摘要

随着互联网技术的持续演进，文本信息数量呈现指数级增长，筛选和阅读这些浩瀚的信息资源变得耗时耗力。现今，迅速地从众多文本中筛选出重要信息并有效利用，成为了一个待解决的紧迫课题。在这种需求推动下，自动化文本摘要技术开始显现其重要性。这项技术能够助力用户迅速把握大量文本的核心要义，显著降低信息处理和阅读的复杂性，因而被视为自然语言处理领域的一项关键技术。尽管如此，该技术仍面临诸多挑战，如关键信息的提取不足，摘要与原文主旨大意重合度低，以及在处理文本时信息遗漏和句子不连贯等问题。鉴于此，对现行文本摘要技术进行深入研究并寻求改进，显得尤为重要。本文的主要研究内容如下：

（1）为了增强Informer模型在处理主题信息时的能力，本文基于经典的注意力机制，加入了主题信息。首先通过LDA（Latent Dirichlet Allocation）主题模型来识别文本中的主题词分布，并据此建立一个主题相似性矩阵。接着，利用这个主题相似矩阵优化注意力权重，让模型掌握文本中词汇的主题关联性，进而生成包含鲜明主题信息的摘要。此外，本文还采用了指针网络机制，实现了从原文中直接复制非词典词汇到摘要中，提高摘要生成的稳定性。经过对比实验数据表明，经过优化的Informer模型在ROUGE评价指标方面的表现超过了基线模型，验证了模型的有效性。

（2）为了克服自动摘要生成中Informer模型由于训练反馈不充分而遇到的困难，本文首先采用ERNIE获取动态的词向量表示，提高了词向量所蕴含的信息。然后设计了一个基于强化学习策略的自动摘要模型。利用强化学习中的自批判策略梯度算法提升模型生成摘要的质量。将生成摘要ROUGE评分和与参考摘要的语义相似性作为模型训练的奖励反馈到模型训练当中，最后在标准数据集上进行实验，实验结果证实了基于强化学习策略梯度的方法能有效提升生成式摘要的表现。

关键词：深度学习；自动摘要；Informer；强化学习；注意力机制

Abstract

With the continuous evolution of Internet technology, the amount of text information has grown exponentially, and screening and reading these vast information resources has become time-consuming and laborious. Nowadays, quickly filtering important information from numerous texts and effectively utilizing it has become an urgent issue to be solved. Driven by this demand, automated text summarization technology has begun to demonstrate its importance. This technology can help users quickly grasp the core essence of a large amount of text, significantly reducing the complexity of information processing and reading, and is therefore considered a key technology in the field of natural language processing. However, this technology still faces many challenges, such as insufficient extraction of key information, low overlap between the abstract and the main idea of the original text, and issues such as information omission and sentence incoherence when processing text. In view of this, it is particularly important to conduct in-depth research on current text summarization techniques and seek improvements. The main research content of this article is as follows:

1.To enhance the Informer model's ability to process topic information. This article is based on the classic attention mechanism and incorporates topic information. Firstly, the LDA (Latent Dirichlet Allocation) topic model is used to identify the distribution of topic words in the text, and a topic similarity matrix is established based on this. Next, use this topic similarity matrix to optimize attention weights, allowing the model to grasp the topic relevance of vocabulary in the text, and then generate summaries containing distinct topic information. In addition, this article also adopts a pointer network mechanism to directly copy non dictionary vocabulary from the original text into the abstract, improving the stability of abstract generation. After comparing experimental data, it was shown that the optimized Informer model outperformed the baseline model in terms of ROUGE evaluation indicators, verifying the effectiveness of the model.

2.In order to overcome the difficulties encountered by the Informer model in automatic summary generation due to insufficient training feedback, this paper first uses NERNIE to obtain dynamic word vector representations, improving the information contained in word vectors. Then, an automatic summarization model based on reinforcement learning strategy was designed. Utilizing the self critical strategy gradient algorithm in reinforcement learning to improve the quality of model generated summaries. The ROUGE score of the generated abstract and the semantic similarity with the reference abstract were fed back into the model training as rewards. Finally, experiments were conducted on a standard dataset, and the experimental results confirmed that the method based on reinforcement learning strategy gradient can effectively improve the performance of the generated abstract.

