Chinese Text Summarization

Team 8: Chengyu Chen, Yuan Ding, Jiyang Ge, Yupei Zhou

Introduction

Problem

1. News have no title
2. Waste time for editor, waste money for company
3. Editor add subjectivity
4. Grab attention, exaggerate/distort true meaning

Solution: Text Summarization model

1. Enable users to target on the interested news
2. Save time and save money
3. Make title more objective

Data Layout

Data source: Large scale Chinese Short Text Summarization dataset*

Data Content:

- 10,666 human labeled (short text, summary) pairs
- Text: ~80 characters; Summary: [10,30] characters
- From Chinese microblogging website Sina Weibo
- Domain of politic, economic, military, movies, etc.

Short Text: 水利部水资源司司长陈明忠今日在新闻发布会上透露，根据刚刚完成的水资源管理制度的考核，有部分省接近了红线的指标，有部分省超过红线的指标。在一些超过红线的地方，将对一些取用水项目进行区域的限批，严格地进行水资源论证和取水许可的批准。

Mingzhong Chen, the Chief Secretary of the Water Devision of the Ministry of Water Resources, revealed today at a press conference, according to the just-completed assessment of water resources management system, some provinces are closed to the red line indicator, some provinces are over the red line indicator. In some places over the red line, It will enforce regional approval restrictions on some water projects, implement strictly water resources assessment and the approval of water licensing.

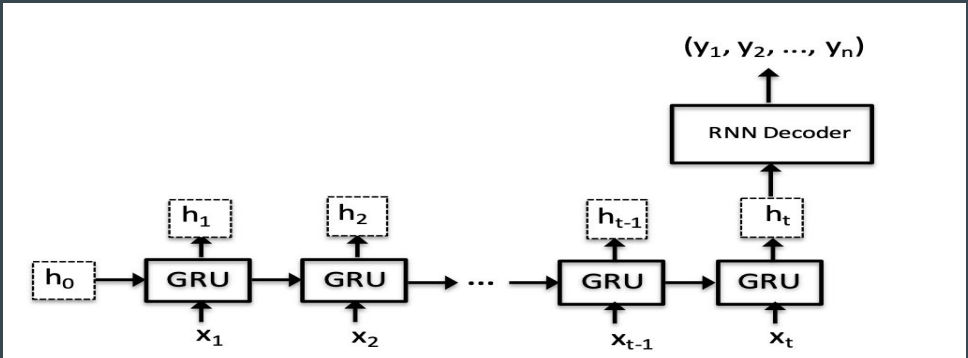
Summarization: 部分省超过年度用水红线指标 取水项目将被限批

Some provinces exceeds the red line indicator of annual water using, some water project will be. limited approved

**source: Qingcai Chen Baotian Hu and Fangze Zhu. 2015. Lcsts: A large scale chinese short text summarization dataset. in proceedings of the 2015 conference on empirical methods in natural language processing. lisbon, portugal, 1967–1972.*

Approach

1. Baseline (RNN)



*source: Qingcai Chen Baotian Hu and Fangze Zhu. 2015. Lcsts:
A large scale chinese short text summarization dataset. in proceedings of the 2015 conference on empirical methods in natural language processing. lisbon, portugal, 1967–1972.*

