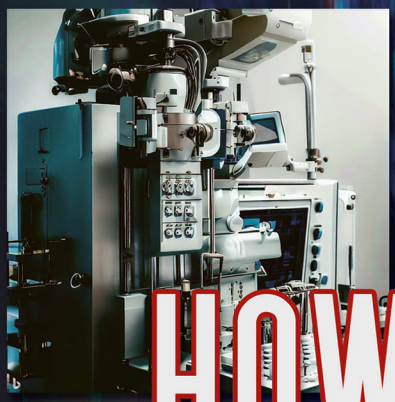
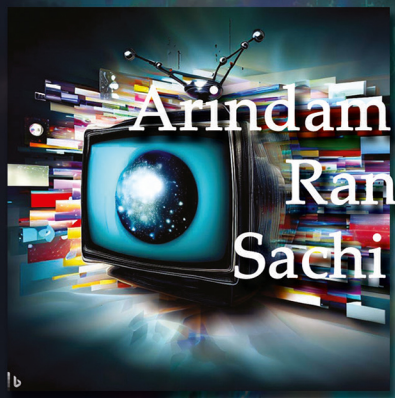
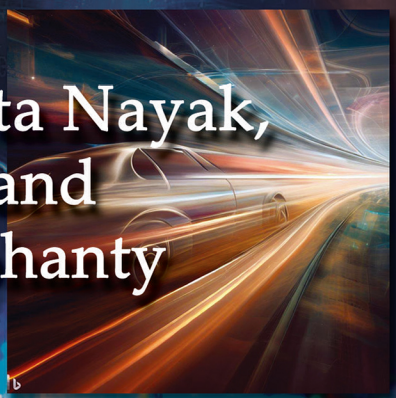


HOW MACHINE LEARNING IS INNOVATING TODAY'S WORLD

A Concise Technical Guide





EDITED BY
Arindam Dey, Sukanta Nayak,
Ranjan Kumar, and
Sachi Nandan Mohanty

Scrivener Publishing

WILEY

How Machine Learning is Innovating Today's World

Scrivener Publishing
100 Cummings Center, Suite 541J
Beverly, MA 01915-6106

Publishers at Scrivener
Martin Scrivener (martin@scrivenerpublishing.com)
Phillip Carmical (pcarmical@scrivenerpublishing.com)

How Machine Learning is Innovating Today's World

A Concise Technical Guide

Edited by

Arindam Dey

School of Computer Science, VIT-AP University, Andhra Pradesh, India

Sukanta Nayak

Dept. of Mathematics, VIT-AP University, Andhra Pradesh, India

Ranjan Kumar

Dept. of Mathematics, VIT-AP University, India

and

Sachi Nandan Mohanty

Dept. of Computer Science, VIT-AP University, Andhra Pradesh, India

WILEY

This edition first published 2024 by John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA and Scrivener Publishing LLC, 100 Cummings Center, Suite 541J, Beverly, MA 01915, USA

For more information about Scrivener publications please visit www.scrivenerpublishing.com.

Wiley Global Headquarters
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Library of Congress Cataloging-in-Publication Data

ISBN 978-1-394-21411-2

Cover image: Pixabay.Com
Cover design by Russell Richardson

Set in size of 11pt and Minion Pro by Manila Typesetting Company, Makati, Philippines

Printed in the USA

10 9 8 7 6 5 4 3 2 1

Contents

Preface

This timely book presents a diverse collection of chapters that delve into the remarkable ways that machine learning (ML) is transforming various fields and industries. It provides a comprehensive understanding of the latest advancements and practical applications of ML techniques.

Machine learning, a branch of artificial intelligence, has gained tremendous momentum in recent years, revolutionizing the way we analyze data, make predictions, and solve complex problems. As researchers and practitioners in the field, the editors of this book recognize the importance of disseminating knowledge and fostering collaboration to further advance this dynamic discipline.

