 DOI: https://doi.org/10.12349/iser.v6i4.5319

Performance evaluation of automatic summary generation algorithm driven by natural language processing in news editing scenarios

Hongqi Zhu

Dezhou Daily, Dezhou, Shandong, 253009, China

Abstract

With the explosive growth of information, the news industry is faced with the challenge of how to efficiently generate summaries in a limited time in order to quickly deliver key information. Traditional manual summary generation is not only time expensive, but also difficult to deal with large amounts of data. The rapid development of natural language processing (NLP) technology provides a new solution for automatic summary generation, especially in the news editing scene, how to apply NLP driven algorithm to generate high quality automatic summary has become the focus of the industry.This paper discusses the application of automatic abstract generation algorithm driven by natural language processing and its performance evaluation in news editing. By comparing the actual performance of several mainstream algorithms (such as rule-based algorithms, deep learning models, etc.) in news editing scenarios, this paper analyzes the advantages and disadvantages of each algorithm, and puts forward improvement strategies to improve the quality of automated summary generation. The research shows that the deep learning-driven algorithm has higher accuracy and effect than the traditional method when dealing with the task of summary generation of news text.

Keywords

natural language processing, automated summary generation, news editing, deep learning, performance evaluation

自然语言处理驱动的自动化摘要生成算法在新闻编辑场景中的性能评估

朱红旗

德州日报社，中国 · 山东 德州 253009

摘　要

随着信息量的爆炸性增长，新闻行业面临着如何在有限时间内高效地生成摘要以便快速传递关键信息的挑战。传统的人工摘要生成不仅时间成本高，而且难以处理大量数据。自然语言处理（NLP）技术的飞速发展为自动化摘要生成提供了新的解决方案，特别是在新闻编辑场景中，如何应用NLP驱动的算法生成高质量的自动化摘要已成为行业关注的焦点。本文探讨了自然语言处理驱动的自动化摘要生成算法的应用及其在新闻编辑中的性能评估。通过对几种主流算法（如基于规则的算法、深度学习模型等）在新闻编辑场景中的实际表现进行比较，本文分析了各算法的优缺点，并提出了提升自动化摘要生成质量的改进策略。研究表明，深度学习驱动的算法，在处理新闻文本的摘要生成任务时，相比传统方法具有更高的准确性和效果。

关键词

自然语言处理，自动化摘要生成，新闻编辑，深度学习，性能评估

1 引言

在信息爆炸的时代，新闻行业每天面临着海量的文本数据，需要在极短的时间内将信息压缩成简洁、有效的摘要，以便快速传递关键信息。人工摘要生成虽然能够保证内容的准确性和可读性，但在处理大量新闻数据时存在明显的时间和人力成本问题。因此，如何高效且准确地生成新闻摘要成为了新闻行业亟待解决的难题。

自然语言处理（NLP）技术的迅猛发展为自动化摘要生成提供了新的解决方案。近年来，基于深度学习的 NLP 方法已经在多个领域取得了显著成果，包括文本分类、情感分析、机器翻译等，其中自动化摘要生成作为其中一个重要任务，得到了广泛的关注。自动化摘要生成技术通过分析和理解输入的新闻文本，自动提取关键信息并生成简洁的摘要，具有较大的应用潜力。尽管如此，当前的自动化摘要生成技术在实际应用中仍面临许多挑战，如摘要生成的准确

【作者简介】朱红旗（1974–），男，中国山东德州人，本科，助理工程师，从事计算机科学与技术研究。

性、内容的连贯性、语法和语义的处理等问题。

本文旨在评估基于自然语言处理的自动化摘要生成算法在新闻编辑场景中的性能，通过对不同算法进行性能对比，分析其优缺点，探索提升摘要生成质量的策略，并为新闻编辑领域的自动化摘要生成提供实践指导。

2 自然语言处理在自动化摘要生成中的应用背景

2.1 自动化摘要生成任务的定义

自动化摘要生成旨在将一篇长文本压缩成简短、准确且具有代表性的内容摘要。根据生成方式的不同，摘要生成可分为提取式摘要和抽象式摘要两类。提取式摘要是通过从原文中选择关键句子或段落来构建摘要，而抽象式摘要则是通过理解原文内容，生成具有概括性、创新性的内容。这两种方法各有优势与局限，提取式摘要较为简便且效果较好，但可能存在内容重复和缺乏创新的问题；而抽象式摘要能够更好地概括原文的核心内容，但由于需要更强的文本生成能力，因此其实现难度较大。

在新闻编辑场景中，通常要求生成的摘要准确、简洁且具有代表性，能够在短时间内传递信息的精髓，便于读者快速了解新闻的核心内容。随着深度学习技术的发展，基于神经网络的抽象式摘要生成方法逐渐成为主流，其能够更灵活地处理复杂的语言结构和语义关系，生成更自然流畅的摘要。

