An empirical study of important keyword extraction techniques from documents

Conference Paper · December 2017

DOI: 10.1109/ICISIM.2017.8122154

4 authors:

H M Mahedi Hasan
University of Regina
2 PUBLICATIONS **19** CITATIONS
SEE PROFILE

Falguni Sanyal
3 PUBLICATIONS **19** CITATIONS
SEE PROFILE

Dipankar Chaki
The University of Sydney
23 PUBLICATIONS **119** CITATIONS
SEE PROFILE

Md. Haider Ali
University of Dhaka
47 PUBLICATIONS **246** CITATIONS
SEE PROFILE

An Empirical Study of Important Keyword Extraction Techniques from Documents

[1]H. M. Mahedi Hasan, [2]Falguni Sanyal, [3]Dipankar Chaki, [4]Md. Haider Ali
Department of Computer Science and Engineering
School of Engineering and Computer Science
BRAC University, Dhaka, Bangladesh
Email: [1]jisan2723@gmail.com, [2]falgunisanyal10@gmail.com, [3]dipankar@bracu.ac.bd, [4]haider@bracu.ac.bd

***Abstract*— Keyword extraction is an automated process that collects a set of terms, illustrating an overview of the document. The term is defined how the keyword identifies the core information of a particular document. Analyzing huge number of documents to find out the relevant information, keyword extraction will be the key approach. This approach will help us to understand the depth of it even before we read it. In this paper, we have given an overview of different approaches and algorithms that have been used in keyword extraction technique and compare them to find out the better approach to work in the future. We have studied various algorithms like support vector machine (SVM), conditional random fields (CRF), NP-chunk, n-grams, multiple linear regression, and logistic regression to find out important keywords in a document. We have figured out that SVM and CRF give better results where CRF accuracy is greater than SVM based on F1 score (The balance between precision and recall). According to precision, SVM shows a better result than CRF. But, in case of the recall, logit shows the greater result. Also, we have found out that, there are two more approaches that have been used in keyword extraction technique. One is statistical approach and another is machine learning approach. Statistical approaches show good result with statistical data. Machine learning approaches provide better result than the statistical approaches using training data. Some specimens of statistical approaches are Expectation-Maximization, K-Nearest Neighbor and Bayesian. Extractor and GenEx are the example of machine learning approaches in keyword extraction fields. Apart from these two approaches, semantic relation between words is another key feature in keyword extraction techniques.**

Keywords*— *conditional random fields; support vector machine; multiple linear regression; logistic regression; semantic learning

I. INTRODUCTION

Keyword extraction regarded as the crucial methods for data analyzation. The primary mission of important keyword extraction is to extract a specific group of words or keywords which highlights the main content of the documents. Automatic clustering, automatic filtering, automatic indexing, automatic summarization, information visualization, topic detection and tracking etc. are the basic data mining applications related to keyword extractions [1]. One of the important methods is the statistical approach that is used to identify the keywords based on statistical data. It does not require any training data. Some common approaches are word co-occurrence, PAT-tree, lexical analysis and syntactic analysis that is term frequency and n-grams [2].

Turney (2000) introduced the automatic keyword extraction. It is considered as a monitored machine learning method. Also, GenEx is suggested by zhang as a machine learning approach and in order to do that vector machine and genetic algorithm has been used. Kupiec, Pedersen, and Chen (1995), Teufel, and Moens (1997) used sentence level features to classify important sentences. Word frequency also used by Kupiec for the classification task. Barzilay and Elhadad (1997) have exhibited that the attributes based on lexical chains are beneficial for text summarization. Turney made a comparison between genetic algorithm and C4.5 decision trees for this task and decided that genetic algorithm provides better results than decision tress. Genetic method unravels both constrained and unconstrained optimization obstacles based on a natural selection procedure and its main task is to divide a population into individual solutions. C4.5 generates a decision tree that can be used for classification. In [3], condition random fields is a recent probabilistic model to label and segment a sequence data that have been applied to identify keywords.

In the next section, we have mentioned different algorithms that have been using in the keyword extraction field. In section 3, we have summarized different approaches and discussed the better approaches. Results and discussions have been made in section 4. At the end of our paper, conclusion along with future works are written. Finally, we have concluded our paper by mentioning some references.

II. LITERATURE REVIEW

Keywords can be viewed as a shortened version of documents. Important keywords can be used for text mining, web-page retrieval, document retrieval etc. Keyword extraction follows different phases and in [2], Pre-Processing Phase: unification and removing Stopwords has been used and AAS (Attention Attractive Strings) as keywords. Unification provides words definitions and gives them a general shape to

978-1-5090-4264-7/17/$31.00 ©2017 IEEE

identify. Post-Processing phase uses different efficient and effective algorithm to find the keywords. In many papers, word occurrence has been considered as one of the important features. The probability of KWNA (threshold), EPKLN and EPKRN have been used to minimize the threshold of words co-occurrences [2]. KWNA, EPKLN and EPKRN is a probabilistic term which identifies the number of keywords using AAS. AAS is a procedure which compares words to a co-occurrence word. In this paper, eight keywords are retrieved from every document for the primary list of keywords and later some of them are removed applying nominated post-processing techniques and 800 datasets have been used [2].

In keyword, NP-chunks accords a superior accuracy than n-grams with including POS-tagging as a further attribute [4]. The number of words and the frequency of a noun phrase, furthermore the frequency of the head noun is used by Barker and Cornacchia (2000) to determine the keywords from documents. Daille (1994) applied statistical filters on the extracted noun phrases to determine the keywords. In [4], a study showed that term frequency is the ultimate filter candidate in keyword extraction technique. N-grams technique removed non-alphanumeric characters that are not keywords in the training dataset. Numbers were taken if they appeared separately and proper nouns were kept. In NP-chunks, nouns contain the content of the documents and manually assigned keywords are happened to be noun or noun phrase with an objective. POS-tag patterns show tagging the word with proper parts of speech such as NOUN (singular or mass), ADJECTIVE NOUN (plural), NOUN (plural), ADJECTIVE NOUN (singular or mass) etc. From the result, it is shown that N-grams with POS-tag has high F-score [test accuracy] and Chunking with POS-tag has high precision as well as Pattern without POS-tag has high recall [4].

In [5], documents were split into words and processed to find out the noun, verb and phrases as keywords and the process of word splitter is 5 different steps as word splitter initialize, POS tagging, place and person names spotted, restart word splitter and restart the previous steps. Stop words have been removed in that process. ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) applies semantic analysis to develop the authenticity of word splitter. ICTCLAS comprises word dissection, unknown words recognition and Parts-Of-Speech tagging.

In keyword extraction technique, authors can set keywords for their documents and those might or might not be occurred in the text. As mentioned before, lexical text chains are effective features for summarization and to make lexical chains, semantic relations and word senses among words must be acknowledged [6]. WordNet is used for lexical chain building algorithm and the WordNet can be used for lexical chain builder to synonym, meronym and hypernym/hyponym. Occurrence of words in the documents is the key feature in different techniques in important keyword extraction. And it could be viewed as first occurrence in the text, a number of appearance of the words in the text and the last appearance of the words. C4.5 algorithm used with different features which decrease the variance and increase the accuracy of extracting keywords from the documents [6]. The lexical chain features develop the accuracy remarkably in the keyword extraction procedure. There are some effective approaches which show better results and description of these methods are discussed briefly here.

A. Conditional Random Fields

Sentence segmenting and labelling is an application of conditional random fields (CRF). It is used in keywords extraction that showed better result than the other approaches. The main approaches of this algorithm is, features extraction and pre-processing, CRF model training and keyword extraction and CRF labelling, results evaluation [3]. In [3], CRF++ tool with POS-tag is used to extract keyword from the documents.

B. SVM, MLR and Logit

Support vector machine (SVM), multiple linear regression (MLR), logistic regression (logit) has been applied in this approach along with TF*IDF (Term frequency and inverse document frequency) as a baseline to extract keywords from the documents [3]. SVM is supervised learning process that categorizes other inputs and produces an optimal output. The result shows that SVM and CRF give better results where CRF accuracy is greater than SVM based on F1 score and according to precision, SVM shows better result than CRF. But, in case of recall, logit shows the greater result. Logistic regression model shows better result than the multiple linear regression models in the process of keyword extraction [3]. And mentioned as future works, conditional random fields and semantic relation between keywords to perform on the greater number of texts.

C. Statistics, and Machine Learning

Keyword or Keyphrase extraction could be categorized into three categories: statistics, semantically matching and machine learning. P. Turney developed Extractor and GenEx which is machine learning based extraction. TRUCKS is the keyphrase extraction technique between statistical and sentimental approaches which show better results in extracting keyphrases from the documents but also reject actual keyphrases [8]. Jordi Vivaldi et.al [8] suggested AdaBoost algorithm to acquire the higher accuracy of keyphrase extraction system. It has also some disadvantages and it categorizes into two groups which are linguistic and statistics. In [8], main functions were split into two group, training mode and recovery mode where training mode initialize the main knowledge base and recovery mode recovers the false rejected keyphrases which are evaluated again to find out keyphrases from the documents. Statistical methods use word frequency, term frequency, and word co-occurrences which provide some good results. In [10], linguistic features increase the performances of extracting keywords from the documents.

Rapid automatic keywords extraction (RAKE) used to extract keywords that include a list of StopWords, phrase delimiters, and word delimiters [10]. In RAKE, stop words and phrase delimiters have been used to divide the document text into candidate keywords and "stemmer function" converts all the plural words to singular words, and removing suffixes based on different features. Adaptive lesk algorithm model will compute the word-to-word resemblance for all word pairs [11]. Statistics-based methods such as Bayesian, K-Nearest Neighbor, and Expectation-Maximization. As mentioned before in both statistical and machine learning approach, Inverse Term Frequency (ITF), Term Frequency (TF), Inverse Document Frequency (IDF)and Word Position (documents, paragraphs and sentences) modified Term Frequency and occurrence(first and last paragraphs, abstract, headings, figures, and tables etc.) has been applied. In Keyword extraction techniques these approaches provide an important impact on huge data set analysis, text analysis, and document summarization etc. fields.

D. Keyphrase Extraction, and Semantic Analysis

KEA (keyphrase extraction approach) used a machine learning methodology and depending on naïve Bayes decision rule [9]. TF-IDF scores can be applied to differentiate between Keyphrase and non-Keyphrase. In [9], noun factors are determined by its frequency in the document, its composition and how distinct these words and sub-phrases are in the domain of the document which is predefined as domain specific keywords or keyphrases in the database. Semantic relatedness can be viewed as numerous relations likely Antonym, Holonym, Hyponym, Hypernym, Troponym and Meronym. It shows the relation between words or meaning of words at word level as well as meaning of sentences and keyphrases at sentence level.

Keywords extraction methodologies can be categorized into two features which are quantitative and qualitative where qualitative techniques based on semantic relation [12]. All these approaches have been applied in this field and get results under different methodologies.

III. MATERIALS AND METHODS

In the previous section, all the approaches and algorithms have been briefly discussed. Among all those approaches, we analyzed Logit, NP-chunks, Ngrams, SVM, CRF and MLR methods. In figure 1, 2 and 3, we have tried to show how CRF, SVM, NP-chunk, Ngrams, Logit, MLF behaves based on precision, recall, F-score on keyword extraction. tp/(tp+fp) is the ratio for precision. Here, total number of true positive and false positive is tp and fp respectively. The precision is a classifier that has the ability not to label negative sample as positive one. The recall is also similar to precision (tp/(tp+fp)) where fn is false positive. It is a classifier that has the ability to identify every possible positive collections. The weighted harmonic means of both precision and recall is F1 score. F1 scores best at 1 and worst at 0. F1 score is the combination of 2*(p*r) / (p + r) (p=precision, r=recall). The differentiation among SVM, CRF, Logit and MLR are considered based on 540 training dataset. And, NP-chunks and Ngrams are considered along with the POS-tagging [1]. In order to extract the important keywords from documents, there is some basic methods that have been used in every algorithm such as Stopwords, Term frequency, Stemming, Parts of speech tagging, N-grams etc. Stopwords doesn't contain any core information of the documents and in order to increase processing speed Stopwords has been removed from the documents at the start of the procedure. Term frequency is calculating the occurrence of words in the documents. Most occurrence keywords certainly contain the core information of the documents and it has been used in the different algorithms to find out the important keywords. Different toolkits provide term frequency feature such as NLTK, RAKE, and TextBlob etc. Stemming means removing the suffixes from words and converts a word plural to singular so that the system would be able to detect the same word from the document. Understanding the words in the sentence is important and the system will POS-tag of every words in the documents which will define the words parts of speech. N-grams of text are being applied often in text mining and natural language processing methods. It is a set of co-occurring words within a given documents and when computing N-grams, the system moves one or more words forward. It depends on the number given in the N-gram calculation.

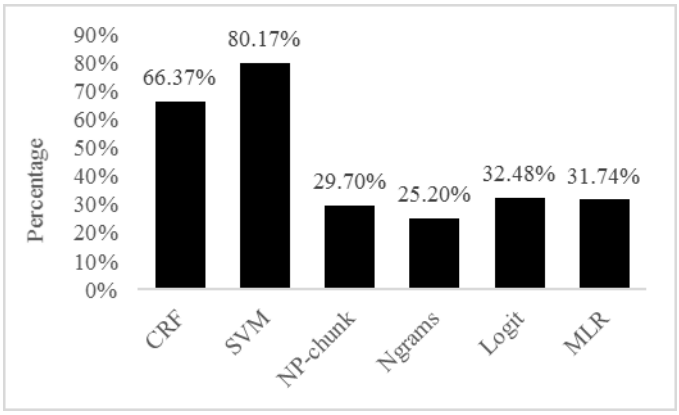

Fig. 1. Precision on keyword extraction

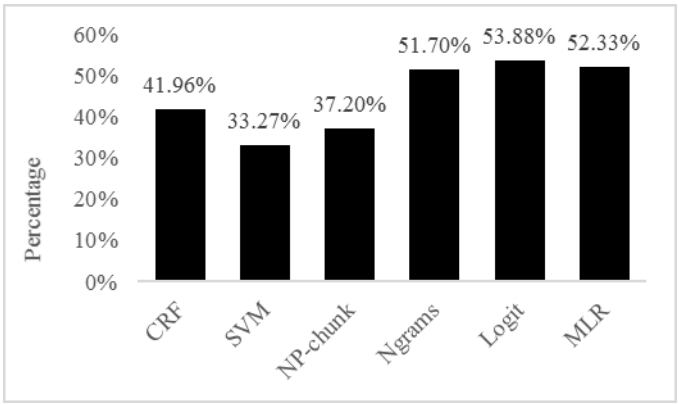

Fig. 2. Recall on keyword extraction

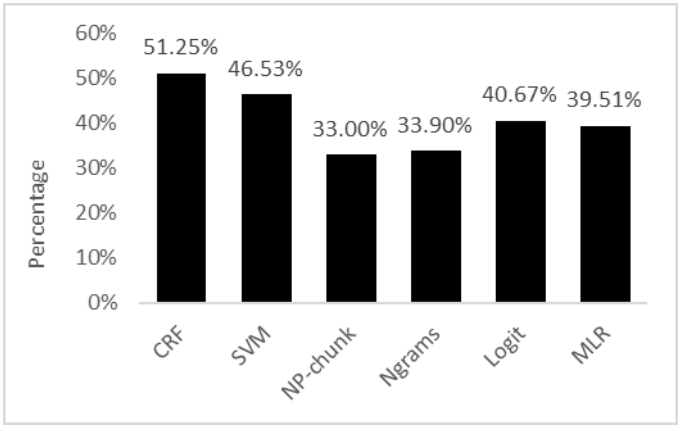

Fig. 3. F-Score on keyword extraction

IV. Experimental results and Evaluation

From above clustered columns and statistics, comparison between NP-chunks, Ngrams, SVM, Logit, CRF and MLR have been shown in the table given below:

TABLE I. COMPARISON OF RESULTS

Names	Precision	Recall	F-score
Logit	32.48%	53.88%	40.67%
NP-chunks	29.70%	37.20%	33%
Ngrams	25.20%	51.70%	33.90%
SVM	80.17%	33.27%	46.53%
CRF	66.70%	49.96%	51.25%
MLR	31.74%	52.33%	39.51%

All the approaches show different results and accuracy. The result shows that SVM and CRF give better results where CRF accuracy is greater than SVM based on F1 score (The balance between precision and recall). According to precision, SVM shows better result than CRF. Logit shows the greater result in case of recall. Ngrams show better result based on precision, recall and F-score. In case of recall, MLR shows the better result than other approaches. Among all the processes and algorithms, these algorithms show better result with higher accuracy in different fields. Above all CRF and SVM showed the better result than the other approaches

V. Conclusion and Future Work

In the paper, we have discussed different approaches of keyword extraction techniques and have made a comparison to find which approaches are better. We have found out that both SVM and CRF give better results. Whereas, based on F1 score (the balance between precision and recall), CRF accuracy is greater than SVM. SVM shows a better result than CRF according to precision. But, in case of the recall, logit shows the best result. In case of huge dataset, machine learning would provide better result than statistical approaches. Machine learning approaches would give better result by using semantic relations, conditional random fields and training data. In future, we will use conditional random fields along with semantic relations for important keyword extraction.

References

[1] C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, B. Wang, Automatic "Keyword extraction from documents using conditional random fields", J. Comput. Inf. Syst. 3 (2008)

[2] H. H. Kian · M. Zahedi, "Improving Precision in Automatic Keyword Extraction Using Attention Attractive Strings", Arab J SciEng DOI 10.1007/s13369-013-0573-6

[3] A. Hulth, "Improved Automatic Keyword Extraction Given More Linguistic Knowledge"

[4] H. Zhao, C., Song Zhu, "Automatic Keyword Extraction Algorithm and Implementation", Applied Mechanics and Materials Vols 44-47 (2011) pp 4041-4049 Online: 2010-12-06 © (2011) Trans Tech Publications, Switzerland doi:10.4028/www.scientific.net/AMM.44-47.4041

[5] G. Ercan, I. Cicekli, "Using lexical chains for keyword extraction", Information Processing and Management 43 (2007) 1705–1714

[6] Y. Matsuo, M. Ishizuka, "Keyword Extraction From A Single Document Using Word Co-occurrence Statistical Information", International Journal on Artificial Intelligence Tools Vol. 13, No. 1 (2004) 157169

[7] R. Kongkachandra; K. Chamnongthai, "Abductive Reasoning for Keyword Recovering in Semantic-based Keyword Extraction", Fifth International Conference on Information Technology: New Generations

[9] Y.B. Wu; Q. Li; R. S. Bot; X. Chen, "Domain-specific Keyphrase Extraction", Information Systems Department New Jersey Institute of Technology Newark, NJ 07102.

[10] A. Dutta, "A Novel Extension for Automatic Keyword Extraction", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 5, May 2016

[11] H. Haggag; A. Abutabl; A. Basil, "Keyword Extraction using Clustering and Semantic Analysis", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Impact Factor (2012): 3.358

[12] H. Haggag, "Keyword Extraction using Semantic Analysis", International Journal of Computer Applications (0975 – 8887) Volume 61– No.1, January 2013

[13] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge", In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'03), 216 – 223, Sapporo, 2003

[14] Y. Suzuki, F. Fukumoto, Y. Sekiguchi, "Keyword extraction of radio news using term weighting with an encyclopedia and newspaper articles", Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 373 - 374, New York, USA, 1998.

[15] M. Andrade and A. Valencia, "Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families", Bioinformatics, 1998, 14(7), 600–607.

[16] A. Hulth , "Combining machine learning and natural language processing for automatic keyword extraction", PhD Thesis, Stockholm University, Faculty of Social Sciences, Department of Computer and Systems Sciences, 2004

[17] L. F. Chien, "PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval", Proceedings of the ACM SIGIR International Conference on Information Retrieval, 1997, pp. 50--59.

[18] B. Hong, D. Zhen, "An Extended Keyword Extraction Method", 2012 International Conference on Applied Physics and Industrial Engineering.

[19] A. Azcarraga, M. Liu, R. Setiono, "Keyword Extraction Using Backpropagation Neural Networks and Rule Extraction", WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012