Algorithm for Bengali Keyword Extraction

Md. Ruhul Amin*, Madhusodan Chakraborty*
* Search Engine Pipilika, Department of CSE
*Shahjalal University of Science & Technology
Sylhet, Bangladesh
{shajib.sust, opuchakraborty}@gmail.com

***Abstract*—We present algorithm for keyword extraction from a Bengali document. In natural language processing (NLP), keyword extraction is the automated process to identify a set of terms that represent the information discussed in a document. A lot of research works have been done for keyword extraction in resource rich languages. Some of those works followed supervised approach using specific corpus whereas the latest techniques use unsupervised approach. Keyword extraction procedure already achieved state-of-the-art performance for the resource rich languages. Only a few works have been done on the keyword extraction for documents in Bengali but none of them could achieve $> 70\%$ accuracy. In this article, we discuss the methods for extracting Bengali keywords from a specific document collection following unsupervised learning approach. Generally, Bengali keyword extraction is difficult in terms of words parsing, stemming, excluding stop words etc. The accuracy of those modules also impact the performance of the keyword extraction procedure. However, we obtained 87% accuracy to identify the correct Bengali keywords from a document. The procedure we have discussed for keyword extraction can also be applied to any language; but here we have provided all of our experimental results specifically for Bengali language.**

***Index Terms*—Term frequency, Inverse document frequency, Co-occurrence matrix, Chi-square distribution and Tseng's keyword extraction algorithm**

I. Introduction

Keyword extraction algorithm spontaneously recognizes a set of terms that best summarize the topics discussed in a document. Keyword identification is extremely important in the field of Text Mining, Information Retrieval and NLP. In searching process, keywords are widely used to categorize search results which help users to find specific data quickly. Keywords are also used for document representation in classification task.

Keyword extraction algorithms have been studied extensively for more than five decades. Those studies can be divided into four broad sections:

- corpus based keyword extraction
- linguistics feature based keyword extraction
- statistical approaches for keyword extraction
- language model based keyword extraction

Despite wide applicability and research, automatic keyword extraction process suffers from poor performance in resource poor languages. To alleviate this situation, this paper discuss the statistical and algorithmic approach to identify keywords form individual document.

This paper is organized as follows. In section 2, we point out different keyword extraction algorithms studied in last five decades; in section 3, we explain the details of our proposed methods; in section 4, we show the results of keyword extraction from Bengali document. In section 5 and 6, we evaluate the proposed algorithms.

II. Different Keyword Extraction Approaches

In this section, we discuss the related works of keyword extraction research that achieved important milestones:

A. Corpus based Keyword Extraction

This approach requires a large collection of word/phrases and their count across thousands of documents. These algorithms contrast the frequency of a particular word from a document against the distribution of word frequencies across all the documents in the corpus to compute the significance of the respective word [1] [2]. But it is very hard to build a corpus for resource poor language to fulfill the objective.

B. Linguistics Feature based Keyword Extraction

This approach use the linguistics feature of the words based on it's use in the sentences and documents [3] [4] [5]. For example, the noun phrases can be considered as the most common sources of keywords in a document. But the linguistic feature annotation, such as: tokenisation, parts of speech tagging, lemmatisation and dependency parsing etc are not very well studied in resource poor language.

C. Statistical Approach for Keyword Extraction

This approach comprises simple methods which are language and domain-independent. The statistics of the words in a document, such as: n-gram statistics, word frequency, TF-IDF, word co-occurrences, etc are used for keyword extraction. Most of these statistical methods also require a wealth of document collection [6].

978-1-5386-8207-4/18/$31.00 ©2018 IEEE

D. Language Model based Keyword Extraction

This approach involves the use of word representation with respect to it's context for keyword extraction. These algorithms rank words/phrases based on the probability distribution of words with respect to the context of given document. Not to mention that computing the language models require a large collection of documents from various sources; hence not suitable for resource poor languages [7].

As most of the above approaches require a large number of documents, we did not follow the footsteps of those respective researchers. We describe the methods for keywords extraction from individual document in the next section. As we did not come across any Bengali keyword identification algorithm that has $> 70\%$ accuracy, we consider those discussion is not mention worthy. We will compare our results with the keyword identification methods applied for English language. Among the methods we described above, some are supervised and others are unsupervised learning. Supervised methods use previously complied corpus and a set of predefined keywords. As in Bengali we do not have any corpus of curated words, we follow unsupervised approach using statistical measures.

