文本摘要简述

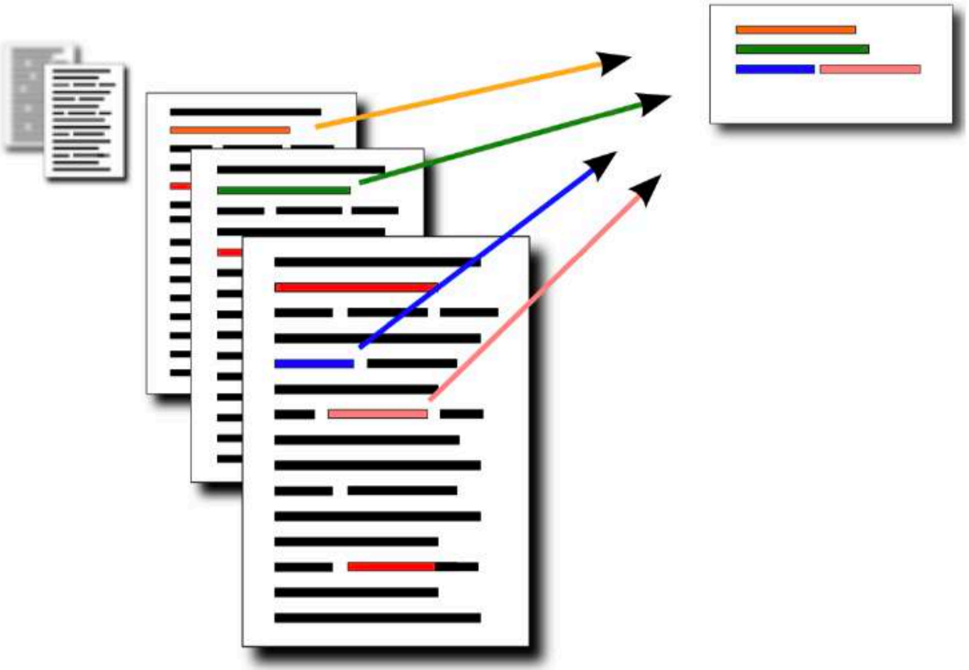
Xiachong Feng

目录

- 简介
- 分类
- 方法
- 评价指标
- 数据集
- 总结

简介

- **文本摘要**旨在将文本或文本集合转换为包含**关键信息**的简短文本。

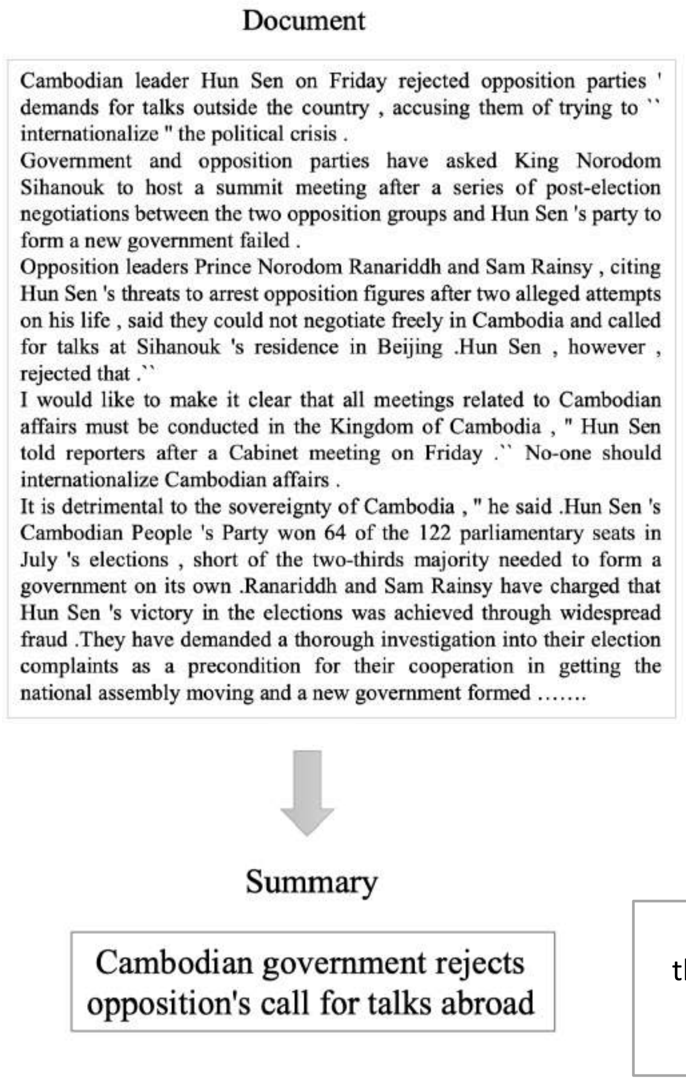


分类

- **输入类型**
  - 单文档摘要（Single-Document）
  - 多文档摘要（Multi-Document）
- **输出类型**
  - 抽取式摘要（Extractive）
  - 生成式摘要（Abstractive）
- **学习方法**
  - 有监督摘要（Supervised）
  - 无监督摘要（Unsupervised）
- **语言种类**
  - 单语言摘要（Monolingual）
  - 跨语言摘要（Cross lingual）
- ……

单文档 && 多文档

1. 减小句子冗余度
2. 确定句子顺序
3. 确定压缩比率



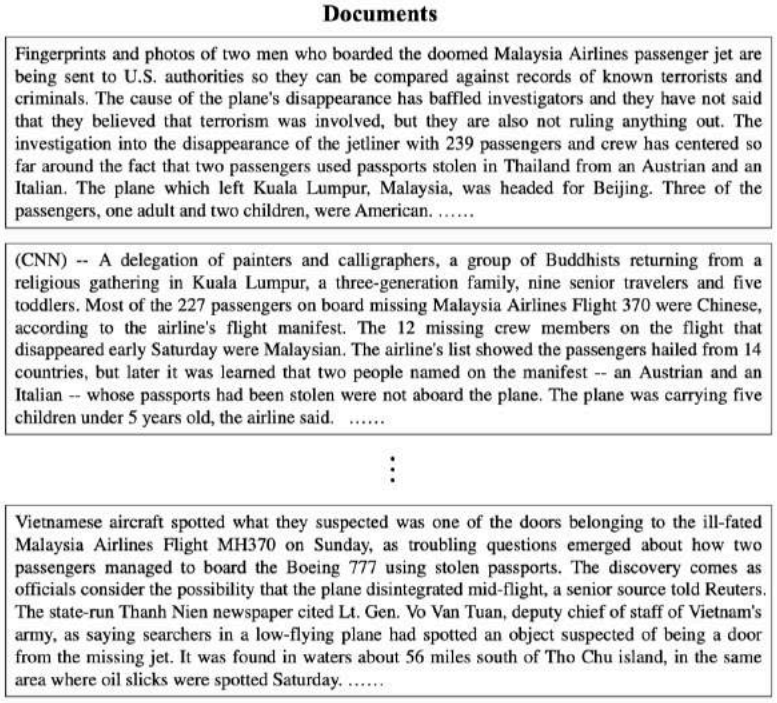
Document

Cambodian leader Hun Sen on Friday rejected opposition parties ' demands for talks outside the country , accusing them of trying to `` internationalize '' the political crisis .
Government and opposition parties have asked King Norodom Sihanouk to host a summit meeting after a series of post-election negotiations between the two opposition groups and Hun Sen 's party to form a new government failed .
Opposition leaders Prince Norodom Ranariddh and Sam Rainsy , citing Hun Sen 's threats to arrest opposition figures after two alleged attempts on his life , said they could not negotiate freely in Cambodia and called for talks at Sihanouk 's residence in Beijing .Hun Sen , however , rejected that .``
I would like to make it clear that all meetings related to Cambodian affairs must be conducted in the Kingdom of Cambodia , '' Hun Sen told reporters after a Cabinet meeting on Friday .`` No-one should internationalize Cambodian affairs .
It is detrimental to the sovereignty of Cambodia , '' he said .Hun Sen 's Cambodian People 's Party won 64 of the 122 parliamentary seats in July 's elections , short of the two-thirds majority needed to form a government on its own .Ranariddh and Sam Rainsy have charged that Hun Sen 's victory in the elections was achieved through widespread fraud .They have demanded a thorough investigation into their election complaints as a precondition for their cooperation in getting the national assembly moving and a new government formed …….

