CCL2016“自然语言处理国际前沿动态综述报告会”

文本自动摘要：现状与未来

万小军

语言计算与互联网挖掘研究室
北京大学计算机科学技术研究所

2016年10月16日，山东烟台

http://www.icst.pku.edu.cn/lcwm/wanxj

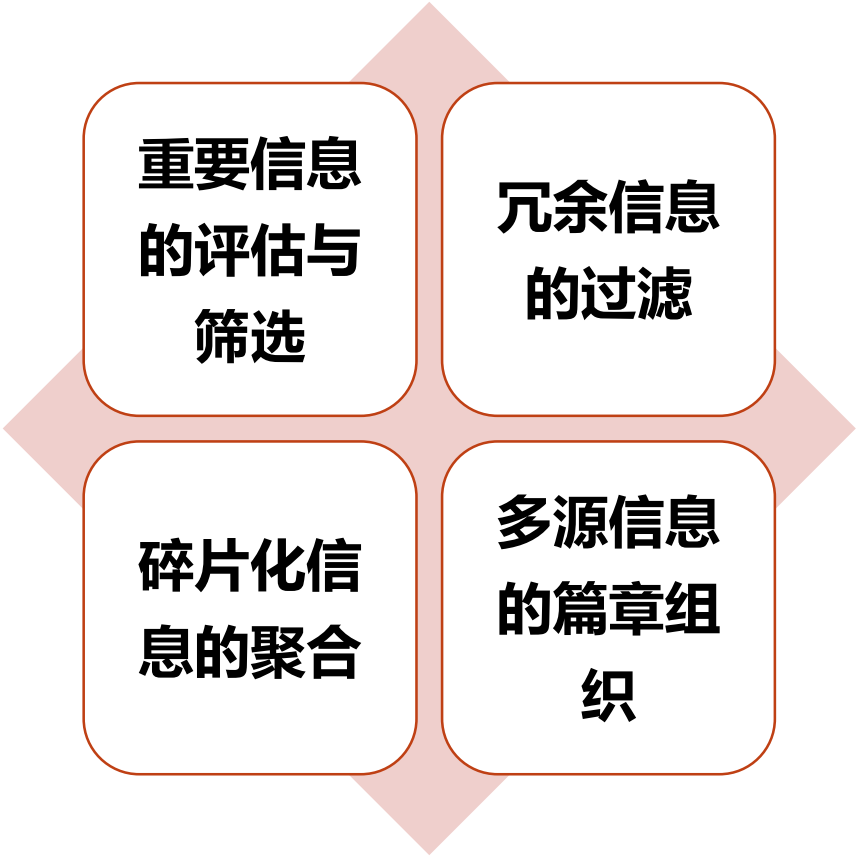
文本自动摘要的重要性

- 信息爆炸
  - 据IDC统计，互联网数据量已跃至ZB级别（1ZB=2^{40}GB），预计2020年达到35ZB

- 搜索引擎不能有效解决信息过载的问题
  - 相关信息过多：冗余、片面、杂质

- 移动设备/智能设备的普及使用
  - 需要新的信息浏览与人机交互方式

文本自动摘要的关键科学问题



文本自动摘要主要方法

- **抽取式方法**
  - 实现简单，保留完整句子，可读性良好
  - 基于启发式规则或机器学习进行句子评估与选择
  - 基于组合优化方法进行摘要提取
- **压缩式方法**
  - 同时进行句子抽取与压缩或融合
  - 能有效提高ROUGE值，但会牺牲句子可读性
- **生成式方法**
  - 直接从意义表达生成摘要句子
  - 难度大，更接近摘要的本质
  - 目前效果不佳，但值得鼓励

2015年以来文本自动摘要相关论文

NLP领域主要会议与刊物	发表数量（长+短）
ACL2015	7
ACL2016	6
EMNLP2015	17
EMNLP2016	2
NAACL2015	7
NAACL2016	4
TACL(2015~2016)	0
CL(2015~2016)	0
TASLP(2015~2016)	4

近两年主要研究进展总结

- 数据集与评测手段
  - 改进文本自动摘要的评价方法
  - 构建大规模文本自动摘要数据集
- 对已有抽取式方法的改进与扩展
  - 整数线性规划(ILP)
  - 次模函数最大化
  - 有监督学习方法
- 提出新的生成式摘要方法
  - 基于短语选择与合并的生成式摘要方法
  - 基于AMR语义图的生成式摘要方法

