Fusing External Language Model in Abstractive Summarization

Abstract

Recent sequence-to-sequence neural network models provide a viable new solution to abstractive text summarization, which aims to rewrite a long text into a short and concise form while preserving the most crucial information. However, these models face significant challenges when generating both semantically and syntactically correct summaries. In this work, we explore the potential approaches to incorporate an external (pre-trained) language model to augment the linguistic quality of text generation. This allows the internal (decoder) language model to focus more on jointly learning summary content selection and generation. Fused with the external language model, our abstractive summarization model achieves the results comparable to state-of-the-art models in terms of ROUGE scores, and meanwhile shows significant improvements in both perplexity and human evaluations.

1 Introduction

Text summarization aims to generate a short natural language summary that compress the information in the original longer text. Summarization approaches fall into two broad categories: extractive and abstractive. Extractive approaches (Cheng and Lapata, 2016; Narayan et al., 2018) typically assemble summaries from passages taken directly from the source text, while abstractive approaches (Chopra et al., 2016; Nallapati et al., 2016; See et al., 2017) are supposed to freely express with new words and phrases not featured in the source text. The recent success of sequence-to-sequence neural network models (Bahdanau et al., 2014) makes abstractive summarization a viable option. The summaries generated by state-of-the-art abstractive summarization models may have high word overlaps when compared against the gold summaries. However, when taking a closer look, the repeated text and the un-grammatical sentences are not uncommon in generated summaries. High ROUGE (Lin and Hovy, 2003) score does not guarantee the good quality and readability of summaries. In light of this problem, (See et al., 2017) introduce a coverage mechanism to address the repeated text issue. Meanwhile (Paulus et al., 2017) propose the intra-decoder attention and (Liu et al., 2018) equip with an additional discriminator to improve summary fluency. (Paulus et al., 2017) also notice that even though the best ROUGE score can be achieved by replacing the maximum likelihood objective with Reinforcement Learning (RL) to directly optimize the ROUGE metric, their RL approach tends to produce non-grammatical text and thus performs the worst in human evaluations. While existing approaches have shown to improve summary readability and fluency to some extent, they share the following limitations. The proposed mechanisms or strategies mainly cope with the language quality problem from very specific points of views, and thus cannot provide the general solution. For example, the coverage mechanism (See et al., 2017) helps generate summaries with less repeated text, but the disfluency and ungrammatical problem still severe. More important, training sequence-to-sequence models to improve the inadequate readability would be a great burden on the decoder. In essence, the decoder has two roles. One is related to summarization, i.e., to copy and fuse different parts in the source sentence using the attention mechanism. The other is related to text generation, i.e., to function like a language model. We refer to the decoder model as the Internal Language Model (ILM) considering it is a component within the sequence-to-sequence model. The supervised-learning nature limits its ability to sufficiently learn the language modeling ability with the current available manually generated summarization training data. In this work, we