


大连理工大学
信息检索研究室
Information Retrieval Laboratory of DUT

Text Summarization

谭金源 2019-11-26

文本自动摘要意义

随着互联网产生的文本数据越来越多，文本信息过载问题日益严重，对各类文本进行一个“降维”处理显得非常必要，文本摘要便是其中一个重要的手段。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摘要。

文本自动摘要分类

- ✓ 按输入类型分类：单文档摘要、多文档摘要
- ✓ 按有无监督数据分类：有监督摘要、无监督摘要
- ✓ 按输出类型分类：抽取式摘要、生成式摘要

不同输出类型摘要

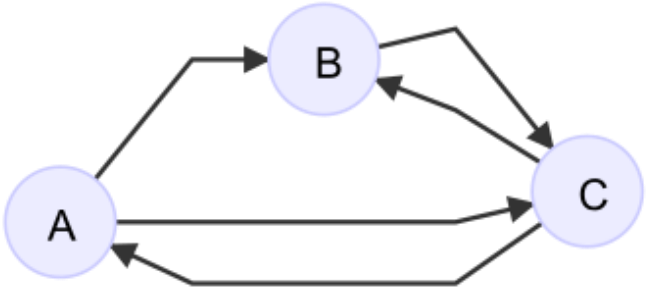
- ✓ 抽取式摘要（Extract）：从源文档中抽取关键句和关键词组成摘要，摘要全部来源于原文。
- ✓ 生成式摘要（Abstract）：不是单纯地利用原文档中的单词或短语组成摘要，而是从原文档中获取主要思想后以不同的表达方式将其表达出来。

常用方法

- ✓ 抽取式摘要（Extract）：句子打分排序
  - ➢ 基于统计学方法：词频、位置（特征工程）
  - ➢ 基于图排序方法：PageRank、TextRank、LexRank
- ✓ 生成式摘要（Abstract）：
  - ➢ Seq2Seq+Attention
  - ➢ Pointer Networks、Pointer-generator

PageRank

- ✓ PageRank：谷歌提出用于计算网页的重要性，当使用搜索引擎搜索网页的时候，越重要的网页越靠前显示。其基本思想是将网页抽象为图结点，网页之间的链接关系抽象为有向边，通过定义在这个有向图上的矩阵来计算各个网页之间的相对重要性。



$$\begin{bmatrix} I(A) \\ I(B) \\ I(C) \end{bmatrix} = \begin{bmatrix} 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 1 & 0 \end{bmatrix} \begin{bmatrix} I(A) \\ I(B) \\ I(C) \end{bmatrix}$$

