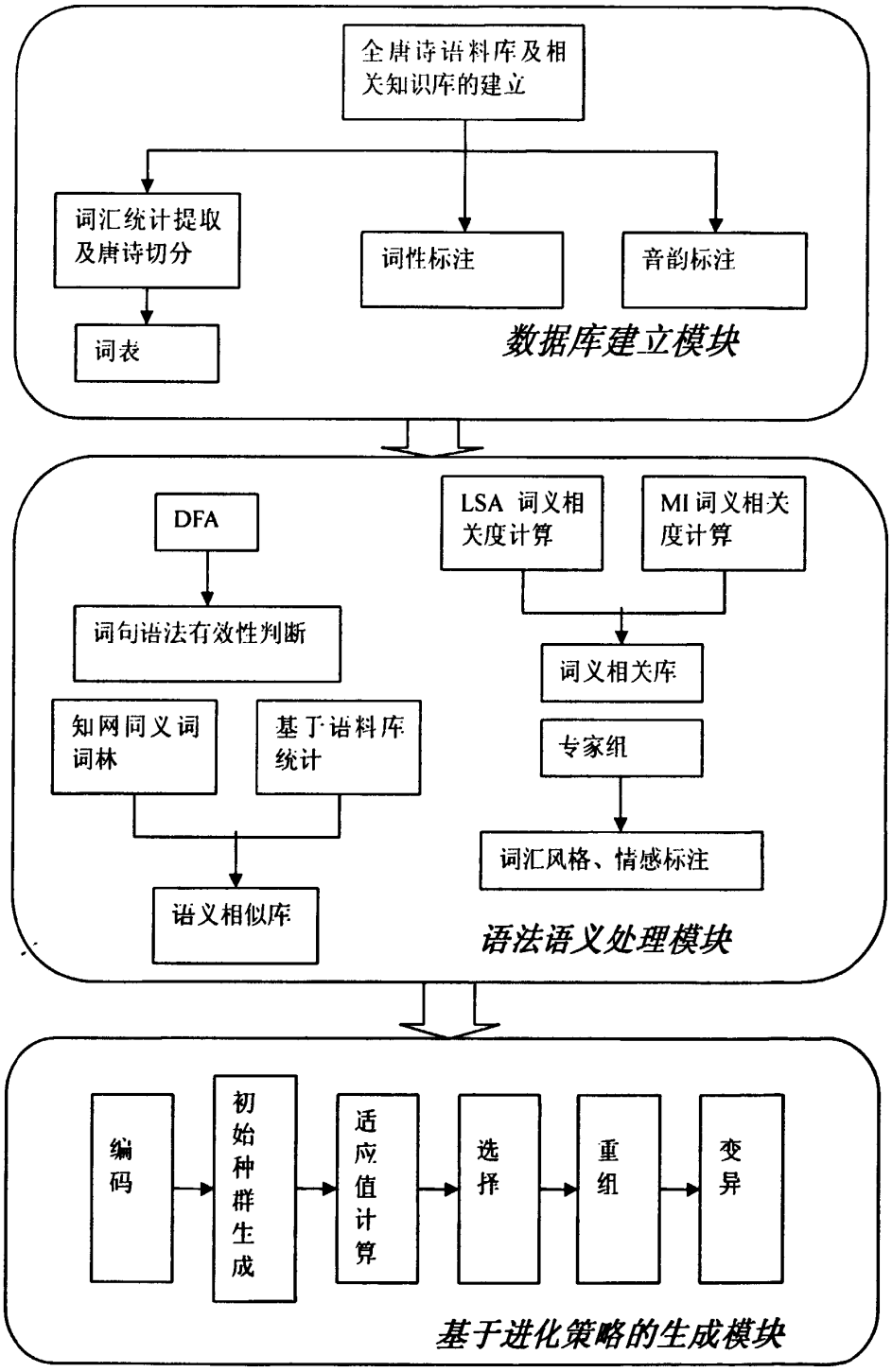

图 6-1.系统总体框架图

Figure 6-1. System framework map

系统总体框架如图 6-1 所示，共分数据库建立、语法语义处理、基于遗传算法的生成三个基本模块。前两个模块的具体计算过程和结果已在第三、四章中做了详细的介绍。本章中，主要进行基于进化策略的生成实验。

