

**课 程 报 告**

**题目： 文本摘要生成技术**

**院（系）：计算机科学与技术学院**

**课程名称：自然语言处理**

**学 号：202215561**

**姓 名：瞿明睿**

**任课老师：魏巍**

**报告日期：2025.5.10**

·目录

[1绪论 1](#_Toc198125438)

[1.1课题背景与意义 1](#_Toc198125439)

[1.1.1分类 1](#_Toc198125440)

[1.1.2 现实应用举例 1](#_Toc198125441)

[1.2现有模型的性能展望 2](#_Toc198125442)

[2抽取式摘要方法 3](#_Toc198125443)

[2.1 传统方法 3](#_Toc198125444)

[2.2.1基于神经网络的句子评分 3](#_Toc198125445)

[2.2.2图与网络的问题 4](#_Toc198125446)

[3生成式摘要方法 6](#_Toc198125447)

[3.1 实现概况 6](#_Toc198125448)

[3.2 深度学习方法 6](#_Toc198125449)

[3.2.1 RNN-based Seq2Seq模型 6](#_Toc198125450)

[3.2.2 注意力机制 6](#_Toc198125451)

[3.2.3 Transformer模型 7](#_Toc198125452)

[3.3现有模型代表 7](#_Toc198125453)

[4总结与展望 8](#_Toc198125454)

[参考文献 9](#_Toc198125455)

# 1绪论

## 1.1课题背景与意义

随着互联网技术的飞速发展和信息量的爆炸式增长，如何从海量文本中快速提取关键信息成为自然语言处理领域的重要研究方向之一。文本摘要作为NLP的核心任务之一，旨在通过算法将冗长的文本压缩为简洁、准确的摘要，同时保留其主要语义和核心信息。

近年来，随着深度学习和大语言模型的发展，文本摘要技术取得了显著进展。特别是基于Transformer架构的预训练语言模型（如BERT、T5、BART、PEGASUS等），极大地提升了生成质量与效率。然而，这些方法仍面临诸多挑战，如生成内容的事实一致性问题、输出稳定性不足、数据偏见影响等。

文本摘要在多个实际场景如新闻媒体，医疗健康，法律金融，教育科研等中具有广泛的应用价值，因此深入研究文本摘要技术，探索更高效、鲁棒的方法，对于推动人工智能在信息处理领域的应用具有重要意义。

### 1.1.1分类

抽取式摘要：从原文中选取重要句子直接组合成摘要。代表方法包括TextRank、TF-IDF、LSA以及基于深度学习的BERTSUM等。这类方法的优点是能够较好地保留原文信息，但生成的摘要可能存在语义不连贯、重复等问题。

生成式摘要：利用语言生成模型创造新的句子来表达原文的核心内容。代表模型包括T5、BART、PEGASUS、GPT系列等。这类方法生成的摘要更加自然流畅，但也可能引入事实错误或幻觉现象。

近年来还出现了融合两者优势的混合式摘要方法，尝试在信息保留与语言生成之间取得平衡。

### 1.1.2 现实应用举例

以下是一些典型的现实应用场景及其成效：

新闻媒体 ：美联社开发了AI摘要系统，可自动生成体育赛事亮点摘要，使记者编辑时间减少70%；KSAT-TV使用AI将视频转录压缩为800字摘要，错误率低于2%。

医疗健康 ：IBM Watson Health系统可以从电子病历中提取治疗方案摘要，医生文档处理时间减少60%，肺结节描述准确率达92%。TRAK方法还可用于消除种族相关术语的过度使用。

法律金融 ：ROSS Intelligence提供法律案例摘要服务，响应速度从小时级提升至秒级，准确率更高。采用双AI+人工复核机制确保关键判决书的可靠性。

这些应用案例表明，文本摘要技术已经在多个行业落地并产生实际效益，未来仍有巨大的发展潜力。

## 1.2现有模型的性能展望

为了更直观地比较不同文本摘要模型的性能，列出下表。

表1.1模型性能表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型 | 类型 | ROUGE-1分数 | 处理速度 | 优点 | 缺点 |
| T5 | 生成式 | 0.45 | 300 | 生成流畅 | 事实错误多 |
| TextRank | 抽取式 | 0.42 | 500 | 无需训练 | 语义不连贯 |
| BERTSUM | 抽取式 | 0.38 | 500 | 高准确性 | 生成内容单一 |
| PEGASUS | 生成式 | 0.47 | 350 | 长文本优化 | 训练成本高 |
| GPT-3 | LLM | 0.5 | 200 | 高质量 | 提示词敏感 |