Summary

Cambodian government rejects opposition's call for talks abroad

Topic "Malaysia Airlines Disappearance"



Documents

Fingerprints and photos of two men who boarded the doomed Malaysia Airlines passenger jet are being sent to U.S. authorities so they can be compared against records of known terrorists and criminals. The cause of the plane's disappearance has baffled investigators and they have not said that they believed that terrorism was involved, but they are also not ruling anything out. The investigation into the disappearance of the jetliner with 239 passengers and crew has centered so far around the fact that two passengers used passports stolen in Thailand from an Austrian and an Italian. The plane which left Kuala Lumpur, Malaysia, was headed for Beijing. Three of the passengers, one adult and two children, were American. ……

(CNN) -- A delegation of painters and calligraphers, a group of Buddhists returning from a religious gathering in Kuala Lumpur, a three-generation family, nine senior travelers and five toddlers. Most of the 227 passengers on board missing Malaysia Airlines Flight 370 were Chinese, according to the airline's flight manifest. The 12 missing crew members on the flight that disappeared early Saturday were Malaysian. The airline's list showed the passengers hailed from 14 countries, but later it was learned that two people named on the manifest -- an Austrian and an Italian -- whose passports had been stolen were not aboard the plane. The plane was carrying five children under 5 years old, the airline said. ……

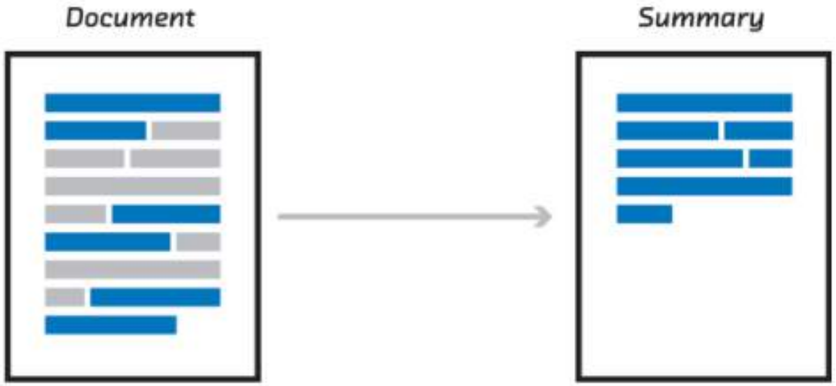
⋮

Vietnamese aircraft spotted what they suspected was one of the doors belonging to the ill-fated Malaysia Airlines Flight MH370 on Sunday, as troubling questions emerged about how two passengers managed to board the Boeing 777 using stolen passports. The discovery comes as officials consider the possibility that the plane disintegrated mid-flight, a senior source told Reuters. The state-run Thanh Nien newspaper cited Lt. Gen. Vo Van Tuan, deputy chief of staff of Vietnam's army, as saying searchers in a low-flying plane had spotted an object suspected of being a door from the missing jet. It was found in waters about 56 miles south of Tho Chu island, in the same area where oil slicks were spotted Saturday. ……

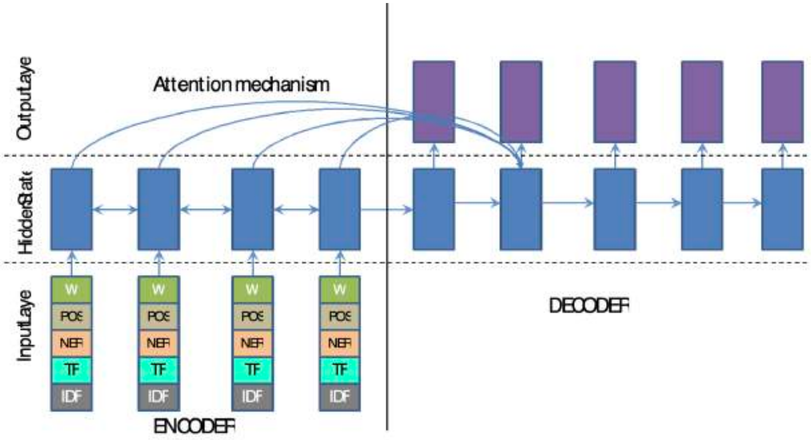
Summary

Flight MH370, carrying 239 people vanished over the South China Sea in less than an hour after taking off from Kuala Lumpur, with two passengers boarded the Boeing 777 using stolen passports. Possible reasons could be an abrupt breakup of the plane or an act of terrorism……

抽取式 && 生成式



抽取式摘要

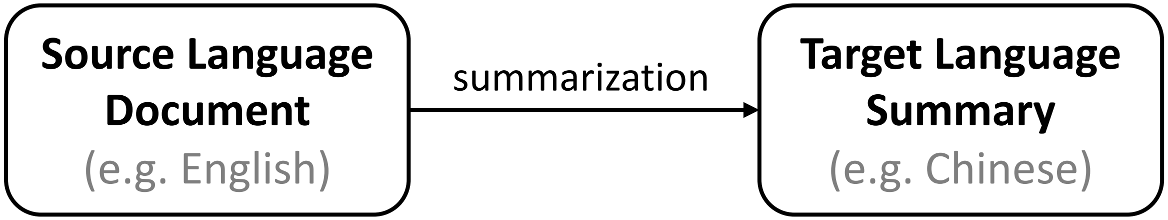


生成式摘要

Nallapati, R., Zhou, B., santos, dos, C. N., Gülçehre, Ç., & Xiang, B. (2016, February 19). Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. arXiv.org.