6.2 进化策略流程及主要参数确定

生成模块是基于进化策略构建的，算法主要由初始种群生成、适应值计算、选择、重组、变异五个主要步骤组成。算法主要流程如下：

生成初始种群，大小为 k1

置代数 gen=0，若 gen<k2 或进化停止，则反复执行以下操作

计算种群中各个个体的适应值；

将适应值最大的个体复制到子代；

置 n=0，循环次数 n<N/2，则反复执行以下操作

进行选择操作，选出两个父代个体；

产生一个随机概率 p

若 p<k3，则执行交叉操作，产生子代；

否则，保持，将父代复制到子代；

对新产生的子代执行下列操作：

计算子代的适应值；

置 m=0，若 m<k4，反复执行以下操作：

对子代执行变异操作，

若新的适应值比原来的小，将适应值置为

新的适应值，更新子代；

否则，以概率 k5 接受；

m++；

计算当前子代的适应值；

若适应值大于相应的父代

则将子代代替父代；

否则，以概率 k6 接受父代；

n++；

gen++;

算法的停止条件有两个：

（1）完成了预先给定的最大进化代数

（2）种群中的最优个体在连续若干代没有改进或平均适应度在连续若干代基本没有改进。

算法中 k1－k6 是可调参数。其中 k1 为种群大小，k3 为重组概率，k5 为变异概率，这三个是进化策略最主要的参数：

（1）种群大小太小时难以求出最优解，太大则增长收敛时间。一般取值在 30－200 之间，本实验中将种群大小定为 100。

（2）重组概率始终控制着进化策略中起主导地位的重组算子，太小时难以向前搜索，太大则容易破坏高适应值的结构。一般取值在 0.3-0.9 之间。本实验中取值为 0.8。

（3）变异概率太小时难以产生新的基因结构，太大则会使遗传算法变成单纯的随机搜索。一般取值范围在 0.001-0.2 之间，本实验取值为 0.15。

另外三个参数中，k2 为设定的最大进化代数，取经验值 5000。k4 为变异操作次数，取值 3000，k6 为父代接受概率，取值 0.3。

6.3 系统的实现与运行情况

6.3.1 开发和运行平台

我们选择 VisualC++6.0 作为系统程序开发工具。根据设计目标,系统程序运行的硬件平台确定为普通 PC 机，运行于 Windows 平台。实验中，我们使用的测试机器基本参数为：CPU 1.83GHz，内存 512 MB。

6.3.2 系统生成实例

系统运行时要求用户输入 1－3 个关键词，并选择一种体裁。（作为初步研究的测试系统，我们仅支持五言绝句和七言绝句）

以下我们用一个生成实例来介绍系统的运行情况。在这个实例中，用户输入的主题关键词为“菊”，体裁为五言绝句。

系统首先提取主题关键词“菊”，在词义相似和相关库中进行查找（如表4），

表 6-1.“菊”的词义相似和词义相关计算结果

Table 6-1. "Chrysanthemum" meaning similar and related calculations Meaning

词义相似计算结果	黄菊 紫菊 嫩菊 槛菊 兰菊 菊花 金菊 菊蕊 野菊 松菊 晚菊 庭菊 细菊 篱菊 赏菊 丛菊 新菊 菊香 白菊	
词义相关计算结果（节选）	一级相关	轻寒 登高 秋色 重阳 晓寒 离恨 雁黄 管弦 香秋 晚秋 微雨 萧疏 零乱 凄然 黯淡 凄楚 憔悴 萦绊 愁颜 梦 影 夜 西风 零落 幽怨 微凉 斜日 馨香 鸿雁 金 祝寿 紫 中秋 新酿 东篱 高歌 醉 残 良辰 庭院
	二级相关	情 舞 携手 竟 金尊 忆 轻轻 朱阑 难忘 红烛 朦胧 寒 烛影 无端 明镜 雁 梧桐 燕 吹 扁舟 故国 潇湘 残荷 露 叠翠 晨星 浩渺 清泪 回首 遥看 人间 笙歌 共舞 冷艳 长亭 相逢 双桨 红颜 暮云 吟 幽愫

