一种自由化的汉字文本生成方法

**一、背景**

**1.1 自动文本生成机**

自动文本生成被认为是当机器智能的重要重要标志，因为人们在创作文字的时候需要极强的想象力，机器必须能够"学习”文章的行文方式并“模拟”人脑的创作能力，因此极为困难。本发明提出一种独特的自动生成文本的方法，与传统方法相比，该方法更灵活，创造力更强。

传统的文本生成主要有两种方式：

1，统计机器翻译（SMT）概率模型

该方法将文字创作看作由前一句对后一句的翻译过程，逐步生成整个需要的文章。该方法的一个缺点是下一句话的生成只依赖于前一句的信息，无法保证生成文本的完整性。同时，SMT模型一般基于字的贡献信息，所以在语义的方面建模能力较差，往往只被应用在语义信息相等或近似的情况下。

2，神经网络(NN)模型

该方法将用户所给的信息（首句）通过神经网络压缩成信息向量，将该信息向量作为初始状态，通过神经网络逐句生成整个文本。

**NN/SMT**

特征

规则

词表

**白日依山尽**

**黄河入海流...**

**图1: 文本生成框架**

**二、问题描述**

传统的文本生成方法存在一个显著缺点：高度依赖于文本中的特定信息，即对输入信息的考虑不周全。传统SMT模型仅仅考虑了字的信息，在语义信息上考虑十分不全，导致生成的文本语义混乱。而传统的RNN模型在训练过程中会变得十分注重语义信息，且在生成每一个字的时候都只考虑同样的全局信息，这使得生成的文本生成文本单一且十分易出错，不能够根据已生成的信息进行调整使得容易产生主题偏差。

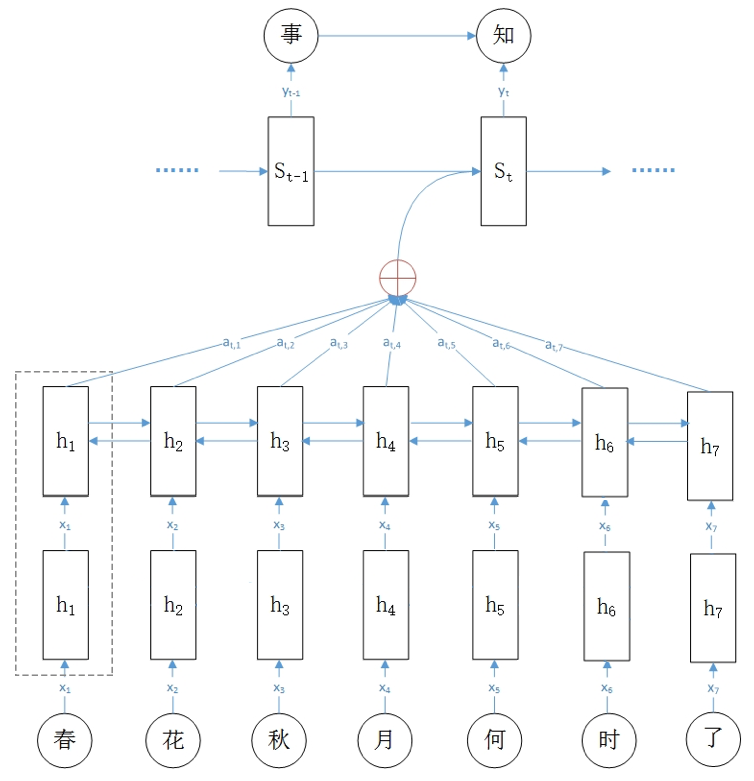
本发明将一种新的汉字文本生成方法。该方法将整个文本看成一个包含断句符号的完整字串（而非若干句子），利用递归神经网络(RNN)对这一字串进行学习，得到文本创作的行文风格，在训练RNN时同时训练字的信息和字的语义信息。在生成时，RNN模型生成候选字，再加入人为设定的规则使生成需要的格式，并且根据生成的信息来重新筛选输入信息，再次生成下一个字。这一方法的优点是：

1. 可以利用多种格式的文本学习这一模型，解决了数据稀疏的问题；
2. 可以利用该模型生成任意长度，任意多个句子的文本；
3. 每一句话与前面所有已经生成的句子相关联，保证整个文本的一致性；
4. 生成过程中不断进行修正，保证了生成效果的鲁棒性。

**三、发明要点**

**3.1 基于RNN的文本生成模型**

本发明基于RNN网络构建文本生成模型，如图1所示。首先，该方法将用户给予的信息（如图1中的“春花秋月何时了”作为首句），经过一个双向RNN网络编码成一组向量（图1下部矩形列），该向量作为用户意图的编码。量。在生成过程中（图1的上部），一个单向RNN网络不断循环运行，生成文本的中的每一个字。在生成每一个字的时候，对用户的意图向量进行查看，找到与当前生成状态最相关的用户意图进行下一字的生成。在生成过程中，强制加入断句、押韵、平仄等限制文本要遵守的限制，这样就保证了生成的“字串”既能最大程度地符合语法和语义规则（RNN的生成符合语法和语义规则），又保证了生成的信息紧紧围绕用户的意图展开。



**图1：基于RNN的文本生成模型**

**3. 2 多种结构文本的灵活生成**

图1所示的模型结构可以生成任意一种格式的文本，只要改变生成过程中的结构限制即可。如当限制为每句5个字后必须有一个断句符时，即是五言诗，当限制为每句7个字后必须有一个断句符时，即是七言诗。只要不停止，可以生成任意多句文本格式，而且所有句字都围绕用户意图生成，不会发生主题发散。

**3.3多种语言风格的文本生成**

基于本发明所示出模型的通用性，可以对该模型用任何语料进行学习，从而得到不同语言风格的文字模型。如图2所示，训练模型所用的语料可以采用现代文，而规定的规则是古文结构，可以生成具有现代风格的古文，这样就实现了大数据的充分利用和语言风格的学习

**红河入大流**

训练模型

特征

规则

词表

党的光辉

**红彤彤...**

**训练好的模型**

特征

规则

词表

**白日依山尽**

**图2. 多种风格文本生成**

**3.5多样化文本生成：**

传统方法随机生成文本的时候一般通过取次优候选句的方式，生成的文本质量下降。基于本发明提出的模型，我们通过对用户输入信息进行随机化以产生不同的生成信息，但是在生成的时候，模型反复从输入信息中筛选重要的信息。这一方法既保证了随机和多样性，同时不影响生成文本信息的综合打分。一种简单的方法是对用户输入进行扩充和乱序。如图3所示，我们在用户输入后加入若干和用户输入相关或相似的词，使得生成的文本在紧扣用户所给的主题的同时，生成不同的文本。

**临水人家深宅院**

**RNN模型**

**梦入蟾光...**

**家深宅院邻水人（田园）**

**RNN模型**

**金缕歌声...**

**图3. 随机文本生成**

**四、方案优势**

* 本发明利用RNN模型，可以在不改变模型结构情况下，通过改变限制条件，生成各种体文本，且生成文本与主题（用户输入的信息）紧密相关。
* 本发明可学习不同语料，从而生成不同语言风格的文本。
* 本发明可在不降低性能的前提下，生成多种与主题紧密相关的文本。