I.J. Information Technology and Computer Science, **2018, 7, 53-60**
Published Online July 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2018.07.06

A Domain Specific Key Phrase Extraction Framework for Email Corpuses

I V S Venugopal
Department of IT, G V P College of Engineering(A), Andhra Pradesh,530048, India
E-mail: venuillinda@gmail.com

D Lalitha Bhaskari
Department of CS&SE, AUCE(A), Andhra Pradesh, Visakhapatnam, India
E-mail: lalithabhaskari@yahoo.co.in

M N Seetaramanath
Department of IT, G V P College of Engineering(A), Andhra Pradesh,530048, India
E-mail: seetaramanath@gmail.com

Received: 08 March 2018; Accepted: 23 May 2018; Published: 08 July 2018

Abstract—With the growth in the communication over Internet via short messages, messaging services and chat, still emails are the most preferred communication method. Thousands of emails are been communicated everyday over different service providers. The emails being the most effective communication methods can also attract a lot of spam or irrelevant information. The spam emails are annoying and consumes a lot of time for filtering. Regardless to mention, the spam emails also consumes the main allocated inbox space and at the same time causes huge network traffic. The filtration methods are miles away from perfection as most of these filters depends on the standard rules, thus making the valid emails marked as spam. The first step of any email filtration should be extracting the key phrases from the emails and based on the key phrases or mostly used phrases the filters should be activated. A number of parallel researches have demonstrated the key phrase extraction policies. Nonetheless, the methods are truly focused on domain specific corpuses and have not addressed the email corpuses. Thus this work demonstrates the key phrases extraction process specifically for the email corpuses. The extracted key phrases demonstrate the frequency of the words used in that email. This analysis can make the further analysis easier in terms of sentiment analysis or spam detection. Also, this analysis can cater to the need for text summarization. The proposed component based framework demonstrates a nearly 95% accuracy.

Index Terms—Email Corpus, Key Phrase Extraction, Domain Specific Extraction, Modified Term Frequency, Modified Inverse Document Frequency.

I. Introduction

The traditional communication methods between the humans were consisting of spoken languages, sign languages and finally the written languages. These communication languages are usually categorised as natural languages [1]. Nevertheless, the communication methods have crossed the barriers of communications between humans and extended over the communication between human and machine and between machines and machines. The communication between the machines as the computers have fewer challenges as the communication is backed up by the binary system. Nevertheless, the communication between the human and computer systems have the major challenge of converting the human understandable languages into the computer understandable language. The well accepted process of language conversion for these purposes are called the natural language processing or NLP [2].

The machine language processing is a widely accepted technique for various reasons like content summarization, information retrieval or the information extractions. The content summarization process is mainly focuses on preparation of summary of any given text. This application of NLP can reduce the time of processing the complete text and regardless to mention has multiple application usages. This method was first introduced in novel work by Jusoh et al. [3] in the year of 2011. Further another application of NLP is the information retrieval process. This process mainly focuses on the query processing and conversion of natural language queries into the content specific terms. This process is elaborated by the notable work of Zukerman et al. [4] in the year of 2002. Yet another parallel research application of NLP is the information extraction. The primary focus of this process is to reduce the time to extract meaningful information from any given corpuses. The benefits of this process is to cater the benefits of correlation based information extraction, where the related terms can be inferred from the corpus and can be considered information gain towards the extraction

process. The notable work by Sekine et al. [5] [6] has demonstrated the use of information extraction with the benefits. The other popular outcomes from the parallel researches on NLP is the querying and answering methods as demonstrated by Bernhard et al. [7], machine based translations as proposed by Zhou et al. [8] [9], text to speech generation as formulated by Kaji et al. [10] and the sentence compression by Zhau et al. [11] [12] [13] [14]. Nonetheless, the base of all these applications and processes are the key phrase extractions.

Thus, it is natural to understand that the key phrase extraction is the major pre-processing analysis for any machine learning tasks ranging from summarization to email filtration. Nonetheless, the key phrase extraction processes are strongly depended on two factors as language and the domain:

- The language influence on the key phrase extraction cannot be ignored due to the inferences present in the languages and grammars.
- Also, the impacts of domain specific vocabularies are strong in terms of key frame extractions.

Thus, this work proposes a domain specific key phrase or key word extraction process for email corpuses.

