A Study on Relevance Measure with Compression Ratio for Text Summarization

P.Vijaya Pal Reddy

Matrusri Engineering College, Hyderabad

Abstract:

Text summarization (TS) is a process condensing the original text into shorter form by retaining the actual content of the text. This paper addresses the process of generating the summary for a given Telugu text document using Maximal Marginal Relevance (MMR) technique. This paper also addresses the influence of the relevance and novelty measure 'λ' on the summary length of the text. The experimental evaluations has performed using F1 score.

Keywords- text summarization, maximal marginal relevance, F1 score, cosine similarity, vector space model.

I. Introduction

Huge amount of information is disseminating into the web day-by-day through varies information services. Due to disparity of the information in the web, retrieving the desired information is a serious problem. Text summarization is a useful tool to efficiently find useful information from immense amount of information according to the user needs. Summarization is the process of condensing a source text into a shorter version while preserving its information content [1]. In general, summarization process divides the original text document into set of sentences and then assign score to each sentence according to some criteria. Then the sentences are extracted based on their scores to form a summary. The number of sentences to be selected is based on the level of compression need to be done on the original text [2]. There are two kinds of summaries namely extractive summary and abstractive summary. Extract summary is a selection of set of sentences from the original text, where as an abstractive summary is a reformulated form of the original document [3].

In this paper we employed extractive summary generation using MMR criteria. Most of the research has done in extractive summarization methods [4,5,6]. Initially, text summarization process has been studied based on frequent words represent in [7]. First paragraph or first sentences of each paragraph contain topic information proposed in [8]. Query-based summarization is studied in[9]. Maximal Marginal Relevance technique is presented in[10] which is followed in our paper for Telugu text single document summarization. Two-step sentence-extraction method for single-document summarization and multi-document summarization is proposed in [11]. TS using Lexical Chain and WordNet proposed in [12]. The nuclei of the discourse structure tree for a text determine salience of information as in [13]. The rest of this paper is organized as follows. Section 2 explains about Maximal Marginal Relevance (MMR) technique and the way how it can be used for text summarization Section 3 describes the proposed model for TS using MMR. Results and discussion are reported in section 4 which discusses about the influence relevance and novelty measure in the process of text summarization using F1 measure. The section 5 gives the conclusions from the results.

2. Maximal Marginal Relevance

When user poses a query onto the search engine, a list of documents are retrieved based on the relevance to user’s query. Marginal relevance is a technique which provides a linear combination of novelty and relevance to the user's query. According to the marginal relevance, if a document is more relevant to the query posed on to the search engine and having maximum dissimilarity to the previously selected documents to the user query then the selected document is said to have maximum marginal relevance. In the process of information retrieval, getting more relevant documents for the user's query having high marginal relevance similarly in the text summarization, getting the most relevant paragraphs or sentences from the document having maximum marginal relevance is a issue of research. We strive to maximize-marginal relevance in retrieval and summarization, labeled as maximal marginal relevance (MMR) method.

$$MMR = Argmax_{D_i E R-S}[\lambda(Similarity_1(D_i,Q) - (1-\lambda) max_{D_j E S} Similarity_2(D_i,D_j))]$$

Where Q is a query posed onto the search engine. R is the list of documents retrieved by an information retrieval system. S is the subset of documents already selected from R, R-S is the subset of documents which are not yet selected from R. Sim 1 is the similarity measure to measure the similarity between the document from R-S and a query and Sim 2 is the similarity measure between the document not yet selected (R-S) and the documents which are already selected by the information retrieval system. MMR computes the relevant set of documents according to the ranking assigned to the documents when the λ becomes one and calculates maximal diversity ranking among the documents when λ becomes zero. For values between 0 and 1 of λ, a linear combination of both criteria is optimized. If information space need to be around the query then λ should be set at smaller values where as if the focus is on reinforcing relevant documents then λ should be closer to 1.