Keywords: Deep Learning; Automatic Summary; Informer; Reinforcement Learning; Attention mechanism

目录

第一章 引言

第一节 研究背景及意义

当前，互联网信息技术的发展速度令人瞩目，特别是文本信息的增长呈现出指数级态势[1]。网络新闻形态各异、层出不穷，其中不乏以吸引眼球为目的的“标题党”新闻。这种现象给读者在海量网络新闻中迅速定位并筛选有价值信息带来了巨大挑战。与人工摘要不同，自动摘要技术天生具有很大的优势。自动摘要技术根据既定规则快速生成摘要，这不仅减少了信息发布者的工作量，也使用户获取信息变得高效。自动摘要技术能够通过算法从文本中自动生成简洁连贯的摘要短文，精准地反映原文的核心内容[2]。这项技术不仅可以帮助研究工作者更好的进行研究，也对日常生活中需要处理很多文本信息的从业者提供了支持。它能够给予这些人们高效的获取新闻、研究报告等文本信息能力，进而可以增强他们的工作和研究效率。在新闻领域，自动文摘技术因其能够应对新闻信息过载问题而得到了广泛应用。读者不可能在他们很有限的时间内阅读大量可能有意义的文本信息。这种需求和适用性激发了大量研究人员的兴趣，他们致力于开发能够持续生成易于理解且具有意义的摘要模型[3]。在人工智能领域，特别是深度学习的推动下，机器自动从大量文本中提炼要点的技术，已经成为研究的热点。目前，这种自动摘要技术根据其生成摘要的方法，主要分为两类：一类是直接从原文中抽取关键句子的抽取式摘要；另一类是基于理解生成全新句子的生成式摘要。

目前，抽取式摘要通过从文本中挑选关键信息，然后依照特定规则进行组织，来有效传达原始内容。这种方式虽然能够确保保留文本的核心信息和语义的准确性，但它可能会带来不必要的信息重复，特别是在处理短文本时，准确性更低。另一方面，生成式摘要通过对原文语义的深入理解，从而能够抽象的概况原文内容。生成式摘要更符合人类阅读的习惯，可以大大减少抽取式摘要中的信息重复问题，实现更精准的信息提炼。然而，生成式摘要涉及更多的语言处理和理解，其算法复杂度更高，对于一些复杂或专业性领域的文章，其效果可能受到限制[4]。为了解决这些问题，学者们提出了使用循环神经网络、长短期记忆网络和门控循

环模块等技术来实现文本自动摘要的生成。但是，尽管早期的自动摘要技术依赖于顺序处理，限制了计算的并行性，使训练效率很低。针对此问题，Gehring等研究者通过引入卷积神经网络来解决这一问题。卷积神经网络虽然优化了并行计算，但它在获取序列特征方面存在不足。为此，在 2017 年，2017 年谷歌的研究团队开发了创新的Transformer模型，该模型摒弃了传统的CNN和RNN，转而采用前馈神经网络和注意力机制来实现序列到序列的转换。Transformer模型可以利用并行机制进行训练，还可以能够捕捉到文本的全局信息，尤其在机器翻译任务中展现出卓越的性能。目前，学术界正致力于探索Transformer和文本摘要的适配性。利用其注意力机制探索文本的深层语义，以期更高效地获得高质量的摘要。但是，Transformer在自动摘要模型的实践中仍面临主题漏洞和语义重复的挑战，这是由于其过度依赖端到端的注意力构造。另外，伴随输入文本数量的增加，所需的模型参数也随之变多，这在编码长文本时需要更多的计算资源的支撑，如果使用截断策略，可能会丢失关键信息，影响摘要的完整性。在 2021 年，Zhou等人提出了Informer模型，这是对Transformer模型的改进，用于长序列时间序列预测任务。他们对电力变压器温度、电力消耗负荷和天气进行了预测，结果显示，所提出的模型预测精度明显提升[5]。

近年来，基于序列到序列(seq2seq)[6]的深度学习方法取得了重大突破，在机器翻译等自然语言处理的研究领域引起了极大关注，因此相关技术和方法在摘要任务中的应用也受到了广泛关注。编码器-解码器架构是生成式文本摘要的主流方法。但是现在这些模型在理解文本的语义上还不够深入，导致在编码过程中无法完全捕捉到文本的关键信息，从而获得的摘要文本可能会遗漏重要内容或与原文事实不一致。为了克服这些困难，本研究将重要信息添加进序列到序列的摘要生成模型中，以提高摘要的准确性并减少错误信息的产生。

第二节　国内外研究现状

自 20 世纪 50 年代末期以来，自动文本摘要技术的发展旨在利用计算机技术缩减冗长文本，转化为简洁的句子或段落。这一过程不仅便于快速把握文本核心思想，而且力求在简化表达的同时，最大限度地保留关键信息和词汇，去除多余细节与重复内容。此技术的应用，不仅提高了阅读效率，也为后续相关

