International Journal of Research in Engineering, Science and Management
Volume-2, Issue-5, May-2019
www.ijresm.com | ISSN (Online): 2581-5792

Improved Keyword and Key Phrase Extraction from Meeting Transcripts

Vishwanath D. Doddamani[1], B. N. Varun[2], Sini Anna Alex[3]
[1,2]*Student, Department of CSE, Ramaiah Institute of Technology, Bengaluru, India*
[3]*Assistant Professor, Department of CSE, Ramaiah Institute of Technology, Bengaluru, India*

***Abstract*: Many a times, the size of the meeting transcript will be so large that identifying the real content in it is quite difficult. 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. This job can be simplified by extracting the keywords and keyphrases from the meeting transcripts. 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. There are various methods that can be employed to extract keywords such as Support Vector Machines(SVM), Max Entropy Classifier (MaxEnt), Latent Dirichlet Approach(LDA), tf-idf method, N-gram (bigram, trigram, …) based approach to identify low frequency words. The quality of the keywords that are extracted can be improved based on sequential pattern mining.**

***Keywords*: keyphrase extraction, meeting transcripts**

1. Introduction

Keyword is a word that occurs in text very frequently with some useful meaning. Keywords provide better understanding of the meeting transcript. It can be used for various processes like text categorization and information retrieval. The difficulty in study are Synonym, Homonym, Hyponymy and Polysemy problems.

- *Synonym:* Synonym means two or more words with the same meaning.
- *Homonym:* Homonym means a word that can have two or more different meanings.
- *Hyponymy:* Hyponymy is a relation between two words in which the meaning of one of the words includes the meaning of the other word.
- *Polysemy:* A Polysemy means word with different, but related senses.

2. Related works

- Extracts keywords and key phrases using SVM and Max ENT classifiers.
- Extracts keywords using clustering methods.
- Extracts keywords based on supervised learning.

3. Methodology

A. *Document preprocessing*

This process or step is required to eliminate noisy and irrelevant data in the documents. Also, to make it compatible with the algorithms. This is the important starting phase which needs to be implemented effectively.

1. *Tokenization:* It is the process of separating words from the sentences, in other words, it chops documents into pieces called tokens. Ex: “Humans are social beings” is chopped into
   [‘humans’,’are’,’social’,’beings’]
2. *Removing punctuations:* All the punctuations are removed in this step as these doesn’t contribute much to the objective and are redundant.
3. *Stopword removal:* The prepositions, conjunctions, articles are removed in this step as they can give misleading information about term frequency in the documents.
4. *Stemming:* This process maps derived words to its root form.
   Ex: interesting, interested will be mapped to interest. Normally porter stemmer algorithm is used for the process. In this paper we suggest a method to overcome some of the problems in porter stemmer algorithm like: for the words axe and axis, the plural form is axes but they both get mapped to the same word ‘ax’ as the rule in the algorithm removes ‘es’ from the words. The solution is, we need to use the homonym dictionary and properly map instead of the fixed rule. Also, we need to build a library of words mapped to it’s noun forms, so that the word ‘better’ gets mapped to ‘good’. But this however comes with tradeoff between speed and accuracy.

However, document preprocessing has some complications:

- *Apostrophe:* “aren’t” and “are not” are not mapped to the same word even though they mean the same.
- *Hyphenations:* The word “state-of-the-art” is separated into different tokens but they should be grouped together.
- *White spaces:* “San Francisco” is the name of the place and has to be considered as a single word but is split

into different tokens.

- *Different formats:* For example, date can be written in yyyy-mm-dd, mm-dd-yyyy, dd-mm-yyyy, dd/mm/yyyy etc, the algorithm doesn't specify whether it's the month or the day.
- *Language specific issues:* Different languages come with different variants of scripts. For example Korean, Japanese, Chinese languages are written without spaces. This causes problems in tokenization.

B. *Topic modeling*

All topic models assume

- Documents are collection of topics
- Topics are collection of words

The objective of this modelling is to identify the latent meaning or hidden topics of the documents and accordingly summarize the document according to the words in the topics identified.

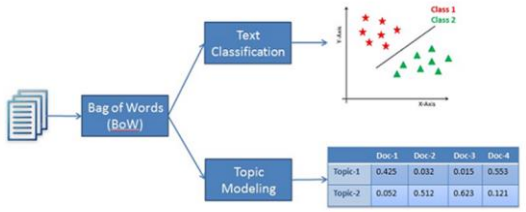

Fig. 1. Topic modelling

The above diagram differentiates text classification and topic modeling.

C. *Topic modeling using LSA*

LSA (Latent Semantic Analysis) also known as LSI (Latent Semantic Index) LSA uses bag of word(BoW) model, which results in a term-document matrix (occurrence of terms in a document). Rows represent terms and columns represent documents. LSA learns latent topics by performing a matrix decomposition on the document-term matrix using Singular value decomposition. The terms and the documents are classified into different topics and from the topics we can extract the keywords and also analyze the sentiments of the transcripts. This gives better insight than text classification.

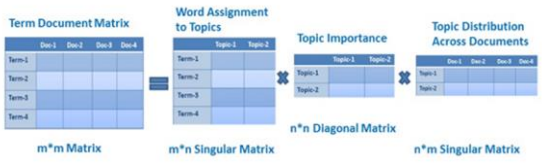

Fig. 1. Topic modelling using LSA

D. *Key phrase extraction using N-gram*

Key phrases are a combination of two or more words that describe the important contents in the document. The extraction of these key phrases are very difficult to be done manually, hence they need to be automated. The main advantage of identifying these key phrases is that they help to group similar documents together. Most of the key phrases are two or three nouns occurring consecutively. So they are extracted with bigram or trigram respectively and they give pretty good insight about the document.

4. Conclusion

This paper has attempted to extract keywords and keyphrases from meeting transcripts. In most of the existing systems keyword extraction is done through text classification but this paper suggests topic modeling instead, as this gives better insight to the transcripts than text classification. Also, for stemming, better method is suggested which however comes with tradeoff between speed and accuracy.

Acknowledgement

We would like to thank our guide Sini Anna Alex, Assistant Professor, Department of Computer Science and Engineering for her valuable comments and suggestions in helping us to write this paper.

References

[1] Immanuvel Rajkumar, Sheeba & Vivekanandan, Kumuda. (2012). Improved Keyword and Keyphrase Extraction from Meeting Transcripts. International Journal of Computer Applications. 52. 11-15.
[2] Feifan Liu, Deana Pennell, Fei Liu and Yang Liu Computer Science Department, The University of Texas at Dallas Richardson, TX 75080, USA.
[3] F. Liu, F. Liu and Y. Liu, "A Supervised Framework for Keyword Extraction from Meeting Transcripts," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 538-548, March 2011.
[4] 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.
[5] Feifan Liu, Deana Pennell and Fei Liu. 2009. Unsupervised Approaches for Automatic Keyword Extraction Using Meeting Transcripts, ACM.
[6] Eibe Frank and Gordon W.Paynter .Domain specific Keyphrase extraction.
[7] Jiajia Feng.2011. Keyword Extraction Based on Sequential Pattern Mining, ICIMCS'11August 5-7, china, ACM.
[8] Kazi Saidul Hasan and Vincent Ng Human Language Technology Research Institute University of Texas at Dallas Richardson, TX 75083-0688.
[9] Ken Barker and Nadia Cornacchia. 2000. Using noun phrase heads to extract document keyphrases. In Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence.
[10] https://www.datacamp.com/community/tutorials/discovering-hidden-topics-python