HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

GRADUATION THESIS

Vietnamese Multi-document Summarization

LÊ HẢI SƠN
son.lh194449@sis.hust.edu.vn

Data Science and Artificial Intelligence
Information Technology

Supervisor: Associate Professor Lê Thanh Hương _____

Signature

Department: Computer Science

School: School of Information and Communications Technology

HANOI, 08/2023

ACKNOWLEDGMENT

First, I would like to express my appreciation to my supervisor, Assoc. Prof. Le Thanh Huong, for her invaluable guidance and support throughout the process of working on this thesis. I am grateful to Prof. Huong for her patience, meticulousness, and her enthusiasm in assisting me.

Furthermore, I want to show my gratitude to my parents and my family for their support. Their encouragement and belief in me served as a source of inspiration, enabling me to overcome the challenges encountered during the completion of this thesis. I would also like to express my sincere appreciation to the lecturers at the School of Information and Communication Technology (SoICT) of Hanoi University of Science and Technology. Their expertise, dedication, and insightful teachings have greatly contributed to my academic and personal growth.

Lastly, I would like to extend my thanks to all my friends and classmates. Their assistance has been invaluable not only throughout the duration of this thesis but also during the four years I spent studying and working at the Hanoi University of Science and Technology.

This work was supported by research grant of the "Vietnamese Multidocument Summarization" project, funded by the CyberIntellect LTD Company. The work related to this thesis - "LatVis: Large-scale pre-trained task-specific language models for low-resource Vietnamese multi-document summarization", has been also already accepted at LRL Workshop - LTC'23: 10th Language Technology Conference.

ABSTRACT

In the field of Vietnamese multi-document summarization, several challenges exist, such as dealing with long input sequences, generating summaries that resemble human-like quality, and the limited availability of labeled data. However, advancements in Transformer-based models, which benefit from parallel computation architecture and attention mechanisms, have partially addressed the issue of long input sequences. Moreover, these models, trained on extensive amounts of text, have demonstrated impressive performance in text generation tasks, approaching the level of human performance. Additionally, employing a pre-training strategy in a self-supervised manner has proven effective in overcoming the scarcity of labeled data.

With these considerations, the objective of my thesis is to utilize a large amount of unlabeled text to pre-train a language model specifically designed for Vietnamese summarization task. Following pre-training phrase, the model is fine-tuned on a small set of text summarization samples, aiming for high performance. In terms of Vietnamese multi-document summarization, my thesis contributes by adapting a promising transformer-based model called PRIMERA [1] to work effectively in the Vietnamese language. Unlike the original Transformer architecture, PRIMERA employs Longformer [2] to handle long input sequences, which are essentially concatenations of multiple documents. Furthermore, the authors of PRIMERA introduce a novel masking strategy named Entity pyramid strategy, which identifies significant information across documents and consolidates it into a single summary.

Experimental results demonstrate that my pre-trained model achieves comparable performance or sometimes higher in both zero-shot and full-finetuning evaluation scenarios. When fine-tuned using approximately 200 samples, our model achieves impressive Rouge scores [3], specifically 76.7%, 78.9%, and 73.9% for Rouge1-F1, and 50.2%, 55.0%, and 46.7% for Rouge2-F1, on the VMDS, ViMS, and VLSP datasets respectively. Evaluated on VNDS, a single-document dataset, the model also achieve comparable result with 63.0%, 33,3% and 42.7% for Rouge1-F1, Rouge2-F1 and Rouge-L-F1 respectively. Our model also achieves very good results in zero-shot evaluation. As far as I know, this is the first publicly available large-scale pre-trained language model specifically designed for Vietnamese multi-document summarization, showcasing its effectiveness in resource-limited languages.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Definition
BERT	Bidirectional Encoder Representations from Transformers
CNNs	Convolutional Neural Networks
LSTM	Long Short Term Memory
NER	Named Entity Recognition
NLP	Natural Language Processing
RNNs	Recurrent Neural Networks

