北京航空航天大学软件学院

工程实践结题报告

课程名称：古诗词生成器的设计与实现

姓 名：王雨清 陈立元 吴迪

组 别：

学 号：

专业方向：人工智能研究方向

所属院系：北京航空航天大学软件学院

指导老师：

实验起止时间：

目录

[1 项目概述 1](#_Toc587218793)

[1.1 项目来源 1](#_Toc1698895986)

[1.2 项目意义 1](#_Toc402266190)

[2 项目需求分析 1](#_Toc609334574)

[2.1 总体需求 1](#_Toc1884156322)

[2.2 重点需求说明 1](#_Toc221445192)

[3 项目总体方案 1](#_Toc240181693)

[3.1 项目总体结构 1](#_Toc1611941538)

[3.2 关键技术及解决方案 1](#_Toc1395222261)

[4 项目测试(实验验证)和效果分析 2](#_Toc1126599034)

[4.1 测试(实验验证)情况说明 2](#_Toc386648839)

[4.2 运行效果分析 2](#_Toc121521251)

[5 个人工作总结 2](#_Toc150717260)

[5.1 吴迪个人总结 2](#_Toc1221769007)

[5.2 陈立元个人总结 2](#_Toc33068035)

[5.3 王雨清个人总结 2](#_Toc1723683319)

1 项目概述

1.1 项目来源

自然语言处理是教会机器如何去处理或者读懂人类语言的系统，目前比较热门的方向，Long Short Term Memory (LSTM)是循环神经网络的改进，在自然语言处理方面有明显的优势，在当前卷积神经网络流行的时期依然担当者重要角色。

中国是诗歌的王国,五千年的优秀文化熔铸了不计其数的脍炙人口的优秀诗篇。许多诗流传千古,哺育了一代代人的成长,是我国极其珍贵的精神财富。研究用自然语言处理技术进行诗歌自动生成，可以帮助普通人进行诗歌创作以及激发创作灵感。但是早起期的中国传统诗歌研究，主要集中在艺术和语言分析领域，直到20世纪90年代才开始有中国传统诗歌自动生成的相关研究。最近几件，以深度学习为代表的人工智能技术取得了越来越多的突破，将人工智能技术用于诗歌自动生成引起了越来越多研究者的注意。本人主要是通过LSTM深度学习技术进行自动生成藏头诗和五/七言律诗的研究。

国外在诗歌自动生成的相关研究开始的比较早，积累了不少方法技术。国内直到20世纪90年代才有中国传统诗歌自动生成的相关研究，起步比较晚，因此还有很多提高和进步的空间。诗歌自动生成不仅可以帮助普通人进行诗歌创作，而且也能以此为契机探索计算机进行艺术创作的可能，因为这方面的研究很有价值和必要。

本研究属于自然语言生成（Natural Language Generation,NLG）领域，它是人工智能在自然语言处理领域的重要应用。Long Short Term Memory (LSTM)是循环神经网络的改进，在自然语言处理方面有明显的优势，在当前卷积神经网络流行的时期依然担当者重要角色[2]。自然语言生成主要以计算语言学和人工智能为基础，研究人类创作自然语言文本的过程[3]。早起的NLG系统一般采用基于模板（Template-Based）的方法来进行文本生成。它的主要原理是先用基于正则表达式的模式匹配技术进行模板匹配，再将模板中的槽位进行替换以得到最终的句子。基于模板的方法由于技术简单并且生成结果非常可靠而得到广泛引用，但是它依然有很多局限性：一方面需要人工来进行繁琐的模板涉及，另一方面人工涉及的模板缺乏灵活性，不能适应复杂的生成情况。为了解决这些问题，人们开始探索其他的NLG方法，以更好的应用不同领域、不同情况下的文本生成任务。

随后有研究者提出基于规则（Rule-Based）的方法[4]。它主要是以树结构来进行文本表达，相比基于模板的系统，可以覆盖更多的自然语言场景和情况，但是基于规则的房费仍然需要人工来决定哪些统计特征可以用于优化目标。整个系统的表现受限于事先定义的语法规则，而在新的应用领域中，仍然需要人工来添加规则。

随着NLP技术的不断发展，基于短语（Corpus-Based）的方法[5]逐渐引起了人们的注意。它定义了一个灵活的学习结构并通过生成模块和重排序模块直接从数据中进行学习，它的最终结果通过对生成器的候选结果进行重排序得到。直接从数据中学习的方法似系统生成的结果更加自然，不然需要提前定义规则并使系统更容易跨领域应用。

近几年有学者提出PPG（Planning based Poetry Generation Model）引入到自动生成诗歌。它包含诗歌规划和诗歌生成两个部分。诗歌规划部分可以利用诗歌语料和外部知识，提前规划处由关键词组成的写作大纲。诗歌生成部分使用改进的基于注意力机制的编码器-解码器，可以同时编码关键词和历史生成内容来主句生成诗歌。

