Flexible Non-Autoregressive Extractive Summarization with Threshold: How to Extract a Non-Fixed Number of Summary Sentences

贾瑞鹏, 曹亚男, 石海超, 方芳, 尹鹏飞, 王石

中国科学院信息工程研究所
中国科学院计算技术研究所
Background

- 我们进入了一个信息爆炸的时代，据 IDC 统计，互联网数据量已经达到 ZB 级别
  - bit, Byte, KB, MB, GB, TB, PB, EB, ZB, BB

图1 互联网信息爆炸
Background

- Massive data contains huge safety threats, including provocative statements, malicious comments, and yellow content, etc. Due to the large scale, complex structure, and diverse forms, the system's automatic detection accuracy is low, which requires professional personnel to perform content manual review, which is a time-consuming and labor-intensive process.
借助自动摘要技术，可以让机器深入理解、分析海量 Web 数据，自动生成文本的摘要，降低了内容审查的复杂度，在提升系统检测准确率的同时能够大大降低人工成本，在信息检索、内容过滤、舆情分析、态势感知等领域具有较高的研究价值和应用需求。
Summarization

- **Text Summary** refers to the automatic extraction/generation of a simple and coherent short text that accurately reflects the central content of the original document.

- **By number of documents:** Single-document automatic summarization / Multi-document automatic summarization
- **By generation method:** Extractive automatic summarization / Generative automatic summarization
- **By different purposes:** Indicative automatic summarization / Reporting automatic summarization
- **By whether context is provided:** Query-based automatic summarization / Ordinary automatic summarization
- ...
Summarization

目前，自动文摘技术应用无处不在，例如新闻标题、论文摘要、评论摘要、查询式摘要、金融报表、24小时实时新闻热点等。

但该课题的研究工作存在众多问题，现有的自动文摘算法不太成熟，整体效果还有很大的提升空间。
Summarization

- Common automatic text summarization algorithms:
  - Extractive summarization's limitations: Generated summaries are redundant and less coherent.
  - Generative summarization's limitations: Long text performance is poor, greedy decoding methods easily generate repetition, and sampling decoding methods are prone to semantic bias.
Deep Learning for Extractive Summarization

- 常用数据集包括:  CNN/DM, NYT, Gigaword

抽取式文摘常用技巧:
- 基于贪心算法将人工摘要转换为原文句子的 0/1 标签 (Nallapati, 2017)
- Trigram-Blocking (Liu, 2019)
Deep Learning for Extractive Summarization

• **研究目标**: 解决目前基于深度学习抽取式摘要中的一些常见的问题

1. 抽取的句子之间存在大量的冗余信息，但现在常用的方法是 Trigram-Blocking算法消除冗余的方法与实际情况不符；例如：在 CNN/DM 测试集中
   • 7.35% 的 Oracle Summary 中存在 Trigram-Overlaps
   • 0% 的基于 Trigram-Blocking 算法抽取的 Summary 中存在 Trigram-Overlaps

2. 常见的基于深度学习的抽取式方法都采用 Top-K 策略来抽取句子
Deep Learning for Extractive Summarization

- 研究内容:
  - 针对被抽取句子间的冗余信息，采用预标签+迭代更新的网络模型
  - 针对 Top-K 的抽取测略，采用基于 Threshold 的方法来增加抽取文摘句子的灵活性
ThresSum

- 模型架构
Algorithm 1: Teacher Algorithm for Soft Labels

Initialize Sentence Set $D = \{s_1, ..., s_M\}$;
Initialize ROUGE $r_1, ..., r_M$, and Iteration Steps $L$;
Sort $D$ by $r_1, ..., r_M$ in descending order;

for $l$ from 0 to $L - 1$ do

    Set the Temperature $T$ as $L - l$;

    for $t$ from $T$ to 1 do

        Temporary Sentence Set: $D_{\text{temp}} \leftarrow \{\}$;
        Temporary ROUGE of $D_{\text{temp}}$: $R_{\text{temp}} \leftarrow 0$;

        for $s_i$ from $D[0]$ to $D[\text{end}]$ do

            $D_{\text{temp}} \leftarrow D_{\text{temp}} + s_i$;
            if $R_{\text{temp}}$ is increasing then
                $D \leftarrow D - s_i$
            else
                $D_{\text{temp}} \leftarrow D_{\text{temp}} - s_i$
            end

        end

    end

    Set the Sentence $s$ in $D_{\text{temp}}$ with Soft Label $\frac{t}{T}$;

end

Set the Sentence $s$ Remained in $D$ with Label 0;
Record these Soft Labels as $(y_1^l, y_2^l, ..., y_M^l)$;
Re-Initialize Sentence Set $D = \{s_1, ..., s_M\}$;
Re-Sort $D$ by $r_1, ..., r_M$ in descending order;

ThresSum

• 知识蒸馏算法获得软标签
Experiments

- 实验结果

| Models                        | CNN/DM |       |       | NYT  |       |
|-------------------------------|--------|-------|-------|------|-------|
|                               | R-1    | R-2   | R-L   | R-1  | R-2   | R-L   |
| **Abstractive**               |        |       |       |      |       |
| ABS (2015)                    | 35.46  | 13.30 | 32.65 | 42.78| 25.61 | 35.26 |
| PGC (2017)                    | 39.53  | 17.28 | 36.38 | 43.93| 26.85 | 38.67 |
| TransformerABS (2017)         | 40.21  | 17.76 | 37.09 | 45.36| 27.34 | 39.53 |
| T5\textit{Large} (2019)       | 43.52  | 21.55 | 40.69 | -    | -     | -     |
| BART\textit{Large} (2019b)    | 44.16  | 21.28 | 40.90 | 48.73| 29.25 | 44.48 |
| PEGASUS\textit{Large} (2019b)| 44.17  | 21.47 | 41.11 | -    | -     | -     |
| ProphetNet\textit{Large} (2020)| 44.20 | 21.17 | 41.30 | -    | -     | -     |
| **Extractive**                |        |       |       |      |       |
| Oracle (Sentence)             | 55.61  | 32.84 | 51.88 | 64.22| 44.57 | 57.27 |
| Lead-3 †                      | 40.42  | 17.62 | 36.67 | 41.80| 22.60 | 35.00 |
| SummaRuNNer † * (2017)        | 39.60  | 16.20 | 35.30 | 42.37| 23.89 | 38.74 |
| Exconsummm † * (2019)         | 41.7   | 18.6  | 37.8  | 43.18| 24.43 | 38.92 |
| PNBERT\textit{Base} † * (2019a)| 42.69 | 19.60 | 38.85 | -    | -     | -     |
| DiscoBERT\textit{Base} (2020)| 43.77  | 20.85 | 40.67 | -    | -     | -     |
| BERTSUMEXT\textit{Large} † * (2019)| 43.85 | 20.34 | 39.90 | 48.51| 30.27 | 44.65 |
| MATCHSUM\textit{Base} † * (2020)| 44.41 | 20.86 | 40.55 | -    | -     | -     |
| ThresSum\textit{Large} † (Ours) | **44.59** | **21.15** | **40.76** | **50.08** | **31.77** | **45.21** |
Future

1. 优化基于 Greedy 的文本解码算法，减少重复性，增加多样性

2. 常见的 TextRank 的无监督文本摘要算法目前有很多的弊端，研究一种全新的无监督摘要框架
THANKS

2020.12.19