The rest of the paper is furnished such as in the Section – II the current outcomes of the parallel researches are elaborated, in Section – III the domain specific extraction methods are discussed, in Section – IV the framework is elaborated, Section – V compresence the driving algorithm of the proposed framework, the results are discussed in the Section – VI and this work finally rests the conclusion in the Section – VII.

II. Current State of Art

The initial attempts for collecting the key phrases were manual as stated by Barzilay et al. [15]. The challenges of key phrase extraction are it is denoted as a complex process by Hasegawa et al. [16] and expected to be a high time consuming process as demonstrated by Ibrahim et al. [17]. During the extraction of key phrases, the possible elaborations in terms of synonyms are also to be considered. The variations of the results with the influence of parallel words are demonstrated by Shinyama et al. [18] [19]. Yet another factor for making the key phrase extraction difficult is making a complete list of parallel words for the key phrases are difficult as shown in the work by Lin et al. [20].

Further in this section of the work, the factors for key phrase extraction are elaborated.

A. Availability of Corpuses

WWW is an instance of free corpora which represents the largest public repository of natural language texts defined by Ringlstetter et al. [21]. This argument is supported by Zhao et al. [11]who write: “First, the web is not domain limited. Almost all kinds of topics and contexts can be covered. Second, the scale of the web is extremely large, which makes it feasible to find any specific context on it. In addition, the web is dynamic, which means that new words and concepts can be retrieved from the web”.

B. Validation of Results

The results of any key phrase extraction process are depending on the validation of the results. It is regardless to mention that the during the extraction process, considering the similar meaning of the words can improve the results. This hypothesis is called the distribution hypothesis introduced first by Harris et al. [22]. The extension to this hypothesis is carried out in the work of Bhagat et al. [23][24].

C. Key Phrase Extraction

It is natural to understand that the most important phase of the extraction process is the extraction of the key phrases from the corpuses. The dilemma in the research attempts is the base of the extraction process as the key phrases can be extracted either from the syntax based features or also from the semantic based features. The work of Ho et al. [25] demonstrates that the use of semantic based features can be useful during the extraction process. Nevertheless these are complementary to each other.

Henceforth, this work summarizes the research challenges in key phrase extraction:

- The extracted key phrases are to be considered for syntax based and semantic based for the difference in accuracy.
- The methods for extraction of key phrases for few popular methods are to be analysed.
- During the extraction process the synonyms and the idioms to be considered.
- The domain specific extraction process are to be analysed and for email based key phrase extraction is to be addressed

Thus in the next section this work analyses the domain specific key phrase extraction processes.

III. Domain Specific Extraction Methods

The domain specific key phrase extraction process is different from the general purpose extraction of key phrases. The domain specific list of text must be available in the framework for referring as the training text rather than the testing text. The novel algorithm elaborated by the Bannard et al. [26] is furnished here:

Algorithm 1: Existing Domain Specific Key Phrase Extraction

Step -1. Calculate the word frequencies from the training text
Step -2. Measure the threshold of word frequency
Step -3. Further analyse the testing corpus
  a. Calculate the word frequencies
  b. Compare with the threshold
  c. If word frequency > threshold
    i. Then Accept the key phrase
  d. Else
    i. Reject the key phrase
Step -4. Present the final list of key phrases

The algorithm is analysed visually as well [Figure – 1].