数据来源：MIT 2024年LLM研究（arXiv:2403.02901），HuggingFace模型库

从上表可以看出，虽然GPT-3在ROUGE-1得分上表现最佳，但其对输入提示词非常敏感，可能影响摘要的稳定性。而TextRank和BERTSUM虽然处理速度快、准确率较高，但缺乏语言生成能力，摘要内容较为机械。相比之下，T5和PEGASUS在生成质量方面表现优异，尤其适合长文本摘要任务，但由于依赖复杂的神经网络结构，训练和推理成本相对较高。

此外，在实际应用中还需关注数据偏见与算法偏见问题。例如，训练数据中性别、种族分布不均可能导致模型在生成摘要时忽略某些群体的信息；注意力机制也可能放大特定群体的表述，造成不公平结果。这些问题将在后续章节中进一步探讨。

# 2抽取式摘要方法

## 2.1 传统方法

传统的抽取式摘要方法主要依赖于手工特征和统计模型，常见的方法包括：

TF-IDF：一种经典的文本特征提取方法，通过计算词语在文档中的出现频率与在整个语料库中的逆文档频率，衡量词语的重要性。

TextRank：一种基于图模型的排序算法，灵感来源于Google PageRank。其基本思想是将每个句子视为图中的一个节点，节点之间的边表示句子间的相似性。通过迭代更新节点的权重，最终选出权重最高的若干句子作为摘要。相似性通常通过余弦相似度计算。

LSA：通过奇异值分解文本矩阵提取潜在语义空间中的关键词和句向量，再基于余弦相似度选择关键句子。

这些方法在早期摘要任务中发挥了重要作用，但由于依赖手工特征，难以应对复杂语义和多义词等问题，逐渐被深度学习方法所取代。

**2.2** 深度学习方法

随着神经网络的发展，抽取式摘要方法也进入了深度学习时代。目前常用的深度学习方法主要包括以下几类

### 2.2.1基于神经网络的句子评分

基于RNN的方法，LSTM和GRU能够有效缓解梯度消失问题，捕捉句子中的长期依赖关系。LSTM结构包括输入门、遗忘门和输出门。遗忘门决定哪些信息需要丢弃，输入门决定哪些新信息需要加入，输出门决定当前状态如何影响输出。通过将句子编码为固定长度的向量后，根据全连接层输出重要性分数选择出重要的句子。GRU是LSTM的变体，它将遗忘门和输入门合并为一个更新门，简化了结构，同时也能有效地捕捉长期依赖关系。

其中对于LSTM/GRU，记忆不断更新，状态也不断更新，遵循后面的公式：

候选记忆：C̃ₜ = tanh(W\_C·[hₜ₋₁,xₜ]+b\_C)

记忆更新：Cₜ = fₜ*Cₜ₋₁ + iₜ*C̃ₜ（f和i为遗忘门和输入门）

隐藏状态：hₜ = oₜ\*tanh(Cₜ)

文本预处理：分词、去除停用词、词干提取

词嵌入：将单词映射为低维向量(Word2Vec/GloVe)

句子编码：通过双向LSTM获取句子表示

文档编码：聚合句子表示形成文档表示

重要性评分：全连接层+sigmoid激活函数

句子选择：按分数排序选取top-k句子

CNN-based通过卷积操作提取句子中的局部特征，适用于快速评分和关键词识别任务。其优势在于并行计算能力强，适合实时场景。CNN模型通常由嵌入层、卷积层、池化层和全连接层组成。嵌入层将词映射为向量；卷积层使用多个卷积核对词向量序列进行卷积操作，提取局部特征；池化层对卷积后的特征进行降维，保留重要信息；全连接层根据池化后的特征输出句子的重要性分数。