The chapters herein cover a wide range of topics, each contributing a unique perspective to the broader landscape of machine learning. First is a comprehensive analysis of various tokenization techniques and the sequence-to-sequence model in natural language processing. Next, Chapter 2 explores the evaluation of English language readability using ML models, followed by a detailed study of text analysis for information retrieval through natural language processing in the subsequent chapter.

Chapter 4 investigates machine learning's role in maximizing cotton yield with a focus on fertilizer selection, and Chapter 5 delves into the application of reinforcement learning approaches to supply chain management. The following chapter examines the performance analysis of converting algorithms to source code using natural language processing in Java, and Chapter 7 presents an alternate approach to solving differential equations utilizing artificial neural networks with optimization techniques.

The exploration of the subject continues with a comparative study of different techniques of text-to-SQL query conversion in Chapter 8, and the next chapter examines ML approaches to catalysis. After that, Chapter 10 presents the systematic study of text generation and classification using tokenization in natural language processing, followed by the classification of livestock diseases using ML algorithms in Chapter 11.

Chapter 12 provides a closer look at the application of ML in image enhancement techniques, and the following chapter demonstrates the prediction of book genres using natural language processing. Additionally, Chapter 14 delves into efficient leader selection for inter-cluster flying ad-hoc networks, and the subsequent chapter provides a comprehensive survey of applications powered by GPT-3 and DALL-E.

Recommender systems' domain of application is discussed in Chapter 16, and the next chapter reviews mood detection, emoji generation, and classification using tokenization and CNN. Chapter 18 delves into a new variation of the exam scheduling problem using graph coloring, and Chapter 19 examines the intersection of software engineering and machine learning applications.

Moreover, Chapter 20 explores ML strategies for indeterminate information systems in complex bipolar neutrosophic environments, and the rise of AI-generated news videos is scrutinized in Chapter 21. The next section highlights ML applications in battery management systems, while the healthcare industry is covered in Chapter 23. The book's final chapter presents how to enhance resource management in precision farming through AI-based irrigation optimization.

This book will serve as a valuable resource for researchers, scholars, and enthusiasts seeking to understand the cutting-edge advancements in ML. The editors extend our gratitude to all the authors who have contributed their expertise, insights, and knowledge to make this book possible. Their commitment to advancing the frontiers of machine learning has greatly enriched the content and depth of this publication.

We also offer our sincere appreciation to Wiley and Scrivener Publishing for their support and guidance throughout the editorial process. Their commitment to publishing high-quality scientific literature has been instrumental in bringing this book to fruition.

We hope that readers find this book insightful, engaging, and thought-provoking. May it inspire you to explore new horizons in machine learning and contribute to the ongoing advancements that are reshaping our world.

The Editors
March 2024

Part 1

NATURAL LANGUAGE PROCESSING (NLP) APPLICATIONS

1

A Comprehensive Analysis of Various Tokenization Techniques and Sequence-to-Sequence Model in Natural Language Processing

Kuldeep Vayadande[1]*, Ashutosh M. Kulkarni[1], Gitanjali Bhimrao Yadav[1], R. Kumar[2] and Aparna R. Sawant[1]

[1]Vishwakarma Institute of Technology, Pune, India
[2]VIT-AP University, Inavolu, Beside AP Secretariat, Amaravati AP, India

Abstract
This research paper provides an in-depth examination of various tokenization techniques and Sequence-to-Sequence (Seq2Seq) models, with an emphasis on the LSTM, Transformer, and Attention-based LSTM models. The process of tokenization, which breaks down text into smaller units, plays a vital role in natural language processing (NLP). This study evaluates different tokenization methods, including word-based, character-based, and sub-word-based methods. It also explores the latest advancements in Seq2Seq models, such as the LSTM, Transformer, and Attention-based LSTM models, which have been successful in tasks like machine translation, text summarization, and dialog systems. The paper compares the performance of different tokenization techniques and Seq2Seq models on benchmark datasets. Additionally, it highlights the strengths and limitations of these models, which helps in understanding their suitability for various NLP applications. The aim of this study is to comprehensively understand the current advancements in tokenization and sequence-to-sequence modeling for NLP, particularly with regard to LSTM, Transformer, and Attention-based LSTM models.

