七言绝句生成项目

陈思元　郭大卫　王乐安　庄敏学　谷雨

联系邮箱：×××××××

指导教师：邓志鸿

北京大学，人工智能引论课程
2019-2020，春季学期

摘要

古体诗是中国国粹，而其比较工整的格式有利于用神经网络学习。通过本次项目，我们以练习为根本目的，也希望借助神经网络生成像样的古诗。

本项目基于 *Chinese Poetry Generation with Planning based Neural Network* 这一篇论文[1]，在继承论文核心模型基础上做出小幅度改动，完成给出主题句生成一首七绝的功能。

项目地址：
https://github.com/Guyutongxue/ChinesePoetryGeneration

关键词： 古诗，神经网络

引言

基于已有的研究成果，我们的生成过程分为规划器（Planner）和生成器（Generator）两部分，两者互相连接形成了以自然语言的主题句作为输入，输出四句七言诗歌的模型。其中：

规划器负责通过主题句和已知的信息推断每行诗句的关键词；

生成器负责根据关键词和已生成的诗句进行选词推断，最终生成整首诗。

这种方法能够较好地提取主题信息，并保证了诗歌的连贯性和语义一致性。

方法

第一部分：数据处理

原始数据为分好词的各朝七绝古诗，共12520首。随后基于TextRank方法提取每句关键字，按 *context-keyword-label* 格式生成训练数据。

第二部分：规划器

生成训练数据：对于已经完成分词的诗句，使用上述关键词表得到每一句的关键词，由此得到以四个关键词为一组的plan_data

关键词向量化：把四个关键词看成一个句子，使用gensim进行向量化，保存模型。由此可得知这些关键词彼此之间的相似性

诗句关键词生成：对给出的主题句，按jieba分词，并根据关键词表提取关键词。根据之前训练的模型，补齐四个关键字，作为后续诗句的关键词。

第三部分：生成器

功能：根据已经生成诗句、关键字生成下一句

模型：应用注意力（Attention）的seq2seq机制

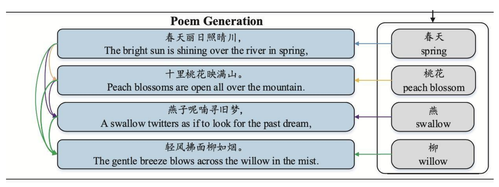
Encoder：关键字和已生成句子的双向GRU

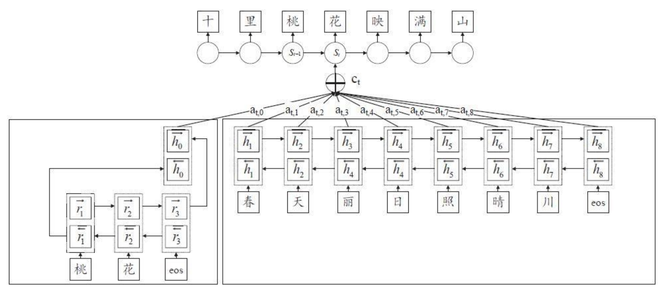
Decoder：注意力机制下的GRU

双向GRU使用*bidirectional_dynamic_rnn*，

注意力机制采用*seq2seq.BahdanauAttention*。

训练：以关键字和已生成句子作为输入， 待生成句子为标签





其它

在押韵处理上，采用较为朴素的方法。当生成到二、四句的最后一个字时，判断概率分布中，每个字与首句第七个字的韵母是否相同。如不同，则对该概率乘以一个较小的权重 k。

平仄处理时，为尽可能保留语义效果，并未严格把控粘、对等平仄要求。仅在生成每句的三、五字时，判断与一、三字平仄是否相同；同时，二、四句的末尾应为平声。如与需求不符，则乘以权重 m。

实验

下面展示实验效果。

输　入： 春天桃花开了
关键词： 唤　乐　人间　故人
结　果：
前欲禁中紫公归，
酒入留客乐春晖。
主人卧对孤臣士，
千年西湖鸟鹊谁。

输　入： 春花秋月
关键词： 秋月　花　月　闲
结　果：
北阙寒花一已无，
云开青处思画图。
空阶月明流觞客，
俯仰纤纤燕子波。

输　入： 云雨江南
关键词： 江南　南　雨　云雨
结　果：
江南村舍须宜诗，
吹送人物艳此时。
哭庙儒酸春梦雨，
明日牵牛已相持。

总结

- 细节决定成败，神经网络
- 深度学习模型在强大算力支撑下有具备强大拟合潜能；但通过实践也感受到它的机械性、不确定性，和训练的笨拙性
- 小组合作的初次尝试，承担、组织、互助、包容的宝贵训练。从素不相识、不敢互相麻烦到打开自己、团结合作、成为伙伴，感到程序员间的人情味。
- 付出不一定收获效果，但在编程能力、对深度学习模型的理解程度上有实打实的提升

参考文献

[1] Z. Wang, W. He, H. Wu, H. Wu, W. Li, H. Wang, and E. Chen, “Chinese poetry generation with planning based neural network,” in Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 1051–1060.

·AIPKU·