而后，系统根据随机生成一个平声韵部“黄”。规定每个个体中至少出现一个与主题词的词义相似词。

生成的初始种群个体举例如下：

霜菊花萎日，悲风陨秋凉。雅望与英姿，零落移新暖。

经过选择、重组、变异等操作，系统最后生成的结果为：

北风携夜雨，东篱独凄凉。待得更漏尽，垂露看朝阳。

分析生成结果，可以看出

（1）在音韵方面，该词很好地满足了五言绝句的平仄、押韵等要求。

（2）在语法方面，没有出现明显的语法错误，语句通顺。

（3）在语义表达方面，我们很高兴地看到了词义相关计算的潜力。由关键词“菊”联想到北风、夜雨、露珠和朝阳，可以看出，有了词义相关的保证，全词具有较好的主题一致性和叙述连贯性。

（4）在风格和情感表达方面，整首诗前面低沉，后面高昂，“北风、夜雨、独、凄凉”等词的使用营造了一种孤苦悲凉的气氛，有明显的婉约派风格，但后面的“垂露、看、朝阳”等词则勾画了一幅含泪笑看希望的画面，是豪放派的风格。两种意境相互对立，造成矛盾反差，使整首诗的意境得到升华，符合绝句起承转合的特点。

6.4 系统性能评价及结果分析

为评价系统的性能，我们对系统进行了运行速度和生成结果满意度两方面的测试：

1、生成仿唐诗的耗时。系统的运行速度是衡量系统性能的一个重要指标，进化策略需要一定时间的运行才能使算法达到收敛。实验中，我们对参数进行一定的调整，尽量把运行时间控制在用户可接受的范围内；

2、用户对所生成仿唐诗的满意率。这是衡量该系统性能的最重要指标，但该项测试具有较大的主观性。这一测试分 3 个指标进行评测：主题相关度评判、风格情感一致性评判和总体质量评判。评判专家组由 5 名中文系本科生组成，评判采用 5 分制。

对 50 次生成实验进行测评，结果如表 5 所示：

表 6-2.系统性能测评结果

Table 6-2. System performance evaluation results

生成仿唐诗的平均耗时	用户对所生成仿唐诗的满意率（5 分制）		
20 分 46 秒	主题相关度	风格情感一致性	总体满意度
	4.25	3.82	4.05

上述实验结果表明，该系统基本实现了自动生成仿唐诗目的，生成作品的质量大部分是可接受的，偶尔也有较为出色的诗作生成。但系统的运行效率和风格情感计算方面还有待改进。

结论与未来展望

本文对仿唐诗的计算机自动生成进行了初步研究。问题的可行与否取决于能否把诗歌需要的规律量化并且输入计算机，并且是否具有可接受的计算代价(例如数据库大小，程序的复杂程度，程序的相应时间等等)。

在对机器作诗的现有方法进行总结和分析的基础上，本文提出了基于进化策略的仿唐诗生成模型。借鉴汉语古诗词计算语言学研究在词汇语义分析方面已取得的成果，建立唐诗切分和音韵标注语料库。针对古诗词与现代汉语的区别，提出了基于 DFA 的语法判定规范，以及基于词义相似度、词义相关度、词汇风格和情感特征的语义度量。根据唐诗特点，设计了编码，适应度计算、选择、重组、变异等进化操作的具体实现算法，并构建系统加以实现。

用进化策略求解诗歌生成问题有以下优势：

1．进化策略从问题解的串集开始搜索，而不是从单个解开始。这是进化策略与传统优化算法的极大区别。传统优化算法是从单个初始值迭代求最优解的；容易误入局部最优解。进化策略从串集开始搜索，覆盖面大，利于全局择优。

2．进化策略求解时使用特定问题的信息较少，容易形成通用算法程序。由于进化策略使用适应值这一信息进行搜索，并不需要问题导数等与问题直接相关的信息。进化策略只需适应值和串编码等通用信息，故几乎可处理任何问题。

3．进化策略有极强的容错能力。进化策略的初始串集本身就带有大量与最优解甚远的信息；通过选择、交叉、变异操作能迅速排除与最优解相差极大的串；这是一个强烈的滤波过程，并且是一个并行滤波机制。故而，进化策略有很高的容错能力。

