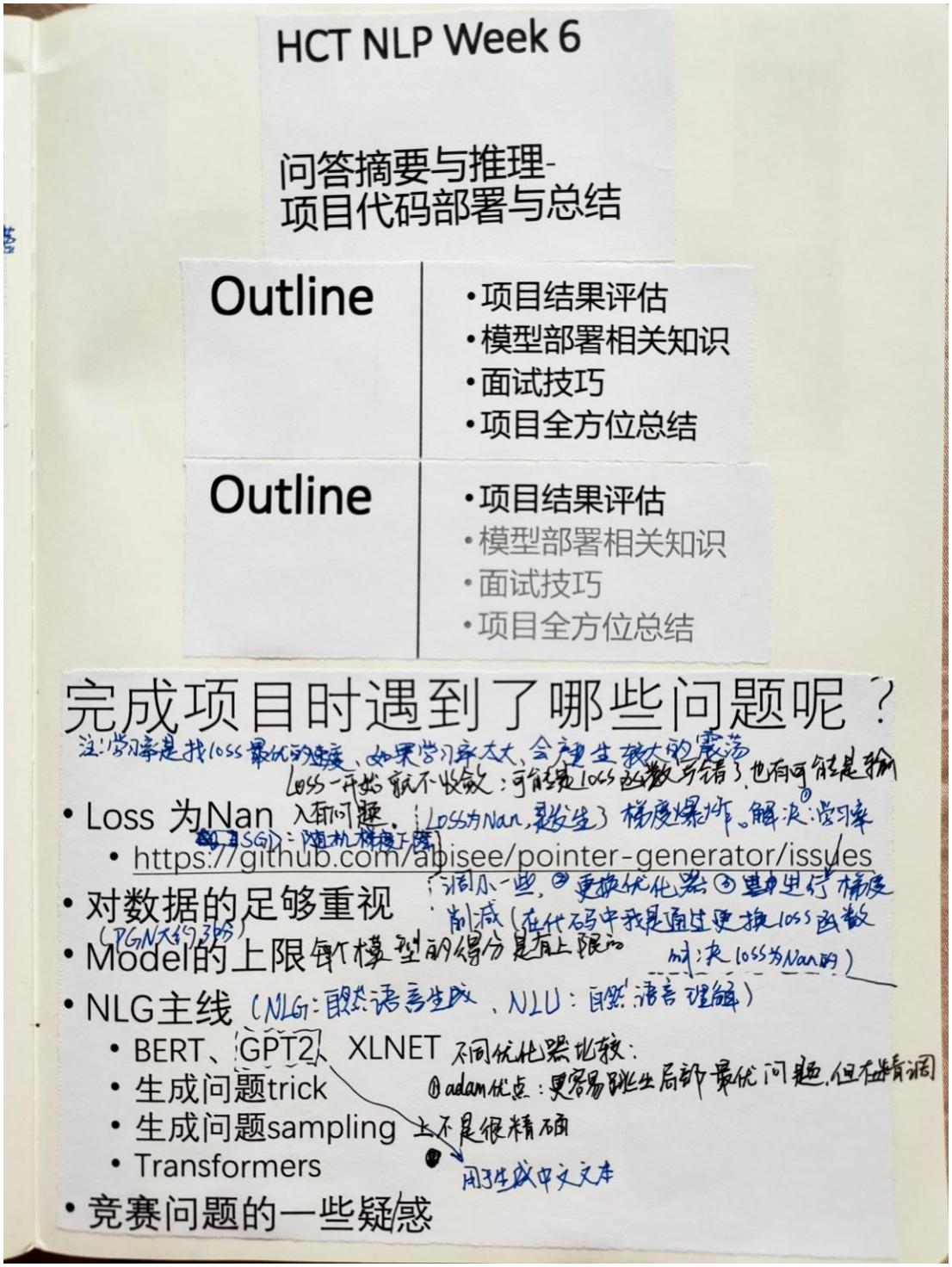
第六章 问答摘要与推理-项目代码部署与总结

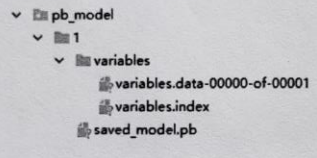


Outline

- 项目结果评估
- **模型部署相关知识**
- 面试技术
- 项目全方位总结

Checkpoint

```
checkpoint
ckpt-10.data-00000-of-00002
ckpt-10.data-00001-of-00002
ckpt-10.index
ckpt-11.data-00000-of-00002
ckpt-11.data-00001-of-00002
ckpt-11.index
ckpt-12.data-00000-of-00002
ckpt-12.data-00001-of-00002
```