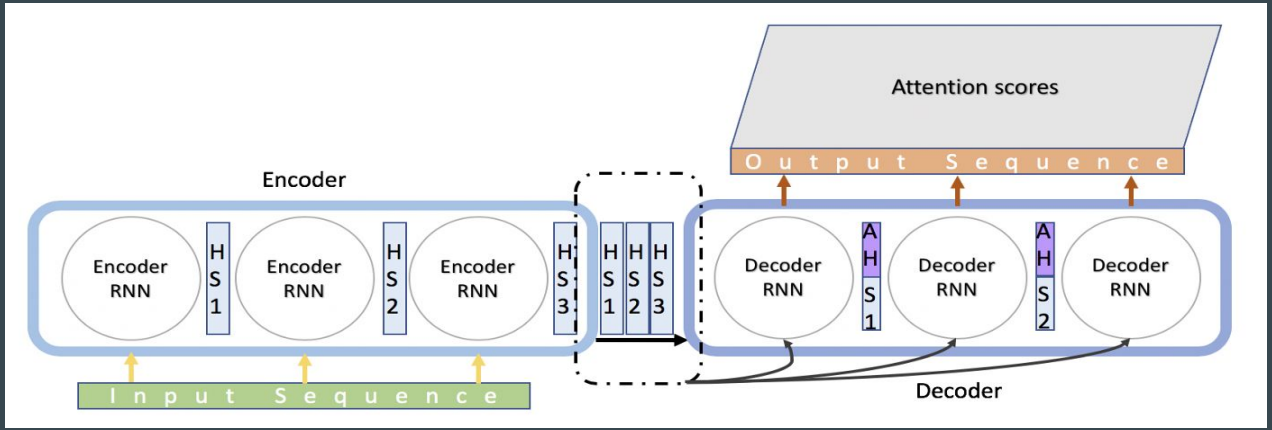
2. Sequence to sequence model
    a. RNN as encoder and decoder, add attention, beam search
    b. BERT as encoder and transformer decoder as decoder, multi-head attention

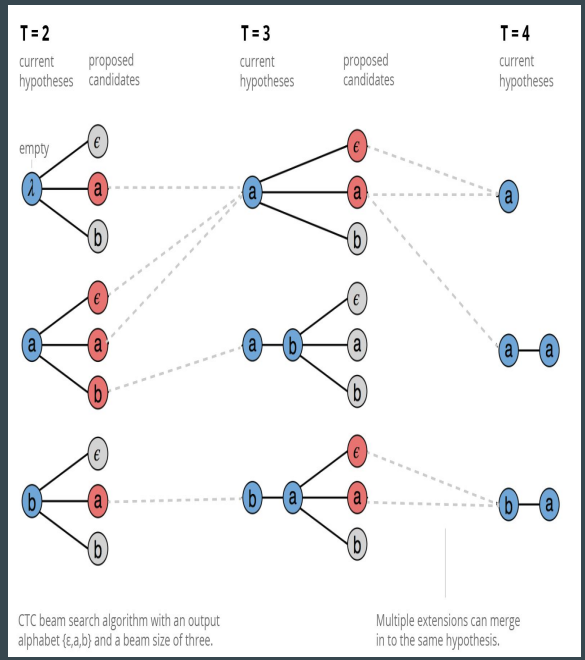
Approach 1

Data Processing:

- Word based method
  - Segmented into Chinese words using package jieba

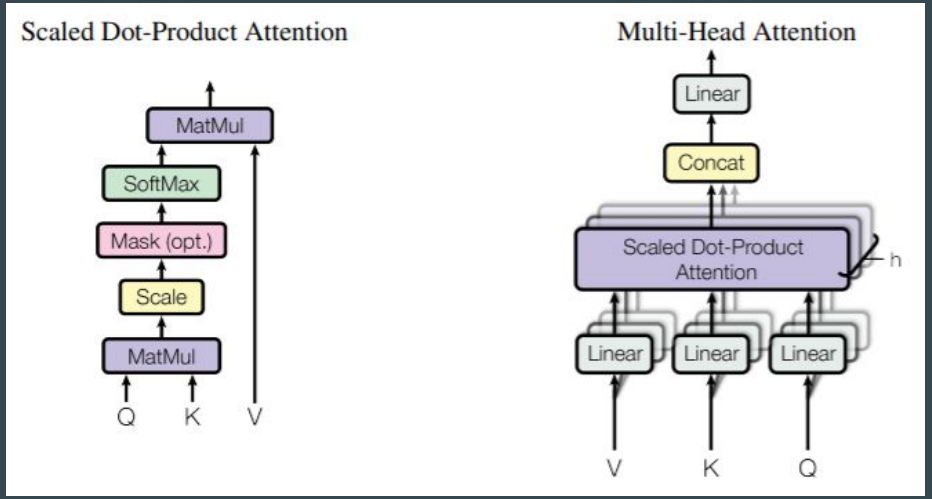
Model Training





Approach 2

- Pretrained BERT model as encoder: bert-base-chinese(https://huggingface.co/bert-base-chinese)
- Transformer decoder