近两年主要研究进展总结

- 对新的摘要任务的研究
  - 跨语言摘要
  - 演化式摘要(Timeline)
  - 观点摘要

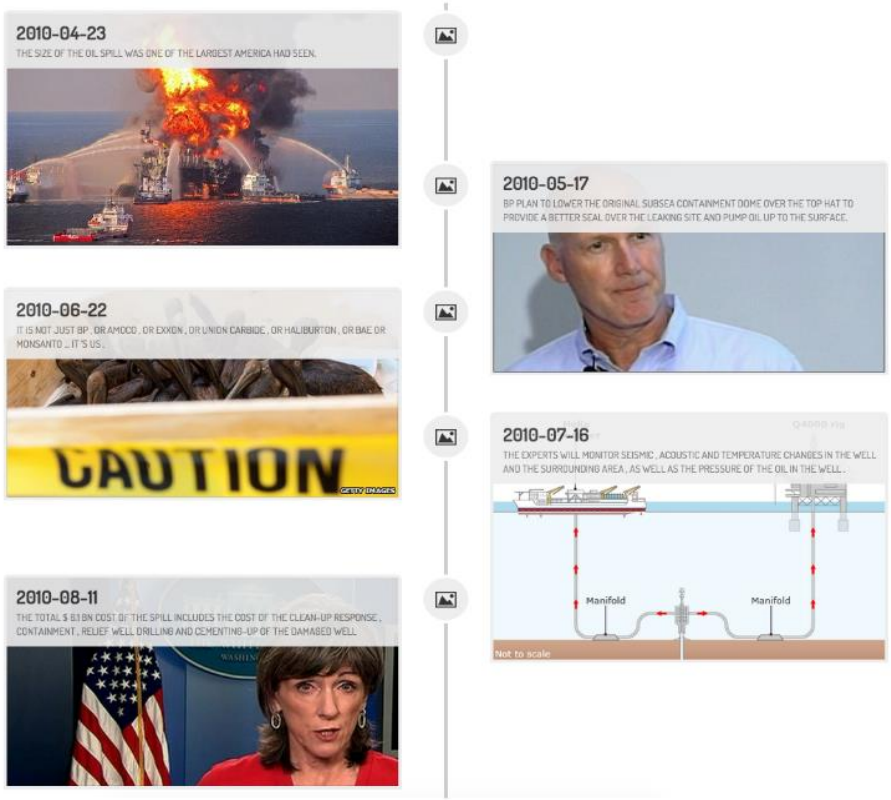


Figure 1: A timeline example for the BP oil spill generated by our proposed method. *Note that we use Yahoo! Image Search to obtain the top-ranked image for each candidate sentence.*

近两年主要研究进展总结

- 针对新的文本类型进行自动摘要
  - 学术文献
  - 会议记录
  - 电影剧本
  - 学生反馈(Student Response)
  - 软件代码
  - 直播文字

```
1. Source Code (C#):
public int TextWidth(string text) {
  TextBlock t = new TextBlock();
  t.Text = text;
  return
    (int)Math.Ceiling(t.ActualWidth);
}
Descriptions:
a. Get rendered width of string rounded up to
the nearest integer
b. Compute the actual textwidth inside a
textblock

2. Source Code (C#):
  var input = "Hello";
  var regEx = new Regex("World");
  return !regEx.IsMatch(input);
Descriptions:
a. Return if the input doesn't contain a
particular word in it
b. Lookup a substring in a string using regex
```

```
3. Source Code (SQL):
SELECT Max(marks) FROM stud_records
  WHERE marks <
  (SELECT Max(marks) FROM stud_records);
Descriptions:
a. Get the second largest value of a column
b. Retrieve the next max record in a table
```

伊万右路45度斜传球到禁区，科茨解围球直接停给了佩德罗，佩德罗将球一拨，左脚抽射，打进！！！	上半场 14'	2-0
威廉前场右路拿球，强突后分边给上来的伊万	上半场 14'	2-0
伊万没机会，回传给到小法	上半场 15'	2-0
小法拿球内切，斜塞禁区找插入的威廉	上半场 15'	2-0
球被回防的范安霍尔特抢先出脚碰出底线，角球	上半场 15'	2-0

近两年主要研究进展总结

- 基于深度学习的文本自动摘要
  - sequence-to-sequence模型/encoder-decoder框架及其变体
    - 应用于句子压缩(句子摘要)任务
      - 词序列=>选择标签[0,1]序列
      - 词序列=>词序列
    - 应用于观点摘要任务
      - 词序列=>词序列
    - 应用于单文档摘要任务
      - 句子序列=>句子选择标签[0,1]序列，并进一步预测词序列

近两年主要研究进展总结

- 基于深度学习的文本自动摘要
  - 相比其它NLP任务，深度学习技术较晚&较少应用于文本摘要任务
    - 任务的特殊性：子集选择问题/压缩问题
    - 数据规模（尤其是多文档摘要任务）
    - 摘要的相对不一致性
    - 长文档的语义编码
  - 目前已有多种尝试，但总体性能提升并不明显

文本摘要技术的发展趋势

- 针对传统摘要任务的进一步探索
  - 最新组合优化模型的使用
  - 基于自然语言生成的文本摘要
  - 基于深度学习的文本摘要
  - 篇章信息和语义信息的有效利用

文本摘要技术的发展趋势

- 新型文本摘要任务与应用
  - 基于文本摘要的自动写稿：综述生成、自动作文
  - 结合文本摘要与语音合成技术的新闻自动播报
  - 多语言/跨语言文本摘要
  - 针对不同领域多类型文本的摘要
  - 面向复杂问题回答与人机交互的文本摘要

谢谢大家！

QA: wanxiaojun (AT) pku.edu.cn