BERT-base通过预训练获得上下文理解能力，将句子编码为包含语义信息的向量。BERT基于Transformer的编码器，采用多层自注意力机制捕捉文本中的依赖关系。其预训练任务包括遮蔽语言模型MLM和下一句预测NSP。在抽取式摘要任务中，将句子输入BERT模型，得到句子的向量表示，然后通过微调分类器预测句子是否应被选入摘要。

端到端模型直接从原始文本中学习复杂特征，输出每个句子的二元标签。采用层次编码器（句子编码器为每个句子生成表示，文档编码器，基于句子表示生成文档级表示，分类器为每个句子分配重要性分数）生成文档级表示，优势在于可以直接从原始文本中学习复杂特征，并通过注意力机制减少摘要中的重复内容，提升整体可读性。

### 2.2.2图与网络的问题

深度学习可以增强图模型中的相似性计算。例如，使用 BERT 生成的句子嵌入计算余弦相似性，从而构建更精确的图模型，提升摘要质量。此外，Swap-Net 等模型利用交替指针网络优化句子选择过程，进一步提升抽取效果。具体来说，Swap-Net 的交替指针网络通过在句子和词汇之间交替选择，使得摘要更具多样性和准确性。

**2.3现有模型代表**

Swap-Net：由Jadhav & Rajan（2018）提出，使用交替指针网络同时选择句子和词汇，提升摘要的多样性与准确性。其网络结构包括编码器、解码器和交替指针网络。编码器将输入文本编码为向量表示；解码器生成摘要；交替指针网络在解码过程中，在句子和词汇之间交替选择，根据上下文信息决定是选择新的句子还是从当前句子中选择词汇。

Neural Extractive with Syntactic Compression ：由 Xu & Durrett（2019）提出，结合句法压缩技术提升摘要质量，使句子更紧凑。该模型通过句法分析树提取句子的主干结构，删除冗余的修饰成分，同时利用神经网络对句子进行重要性评分，选择关键句子。

Ranking Sentences with Reinforcement Learning ：由 Narayan 等人（2018）提出，使用强化学习优化句子选择策略，以最大化整体摘要质量为目标。其强化学习框架包括状态空间、动作空间和奖励函数。状态空间表示已选择的句子和当前候选句子；动作空间表示选择或不选择当前候选句子；奖励函数根据摘要的质量指标（如 ROUGE 分数）评估选择策略的好坏。

这些模型在多个标准数据集（如 CNN/DM、Newsroom）上都取得了较好的实验结果，证明了深度学习在抽取式摘要中的强大潜力。

# 3生成式摘要方法

## 3.1 实现概况

生成式摘要的目标是模仿人类写作者的行为，通过对原文的理解，生成全新的、语义连贯的摘要句子。与抽取式摘要相比，生成式摘要更接近人类写作方式，能够有效避免原文句子的机械拼接，提升摘要的自然性和概括性。

生成式摘要的核心思想是利用序列到序列（Seq2Seq）框架，将输入文本编码为中间表示，再解码生成目标摘要。这一过程类似于机器翻译任务，但在摘要生成中更注重对信息的提炼和重构。

尽管生成式摘要具有更高的灵活性和表达能力，但生成内容可能偏离原文事实、语法错误较多、生成稳定性差等。

## 3.2 深度学习方法

### 3.2.1 RNN-based Seq2Seq模型

最早的生成式摘要模型基于 RNN 结构，如 LSTM 或 GRU。它们通过编码器将输入文本转化为上下文向量，再由解码器逐步生成摘要句子。编码器通常使用双向 RNN，以便能够捕捉输入序列的上下文信息；解码器使用单向 RNN，根据上一时刻的输出和上下文向量生成当前时刻的输出。

