An Approach for Selecting Text from e-mail in Satellite-based Education

Raghuveer.K[a], Dr. G. Raghavendra Rao[b], Dr. G. L. Shekar[c], Dayanand. R[d]

[a]Research Scholar, Department of Information Science, National Institute of Engineering, Mysore-570 008, India.
Tel : 0821-2482317, Fax : 0821-2485802, E-mail : raghunie@yahoo.com,

[b]Principal, National Institute of Engineering, Mysore-570 008, India
Tel : ,0821-2480475, Fax : 0821-2485802, E-mail : principal@nie.ac.in

[c]Professor & Special Officer, e-Learning Center, Visvesvaraya Technological University, Mysore-570 006, India
Tel : 0821-248220, Fax : 0821-2485802, Email : gl_shekar@yahoo.co.in

[d]Senior Software Engineer, GE, Bangalore, Tel : 91-09448616468 E-mail : dayanand_7@yahoo.com

ABSTRACT

It was in September 2004 that Indian Space Research Organization (ISRO), in its endeavor to provide sustainable distance education service in India, launched EDUSAT that is designed to supplement the present teaching system in the country. The EDUSAT project has sought to wipe out disparities in quality between the various educational institutions by addressing itself to key issues of access, interactivity, cost-effectiveness and consistency of information to students. One subject expert can simultaneously teach hundreds of students in multiple locations across a vast geographical area. The major issue of paucity of faculty members in higher education has been tackled in one single stroke. In this regard, VTU has set up a Forum (http://forum.vtu.ac.in), which provides a platform for students to exchange their view, share resources and also get answers from subject experts by sending their queries into this Forum by e-mail. As this VTU-EDUSAT network is being used by a whopping 150,000 students, there will be thousands of questions from students across different subjects that will be handled through the network by hundreds of subject experts.

The questions will be aired by the students during the live sessions of the satellite-based programme or by sending e-mail post transmission of sessions. Further, there will be number of questions by the students after they go through the e-Learning content. It has become impossible to answer all these questions by the subject experts because of paucity of time. It is also observed that there will be many questions that are similar or repetitive or these questions would have been answered earlier by the subject experts. Also, it is impossible for the subject expert to go through all the e-mails completely.

Therefore, it is required to develop a solution that involves summarization of e-mail text sent by the students, which means selection of important sentences or text so as to compare them with the current data base of questions. Once comparison is made, only the non-repetitive questions are extracted and place them before the subject expert. Further, if the answer is already available for questions, the system advices students to refer the FAQ data base.

Here, it is necessary to distinguish the important information with less important information within the original text (e-mail). For this, we need the combination of two steps: information filtering and text reduction. The on-going research is to address such an issue by domain-specific text summarization. We have made an attempt to develop a model that examines content-based text selection.

Keywords

Information retrieval, Text mining, Domain-specific summarization

1.0 INTRODUCTION

One of the main of challenges of the Visvesvaraya Technological University (VTU), Belgaum, India, that manages 120 engineering colleges spread across a vast area in the state of Karnataka with a total student strength of a whopping 150,000 students is to address the issues of access, interactivity, cost-effectiveness and consistency of information to students. This challenge is being largely addressed by VTU-EDUSAT network of Distance Education using satellite. One subject expert can simultaneously teach hundreds of students in multiple locations across a vast geographical area. The advantage of this a students who are located at rural places can able to get the best training, material and other information resources. VTU is providing a platform, called Forum, for students to exchange their view, share resources and also get answers from subject experts by sending their queries into the Forum by e-mail. There will be number of questions by the students after they go through the live satellite-based sessions or web-based e-Learning content. It has become impossible to answer all these questions by the subject experts because of paucity of time. It is also observed that there will be many questions that are similar or repetitive or these questions would have been answered earlier by the subject experts. Many times it is impossible for the subject expert to go through all the e-mails completely.

In this paper we have proposed a model that examines content-based text selection. We have proposed three approaches for selection of text from the e-mail received from the students through Forum facility that has been built in VTU-EDUSAT network.

In the first approach, user specifies his interests by marking texts or text passages. From this marked collection, a weighted word list ranking the most important words for the topics of interest is computed. Using the word weights as a user model, sentences from the original text are selected if their weight sum is above a threshold. This amounts to a filter for interesting text passages. In the second approach, we take clue from the title of the e-mail that contains candidate words, which is the most important bit of information. Taking title of the e-mail, we build the related words with the help of WordWeb, rank the sentences depending on the number of candidate words and related words and finally the sentences are extracted. In the third approach, we take clue from the title of the e-mail, where title contains candidate words. Depending on the relative position of the words in sentences it will be ranked and sentences are extracted.

