Abstract

Generating a hierarchical structure of topic-information (HST) for multiple documents written about the same topic is a new task in natural language processing. In a HST, topic-information is represented in a phrase and it can be seen as a title. Intuitively, a HST looks like a table-of-contents which is normally presented at the beginning of a book. It could play as a navigation tool to help readers quickly locate interesting parts. In addition, readers could look through a HST to get an overview of the topic of the document set. In this study, we propose a framework for generating a HST for multi-documents which involves three sequential tasks:

Text segmentation is a task of splitting a document into topically coherent segments. All documents in the set are put into a text segmentation system to get a collection of segments.

Segment combination is a task of merging and combining all the segments to form a hierarchical structure of segments (a tree of segments) which reflects the hierarchical structure of information.

Title generation is a task of generating a title for each node in the tree of segments. A title is a phrase which reflects the content of segments belonging to the node.

In the last decade, the text segmentation and title generation tasks have been received much attention from the research community. Although there are many studies investigated on various methods for these problems, the performance of available systems or published results are limited. Therefore, they are still open problems and challenges in the natural language processing field. Besides, the segment combination task is a particular task which is raised from our model. The literature related to that task is relatively sparse. Those open challenges are reasons for us to make a study on generating a HST for multi-documents. In addition, due to the dramatically improvement of computing power, people can now deal with problems which use large corpora and need high speed computation.

In this study, we aim to improve the performance of the above three tasks by using supportive knowledge in terms of semantic and topic information. The supportive knowledge which is a kind of semantic knowledge has been acquired from a large collection of texts by unsupervised learning algorithms such as word clustering and topic modeling.

The major research problems and our contributions are summarized as follows.

- First, the task of generating a HST for multi-documents is new. Therefore, we propose a framework which integrates the above three tasks in a pipeline to receive a set of documents as the input and produce a HST as the output. This framework allows us to improve the performance of tasks individually.

- Second, we focus on improving the performance of the unsupervised linear text segmentation. The current works on the task are mainly based on the assumption of lexical cohesion which consists of reiteration and collocation relations. However, they only take into account the first type of relations which can be easily recognized by observing the repetition of words. The second type of relations includes systematic and non-systematic semantic relation, which are the most complex relations to be recognized. In this study, we investigate on linguistics phenomena to find that supportive knowledge could be used to recognize these relations effectively. In addition, we also generalized current unsupervised text segmentation methods in a unique framework. The evaluation on public corpora shows the advantages of our model over the current state-of-the-art models.

- Third, the current learning models for the title generation task are still using non-semantic features about words in a text such as frequency, position, part-of-speech, syntactic function, and so on. That may be reason of the low quality of generated titles of current models. In this study, we investigate on a method to integrate semantic and topic information to the title generation learning model by using supportive knowledge. In addition, due to the lack of training data, we also investigate on using the word clustering to avoid the sparseness of data. We evaluated our proposed approach on a public dataset and get potential results.

- Finally, we investigate on the segment combination task which is raised from our framework for HST generation for multi-documents. In this study, we proposed a combination algorithm which is based on the hierarchical agglomerative clustering (HAC) method. This algorithm combines segments by the degree of topic relation between segments. The output of the algorithm is a tree which reflects the hierarchical structure of information. We also propose a heuristic algorithm to flatten the binary tree which is the output of the HAC-based algorithm to make the output look more realistic.

In summary, main contributions of this study are to propose a framework for generating a HST for multi-documents and to investigate on using supportive knowledge to improve the performance of the text segmentation and title generation tasks. The improved systems have been evaluated on the public datasets in comparison to the current state-of-the-art methods. We also did experiments on real datasets to verify the practical use of the framework.

Keywords: text summarization, multi-document summarization, text segmentation, title generation, supportive knowledge, topic modeling word clustering, lexical cohesion, semantic relation, semi-supervised learning, incremental perceptron.