然而，RNN 结构存在梯度消失、长程依赖难以建模等问题，限制了其在长文本上的表现。

### 3.2.2 注意力机制

注意力机制的引入显著提升了生成式摘要的质量。它允许解码器在生成每个词时动态关注输入文本的不同部分，解决了 RNN 在长文本中注意力分散的问题。Rush 等人（2015）首次在单句摘要任务中应用注意力机制，其注意力模型通过计算解码器当前状态与编码器各个隐藏状态的相似度，得到注意力权重，然后对编码器的隐藏状态进行加权求和，得到上下文向量。随后发展出全局注意力和局部注意力等变体，进一步增强了模型的语义理解能力。

### 3.2.3 Transformer模型

Transformer 模型通过自注意力机制完全取代了 RNN 结构，成为当前生成式摘要的主流方法。其优势在于：并行计算能力强，训练效率高；自注意力机制能有效捕捉长距离依赖关系；可灵活应用于多种 NLP 任务。

Transformer 模型由编码器和解码器组成，编码器包括多层自注意力机制和前馈神经网络，用于提取输入文本的特征；解码器包括掩码自注意力机制、交叉注意力机制和前馈神经网络，用于生成摘要。

对于transformers，采用自注意力机制，多头注意力和位置编码。

输入嵌入=>词嵌入+位置编码

编码器堆叠=>多头自注意力层+前馈神经网络(FFN+)残差连接+层归一化

解码器堆叠=>掩码多头自注意力+编码器-解码器注意力+FFN层

输出层=>线性变换+softmax

堆叠如上所示。

## 3.3现有模型代表

BERT ：通过添加解码器模块可用于生成任务，擅长理解上下文。BERT 的编码器基于 Transformer，采用多层自注意力机制捕捉文本中的依赖关系。在生成式摘要任务中，通常将 BERT 的输出作为解码器的输入，生成摘要。

T5 ：将摘要视为文本到文本转换任务，支持多种任务统一处理。T5 模型基于 Transformer 架构，使用统一的文本到文本框架，将各种 NLP 任务（如文本分类、问答、摘要生成等）转化为文本生成问题。其预训练任务包括掩码语言模型和跨句子掩码语言模型。

BART ：结合双向编码与自回归解码，特别适合摘要任务。BART 的编码器使用 Transformer 的双向编码器，能够充分利用上下文信息；解码器使用自回归的 Transformer 解码器，生成摘要。

PEGASUS ：提出针对摘要的预训练目标（如提取间隙句子），显著提升生成质量。PEGASUS 的预训练目标是最大化句子之间的互信息，通过在原文中随机抽取出若干句子作为目标摘要，训练模型预测这些句子。

# 4总结与展望

文本摘要作为自然语言处理领域的重要任务之一，其核心目标是通过算法自动提取或生成一段简洁、准确且具有代表性的文本片段，以帮助用户快速获取关键信息。本文系统地介绍了文本摘要的基本分类（抽取式与生成式），并结合近年来深度学习的发展，重点分析了主流模型的技术原理、性能表现及其在实际应用中的优劣。

在抽取式摘要方面，传统方法如 TF-IDF、TextRank 和 LSA 等依赖手工特征和统计手段，虽然实现简单、计算效率高，但难以应对复杂的语义关系和上下文理解。随着深度学习的发展，基于 LSTM、CNN 和 BERT 等神经网络模型的抽取式方法显著提升了句子评分和选择的准确性，尤其是在大规模数据集上的表现更为稳定和可靠。

在生成式摘要方面，序列到序列（Seq2Seq）模型的引入为摘要生成带来了突破性进展。注意力机制进一步增强了模型对输入文本的关注能力，使得生成结果更贴合原文语义。而 Transformer 架构的广泛应用，尤其是 T5、BART、PEGASUS 等预训练语言模型的出现，极大地提升了生成式摘要的质量和多样性，使其在多个标准数据集上取得了领先的成绩。

此外，文本摘要在现实场景中也展现出广泛的应用价值，包括新闻媒体、医疗健康、法律金融等多个行业。然而，当前技术仍面临一些挑战，如生成内容的事实一致性问题、幻觉现象、数据偏见与算法偏见等，这些问题限制了其在敏感领域的广泛应用。

# 参考文献

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[2] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[3] Mihalcea R, Tarau P. Textrank: Bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. 2004: 404-411.