古诗生成系统的设计与实现

摘要

诗歌具有很高的文学艺术价值，是人类智慧的结晶。中国传统诗歌的自动生成是自然语言处理领域中一项非常有挑战性的工作[1]。本文采用两层的LSTM网络，设计了基于LSTM网络的藏头诗自动生成框架，文中详细介绍LSTM网络的结构以及原理，对数据预处理过程分为数据清洗、统计词频、生成字库、根据字库把每一首诗转化为一个向量。对于采用的one-hot编码原理，优缺点文中做了细致的分析。从最后的生成结果看，生成的藏头诗和古诗基本符合唐诗的形式，但在诗歌意境方面还有很大的提升空间。并针对此模型设计了前端用户交互系统，无需了解内部设计便可直观地获取藏头诗及五/七言诗词的生成结果。

关键词： 古诗 藏头诗 五言诗 LSTM网络

**第1章 绪论**

1.1 课题背景与意义

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。

1.2 国内外研究现状及对比分析

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。

1.3 研究目标及研究内容

本研究目的是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

希望模型可以做到以下几个要求：

* 给出一个关键词，就能生成不同风格的诗歌。
* 通过无监督的方式实现这个功能。
* 生成的诗在其他指标上尽量减少损失或者没有损失（流畅性、通顺性……）

**第二章 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。

2.2 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。

参考文献：

[1] 王哲. 基于深度学习技术的中国传统诗歌生成方法研究[D]. 2017.

[2] 姜力. 基于深度学习的散文诗自动生成方法[D].

[3] 张冬茉, 葛永, 姚天. 多语种自然语言生成系统中的预映射句子规划器[J]. 计算机研究与发展, 2001, 38(4).

[4] Guzzoni D, Baur C, Cheyer A. Modeling Human-Agent Interaction with Active Ontologies[C]//AAAI Spring Symposium: Interaction Challenges for Intelligent Assistants. 2007: 52-59.

[5]

Oh A H, Rudnicky A I. Stochastic language generation for spoken dialogue systems[C]//ANLP-NAACL 2000 Workshop: Conversational Systems. 2000.