**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**

***IJCSMC, Vol. 3, Issue. 2, February 2014, pg.382 – 388***

**RESEARCH ARTICLE**

# Improvement of Email Summarization Using Statistical Based Method

**Mithak I. Hashem**

Department of Computer Science, Thi-Qar University, Iraq

Mithakh@yahoo.com

***Abstract - Automatic text summarization is undergoing wide research and gaining importance as the availability of online information is increasing. Email is one of the most important online tools that many of us depends on in his everyday life. Finding Email summaries may be crucial for many users. We deal with email text as a single-document in this research. Text summarization can be classified into two approaches: extraction and abstraction. This research focuses on extractive one. The goal of text summarization based on extraction approach is sentence selection. Our proposed method to obtain the suitable sentences is to assign some numerical measure of a sentence (statistically) for the summary called sentence score and then select the best ones to be included within Email summary. The most important step in summarization by extraction is the identification of important features. In our experiment, we used 130 test Email text from Enron_Sent_Mail_Sample data set. Each Email document is prepared by preprocessing process: sentence segmentation, tokenization, removing stop word, and word stemming. Then, we used 7 important features and calculate their score for each sentence. The results show that the best average similarities with the reference summary (gold summary) were obtained by our method.***

***Keywords- email summarization; single-document summarization; sentence feature; sentence score; similarity***

## I. INTRODUCTION

The goal of text summarization is to present the most important information in a shorter version of the original text while keeping its main content and helps the user to quickly understand large volumes of information. Email text summarization addresses both the problem of selecting the most important sections of text and the problem of generating coherent summaries. This process is significantly different from that of human based text summarization since human can capture and relate deep meanings and themes of text documents while automation of such a skill is very difficult to implement. With the ever increasing popularity of emails, email over-load becomes a major problem for many email users [1]. Users spend a lot of time reading, replying and organizing their emails. To help users organize their email folders, many forms of support have been proposed, including spam filtering [2], email classification [3] and email visualization [4]. In this research, we discuss a different form of support email summarization. The goal is to provide a concise, informative summary of email conversation. Email summarization can also be valuable for users reading emails with mobile devices. Given the small screen size of handheld devices, efforts have been made to re-design the user interface. However, providing a concise summary may be just as important.

 

## II. RELATED WORKS

Muresan et al.[5][6] applied machine learning methods for single email summarization with linguistic features. The authors extracted noun phrases from an email as a representation of the content. Rambow [7] and Wang et al. [8] proposed a sentence extraction summarization approach for email threads. They described each sentence in an email conversations by a set of features and used machine learning to classify whether or not a sentence should be included into the summary. Wan [9] and Feng et al. [10] proposed a summarization approach for decision-making email discussions. They extracted the issue and response sentences from an email thread as a summary. Similar to the issue-response relationship, Shrestha [11] and McKeown et al. [12] proposed methods to identify the question-answer pairs from an email thread. Similar results were obtained by Corston-Oliver et al. They studied how to identify "action" sentences in email messages and use those sentences as a summary [13]. All these approaches used the email thread as a coarse representation of the underlying conversation structure.

Lewis and Knowles et al. [14], also described a few methods applied to construct the email threads. Most of those methods use the "header" of emails to construct the email threads, e.g., "Subject", "In-reply-to" and "References". Conceptually, an email conversation is based on the content, not on the header. The header just provides some clues for how an email is related to others. Yeh et al. also found that simply using the header information is less accurate than using the content analysis based on the email body [15]. Other works fall in the larger field of summarization by using NLG means, a discipline that has received significant attention of late Belz [16] and Duboue et al. [17].

## III.THE PROPOSED METHOD

In this section, we will focus on our work on the email text to obtain the proposed email summary. We propose a statistical method to extract the best sentences as a summary candidate based on features scores for each sentence. Therefore, the features score of each sentence that will be described in this section are used to obtain the significant sentences as shown in Fig 1. This method consists of the following main steps:

a) Separation of sentences.
b) Perform Tokenization, the Removal of stop words and Stemming process.
c) The features are calculated to obtain the sentence score base on our proposed method.
d) A set of highest score sentences are extracted as document summary.

