ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering
(IJERCSE)
Vol 3, Issue 11, November 2016

Concept Based Sentence Modeling for Extractive Speech Summarization

Guru keerthana Gaddam,
M.Tech student , Department of Computer Science and Engineering
Jawaharlal Nehru Technological University, Anantapur India

***Abstract:* -- With large volumes of multimedia data and speech recordings available over internet,there is need to efficiently process these data so that users can quickly review important information. So the research is mainly focused on automatic processing of transcripts.In past decades,lot of methods have been proposed for summarization of text. The solution for speech summarization is to transliterate the spoken documents to texts, and apply some well-defined text summarization methods. When these methodologies are applied to spoken documents, they doesn't work good for text processing.**

Shih-Hung Liu et al. performed speech summarization by combiningClarity Measure andRelevance Language Modelling(RLM). Clarity measure is used for important sentence selection, which helps to identify the individual sentences which reflect the main theme of the document. The experimental evidence of this model indicated that the various formulations instantiated are better than few existing methods for extractive speech summarization. The sentence-level clarity measure in combination with RLM indeed benefits speech summarization significantly.

The limitations observed in this model are that, it is purely term based and doesn't consider concept relevance. So this project aims at proposing the use ofCorpus or Knowledge base for Extractive speech summarization,where a subset of sentences will be selected to cover as many important concepts as possible.

***Keywords:*-- Speech summarization, Corpus, Term frequency, Similarity score.**

I. INTRODUCTION

Speech is the most common and effective method of communication between human beings, but it is not easy to quickly assess, retrieve and use speech documents if they are simply recorded. So for effective utilization of these data it should be summarized which consists of following problems.1.Spontaneous speech is very different from written text due to restricted knowledge. 2.Term sensitivity occurs as speech documents have constrained exposure of vocabulary i.e, contains mistakes as it is spoken by humans and restricted knowledge about particular domain of speech. A solution for this is to transliteraterecordings to texts, and apply some text summarization approaches. Summary can be generated in either abstractive orextractive form. As abstractive method needs more sophisticates techniques like semantic representation and natural language generation, the research is constrained to Extractive speech summarization [1]. However, usually when the traditional text summarization methods are directly applied to speech transcript, the performance is not as good as for text processing.

The previous work include different methodologies developed depending upon several statistical features such as the word frequency, likelihood measure, line position and sentence ranking [2]. The main methods consisting of these features are vector space methods (VSM),Latent sematic Analysis (LSA) methods[3], the Markov random walk methods (MRW)[4] and Maximum marginal relevance(MMR) methods[5]. The limitation observed is word frequency which is calculated based on occurrence of words (no of times word appears), some terms which are least used but more relevant to domain are missed. As a result the performance is greatly affected.

The remaining paper is formulated as follows. Section II discusses the previous research. Section III presents the enhanced work in detail. Section IV presents results while section V concludes the paper.

II. RELATED WORK

The techniques for automatic speech summarization and evaluation results for summarizing spontaneous speech [7] or presentations were first proposed by SadaokiFuruiet.al. To represent summarization results text or speech is used. For speech-to-text summarization, a two-stage automatic speech summarization method was proposed.It consistsof sentence extraction and compaction. Before sentence compaction(grouping similar sentences) the sentences which consists of recognition errors and less importance are automatically removed. The combination of sentence compaction and extraction is effective and achieves better performance at 70% and 50% summarization ratios when compared with previous one-stage methods. Sentence extraction includes three scores, the linguistic score, the word significance score and the word confidence score, which are effective for extracting importance sentences. Theratios of sentence extraction and sentence compaction mainlydepends on the summarization ratio and features of presentation utterances. To present the summaries by speech-to-text summarization, three kinds of units are considered. They are sentences, phrases and filler units which are to be extracted from speech and concerted to generate the summaries. Similar measures are used for finding extracted units which are combined to produce the summaries. In order to elude acoustic incoherence,amplitudes of speech waveforms at the margins are slowly weakened and pauses are interleaved before concatenation. Theevaluation ofresultsover three scores for the summarization ratio of 50% indicated that between-filler units are expected to achieve good performance.when the summarization ratio becomes smaller it gains additional benefit.

Berlin Chen et al projected a risk-aware modelling framework [1],which is used to select summary sentences list wiseincludingaggregation of either a generative modelling pattern or a direct-modellingpattern. This process includes integratingseveral existing summarization procedures into the enhanced framework. The experimentalresults witness consequentialraise in performance of the summarization methods.To achieve best results for either the manual or spoken documents, list-wise selection strategy in conjunction with the generative modelling hypothesis was used. Some other possible futureextensionsto implement the list wise selection strategy more effectively are listed: 1) explore extra information cues and sophisticated modelling paradigms 2) probing different training criteria for preparing the constituent models of this framework; 3) ranging and applying the proposed framework to multidocument summarization tasks.