工作打下了坚实基础。目前，抽取式与生成式摘要技术已成为该领域的两大主流方向。本文将认真研究这两种技术的研究过程，并研究深度学习模型在自动文本摘要中的结合状况。

一、抽取式文本摘要模型

在自动文本摘要的初期阶段中，以统计和规则为基础的抽取式摘要技术是主流的研究方法，这些方法侧重于从文本中选择关键句子或短语来形成摘要，而不涉及对原文内容的深入理解或重写。Luhn[7]在 1958 年首次提出了一个依据词频等统计数据来确定句子重要性的方法，并根据这些数据来选取关键句形成摘要。受Luhn方法的启发，研究者们进一步将句子在文本中的位置和出现的关键词纳入评价标准，以此来判定句子的重要性。这种方法认为，文本中靠前的位置和含有高频关键词的句子通常承载着文章的主要信息，因此在摘要中应被优先考虑。例如，Baxendale[8]发现文章的开头和结尾往往包含总结性的描述，因此在摘要中应当优先考虑这些部分的句子。而Edmundson[9]则在词频和句子位置的基础上，进一步引入了线索词、文章标题和句子结构等特征来计算句子的权重。这些早期方法为自动文本摘要技术奠定了基础，并对后续的发展产生了深远影响。

自 1990 年代起，文本数据量的激增促进了机器学习技术在自动文本摘要领域的进行尝试研究。这些技术通过分析庞大的数据集，能够有效地提取关键信息，生成准确的摘要，从而在信息过载的时代中发挥着至关重要的作用。Kupiec等人[10]采用朴素贝叶斯模型来预测句子成为摘要的可能性。Aone[11]在 1999 年的研究中指出，高频词汇并不总是与主题密切相关，而某些低频词汇可能更具指示性，因此他引入了TFIDF（词频-逆文档频率）以提高摘要的质量。Conroy[12]基于隐马尔科夫模型，考虑了句子中的词汇数量和词汇间的关联性等特征进行摘要生成。2004年，Mihalcea和Tarau[13]基于Google的PageRank[14]算法，提出了TextRank算法，该算法通过构建文本的图表示，将句子视为节点，通过句子间的相似度计算边的权重，并通过迭代计算直至收敛，最终根据节点权重对句子进行排序，以产生文本摘要。TextRank算法它利用图论中的排名算法来确定文本中句子的重要性，从而选出最能代表原文意义的句子进行摘要。这种方法不依赖于复杂的语言模型，因此在处理摘要任务时既快速又有效。

深度学习的发展极大地推动了自然语言处理技术的革新，这其中也包括了抽取式文本摘要的进步。这些技术通过模仿人脑处理信息的方式，提高了机器对文本的理解能力，从而更有效地从大量文本中提取关键信息，生成准确的摘要。2015 年，Alexender等[15]引入了TDNN和注意力机制来从文本中提取语义特征。到了2017年，Nallapati等[16]运用循环神经网络建立了一个句子分类的摘要模型，该模型结合了词语和句子级别的向量表示，以分类句子并生成摘要。同年，Google的Transformer[17]模型在多个自然语言处理任务中取得了显著成就，特别是在机器翻译、文本生成和问答系统等方面表现出色。Transformer模型在处理长序列数据方面的并行处理能力和捕捉丰富语义信息的能力，使其优于传统的循环神经网络和卷积神经网络。此后，以BERT[18]为代表的预训练模型被应用于文本摘要任务。Liu等[19]通过对BERT模型进行微调，使其能够从长文本中提取关键句子，生成简洁的摘要。同时，Zhou等[20]提出了一种基于BERT的摘要模型，该模型通过句法分析技术切分文本，作为摘要的抽取单元，有效减少了冗余信息，提高了摘要的精确度和可读性。

二、生成式文本摘要模型

生成式摘要在早期阶段主要依靠句子压缩技术，这种方法通过删除句子中的非关键部分来缩短句子长度，从而生成摘要。这一技术是生成式摘要发展的基础，为后续更复杂的摘要技术奠定了基础。Jing等[21]指出，专业摘要编写者在创建摘要时，通常会对原文进行压缩处理。他们的研究分析了人工摘要与原文的对应关系，发现大多数摘要是通过压缩原句生成的。基于这一发现，Jing[22]进一步提出了一种通过删减句子成分来生成摘要的方法。这一方法首先保留句法树中的关键节点以确保语法正确性，然后根据上下文信息评估单词的相关性，并删除相关性低的部分。尽管句子压缩技术能够有效减少冗余，但其灵活性有限。Barzilay等[23]则通过提取多个相似句子的共同部分来形成摘要，这一方法在多文档摘要任务中表现出色。

在规则驱动的摘要生成方法显示出其局限性后，研究者们开始利用机器学习算法来更精确地辨识并去除句子中的次要成分，这样做可以更高效地进行句子压缩，从而改善摘要的质量。这种方法提高了生成式摘要的准确性和可读性。。