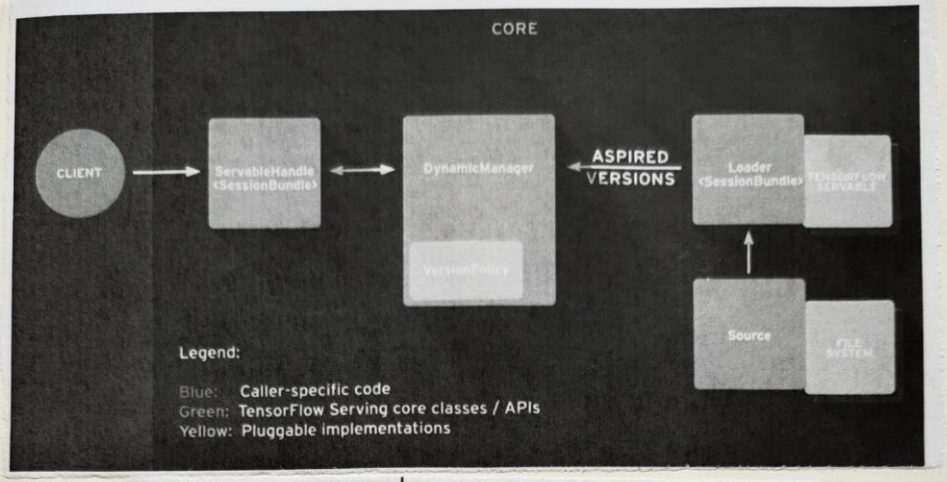


代码部署

方法：① Flask

② Tensorflow Serving

TF Serving



Outline

- 项目结果评估
- 模型部署相关知识
- **面试技巧**
- 项目全方位总结

近几年的招聘动态

- 2016年秋招
  - 毕业生offer不愁
- 2017年秋招
  - 大厂争夺更加激烈，毕业生offer薪资普遍升高。抢人时间甚至提前到了3，4月
- 2018年秋招
  - 大厂的职位饱和，新的HC有所限制，新闻上还曾爆出了某些创业公司毁约毕业生的事情；另一方面，对毕业生的动手能力要求也更高
- 2019年秋招
  - 这一年的情况比较差，offer比较难，可能期望值提的太高，所以算法岗位的竞争更加激烈，不仅要和同届的同学竞争，还要和从AI公司离职的工程师竞争，而且19年招聘更看重动手能力，单凭算法能力已经无法拿到好的offer了。

- 反映到毕业生这边，核心还是看动手能力，其次才会考量对NLP算法的理解。尤其是在深度学习普及的今天，深度学习能力（不仅是调参，更重要的是对model的理解和优化）也会成为考量的重要因素。
- 好人才总归是缺的，无论是NLP，CV，还是DL，足够优秀的同学一定是我们所渴求的。那些不仅有调参经验，而且编程能力超强，对数学的掌握足够深厚的同学，可以根据实际问题调整网络结构，利用各种方法满足工程需求，并能熟练运用各种编程语言（例如C和C++）进行工程化改造。这样的人也能够拿到更好的offer。

Timit语音识别
目的: 实现嘈杂环境下自动语音识别（ASR）
职责: 在Timit数据集上建立三音素模型
❖ 用DNN网络代替传统GMM-HMM模型中的GMM层
成果：实现5dB嘈杂噪声环境下54%WER，较基准提高20%