Xiaojun Wan et al made use of two unprecedent models for incorporating theme clusters in document.The first model assimilate the clusters information in the Conditional Markov Random Walk Model[3] and the second model uses the HITS algorithm in which clusters are considered as hubs and sets are extracted from sentence clustering.As extension work finding theme clusters which are meaningful using different theme detection methods can be considered. To include the cluster-level informationlink analysis methods are used.

Research work on spoken document summarization in unobstructed domains focused on Broadcast News[8] and text of voice mail speech [8].This work uses large units, like sentences or speaker turn, as basic units for summarization.

The evaluation results are performed based onspontaneous utterances in the Spontaneous Speech Corpus and Processing Project [7]. This project is initiated by building a large spontaneous speech corpus.The corpus consists of coarsely 7 M words with 700h speech length. It mainly records epilogues such as lectures, presentation and news annotations. The recording with low impulsiveness, such as those from read text, isexcepted from the corpus. The utterances are automaticallytransliterated and some of them are tagged manually and used for morphological analysis and part-of-speech (POS) tagging.

III ENHANCED SPEECH SUMMARIZATION FRAMEWORK

Concept based sentence modelling performs summarization in 2 levels

(i)Corpus level
(ii)Document level

***(i)Corpus level**:*

The corpus is loaded with standard documents related to domain on which we are going to extract summary.Later on it undergoes following calculations:

- Pre-processing and clustering
- Word frequency
- Similarity score
- Likelihood measure
- Position weights

1. Pre-processing and clustering:

A preliminary processing of data in order to prepare it for further analysis.It involves tokenization,stemming,stop word removal.Clustering is a process of identifying interesting patterns within a document or group of documents.This whole process is done using OpenNLP toolkit which consists of advanced text processing services.

***2.Word frequency**: (weight factor)*

Frequency(No. of occurrences) of a word in a document with respect to overall terms. TF(w)=(No. of times a term 'w' appears in a document) / (Total no. of terms in the document) It is a statistical feature that reflects how important a word is to a document in corpus.

3.Similarity score:

Real valued function that quantifies similar objects. Cosine similarity is one of the mostly used similarity measure .It represents each and every phrase asa vector.

$$\text{Sim } (D_1,D_2)= \frac{\Sigma_i t_{1i} t_{2i}}{\sqrt{\Sigma_i t_{1i}^2} \times \sqrt{\Sigma_i t_{2i}^2}}$$

Wheret_i is the term weight.

***4.Likelihood measure*:**

It is a probabilistic measure used for sentence ranking.Probability here is calculated using Bayes' rule:

$$P(S/D) = \frac{P(D/S)\,P(S)}{P(D)}$$

Where $P(D/S)$ is probability with respect to S,i.e likelihood of D being generated by S. P(S) ,P(D) is prior probability of S,D.

$$P(D/S) \approx \pi_{l=1}^{L} P(w_i/S)$$

Where L denotes length of the document. $P(w_i/S)$is frequency of each distinct word w occurring in the sentence.

$$P(W/S) = \frac{C(W,S)}{S}$$

Where c is count.

5.Position weights:

It determines position of a sentence where to be placed in the summary as it should be concise and should consider some order.

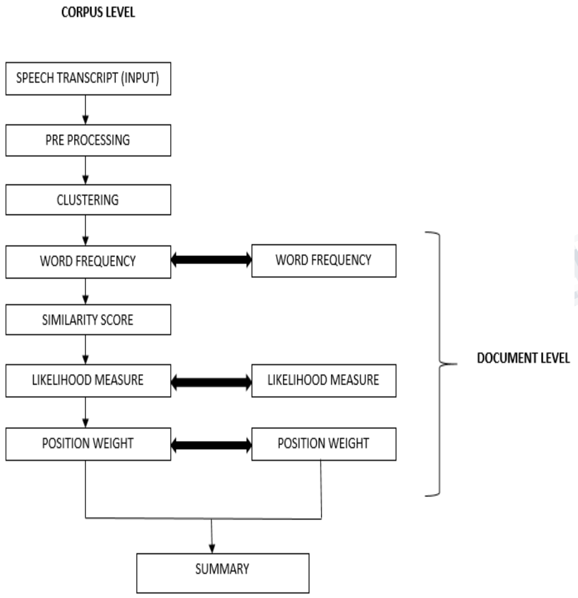

Fig 1: Enhanced speech summarization framework.

(ii) Document level:

The calculations here include

- Term frequency
- position weight

The whole process is first performed on corpus and then with respect to document.. Even if the term which reflects the main theme is given less frequency in document but more in corpus it will be retained eliminating nonfunctionalcontent.As position weights are also compared readability increases. This increases the summarization performance.

IV RESULTS AND DISCUSSION

A.Dataset

The dataset is prepared from the speech samples, which are collected from the previous articles.The total number of speech samples considered were 180.These speech samples are associated with manual transcript summaries that are used for cross validation.

