Report on Topic-based Extractive Summary with Reduced Redundancy by Clustering

Minjune Hwang
Department of Electrical Engineering and Computer Science
University of California, Berkeley
mjhwang@berkeley.edu

***Abstract*—Extractive Summarization is an effective method to summarize a given set of documents by extracting a few sentences from corpus. While recent works in summarization field focus on the usage of neural network, there is a trade-off regarding computational efficiency and training time. Also, extractive summarization can be completely based on unsupervised methods. However, unsupervised extractive summarization often fails to detect redundant sentences and lowers the quality of the summary. In this report, we introduce a new topic-based extractive summarization method with reduced redundancy by clustering and selecting a representative from each cluster in topics. we also propose a new topic-based metric that can resolve the issue that Rouge score has.**

***Index Terms*—extractive summarization, clustering, topic modeling, summarization metrics**

I. Introduction

With large text corpora in which similar or identical sentences are repeated, unsupervised extractive summarization methods often fail to selectively choose sentences from different topics or sub-topics and rather repeat similar sentences in the summary. This is especially true when we select a certain number of sentences based on their cosine similarity to a given topic, as sentences with high similarity will simply contain multiple topic keywords, and it is likely for them to contain similar word combination or to be almost identical. Since the length of summary is a limited resource, repetition lowers the quality and succinctness of the generated summary.

Another problem is that ROUGE, a widely-used evaluation metrics for summarization methods, neither penalizes repetition nor imposes weights based on importance of words. In this paper, we propose a new topic-based extractive summarization method with reduced redundancy by clustering and selecting a representative from each cluster in topics. we also propose a new topic-based metric that can resolve the issue that Rouge score has.

II. Summarization Metrics

A. Problems with ROUGE score

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics and a software package used for evaluating automatic summarization and machine translation task. The metrics compare an automatically produced summary against a reference summary based on overlaps and co-occurrence of n-grams. ROUGE is currently the most widely used summarization metrics.

However, due to its simplicity, ROUGE has following problems.

Article: (...) He will miss the first # minutes of the opening practice session for the NAPA California # on Friday as a penalty for being late to the pre-race inspection last Sunday at Alabama 's Talladega Superspeedway for the DieHard # . It will mark the second straight race in which Benson has missed practice time because of a rules infraction . (...)
RL: benson to the for ford
Ground truth: benson penalized for his bad timing
ROUGE-L score: 0.358
Article: (...) Vikram S. Pandit is doing some serious spring cleaning at Citigroup . Since becoming chief executive in December , Pandit has been clearing out the corporate attic of weak businesses and unloading worrisome assets at bargain-basement prices . (...)
RL: sports column : citigroup to citigroup at citigroup
Ground truth: citigroup embarks on plan to shed weak assets
ROUGE-L score: 0.25

Fig. 1. Example of bad headlines with high ROUGE scores

- Fails to penalize repeated words or sentences which lower the quality of summaries, as shown in the above table.
- Fails to differentiate how important or relevant words are regarding the content or topics in documents.
- Fails to penalize lengthy documents (similar to the first problem).
- Requires a reference summary.

B. Topic-Based Metrics

In order to solve previously shown problems of ROUGE, we propose a topic-based metrics that can resolve those issues. With any topic modeling method, we can use topic weights and keywords to impose weights on importance of words in each sentence in the generated summary (to resolve the second issue of ROUGE).

For any arbitrary topic modeling method (e.g. Latent Dirichlet allocation, sparse PCA), let's say we have m topics and n keywords for each topic. We define t_i to be the strength of topic i, and w_{ij} to be the weight of jth word of topic i. Subsequently, $\sum_{i=1}^{m} t_i = 1$ and $\sum_{j=1}^{m} w_{ij} = 1 \quad \forall i \in [1, m]$. (*For the greater word coverage, it is recommended to use a relatively large number for n.)

Then, we define the topic-based metric, s, as following.

$$s = \sum_{i=1}^{m} \sum_{j=1}^{n} t_i w_{ij} a_{ij} \quad (1)$$

Here, a_{ij} is the boolean variable that shows whether jth word of topic i exists in the generated summary. Because $a_{ij} \in \{0, 1\}$, $s \in [0, 1]$. Since we use the boolean vector for the appearance of each word, our metric does not encourage to generate a summary with redundant sentences and words, and penalize repeated sentences and words, given that we limit the number of sentences or words in the summary. This resolves the first and the third issue of ROUGE score.

C. Redundancy Metrics with Jaccard Index

Work in progress

III. Methodology

A. Topic-Based Extractive Summarization

Our base method is an unsupervised extractive method with topic modeling. After computing topics with any topic modeling method, we select sentences in the original document based on their cosine similarity to each topic. Like the previous section, we define t_i to be the strength of topic i and w_i to be the weight vector of words in topic i; w_{ij} will be the weight of jth word of topic i as before.

If the desired number of sentences is l, we select top $\lfloor l * t_i \rfloor$ sentences with highest cosine similarities, for topic i.

As discussed above, this method is prone to select repetitive sentences if the original corpus contains multiple similar sentences that repetitively have keywords with large weights.

B. Clustering for Reducing Redundancy

To solve the above issue, we first select a number of sentences for each topic. Then, we partition sentences into k clusters, where k is the desired number of clusters. We select $\left\lfloor \frac{l*t_i}{k} \right\rfloor$ sentences from each cluster. We expect almost identical sentences to be classified into a same cluster so that our method will avoid to select those sentences to be in the final summary.

For K-means, LDA,

QDA

C. Other Adjustments

After extracting sentences from each cluster in sentences with high similarity with each topic, we sort them in the order they appeared in the original document in order to preserve the structural flow of the document.

Also, we

IV. Experiment

We use Philippuine Earthquake Article Dataset from multiple sources, collected in PEER project. We compare the performance of different summarization methods, including Gensim (TextRank), topic-based method (base method) with LDA and SPCA, and our method with different clustering methods.

We use both including generative methods like LDA and QDA, and discriminative methods like K-means for clustering.

V. Result

A. Empirical Result

In general, we observe that general methods our method generates summary with less or no redundancy. Below is an excerpt from a generated summary without clustering.

B. Metric-based Result

VI. Future Improvements

There are following directions we can take.

- Summarization by merging summaries of each section in the document for the sake of flow or naturalness of summary. This will also prevent repetition since most similar sentences exist in the same section.
- Fails to penalize lengthy documents (similar to the first problem)