Extractive Text Summarization Implemented with SKG Formula

Nann Kavdavid Ryuichi Matoba
National Institute of Technology, Toyama College
{i1421442, rmatoba}@nc-toyama.ac.jp

1 Introduction

The amount of information available today is tremendous and the problem of finding the relevant pieces and making sense of these is becoming more and more essential. Nowadays, a great deal of information comes from the Internet in a textual form. The challenge of finding relevant documents on the web is mainly handled by information retrieval techniques.

Summary generated from automatic summarization system can be used as the replacement for the original content or help to identify the events that a person is particularly interested in. The straightforward way to generate a summary is to select several sentences from the original text and organize them in way to create a coherent text. The main idea of summarization is to find a representative subset of the data, which contains the information of the entire set.

Therefore, the purpose of this research is to extract important information, and to generate informative and short summary from a text.

2 Method

2.1 Preprocessing

Use a raw text as input and apply some basic elements such as stop words removal and stemming. The output is a group of sentences was stemmed without stop words containing. Because stop words does not have a lot of meaning in the text.

2.2 Sentence representation

Textual contexts are often represented in terms of features. From the output of the preprocessing, we can analyze the important words by using TF-IDF to calculate the weight of each word in the sentences.

The TF-IDF weight is the product of two components; the term frequency TF and the inverse document frequency IDF. The TF determines the importance of a word for a given document, while the IDF indicates the importance of a word over the whole set of documents. A word that occurs often in a specific document but rarely in other documents is considered to be relevant for this document and, consequently, receives a high weight value. The TF-IDF weight for word w in document d is calculated as follows:

$$tf \cdot idf(w,d) = tf(w,d) \times idf(w) \tag{1}$$

$$tf(w,d) = \frac{tc(w,d)}{|d|} \tag{2}$$

$$idf(w) = log\frac{n_d}{df(w,d)+1} \tag{3}$$

Where tc(w,d) (term count) is the number of times word w occurs in document d, $\mid d \mid$ is the number of words in document d, $\mathrm{n_d}$ is the number of documents in a collection of documents. df(w,d) is the number of documents in d that contain word w.

2.3 Similarity measurement

2.3.1 Latent semantic analysis (LSA)

LSA is also known as latent semantic indexing, LSI, though strictly that refers to its use in persistent indexes for information retrieval purposes [2]. LSA helps finding the similarity between words and sentences. It uses a matrix factorization method called singular value decomposition (SVD) to approximate the initial term-context matrix by a matrix of a much smaller size.

Mathematically, truncated SVD applied to training samples X produces a low-rank approximation X. The formula is as follows:

$$X \approx X_k = U_k \Sigma_k V_k^T \tag{4}$$

Where, $\mathrm{U_k}$ is the SVD term matrix, Σ_k is the singular values, and V_k^T is the SVD document matrix. In this experiment, we use k=2 to get a two dimensional vector. The result returns, the term are represented by the row vectors of the m×k matrix $U_k \Sigma_k$,Whereas the documents by the column vectors the k×n matrix $\Sigma_k V_k^T$.

2.3.2 Cosine similarity

Textual similarity is a complex concept that can be defined as the semantic relatedness of two textual contexts. Two contexts are considered similar if they focus on the same or related concepts, actors or actions.

The cosine similarity metric based on the angle between two vectors is one of the most widely used similarity metric. The cosine similarity is calculated as follows:

$$similarity = cos(\theta) = \frac{XY}{\|X\| \, \|Y\|} \tag{5}$$

2.4 Content selection

The goal of the selection procedure is to identify a set of sentences that contain important information. Three criteria are optimized when selecting the sentences: relevance, redundancy and length.

This step is to calculate total score for each sentence and each word. So we can use this score to find which sentences are important for the text. There are 3 types of score. They are Document centrality score, Query relevance score, and Compactness score.

2.4.1 Document centrality score

The idea behind document centrality score is to select sentences that contain informative words, also referred to as topic signatures. Stop words like articles and pronouns are usually ignored.

$$score(s) = \frac{1}{|s|} \sum_{w \epsilon s} weight(w) \tag{6}$$

Where w is the word that appears in a sentence S and $|S|$ is the number of words in the sentence S.

2.4.2 Query relevance score

The sentences that are similar to many other sentences are likely to contain the information that should be included in the summary.

$$score(s) = \frac{1}{|D|} \sum_{U \epsilon D} similarity(S, U) \tag{7}$$

Where U is a sentence other than S from a set of sentences D, $|D|$ is the total number of sentences in the input documents excluding S.

2.4.3 Compactness score

The intuition is that a good summary should contain short but informative and relevant sentences. By using the length of a sentence; longer sentences get lower scores.

$$score(s) = \frac{1}{|S|} \tag{8}$$

Where $|S|$ is the length of sentences S.

2.5 Sentences shortening

Up until this step, we could extract the sentences that are informative for the context. However, in some of the sentences there should be some words or clauses that are not so important for the text. Therefore, in this step, by using Enju parser and SKG (Short Keywords and Grammatical) formula, the system can shorten the sentences with grammatically correct and meaningful sentences.