3. Proposed Summarization Model

Proposed Text Summarization model has four phases as shown in Figure 1. Initially the given text document is preprocessed and represented in the form of vector space model. Weights are assigned to the features in the vector space by using the TF-IDF measure. Adopting the maximal marginal relevance (MMR) technique to the problem of text summarization is by viewing set of sentences in a document as a set of documents in the corpus and selecting a sentence with linear combination of relevance with the total document and dissimilar with the already selected set of sentences in the summary. The relevance measure between sentence and document and sentence with already selected sentences is performed using cosine similarity measure. Then the selected set of sentences are arranged based on the scores assigned using MMR in the score decreasing order. The number of sentences to be in the summary is chosen based on λ value which is evaluated using the F1 measure. The different phases of the model are briefly explained below:

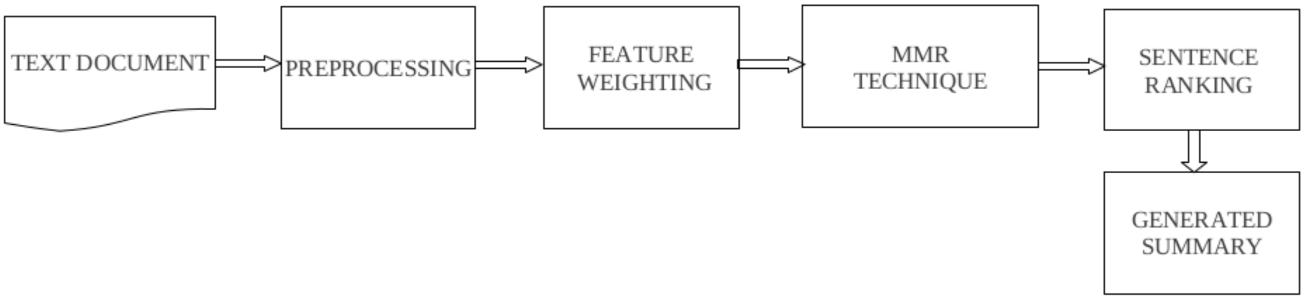

Figure 1: Summarization Model Using MMR Technique

A. Preprocessing

Document need to be preprocessed before processing through the machine. The pre-processing contains removing the unnecessary content from the document which is not useful for TC like punctuation marks, numbers, dates and symbols etc. Secondly, features which can create noise to the TC process called stop words which are used to give meaning to the sentence need to eliminate. For Telugu text stop words are identified by using the NLTK tool kit. As Telugu is complex morphological variant language, reducing the features of document into their root form can greatly reduce the dimensionality space of the document. Hence features of the document are converted into their root form using TMA tool. After pre-processing the feature space of the document contains only stemmed form of the features.

B. Feature Weighting

We followed the Term Frequency-Inverse Document Frequency (TF-IDF) approach for feature weighing which is proved to be more prominent in the TC literature. The product of TF and IDF is given as:

$$W(d,t)=TF(d,t).IDF(t)$$

where TF(d,t) is the number of times the term t occurs in the document d and

$$IDF(t)=\log\left(\frac{N}{n}\right)$$

where N is the total number of documents in the training dataset and n is the number of documents that
contain the term t.

C. MMR Technique

In this paper, we used Maximal Marginal Relevance (MMR) technique to generate summary of a Telugu text document. In view of MMR technique, summarization is a process of ranking the sentences by its relevance to the overall document and also reducing the redundancy among the selected sentences. The summarization process was as follows:

Step1: Generally in news articles the first sentence is considered as a most content sentence. Hence pick up the first sentence of the article and add it into summary S.
Step2: Calculate Marginal Relevance for each sentence in D-S where D is the document to be summarized
and D-S is the set of sentences in D which are not yet included into S. The definition of Marginal Relevance for sentence s i is as follows:

$$MMR(S_i)=\lambda Similarity(S_i,D)-(1-\lambda)max_{S_i E S} Similarity(S_i,S_j) \quad \text{where } 0 \le \lambda \le 1$$

Sim(s i ,D) is the similarity between s i and the document D, which gives the representativeness of the sentence s i for the document D. Sim(s i , s j) is the similarity between s i and s j where s j is any sentence from summary S. This measure infers the repetitiveness of sentence s i with respect to the sentences which are already included in the summary S. To find the similarity between sentence and document and similarity between sentences, we used the similarity cosine similarity measure.

Step3: Pick up the sentence with maximal Marginal Relevance value and add it into summary set S.
Step4: Repeat step2 and step3 until expected summary length is reached. In this experiment, summary length is set to 10%, 20%, 30%, 40% and 50% of the original article. Then we experimented with λ value which actually indicates the level of redundancy contained in the final summary. In this experiment, we generated summary for the article based on λ values of 0, 0.3, 0.7 with different compression ratios and evaluated the accuracy of the summary using F1 measure.