In the following text, Section-2 relates previous work done to our proposed approach. Section-3 describes all the methods, while Section-4 deals with the data sets and experiments carried out with corresponding outputs. The Section-5 gives concluding remarks.

2.0 RELATED WORK

Important sentences extraction is a standard approach for the summary based on characteristic words (Mani & Maybury, 1999). Text summarization consists in compressing a document into a smaller précis. Its goal is to include in that précis the most important facts in the document. Alternatively, it can be done by extracting from a text those elements, usually sentences, best suited for inclusion in a summary (Copeck, Japkowicz, & szpakowicz, 2000). Normally summarization is a two step process: the first step is to build a source text representation from the source. The second step consists of summary generation – forming a summary representation from the source representation built in the first step and synthesizing the output summary text (Barzilay & Elhadad, 1998). In the location and cue phrase approach, location and cue phrases produce better results than the word frequency method and can be accurately computed. Also learning in order to combine several shallow heuristics (cue phrases, location, sentence length, word frequency and title), using a corpus of the research papers with manually produced abstracts (Chuang & Yang, 2000).

Considering domain-specific summarization, words that are characteristic for a given topic, not for a text as a whole, must be found. In the context of information extraction words from the user query indicate topic for specific summarization (Tombros &.Sanderso, 1998).

Some systems considered features such as frequency, sentence location and bonus/stigma words in order to extract sentences to make a summary. Their approaches performed fairly well despite simplicity and have been the basis in the area of automatic text summarization. However, their approaches ignored the structural aspect of the text which may contain significant information for summarization (Luhn, 1958).

3.0 SENTENCE SELECTION APPROACHES

This section describes how word lists can be used to extract from the e-mail text.

3.1. Weight Based Method

Several ways to compute a word ranking are possible. According to (Hovy, 2002), the algorithm for this method is given below.

1) The given context is taken as input for extraction. The sentences that are selected at the initial stage of extraction are termed as relevant sentences; otherwise they are termed as irrelevant sentences.

2) Given collection of r relevant and i irrelevant passages, let f_r and f_i be the absolute frequencies of a word in them, respectively. Then its weight w is computed as

$$w = \frac{r/f_r}{i/f_i}$$

This weight can be interpreted as an approximation to the probability that a text containing this word deals with the relevant topic.

3) The weight for the stop words (those words that appear at a high frequency in the English language), propositions, and conjunctions are not calculated.

4) The relevant sentences are put into a file. The weight of each sentence present in the file is calculated, which is the summation of word weight of all those words in that sentence.

5) For each sentence, the sum of its word weight is divided by its word count so as not to bias for longer sentences.

6) The result is compared to a threshold to decide whether the sentence should be in the domain- specific extract.

$$Threshold = \frac{\min imum + \max imum}{2}$$

(of sentence weight)

3.2. Key Word Based Method

In this approach, we consider the information given in the title of the e-mail or core content raised by the students. We consider the title which is composed of candidate words. With Wordweb we find out for each candidate word, the list of its related words. The algorithm is given below

1. The first sentence in the document is considered as the title.

2. For each sentence, the frequency of keywords and related words is computed.

3. The user is allowed to specify the amount of text that he wants in the summary.

4. The sentences are extracted depending upon the following criteria:

   a) The sentence with the highest number of keywords is extracted.

   b) If any two sentences have equal number of keywords then the one with more number of related words among the two is extracted.

   c) If any two sentences have equal number of keywords and equal number of related words then the sentence that appears first in the discourse are selected and extracted.

3.3 Relative Word Based Method

Here, we consider relative ranking of the words distribution in the title of the e-mail. The algorithm is as

1. The first sentence in the document is considered as the title.

2. We skip the list of preposition, conjunctions so that it is not going to given the weightage for domain.

3. The words that present in the title are ranked with 1 and decreased by a factor of 2 (0.5, 0.25, 0.125 and so on).

4. Rank the sentence according the word distribution.

5. The sentences that score more weightage are extracted depending on the selector choice.

4.0 DATA SET

We selected about 50 e-mails sent by the students to subject experts through Forum. The range of sentences in e-mail is from four sentences to thirty sentences. We applied all the three methods mentioned in Section-3. In weight based method, the final extraction of the sentences used to depend on the number of sentences in the document and number of relevant sentences are marked. We find that if number of relevant sentences were less compared to the irrelevant sentences, the final selection of sentences is to be more precise. Suppose, if relevant sentences are more, then selection of sentences, which are used to contain some more additional information also. In the key word based approach, we find sentences which are used to depend on the distribution of candidate words and related words are selected. This we find it, is more accurate in most of e-mails. In the relative word based approach we find that relative position of the word used to extract a more precise sentence.