 

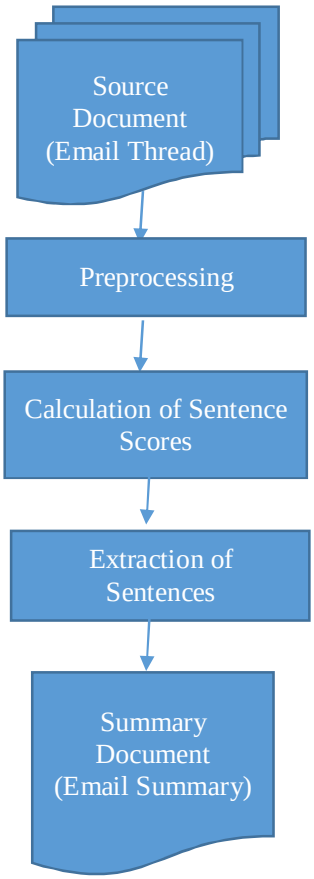


Fig. 1 Email summarization based on statistical method

*A. Separation of Sentences*

Our proposed method deals with email as a plain text (it deals with the text of an email only); i.e. the treatment of graphics, sounds files and other data on a Web page is beyond the scope of our current work. If we are to extract key sentences from a document it is imperative that we first separate them. The separation process is simple. We scan through the entire document and search for characters and punctuations that signify the end of a sentence, for e.g. ".", "-", ";", ":", etc. this breaks the entire document into a list of sentences so to say. It becomes easier to score sentence individually rather than the complete document.

*B. Preprocessing*

There are three main activities performed in this step: Tokenization, Removing Stop Word, and Word Stemming. Tokenization is separating the input document into individual words. Next, Removing Stop Words. To effectively score the sentences we need to only consider the words in the document which are contributing to the value of the sentence. All the unimportant words will have to be eliminated. Such words are called "stop words" in any language. Words such as "a", "an", "and", "as", "at", "by", "for", "from", "if", "in", "into", "on", "or", "of", "the", "to", "with" are examples of stop words for English. Thus, the next important phase will be to remove all these words from the document. After this phase we will have a list which consists of the sentences with only the words that are meant for further processing

 

by the method. The last step for preprocessing is Word Stemming; Word stemming is the process of removing prefixes and suffixes of each word.

*C. Feature Extraction and Sentence Score Calculation*

After this preprocessing mentioned above, each sentence of the document is represented by a vector of features. We use eight features for each sentence. Each feature is given a value between (0) and (1). Each score assigned to sentence feature will be denoted S.S.(k) where k is one of seven features. The seven sentence features are being discussed as follows:

*1. Title Feature*

This feature gives the measure of the similarity between the title sentence and every other sentence of the document. We would like to intimate here that the Subject field in every Email would correspond the title of the document. This feature score is determined by counting the number of matches between the content words in a sentence and the words in the title. The score for this feature is the ratio of the number of matches between a sentence and title sentence over the number of words in title [19]:

$$S.S.(1) = \frac{No.of\ Title\ Words\ in\ Sentence\ S}{No.of\ Words\ in\ title} \qquad (1)$$

*2. Term weight Feature*

The frequency of term occurrences within a document has often been used for calculating the importance of sentence. The score of a sentence can be calculated as the sum of the score of words in the sentence. The score of important score *wi* of word *i* can be calculated by the traditional *tf.idf* method as follows. We applied this method to *tf.isf* (Term frequency, Inverse sentence frequency).

$$Wi = tfi\ x\ logN/n \qquad (2)$$

where *tfi* is the term frequency of word *i* in the document, *N* is the total number of sentences, and *n* is number of sentences in which word *i* occurs. This feature can be calculated as follows [19]:

$$S.S.(2) = \frac{\sum_{i=1}^{m} Wi(S)}{Max[\sum_{i=1}^{m} Wi(S)]_{i=1}^{n}} \qquad (3)$$

Where *m* is number of words in sentence.

*3. Sentence Position Feature*

Position of the sentence in the text, decides its importance. This feature can involve several items such as the position of a sentence in the document, section and paragraph etc. It should be noted here that if the implementation of the position score is a design choice based on summarizing news articles, which are typically written using the "inverse pyramid" structure, in which more important information is given to the reader first, followed later by finer-grained details of a story then we will give the first sentences (in position) high rank [18]. But we modify this parameter to fit with the new fact that we deal with personal email now. In the case of summarization of personal email, where we might expect the first one or two sentences of the text to express a courtesy, greeting etc. rather than the main point of the communication, we should tune the sentence position score to take this into account. We proposed here a position ranking from 1 to 5 for the position score as shown below:

$$S.S.(3) = \frac{1}{5}\ for\ the\ fist\ sentence, \frac{5}{5}\ for\ the\ second, \frac{4}{5}\ for\ the\ third,$$
$$\frac{3}{5}\ for\ the\ fourth, \frac{2}{5}\ for\ the\ second, \frac{0}{5}\ for\ the\ others. \qquad (4)$$

 

4. *Where and When Feature (Event Feature)*

We propose this feature for the sentence within the document of personal email; the correlation between the place and time (the existence) in the single sentence tells us about some event such as meeting, conference etc. We proposed here the following scoring method:

$$S.S.(4) = \frac{2}{2}\ if\ place\ and\ time\ have\ been\ mentioned\ in\ the\ sentence\ S, or$$
$$\frac{1}{2}\ if\ one\ factor\ (place\ or\ time)\ have\ been\ mentioned\ in\ the\ sentence\ S, or$$
$$\frac{0}{2}\ for\ others. \quad (5)$$