4．进化策略中的选择、重组和变异都是随机操作，而不是确定的精确规则。这说明进化策略是采用随机方法进行最优解搜索，选择体现了向最优解迫近，重组体现了最优解的产生，变异体现了全局最优解的覆盖。

实验结果证明了进化策略模型的有效性和较好的通用性。研究工作虽然取得了一定的进展，但在实验过程中也突现出一些问题。

1、在语法合法性的判定方面，现代汉语已有较成熟的判定方法，如严格的规则过滤和句法分析。但对于诗歌，尤其是古典诗词，由于其用语高度凝练，多使用典雅的“诗家语”等特殊性，很难找到可供借鉴的现成语法和标准化评判规则。实验中采用的基于 DFA 的判断，是基于对大量实例的总结。但这样的总结难以覆盖所有可能的合法组合：一则，大量实例不等同于所有实例；二则，没有出现过的组合并非都是不合法的组合。因此，实验中有出现误判的可能性。这需要我们进行更全面的统计，或者设法将已经较为完善的现代汉语语法规则加以利用和改造，使其适用于诗词的语法判定。

2、词语搭配准确性的判断是语义度量最重要的一个环节。随着计算机运算能力的大大提高，基于语料库的计算语言研究受到了很大的关注。它为解决语言处理问题先建立统计模型，并由训练数据（语料库）来估计统计模型中的参数。采用语料库频率统计的方法可以细致的刻画搭配关系，但它也存在固有的缺陷，例如在某词汇出现稀疏的情况下，统计出的词义相关可能难以全面反映该词汇的潜在语义[49]。因此，本实验中，我们仅对词库中的部分高频词进行了词义相关的统计，这导致在适应度计算中部分词义相关信息的缺失，从而影响了计算的精度。在今后的研究中，我们将尝试对不同的语料库进行更全面的统计，减少词汇出现稀疏的可能性，使计算出的词义相关度信息更加精确和完整。

3、对候选诗作的语法和语义评判多以词和句为单位，对于句间的逻辑组织考虑不够。篇章内容的合理安排是人类逻辑思维能力的重要体现，也是计算机模拟的一个难点。句序安排的不当会使篇章缺乏严整的逻辑结构，这个问题在适应值函数有效性检验实验中得以反映。在将人类诗作以词为片段拆分打乱后，适应值判定函数在句间词序的组织上体现出了优势，但在句序的安排上则暴露了评判准则的不足。因此，将句间逻辑关系规则化并量化，使其成为评判准则的一部分是我们今后研究的一个重点。

4、在对词语进行风格和情感判定时，我们对每个词语给出了一个固定的评判值。但事实上，一个词语体现的风格和情感意义并不总是唯一的，准确的分析必须在特定的语境中进行。在今后的研究中，我们将尝试将判定过程动态化，即在获得人工的预判值后，在诗词生成的过程中根据上下文语境动态修正预判值。考虑到动态判定的计算开销，设计高效的算法是解决这一问题的关键。

5、生成诗词作品的质量不够稳定。这主要是由于评判标准难以全面和量化。诗词的鉴赏历来是个“仁者见仁，智者见智”的过程，而且大多局限于文学领域，要形成一套系统的，可供计算机量化操作的评判标准相当困难。不全面的评判标准可能导致某些有缺陷的个体在适应值计算中依然获得高分，从而使算法过早收敛到一个局部最优解。这个问题可以说是实验过程中遇到的最突出，也最棘手的问题。其解决有待于我们及各相关领域专家的共同努力。

6、作为初步探索，我们仅对唐诗的绝句进行了具体的算法设计和实现。虽然我们建立的计算模型具有一定的通用性，但要使模型的移植性更好，生成形式更为多样的诗句，仍需要进行更深入的研究和改进。

在诗词生成的研究中，我们更想表达一种人工智能模拟的创造力。展望我们的工作，还可以在多方面多层次展开，如改进算法，增强系统的自学习能力；完善诗词自动评价体系；建立相关的修辞数据库，如用典的收集和理解，使生成的诗歌更有文采；创作种类更为丰富的诗词作品等。相信这些研究会推动诗歌生成的发展起到积极的推动作用。