5.0 RESULTS

Here, we have taken data set comprised of e-mail text sent by the students to the university forum.This has been set up by the university to accept the doubts raised by the students. This will be checked by the subject matter expert by the logging into the forum. The sample data set has been taken from data communication subject forum. We applied our methods to extract the core contents from the e-mail. The following section shows the documents selected and final selected sentences for a e-mail.

Context Switching

> From the basics I came to know that the traditional telephone systems makes use of circuit switching.internet based cumminication method uses packet switching.
> I am interested in knowing more about the circuit switching technique. I also read many books regarding this technique. I hope that this technique may be used in some good applications which may result in increase of efficiency of these application. Even I tried to gather information about this topic using world wide web. But, I couldnot gather much info from it. Sir can u provide some links to the documents. So that it will clear my doubts. Expecting reply as soon as possible.

Table 1: Statistics from Weight based method

A	B	C

10	5	1

A - Number of Sentences in the original document
B – Number of relevent sentences
C – Number of sentences selected

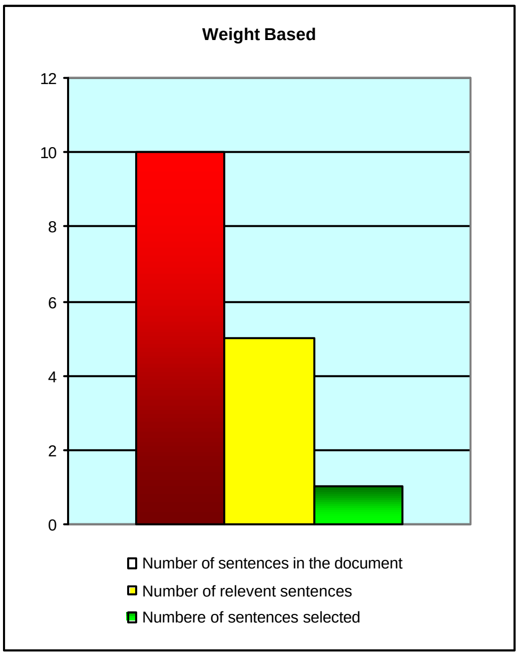

Figure 1: Results obtained from weight based method

Relevant Sentences

Internet based cumminication method uses packet switching.
I am interested in knowing more about the circuit switching technique.Sir can u provide some links to the documents. So that it will clear my doubts. Expecting reply as soon as possible.

I am interested in knowing more about the circuit switching technique.

A - Number of Sentences in the original document

Table 2: Statistics from Keyword based method

A	C
10	2

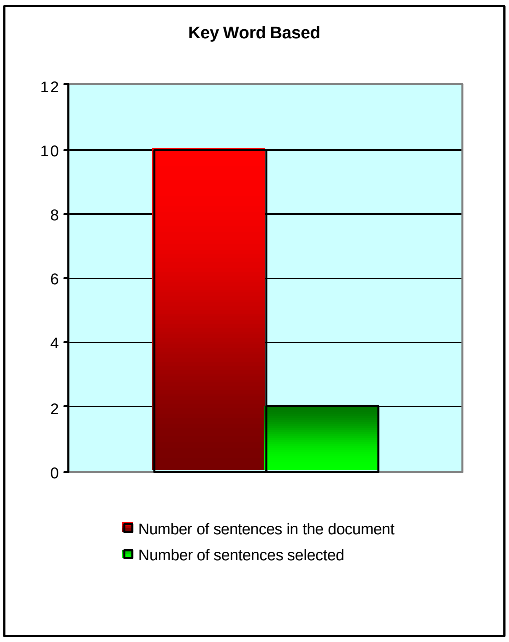

Figure 2: Results Obtained from Keyword based method

Selected Sentences

From the basics I came to know that the traditional telephone systems makes use of circuit switching.
I am interested in knowing more about the circuit switching technique.