D. Sentence Ranking

Based on the MMR scores of the sentences, sentences are arranged in the descending order of the scores. The number of sentences to be included in the summary of the document is decided based on the relevance and novelty measure λ and the value F1 measure

4. Results And Discussions

A. Evaluation Measures

There are different methods to evaluate the performance of a text summarization system. In this study, we have chosen intrinsic evaluation. Intrinsic evaluation judges the quality of a machine generated summary based on the correspondence between the generated summary and the human generated summary. We have used F1 measure which is based on precision (P) and recall (R) to judge the coverage between manual and machine generated summaries. Assume T is a manual summary and S is a machine generated summary, the precision can be defined as:

$$\text{Precision (P)} = |S \cap T| \ / \ |S|$$

recall can be defined as:

$$\text{Recall (R)} = \quad |S \cap T| \;\; / \;\; |T|$$

and F1 measure is defined as:

$$F_1 = \frac{2 * P * R}{P + R}$$

Compression ratio λ	10%	20%	30%	40%	50%
0.0	0.658	0.693	0.731	0.719	0.682
0.3	0.759	0.771	0.784	0.762	0.748
0.7	0.663	0.679	0.697	0.680	0.665
1.0	0.657	0.670	0.682	0.667	0.643

Table 1: F1 measure for λ Vs. compression ratio on the original document

In this experiment, summary length is set to different ratios i.e 10%, 20%, 30%, 40% and 50% of the original article. As the summary length increases upto 30% and the λ increases upto 0.3 the accuracy of the generated summary is increasing. From 30% to 50% and from 0.3 to 1.0 of the λ value the F1 value of the generated summary is decreasing. Under this compression ratio of 30% and the λ value of 0.3, summaries for the most articles contain 5 or more sentences. Experiments showed that summaries with compression ratio of 30% achieved better F1 performance than those with higher compression ratios and then those of very short summaries. Therefore, compression ratio was fixed at 30% and we tuned λ value to 0.3 to get most appropriate summary of the document.

5. Conclusions

In this paper, we have shown that MMR technique which provides information to the user by allowing the user to minimize redundancy in the summary of a document. A single document news summarization has been presented that picks up sentences which is aligned closely with the user ranking. λ measure which is a linear combination of relevance of the summary sentences with the original document and the dissimilarity among the sentences with in the summary is observed with different summary lengths. For the experimental analysis it is observe that the suitable values are suggested both for λ and summary length to optimize the performance of the summarization system. The λ value at 0.3 and at the compression ratio of 30% of the original document provides most suitable summary of the document with the F1 measure of 0.784. The Work is in progress to extend the scope of the automatic summarization to multiple documents.

References

[1] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," in Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, 1997, pp. 10-17.

[2] D. Marcu, "Discourse trees are good indicators of importance in text," Advances in Automatic Text Summarization, 1999, pp. 123- 136.

[3] E. Hovy and C. Y. Lin, "Automated text summarization in SUMMARIST," Advances in Automatic Text Summarization, 1999, pp. 81- 94.

[4] Kim, J., Kim, J., Hwang, D., 2001. Korean text summarization using an Aggregation Similarity. In: Proc. 5th Internat. Workshop Information Retrieval with Asian Languages, pp. 111–118.

[5] Nomoto, T., Matsumoto, Y., 2001. A new approach to unsupervised text summarization. In: Proc. ACM SIGIR'01, pp. 26–34.

[6] Wang, J.C., Wu, G.S., Zhou, Y.Y., Zhang, F.Y., 2003. Research on automatic summarization of web document guided by discourse. J. Comput. Res. Develop. 40 (3), 398–405.

[7] Luhn, H.P., 1959. The automatic creation of literature abstracts. IBM J. Res. Develop., 159–165.

[8] Edmundson, H.P., 1968. New methods in automatic extraction. J. ACM 16 (2), 264–285.

[9] Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J., 1999. Summa rizing text documents: Sentence selection and evaluation metrics.In: Proc. ACM-SIGIR'99, pp. 121–128.

[10] Carbonell, J., Goldstein, J., 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proc. 21th ACM SIGIR Internat. Conf. on Research and Develop ment in Information Retrieval.

[11] Jung, W., Ko, Y., Seo, J., 2005. In: Automatic Text Summarization Using Two-step Sentence Extraction, LNCS, vol. 3411, pp. 71–81.

[12] Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., 1990. Introduction to WordNet: An on-line lexical database (special issue). Internat. J. Lexicogr. 3 (4), 234–245.

[13] Marcu, D., 1996. Building up rhetorical structure trees. In: Proc. 13Th National Conf. on Artificial Intelligence, vol. 2, pp. 1069–1074.