2019.04~至今　　凤凰网　　自然语言处理实习生
✧ **项目描述：**凤凰新闻 APP 的推荐。
✧ **工作内容：**（1）负责百度词条中命名实体的提取，构建人名、地名、机构名三个词表；（2）主要负责新闻标题党的识别，独立实现各种监督学习的训练代码，并在测试集上对训练效果进行评估。
✧ **工作进展：**目前模型在测试集上的召回率达到 95%，准确率达到 85%，现在还在不断用线上数据测试训练模型，增加语料的覆盖度，提高模型的泛化能力，在保证召回率的同时提高标题党的准确率。

智能视觉问答(Visual Question Answering, VQA)　核心成员 2019.01-2019.07
- 开发基于深度学习的图片信息和自然语言信息多模态融合算法模型，实现图片问答功能；
- 问题信息使用 RNN、LSTM、Transformer、BERT 等方式编码，图片信息使用 Faster RCNN、YOLO 等模型提取特征，利用 Attention 的方式将问题信息与图片信息进行多模态信息融合，最后接入 FC 分类器输出答案；
- Attention 方案有多种，包括 TopDownBottomUp Attenton、Bilinear Attention、CoAttenion、Self Attention 等多种融合方案，实现更好的效果；
- 对问题信息、图像信息进行多种方式的数据增强，提升算法性能；
- 引入图像的 Image Caption 信息，与问题信息进行融合，提升模型性能；
- 设计答案分类掩模，包括预定义和动态自适应两种方法，对明显不相关答案进行过滤；
- 申请国家发明专利 2 项（均为第 1 作者）；
- 获得 CVPR 2019 VQA Challenge 第 6 名。

1、研发聊天系统中的问答相关性建模、命名实体识别、生成式聊天等技术。
- 负责设计和实现聊天系统 (Chatbot) 中的 query-post 相关性以及 query-reply 相关性模型，对词向量（Word Embedding）加权模型、RNNLM (GRU、Bi-GRU)、Pairwise-CNN、Seq2Seq 等模型进行效果测试。
- 支持多个项目中的命名实体识别 (NER) 组件的研发，对条件随机场 (CRF)、双向长短时间记忆 (Bi-LSTM)、Bi-LSTM+CRF 等模型进行效果评估。

Resume

- 自动文本摘要项目
- 生成式方案
  - Seq2seq+attention，通过构建一些专有词汇表和文本数据的处理，Rouge-L最高能到30，但是会有大量OOV和词语重复问题
  - 通过Point-generater Network不但解决OOV问题，同时可以进一步减少Vocab大小，加快计算收敛速度；采用coverage机制，减少文本摘要生成词语重复问题。Rouge-L分数达到45左右，提升xx%
  - 在训练策略上做了些调整，采用schedule sampling策略，同时为了生成的多样性，对beamsearch进行改进，效果提升xx%
- 抽提式方案
  - BertSum

Outline

- 项目结果评估
- 模型部署相关知识
- 面试技巧
- 项目全方位总结

Datesets

CNN/Daily Mail数据集

First highlight: Argentina coach Sabella believes Messi's habit of being sick during games is down to nerves.

First 2 sentences: Argentina coach Alejandro Sabella believes Lionel Messi's habit of throwing up during games is because of nerves. The Barcelona star has vomited on the pitch during several games over the last few seasons and appeared to once again during Argentina's last warm-up match against Slovenia on Saturday.

A standard benchmark dataset

More than 300K news articles

cnn-dailymail数据集地址

Text summarization

Extractive text summarization

Source Text: Peter and Elizabeth took a taxi to attend the night party in the city.

While in the party, Elizabeth collapsed and was rushed to the hospital.