III. PROPOSED METHOD

Matsuo and Ishizuka [8] applied a chi-square measure to compute the significance of words and phrases based on the co-occurrences within the sentences in a particular document. The chi-square measure, which determines the bias of word co-occurrences in the document, is used to rank words and phrases as keywords of the document. As the chi-square measure is language independent and it can be computed from the word co-occurrences of the given document, we use this procedure for extracting significant terms from document. Those words are then used by Tseng's keyword extraction algorithm that repeatedly merges back nearby words based on three simple merging, dropping and accepting rules to generate keywords [9].

The proposed methods not only focus on the terms that are important but also consider term biasness with important terms. It then suggests the most likely keywords of document through filtering process. Here we discuss important terms and explain their usability and reasons in our procedure. We define the important terms below and then provide the algorithm for keyword extraction.

A. TF-IDF

TF-IDF is defined as the product of term frequency and inverse document frequency for a specific term in a given document [10]. It represents the importance of a word in the given document with respect to the whole corpus. So, for a specific term t and a particular document d, the term frequency will be,

$$TF(t,d) = \frac{f}{T} \tag{1}$$

f = number of times term t occurs in document d.
T = total number of different terms in document d.

Inverse document frequency of a term can be measured by calculating the total number of documents in collection divided by the number of documents in which the term occurs and taking the logarithm of this result. For a specific term t and a particular document d, the inverse document frequency will be

$$IDF(t,d) = \log \frac{N}{n} \tag{2}$$

Where, N = total number of documents.
n = number of documents in which term t occurs.

In our methods firstly, we calculate TF-IDF of different terms and then sort them in descending order according to TF-IDF score and collect 30% of them to generate a set of important terms.

$$TF - IDF(t,d) = TF(t,d) \times IDF(t,d) \tag{3}$$

Where, $d\ \varepsilon\ D$ (all documents)

B. Chi-square Distribution

Chi-square (χ^2) distribution [11] is used to find the co-occurrence bias between a term and an important term. It can be computed based on the co-occurrence frequency of two adjacent words. In our approach, we use co-occurrence matrix for computing the relevancy between two terms.

Now, if probability distribution of co-occurrences, between term t and the important terms I (based on TF-IDF score discussed above), is biased to a particular subset of important terms, then term t is likely to be a keyword. The statistical value of χ^2 for a term t can be defined as,

$$\chi^2 = \sum_{i \varepsilon I} \frac{((freq(t,i) - (n_t \times p_i))^2}{n_t \times p_i} \tag{4}$$

Where,
I = Set of all important terms according to their TF-IDF.
i = an important term from set I, $(i \epsilon I)$
n_t = total number of co-occurrence of term t and i
p_i = expected probability of i
$freq(t,i)$ = frequency of co-occurrence of term t and i

Here, if the chi-square value of a term exceeds critical value according to **degree of freedom** and the **significance level** then that term is transferred to Tseng's keyword extraction algorithm for suggesting multiple terms keywords.

C. Tseng's Keyword Extraction Algorithm

Tseng's keyword extraction method assumes that a document concentrating on a topic is likely to mention a set of terms a number of times. Maximally repeated terms in the text are thus extracted as keyword candidates. We describe the algorithm below:

Algorithm 1 Tseng's Keyword Extraction Algorithm

1: Calculate inverse document frequency (IDF) of all different term's root form.
2: **for** each individual document to entire document size **do**
3: Calculate TF-IDF value of all different terms in individual document.
4: Sort all different terms according to TF-IDF score in descending order.
5: Collect top 30% TF-IDF scoring terms and consider them as important terms.
6: Compute co-occurrence matrix of filtered terms.
7: Measure chi-square distribution for all different terms.
8: Collect the terms that support the null hypothesis of being keyword using chi-square test.
9: Generate multiple terms keyword suggestions from single terms by repeatedly merging nearby words based on Tseng's merging, dropping and accepting rules.
10: Remove the keywords with lower TF-IDF.
11: Sort the suggested keywords with respect to their frequency in the document.
12: Use filtering process and take the high frequency keywords as final keywords.
13: **end for**

IV. KEYWORD IDENTIFICATION

Firstly, we generate IDF value for the root forms of all terms for the documents in collection. To find the root form for each term, we use a Bengali stemming procedure which has its own dictionary list. We calculate the TF for each term as well as its root form. Then we find out TF-IDF value and save this value against the original term as well as its root form. Here we use one assumption which is any keyword whether in root form or not will get the same TF value. The reason for this assumption is key phrases are not always in root form. Hence for a better phrase detection, we need to analyze both the original and root form of a word using chi-square test.

