Volume 2, Issue 7, July 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Concept Frequency: A Feature Set Based Text Compression Model

P.Naveen Kumar[1], Dr. A.P. Siva Kumar[2]
[1]M.Tech, Department of CSE
[2]Assistant Professor,Department of CSE
JNTUA College of Engineering
Anantapur, India

Abstract— **A summary is a shorter version of the original. Such a simplification highlights the major points from the much longer subject, such as a text, speech, film, or event. The purpose is to help the audience get the gist in a short period of time. Automatic summarization involves reducing a text document or a larger corpus of multiple documents into a short set of words or paragraph that conveys the main meaning of the text. In this paper, we propose a new multi-document summarization approach that makes use of a feature called 'concept frequency' that verifies sentences nearness with a cluster opted. Hence the summarization relevance would be more effective. A novel feature called exemplar is to help simultaneously deals with sentence ranking. A fuzzy medoid-based clustering approach is used to produce sentence clusters or subsets where each of them corresponds to a subtopic of the related topic.**

Keywords— **Concept frequency, exemplar, text Summarization**

I. INTRODUCTION

The data in the internet is increasing every day and thousands of documents are produced and made available in the internet. The amount of information available in documents exceeds our capacity to read them. We need to get the right information without having to go through the whole document. Therefore, we need a summary of the document so that we can get the gist of the whole document.

Text compression is the process of decreasing the large content to a shorter form which will be easy to read.

There are two different approaches to generate summaries from text: extractive and abstractive. The *extractive* approach extracts sentences or parts of sentences from text, and is the most common way to perform summarization. The second and substantially more difficult approach is called *abstractive*, and involves generating summary text using natural language processing techniques.

There are different types of summaries depending what the summarization program focuses on to make the summary of the text, for example generic summaries or query relevant summaries (sometimes called query-based summaries). Summarization systems are able to create both query relevant text summaries and generic machine-generated summaries depending on what the user needs. Summarization of multimedia documents, e.g. pictures or movies, is also possible. Some systems will generate a summary based on a single source document, while others can use multiple source documents (for example, a cluster of news stories on the same topic). These systems are known as multi-document summarization systems.

Multi-document summarization differs greatly from single-document summarization. In fact, single-document summarization can be considered as one of the critical sub-tasks of multi-document summarization. Single document summaries are produced with the goal of presenting the most important information in the document. Whereas the multi-document summaries are produced by extending the techniques of single document summarization to multiple documents. In this paper, we focus our efforts on multi-document summarization through sentence extraction.

II. RELATED WORK

Various methods for single document summarization will extracts the important sentences from the document by ranking the sentences based on sentence position, term frequency, inverse document frequency, etc. Multi document summarization is different to single document summarization. In single document summarization, only the relation between the sentences in that document is verified. Whereas in multi-document summarization, the relation between sentences from different documents should be verified.

Gees C. Stein proposed an approach for multi-document summarization that summarizes each document individually and then it groups all the summaries into clusters. It picks up the most important passage(s) or sentence(s) from each cluster and forms a final summary.

Jian-Ping proposed a novel sentence extractive summarization SumCR using anew subtopic-level feature Exemplar and a document-level feature Position. This system presents overview of the generic multi-document summarization. The overall system structure of the SumCR is, Given a set of documents $D = \{d1, d2, \ldots, dn\}$ related to the same topic, SumCR produces a short summary for the document set with several procedures given as follows:

1. Sentence similarity matrix: First, every document is parsed into sentences to get a set of sentences $X = \{x_i,$

$x_2, \ldots, x_N\}$, where *N* is the total number of sentences from *n* documents. After that, the similarity matrix $S_{N\times N}$ that contains similarities between every pair of sentences is calculated.

***2. Exemplar* score**: After getting the similarity matrix, fuzzy clustering method is applied to group the sentences into clusters. At the same time, prototype weight of each sentence with respect to a cluster is also produced at the end of the clustering process. The *Exemplar* weight of a sentence considers both the closeness to sentences in the same cluster and the prototype weight of this sentence in the cluster it belongs to.

3. Summary generation: Sentences are scored with the exemplar weight. Sentences are ranked with their scores in descendent order and those top ones are selected to form the summary. The pairwise similarities of sentences being selected need to be smaller than a given threshold. By selecting sentences on scoring to form summary gives redundancy. Various existing summarization systems for multi-document summarization are listed in table 1.

