Journal of Engineering Science and Technology
Special Issue on ICCSIT 2018, July (2018) 144 - 151

TEXT SUMMARIZATION EVALUATION BASED ON SENTENCE SCORING AND CLUSTERING

MUHAMMAD AZHARI[1,*], YOGAN JAYA KUMAR[2], ONG SING GOH[3], BASIT RAZA[4]

[1,2,3] Faculty of Information and Communication Technology,
Universiti Technical Malaysia Melaka,
76100 Durian Tunggal, Melaka, Malaysia
[4] Department of Computer Science, COMSATS Institute of Information Technology,
Islamabad, Pakistan
*Corresponding Author: muhd.azhari.ayie@gmail.com

Abstract

A summary system comprises a subtraction of text document contents to generate a new form that delivers the essentials contents of the text documents. Due to the hassle of documents overload, getting the right information and effectively-developed summaries are essential in retrieving information. Reduction of information allows users to find the information needed quickly without the need to read the full document collection. The purpose of this paper is to evaluate the performance of sentence scoring and clustering in the process of generating text summaries. Adaptive Neuro-Fuzzy Inference System (ANFIS) has been used to score the sentences and clustering were performed using K-Means and Hierarchical Clustering (HC) approaches. From the experimental findings, it was found that no improvements in the quality of the generated summaries obtained by simply performing clustering. This paper proposes some recommendations to improve text summarization.

Keywords: Adaptive neuro-fuzzy inference system, Hierarchical clustering, K-means clustering.

1. Introduction

Recently, studies in the field of text summarization have grown because the capacity of information retrieval has increased with respect to the current masses of information load. Whilst a user enters a query online, the response includes several web pages regarding the query, and it is difficult and time consuming for readers to digest them. Research in automated text document summarization has received tremendous attention due to the ever-increasing information available on the web. A summary refers to a textual content that is produced from single or multiple documents, which incorporates an essential portion of the content from the source documents. There are various types of summaries which can be produced namely domain specific, generic, query-based and abstractive summaries [1, 2].

Human-generated summaries are usually produced by domain experts who can synthesize and assess the important concepts in the documents. Since the produced summary is an abstract, not explicitly contained within the input, this turns to be a challenge for computers to generate human-like summaries. Since computer systems now do not yet have the language abilities of humans, alternative techniques have been taken into consideration, which mainly uses sentence features to produce extractive summaries.

In this paper, we intend to evaluate text summarization using sentence scoring and clustering. ANFIS has been used to score the sentences based on the sentence features. The clustering model is implemented using the term frequency-inverse document frequency (tf-idf) score to cluster the sentences. We compare the performance of clustering-based summarization models. The rest of this paper is prepared as follows: Section 2 discusses the related works. Section 3 presents clustering-based summarization models. The experimental results and discussion are given in Section 4. Lastly, the conclusion is presented in Section 5.

2. Related Work

The approach presented in [2] proposed a cluster-based multi-document summary using feature extraction approach. The similar documents are clustered into the same cluster using threshold-based document clustering algorithm. The features that have been used are word weight, sentence title, sentence length, sentence position, proper nouns and numerical facts in the sentence. The sentence overall score is calculated by using its features score. Sentences are then picked from every cluster based on its sentence score. Selected sentences are organized in sequential order as it appears the source documents to form the summary.

There are research works done by others which used similar approach [3, 4]. They considered the clustering sentences, cluster order, and selection of representative sentences from the clusters. Similarity histogram based on incremental clustering approach has been used for sentence clustering. It is an unsupervised sentence clustering technique. Several important words contain will be used to measure the importance of the cluster. Additionally, the cluster will be ordered based on the cluster rank and top sentences in clusters are then selected.

Radev et al. [5] used cluster centroids and top ranking tf-idf to represent the cluster. The sentence that had similarity to these centroids will be selected and used to form the summary. Another commonly used clustering algorithm is the k-means

clustering algorithm. The k-means clustering algorithm has been widely used to represent diversity and reduce redundancy within the data source.

The abovementioned studies have incorporated diversity that was achieved by clustering to produce summaries from input document content. However, it should be investigated if diversity is significant for a summary generation, especially when we deal with the summarization of news articles. This study will investigate the effect of sentence clustering on the summary quality. Furthermore, both hierarchical and non-hierarchical based clustering methods will be implemented with our ANFIS based summarization model.