单语言 && 跨语言

- Cross-**Language** Summarization (*-2018)
- Cross-**Lingual** Summarization (2019)



Source Language Document
(e.g. English)

summarization

Target Language Summary
(e.g. Chinese)

目录

- 简介
- 分类
- **方法**
  - **抽取式、生成式、结合式、多任务学习、强化学习、图神经网络、基于预训练、无监督、跨语言**
- 评价指标
- 数据集
- 总结

传统抽取式摘要方法

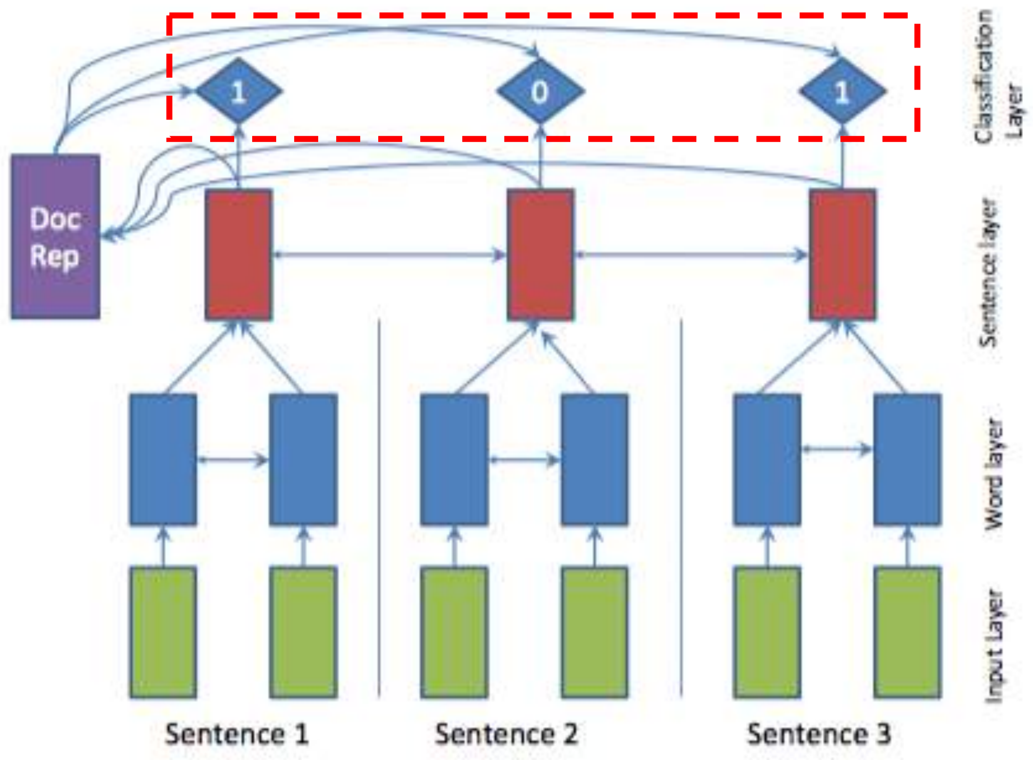
- **Lead-3**
  - 选取前三个句子作为摘要。
- **TextRank**
  - 仿照PageRank，句子作为节点，构造无向有权边,权值为句子相似度。
- **Clustering**
  - 句子级别向量表示
  - 聚类算法
    - K均值聚类……
  - 抽取式摘要
    - 从每一个聚类中，选择距离“质心”最近的句子
- **Maximal Marginal Relevance(MMR)**
  - 最大边界相关法
  - 根据句子位置、句子长度、句子关键词、是否包含标题词等特征打分，引入新颖性特征（避免相似性），使得抽取结果“句句重要，句句不同”

基于神经网络的抽取式摘要

- **序列标注**
  - 为原文中的每一个句子打一个二分类标签（0或1）；
  - 1代表该句属于摘要；
  - 0代表该句不属于摘要；
  - 最终摘要由所有标签为1的句子构成。
- **句子级标签获取**
  1. 首先选取原文中与标准摘要（Golden）计算ROUGE得分最高的一句话加入候选集合；
  2. 接着继续从原文中进行选择，保证选出的摘要集合ROUGE得分增加，直至无法满足该条件；
  3. 得到的候选摘要集合对应的句子设为1标签，其余为0标签。

基于神经网络的抽取式摘要

- SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents ***AAAI17***



Output

Sentence-level

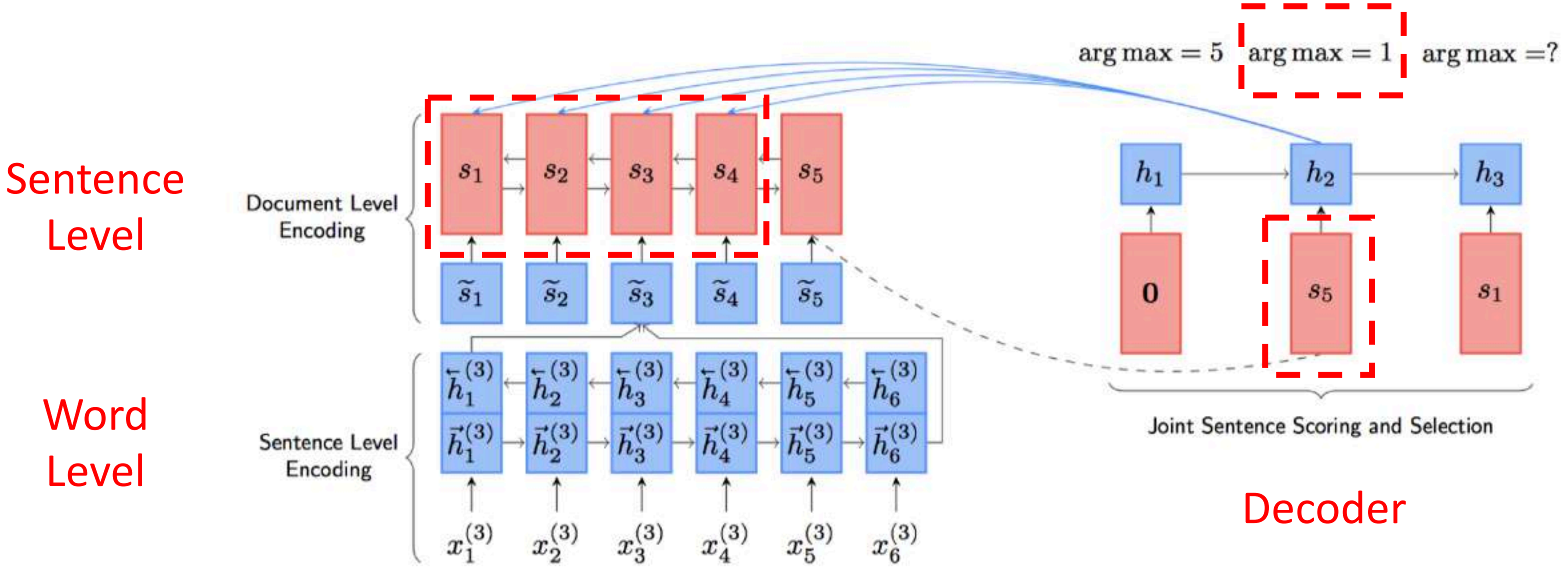
Word-level

Input

基于神经网络的抽取式摘要

- Neural Document Summarization by Jointly Learning to Score and Select Sentences ***ACL18***
- 创新：计算选择**句子的收益**作为打分方式。

$$g(S_t|\mathbb{S}_{t-1}) = r\left(\mathbb{S}_{t-1} \cup \{S_t\}\right) - r(\mathbb{S}_{t-1})$$