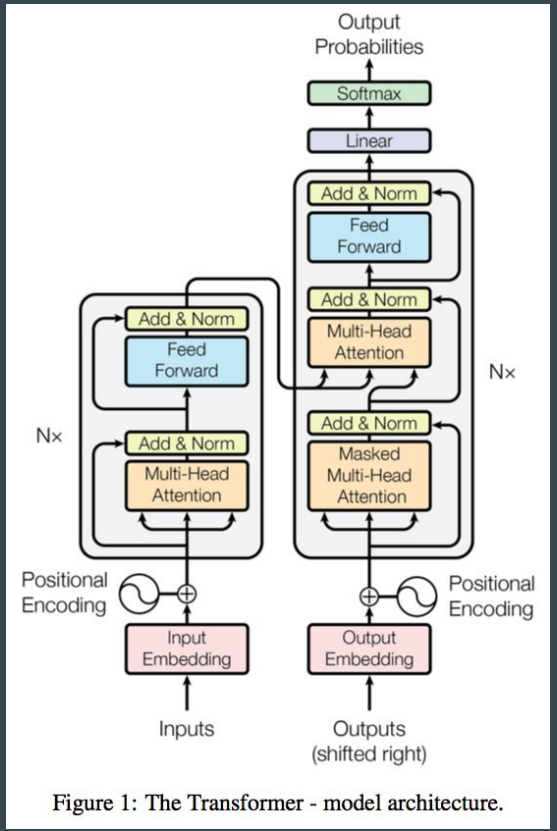


Figure 1: The Transformer - model architecture.

Results

Evaluation metrics:

- ROUGE-n: refers to the overlap of **n-gram** between the prediction and reference summaries.
- ROUGE-L: **Longest Common Subsequence** (LCS) based statistics, taking into account sentence level structure similarity and identifies longest co-occurring in sequence n-grams automatically.

Scores	Rouge-1	Rouge-2	Rouge-L
RNN without attention (Chinese)	0.060	0.010	0.055
RNN with attention (Chinese)	0.163	0.080	0.150
BERT encoder + transformer decoder (Chinese)	0.270	0.108	0.253
BERT encoder + transformer decoder (English)	0.313*	0.134*	0.295*

*sources: Aksenov, Dmitrii, et al. "Abstractive Text Summarization based on Language Model Conditioning and Locality Modeling." arXiv preprint arXiv:2003.13027 (2020).

Results(Examples)

1. Summary: 10门酷毙了的云计算编程语言
   Prediction: 10门酷毙了的云计算编程语言
   Summary: Ten cool cloud computing programming languages
   Prediction: Ten cool cloud computing programming languages

2. Summmary: 海南政府服务热线将统一为:12345
   Prediction: 海南整合各类服务热线将统一为1234512345
   Summary: Hainan government service hotline will be unified as: 12345
   Prediction: Hainan integrates various service hotlines and will unify them to 1234512345

3. Summary: 望城区一届人大三次会议将于12月29-31日召开
   Prediction: The Third Session of the First People's Congress of Wangcheng District will be held on December 29-31
   Summary: 望城区召一届人大代表召开会召开会议议议
   Prediction:Wangcheng District called the first People's congress to hold a meeting, hold a meeting, meeting

We are open to questions, please feel free to discuss with us any time!

References

- https://www.aclweb.org/anthology/D19-1301.pdf
- https://arxiv.org/pdf/1506.05865.pdf
- https://arxiv.org/pdf/1506.01057.pdf
- https://www.aclweb.org/anthology/W04-1013
- https://www.aclweb.org/anthology/P18-2115
- https://towardsdatascience.com/day-1-2-attention-seq2seq-models-65df3f49e263
- https://arxiv.org/pdf/1706.03762.pdf
- https://arxiv.org/pdf/1703.03130.pdf