Summary by % of sentences	Accuracy (%)
2	97
5	97
7	97
10	97
15	96
20	96
25	95

Table 1:The Summarization Accuracy observed from the experiments for different summary ratios

B.Performance Analysis

The results obtained from experiments were visualized as performance graphs and tables.The accuracy of the proposal under different ratios of summary size were explored in table1.The summarization accuracyis found to be stable for the ratio of summary between 2 to 25%

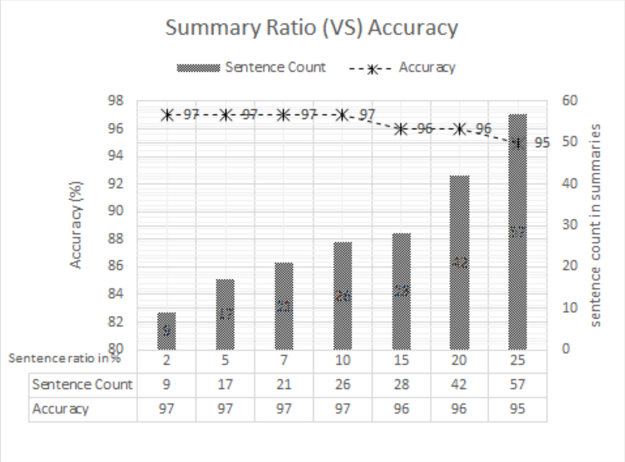

Figure 1:The Accuracy for different summary ratios.

The resource utilization of proposed model is optimal since the completion time is found to be linear (see fig 2) and also the ratio of memory usage is linear for the given distinct speech documents.

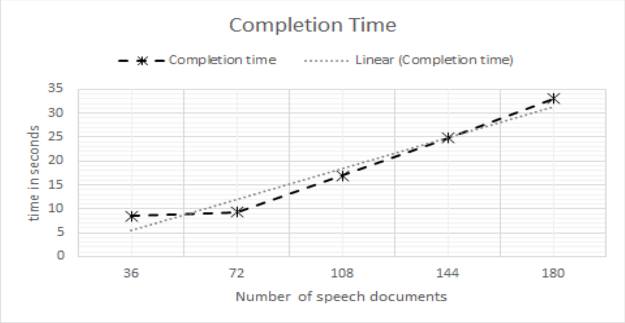

Fig 2:Process completion time vs linearity

Theproposed speech summarization technique is justified as optimal as the prediction accuracy is stabilized against the demand of divergent summary ratios and the prediction accuracy is above 95%, which is substantially good that compared to existing models.

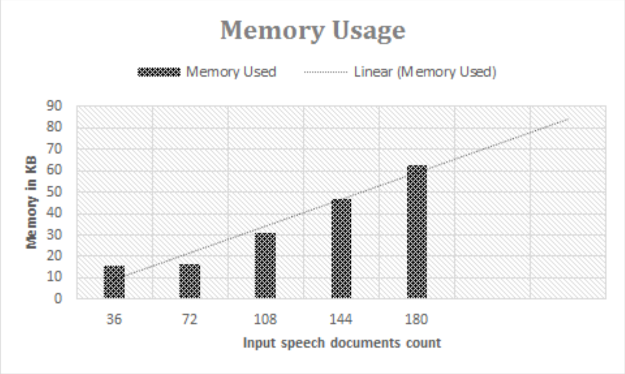

Fig 9:Memory usage vs linearity

V. CONCLUSION

This paper performs extractive speech summarization by introducing a corpus or knowledgebase.All the statistical features are compared with respect to both document and corpus .As the corpus consists of all important concepts regarding the domain of summary,it retains as many important concepts as possible thereby increasing performance of summarization.As to future work, there is a chance of investigating abstractive summarization which reflects the main theme of document. But it requires some difficult implementations of NLP and NLG techniques along with semantics.

REFERENCES

[1] J.J Zhang, R.H.Y. Chan and P. Fung, "Extractive speech summarization using shallow rhetorical structure modeling,"IEEETrans.Audio,Speech,Lang.Process.,vol.18,no.6,pp.1147-1157,Aug.2010

[2] B. Chen, H.C. Chang, and K.Y. Chen, "Sentence modeling for extractive speech summarization, " in Int. Conf. 2013,pp. 1-6.

[3] X. Wan and J. Yang, "Multi-document summarization using clusterbasedlink analysis," in Proc. Int. ACM SIGIR Conf. Res. Develop.Inf. Retrieval, 2008, pp. 299–306.

[4] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1998,pp. 335–336.

[5] B. Chen and S.-H. Lin, "A risk-aware modeling framework for speechsummarization," IEEE Trans. Audio, Speech, Lang. Process., vol. 20,no. 1, pp. 199–210, Jan. 2012.

[6] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-text andspeech-to-speech summarization of spontaneous speech," IEEE Trans.Speech and Audio Process., vol. 12, no. 4, pp. 401–408, Jul. 2004.

[7] Corpus based web document Summarization using Statistical and Linguistic Approach

[8] Shih-hung liu,Berlinchen,"Combining Relevance language modeling for extractive Speech summarization"IEEE/ACMTrans.Speech and language processing, vol. 23, no. 6, pp. , June. 2015