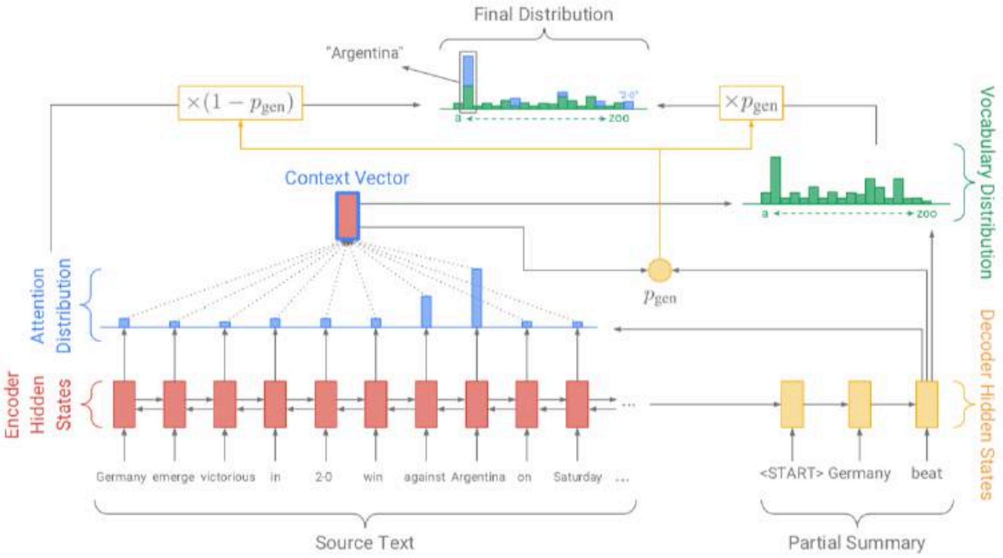

Sentence Level
Word Level
Document Level Encoding
Sentence Level Encoding
s_1 s_2 s_3 s_4 s_5
$\tilde{s}_1$ $\tilde{s}_2$ $\tilde{s}_3$ $\tilde{s}_4$ $\tilde{s}_5$
$\overleftarrow{h}_1^{(3)}$ $\overleftarrow{h}_2^{(3)}$ $\overleftarrow{h}_3^{(3)}$ $\overleftarrow{h}_4^{(3)}$ $\overleftarrow{h}_5^{(3)}$ $\overleftarrow{h}_6^{(3)}$
$\overrightarrow{h}_1^{(3)}$ $\overrightarrow{h}_2^{(3)}$ $\overrightarrow{h}_3^{(3)}$ $\overrightarrow{h}_4^{(3)}$ $\overrightarrow{h}_5^{(3)}$ $\overrightarrow{h}_6^{(3)}$
$x_1^{(3)}$ $x_2^{(3)}$ $x_3^{(3)}$ $x_4^{(3)}$ $x_5^{(3)}$ $x_6^{(3)}$
$\arg\max = 5$ $\arg\max = 1$ $\arg\max = ?$
h_1 h_2 h_3
$\mathbf{0}$ s_5 s_1
Joint Sentence Scoring and Selection
Decoder

生成式摘要

- 抽取式摘要问题：**内容选择错误、连贯性差、灵活性差。**
- **生成式摘要允许摘要中包含新的词语或短语**，灵活性高。
- 随着近几年神经网络模型的发展，序列到序列（Seq2Seq）模型被广泛的用于生成式摘要任务，并取得一定的成果。
- **Seq2Seq问题：**
  - **重复**描述某些事实性信息
  - **未登录词**问题（OOV）

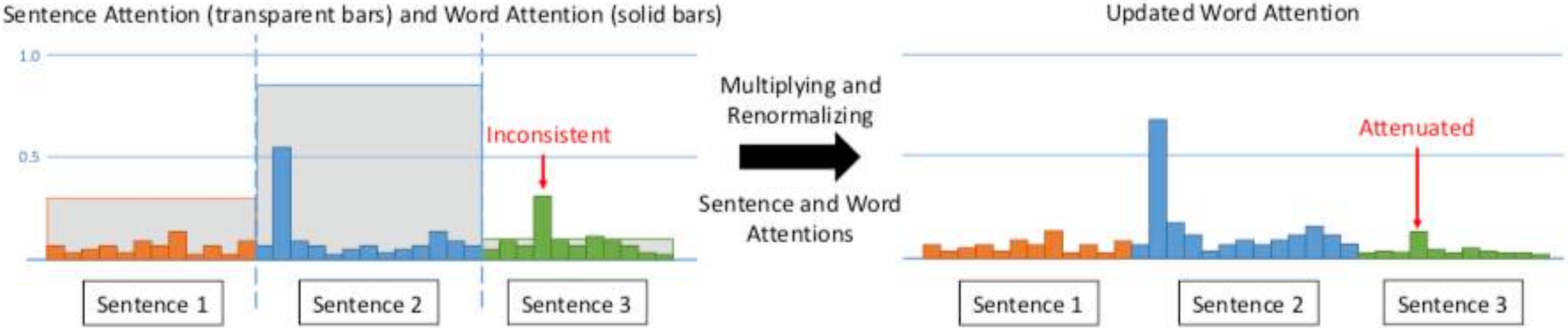
生成式摘要

- Get To The Point-Summarization with Pointer-Generator Networks ***ACL17***
- **Copy机制**，在解码的每一步计算拷贝（pointer）或生成（generator）的概率，该机制可以选择从原文中拷贝词语到摘要中，**有效的缓解了未登录词（OOV）的问题。**
- **Coverage机制**，在解码的每一步考虑之前步的attention权重，结合coverage损失，避免继续考虑已经获得高权重的部分。该机制可以**有效缓解生成重复的问题。**



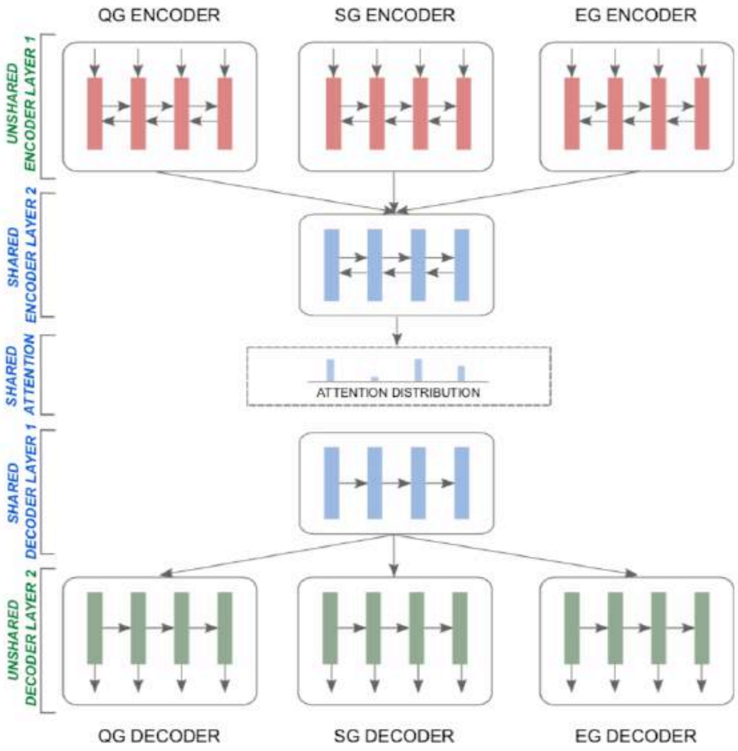
抽取式结合生成式

- A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss ***ACL18***
- 方法
  - 使用句子级别的抽取概率作为句子级别的attn权重
  - 使用句子级别权重来re-weight词语级别的attn权重



多任务学习

- Soft Layer-Specific Multi-Task Summarization with Entailment and Question Generation ***ACL18***
- 摘要生成 + 问题生成 + 蕴含生成

