中国中文信息学会自然语言生成与智能写作专业委员会（筹）

评测任务二：基于大纲的条件故事生成

Transfer and Denoising Learning For Story Generation

Bin Li[1], Minjun Zhu[2], Fei Xia[2], Yixuan Weng[2]

[1]College of Electrical and Information Engineering, Hunan University
[2]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy Sciences

Reporter: Bin Li

May 1, 2022

Contents

Problem Introduction

Outline-based Story Generation:

Given:

The outline of the story

Goal:

To generate the story according to the outline, where story shall be

- **Accurate**
- **Smooth**
- **Logical**

标题：
"运神像的驴子"

输入大纲：
"对着神像顶礼膜拜", "再也不肯往前走", "神像放在驴子", "赶着进城", "驴夫狠狠", "洋洋得意", "大喊大叫", "遇见"

输出故事：
有个人把神像放在驴子背上，赶着进城。凡是遇见他们的人都对着神像顶礼膜拜。驴子以为人们是向它致敬，便洋洋得意，大喊大叫，再也不肯往前走了。结果挨了驴夫狠狠的一棍。

Figure 1: An illustration of outline-based story generation

Dataset Introduction

The dataset of the outline-based story generation is presented, which is shown in Table 1. The dataset is divided into training (1456), development (242), and testing (729) according to the data set.

Item	Training Set	Development Set	Test Set
Number	1,456	242	729
Vocab Size	19,320	5,698	12,488
Outline Number	19.20	19.05	19.47
Averaged Outlines	8.00	8.00	8.00
Averaged Words in Story	108.91	108.68	109.04
Averaged Sentences in Story	7.20	7.11	7.15

Main Method

RAKE Training

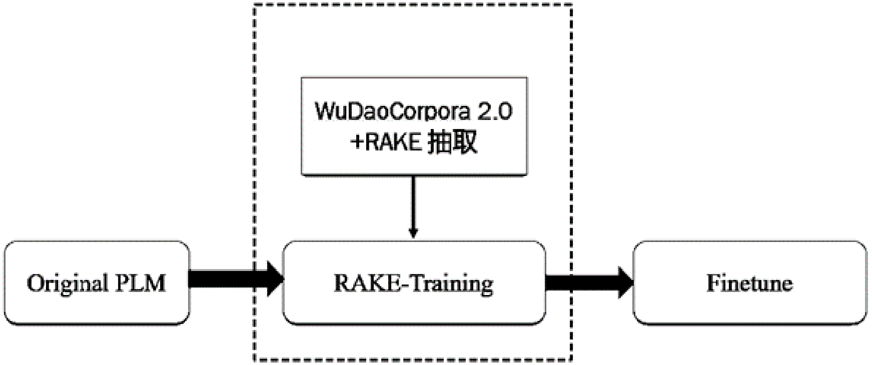
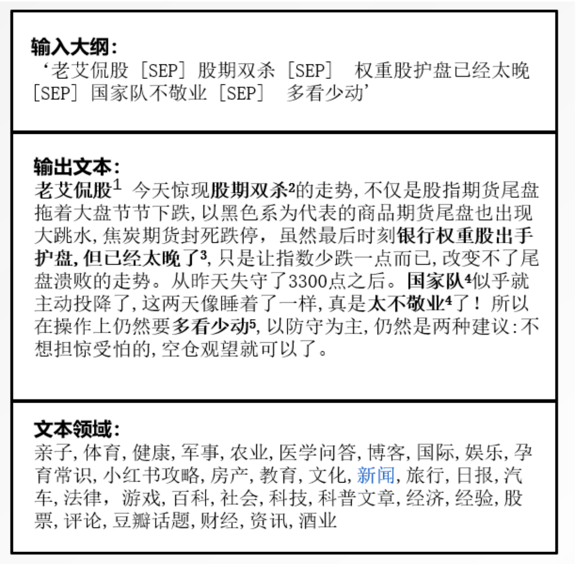


Figure 2: Rake-training



输入大纲:
'老艾侃股［SEP］股期双杀［SEP］ 权重股护盘已经太晚［SEP］国家队不敬业［SEP］ 多看少动'

输出文本:
老艾侃股[1] 今天惊现**股期双杀**[2]的走势，不仅是股指期货尾盘拖着大盘节节下跌，以黑色系为代表的商品期货尾盘也出现大跳水，焦炭期货封死跌停， 虽然最后时刻**银行权重股出手护盘，但已经太晚了**[3]，只是让指数少跌一点而已，改变不了尾盘溃败的走势。从昨天失守了3300点之后。**国家队**[4]似乎就主动投降了，这两天像睡着了一样，真是**太不敬业**[4]了！ 所以在操作上仍然要**多看少动**[5]，以防守为主，仍然是两种建议：不想担惊受怕的，空仓观望就可以了。

文本领域:
亲子，体育，健康，军事，农业，医学问答，博客，国际，娱乐，孕育常识，小红书攻略，房产，教育，文化，新闻，旅行，日报，汽车，法律， 游戏，百科，社会，科技，科普文章，经济，经验，股票，评论，豆瓣话题，财经，资讯，酒业

Figure 3 : Rake-training text examples

RAKE-Training method is shown in **Figure2**, we observe the discrepancy between the pretraining task and the downstream task will seriously affect the performance of the model. We propose rake-training as a supplementary pre-training method on public PTMs, specifically, the task-conditional text generation based on the input outlines (examples shown in **Figure 3**). So that models can transfer the knowledge stored in the parameters to the story generation task preferably.