2.2 自然语言处理技术的发展与应用

自然语言处理作为人工智能领域的重要分支，致力于研究计算机与人类语言之间的交互，涵盖了语言理解、语言生成、文本分析等多个方面。近年来，深度学习尤其是神经网络的应用，极大推动了 NLP 技术的发展。在文本生成、文本摘要等任务中，深度学习模型的表现优于传统的规则驱动模型，成为了学术界和工业界的研究热点。

在自动化摘要生成方面，NLP 技术的应用已经从早期的基于规则的模型发展到如今的基于神经网络的深度学习模型。例如，长短期记忆网络（LSTM）、卷积神经网络（CNN）以及近年来表现优异的变换器（Transformer）模型，均已被广泛应用于文本摘要生成任务中。这些技术通过对大规模语料的训练，使得模型能够学习到语言的深层语法、语义信息，提升了自动化摘要的质量。

2.3 新闻编辑场景中的自动化摘要需求

新闻行业对自动化摘要生成有着极高的需求，尤其是在新闻传播的实时性和广泛性要求下，如何快速、准确地生成新闻摘要成为了关键问题。传统的人工摘要生成虽然能够保证摘要的质量，但面对海量的新闻文本，人工工作量巨大，无法满足快速传播的需求。通过自动化技术生成摘要，可以大幅提高工作效率，保证新闻内容的及时性。

然而，新闻摘要生成的难度较大，主要体现在以下几个方面。首先，新闻文本内容通常涉及多种信息类型，包括事实、数据、评论等，如何准确提取新闻的核心信息并生成简洁明了的摘要，要求模型具备较强的理解能力。其次，新闻摘要需要具备良好的连贯性和语法结构，使得读者能够轻松理解摘要的内容。此外，新闻领域的文本内容具有较强的时效性和领域性，如何在自动化摘要生成中考虑时效性和背景信息，也是一个亟待解决的问题。

3 自动化摘要生成算法的分类与特点

3.1 提取式摘要生成算法

提取式摘要生成算法是自动化摘要生成任务中最早应用的方法之一，其核心思想是从原文中提取出最具代表性、最关键的句子或段落，直接拼接成简短的摘要。这类算法的优势在于简便、高效，能够快速从大量文本中生成简明扼要的摘要。常见的提取式算法包括基于频率的词汇选择算法（如 TF-IDF）、图模型算法（如 TextRank）、以及支持向量机（SVM）等。

TF-IDF（Term Frequency-Inverse Document Frequency）是一种基于词频与逆文档频率的权重计算方法，能够有效识别文本中最具代表性的关键词或句子，进而选取最具信息量的部分作为摘要内容。TextRank 算法则通过构建句子之间的图模型，将句子视作节点，句子之间的相似度作为边，通过图的结构计算出各个句子的“重要性”，从而选取权重较大的句子作为摘要。支持向量机（SVM）则是通过训练样本中的正负样本，利用分类算法识别最具代表性的句子，来生成摘要。

3.2 抽象式摘要生成算法

与提取式摘要不同，抽象式摘要生成算法更侧重于理解和概括原文的核心内容，生成具有创新性和概括性的摘要。这类算法的目标不仅仅是从原文中提取信息，而是通过深度理解原文，生成新的、简洁的内容。基于深度学习的抽象式摘要生成算法，特别是使用神经网络模型，如长短期记忆网络（LSTM）、循环神经网络（RNN）以及变换器（Transformer）等，已经在自动化摘要生成领域取得了显著的进展。

深度学习模型可以通过大规模的训练数据学习语言的深层次语法结构和语义关系，从而生成更具概括性和创新性的摘要。例如，LSTM 和 RNN 网络通过逐步处理输入文本中的每个单词，并且能够捕捉到文本中的长程依赖关系，使得生成的摘要不仅能准确反映文章的主要信息，还能灵活调整表达方式，形成创新性的总结。近年来，Transformer 模型凭借其并行处理能力和自注意力机制，取得了更为出色的表现，尤其在处理复杂语法结构和长文本时，能够更好地捕捉文本中的深层次语义关系，生成更加准确和流畅的摘要。

然而，抽象式摘要生成模型也面临一些挑战。首先，这类模型通常需要较大的计算资源和较长的训练时间，这对

于资源有限的应用场景可能是一个障碍。其次，尽管神经网络模型能够学习到更深层的语义关系，但在处理复杂的语法结构和长文本时，仍然存在一定的困难。模型可能无法完美地理解长篇文章中的细节，或者可能出现生成的摘要偏离原文的情况，影响摘要的准确性和可信度。