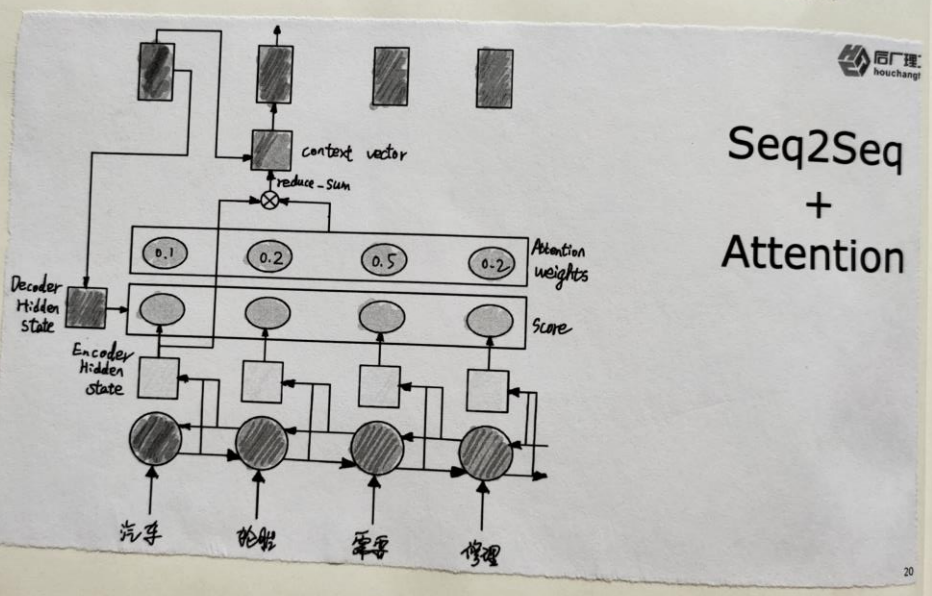
Summary: Peter and Elizabeth attend party city. Elizabeth rushed hospital.

Abstractive text summarization

Source Text: Peter and Elizabeth took a taxi to attend the night party in the city.

While in the party, Elizabeth collapsed and was rushed to the hospital.

Summary: Elizabeth was hospitalized after attending a party with Peter.