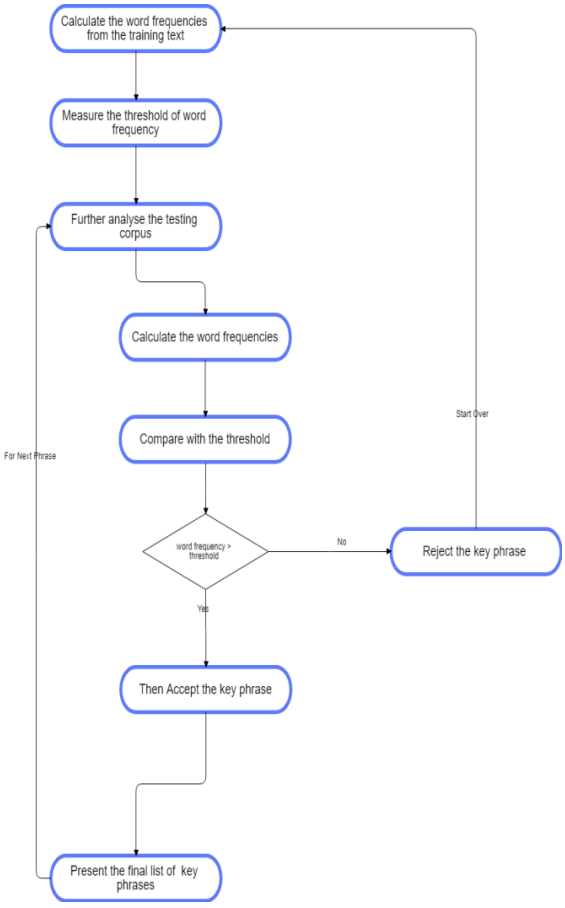

Fig.1. Existing Domain Specific Key Phrase Extraction

Thus, with the understanding of domain specific key framework extraction process and with the knowledge of no availability of the existing methods for key phrase extraction process for email, in the next section this work furnishes the framework for the intended purpose.

IV. Framework Elaboration

The major motivation of this work is the minimal availability of domain specific extraction of key words or key phrases and at the same time no availability for email key phrase extraction. This results into the proposed framework furnished in this section.

Considering the limitations of the parallel research outcomes, this work elaborates the components of the proposed framework [Figure – 2].

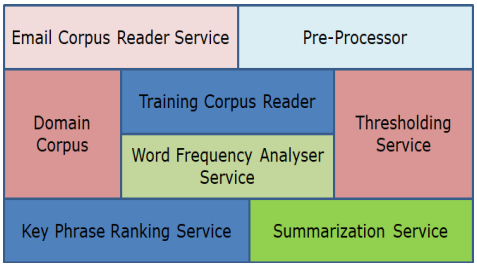

Fig.2. Domain Specific Key Phrase Extraction Process

The components of the framework are elaborated here:

A. *Email Corpus Reader Service*

The first component of the framework is the email reader service component. Due to the component based nature of this framework, any email service can be connected to this framework. This service needs to be preconfigured with the following parameters [Table – 1].

Table 1. Email Reader Service Configuration Parameters

Configuration Parameters	Purpose
Email_Address	Email Address of the receiver
Passwd	Password for the email account of the receiver
Server_Name	Name of the email server
Port	The port for the receiving email
Record_Size	Number of emails to be fetched per minute

The purpose of this component is to read the emails and build the testing corpus.

B. *Pre-Processor*

The second service or component in the framework is the pre-processor component. This component is responsible for cleaning the text and remove stop words. After the initial pre-processing, this component converts the complete text into tokenized set of text. The algorithm used in this component is elaborated in the next section.

C. *Domain Corpus*

The incorporated domain specific email corpus is used during the extraction of the key phrase frequency and further calculates the weighted average for the threshold. The description of the email domain corpus used in this work is elaborated here [Table – 2].

This corpus is a training corpus rather than testing corpus.

D. *Thresholding Service*

The thresholding service is the component in the framework to calculate the threshold of each word present in the training corpus. The algorithm for threshold calculation is elaborated in the next section of this work.

Table 2. Email Domain Corpus Description

Meta Information	Description
Number of users	158
Number of Emails	619446
Number of Email Threads	7520
Number of Emails per user	3920
Corpus Major Properties	• To • From • Text • Date_Time

E. Training Corpus Reader

The training corpus reader component is responsible for reading the tokenized words and passes the words to the word frequency analyser service.

F. Word Frequency Analyser Service

The word frequency analyser component is the implementation of term frequency and inverse document frequency calculator. The elaborated algorithm is analysed in the next section of the work.

G. Key Phrase Ranking Service

The final ranking of the keywords are given based on the thresholds obtained from the thresholding service. If the thresholds of the extracted key words are nearing to the value of the thresholds of the extracted key words from training corpuses then the keywords are listed in the final summarization service.

H. Summarization Service

The final service or component in this framework is the summarization service. This service provides the key phrases or the key words in terms of actual phrase and ranks. This information can further be used to calculate the sentiment or the spam factors of the emails.

V. Proposed Algorithm

This section of the work elaborates on the driving algorithms for the framework. The four fold algorithm is elaborated and analysed in this section.