***Keywords*:** RNN, CRNN, LSTM, bidirectional-LSTM, text augmentation, tokenization, attention-based LSTM

1.1 Introduction

Tokenization is a fundamental step in natural language processing (NLP) that entails breaking down text into smaller units, such as words or characters. This process is critical for many NLP tasks, including text classification, machine translation, and text summarization. Different levels of granularity, such as word-level, character-level, and sub-word-level, can be used for tokenization.

In recent years, various tokenization techniques have been proposed, each with their unique advantages and disadvantages. The Multi-head Self-attention Mechanism in [1] is

**Corresponding author*: kuldeep.vayadande1@vit.edu

Arindam Dey, Sukanta Nayak, Ranjan Kumar and Sachi Nandan Mohanty (eds.) How Machine Learning is Innovating Today's World: A Concise Technical Guide, (3–12)

a type of attention mechanism that allows the model to concentrate on multiple parts of the input text simultaneously. Tokenization is the most straightforward approach, and it is widely used in many NLP tasks. However, it may not be as effective for languages with complex morphological structures, such as agglutinative languages. On the other hand, character-level tokenization can handle such languages better, but it may also introduce more noise into the data. Sub-word-level tokenization, such as byte-pair encoding (BPE) and unigram language modeling (ULM), has been proposed as a compromise between word-level and character-level tokenization.

The goal of this study is to provide a thorough understanding of the tokenization methods that have been introduced recently. The research will evaluate different tokenization methods, including word-based, character-based, and sub-word-based methods, and compare their performance on a set of benchmark datasets. Furthermore, the research will delve into the details of these techniques, their working principle, and their performance on various NLP tasks. Additionally, the research will also analyze the advantages and limitations of these techniques, which will assist in understanding their suitability for different types of NLP applications. The objective of this research is to gain a complete insight into the latest developments in tokenization for NLP. This research will be a valuable resource for researchers and practitioners in the field of NLP, supplying students with a thorough comprehension of the most advanced tokenization algorithms available at the time. The popularity of Sequence-to-Sequence (Seq2Seq) models in NLP has grown in recent times because of their capability to process input and output sequences of varying lengths. Seq2Seq models, also known as encoder–decoder models, have produced noteworthy outcomes in a number of NLP applications, including dialogue systems, machine translation, and text summarization. Different Seq2Seq models have been put forth through time, and each has merits and faults of its own. One such model is the Long Short-Term Memory (LSTM), a popular Seq2Seq model that has shown promise in a variety of NLP applications. Its limitation is that it is computationally expensive. An alternative to the LSTM model that has proven to be more effective is the Transformer model, which is built on the attention mechanism. Attention-based LSTM models are also proposed, which combine the advantages of LSTM and attention mechanisms. With an emphasis on the LSTM, Transformer, and Attention-based LSTM models, this survey seeks to provide a thorough understanding of the many Seq2Seq models that have been suggested in recent years. We will cover the details of these models, their working principle, and their performance on various NLP tasks. Furthermore, we also cover the advantages, limitations, and performance comparison of these models, which helps in understanding their suitability for different types of NLP applications.

Also, some non-tokenization technique as mentioned in [4] is focused on pre-training an efficient encoder that can operate without tokenization.

1.2 Literature Survey

The Multi-head Self-attention Mechanism [1] is a type of attention mechanism that enables the model to concentrate on multiple sections of the input text simultaneously. This is achieved by utilizing multiple "heads" to attend to different segments of the input. This can aid the model's understanding of the context and relationships between different parts of the input text, resulting in more accurate and coherent summaries.

The pointer network is a type of encoder–decoder [2] model that uses an attention mechanism to point to the part of the input text that should be included in the summary. It helps the model to generate the summary by copying relevant words from the input text, rather than generating them from scratch.