1.2 项目意义

人与人之间的日常交流和沟通一定离不开自然语言以及各种符号表达类语言。因此语言在日常的生活,学习以及工作等活动中有着不可或缺的作用。另一方面,计算机已经开始在多个领域逐步的替代或者简化一些工作,目的就是可以减轻人类的负担。所以人们也在思考计算机能否在自然语言处理与表达上达到模仿人类,或者超过人类的语言表达水平,从而起到辅助人类的作用。

诗歌是一种特有的表达内心情感的文学体裁。通过使用高度凝练的语言、跳

跃的语句结构以及一定的节奏和韵律,从而集中反映出当时的社会生活以及表达出个人思想感情,是人类智慧和文明发展在语言表达层面的优秀展现。中国素来享有诗歌之国之称,诗歌文化也是源远流长,但是总结之前工作,主要是从人类进行文学创作的方向上研究,直到20世纪90年代中期才逐渐着眼于从计算语言学角度对进行研究山。因为诗歌创作有着一定的难度并且对创作者的自身文化功底有一定基础要求,所以一直以来诗歌创作就是文人墨客的专属,也是人类智慧的体现。然而如今计算机技术的逐步前进发展,人工智能产品应用越来越广,其正在成为一种潮流,特别是在自然语言处理方面取得的巨大进展,让人们不禁开始思考是否可以让计算机自动或者辅助创作诗歌。

深度学习技术旨在模仿人脑的机制,从数据中学习岀高层抽象的分布式表示。近些年深度学习技术在语音,图像和自然语言处理领域都取得了很大的突破,探索将深度学习技术应用到诗歌生成中有很大的研究价值。我们希望借助深度学习技术,设计出一套比较完善的诗歌生成模型,让计算机具有人一样的表达和写作能力，即能够根据一些关键信息以及在机器中的表达形式，经过一个规划的过程，从而自动生成高质量的藏头诗和五言/七言古诗。它可以为广大诗歌文学爱好者提供很大便利，同时也有助于汉语言的传播与宣扬。

2 项目需求分析

古诗生成系统支持用户录入关键字，通过后台搭建的LSTM训练出的网络生成藏头诗和古诗。

2.1 总体需求 //TODO

给出研究开发项目功能需求的描述，最好能采用一些需求分解图、算法原理、应用方式 进行说明

2.2 系统前端功能实现

用户可以根据类型进行选择藏头诗、五言还是七言古诗，并且录入需求诗歌的关键词如江城美景，

若已选中藏头诗：

江焰红花里，风经雨起烟。

城西深夜后，叶满不胜经。

美洁漏将受，出门临碧池。

景间陪待罢，佳景尽依依。

若已选中五言诗：

花偏君亦长，一别少看花。

项小黔州路，天边山已深。

渡长淮河上，月夜南山分

情世不可识，歌枝一少愁。

若已选七言诗：

彼处闻寒溜泉频，竹峰蛛网木浓阴。

先生成性思成远，白发幽人事肯醒。

惟有月圆心便寝，起经徒到意无身。

跂襟藉笋丛青菊，声价同亲奈欲何。

2.3 后台功能实现

后台实现的主要功能是针对前端传入的类型进行相应的处理生成对应的诗歌。生成的模型是提前训练好的。因此后台主要包括模型训练和生成实例诗歌两部分，其中模型训练是整个系统的实现重点。

本系统输入的样本数据是一共包含34646首的全唐诗，但是原始数据并不能直接满足训练需求，是需要进行数据预处理。数据预处理的过程，实际上是文本处理的过程，先对诗歌进行拆分，提取诗歌内容。步骤包括数据清洗、统计词频、生成字库、根据字库把每一首诗转化为一个向量。

根据已输入的诗歌类型和关键字，再根据训练模型存入的Model,生成最终的预测诗歌。

3 项目总体方案

本节是对人工智能软件内容设计文档的综述，说明该项目所涉及的总体设计方案。

3.1 项目总体结构

说明人工智能软件结构设计、算法原理

3.2 关键技术及解决方案

对系统的原理及其关键算法的解决方案，或系统的关键技术和重要模块的设计与实现。

4 项目测试(实验验证)和效果分析

本节是对软件系统实现内容测试情况的说明，并介绍软件实际运行效果。

4.1 测试(实验验证)情况说明

介绍针对项目所作的测试情况:实施了哪些测试、通过测试发现的问题已经如何解决。

4.2 运行效果分析

介绍该项目实际的运行情况，可以截取 1~2 个界面有针对性地进行说明。

5 个人工作总结

针对项目组的每一个成员通过一小节来说明该成员在该项目中的工作情况。

5.1 吴迪个人总结

说明个人在项目中所承担的角色，主要负责完成的工作，以及对该项目的体会。

5.2 陈立元个人总结

说明个人在项目中所承担的角色，主要负责完成的工作，以及对该项目的体会。

5.3 王雨清个人总结

说明个人在项目中所承担的角色，主要负责完成的工作，以及对该项目的体会。