A. Pre-Processing

The algorithm used in this component is elaborated here:

Algorithm 2: Pre-Processing Algorithm

Step -1. Accept the Email Corpus
Step -2. For Each Sentence Convert the stop words into ","
  a. Convert punctuations
  b. Convert the Braces
  c. Convert the question marks
  d. Convert the forward and backwordslace
  e. Converts "and" and "or"
Step -3. Convert all words into lower case
Step -4. Find the initial token
Step -5. For Each Sentence
  a. Extract the tokens based on separator
  b. Build the final token sets
Step -6. Generate the final token set for the corpus

The algorithm is visualized graphically [Figure – 3].

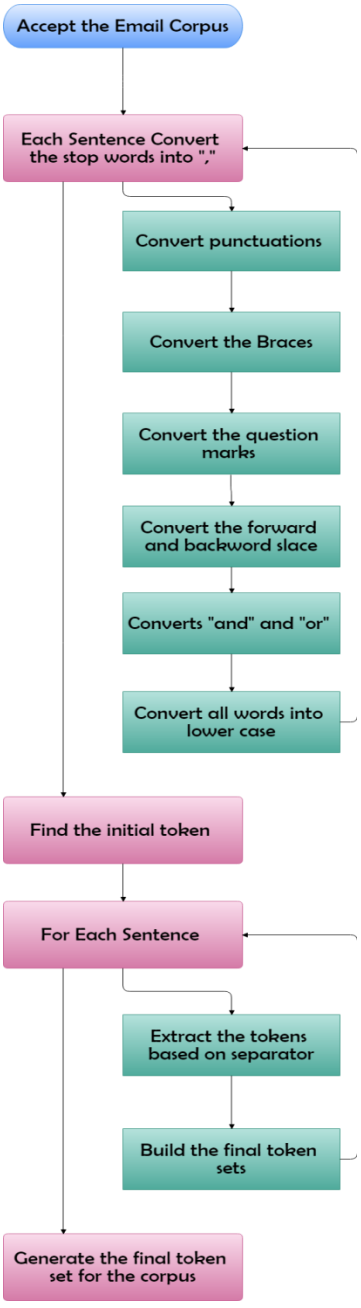

Fig.3. Pre-Processing Algorithm

B. TF-IDF

The second driving algorithm of this framework is the modified term frequency and inverse document frequency algorithm as elaborated here:

Algorithm 3: Modified TF-IDF Algorithm

Step -1. Generate the term count in the email corpus
Step -2. For each term in the list
  a. Calculate the term frequency as (term count / total terms count in the document)
Step -3. For each document in the corpus
  a. Calculate the inverse document frequency as log (documents includes the term / total number of documents)
Step -4. For each term in the document
  a. Calculate the term frequency with respect to inverse document frequency as term frequency X inverse document frequency
Step -5. Present the final list of terms per document

The algorithm is visualized graphically here [Figure – 4].

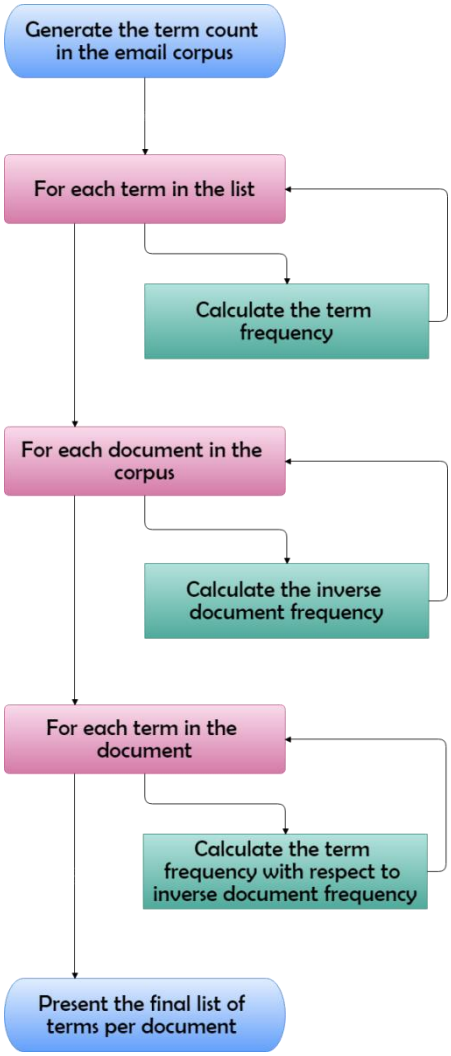

Fig.4. Modified TF – IDF Algorithm

C. Thresholding

The next algorithm is the Threshold calculation algorithm from the training corpus:

Algorithm 4: Thresholding Algorithm

Step -1. For each document in the training corpus
  a. Accept the TF-IDF values for each keyword
  b. Calculate the moving average for the keywords
Step -2. Build the weighted average for all the terms

D. Ranking

The final algorithm in this framework is the ranking algorithm. As a outcome of this algorithm, the documents will be summarized for further analysis. The algorithm is elaborated here:

Algorithm 5: Ranking Algorithm

Step -1. Accept the TF - IDF for each term in the testing corpus
Step -2. For each document
  a. Build the Array List with all the terms
  b. Sort the elements in the array list
Step -3. Generate ranking for all the key words or key phrases

Thus this fourfold algorithm in the framework generates the final ranking of the key phrases for further analysis.

The results obtained from this framework are discussed in the next section.

VI. Results and Discussion

Results obtained from this framework are highly satisfactory and discussed here in this section. This work evaluates three major corpuses collected from the spam filtration sample domain of google.

A. Corpus Length

Firstly the Corpuses used for generating the results are evaluated here [Table – 3]:

Table 3. Corpus Description Analysis

Parametric Information	Email - 1	Email - 2	Email - 3
Word Count	394	73	89
Number of Lines	53	12	20
Number of Paragraphs	19	5	9
Time to Pre-Process (Sec)	0.2	0.11	0.10

Thus it is natural to understand that the proposed pre-processing algorithm is time efficient and cater to the need of reduction in time complexity.

The results are been analysed graphically [Figure – 5].

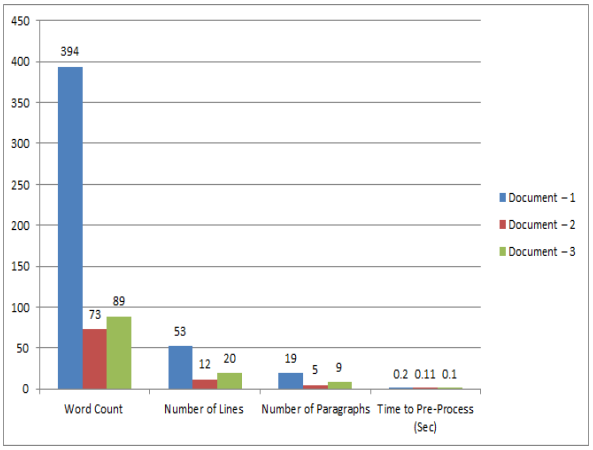

Fig.5. Corpus Analysis

B. *TF – IDF*

Secondly for the used corpuses the term frequency and inverse document frequency is analysed [Table – 4].

Table 4. TF – IDF Analysis

Key Phrases	TF – IDF		
	Email - 1	Email - 2	Email - 3
last	0.002726	0	0.037032
april	0	0	0.037032
your	0.043617	0	0.024688
payment	0.008178	0	0.024688
date	0	0	0.024688

Here this TF – IDF analysis demonstrates the stability of the proposed framework to extract the key phrased based on the term frequency.

The results are been visualised graphically [Figure – 6].

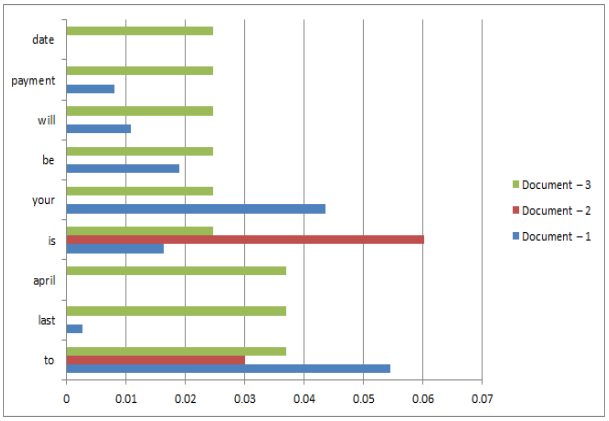

Fig.6. TF – IDF Analysis

C. *Ranking*

Further, the extracted key words are been ranked for each documents in the corpus [Table – 5].

Table 5. TF – IDF Analysis

Rank	Key Phases		
	Email - 1	Email - 2	Email - 3
(1)	YOU	TEXT	LAST
(2)	ATM	MANY	APRIL
(3)	CARD	QUALITY	DATE
(4)	THIS	WHAT	APPLICATION
(5)		DOCUMENTS	PAYMENT
(6)			YOU

Henceforth based on the key phrase analysis, the nature of the emails can be identified [Table – 6].

