

MASTER'S THESIS

A Redundancy-Aware Length Control Framework for Extractive Summarization

LI, Shuxin

Date of Award:
2021

[Link to publication](#)

General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

Abstract

While extractive summarization is an important approach of the NLP text summarization task, redundancy in the generated extractive summary is a big problem. Previous works usually set the length of the output summary to a fixed number, which might be appropriate for some of the documents while too long for others. At the same time, though extractive summarization possesses high readability as it directly selects sentences from the document, the unimportant parts within sentences are also selected. We propose a length control framework for extractive summarization, named *LenC*, in a two-stage pipeline to reduce the redundancy in the output. We first use a pretrained BERT-based summarizer to select smaller units (i.e., EDUs) than original sentences to abandon the insignificant parts of a sentence. Then a light-weighted length controller which could be attached to any summarization model is implemented to prune the output summary to an appropriate length. Experiments show that the proposed model outperforms the state-of-the-art baseline models and successfully reduces the redundancy in the extractive summaries.

Keywords: Single document Summarization, Redundant Information, Length Control

Table of Contents

DECLARATION	i
Abstract	ii
Acknowledgements	iii
Table of Contents	iv
List of Tables	vii
List of Figures	viii
Chapter 1 Introduction	1
1.1 Background	2
1.2 Motivation	3
1.3 Contributions	4
Chapter 2 Related Works	6
2.1 Extractive summarization	6
2.2 BERT-based model	7

2.3	Redundancy-awareness	7
2.4	Two-stage summarization	9
2.5	Dataset	9
Chapter 3 Methodology		13
3.1	Sentence Segmentation	13
3.1.1	Universal Conceptual Cognitive Annotation [1]	13
3.1.2	Rhetorical Structure Theory [16]	15
3.2	Data Preprocessing	16
3.2.1	Document slicing	17
3.2.2	Tokenization	17
3.3	Word Representation	19
3.3.1	One-hot encoding	19
3.3.2	Word embedding	19
3.4	Encoder-Decoder Framework	20
3.5	LenC Framework	22
3.5.1	Summarizer	23
3.5.2	Length Controller	24
3.6	Chapter Summary	26
Chapter 4 Experiments		27
4.1	Baseline Models	27
4.2	Experimental Setup	29
4.2.1	Metrics	29

4.2.2	Rule-based system	32
4.2.3	Dataset setup	34
4.2.4	Data analysis	35
4.2.5	Summarizer setup	36
4.2.6	Length controller setup	36
4.3	Experiment Results	37
4.3.1	Example analysis	40
Chapter 5 Conclusions and Future Works		42
5.1	End-to-end Model	42
5.2	Loss Function	44
5.3	Semantic Segmentation	44
Bibliography		45
CURRICULUM VITAE		52

List of Tables

2.1	Overview of common English datasets used in text summarization.	11
4.1	Results on CNNDM test set. F1-score for ROUGE-1, ROUGE-2, and ROUGE-L are reported. Results for models with an asterisk symbol (*) are captured from the corresponding papers. Models with symbol † are trained and tested on Disco-CNNDM.	38
4.2	The average number of words for the summary of different models.	39
4.3	Example of output summaries from LenC for Disco-CNNDM dataset	41

List of Figures

2.1	Flowchart for extractive summarization based on deep learning	9
3.1	Flowchart for extractive summarization learning in LenC	14
3.2	Example of UCCA annotation graph.	15
3.3	Example of RST diagram.	16
3.4	Overview of Encoder-Decoder Framework	21
3.5	The two-stage pipeline of <i>LenC</i> . Each s_i^k is wrapped with $[CLS]$ and $[SEP]$. Embedding layer processes texts with token embedding, interval segment embedding, and position embedding. A BERT-based summarizer is illustrated in the gray box. The length controller is attached to the discourse-level embeddings from the BERT-based encoder to control the output number of EDUs from the ranked s^k	22
4.1	Confusion Matrix for Binary Classification. Predict refers to generated text/summary. Actual refers to reference text/summary.	29

4.2	Distribution of the oracle length for both training and validation set in Disco-CNNNDM.	35
4.3	Curve of <i>ROUGE</i> – F_1 under different output lengths for short summary in validation set	40
5.1	Overview of an end-to-end model of <i>LenC</i> . $[CLS]$ token is placed in front of the input sequence s^k . each sentence representation in e^k goes through a sigmoid classifier to get the probability to be in the output summary. The length representation e_0^k for the first $[CLS]$ token will be further "analyzed" in the fully connected layer and a softmax classifier to calculate the length of output summary for the input article.	43