Main Method

In trust Loss

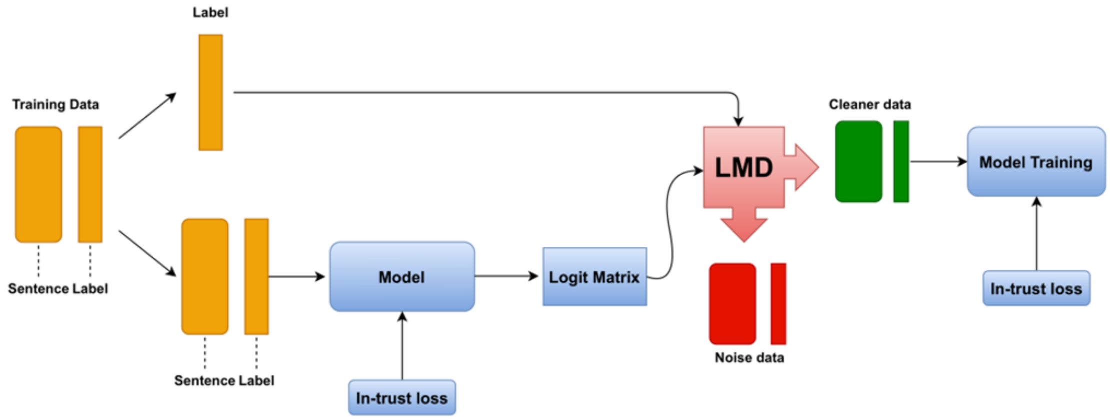


Figure 4: In trust Loss

$$L_{DCE} = -p\log(\delta p + (1-\delta)q)$$

$$L_{In-trust} = \alpha L_{CE} + \beta L_{DCE}$$

To reduce the influence of noise and make better use of difficult samples, we introduce the incomplete-Trust (In-trust) loss function, which trusts both model output and label information to utilize both the model output which obtains the relative correctness after enough training, and the incomplete correctness of labels may be mislabeled.

Main Method

Child Tuning

The way to reduce the difference between model fine-tuning and reality. By freezing a part of the parameters, catastrophic forgetting is avoided and the generalization effect is improved. Specifically, in the first round, the model uses full parameter fine-tuning, and then finds the top 20% parameters with the largest gradient (or directly finds some parameters through Bernoulli distribution). The parameters of this part are fine-tuned, and the remaining parameters are involved in the calculation but not updated.

$$w_{t+1} = w_t - \eta \frac{\partial \mathcal{L}(w_t)}{\partial w_t} \odot M_t$$

$$M_t \sim \text{Bernoulli}(p_F)$$

Experimental Setting

PLM	Methods	Size
CPT `https://github.com/fastnlp/CPT`	DAE MLM	16 Heads and 1024 Model dim.
BART-large `huggingface.co/facebook/bart-large`	DAE	370M
LongLM `https://github.com/thu-coai/LOT-Benchmark`	LOT	223M

Figure 5: Baseline Method

(1) BLEU-1/2: Measure the similarity between the output story and the standard answer.

(2) Distinct-3/4: Measure the diversity of output stories.

(3) Coverage: Calculate the Rouge-L recall score between the generated story and the given episode sequence to measures how many episodes in a given outline are included in the output story.

(4) Order: Divide the number of phrase pairs by the total number of phrase pairs by subtracting 1 from the reverse order (the order in which the phrases of the standard answer appear as the correct order), where the occurrence position of the plot is determined by aligning the longest common sequence. Measure whether the phrases appear in the proper order.

Experimental Setting

Framework： Pytorch+Huggingface
Optimizer： AdamW
MaxLen： 512

Pre-Training	Finetune
Batchsize： 64 GPU： 8*RTX3090 Lr： 5e-5	Batchsize： 8 GPU： RTX3090 Lr： 1e-5

Figure 6: Experimental Setting

Results & Analysis

Method	BLEU1	BLEU2	Dist-3	Dist-4	Coverage	Order	Overall
LongLM+ Small	26.6	16.0	17.9	31.4	83.6	63.2	30.2
LongLM+ Base	29.2	17.8	18.5	32.6	88.8	63.7	32.1
LongLM+ Large	30.2	17.8	18.9	33.9	89.9	65.4	32.8
BART-L + In-trust	24.4	13.4	**28.8**	**47.0**	93.3	66.8	31.9
CPT-L + In-trust	24.1	14.4	28.4	46.2	94.7	66.3	32.2
BART-L + In-trust + RAKE	30.4	20.8	15.8	28.1	95.7	71.7	34.7
CPT-L + In-trust+ RAKE	30.5	20.9	16.0	28.5	96.0	**71.9**	34.7
BART-L + In-trust + RAKE + Child	31.2	21.4	15.9	28.6	96.1	70.7	34.9
CPT-L + In-trust + RAKE + Child	**31.3**	**21.5**	23.5	39.5	**97.8**	69.9	**36.5**

Figure 7: Experimental results in validation

Our model surpasses the baseline model to varying degrees in terms of answer similarity, diversity, the inclusion of outline plot, and proper order.

- Similarity： Compared with the LOT, the BLUE score of our model improved by 4.8;
- Diversity： After using In-trust, our model improves the Dist metric by 11.5;
- Inclusion of outline： Compared with the baseline model, our model improved by 7.9 on the Coverage metric;
- Order： Rake-training improves our model by 6.5 in the Order metric.

Conclusion

- In this paper, we proposed a framework for story generation.
  - We have pioneered the rake-training method to bridge the gap between pre-training tasks and downstream tasks.
  - We introduce the child-tuning method to learn task-related PLMs and also take denoising learning to get integrated and logical smoothly stories.
  - We conduct extensive experiments in 6 metrics. The experimental results prove the effectiveness of our approach.
- In the future, we will further explore how to integrate more logical knowledge and common sense into existing model for generating discourse-level stories.

Thanks for listening!