TABLE 1
EXISTING SUMMARIZATION SYSTEMS

System ID	Description
Concept freq	A feature set based approach
SumCR	A new subtopic-based extractive approach
MEAD	A centroid-based approach
SNMF	A clustering-based approach
HybHSum	A probability model-based approach
DrS-G, DrS-Q	A graph-based approach
HIERSUM	A hierarchical LAD-style model based approach
Human-letter	The worst human performance provided by DUC
System-number	Top 3 automatic systems in DUC competition
Baseline	The baseline system used in DUC

III. THE PROPOSED METHOD

We propose a multi-document summarization method that generates a generic summary for the multiple documents given as input. This method will works as follows:

First, each of the documents is segmented into sentences to get a set of sentences where *N* is the total number of sentences from *n* documents. After that, the similarity matrix $S_{N\times N}$recording similarities between every pair of sentences is calculated or estimated with any proper similarity measure techniques. Once the similarity matrix is obtained then fuzzy clustering method is used to group the sentence to form clusters. Then, the system proposes a new measuring attribute called concept frequency. The main purpose of this feature is that it is used to remove the redundant sentences occur in the cluster where a sequence of word concept occurred frequently in two or more sentences. At the same time, Prototype weight of each sentence with respect to a cluster is also produced at the end of the clustering process. The *Exemplar* weight of a sentence considers both the closeness to sentences in the same cluster and the prototype weight of this sentence in the cluster it belongs to.

CONCEPT FREQUENCY

First, calculate the sentence similarity score Sm_i and then the sentences are divided into groups called clusters using fuzzy clustering approach. After cluster formation, the concept frequency feature is applied on each cluster to remove the redundant sentences by matching the sequence of words in each sentence when comparing to other sentence. Concept frequency should be compared for every two sentences in the cluster. If the frequency of words will be more, the other sentence will be removed from the cluster for decreasing the redundancy in each cluster.

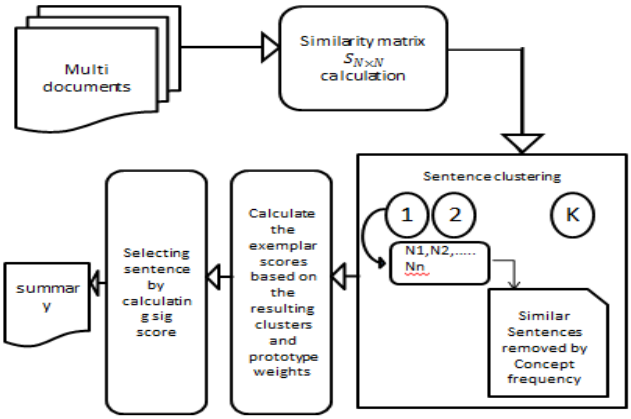

Fig.1 Architecture of the proposed Approach

In a cluster, s_1, $s_2 \ldots s_n$ are the sentences and n_1, $n_2 \ldots n_n$ are the no. of words in each sentence respectively. Consider two sentences s_1 and s_2 with no. of words n_1 and n_2.If the no. of words in a concept that matched between the sentences is greater than or equal to the value of C, then one of the sentences with more no. of words is removed from the cluster. C value is calculated by

$$C=\frac{min(n1,n2)}{2}$$

Where,

n_1 is no. of words in s_1,
n_2 is no. of words in s_2.

Fuzzy clustering with prototype weights: PFC

Given the similarity matrix $S_{N\times N}$ with each element $S_{ij} \in S$ denoting the similarity between sentence x_i and x_j, the objective of PFC is to maximize the following criterion:

$$J_{PFC} = \sum_{c=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N} u_{ci}v_{cj}s_{ij} - \frac{T_u}{2}\sum_{c=1}^{K}\sum_{i=1}^{N} u_{ci}^2 - \frac{T_v}{2}\sum_{c=1}^{K}\sum_{i=1}^{N} v_{cj}^2$$

Where K is the number of clusters u_{ci} is the fuzzy membership representing how well x_i belongs to cluster c, and v_{cj} is the prototype weight which reflects how much x_j is weighed as a representative in cluster c. It can be seen that the objective function of PFC consists of three terms. The first term measures the total compactness of all clusters and controls the main direction of the clustering process; while the last two terms are the regularization of u and v used to prevent u and v from singular values. Parameters T_u and T_v control the trade off between the main term and the regularization terms. Goal is to find all $u_{ci} \in U$ and $v_{cj} \in V$ to maximize J_{PFC}. Updating rules of u_{ci} and v_{cj} are derived as below with the Lagrange multiplier method.

$$u_{ci} = \frac{1}{K} + \frac{1}{T_u}\left[\sum_{j=1}^{N} v_{cj}S_{ij} - \frac{1}{K}\sum_{f=1}^{K}\sum_{j=1}^{N} v_{fj}\,s_{ij}\right]$$

$$v_{cj} = \frac{1}{N} + \frac{1}{T_v}\left[\sum_{i=1}^{N} u_{cj}S_{ij} - \frac{1}{N}\sum_{q=1}^{N}\sum_{i=1}^{N} u_{ci}\,s_{iq}\right]$$

Exemplar: a subtopic-based feature

Based on the fuzzy membership $u_{ci} \in u_{K*N}$ and prototype weights $v_{cj} \in v_{K*N}$ produced by PFC, two possible ways might be used to calculate the *Exemplar* weight, the one based on truncated clusters and the one based on original fuzzy clusters. And we obtain the prototype weight of this sentence in the associated cluster $p_{wi} = v_{Ki}$. After that, we calculate the sum of similarities between this sentence to all other sentences in the same cluster,

$$S_{mi} = \sum_{j\neq i, j\in C_k} s_{ij}$$

Where sentence j is in the same cluster C_k as sentence i. For a sentence i, both S_{mi} and P_{wi} reflect how central this sentence is in the related cluster. We hybridize these two measures to get *Exemplar* as below

$$Exemplar_i = (1-\alpha) \times P_{wi} + \alpha \times S_{mi}$$

Where α is the hybrid parameter. To make effective use of both P_w and S_m, α is set in (0, 1).