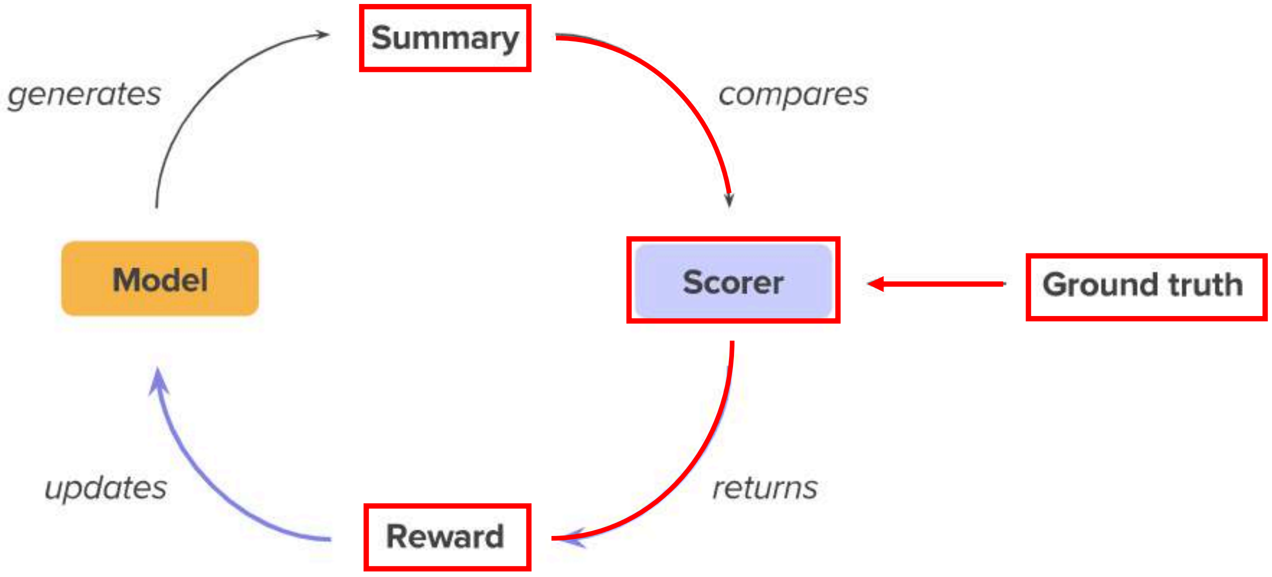


强化学习

- A Deep Reinforced Model for Abstractive Summarization

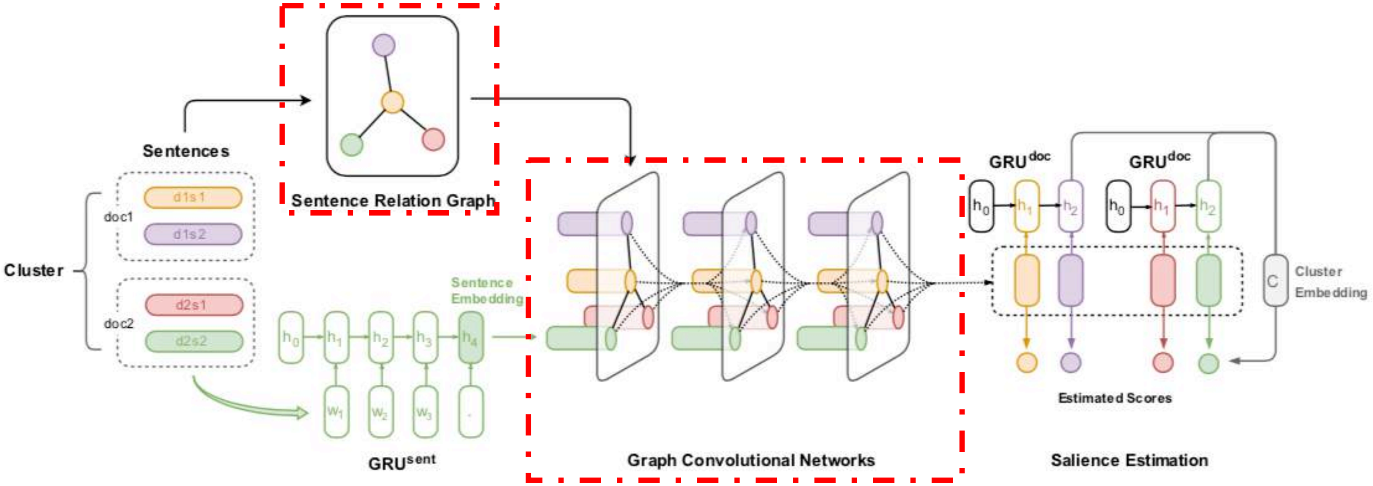
  ICLR18

- 奖励 : **ROUGE**
- 学习策略: **Self-critical** policy gradient training algorithm



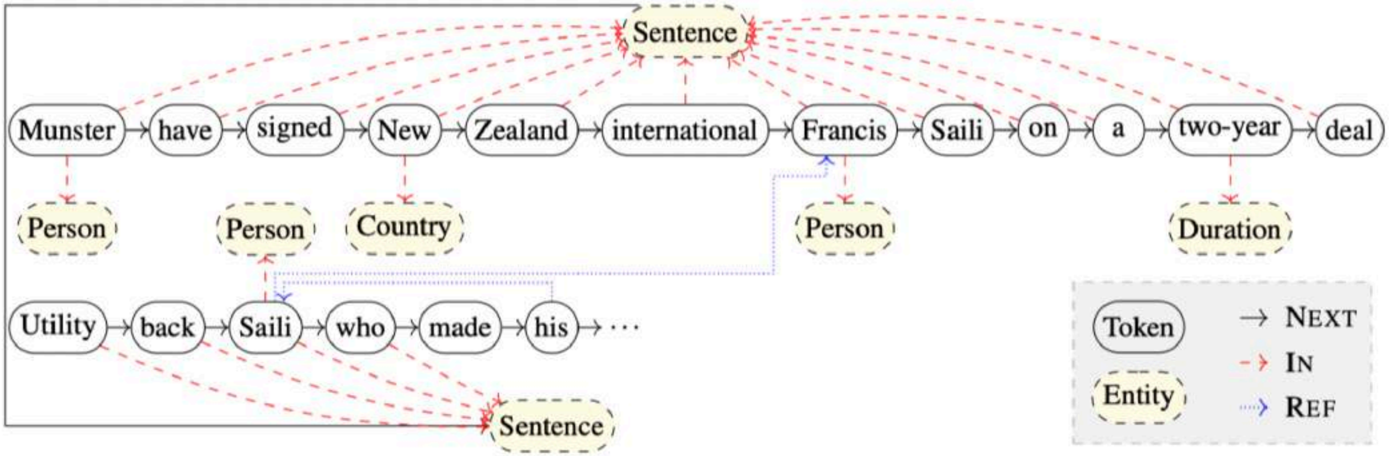
图神经网络

- Graph-based neural multi-document summarization ***CoNLL 2017***



Sentence Level

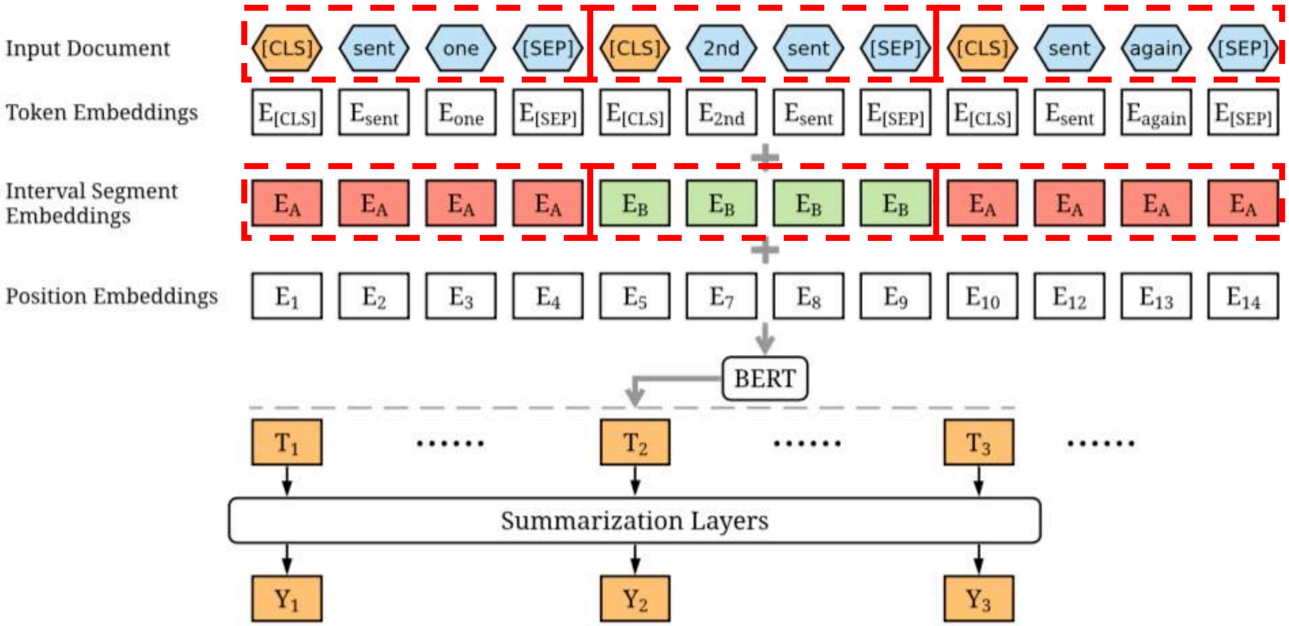
- Structured Neural Summarization ***ICLR19***



Word Level

基于预训练的方法

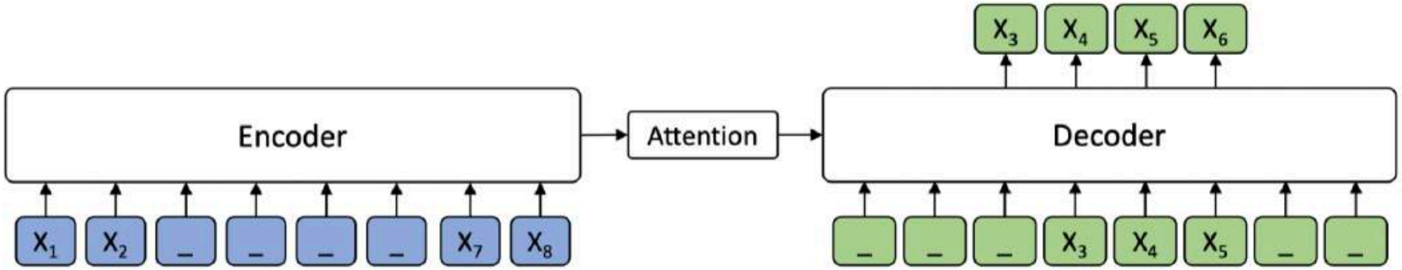
- Fine-tune BERT for Extractive Summarization
- 每个句子的开始添加CLS，结尾添加SEP
- 分句向量使用AB交替
