Improved Keyword and Keyphrase Extraction from Meeting Transcripts

J. I. Sheeba
Assistant Professor, Department of Computer Science & Engineering, Pondicherry Engineering College, Puducherry, India.

K. Vivekanandan
Phd, Professor, Department of Computer Science & Engineering, Pondicherry Engineering College, Puducherry, India.

ABSTRACT
Keywords play a vital role in extracting the correct information as per user requirements. Keywords are like index terms that contain the most important information about the content of the document. Keyword Extraction is the task of identifying a keyword or keyphrase from a document that can help users easily to understand the documents . Meeting transcripts is significantly different from document or other speech domains. This paper aims to extract keywords and keyphrases from meeting transcripts and also to add some additional features for improving the keyword and keyphrase extraction method . Here, this method is performed by both human transcripts and ASR transcripts and the keywords are extracted through MaxEnt and SVM classifier and Extraction of bigram and trigram keywords retrieval using N-gram based approach efficiently and also to identify the low frequency keywords using LDA (Latent Dirichlet Approach).Finally, the quality of the Extracted keywords is improved using pattern features through sequential pattern mining.

General Terms
Data Mining, Text Mining, Keyword Extraction

Keywords
Keywords, Keyphrase, Meeting Transcripts, LDA, SVM, MaxEnt.

1. INTRODUCTION
Now a days thousands of magazines ,articles and papers are published through online which makes it very difficult to go through all the documents. So there is a need of keyword Extraction methods which provide the main contents of a given document. It can also be used for other fields such as automatic indexing , information retrieval , classification , clustering , filtering ,tracking , text summarization, topic detection and report generation , information visualization, web searches etc[1].In the internet, all the information is available up to date it is becoming very important to search effectively and manage the available information. Keywords are very important information about the content of the document, usually keywords are manually written the number of documents is quickly growing every day so it is need to change this manual procedure to an automatic one[2].

Keyphrases as a brief summary of a document provide a solution to help organize, search and retrieve documents are very fast Keyphrases are noun phrases (NPs) that can reflect the main content of documents. Keyphrase extraction method aims to select a set of terms like bigram, trigram and N-gram words from a given document .It can be used for various Natural Language Processing (NLP) applications such as summarization and Question-Answering (QA) and search engines field, Keyphrases are mainly used for full-text indexing and assist users creating good queries[3][4].In this paper, some ideas are discussed to improve the keyword and keyphrase extraction from meeting transcripts.

Generally spoken, documents do not have keywords like meeting transcripts, so it is required to generate keywords automatically in the large amount of audio and video files. In this paper, the domain meeting transcripts has been focused. Meeting transcript is significantly different from written text and other audio data. For example, in meeting transcripts many people can participate even the deliberations also not well organized, and the speech is spur-of-the-moment and contains disfluencies and not well formed sentences and people may have different speaking styles,various pronunciations and they use different types of words and slangs in the meetings. People can act as different types of roles and topics in the meeting transcripts [5].So extracting keyword or keyphrases in the transcripts is difficult to compare with documents.

The aim of this paper is to extract low frequency keywords and keyphrases for every sentence in the meeting transcripts. In this proposed method it has been added some additional features to improve the keyword extraction methods like N-Gram based method, Capital letters, Double quotes method, LDA method, Sequential pattern mining and also it includes TFIDF method. Experiments are conducted using both the human transcripts and the Automatic Speech Recognition (ASR) out put. Results are evaluated by two classifiers like Max Ent, SVM and it is concluded that MaxEnt can extract more keywords .

2. RELATED WORKS
In(Feifan liu, Fei liu 2008) authors proposed automatic keyword Extraction for the meeting corpus using supervised approach and bigram expansion .In that paper, they extracted "entity bigrams" using bigram expansion compared to unsupervised TFIDF selection with POS filtering it performs well[6].

In (Feifan Liu, Deana Pennell, Fei Liu 2009) they introduced keyword extraction using TFIDF framework, it incorporates other methods like POS and word clustering and also it evaluates the importance of a word using graph based method[5] ,In (Fei Liu, Feifan Liu 2011)authors proposed Single-loop feedback strategy for keyword extraction. In addition of traditional frequency or position-based clues, term specificity features, decision-making sentence-related features, as well as a group of features derived from summary sentences. To generate better system summaries, they proposed a feedback loop mechanism under a supervised framework to leverage the relationship between keywords and summary sentences[7].