参考文献

[1] 张建华，陈家骏.自然语言生成综述[J].计算机应用研究，2006，8
Daniel Jurafsky,James H.Martin 著.冯志伟孙乐译.自然语言处理综论[M].北京：电子工业出版社，2005.471－491

[2] 张冬茉，葛永，姚天昉.多语种自然语言生成系统中的预映射句子规划器.计算机研究与发展，2004，34（7）：467－474

[3] Bailey,R.W.Computer-assisted poetry:the writing machine is for everybody.[M]InMitchell,J.L.,editor,Computers in the Humanities,pages 283–295.EdinburghUniversity Press,Edinburgh,UK.1974

[4] Gerv′ as,P.Exploring quantitative evaluations of the creativity of automatic poets[C]. InProceedings of the 2nd. Workshop on Creative Systems,Approaches to Creativity inArtificial Intelligence and Cognitive Science,15th European Conference on ArtificialIntelligence(ECAI 2002),Lyon,France.2002

[5] van Mechelen,M.V.Computer poetry[EB/OL]. http://www.trinp.org/Poet/ComP/ComPoe.HTM.1992

[6] Hartman,C.O.Virtual Muse:Experiments in Computer Poetry[M].Wesleyan University Press.1996

[7] Boden,M.A.The Creative Mind:Myths and Mechanisms[M].Weidenfeld and Nicolson,London,UK.1990

[8] Gerv′ as,P.An expert system for the composition of formal spanish poetry.Journal ofKnowledge-Based Systems,2001.14(3-4):181–188.

[9] Kurzweil,R.Ray kurzweil ‘s cybernetic poet[EB/OL]. http://www.kurzweilcyberart.com/poetry.2001.

[10] Rubaud,J.,Lussonnal,P.,and Braffort,P.ALAMO:Atelier de litt′ erature assist′ e parla math′ ematique et les ordinateurs[EB/OL]. http://indy.culture.fr/alamo/rialt/pagaccalam.html.2000.

[11] Kempe,V.,Levy,R.,and Graci,C.Neural networks as fitness evaluators in genetic algorithms:Simulating human creativity[C].In Moore,J.D.and Stenning,K.,editors,Proceedings of the 23rd Annual Conference of the

Cognitive Science Society,Edinburgh,UK.2001.

[12] Gruber,H.and Davis,S.Inching our way up mount olympus:The evolving systemsapproach to creative thinking[M].In Sternberg,R.J.,editor,The Nature of Creativity,.Cambridge University Press, New York,USA.1988.pages 243–26952

[13] Sims,K.Artificial evolution for computer graphics[J].Computer Graphics, 25(4):319–328.1991

[14] Diaz-Agudo,B.,Gerv′ as,P.,and Gonz′ alez-Calero,P.Poetry generation in COLIBRI[C].In Proceedings of the 6th European Conference on Case Based Reasoning(ECCBR 2002),Aberdeen,UK.2002.

[15] Luger,G.F.and Stubblefield,W.A.Artificial Intelligence:Structures and Strategiesfor Complex Problem Solving.Addison Wesley Longman,Inc.,third edition.1998.

[16] Aamodt, A.and Plaza, E.Case-based reasoning: Foundational issues,methodologicalvariations, and system approaches. AI Communications,7(1):39–59.1994.

[17] Hisar Maruli Manurung.An evolutionary algorithm approach to poetry generation[D]University of Edinburgh,2003