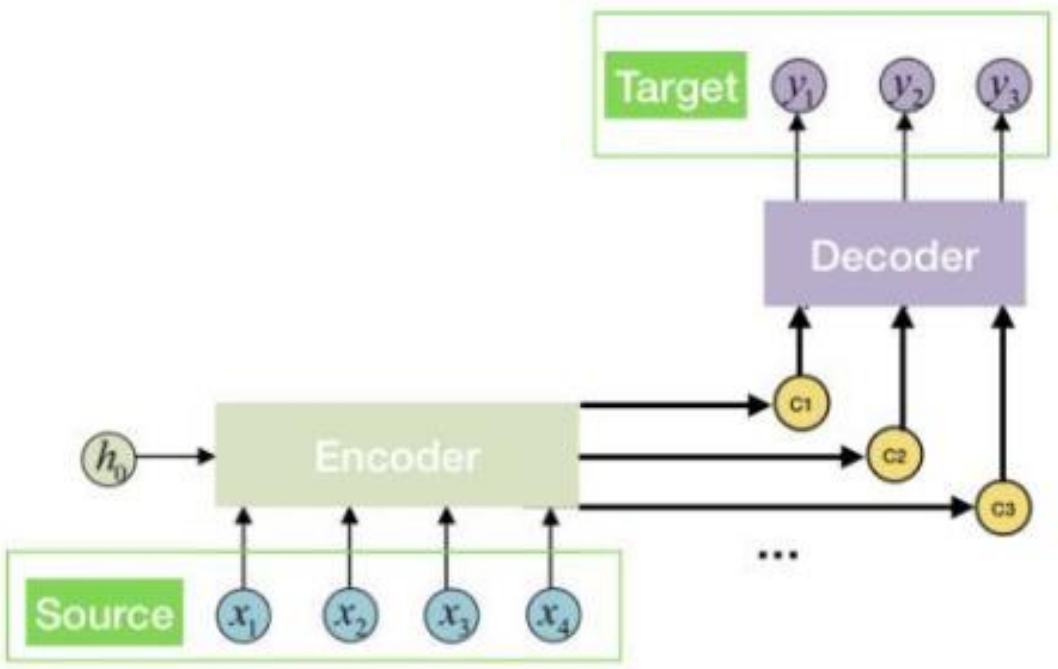
TextRank

- ✓ TextRank：如果一个单词出现在很多单词后面的话，那么说明这个单词比较重要。一个TextRank值很高的单词后面跟着的一个单词，那么这个单词的TextRank值会相应地因此而提高。

$$S(v_i) = (1-d) + d \sum_{(j,i)\in\varepsilon} \frac{w_{ji}}{\sum_{v_k \in out(v_j)} w_{jk}} S(v_j)$$

Seq2Seq+Attention

- ✓ 解决句子过长时效果不佳的问题。
- ✓ 采用beam search贪心策略



Pointer Network

$$
\begin{aligned}
u_j^i &= v^T \tanh(W_1 e_j + W_2 d_i) \quad j \in (1, \ldots, n) \\
a_j^i &= \text{softmax}(u_j^i) \quad j \in (1, \ldots, n) \\
d_i' &= \sum_{j=1}^{n} a_j^i e_j
\end{aligned}
$$

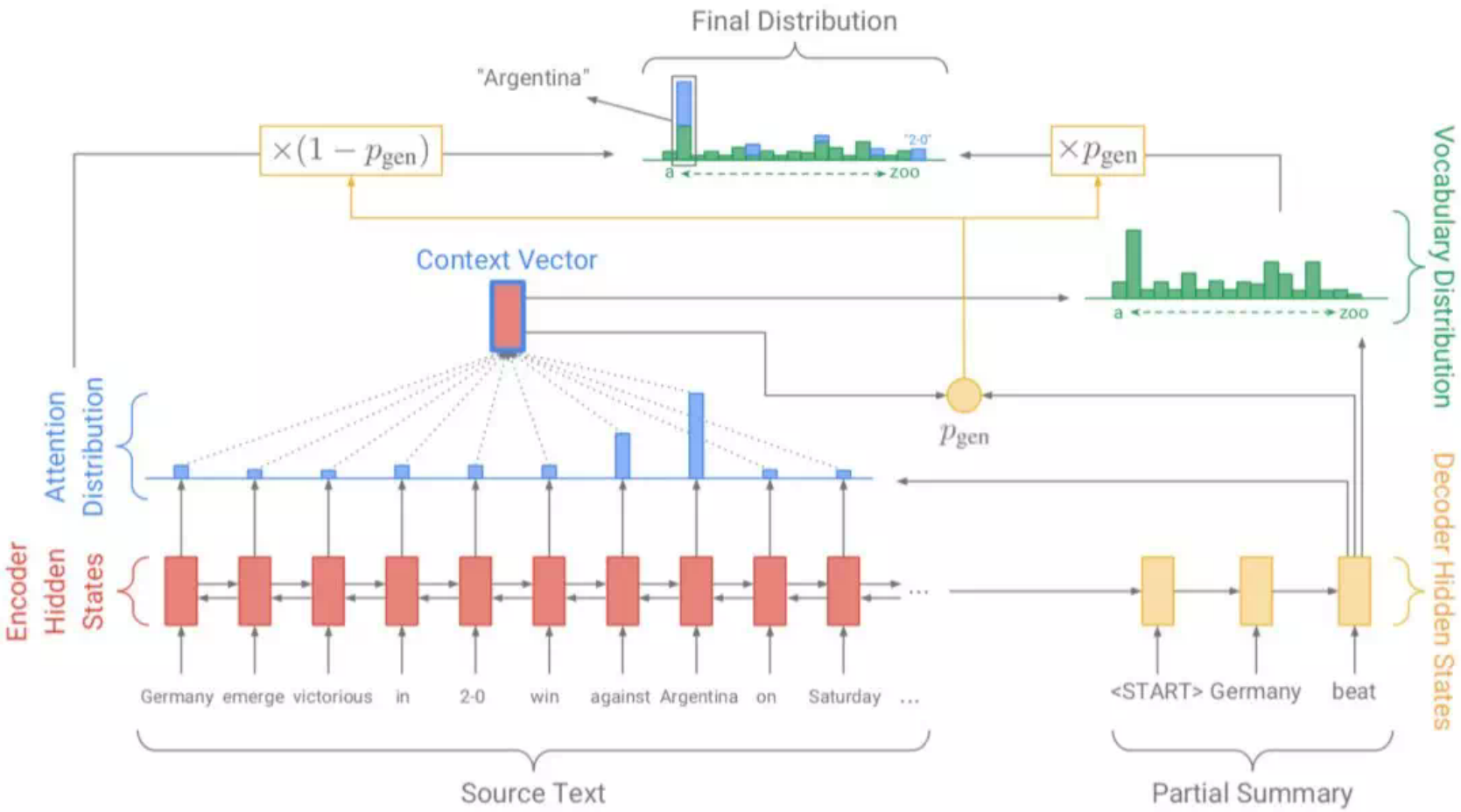
- ✓ Pointer Network:将权重关系a和隐式状态整合为context vector，而是直接进行通过softmax，指向输入序列选择中最有可能是输出的元素。