In (Xuan-Hieu Phan,2011) proposed for Hidden Topic-Based Framework with Short Web Documents using LDA model. In this frame work, they solved two problems like data

sparseness, synonyms problem using LDA method through MaxEnt classifier[8] .

In this proposed framework, it includes extracting low frequency keywords and keyphrases from the meeting transcripts using proposed model and also it includes some additional features for improving keyword, keyphrase extraction and to increase the quality of the keywords.

3. OVERALL ARCHITECTURE

The investigation of keyword extraction problem on meeting domain is done. So, this process is done step by step for best result in keyword extraction process. In this unsupervised approach, Maximum Entropy (MaxEnt) and SVM classifier are used to determine whether a word is a keyword. This is done by using both human and ASR transcripts. Feature extraction such as N-gram based approach for extracting bigram and trigram keywords, extracting Double quoted words and capitalized words are done and also low frequency keywords are extracted very effectively. Using pattern features in sequential pattern mining, the quality of the extracted keywords are improved.

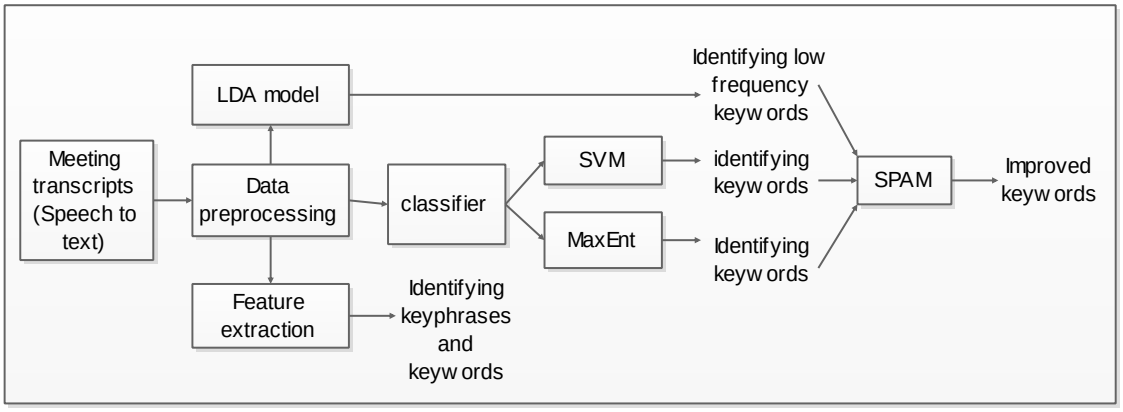

Figure 1.Proposed Unsupervised Framework for Extracting keywords and keyphrases from Meeting Transcripts

The figure 1 shows a Unsupervised frame work for Extracting keywords and keyphrases from the Meeting Transcripts . In this task, it includes 4 Steps

1. Extracting keywords from the Meeting Transcripts using Classifiers
2. Extracting Low frequency words using LDA model
3. Identifying keywords using Feature Extraction
4. Improving the quality of keywords using SPAM

3.1 Extracting Keywords from the Meeting Transcripts using Classifiers

3.1.1 Speech to Text

Speech to text is a technique of converting voice to text automatically. Speech to text is otherwise known as automatic speech recognition or voice recognition or computer speech recognition. Here, the term "speech recognition" is sometimes used to refer to recognition systems that must be trained to a particular speaker in the desktop recognition software. Here, Dragon naturally speaking software to provides better accuracy compared to the windows 7 speech recognition software and it has been used to provide the powerful software for converting voice to text which is used here.

3.1.2 Data Preprocessing

Data preprocessing is mainly used for removing irrelevant and redundant information present or noisy and unreliable data in the meeting transcripts. It includes cleaning, selection and feature extraction , normalization, transformation etc . This paper has undertaken stemming and stop word removal process for data pre-processing.

Stemming is the process to identify a word by its root which attempts to reduce a word to its *stem* or root form. Generally, the key terms of a document are represented by stems rather than by the original words. The number of discrete terms are needed for representing a set of documents. The stemming process makes a word shorter by removing such things as prefix or suffix. Examples for stemming the words flying, flew, flies can be changed into fly after the stemming process. The porter stemming algorithm is used to remove all affixes here.