5. *Proper Nouns Feature*

The sentence that contains more proper nouns (name entity) is an important and it is most probably included in the document summary. The score for this feature is calculated as the ratio of the number of proper nouns that occur in sentence over the Summation of all proper nouns in the email [19]:

$$S.S.(5) = \frac{No.of\ Proper\ Nouns\ in\ the\ single\ sentence\ S}{Maximum\ number\ of\ Prper\ Nouns\ within\ Email} \quad (6)$$

6. *Numerical Data Feature*

The number of numerical data in sentence, sentence that contains numerical data is important and it is most probably included in the document summary. The score for this feature is calculated as the ratio of the number of numerical data that occur in sentence over the sentence length [19]:

$$S.S.(6) = \frac{No.of\ Numerical\ Data\ in\ the\ single\ sentence\ S}{Length\ of\ the\ Sentence\ S} \quad (7)$$

7. *Topical words*

The number of topical word in sentence, this feature is important because terms that occur frequently in a document (email) are probably related to topic. The number of topical words indicates the words with maximum possible relativity. Because of the shortage nature of email threads, we used the top 3 most frequent content word for consideration as topical. The score for this feature is calculated as the ratio of the number of topical words that occur in the sentence over the maximum summary of topical words in the sentence:

$$S.S.(7) = \frac{No.of\ topical\ words\ in\ the\ single\ sentence\ S}{maximum\ No.of\ topical\ words\ in\ the\ Sentence\ S} \quad (8)$$

8. *Sentence Score Calculation*

In this step we calculate the total score of the sentence S by summing all features subscores:

$$Score(S) = \sum_{k=1}^{7} S.S.(k) \quad (9)$$

*D. Sentence Extraction*

Then we will extract the best sentences as email summary according to the highest scores sentences.

## IV. RESULTS

We used 130 email documents from Enron_Sent_Mail_Sample from Enron email dataset as an input to the system. Here the human generated summaries are used as reference summaries for evaluation of our results. The human generated summary acts as a useful standard summary since humans can capture and relate deep meanings of the text as compared to machines. We received human generated summaries for our input documents from many different Experts. Here we call the summaries of our statistical summarizer, Copernic summarizer and MS Word summarizer as the candidate summaries and human

 

generated summaries as reference summaries. The chart below shows average similarity between candidate summaries and reference summary based on three factors: first; (Content similarity factor) which means the number of sentences from candidate summary that are similar to the sentences in the reference summary, second; (Position similarity factor) which means the similarity between position of the sentences in candidate summary and reference summary, while the third; (Topical similarity factor) which means the number of topical words that have occurred in the reference summary which are similar to the topical words that have occurred in the candidate summary. Results shows that the average similarity of our approach is 83% for the content factor, 81% for the position factor and 77% for the topical factor. MS Word 2007 Summarizer shows 82%, 80% and 74% respectively while for Copernic Summarizer it was 79%, 77% and 72% respectively.

TABLE I
AVERAGE SIMILARITY BETWEEN CANDIDATE AND REFERENCE SUMMARY

| Summarizer | Average | | |
|---|---|---|---|
| | Content Similarity | Position Similarity | Topical Similarity |
| Statistical | 0.83 | 0.81 | 0.77 |
| MS Word 2007 | 0.82 | 0.80 | 0.74 |
| Copernic | 0.79 | 0.77 | 0.72 |