Table 6. Corpus Description Analysis

	Email - 1	Email - 2	Email - 3
Nature of the Email	Email from Bank regarding ATM Card	Email regarding text quality	Email regarding payment or application

Thus this key phrase extraction process can be helpful in many domains for emails corpus analysis.

D. *Accuracy Analysis*

Finally, accuracy of the key phrase extraction is evaluated for this framework [Table – 7].

Table 7. Accuracy Analysis

	Email - 1	Email - 2	Email - 3
Number of Actual Key Phrases	190	60	65
Number of Extracted Key Phrases	188	53	65
Accuracy (%)	98.94	88.33	100

It is natural to understand that the framework provides 95% accuracy in extraction.

The result is evaluated graphically [Figure – 7].

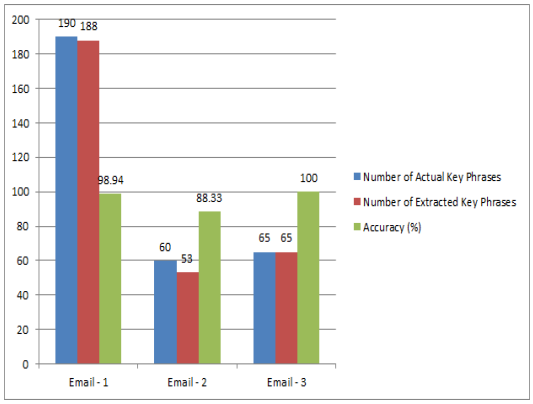

Fig.7. Accuracy Analysis

VII. Conclusion

The key phrase or the key words extraction is a primary task for further processing of a corpus to analyses the meaning, summary or the clustering. The extracted key phrases can be justified by the term frequency in the corpus. Nevertheless, the term frequencies depend on the writing style for each author and must be validated against the domain specific terms. This work provides a framework for email domain specific key frame extraction process. The accuracy demonstrated by this framework is highly satisfactory and nearly 95%. The final outcome of this work is to provide the email domain specific key phrase extraction and making the world of email analysis better.

References

[1] Azmi Murad MA, Martin TP, "Using fuzzy sets in contextual word similarity", *Intell Data Eng Automa Learn (IDEAL)*, LNCS 3177 pp.517–522, 2004.

[2] Bannard C, Callison-Burch C, "Paraphrasing with bilingual parallel corpora" *In Proceedings of the 43rd annual meeting of the Association for Computational Linguistics*, pp. 597–604, 2005.

[3] Jusoh S, Masoud AM, Alfawareh HM, "Automated text summarization: sentence refinement approach", *Commun Comput Inf Sci Digit Inf Process Commun 189(8)*,pp.207–218,2011.

[4] Zukerman I, RaskuttiB,Wen Y, "Experiments in query paraphrasing for information retrieval", *Adv Artif Intell, LNCS 2557*, pp.24–35,2002.

[5] Sekine S, "Automatic paraphrase discovery based on context and keywords between NE pairs", *In Proceedings of IWP*, 2005.

[6] Sekine S, "On–demand information extraction", *In Proceedings of the COLING/ACL onmain conference poster sessions*, pp. 731–738, 2006.

[7] Bernhard D, Gurevych I, "Answering learners questions by retrieving question paraphrases from social Q&A sites*", In Proceedings of the 3rd workshop on innovative use of NLP for building educational applications*, pp. 44–52,2008.

[8] Zhou L, Lin C, Munteanu DS, Hovy E, " ParaEval: using paraphrases to evaluate summaries automatically", *In Proceedings of the human language technology conference of the North American chapter of the ACL*, pp. 447–454,2006.

[9] Wu H, Zhou M, "Optimizing synonym extraction using monolingual and bilingual resources", *In Proceedings of the second international workshop on paraphrasing (IWP)*, pp. 72–79,2003.

[10] Kaji N, Kurohashi S, "Lexical choice via topic adaptation for paraphrasing written language to spoken language", *InfRetrTechnol LNCS 4182*, pp.673–679,2006.

[11] Zhao SQ, Wang HF, Liu T, Li S, "Pivot approach for extracting paraphrase patterns from bilingual corpora", *In Proceedings of ACL–HLT*, pp.780–788,2008.