- 输出：线性分类器、Transformer、LSTM

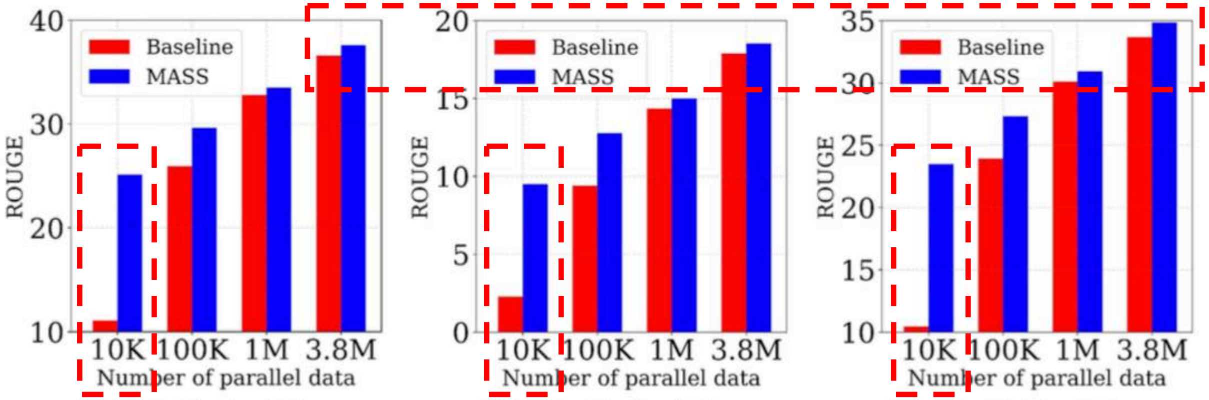


Input Document
[CLS]
sent
one
[SEP]
[CLS]
2nd
sent
[SEP]
[CLS]
sent
again
[SEP]
Token Embeddings
$E_{[CLS]}$ E_{sent} E_{one} $E_{[SEP]}$ $E_{[CLS]}$ E_{2nd} E_{sent} $E_{[SEP]}$ $E_{[CLS]}$ E_{sent} E_{again} $E_{[SEP]}$
Interval Segment Embeddings
E_A E_A E_A E_A E_B E_B E_B E_B E_A E_A E_A E_A
Position Embeddings
E_1 E_2 E_3 E_4 E_5 E_7 E_8 E_9 E_{10} E_{12} E_{13} E_{14}
BERT
T_1 ...... T_2 ...... T_3 ......
Summarization Layers
Y_1 Y_2 Y_3

基于预训练的方法

- MASS: Masked Sequence to Sequence Pre-training for Language Generation ***ICML19***