2.5.1 Enju parser

In order to make a shorter and grammatical correct sentence, a Head Driven Phrase Structure Grammar (HPSG) is needed.

Enju is a syntactic parser for English. The grammar used by the parser is based on HPSG [3]. Enju can analyze syntactic or semantic structures of English sentences can output phrase structure and predicate-argument structures.

2.5.2 SKG (Short, Keywords and Grammatical) formula

SKG is a formula that we come up by combining the length of sentence, weight and the similarity of the words in the sentence. It is an algorithm to get a shorter part of a sentence which is holding the most important information of the sentence. The formula is as follows:

$$SKG(s) = \frac{1}{L^2} \sum_{i=1}^{n} (weight(w_i) \cdot \sum_{j=1, j \neq i}^{n} (similarity(w_i, w_j))) \tag{9}$$

Where L is the length of sentence S, w is the word in the sentence. Weight is the weight calculated by TF-IDF, and Similarity is calculated by LSA and cosine similarity in the previous step.

From the value of SKG, there is a possibility that the value of SKG of the original sentence is greater than the shorter one. It means that the sentence should not be reduced to be shorter, since it holds important meaning to the context.

Suppose that we have one of the important sentences as in the example below.

Example: we have two sentences. The First sentence is the sentence Before using SKG formula, and the second sentence is the sentences after using SKG formula.

1. **For example, a case in which participants used sealed envelopes to place their bids on a piece of real estate represents this type of auction.**
2. **Participants used sealed envelopes to place their bids on a piece of real estate represents this type of auction.**

As in the example some words like “For example, a case in which” is reduced to make the sentence shorter but still contain the meaning for the original sentences.

3 Experiment

3.1 Setup

TAC 2011 Guided Summarization Task data set has been used in our experiment to compare our system summary and ideal summary for the evaluation. The data set is to be summarized in a maximum of 100 words. There are 4 human reference summaries, against which an automatically generated summary is compared. And, ROUGE (Lin, 2004) is used to evaluate the summarization results.

ROUGE is a method to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans [4]. In Rouge score, Rouge-1 which performed great in evaluating very short summaries, and Rouge-2 which worked well in single document summarization tasks, are used in the evaluation.

3.2 Result

	Recall	Precision	F-score
Rouge-1	0.45684	0.28549	0.35139
Rouge-2	0.11239	0.06935	0.08578

Table 1: average Rouge score before using SKG

	Recall	Precision	F-score
Rouge-1	0.42477	0.30070	0.35213
Rouge-2	0.10465	0.07353	0.08637

Table 2: average Rouge score after using SKG

Table 1 and Table 2 show the average Rouge score before and after using SKG, respectively.

The table shows that after using SKG formula, the average Precision and average F-score is higher than before using SKG formula. However, average Recall is decreasing. Since average F-score is considered both the precision and the recall of the test to compute the score. The evaluation can be judge by the F-score. For Rouge-1 score, there is an increasing rate of 0.21% between using and not using SKG formula. Indeed, for Rouge-2 score, the rate is about 0.68%.

4 Discussion

From the experiment result, we notice that there is a drop in the average recall but there is a rise in the average Precision and the average F-score. Since the purpose of using SKG formula is to help reduce some not important words in sentences, so the number of words in the summary is also reduced compare to the one before using the SKG formula.

The average recall is the number of sentences occurring in both system and ideal summaries divided by the number of sentences in the ideal summary. By reducing the number of words in sentences, it is naturally that the average recall score is decreased. On the other hand, the average precision is the number of sentences occurring in both system and ideal summaries divided by the number of sentences in the system summary. We can think that it is naturally increase if the number of words in the sentence is reduced.

But there is the average F-score which is the combination of the average recall and average precision that is often used in the field of information retrieval for measuring search, document classification, and query classification performance. From the result, we can say that SKG formula help improve the precision of the summary and make the sentences shorter.

5 Conclusion

In this research, we extract the important part of the sentences from a context to create an extractive summary. Also using the SKG formula to reduce some unimportant part of the sentences to make sentences even shorter but still hold the meaning as the original sentences. We can conclude that even after using SKG formula, there is only a small increasing rate in Rouge score but still it can help to reduce more and more unimportant information in textual information.

References

[1] Gleb Sizov, Extraction-Based Automatic Summarization Theoretical and Empirical Investigation of Summarization Techniques, NTNU

[2] Markovsky I. (2012) Low-Rank Approximation: Algorithms, Implementation, Applications, Springer, 2012

[3] Kenji Sagae, Yusuke Miyao, and Jun’ichi Tsujii. 2007. HPSG Parsing with Shallow Dependency Constraints. In Proceedings of ACL 2007.

[4] Chin-Yew Lin, ROUGE: A Package for Automatic Evaluation of Summaries, Information Sciences Institute, University of Southern California