Abstract

Text summarization is a challenging task of artificial intelligence and natural language processing, in that it reduces the size of an input document (or a set of documents) while preserving its meaning. The task has a long history dating back to 1950s. It mainly falls into two directions: extraction and abstraction. While extraction selects important information from a document, abstraction generates summaries, which are close to the writing style of humans. Even text summarization has extensively investigated by many studies in both directions, outputs of summarization systems are still far from human satisfaction, especially with abstractive summarization.

In the context of social media, users can freely reveal their viewpoints on an event and topics mentioned in a Web document published by a news provider. They tend to discuss an event by writing their comments on the web interface of a provider or posting relevant information on their timeline on social networks, e.g. Twitter. Such information has two important characteristics: (i) it reflects the content of an event and (ii) it includes viewpoints of readers. This observation suggests an interesting idea that the relevant social information of a Web document can be exploited to improve summarization. While traditional text summarization has deeply investigated, exploiting the support from social information for Web document summarization is still in an early stage, which requires more research. The objective of this thesis is to improve the quality of extractive summarization for single Web documents[2] by exploiting their social context.

First, we introduce three unsupervised ranking models, which formulate relationships between Web documents and their social context. The first model uses many lexical similarity features to measure the importance of a sentence and a user post in a mutual support fashion. More precisely, we encode the intra-information and inter-information of a sentence (or a user post) in a model, which ranks to extract summaries. The intuition behind this model is that important sentences include essential words or phrases which also appear in representative user posts. We show that by using a simple greedy selection method, this model obtains competitive results with state-of-the-art systems in term of ROUGE-scores. We also highlight that the number of social messages affects the importance estimation. The second model takes advantage of semantic similarity between sentences and user posts with an assumption that sentences and user posts share common topics denoted in the form of common words or phrases, which are in a variation writing style. From this, we present another ranking model, which combines intra-information and inter-information under a semantic similarity calculation. By using a greedy or an

[2]We consider main documents as single ones.

integer linear programming method, we show that this model obtains the best results in many cases. Among aspects affecting this model, we point out that the number of important words significantly influences the extraction of summaries. The third model explores the nature of sentences and user posts in sharing hidden topics presented in the form of common words or phrases. It encodes sentences and user posts in a unified ranking algorithm, which uses our proposed non-negative matrix co-factorization. It measures the importance of sentences and user posts by estimating their influence on hidden topics in term-sentence matrices. Experimental results indicate that this model obtains promising ROUGE-scores. We show that a joint optimization algorithm produces better results than an individual one.

Second, we present two learning-to-rank models to estimate the importance of sentences and user posts. We exploit social context by introducing many indicators extracted from three channels: local features, user-generated features, and third-party features for training summarizers. Following a supervised learning-to-rank algorithm which uses a greedy or majority voting method for sentence selection, we show that our models achieve the best results in many cases. Our analyses indicate that local features extracted from sentences in primary documents play an important role and features collected from their social context support local ones to improve the quality of the importance estimation step. Among our features, we point out that those from relevant Web articles are very useful in measuring the importance of sentences. We find that since sentences differ from user posts in term of writing style, different features should be used when modeling them.

Finally, we adapt deep learning for our task because it recently has achieved impressive results in many research fields, including text summarization. However, there are very little studies in applying this technique to our task. We first describe a well-known basic deep learning model, Convolutional Neural Network, for classification. Based on that, we adapt and extend it for our ranking purpose. Our model formulates relationships among $n-$grams in a sequence to enrich its representation. It also uses many our features to integrate social context into the ranking step. By doing that, our model obtains improvements compared strong baselines. Analyses show that using all features is inefficient due to the conflict when combining many different ones.

We apply our models to the task of sentence and highlight extraction on three datasets in two languages, English and Vietnamese. Promising results indicate that they can be viable alternative to extraction-based systems. Our findings and results contribute to the literature of text summarization as well as the task of summarizing Web documents by taking advantage of their social context.

Keywords: social context summarization, ranking, feature extraction, integer linear programming, deep leaning.