সংবিধান	মূলনীতি	বিধান	পাঠক	শেখ
অনুচ্ছেদ	হইল	রাষ্ট্রপতি	প্রিয়	বঙ্গবন্ধু
হইবে	বাংলাদেশ	প্রজ্ঞাপন	সাল	অনুরূপ
সংশোধন	আইন	স্বৈরশাসক	সংসদ	এবং
প্রস্তাবনা	মুজিবুর	ধারাবাহিকতায়	বিভাগ	পরিচালনা
প্রজাতন্ত্র	অতঃপর	সামরিক	আপিল	তার
ঘোষণাপত্র	সম্মতিদান	ক্ষমতা	সুপ্রিম	দ্বারা
তাহা	পাকিস্তান	সূত্র	কোন	এই
স্বাধীনতার	গৃহীত	রাষ্ট্র	জন	গ্রহণ
বিল	বলিয়া	নিকট	কোর্ট	রায়

Fig. 1. Important terms from a document

In our experiment, we consider 1000 documents related to political news from online newspaper www.prothom-alo.com. From these documents, the IDF value of each unique term has been calculated. We take a single document from which terms with top 30% TF-IDF score are taken as important terms set (Figure 1). The snippet of this document along with the experimental results of keywords extraction is shown in the appendix at the end of this paper.

Also, we define the TF-IDF threshold value as the lowest score of the top 30% important terms. Later during keyword identification, we will remove all the candidate keywords under this threshold.

Hence, TF-IDF threshold = The least TF-IDF score of important terms

For our experiment, the number of classes is the total important terms and number of restriction is 1 because we assume that at least one of the important terms can bias a term.

Hence, Degree of freedom = Important terms list size-1

After threshold calculation chi-square distributions for different terms are calculated. To measure the chi-square distribution, firstly we generate a co-occurrence matrix. Using the co-occurrence matrix and the important terms, the chi-square value for each term is calculated. In our approach, asymmetric co-occurrence matrix is used where the co-occur frequency resides in cell (t_1, t_2) defines the frequency of occurrence of one term t_2 after another term t_1 (Figure 2).

Words	সংবিধান	বাংলাদেশ	শাসন	কোর্ট	অনুচ্ছেদ	প্রস্তাবনা	রাষ্ট্র
সংবিধান	0	0	0	0	12	9	0
বাংলাদেশ	32	0	12	27	0	11	23
সামরিক	22	0	67	0	2	0	43
সুপ্রিম	0	0	0	39	4	0	13
কোর্ট	14	2	0	0	0	23	0

Fig. 2. Co-occurrence matrix

Here In our experiment, a (NXM) co-occurrence matrix is used to calculate co-occurrence frequency where N = all different terms, and M = all important terms. Now, using co-occurrence matrix, we can calculate the chi-square distribution for all the terms with respect to the important terms. The following table shows the chi-square value of some important terms (Figure 3).

Term	Chi square value
সংবিধান	1086.4720347565776
অনুচ্ছেদ	508.9198424812262
সংশোধন	468.51875919464106
প্রস্তাবনা	282.5029233383468
প্রজাতন্ত্র	235.42065570509646
মূলনীতি	241.16636424668312
বাংলাদেশ	459.6678857948162
ঘোষণাপত্র	253.06204850092578
স্বাধীনতার	334.560694044899
বঙ্গবন্ধু	142.99028694923007
আইন	286.4838384100391
স্বৈরশাসক	254.8788853523754

Fig. 3. Chi-square value of terms

After chi-square test, we send the terms with their TF value in Tseng's keyword extraction algorithm which will generate candidate keywords. To choose the best possible keywords from this candidate set, we follow the following filtering steps:

1) Using Bengali stop word list for removing terms which are stop words.
2) Remove keywords under TF-IDF threshold.
3) Consider single term keyword in root form.
4) For multiple terms keyword, consider last term in root form.

After Tseng's keyword extraction algorithm and filtering process, the result is shown in the following Figure 4.