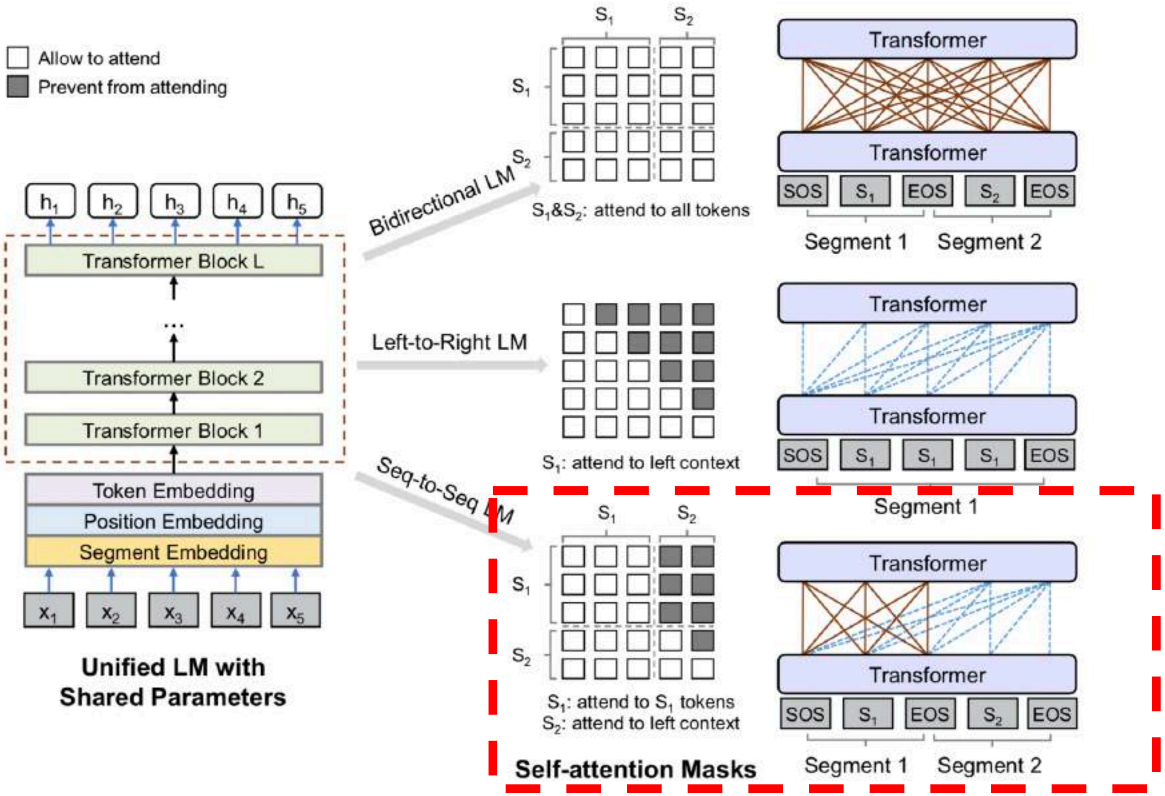




(a) RG-1 (F) (b) RG-2 (F) (c) RG-L (F)

基于预训练的方法

- Unified Language Model Pre-training for Natural Language Understanding and Generation



跨语言摘要

- A Robust Abstractive System for Cross-Lingual Summarization

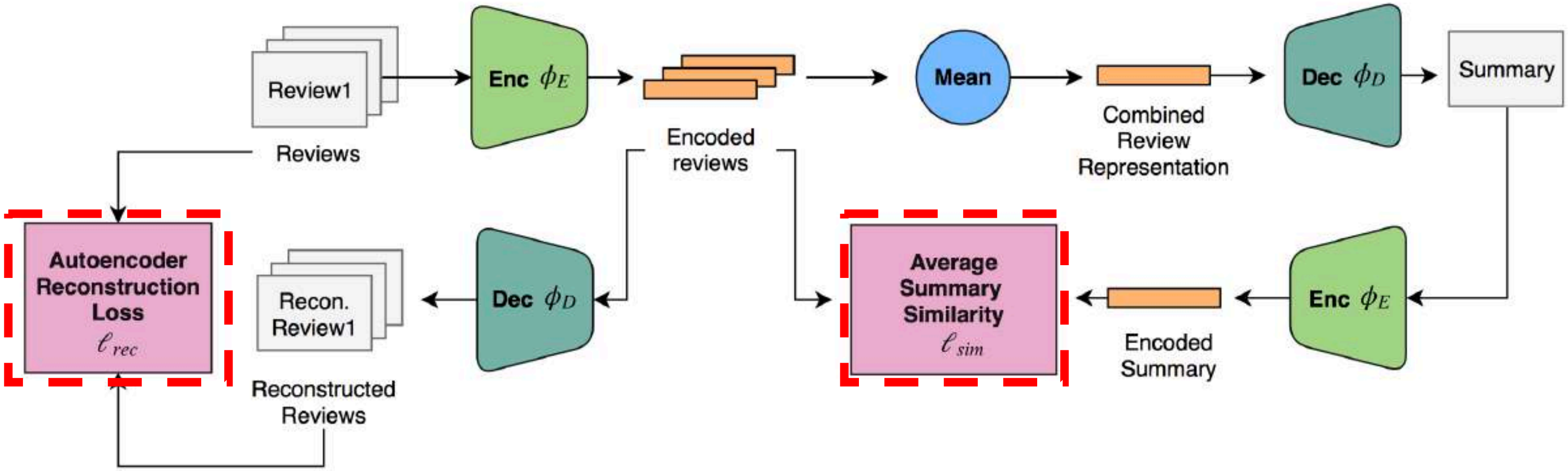
  NAACL19
- 问题
  - **先翻译后摘要：**计算代价、翻译错误
  - **先摘要后翻译：**对于资源稀缺的语言无法使用
- 方法
  - 英语→索马里语→带噪声的英语（索马里语语境）

    →斯瓦希里语→带噪声的英语（斯瓦希里语语境）

    →塔加拉族语→带噪声的英语（塔加拉族语语境）
  - **带噪声的英语**与**标准摘要**训练摘要系统

无监督生成式摘要

- MeanSum : A Neural Model for Unsupervised Multi-Document Abstractive Summarization ***ICML19***



Auto-encoder

目录

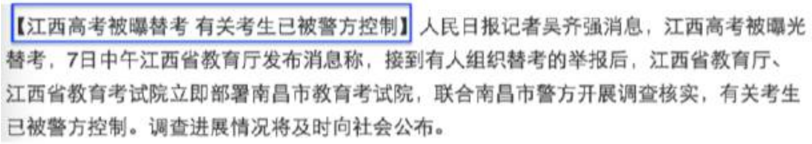
- 简介
- 分类
- 方法
- **评价指标**
- **数据集**
- 总结

评价指标

- Rouge (Recall-Oriented Understudy for Gisting Evaluation)
- 一种基于召回率的相似性度量方法，是评估自动文摘以及机器翻译的一组指标，考察翻译的**充分性**和**忠实性**。
- 用于文本摘要的评价指标主要有ROUGE-1、ROUGE-2、ROUGE-L三个指标，其计算分别涉及Uni-gram、Bi-gram和Longest Common Sub-sequence。

数据集

- **DUC**
  - Multiple标准摘要。
  - 数据规模小，因此常被用作测试集。
- **New York Times**
  - 1996至2007年期间的文章。
- **CNN/Daily Mail**
  - 多句摘要数据集。
  - 匿名版本和未匿名版本，未匿名版本包括了真实的实体名，匿名版本将实体使用特定的索引进行替换。
- **Gigaword**
  - 单句摘要数据集。
- **LCSTS**
  - 中文短文本摘要数据集，由新浪微博构建得到。

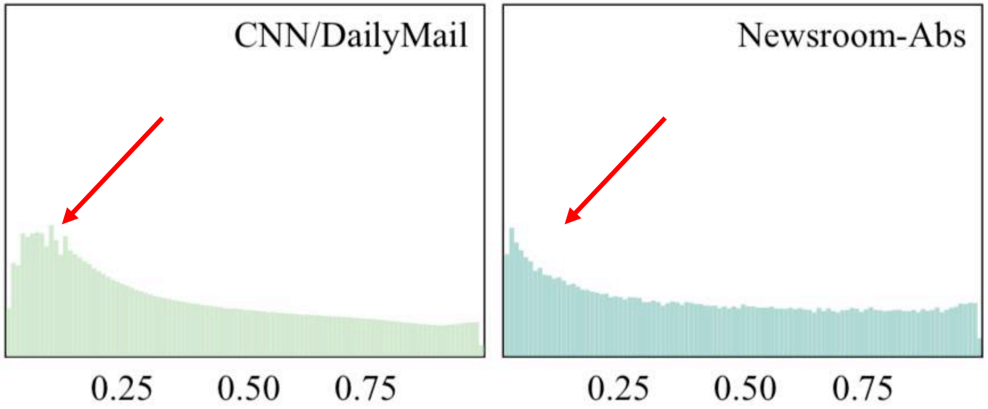


【江西高考被曝替考 有关考生已被警方控制】人民日报记者吴齐强消息，江西高考被曝光替考，7日中午江西省教育厅发布消息称，接到有人组织替考的举报后，江西省教育厅、江西省教育考试院立即部署南昌市教育考试院，联合南昌市警方开展调查核实，有关考生已被警方控制。调查进展情况将及时向社会公布。