[18] 刘岩斌,俞士汶,孙钦善.古诗研究的计算机支持环境的实现[J].中文信息学报，1996，11（1）：27－35

[19] 穗志方,俞士汶,罗凤珠.宋代名家诗自动注音研究及系统实现[J].中文信息学报，1998，2：44-53

[20] 罗凤珠，李元萍，曹伟政.中国古代诗词格律自动检索与教学系统[J]，中文信息学报 1999.1:35-42

[21] 胡俊峰.基于词汇语义分析的唐宋诗计算机辅助深层研究[D]北京大学博士学位论文北京 2001.5

[22] 易勇.计算机辅助诗词创作中的风格辨析及联语应对研究[D]重庆大学博士学位论文重庆 2005.6

[23] 李良炎.基于词联接的自然语言处理技术及其应用研究.[D].重庆大学博士学位论文.重庆 2004.10

[24] 周昌乐.心脑计算举要[M].北京：清华大学出版社，2003

[25] 费越.汉语语义的多层次集成研究—及春联艺术系统设计[D].中国科学

院自动化研究所博士学位论文.北京.1999.7

[26] 周明.微软对联生成系统[EB/OL].微软亚洲研究院自然语言组.北京 http://duilian.msra.cn/2006

[27] 俞士汶,胡俊峰.唐宋诗之词汇自动分析及应用[J].语言暨语言学，2000，4(3)：631-647

[28] 罗凤珠.诗词语言切分与语意分类标记之系统设计及应用[C].第四届数位典藏技术研讨会，2005

[29] 张忠纲.全唐诗大辞典[M]. 语文出版社,2000

[30] 范之麟 吴庚舜. 全唐诗典故辞典[M]. 崇文书局,2001

[31] 徐青. 唐诗格律通论[M]. 当代中国出版社,2002

[32] 周勋初.唐诗大辞典[M]. 江苏古籍出版社,2003

[33] 曹寅、彭定求.全唐诗[M].清

[34] 俞士汶,段惠明.北京大学现代汉语语料库基本加工规范[J].中文信息学报,2002,16(5):49-64

[35] 俞士汶,段惠明.北京大学现代汉语语料库基本加工规范(续)[J].中文信息学报，2002,16(6):58-64

[36] 游维.基于语料库的汉语隐喻生成研究[D].厦门大学学士学位论文.厦门 2004

[37] Pantel,Patrick.Lin,Dekang.Discovering Word Senses from Text[C].In Proceedings ofACM SIGKDD Conference on Knowledge Discovery and Data Mining.Edmonton,Canada.2002,613-619

[38] Church,Kenneth W.,Hanks,P.Word Association norms,MutualInformation and Lexicography[J]. Computational Linguistics,1990,Vol.16(1).

[39] 鲁松,白硕,黄雄.基于向量空间中义项词语的无导词义消歧[J].软件学报，2002，Vol.13(6):1082-1089

[40] 鲁松,白硕,黄雄等.基于向量空间模型的有导词义消歧[J].计算机研究与发展，2001，Vol.38(6)

[41] 张威.汉语语篇理解中元指代和隐喻的机器消解研究[D].浙江大学博士学位论文，浙江 2003

[42] 胡俊峰,俞士汶.唐宋诗中词汇语义相似度的统计分析及应用[J].中文信息学报.2002，16(4)：39－44

[43] 刘群,李素建.基于《知网》的词汇语义相似度计算[C].第三届中文词汇语义学研讨会论文集.http://www.keenage.com/html/paper.html.

[44] 董振东.汉语知识词典及词汇内部语义描述研究.语言文字应用[J].2000.1

[45] 陈望道.修辞学发凡[M].上海:上海教育出版社,2001.264

[46] 李良炎何中市易勇.基于词联接的诗词风格评价技术.中文信息学报[J].2005.19(6)：98－104

[47] 应英,周峰,周昌乐.汉语情感意义的机器标注研究初探[J].中文信息学报2002.2

[48] Christopher D.Manning.Hinrich Schutze 著，苑春法李庆中等译统计自然语言处理基础[M].北京：电子工业出版社，2005.182－188

攻读学位期间发表论文

曹卫华，张灵，刘军. 模糊自动机的一种新模型. 现代计算机，2011

致谢

在论文完成之际，我要特别感谢我的指导老师张灵老师的热情关怀和悉心指导。在我撰写论文的过程中，张老师倾注了大量的心血和汗水，无论是在论文的选题、构思和资料的收集方面，还是在论文的研究方法以及成文定稿方面，我都得到了张老师悉心细致的教诲和无私的帮助，特别是她广博的学识、深厚的学术素养、严谨的治学精神和一丝不苟的工作作风使我终生受益，在此表示真诚地感谢和深深的谢意。

在论文的写作过程中，也得到了许多同学的宝贵建议，同时还到许多在工作过程中许多同事的支持和帮助，在此一并致以诚挚的谢意。感谢所有关心、支持、帮助过我的良师益友。最后，向在百忙中抽出时间对本文进行评审并提出宝贵意见的各位专家表示衷心地感谢。