3. Clustering based Text Summarization

In this paper, we will evaluate text summarization using sentence scoring and clustering. For sentence scoring, we will be using ANFIS [1]. It combines both fuzzy reasoning and neural networks concepts and has the capability to combine the advantages of both techniques. Its inference engine comprises a set of fuzzy IF-THEN rules which have learning functionality to estimate sentence importance [4]. Clustering is then used to group the similar sentences [6]. The approach assumes that summarization result not only relies upon at the sentence features but also relies upon on the diversity in the document contents. It allows to cluster the sentences from the document and extracts the topic sentences to generate the summary. In this study, we will be using both the widely used non-hierarchical K-means clustering method [3] and Hierarchical Clustering (HC) method [7, 8] as our evaluation for clustering-based text summarization. Figure 1 depicts the structure of the summarization model which is based on sentence scoring and clustering.

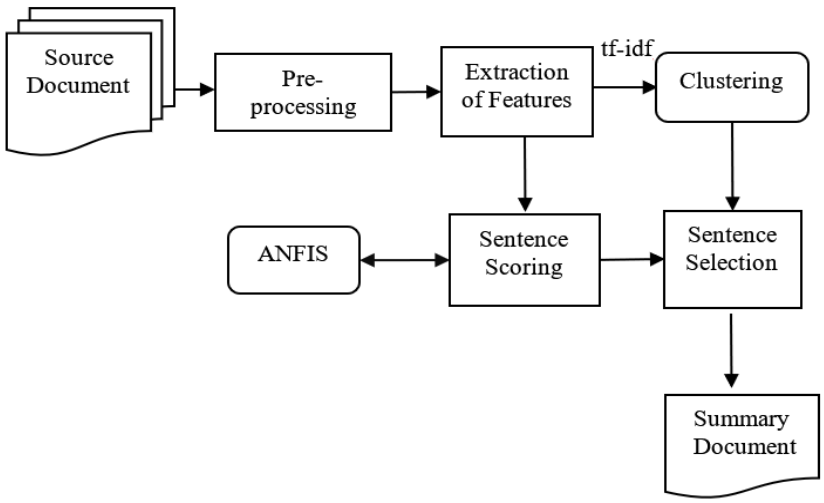

Fig. 1. Text summarization model structure.

3.1. Algorithm

The algorithm for the text summarization model evaluation is given here:

1. Source document set obtained from DUC 2002.
2. Pre-processing: Removing stop words and stemming word in the texts.

3. Features extraction: five features were extracted from each sentence *fn = {f1, f2, f3, f4, f5}* [1].
4. Sentence scoring: ANFIS model will be used to score the sentences.
5. Clustering: tf-idf score will be used for clustering the sentence.
    a. 5 clusters will be set (average length 20 words per sentence and max summary 200 words, 20/200=5 clusters).
    b. Removal of redundant sentences using word overlap check.
    c. The sentence will be ranked by position.
6. Sentence selection:
    a. Select one sentence from each cluster starting from cluster 1.
    b. Selected sentence score must equal or more than threshold value 0.6.
    c. If total length < 200 word, repeat step 6 (a).
7. Summary document: the final summary is obtained.
8. Validation: Rouge-n [9].

3.2. Data set

For this study, Document Understanding Conference (DUC) 2002 dataset for multi-document summarization task has been used. The dataset comprises of news articles related to natural disaster events. It also provides annotations of human selected sentences as summaries. In 2008, DUC has become a summarization track in the Text Analysis Conference (TAC). TAC is a sequence of assessment workshops prepared to encourage research in natural language processing and associated applications, by offering a massive test collection and common assessment procedures. TAC is prepared by the Retrieval group of the Information Access Division (IAD) within the Information Technology Laboratory (ITL) at the National Institute of Standards and Technology (NIST).

3.3. Extraction of features

The pre-processed input documents which are segmented into sentences are represented as feature vectors. These features are properties used to represent summary to determine its important contents. Many features were computed and used in related research works [10-14]. The score for these features is in the range of [0, 1]. Following are the five features used as inputs to ANFIS:

3.3.1. Title feature

Sentences containing the words in the title of the document will be given a high score, as shown in Eq. (1)

$$f_1 = \frac{title\ words\ appearing\ in\ sentence}{total\ title\ words} \quad (1)$$

3.3.2. Length of sentence

A sentence that is long is considered to acquire essential data, as shown in Eq. (2)

$$f_2 = \frac{total\ sentence\ words}{maximum\ length\ of\ sentence} \quad (2)$$

3.3.3. Proper noun

A sentence containing proper nouns, place or thing is thought to be an imperative sentence, as shown in Eq. (3)

$$f_3 = \frac{proper\ nouns\ appearing\ in\ sentence}{length\ of\ sentence} \tag{3}$$

3.3.4. Thematic word

This feature is utilized to decide the regularity of a term. A term that is utilized as often as possible is presumably identified with its appearance in the texts. Ten frequent sentence words were used as thematic terms, as shown in Eq. (4)

$$f_4 = \frac{frequent\ sentence\ words}{\max(frequent\ sentence\ words)} \tag{4}$$

3.3.5. Term weight

Each sentence can be weighted based on term frequency-inverse sentence frequency (*tf-isf),* as shown in Eq. (5)

$$f_5 = \frac{\sum_{i=1}^{k} W_i(S)}{Max(\sum_{i=1}^{k} W_i(S))} \tag{5}$$

where $w_i(s)$ is the *tf-isf* score of term w_i in sentence *s*.