CHAPTER 1. INTRODUCTION

1.1 Problem Statement

Text Summarization is the task of condensing lengthy texts into concise and coherent summaries that effectively conveys the main ideas and important information. This task is exceptionally difficult, even for human beings. With the massive amount of textual content available, especially due to the ever-expanding Internet, comprising web pages, news articles, status updates, blogs, and more, it becomes increasingly challenging to navigate through this vast amount of unstructured data. As a result, we often rely on search engines and skim through the search results. However, there is a need to reduce the length of this textual data by generating focused summaries that capture the important details. This would enable us to navigate the content more efficiently and determine whether larger documents contain the information we are looking for. Additionally, it is impractical to manually create summaries for all documents.

The task of Multi-document Text Summarization is even more challenging as it involves clusters of related documents that can be lengthy, containing overlapping content and spanning thousands of words. In my research on developing a deep neural network for Vietnamese Multi-document Summarization, I encountered a major obstacle due to the scarcity of labeled data. Until recently, only three small publicly available datasets, each comprising approximately 300 samples, existed. To overcome this limitation, recent studies such as Phobert[4], BartPho [5], and ViT5 [6] have utilized a large amount of unlabeled data for pre-training language models, followed by fine-tuning them on a smaller set of labeled data specifically for multi-document summarization.

However, these pre-trained language models have two common limitations. Firstly, they are designed as general-purpose language models, meaning that they are trained with the objective of modeling the probability distribution over sequences of words. For instance, BARTpho [5] is based on the BART model [7], which corrupts input text using a noising function and then learns to reconstruct the original text. Secondly, these models have architectures that are not well-suited for processing long input sequences, which poses a challenge for multi-document summarization tasks. Additionally, studies on multi-document text summarization in Vietnamese texts are still in the early phases and there are not many Vietnamese language models designed especially for this task because the resources are quite scarce.

Motivated by these limitations and my willing to contribute to the development

of research on abstractive multi-document summarization for Vietnamese text, this Bachelor thesis focuses on building a Vietnamese task-specific language model using an objective function specifically designed for Multi-document Summarization, which is modified from a English model to be suitable for Vietnamese contexts. This model will be then applied to a variety of Vietnamese summarization datasets and texts.

1.2 Background and Problems of Research

Recent advancements in deep learning techniques have significantly improved the performance of text summarization models over the past few decades. State-of-the-art approaches can be categorized as extractive (MatchSum [8], DiscoBERT [9], BertSumExt [10], and PNBert [11]), abstractive (BRIO [12] and SimCLS [13]), or hybrid (EASE [14]). The task is challenging due to the difficulty in data creation and performance evaluation. For Vietnamese, there has been little attention on Text Summarization, especially in the context of Multi-document Summarization.

Early studies in Vietnamese Text Summarization primarily focused on Single-document Summarization, predominantly employing traditional statistical methods. For instance, [15] combined various techniques such as word co-occurrences, TF-IDF, position-based, title-based, and proper noun-based methods to select salient sentences and generate the summary. Another study by [16] introduced a graph-based system that utilized a self-organizing map to cluster documents and extract the main idea.

In terms of Vietnamese Multi-document Summarization, there have been limited studies. One of the earliest research works by [17] presented an extractive system comprising three phases: pre-processing, score computation, and summarization generation. The system employed a set of manually selected features at both word and sentence levels, specifically tailored for Vietnamese news text, to compute sentence scores. These features included word frequency, word location, sentence position, time, and PageRank-based sentence features.

Recent research efforts have predominantly focused on optimizing the capabilities of transformer-based models for Vietnamese Text Summarization. For instance, [18] introduced extractive models that utilize variations of BERT to generate sentence embeddings. These models concatenate multiple documents into a single paragraph, employ BERT for sentence encoding, and utilize K-means clustering to rank and select salient sentences for summarization. More recently, [5] presented BARTpho, a large-scale Vietnamese sequence-to-sequence model based on the BART architecture [7]. BARTpho includes two versions: $\text{BARTpho}_{\text{word}}$ and $\text{BARTpho}_{\text{syllable}}$. These