Stop words are natural language words which have been filtered out after processing. To save disk space or to speed up search results in order to remove the common words are as called as stop words. Based on human input, only the stop words are removing and it is not an automated one .The list of stop words will be changed depending on their user input.

Examples of Stop words: A, An, About, Being, Can, The ,You etc.

3.1.3 Keyword Extraction Through MaxEnt Classifier

The MaxEnt classifier is a Maximum Entropy classifier and it is useful in extracting the keywords for the given input file based on frequency. Here, the input file format is given as .csv. Also with the frequency, some of the features are included and these features are produced by the environmental layers.

It works based on the formula

$$P(x) = \exp(c1 * f1(x) + c2 * f2(x) + c3 ** f3(x) ...) / Z$$

Where C1, C2, ….. → constant. f1, f2, ….. → features. Z → scaling constant

In Linear features, for each species, the output distribution has the same expectation in quadratic feature same expectation and variance of the environmental variables are the samples. Product feature produces mean value which

is the product of two continuous environmental variables. Environment variable is resultant from a continuous environmental variable. It produces binary values 0 and 1. When variable value >1, it produces 1. Hinge feature is same as the linear feature which has been derived from a continuous environmental variable.

Many files are produced by the MaxEnt for every species. For a species called *file1*, it produces files *file1.csv*. *file1.asc*(or*mySpecies.grd*). *file1.lambdas* containing the computed values of the constants *c1, c2, …. file1.png* is a picture of the prediction of each of the continuous environmental given by a directory containing the layers

3.1.4 Keyword Extraction Through SVM Classifier

A Support Vector Machine is used for classification purpose. It takes a set of input data and predicts for each given input to which category does the input fall. The training examples of one or two categories are taken. And the SVM training algorithm builds a model using those categories. It assigns new examples into one or the other categories.

SVM classifier takes a set of input data and predicts for each given input, to which category does the input fall. The SVM classifier can be trained by taking some sample documents per each category. The sample documents can be prepared by browsing the web pages which contain topic and category. The new input documents can be classified based on predefined category. As the number that stands for each category the result from the SVM classifier is extracted

3.2 Extracting low frequency words using LDA model

LDA is a generative probabilistic model of a corpus, the documents are represented as random mixtures over latent topics and each topic is characterized by a distribution over words. In the existing work, the keywords are identified only through the features such as position based, location, concept based, tf-idf, term specifity, etc[7].In LDA model, the words are grouped based on probability frequency then it takes many iterations to find more relevant words and hidden words as Output and Output is generated randomly, using gibb sampling algorithm and it has been identified the low frequency keywords from the meeting transcripts.

3.3 Identifying Keywords Using Feature Extraction

3.3.1 N-gram based Keyphrase Extraction

Keyphrases are the combination of 2 or more words which describe a meaningful and important content in a document. Only minimal documents have the author assigned keyphrases. Extracting the keyphrases manually is a difficult task. So, it must be automated. Keyphrases give the high-level description of the documents content and it is mainly used for users to decide whether the particular document is relevant or not.

Keyphrases summarize documents very quickly and it can be used as a low-cost measure of similarity between documents, additionally, it is used to cluster documents into groups .If the keyphrase is entered into a search engine , all documents with this particular keyphrase attached will be returned to the user[9].

The keyphrases are noun phrases which represent the main content of the documents. Noun phrases describes the meaningful phrases and it may be used in fields like intrusion detection, quality of service, text summarization etc. Keyphrases can be simple words or combination of 2 or more words. It may also contain hyphens(e.g sensor grouping) and apostrophes (e.g Bayes' theorem).The Part of Speech (POS) is to be assigned for each word in the document and if the words in the document have 2 and 3 nouns consecutively and it is taken as bigram and trigram respectively.

3.3.2 Double Quoted keywords

Important words in the transcript are referred by using double quotes. This feature checks whether the word starts and ends with double quotes. Hence, it is taken as a keyword.

3.3.3 Capital Words :Capitalization

This feature has a binary value showing whether there is an upper case letter in a keyword candidate. Generally the upper case letter words can represent important words and abbreviations. This feature checks whether the word is in uppercase. Hence it is taken as a keyword.