- **Newsroom**
  - 130w
  - 社会媒体：新闻，体育，娱乐，金融，其他

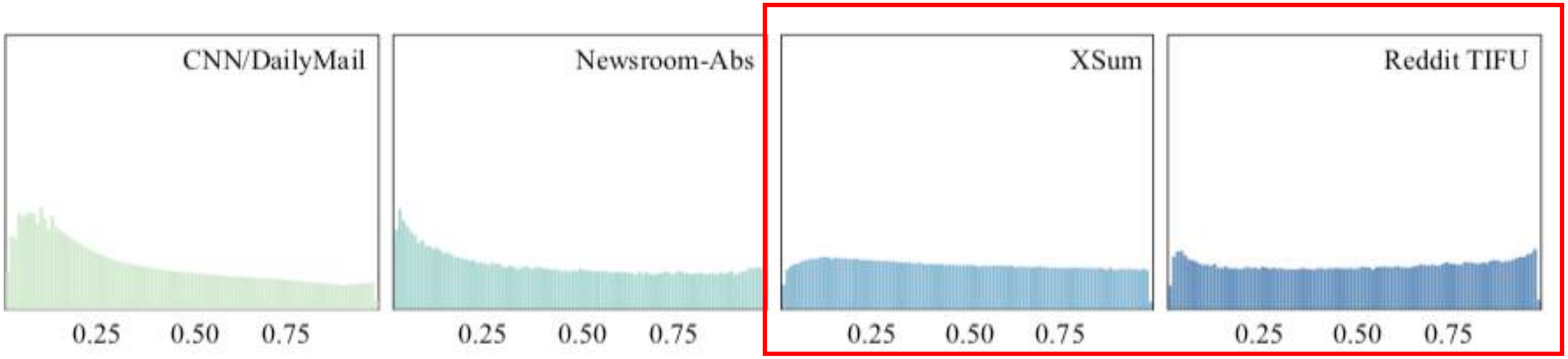
数据集

- Content Selection in Deep Learning Models of Summarization ***EMNLP18***
- **Lead Bias（摘要往往集中在文档的开始）**



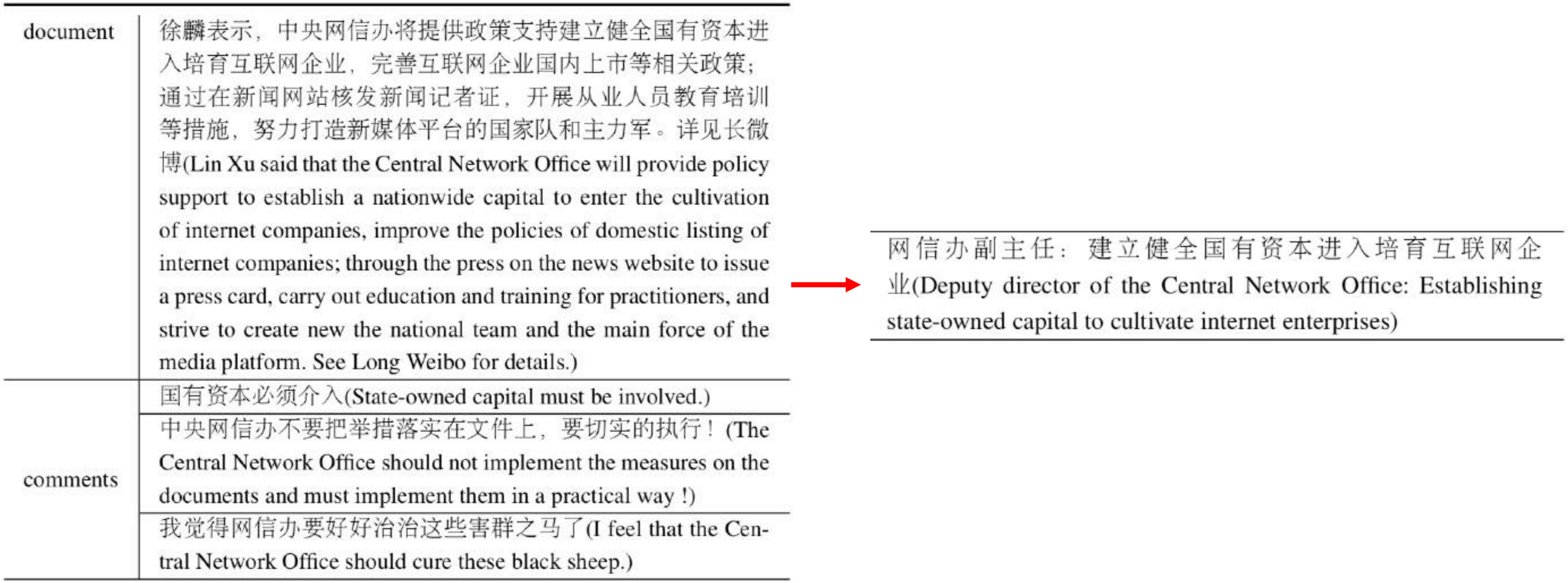
数据集

- **Xsum**
  - Don’t Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization ***EMNLP18***
  - BBC新闻
- **Reddit TIFU**
  - Abstractive Summarization of *Reddit* Posts with Multi-level Memory Networks ***NAACL19***
  - Reddit板块



新任务数据集

- Abstractive Text Summarization by Incorporating Reader Comments ***AAAI19***
- 任务：Reader-aware abstractive text summarization



document	徐麟表示，中央网信办将提供政策支持建立健全国有资本进入培育互联网企业，完善互联网企业国内上市等相关政策；通过在新闻网站核发新闻记者证，开展从业人员教育培训等措施，努力打造新媒体平台的国家队和主力军。详见长微博(Lin Xu said that the Central Network Office will provide policy support to establish a nationwide capital to enter the cultivation of internet companies, improve the policies of domestic listing of internet companies; through the press on the news website to issue a press card, carry out education and training for practitioners, and strive to create new the national team and the main force of the media platform. See Long Weibo for details.)
comments	国有资本必须介入(State-owned capital must be involved.)
	中央网信办不要把举措落实在文件上，要切实的执行！(The Central Network Office should not implement the measures on the documents and must implement them in a practical way !)
	我觉得网信办要好好治治这些害群之马了(I feel that the Central Network Office should cure these black sheep.)

网信办副主任：建立健全国有资本进入培育互联网企业(Deputy director of the Central Network Office: Establishing state-owned capital to cultivate internet enterprises)

目录

- 简介
- 分类
- 方法
- 评价指标
- 数据集
- **总结**

总结

- 文本摘要作为传统的自然语言处理任务，至今依旧有新的发展和创新。一方面得益于模型、 方法、语料的支撑，另一方面也是由于摘要任务自身的重要性。
- 摘要生成作为文本生成的一种，除了有着重复、冗余、不连贯、生成较短等问题，还有着摘要任务特定的问题，其中最核心的为：如何确定关键信息。当下的文本摘要更关注“什么是真正的摘要”，而不仅仅是简单地句子压缩。
- 在预训练模型兴起之后，摘要应该何去何从?

谢谢！