4. Experimental Findings and Discussion

Here, we compared the results generated by both sentence scoring (ANFIS) and clustering methods (both hierarchical and non-hierarchical based clustering methods) which we have been implemented as a part of our evaluation study. The evaluation is done using ROUGE: Recall-Oriented Understudy for Gisting Evaluation [9]. ROUGE measures the similarity between the model and human summaries. Among ROUGE assessment measures, ROUGE-1, ROUGE-2, ROUGE-S*, and ROUGE-SU are commonly applied in the multi-document summary evaluation [15]. Each measure provides the average precision, recall, and F-measure. Refer to Table 1 and Fig. 2; it can be observed that no improvements in the quality of the summaries obtained by simply performing clustering and selecting a sentence from the generated clusters. In fact, the ROUGE scores obtained from clustering have reduced further.

Table 1. F-Measure comparison between sentence scoring and clustering.

	ROUGE Evaluation					
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-W**	**ROUGE-S***	**ROUGE-SU**
ANFIS	0.34684	0.16462	0.33609	0.16804	0.10682	0.11105
K-mean	0.22625	0.06487	0.21877	0.11193	0.04693	0.05005
HC	0.22321	0.10869	0.21734	0.11221	0.06333	0.06616

During our experimental observation, we found that by lowering the threshold value of sentence score will improve its recall and lower its precision. We assume that by improving the recall percentage, we can improve the summary result. However, by doing so, it will lower the chance of selecting the important sentence. For example, there were 20 sentences selected by a human expert (from DUC annotation) and 18 sentences were recalled by the system. Out of 18 sentences, only 10 sentences matched the human expert's selection, thus lowering its precision. Further analysis shows that sorting sentences by its position had better chance to be selected, similar as it would be chosen by a human because most of these important sentences appear in the beginning segments of the news article. This could explain why clustering could not improve the results, as the clusters group similar sentences together and do not necessarily have sentences with a high score in each cluster. By considering the sentences score with appropriate score threshold together with its position in the original document, it could possibly improve the selection of summary sentences.

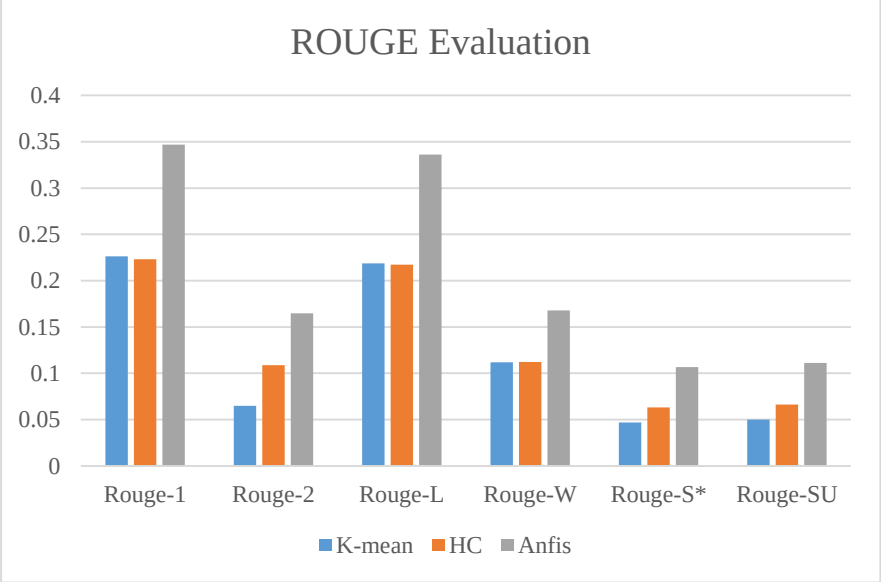

Fig. 2. ROUGE comparison between sentence scoring and clustering.

We additionally performed T-tests to demonstrate the distinction in comparison amongst ANFIS, K-means and HC for the average precision, recall and F-measure obtained from ROUGE evaluation. The T-test is utilized to decide whether there is a significant contrast between the scores of two groups. The T-test result given in Tables 2 and 3 show that the results of the two comparison methods were statistically significant. The comparison methods display the significance level, 0.05. Thus, it could be concluded that the results did not improve even when tested on both hierarchical and non-hierarchical based clustering methods.