In summary, the research paper [3] focuses on improving the language generation performance by calibrating the likelihood of the generated sequences, while the research paper [4] focuses on Long Document Summarization and uses a combination of top-down and bottom-up inference to extract high-level concepts and specific details from the input text. Both papers provide different approaches to improve the performance of NLP tasks.

CANINE [4] is focused on pre-training an efficient encoder that can operate without tokenization, making the training process faster and more scalable. It uses a simple linear-layer-based architecture and employs a binary masking strategy to hide specific words during training, in order to predict them during inference.

FNet [5], on the other hand, introduces a new method of mixing tokens with Fourier transforms to capture long-range dependencies. This method can produce highly expressive representations and has the advantage of being computationally efficient. The paper shows that the approach outperforms traditional pre-training methods on various NLP tasks.

Charformer [6] focuses on the tokenization stage of pre-processing and proposes a novel method of sub-word tokenization that utilizes gradient information to identify the best sub-word splits. The authors show that their method is fast and results in improved performance on several NLP tasks.

The paper [7] proposes a new pre-training method for language models that leverages retrieval-based techniques. The authors show that this approach can effectively pre-train models on large-scale text corpora, leading to improved performance on a variety of NLP benchmarks.

The paper [8] focuses on enhancing the training efficiency of large-scale transformers, which are frequently employed in NLP applications like language modeling, text classification, and machine translation. The authors propose a new training method, "Random-LTD," which involves randomly dropping tokens and layers during the training process to speed up convergence and reduce memory requirements. The authors show that this method can effectively train large-scale transformers with improved efficiency.

Detecting Label Errors in Token Classification Data [9] focuses on a different challenge in NLP, which is detecting label errors in token classification data. The authors propose a method for detecting label errors in token classification datasets, which can negatively impact the performance of NLP models. The authors present experiments demonstrating the effectiveness of their method in detecting label errors in real-world datasets.

1.3 Sequence-to-Sequence Models

1.3.1 Convolutional Seq2Seq Models

Convolutional Seq2Seq (ConvSeq2Seq) models [10] are a variant of Seq2Seq models that incorporate (CNNs) into model architecture. ConvSeq2Seq models are particularly useful for processing sequences of data with a grid-like structure, such as image sequences or spectrograms.

Compared to traditional Seq2Seq models that use recurrent neural networks (RNNs), ConvSeq2Seq models have the advantage of being able to process sequences in parallel, which can lead to faster training and inference times. However, they may be less effective at capturing long-range dependencies in the data compared to RNN-based models.

1.3.2 Pointer Generator Model

Pointer Generator [11] models are a type of Seq2Seq model used for text summarization and other tasks where the output sequence is a subset of input sequence. The Seq2Seq model comprises of a decoder that generates the target sequence and an encoder that analyzes the input sequence and generates a fixed-length vector representation.

Pointer Generator models have been utilized for a wide range of tasks such as text summarization, headline generating, and question answering. Compared to traditional Seq2Seq models, attention-based models offer a more adaptable approach to handle the issue of Out-of-Vocabulary (OOV) words and can produce output sequences that are more reflective of the input data.

1.3.3 Attention-Based Model

Models that rely on attention [12, 13] are a category of Seq2Seq models that have recently become widespread for various NLP jobs such as machine translation, text summarization, and answering questions. The attention mechanism, which enables it to concentrate on various segments of the input sequence while producing the output sequence, is a key aspect of these models. The attention mechanism computes a score for each element in the input sequence that reflects its importance for the current task, and the decoder uses these scores to weight the contribution of each element when generating the output.

However, attention-based models can be computationally expensive and difficult to train, especially for large sequences, due to the need to compute attention scores for every element in the input sequence. Additionally, the attention mechanism can sometimes be unstable, leading to poor performance in some cases.

1.4 Comparison Table

Table 1.1 shows different Approach-based comparison [1–9]. Table 1.2 shows advantages and disadvantages of Tokenization and Seq2Seq Model [1–9]. Table 1.3 shows model performance [1–9].