Score for generic summarization

After both scores of *Exemplar* and *Position* being calculated, we combine them to obtain the overall *Significance (Sig)* score of each sentence as

$$Sig_i = w_e \times Exemplar_i + w_p$$

Selection criterion

Sort the sentences in descending order by sig score computed. Add the first sentence in the list to the summary. Repeat until the summary length does not exceeds the required limit.

IV. PERFORMANCE EVALUATION

Here we evaluate the proposed approach compared with existing summarization system.

For evaluation of our summarization by concept frequency approach we make use of below method. Our approach for evaluating summary will be done by generating a reference summary, produced by Microsoft summarizer. The output of our summarizing system is then compared with the reference Microsoft summary. The summaries are compared by counting overlapping units such as 1-gram, 2-gram and we accomplish best results. The recall value is calculated by using the following formula [1]:

$$Recall-n = \frac{\sum_{S\in(RefSum)}\sum_{n-gram\in S} Count_{match}(n-gram)}{\sum_{S\in(RefSum)}\sum_{n-gram\in S} Count(n-gram)}$$

Where,

n is the length of the n-gram,

$Count_{match}$ (n-gram) is the no. of n-grams matched with the reference summary.

The proposed system uses concept frequency to remove the redundant sentences by matching the sequence of words in each sentence with other sentences. The comparison with the existing approach that uses only the similarity values and it is shown in Table 2 and graphically represented in Fig. 2.

TABLE 2
COMPARISON OF RESULTS

n-gram	Proposed system	Existing system
1-gram	0.1846	0.1465
2-gram	0.0995	0.0965

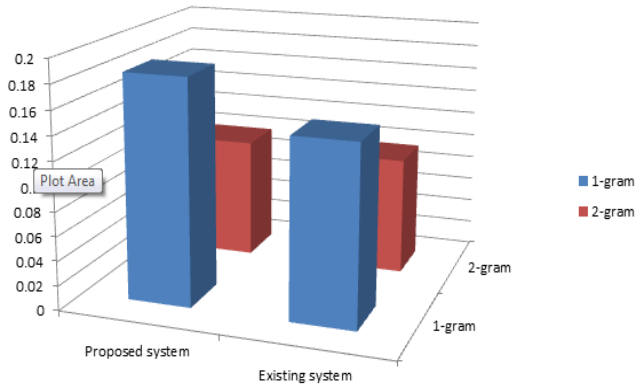

Fig.2 Results comparison

V. CONCLUSION

A summary is a shorter version of the original. The purpose is to help the audience get the gist in a short period of time. Automatic summarization involves reducing a text document or a larger corpus of multiple documents into a short set of words or paragraph that conveys the main meaning of the text. In this paper, we propose a new multi-document summarization approach that makes use of a feature called 'concept frequency' that verifies sentences nearness with a cluster opted. Hence the summarization relevance would be more effective.

REFERENCES

[1] Jian-Ping Mei · Lihui Chen, "SumCR: A new subtopic-based extractive approach for text summarization", Springer, Knowledge inf Systems, 06, August 2011.

[2] Celikyilmaz A, Hakkani-Tur D (2010) A hybrid hierarchical model for multi-document summarization. In: Proceedings of the 48th annual meeting of the association for computational linguistics (ACL 2010), pp 1149–1154

[3] ErkanG,RadevDR(2004) LexPageRank: prestige inmulti-document text summarization. In: Proceedings of empirical methods in natural language (EMNLP 2004), pp 365–371

[4] Haghighi A, Vanderwende L (2009) Exploring content models for multi-document summarization.In: Proceedings of human language technologies: the 2009 annual conference of the North American chapter of the association for computational linguistics (NAACL'09), pp 362–370

[5] Long C, Huang M, Zhu X, Li M (2009) Multi-document summarization by information distance. In: Proceedings of the 2009 Ninth IEEE international conference on data mining (ICDM'09), pp 866–871

[6] Aliguliyev RM (2006) A novel partitioning-based clustering method and generic document summarization.In: Proceedings of the 2006 IEEE/WIC/ACM international conference on web intelligence andintelligent agent technology, pp 626–629

[7] Mei J-P, Chen L (2010) Fuzzy clustering with weighted medoids for relational data. Pattern Recognit43(5):1964–1974

[8] Neto JL, Santos AD, Kaestner CA, Freitas, AA (2000) Document clustering and text summarization.In: Proceedings of the 4th international conference on practical applications of knowledge discovery anddata ming (PAKDD'00), pp 41–55

[9] Aggarwal CC, Yu PS (2010) On clustering massive text and categorical data streams. Knowl Inf Syst24(2):171–196