3.4 Improving the quality of keywords using SPAM

Sequential pattern mining is used to improve the quality of extracted keywords. Pattern features are acquired by sequential pattern mining from the sentence sequence. The keywords are extracted from the original meeting transcripts using different classifiers. That specified keyword list is taken as an input file to improve the quality of those extracted keywords. The seq_length specifies the keyword's longest pattern. The keyword is compared with the meeting transcript and the length is founded. Here, the seq_sup specifies the support of the longest pattern. The support value is calculated by the number of times and the specific keyword occurrence is divided by the total number of words in the longest pattern. The final value of seq_sup and seq_length are multiplied and the resultant is stored in Seq_sup_len. The threshold value is set to *min_sup*. Finally, the seq_sup_len value which falls on or above the min_sup value is taken as keywords using pattern features[10].

4. EXPERIMENTAL RESULTS

The keywords are extracted through MaxEnt and SVM classifier. The precision, recall and F-measure value are calculated to compare MaxEnt and SVM classifier's output. The Comparison of both human and ASR transcripts shows the extraction of more keywords through Dragon naturally speaking software 11.0 since reduced ASR error rate.

The transcripts are symbolically mentioned as T1, T2, T3, T4 and T5 which have been referred to as transcript T1 to transcript T5. The keywords obtained from Max Ent and SVM classifier and the results are compared. Finally, MaxEnt classifier shows better performance when comparing with the SVM classifier.

Table 1. Comparison of Keywords using MaxEnt and SVM classifiers

Transcripts	Max Ent classifier			SVM classifier		
	Precision	Recall	F- Measure	Precision	Recall	F- Measure
T1	1	0.6	0.75	0.8	0.4	0.53
T2	0.84	0.45	0.58	1	0.27	0.43
T3	0.67	0.36	0.47	0.33	0.09	0.14
T4	0.84	0.56	0.67	1	0.44	0.61
T5	0.84	0.56	0.67	1	0.44	0.61
Avg	0.84	0.55	0.63	0.84	0.33	0.46

The graph shows the comparison of MaxEnt classifier with SVM classifier having 5 meeting transcripts.

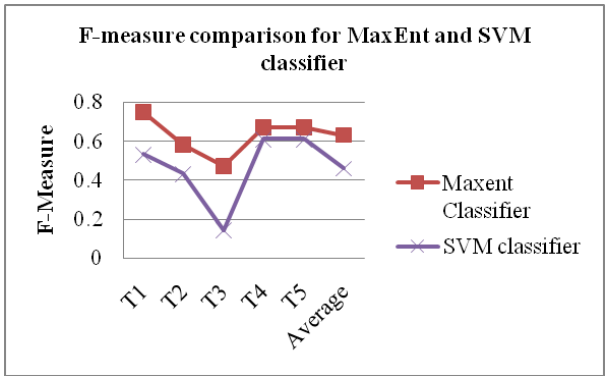

Fig 2: F-Measure comparison graph for Keywords extracted using MaxEnt and SVM classifiers

The table of Comparison among Human, windows7speech recognition software & Dragon naturally speaking software which shows that MaxEnt classifier performs better when comparing with the SVM classifier. Also, Dragon software is better than windows 7 naturally speaking software and extracts more keywords for the transcripts because of low errors.

The Number of keywords extracted from Max Ent classifier, SVM classifier and LDA method is given below.

Table 2. Comparison of Human, windows7speech recognition software & Dragon naturally speaking software

Meeting Transcripts	Human Annotated keyword	Keyword extraction human transcript		Keyword Extraction using Windows7 Speech recognition software		Keywords extraction using Dragon Naturally speaking software	
		MAX ENT	SVM	MAX ENT	SVM	MAX ENT	SVM
T1	10	6	5	4	4	5	4
T2	11	6	3	3	1	4	1
T3	11	7	3	3	2	5	3
T4	9	6	4	4	3	5	3
T5	9	6	4	3	2	5	2

Table 3: Keywords extracted through MaxEnt, SVM and LDA

Meeting Transcripts	Keywords Extracted		
	Max Ent	SVM	LDA
T1	6	5	8
T2	6	3	8
T3	7	3	9
T4	6	4	8
T5	6	4	8

Through LDA (Latent Dirichlet Allocation) approach, more number of keywords are extracted. Then the MaxEnt classifier performs better when compared with SVM classifier. The comparison is illustrated in the figure given below.

