This work aims to translate sentences in one language into the different-style sentences in another language.

I use the Chinese-English translation with the English poetry style as a case study.

In this case, I use the attention-based NMT system to conduct the translation task and an RNN language model to do the style transfer task. I will denote the translation model as NMT and the language model as LM.

Merge function

$$p(y) = p_{NMT}(y) + \beta * p_{LM}(y)$$

Dataset

NMT: 1M news-domain sentence pairs

> 江泽民 对 杜苏率团 来访 表示 欢迎 .

> jiang zemin extended his welcome to tozzoli for leading the delegation to visit china .

> 那些 自焚 惨剧 的 参与者 受到 李洪志 蒙蔽 , 毒害 , 控制 , 最后 却 成为 李洪志 妄图 达到 罪恶 目的 的 " 炮灰 " 和 牺牲品

> the participants of the self - immolation tragedy have been deceived , poisoned , and controlled by li hongzhi and , in the end , served as his cannon fodder and victims in his vain attempt to achieve his criminal purpose .

LM: 230,000+ sentences after preprocessing 55,000+ English modern poems

> we've shared so much , we two , |^ a multitude of joyous things . |^ years of wild and wonderful passions , |^ laughter , tears . . .

> goodbyes that shook the heavens , |^ only to embrace again with a love |^ so profound we must

> have been the envy |^ of even ancient lovers .

> cut off the |^ phone |^ turned the |^ music

Experiments

The number of test sentence is 747. Here is a table showing the number of influenced sentence under different settings.

β	**# influenced sentence/# total sentence**
0	0/747
0.1	0/747
0.2	0/747
0.5	2/747
1.0	4/747
100.0	199/747
1000.0	607/747

- β = 0, 0.1, 0.2: BLEU = 31.44
- β = 0.5: BLEU = 31.42

  examples:

  > 英国 媒体 今天 报导 , 席哈克 与 英国 首相 布莱尔 明天 将 在 法国 北部 le touquet 晤谈 , 就 伊拉克 危机 的 最新 发展 彼此 交换 意见 。
  >
  > **w/o style**: british newspapers and british prime minister blair today vowed to exchange views with the british parliament on the latest development of the iraq crisis .
  >
  > **w/ style**: british newspapers and british prime minister blair today vowed to exchange views with the british parliament on the latest development of a crisis in iraq .
  >
  > 加拿大 太空署 主任 , 加拿大 首 位 宇航员 加诺 表示 , 美方 对 " 哥伦比亚 " 号 事件 的 的 调查 可能 使 国际 空间站 已 定下 的 任务 受阻 , 有些 计划 可能 会 延期 执行 甚至 取消 。
  >
  > **w/o style**: canada 's chief administrator of canada 's first astronaut , canada 's first astronaut , said that the investigation conducted by the us side on the " el salvador incident " may cause the implementation of the international space station , and some may be postponed or even cancelled .
  >
  > **w/ style**: canada 's chief administrator of canada 's first astronaut , canada 's first astronaut , said that the investigation on the " el salvador incident " may affect the missions set by the international space station , and some may be postponed or even cancelled .

- β = 1.0: BLEU = 31.43
- β = 100.0: BLEU = 31.27

  examples:

  > 他 同时 宣布 , 格 强力 部门 近期 还 将 会 继续 在 潘基西 峡谷 展开 清剿 非法 武装 的 行动 。

> w/o style: he also announced that the turkish government would continue to carry out anti - illegal armed operations in the southern part of _UNK .
>
> w/ style: he also announced that the turkish government would continue to carry out illegal armed operations in the _UNK valley in the near future .
>
> 届时 , 游客 可以 搭乘 埃及 航空 公司 的 班机 由 北京 起飞 , 在 开罗 游玩 0 天后 , 再 飞往 马耳他
>
> w/o style: by that time , tourists can take part in the flight of the egyptian airlines flight from beijing for a couple of days to fly back to malta .
>
> w/ style: by then , tourists can take part in the flight of the air services of egypt to fly from beijing to _UNK , and then fly to malta .

- β = 1000.0: BLEU = 29.09

  examples:

  > 去年 巴西 国家 风险 指数 曾 超过 0000 点 , 债券 价格 下降 到 面值 的 00% 。
  >
  > w/o style: last year , the brazilian national league index fell more than 0,000 points , and the price drop was down
  >
  > w/ style: last year , the |^ index exceeded 0,000 points , down from 00 percent to 00 percent .
  >
  > 近来 外国 投资者 认为 巴西 债券 价格 便宜 , 回报率 高 , 因而 需求 大幅 增加 , 使 巴西 国家 风险 指数 下降 加快 。
  >
  > w/o style: foreign investors have expressed their interest rates and the volume of crude oils has risen sharply in recent years .
  >
  > w/ style: in recent years , investors have expressed interest in _UNK , _UNK , _UNK , _UNK , _UNK , _UNK , _UNK , _UNK , _UNK , _UNK

Analysis

The main problem is that NMT dataset and LM dataset don't have many shared prefix or shared sentence segment. So LM can't generate confident result to influence the final result. That's why there aren't many influenced lines when beta is small (0.1, 0.2, 0.5, and 1.0).

As the β increases large enough, the poem phase separator "|^" and "_UNK" symbols come up (because news dataset has too many domain-specific words that LM data doesn't have?)

NEXT:

1. Try to renormalize the softmax of the LM without the "_UNK" symbols to alleviate the too-many-UNK problem.
2. Handle the problem of less shared prefix.