Seq2Seq 歌词生成

对于自然语言生成任务，除了一些比较高级的学术和商业用途，网络上有很多有趣的小工具，如 Shakespeare sonnet generator，藏头诗生成器，小红书文案生成器等。根据 Seq2Seq 模型，使用 PyTorch 和提前训练好的 Word2Vec，实现中文歌词输入上句生成下句，输出用于生成再下一句，如此循环往复得到一首歌。

```
encoder输入: ['今天', '是', '我', '仅', '有', '的', '时间']
decoder输入: ['<go>', '只', '想', '让', '你', '看见', '<eos>']
target目标: ['只', '想', '让', '你', '看见', '<eos>']
预测结果: ['就', '怪', '给', '我', '走进', '<eos>']


encoder输入: ['繁华', '的', '夜']
decoder输入: ['<go>', '失眠', '的', '街', '<eos>']
target目标: ['失眠', '的', '街', '<eos>']
预测结果: ['失眠', '的', '街', '<eos>']


encoder输入: ['爱', '终', '有', '一天', '会', '拯救', '我', '所有', '恐惧']
decoder输入: ['<go>', '就', '像', '你', '对', '我', '期待', '的', '一样', '<eos>']
target目标: ['就', '像', '你', '对', '我', '期待', '的', '一样', '<eos>']
预测结果: ['就', '像', '你', '对', '我', '期待', '的', '一样', '<eos>']
```

使用来自 40 余位华语歌手、乐队、组合的 7000 多首歌曲进行预训练，得到通用模型。注意到这几个现象：首先，通用模型很难看出具体化用了哪位歌手的哪首歌，风格比较杂糅；其次，不管出什么题目让它接，绕不开情情爱爱，可见华语乐坛情歌很多；最后，虽然语句还算通顺，单看相邻的两句不能说毫不相干，但总是写着写着就跑题了，最后整首歌看起来跟开始出的题目不怎么相干。

夏夜晚风

是谁的侧脸
火在我的窗
漆黑的夜里
漂在跟随
走过多少回忆
挥着我的手
你的出现
我捕捉
精采的画面
你的关怀
我的心我的心
我的心你的心
我的爱总不能把你的心
为什么你的心我的心
为什么你的心我的心
却偏偏藏在世界里面

一场游戏一场梦

还没学会
你说过的再见
我们说过的想法
不知道谁会说得太多
但我不太容易
我的心不懂
曾在你的面前
在漫漫夜里一分泪
望着你的一切渺渺
怎能忘记这段情
你的吻
你的心
在我的心里
偶尔争执的机会
我当然
我愿意

在微调阶段，选择收入训练集歌曲最多的歌手（邓丽君，531 首）和歌曲最少的乐队组合（果味 VC，41 首），加载预训练的通用模型，用这两位的歌词进一步训练，使生成的歌词偏向她们的风格，可以部分看出化用自哪一首歌曲。

邓丽君风格的歌词生成：

小城故事

小城故事多
充满喜和乐
若是你到小城来
收获特别多
看似一幅画
听像一首歌
人生境界快乐总是在一起
要去孤独就把它遗忘
多少次还会有谁的春天
谁能了解我
只有那春风知道我
知道我寂寞
我寂寞
谁人
谁像那悲欢离合
别流泪

这里以邓丽君真实存在的一首歌为题，可见前几句完全复现了《小城故事》原本的歌词，后半段（从“人生境界”开始）则融合了《我要对你说》。

漫步人生路

在你身边路虽远
末疲倦
伴你漫行
一段接一段
越过高峰
另一个阳光
叫我去到海边
是那美丽的流水
盼望你
想你
想你
想你
天上白云有几颗
有谁能了解我
只有时光能找回甜蜜
请你别把我的心相遇

这首歌的前几句也复刻了《漫步人生路》，后半段（从“越过高峰”往后）出处比较模糊，“只有时光能找回甜蜜”这一句出自《留下我自己》。

轻声细语

我问过彩霞
何处是你的家
爱情
若我说我爱你
这就是欺骗了你
如果我不能够
给我一吻
但愿这是我
同我孤独相系
但愿你共我
接受我的美梦
的心印在心里
不会问我
爱情的春天

这首是随机命题，开头两句出自《彩霞回来吧》，接下来几句出自《爱情》。

果味 VC 风格的歌词生成：

陌生的眼睛

我向往的眼睛
我的光让你照亮
给我一个时间
等待你出现
你一定会看见
我们的路
当我们在等待中沉没
直到会有一天
你会相信
如果你需要
微笑属于我
我感到无数彩虹
可忽然看见
我的眼睛
无力隐藏
无力抵抗

可以看出歌词部分融合了《罗素广场》和《模糊》。

最后一盏灯

你说你将去
享受浮云的气息
甜蜜的奇遇
留下清晰的足迹
你宛若独立的城
超音速列车中的光
你的我看
你说这夜色
就像你的影子
你看这世界
有一天你会出现
时间会停在你的
所有的眼泪
还在
一个人
你指尖在一起

可以看出歌词前半段融合了《超音速列车》。

直到世界尽头

你我
你说你梦见过此刻
甜蜜让你迷惑
午后三点阳光洒落
沉默被你打破
我开始让你看
在日落之前
如果看不到你眼睛
我的今天
从未开始失去意义
日出之前
让我看见
却无法看清
你是空中复杂的旅行
你不会去看我的
时间会在尽头

歌词前半段融合了《午后三点》，后半段融合了《木星三号》。

总的来说，通过 Seq2Seq 训练一个模型来根据上句预测下句，循环得到像歌词这样逻辑性不强（毕竟真人写的词有时也会很意识流）的文本是可行的，而且如果某个歌手的歌曲数量不多，也可以先加入其他歌手的歌曲进行预训练，再对得到的通用模型进行微调，从而一定程度上实现对该歌手写作风格的模仿。

然而，由于每句歌词都只由紧邻的上一句决定，不存在“点题”，很难保证整首歌围绕一个固定的主题展开。其他许多让 AI 生成小说、剧本（如续写《哈利波特》、《蝙蝠侠》）的尝试也遇到了类似的问题，就是写着写着会“忘记”上文的剧情，导致故事情节不怎么连贯。要想在生成的过程中对上文保持更长期的“记忆”，从而实现主题的统一和连续，这样简单地通过上句生成下句是不够的。