Table 3 : Statistics from Relative word method

A	C
10	1

Selected Sentence

I am interested in knowing more about the circuit switching technique.

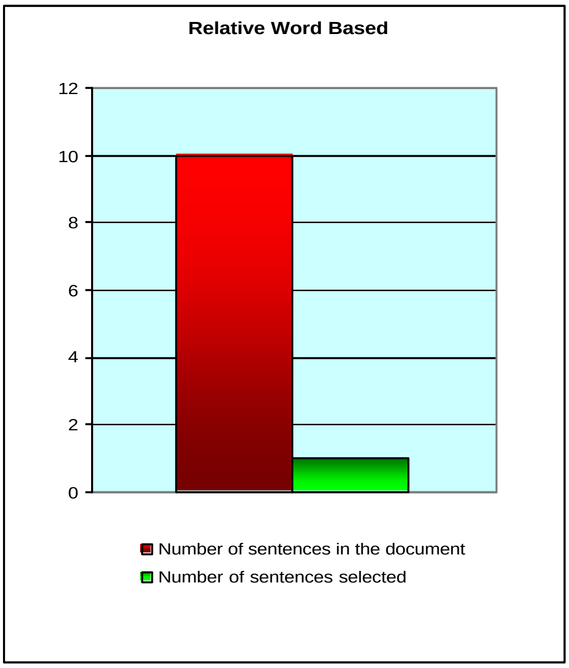

figure 3: Results obtained from Relative word based method

From the above, we conclude the results that are obtained after processing in each of the method gives a result with varying sized length. The conclusion about the result is presented in section 7.

6.0 EVALUATION

As mentioned earlier, core content extractions are easier provided that word distribution in the body of the text is uniform. If title of the text is not going to give more information, then the method may not give the expected result. It meas that important information is often preserved in the messages. It seems better to leave some sentences out of extracts so that it is not going to play any crucial role in the content of the e-mail.

7. 0 CONCLUSION

In this paper, it was shown how ranked list of words related to a topic can be used to extract from the e-mail. No explicit knowledge about semantic relations between words on the list is needed, instead a semantic relation is assumed implicitly between the words with high weight as they are significant for the positively labeled texts.

We would like to extend the work to identify same or similar question raised by the students, so that the subject expert need not see or give answer to such questions that have been addressed earlier. Further, if the answer is already available to question/s, the system advices students to refer the FAQ data base. Finally the quality of results obtained from the automation process will be tested against natural intelligence of a human. The attempt here is to reduce response time and to build question-answer repository.

ACKNOWLEDGEMENTS

The authors would like to thank the support extended by Dr.K.Balaveera Reddy, Vice Chancellor and Dr.M.S.Shivakumar, Registrar of Visvesvaraya Technological University, staff members of VTU e-Learning Center, Indian Space Research Organisation and Department of Information Science & Engineering, National Institute of Engineering, Mysore.

REFERENCES

Barzilay, R. & Elhadad, M. (1998). *Using Lexical Chains for Text Summarization.*

Brunn, M. (2001). *Text summarization using Lexical chains.*

Choi (2000). Advances In Domain Independent Linear Text Segmentation. *In proceedings of 1 North American chapter of the Association for computational Linguistics Seattle*, pp 26-33.

Chuang, T. W. & Yang, J. (2000). *Extracting Sentences for Text Summarization: A machine learning approach.*

Copeck, T., Japkowicz, N., & szpakowicz, S. (2000). *Text Summarization as Controlled search.*

Euler, T. (2002). *Tailoring Text using Topic words: Selection and compression NLIS.*

Hovy, E. (2002). *Automated Text summarization.* DUC.

Luhn, H. (1958). The Automatic Creation Of Literature Abstracts. *IBM Journal Research and Development.*

Mani &.Maybury, M. (1999). editors Advances in Automated Text Summarization, *MIT Press.*

Moen, M. & De Busser, R. (2001). Genreric Topic Segmentation Of Document Text. *Proceedings of 24 the Annual ACM SIGIR conference on Research and Development in information Retrieval , ACM , New York* .pp 418-419.

Morris (2000). Lexical cohesion computed by thesaural relations as an indicator of the structure of text, *Computational Linguistics*, 21-48.

Tombros, A. & Sanderso, M. (1998). Advantages Of Query Biased Summaries In Information Retrieval. ACM SIGIR .