Seq2Seq
+
Attention

Problems

The encoder is not well trained via back propagation through which is

从模型的路径上看，encoder到实际输出有一定距离，从此限制了反向传播。

OOV（Out-of-vocabulary未登录词）

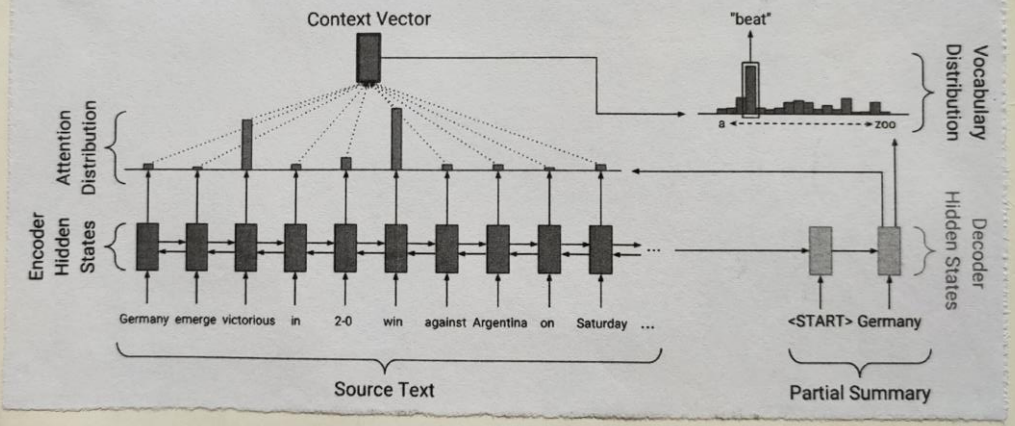
PGN 解决

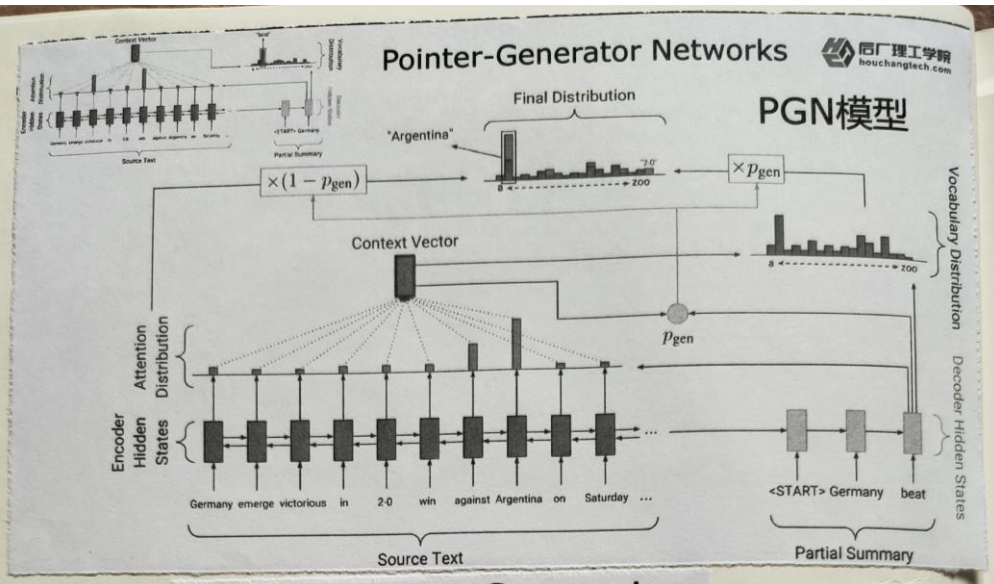
摘要总结的结果有的时候并不准确，比如摘要的结果可能输出德国队以2-1比分击败阿根廷，但是实际比分是2-0，出现这个的原因是out-of-vocabulary words（OOV）的出现

Word-repetition问题 => Coverage 解决

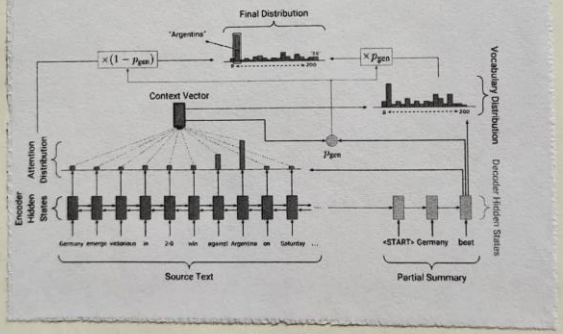
摘要结果会出现repeat重复的信息，比如重复出现德国队击败阿根廷队

Baseline Seq2seq





PGN (Pointer-Generator Networks)



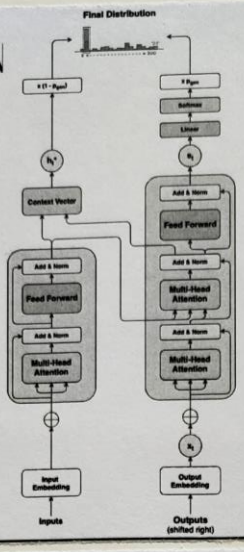
Repetition Handling

model generated summaries suffer from both word-level and sentence-level repetitions.