$$
\begin{aligned}
u_j^i &= v^T \tanh(W_1 e_j + W_2 d_i) \quad j \in (1, \ldots, n) \\
p(C_i | C_1, \ldots, C_{i-1}, \mathcal{P}) &= \text{softmax}(u^i)
\end{aligned}
$$

Pointer-generator

- ✓ 仅使用 Seq2Seq 摘要存在问题：未登录词问题、生成重复
- ✓ 在基于注意力机制的 Seq2Seq 基础上增加了 Copy 和 Coverage 机制
- ✓ Copy 机制：需要在解码的每一步计算拷贝或生成的概率，因为词表是固定的，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：需要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。

Pointer-generator



数据集——Gigaword（2012）

- News article（title+headline）
- 使用Rush，Chopra（2015）脚本进行处理
- 删除重复条目，标点符号， 过长摘要，’ 缩写
- Github: https://github.com/facebookarchive/NAMAS

Dataset	Train	Dev.	Test
Count	3.8M	189K	1951
AvgSourceLen	31.4	31.7	29.7
AvgTargetLen	8.3	8.3	8.8

CNN/Daily Mail（2015）

- 是否替换命名实体（named entities）
- 阅读理解任务——文本自动摘要任务
- Nallapati（2017） non-anonymized不匿名版本
- Dataset: https://cs.nyu.edu/~kcho/DMQA/
- Preprocess: https://github.com/abisee/cnn-dailymail

Dataset	Train	Dev.	Test
Count	287113	13368	11490

LCSTS（2015）

- 新浪微博爬虫（哈尔滨工业大学智能计算研究中心深圳研究生院）
- Part II、Part III手工标记打分（>=3）
- 网站申请数据集：

http://icrc.hitsz.edu.cn/Article/show/139.html

中文

Part I	2,400,591	
Part II	Number of Pairs	10,666
	Human Score 1	942
	Human Score 2	1,039
	Human Score 3	2,019
	Human Score 4	3,128
	Human Score 5	3,538
Part III	Number of Pairs	1,106
	Human Score 1	165
	Human Score 2	216
	Human Score 3	227
	Human Score 4	301
	Human Score 5	197

Table 1: Data Statistics

LCSTS（2015）

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

评价指标——ROUGE

- ROUGE-N：比较生成的摘要和参考摘要的n-grams（连续的n个词）。

  常用的有ROUGE-1，ROUGE-2，ROUGE-3。

- ROUGE-L：基于最长公共子序列（LCS）评价方式。生成的摘要和参考摘要的LCS越长，则生成的摘要质量越高。不足：n-grams一定是连续的。

- ROUGE-SU：允许bi-grams的第一个字和第二个字之间插入其他词，比ROUGE-L更灵活。

谢谢！