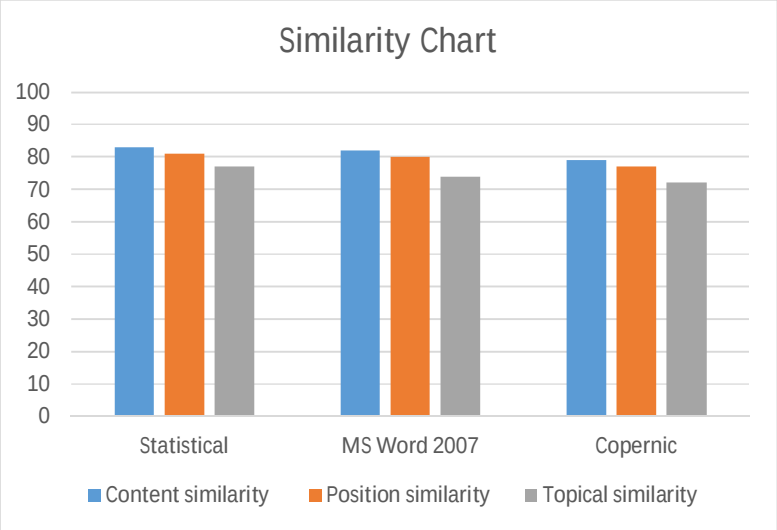


Fig. 2 Average similarity between candidate summary and reference summary

## V. CONCLUSION AND FUTURE WORKS

In this research we implement automatic text summarization which involves feature based extraction of key sentences using statistical method. The system is tested with 130 email conversation documents and compared with MS Word 2007 summarizer and Copernic summarizer. The results show that the use of statistical method with modifying the criteria of some features of sentence extraction in text summarization improves the quality of summaries. We applied our method for single document summarization which could be extended for multi-document summarization for future work.

## ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their valuable comments and suggestions. We would like to thank all the Experts who support this research by their reference summary and valuable comments.

 

References

[1] Steve Whittaker and Candace Sidner. Email overload: exploring personal information management of email. In CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems, pages 276–283, 1996.

[2] Gordon Rios and Hongyuan Zha. Exploring support vector machines and random forests for spam detection. In First Conference on Email and Anti-Spam (CEAS), July 30 - 31, 2004.

[3] Jihoon Yang and Sung-Yong Park. Email categorization using fast machine learning algorithms. In Discovery Science, pages 316–323, 2002.

[4] Gina Danielle Venolia and Carman Neustaedter. Understanding sequence and reply relationships within email conversations: a mixed-model visualization. In CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems, pages 361–368, 2003.

[5] Evelyne Tzoukermann Smaranda Muresan and Judith L. Klavans. Combining linguistic and machine learning techniques for email summarization. In Walter Daelemans and R´emi Zajac, editors, Proceedings of CoNLL-2001, pages 152–159. Toulouse, France, 2001.

[6] Evelyne Tzoukermann, Smaranda Muresan, and Judith L. Klavans. Gistit: summarizing email using linguistic knowledge and machine learning. In Walter Daelemans and R´emi Zajac, editors, Proceedings of HLT/KM 2001 Workshop at ACL/EACL 2001 Confenrence. Toulouse, France, 2001.

[7] Owen Rambow, Lokesh Shrestha, John Chen, and Chirsty Lauridsen. Summarizing email threads. In HLT/NAACL, May 2–7 2004.

[8]Wang, and Bo Li. Adaptive maximum marginal relevance based multi-email summarization. In *AICI*, volume 5855 of *LNCS*, pages 417–424. Springer. 2009.

[9] Stephen Wan and Kathleen McKeown. Generating overview summaries of ongoing email thread discussions. In Proceedings of COLING'04, the 20th International Conference on Computational Linguistics, August 23–27 2004.

[10] Donghui Feng, Erin Shaw, Jihie Kim, and Eduard Hovy. Learning to detect conversation focus of threaded discussions. In *Proc. HLT-NAACL*, pages 208–215. 2006.

[11] Lokesh Shrestha and Kathleen McKeown. Detection of question-answer pairs in email conversations. In Proceedings of COLING'04, pages 889–895, August 23–27 2004.

[12] Kathleen McKeown, Lokesh Shrestha, and Owen Rambow. Using question-answer pairs in extractive summarization of email conversations. In *CICLing*, volume 4394 of *LNCS*, pages 542–550. Springer. 2007.

[13] Simon Corston-Oliver, Eric K. Ringger, Michael Gamon, and Richard Campbell. Integration of email and task lists. In First Conference on Email and Anti-Spam, Mountain View, California, USA, July 30-31 2004.

[14] David D. Lewis and K. A. Knowles. Threading electronic mail - a preliminary study. Information Processing and Management, 33(2):209–217, 1997.

[15] Aaron Harnly Jen-Yuan Yeh. Email thread reassembly using similarity matching. In Third Conference on Email and Anti-Spam (CEAS), July 27 - 28 2006.

[16] Anja Belz, Roger Evans, and Sebastian Varges, editors. Proc. of the 2009 Workshop on Language Generation and Summarization (UCNLG+Sum 2009). ACL, Suntec, Singapore, August. 2009.

[17] Pablo Ariel Duboue. Extractive email thread summarization: Can we do better than He Said She Said?. In Proceedings of the 7th International Natural Language Generation Conference, pages 85–89. 2012.

[18] Jahna Otterbacher, Dragomir Radev and Omer Kareem. Hierarchical Summarization for Delivering Information to Mobile Devices, Preprint submitted to Elsevier Science, 2007.

[19] Ladda Suanmali, Naomie Salim and Mohammed Salem Binwahlan. Fuzzy Logic Based Method for Improving Text Summarization. (IJCSIS) International Journal of Computer Science and Information Security, Vol. 2, No. 1, 2009.

 