3.3 混合型摘要生成算法

近年来，结合提取式和抽象式摘要生成的混合型算法逐渐成为研究的热点。混合型摘要生成算法尝试融合提取和生成两种方法的优点，通过首先从原文中提取出重要的信息，然后通过生成模型对提取的内容进行优化、改写，最终生成一个更加简洁、流畅且具有创新性的摘要。

这种算法的基本流程通常是先使用提取式方法进行初步的摘要生成，通过识别出文本中最关键的句子或段落，构建一个粗略的摘要。接着，利用生成模型对这个粗略摘要进行优化和再生成，使其更具连贯性、简洁性和创新性。通过这种方式，混合型算法能够在提取式算法的高效性和抽象式算法的创新性之间找到平衡。混合型算法通常能够生成质量更高的摘要，避免了纯粹提取式方法中可能出现的重复性和信息不连贯的问题，同时又弥补了纯粹抽象式方法中计算资源需求过大和长时间训练的问题。

然而，混合型算法也面临着一定的挑战。首先，这种方法的计算复杂度较高，涉及到提取和生成两个步骤，需要消耗较多的计算资源。其次，尽管混合型算法在提高摘要质量方面表现出了较好的性能，但其仍然存在一定的优化空间，尤其是在自动化摘要的准确性、连贯性和创新性方面。随着技术的不断进步，混合型算法有望进一步提高其效率和准确性，为自动摘要生成提供更优的解决方案。

4 自动化摘要生成在新闻编辑中的性能评估

4.1 评估指标

自动化摘要生成算法的性能评估通常依赖于多个指标，包括抽取信息的准确性、生成摘要的简洁性、连贯性以及语法的流畅性等。常用的评估指标包括 ROUGE（Recall-Oriented Understudy for Gisting Evaluation）、BLEU（Bilingual Evaluation Understudy）等，这些指标能够有效衡量生成摘要与人工摘要的相似度。

ROUGE 是目前最为常用的自动化摘要评估指标，主要通过计算生成摘要与人工摘要之间的重叠度，如 n-gram 重叠、最长公共子序列等。ROUGE 分为多个版本，包括 ROUGE-N、ROUGE-L 等，分别衡量不同类型的重叠信息。通过 ROUGE 指标，研究者可以客观评估自动化摘要生成的质量和效果。

4.2 算法性能对比

在实际应用中，不同算法在新闻摘要生成中的表现差异较大。基于深度学习的抽象式算法通常能够生成更加简洁且有创新性的摘要，但由于其计算复杂度较高，训练和推理时间较长，因此在实时性要求较高的新闻编辑场景中，可能存在一定的性能瓶颈。而提取式摘要生成算法虽然计算效率较高，但生成的摘要缺乏创新性和深度，往往无法完全体现新闻的核心价值。混合型算法在性能上表现出较好的平衡，能够在保证摘要质量的同时提高计算效率，适应新闻编辑场景的需求。

4.3 应用案例分析

通过具体的案例分析，本文评估了不同算法在新闻编辑中的实际应用效果。以一组新闻数据集为例，分别使用提取式、抽象式和混合型算法生成摘要，并通过 ROUGE 评估指标对比其生成效果。结果表明，抽象式算法在摘要的创新性和简洁性上表现优越，但在处理长文本时仍存在一定的不足。提取式算法在精度和效率上表现较好，适用于时间紧迫的新闻编辑场景。混合型算法则在各方面表现均衡，尤其在处理复杂语境和多样化新闻内容时，能够生成较为理想的摘要。

5 结语

自然语言处理驱动的自动化摘要生成算法在新闻编辑场景中的应用，已经取得了显著的进展。通过各种算法的比较与评估，本文总结了不同算法的优缺点，并探讨了其在实际应用中的效果和改进空间。未来，随着 NLP 技术的进一步发展，自动化摘要生成算法将变得更加智能和高效，能够更好地适应新闻行业对摘要生成质量和时效性的要求。通过结合领域知识、优化算法模型，未来的摘要生成技术将能够为新闻编辑提供更加精确、创新和高效的支持，推动新闻行业的智能化发展。

参考文献

[1] 孙财茂.材料科学文本挖掘软件的开发及其应用于新型低热导率材料预测[D].吉林大学,2024.

[2] 边慧聪.基于深度强化学习的游戏阵容角色推荐方法研究[D].齐鲁工业大学,2024.

[3] 毛寅辉.基于图卷积网络的短文本分类方法研究[D].大连交通大学,2024.

[4] 袁琳.基于图模型表达的报道性新闻自动摘要研究[D].中国农业科学院,2023.

[5] 黄蕙.融合句子情感的新闻文档自动摘要提取[D].北京印刷学院,2023.