Coverage

Beyond RNN

Transformer

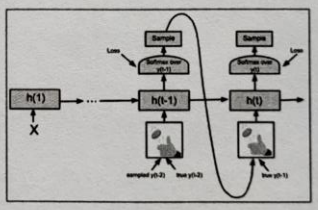


Transformers and Pointer-Generator Networks for Abstr
Summarization

TRAINING STRATEGIES

1) Teacher Forcing

2) Scheduled Sampling

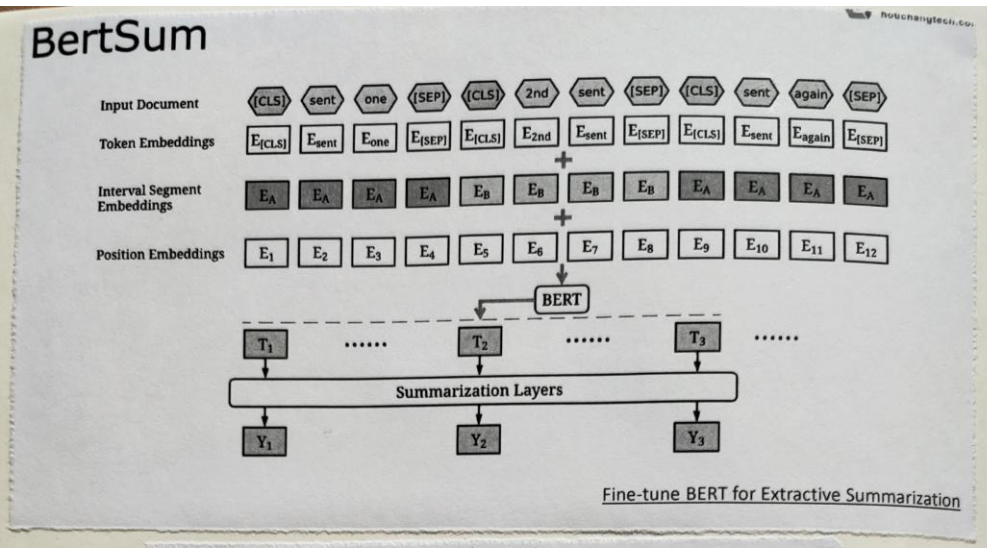


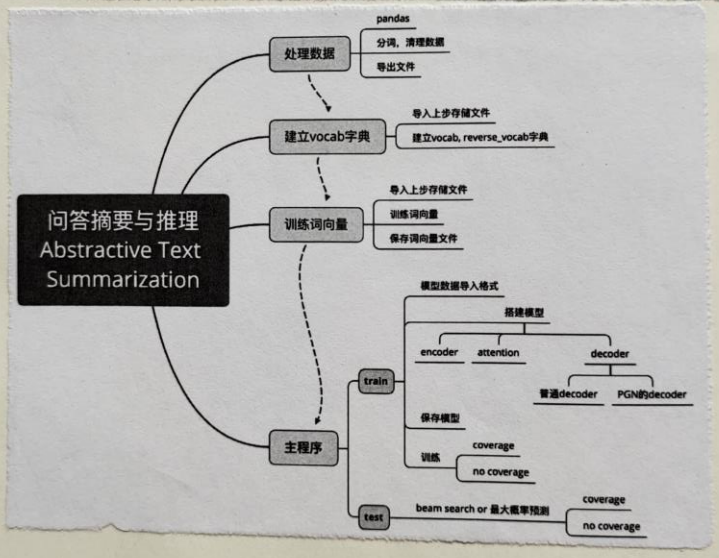
Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks

SUMMARY GENERATION

Diverse Beam Decoding

the top-B hypotheses may differ by just a couple tokens at the end of sequences, which not only affects the quality of generated sequences but also wastes computational resources


BertSum
Input Document
Token Embeddings
Interval Segment Embeddings
Position Embeddings
BERT
Summarization Layers
Fine-tune BERT for Extractive Summarization


问答摘要与推理
Abstractive Text Summarization
处理数据
pandas
分词，清理数据
导出文件
建立vocab字典
导入上步存储文件
建立vocab, reverse_vocab字典
训练词向量
导入上步存储文件
训练词向量
保存词向量文件
主程序
train
模型数据导入格式
搭建模型
encoder
attention
decoder
普通decoder
PGN的decoder
保存模型
训练
coverage
no coverage
test
beam search or 最大概率预测
coverage
no coverage

附：

（1）中文预训练模型牛人 github：https://github.com/bojone/bert4keras
（2）博客：http://jalammar.github.io/illustrated-transformer/
（3）nlp 顶会论文：THE CURIOUS CASE OF NEURAL TEXT DeGENERATION
（4）部署模型：tensorflow.org/tfx/tutorials/serving/rest_simple
（5）牛人（brightmaret）github：https://github.com/bojone/albert_zh
https://github.com/brightmart/albert_zh
（6）transformer 作者 github：
https://github.com/huggingface/transformers/blob/master/examples/summarization/bertabs/modeling_bertabs.py
https://github.com/huggingface/transformers
（7）docker：https://hub.docker.com/
（8）GPT2: https://github.com/qingkongzhiqian/GPT2-Summary
（9）opennmt：https://opennmt.net/
（10）论文：the_curious_case_of_neural_text_degeneration.pdf