Table 2. T-test between K-Means and ANFIS.

Paired	ROUGE	Sig.
	Precision	0.000

K-Means- ANFIS	Recall	0.000
	F-Measure	0.000

Table 3. T-test between hieratical clustering (HC) and ANFIS.

Paired	ROUGE	Sig.
HC - ANFIS	Precision	0.006
	Recall	0.000
	F-Measure	0.000

5. Conclusions

The purpose of this paper is to evaluate the performance of sentence scoring and clustering in the process of generating text summaries. The study investigates the effect of sentence clustering towards the quality of summary to determine if diversity really matters for a summary generation, especially when we deal with the summarization of news articles. Here, both hierarchical and non-hierarchical based clustering methods were implemented with the summarization model. ANFIS was used to score the sentences and clustering were performed using K-means and HC approaches. The experimental findings show that performing clustering and selecting sentences from each cluster will not necessarily improve the results. Thus, recommendations for improvement have been highlighted.

Acknowledgements

This research work supported by Universiti Teknikal Malaysia Melaka (UTeM) and Ministry of Higher Education (MOHE), Malaysia Grant No. RAGS/1/2015/ICT02/ FTMK/02/B00124.

Nomenclatures

tf-idf	Term frequency - inverse document frequency
tf-isf	Term frequency - inverse sentence frequency
f_i	Sentence feature
w_i	Term weight

Abbreviations

ANFIS	Adaptive Neuro-Fuzzy Inference System
DUC	Document Understanding Conference
HC	Hierarchical Clustering
IAD	Information Access Division
ITL	Information Technology Laboratory
NIST	National Institute of Standards and Technology
TAC	Text Analysis Conference
ROUGE	Recall-Oriented Understudy for Gisting Evaluation

References

1. Azhari, M.; and Kumar, Y.J. (2017). Improving text summarization using neuro-fuzzy approach. *Journal of Information and Telecommunication*, 1(4), 367-379.

2. Al-Hashemi, R. (2010). Text summarization extraction system (TSES) using extracted keywords. *International Arab Journal of e-Technology*, 1(4), 164-168.
3. Jain, H.J.; Bewoor, M.S.; and Patil, S.H. (2012). Context Sensitive text summarization using k means clustering algorith. *International Journal of Soft Computing and Engineering*, 2(2), 301-304.
4. Babar, S.A.; and Patil, P.D. (2015). Improving Performance of text summarization. *Procedia Computer Science*, 46, 354-363.
5. Radev, D. R.; Jing, H.; Styś, M.; and Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6), 919-938.
6. Fattah, M.A.; and Ren, F. (2008). Automatic text summarization. *international journal of computer. Electrical, Automation, Control and Information Engineering*, 2(1), 90-93.
7. Balcan, M.F.; Liang, Y.; and Gupta, P. (2014). Robust hierarchical clustering. *The Journal of Machine Learning Research*, 15(1), 3831-3871.
8. Gollapudi, S.; Kumar, R.; and Sivakumar, D. (2006). Programmable clustering. *Prococeedings of the 25th ACM Symposium on Principles of Database Systems*, 348-354.
9. Lin, C.Y. (2004). ROUGE : A package for automatic evaluation of summaries. *In Proceedings of Workshop on Text Summarization of ACL, Spain*. 74-81.
10. Litvak, M.; and Last, M. (2008). Graph-based keyword extraction for single-document summarization. *Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization*, 17-24.
11. Mahdipour, E.; and Bagheri, M. (2014). Automatic Persian text summarizer using simulated annealing and genetic algorithm. *International Journal of Intelligent Information Systems*, 3(6), 84-90.
12. Moradi, M.; and Ghadiri, N. (2016). A Bayesian approach to biomedical text summarization. *arXiv Preprint arXiv:1605.02948*, 1-25.
13. Saleem, S.M.; Krithiga, R.; Rani, S.K.; and Sindhya, S.C. (2015). Study on text summarization using extractive methods. *International Journal of Science, Engineering, and Technology Research (IJSETR)*, 4(5), 1399-1405.
14. Shinde, R. D.; Routela, S.H.; Jadhav, S.S.; and Sagare, S.R. (2014). Enforcing text summarization using fuzzy logic. *International Journal of Computer Science and Information Technologies*, 5(6), 8276-8279.
15. Lin, C.Y. (2004). Looking for a few good metrics: Automatic summarization evaluation-how many samples are enough? *Proceedings of NTCIR Workshop* 4, Tokyo, Japan.