自然语言处理与文本挖掘 作业二
自动摘要

2020.04.09

本次实验希望大家理解和应用常见的自动摘要方法。

数据集说明：

我们使用 LCSTS[1] 数据集的 Part I 部分来进行模型的巡礼和测试。该数据集包含了 200 万真实的中文短文本数据和每个文本作者给出的摘要，同时作者团队也手动标注了 10666 份文本的摘要。

./DATA/PART_I.txt 包含 2,400,591 个(short_text, summary)对，示例如下：

```
<doc id=0>
    <summary>
        修改后的立法法全文公布
    </summary>
    <short_text>
        新华社受权于18日全文播发修改后的《中华人民共和国立法法》，修改后的立法法分为“总则”“法律”“行政法规”“地方性法规、自治条例和单行条例、规章”“适用与备案审查”“附则”等6章，共计105条。
    </short_text>
</doc>
```

合理划分训练、验证和测试集后，可将<short_text>字段用作模型输入，<summary>字段用作 ground_truth，训练模型，进而完成实验。

数据集中还包含了原作者的工作 paper，供大家参考。其他具体描述详见 http://icrc.hitsz.edu.cn/Article/show/139.html。

结果评价指标应至少包括 ROUGE-2、ROUGE-L，自行实现测试结果评价脚本。

作业提交说明：

需要提交的内容有：报告文档（包括结果数据，分析等，见题目具体要求），程序源代码及其运行方法（可以复现报告中的结果数据）

编程所使用的语言不限。

1. 基于特征打分排序的抽取式方法
   实现一种基于特征打分排序的抽取式文本自动摘要方法，例如：SumTF-IDF、SumBasic 等（特征选择不局限于 TF-IDF），并在测试数据上评价其性能。从特征和排序方法两方面讨论如何提高方法的性能，实验验证你的想法。
2. 基于图排序的抽取式方法
   实现一种基于图排序的抽取式文本自动摘要方法，例如：PageRank、TextRank 等，并在测试数据上评价其性能。比较该类方法与特征打分排序方法的不同。
3. （选做）基于神经网络的概括式方法
   实现一种基于神经网络的概括式文本自动摘要模型，给出你的模型设计、训练和测试的流程，以及测试集上的性能指标，尝试分析注意力机制的作用，并与前面两个模型进行比较，分析不同模型的特点。

[1] Hu B, Chen Q, Zhu F. LCSTS: A Large Scale Chinese Short Text Summarization Dataset[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1967-1972.