[12] Zhao SQ, Lan X, Liu T, Li S, "Application-driven statistical paraphrase generation", *In Proceedings of the 47th annual meeting of the ACL and the 4th IJCNLP of the AFNLP*, pp.834–842,2009a.

[13] Zhao SQ, Wang HF, Liu T, Li S, "Extracting paraphrase patterns from bilingual parallel corpora", *Nat Lang Eng* 15(4),pp.503–526,2009b.

[14] Zhao SQ, Wang HF, Liu T, "Paraphrasing with search engine query logs", *In Proceedings of the 23rd international conference on computational linguistics (COLING)*, pp.1317–1325,2010.

[15] Barzilay R, McKeown KR, "Extracting paraphrases from a parallel corpus", *In Proceedings of the 39th annual meeting on Association for Computational Linguistics*, pp. 50–57,2001.

[16] Hasegawa T, Sekine S, Grishman R, "Unsupervised paraphrase acquisition via relation discovery", *Technical Report 05-012, Proteus Project, Computer Department, New York University*,2005.

[17] Ibrahim A, Katz B, Lin J, "Extracting structural paraphrases from aligned monolingual corpora*", In Proceedings of ACL*, pp.10–17,2003.

[18] Shinyama Y, Sekine S, Sudo K, "Automatic paraphrase acquisition from news articles", *In Proceedings of HLTR*, pp. 313–318,2002.

[19] Shinyama Y, Sekine S, "Paraphrase acquisition for information extraction", *In Proceedings of IWP*, pp. 65–71,2003.

[20] Lin D, Pantel P, "DIRT—discovery of inference rules from text", *In Proceedings of ACM SIGKDD*, pp. 323–328,2001.

[21] Ringlstetter C, Schulz KU, Mihov S, "Orthographic errors in web pages: toward cleaner web corpora", *J Comput Linguist 32(3)*,pp.295–340,2006.

[22] Harris Z, "Distributional structure. Structural and transformational linguistics", pp.775–794,1970.

[23] Bhagat R, Ravichandran D, "Large scale acquisition of paraphrases for learning surface patterns", *In Proceedings of ACL–HLT*, pp.674–682, 2008.

[24] Bhagat R, Hovy E, Patwardhan S, "Acquiring paraphrases from text corpora", *In Proceedings of the 5th international conference on knowledge capture (K-CAP)*, pp.161–168, 2009.

[25] Ho CF, Azmi Murad MA, Doraisamy S, Abdul Kadir R,"Comparing two corpus-based methods for extracting paraphrases to dictionary-based method", *Int J Semant Comput (IJSC)* 5(2), pp.133–178, 2011.

[26] Colin Bannard and Chris Callison-Burch, "Paraphrasing with bilingual parallel corpora*", In ACL*, pp.597–604, 2005.

Authors' Profiles

Mr.I.V.S. Venugopal received M.Tech degree in Software Engineering from JNT University Kakinada in 2010. He is pursuing Ph.D in JNTUK,Kakinada. Presently he is working as Assistant Professor in Department of IT at Gayatri Vidya Parishad College of Engineering(A), Visakhapatnam, Andhra Pradesh, India. His research interests include data and cyber security.

Dr. D.Lalitha Bhaskari is a Professor in Department of computer science & Systems Engineering in Andhra University College of Engineering(A), Andhra Pradesh, Visakhapatnam, India. she has received her Ph.D from JNT University Hyderabad in 2009. 6 Phds were awarded under her guidance and she is presently guiding more than 20 research scholars. Her research interests include

Cryptography & Network Security, Stenography & Digital Watermarking, Pattern Recognition, Image Processing, Cyber Crime & Digital Forensics.

Dr. M. N Seetaramanath is a Professor in Department of IT at Gayatri Vidya Parishad College of Engineering(A),Visakhapatnam, Andhra Pradesh, India. He received his PhD from Andhra Univerity in 1984. He has guided several research scholars. His research interests are in resource and mobility management for wireless adhoc networks, wireless sensor networks, and Network Security.

How to cite this paper: I V S Venugopal, D Lalitha Bhaskari, M N Seetaramanath, "A Domain Specific Key Phrase Extraction Framework for Email Corpuses", International Journal of Information Technology and Computer Science(IJITCS), Vol.10, No.7, pp.53-60, 2018. DOI: 10.5815/ijitcs.2018.07.06