Table 1.1 Approach-based comparison.

Paper	Approach used	Performance
[1]	Multi-head self-attention mechanism	Improved understanding of context and relationships in input text
[2]	Pointer network	Use of attention mechanism to point to relevant input text for summary generation

(*Continued*)

Table 1.1 Approach-based comparison. (*Continued*)

Paper	Approach used	Performance
[3]	Improved language generation performance by calibrating likelihood of generated sequences	[Performance not specified]
[4]	Long Document Summarization using top-down and bottom-up inference	[Performance not specified]
[5]	Pre-training an efficient encoder with binary masking strategy	Improved performance on NLP tasks
[6]	Novel sub-word tokenization method using gradient information	Improved performance on several NLP tasks
[7]	Retrieval-based pre-training method for language models	Improved performance on a variety of NLP benchmarks
[8]	Improving training efficiency of large-scale transformers	Improved efficiency in training large-scale transformers
[9]	Detecting label errors in token classification data	Effective method for detecting label errors in real-world datasets

Table 1.2 Advantages and disadvantages based on tokenization and Seq2Seq approach used.

Paper	Approach used	Performance
[1]	Multi-head self-attention mechanism	Improved understanding of context and relationships in input text
[2]	Pointer network	Use of attention mechanism to point to relevant input text for summary generation

(*Continued*)

Table 1.2 Advantages and disadvantages based on tokenization and Seq2Seq approach used. (*Continued*)

Paper	Approach used	Performance
[3]	Improved language generation performance by calibrating likelihood of generated sequences	[Performance not specified]
[4]	Long Document Summarization using top-down and bottom-up inference	[Performance not specified]
[5]	Pre-training an efficient encoder with binary masking strategy	Improved performance on NLP tasks
[6]	Novel sub-word tokenization method using gradient information	Improved performance on several NLP tasks
[7]	Retrieval-based pre-training method for language models	Improved performance on a variety of NLP benchmarks
[8]	Improving training efficiency of large-scale transformers	Improved efficiency in training large-scale transformers
[9]	Detecting label errors in token classification data	Effective method for detecting label errors in real-world datasets

Table 1.3 Model performance comparison w.r.t accuracy and RMSE.

Model	Accuracy
BERT model	94
Multi-head self-attention pointer model	96
Bidirectional LSTM model	97.5
Encoder–decoder model	97
Encoder–decoder model with attention layer	98

1.5 Comparison Graphs

This section contains comparison graphs of various techniques. Figure 1.1 shows accuracy comparison graph pf five different sequence-to-sequence models [1–9] and Figure 1.2 shows RMSE comparison graph of five different sequence-to-sequence models [1–9].

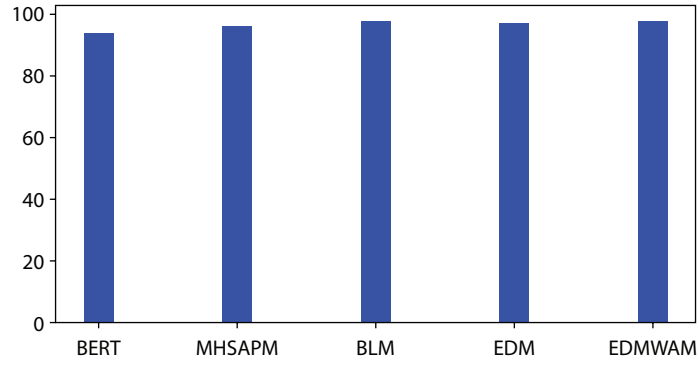


Figure 1.1 Accuracy comparison graph of five different sequence-to-sequence models [1–9].

