MASTER'S THESIS

A Redundancy-Aware Length Control Framework for Extractive Summarization

LI, Shuxin

Date of Award:
2021

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

List of Tables

List of Figures