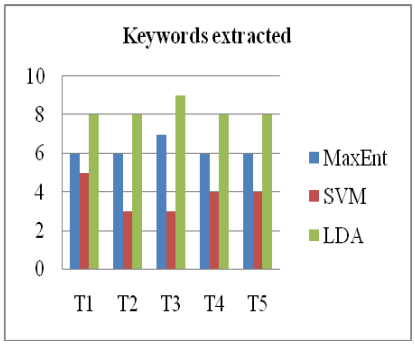

Fig 3: Comparison graph for MaxEnt Classifier, SVM classifier and LDA approach

Using feature extraction, the bigram, trigram, Double quoted and capitalized keywords are extracted. The bigram keywords are two combinational keywords and trigram keywords are three combinational keywords.

Table 4: Keywords extracted using feature extraction

Meeting transcripts	Keywords extracted			
	N-gram based keywords retrieval		Double quoted keywords	Capitalized keywords
	Bigram Keywords	Trigram Keywords		
T1	5	1	2	3
T2	11	2	2	3
T3	5	2	3	2
T4	10	8	3	1
T5	5	1	1	1

5. CONCLUSION

The paper has attempted to investigate extracting keywords from the meeting transcripts . In the existing system, the extraction of only less keywords are possible and also there occurs high word error rate due to ASR technique and only keywords are extracted not key phrases. But in this system, extraction of more keywords is possible with the Maxent classifier. Thus Maxent classifier is found better when comparing with the SVM classifier. The experiments are performed using both human transcripts and ASR transcripts. Finally, this method provides the best result by extracting key phrase, more keywords, minimized ASR error rates and N-gram based extraction is also used for bigram and trigram expansion of keywords. In addition , low frequency keywords are also identified using LDA (Latent Dirichlet allocation) model. The pattern features from sequential pattern mining is used to improve the quality of extracted keywords.

6. ACKNOWLEDGMENTS

I would like to thank my guide Dr. K.Vivekanandan, Professor, Department of Computer Science and Engineering, Pondicherry Engineering College for his valuable comments and suggestions in helping me to write this paper.

7. REFERENCES

[1] Jasmeen Kaur and Vishal Gupta.2010. Effective Approaches For Extraction Of Keywords**.** IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 6, November 2010.

[2] Joaquim Silva and Gabriel Lopes.2010. Towards Automatic Building of Document Keywords. Coling 2010: Poster Volume, pages 1149–1157,Beijing, August 2010.

[3] Zhiyuan Liu, Xinxiong Chen and Yabin Zheng .2011. Automatic Keyphrase Extraction by Bridging Vocabulary Gap .Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 135–144,Portland, Oregon, USA, 23–24 June 2011.

[4] Su Nam Kim, Timothy Baldwin and Min-Yen Kan.2010 Evaluating N-gram based Evaluation Metrics for Automatic Keyphrase Extraction. Proceedings of the 23rd International Conference on Computational Linguistics , pages 572–580, Beijing, August 2010.

[5] Feifan Liu, Deana Pennell and Fei Liu.2009.Unsupervised Approaches for Automatic Keyword Extraction Using Meeting Transcripts, ACM.

[6]Fei Liu, Feifan Liu, Yang Liu.2008.Automatic Keyword Extraction for the Meeting corpus using Supervised Approach and Bigram Expansion,2008 ACM.

[7] Fei Liu, Feifan Liu, and Yang Liu .2011.A Supervised Framework for Keyword Extraction From Meeting Transcripts, IEEE Transactions On Audio, Speech, And Language Processing, VOL. 19, NO. 3, March 2011,pp.538-548.

[8] Xuan-Hieu Phan and Cam-Tu Nguyen.2011.A Hidden Topic-Based Framework toward Building Applications with Short Web Documents, IEEE Transactions On Knowledge AndDataEngineering,VOL23.

[9] Eibe Frank and Gordon W.Paynter .Domain specific Keyphrase extraction.

[10]Jiajia Feng 2011.Keyword Extraction Based on Sequential Pattern Mining, ICIMCS'11August 5-7,china ,ACM.