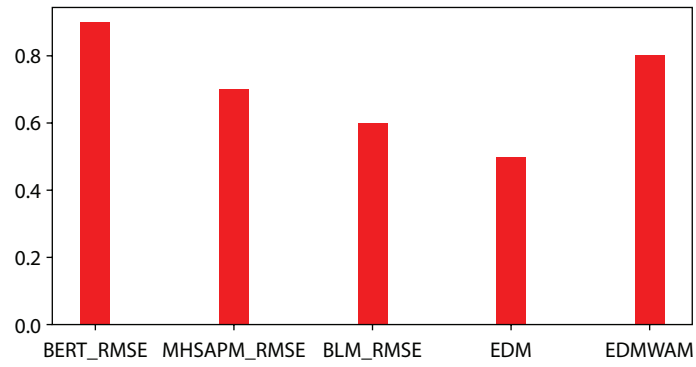


Figure 1.2 RMSE comparison graph of five different sequence-to-sequence models [1–9].

1.6 Research Gap Identified

The current available models lack multi-sentence summarization tasks, and the existing model does not work for large datasets and is limited to character-level data. Handling long sequences: Seq2Seq models tend to struggle with long sequences, especially when the sequence is much longer than the training data. This leads to a drop in performance, and a need to address this challenge. Despite the impressive results achieved by Seq2Seq models, it is often difficult to understand how the model makes its predictions. Improving the interpretability of these models is an important research area. Seq2Seq models are trained on large amounts of data, and any biases in the data are likely to be reflected in the model's predictions. Addressing data bias is an important research area in Seq2Seq models.

Overall, the research gap in Seq2Seq models is to continue to improve the models' accuracy, interpretability, scalability, robustness, and ability to handle diverse data types, while addressing data bias and other challenges.

1.7 Conclusion

Tokenization is an important step in pre-processing text data for NLP tasks. It involves dividing a text into smaller units, called tokens, which can be words, characters, sub-words, or even bytes. There are different tokenization techniques that have been proposed in the literature, each with its own advantages and limitations.

Character-level tokenization, for example, can handle rare and out-of-vocabulary words effectively, but is computationally expensive and can result in a large number of tokens. Word-level tokenization is computationally efficient, but may not handle rare or out-of-vocabulary words effectively. Sub-word tokenization combines the advantages of both character-level and word-level tokenization, and has become a popular choice in many NLP tasks.

In addition, recent research has proposed novel tokenization techniques such as gradient-based sub-word tokenization and token dropping to address the challenges of processing large amounts of text efficiently. These techniques show promise in improving the efficiency of NLP models while maintaining their accuracy.

In conclusion, tokenization is a critical step in NLP pre-processing, and different tokenization techniques have been proposed to address the challenges of processing large amounts of text effectively and efficiently. The choice of tokenization technique depends on the specific NLP task, the size of the text corpus, and the computational resources available.

References

1. Qiu, D. and Yang, B., Text summarization based on multi-head self-attention mechanism and pointer network. *Complex Intell. Syst.*, 8, 555–567, 2022, https://doi.org/10.1007/s40747-021-00527-2.
2. Li, Z., Peng, Z., Tang, S., Zhang, C., Ma, H., Text summarization method based on double attention pointer network. *IEEE Access*, 4, 1–1, 2016.
3. Zhao, Y., Khalman, M., Joshi, R., Narayan, S., Saleh, M., Liu, P.J., Calibrating sequence likelihood improves conditional language generation. *ArXiv*, 2022, https://doi.org/10.48550/arXiv.2210.00045.
4. Clark, J.H., Garrette, D., Turc, I., Wieting, J., CANINE: Pre-training an efficient tokenization-free encoder for language representation. 5, 11 March, 2021.
6. Lee-Thorp, J., Ainslie, J., Eckstein, I., Ontanon, S., FNet: Mixing tokens with fourier transforms, NAACL, in: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4296–4313, 2022.
7. Tay, Y., Tran, V.Q., Ruder, S., Gupta, J., Chung, H.W., Bahri, D., Qin, Z., Baumgartner, S., Yu, C., Metzler, D., Charformer: Fast character transformers via gradient-based subword tokenization, ICLR, 2022.
8. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan,