- 文本摘要
  - 抽取式
    - 传统
      - Lead-3：抽取文章的前三句作为文章的摘要。
      - TextRank：仿照 PageRank，句子作节点，用句子间相似度构造无向有权边。使用边上的权值迭代更新节点值，最后选取 N 个得分最高的节点，作为摘要。
      - 聚类：编码得到句子的向量，再用 K 均值和 Mean-Shift 聚类进行句子聚类，得到 N 个类别。从每个类别中选择距离质心最近的句子，得到 N 个句子，作为最终摘要。
    - 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），最终摘要由所有标签为 1（属于摘要）的句子构成。
      - SummaRuNNer 模型：需要监督数据，现有数据集往往没有对应的句子级别的标签，因此需要通过启发式规则进行获取。具体方法为：首先选取原文中与标准摘要计算 ROUGE 得分最高的一句话加入候选集合，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。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。
      - Latent 模型（序列标注结合Seq2Seq）：摘要数据集往往没有对应的句子级别的标签，需要通过启发式规则获取，然而仅仅利用这些标签训练模型会丢失很多标准摘要中重要的信息。因此 Latent 模型不采用序列标注方法计算标签级别的损失来训练模型，而是将序列标注作为中间的步骤。在得到序列标注的概率分布之后，从中采样候选摘要集合，与标准摘要对比计算损失，可以更好地利用标准摘要中的信息。
    - Seq2Seq：直接使用 Seq2Seq 模型来交替生成词语和句子的索引序列来完成抽取式摘要任务。其模型 SWAP-NET 在解码的每一步，计算一个 Switch 概率指示生成词语或者句子。最后解码出的是词语和句子的混合序列。最终摘要由产生的句子集合选出。除了考虑生成句子本身的概率之外，还需要考虑该句是否包含了生成的词语，如果包含，则得分高，最终选择 top k 句作为摘要。
      - 考虑生成的词语
    - 句子排序：序列标注对于每一个句子表示打一个 0、1 标签，而句子排序则是针对每个句子输出其是否是摘要句的概率。依据概率，选取 top k 个句子作为最终摘要。
      - NeuSUM 模型：之前的模型都是在得到句子的表示以后对于句子进行打分，后根据得分进行选择。没有利用到句子之间的关系。这个模型使用句子受益作为打分方式，句子编码部分与之前相同，打分和抽取部分使用单向 GRU 和双层 MLP 完成。单向 GRU 用于记录过去抽取句子的情况，双层 MLP 用于打分，逐步选择使得 g 最高的句子。
  - 生成式
    - Seq2Seq
      - Pointer-Generator 模型：模型基本部分为基于注意力机制的 Seq2Seq 模型，使用每一步解码的隐层状态与编码器的隐层状态计算权重，最终得到 context 向量，利用 context 向量和解码器隐层状态计算输出概率。
        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。
      - 利用外部信息：基于 Seq2Seq 的模型往往对长文本生成不友好，对于摘要来说，更像是一种句子压缩，而不是一种摘要。其核心想法在于：相似句子的摘要也具有一定相似度，将这些摘要作为软模板，作为外部知识进行辅助。其模型：
        Retrieve 部分主要检索相似句子，获得候选摘要。
        Rerank 部分用于排序候选模板，在训练集中，计算候选与真实摘要的 ROUGE 得分作为排序依据，在开发集与测试集中，使用神经网络计算得分作为排序依据。训练过程中，使得预测得分尽可能与真实得分一致。
        Rewrite 部分，结合候选模板与原文生成摘要。
      - 多任务学习：将摘要生成作为主任务，问题生成（要求模型具有选择重要信息的能力）、蕴含生成（要求模型具有逻辑推理能力）作为辅助任务进行多任务学习定位原文中的关键信息。根据原文生成摘要又具有一定的逻辑推理能力，使得生成的摘要与原文不违背，无矛盾。
      - 生成对抗摘要：利用生成模型 G 来生成摘要，利用判别模型 D 来区分真实摘要与生成摘要。使用强化学习的方法，更新参数。

Summary：当前最流行、效果最显著的是基于深度学习的生成文本摘要。
利用了计算机强大的性能对文档的局部以及上下文的多维特征同时学习，对特征进行编码向量化，使文档的上下文特征、句法特征、语义特征等多维特征转化为能够进行计算的向量特征，方便利用深层网络对其进行训练学习，在文本摘要质量上实现了许多最优的实验结果。

表 1　各文本摘要方法的优缺点

方法	优点	缺点
基于统计学方法	依据文本形式上的规律，简单直观，避免考虑复杂的句法、语法结构，易于实现且应用广泛，无需训练数据，执行速度快	只是单纯利用了单词表层特征，没有充分挖掘词义关系和语义特征，存在较大局限性
基于外部语义资源方法	在统计学方法的基础上利用词间关系、词义关系进行了改进，使文本摘要的语义性能得到了一定的提高	受收录词汇的限制比较大；对于文章题目依赖程度较高；分词对关键词的影响较大；相似度阈值的选取对构建词汇链有影响。语法语义结构不连贯
基于图排序方法	适用于结构较为松散且涉及主题较多的结构；计算句子权重的同时可以充分考虑词汇之间、词组之间或句子之间的全局关系；无监督，语言独立，不需要对大量语料进行处理	通常只考虑了句子节点间的相似性关系，而忽略了文档篇章结构以及句子上下文的信息；相似度计算的好坏决定了关键词和句子重要性排序的正确与否；对数据的利用不够充分；没有考虑信息冗余
基于统计机器学习方法	特征选择和训练分类器的选择上有较大的可供选择范围，还可以综合一些开放性特征提高分类的精度	需要人工标注的数据集；效果严格依赖于训练数据质量的好坏；监督或半监督，执行速度较无监督的方法慢
基于深度学习方法	降低了对人工的依赖，可以高效地进行训练；可以与多种神经网络结构和 Sequence-to-Sequence 模型结合，生成文本摘要的可读性和准确度高	可解释性差；需要大量人工标注的数据集；由于有复杂的神经网络结构的引入，执行速度慢，需要花费相对较长的时间；对计算机性能有一定的要求