বিভাগ	স্বাধীনতার ঘোষণাপত্র	সামরিক	প্রজাতন্ত্র
রাষ্ট্র	প্রস্তাবনা	বাংলাদেশ	দ্বারা
হইবে	প্রজ্ঞাপন	বিধান	প্রিয়
স্বৈরশাসক	আপিল বিভাগ	বঙ্গবন্ধু	মূলনীতি
সম্মতিদান	অনুচ্ছেদ সংশোধন	কোর্ট	বলিয়া
ঘোষণাপত্র	সংবিধানের প্রস্তাবনা	সূত্র	ক্ষমতা
সামরিক আইন	রাষ্ট্রপতির নিকট	রাষ্ট্রপতি	

Fig. 4. Keyword list

V. PERFORMANCE ANALYSIS

Here, we take a small collection of text documents which are articles of Prothom Alo newspaper as our document collection for testing the procedure. It is a collection of 1000 text files. For the example document (see appendix) we define the keywords manually and compare them with our extracted keywords. The terms used for this calculation are the following:

- True Positive: Keyword detected as Keyword
- True Negative: Not Keyword and not detected as Keyword
- False Positive: Not Keyword but detected as Keyword
- False Negative: Keyword but not detected as Keyword

For measuring the performance of our described method, we use some known measurement here. We show the Precision, Recall, Accuracy and F-Measure for the generated result. Table I and II shows the performance of keyword extraction method for the example document where actual keywords are extracted manually.

TABLE I
EXPERIMENTAL RESULTS

Actual keywords	Total keywords found	Actual keywords found	Missed keywords
59	44	31	13

TABLE II
PRECISION, RECALL, ACCURACY AND F-MEASURE

Precision	Recall	Accuracy	F-measure
70.45%	52.54%	87.21%	59.67%

VI. PERFORMANCE EVALUATION

There are some inappropriate keyword suggestions generated by Tseng's algorithm which have a low TF-IDF value but a high chi-square value. These terms are coming in the keyword list for their high co-occurrences with the terms having a greater TF-IDF value. To solve this problem, we can remove the words with very low TF-IDF score from the given document in the beginning. These considerations will help us to increase the performance of our described method. We may also try to improve the filtering process to get better result.

We have used a Bengali stemmer for stemming the words. Its accuracy is 72%. Again, we did not use any corpus. We have used a generic parsing method to extract data from the online newspaper sites and its inaccuracy also decreases the performance of the given procedure. Hence the better performance of these modules will also help the keyword extraction procedure to provide much better result.

Tseng's algorithm used in [12] to identify keywords from English documents achieved 85.84% accuracy compared to the accuracy of 80.04% by Alchemy API [13] on the same contents. Whereas, for keywords identification from Bengali document achieved maximum accuracy of 87.21%. If we consider more documents to compute the TF-IDF of Bengali words then this accuracy will be much higher to identify the keywords from a given document. So, we can conclude that Tseng's algorithm on top of chi-square measurement of word co-occurrence bias can be used successfully for Bengali keyword identification.

VII. CONCLUSION

By using the proposed procedure explained so far, we will be able to find the Bengali keywords from multiple document collection. As the size of document collection increases, the accuracy with the TF-IDF value to identify important terms from a document also increases. The important terms then help us to determine the biased word set more precisely based on the chi-square distribution. Then the co-occurrence value of these words can be used to identify the keywords with more accuracy. In future, we will combine this statistical approach with large scale Bengali N-grams analysis and word embeddings to improve keyword suggestions in Pipilika search engine [14] [15] [16]. We believe this paper will also help to achieve better results in downstream NLP applications such as identifying trending topics from the daily newspapers or from social media discussion etc.

ACKNOWLEDGMENT

This work has been done using support from Pipilika Search Engine, a research initiative of Shahjalal University of Science and Technology, Sylhet, Bangladesh.

APPENDIX

We present the detail example of keyword identification from a document using the proposed method (Figure 5 - 8).

পশ্চাত্পট ২৬ মার্চ, ১৯৭১, রাত ১২টা। ‘দোজখের সব কটি ফটক খুলে দেওয়া হলো। যখন প্রথম গুলিটি ছোড়া হলো, শেখ মুজিবুর রহমানের ক্ষীণ কণ্ঠস্বর ভেসে এল পাকিস্তানের সরকারি বেতারের কাছাকাছি তরঙ্গ দৈর্ঘ্যে। এটি নিশ্চয়ই এবং সে মতোই শোনা গেল পূর্ব টেপরেকর্ডকৃত বাণী–শেখ সাহেব পূর্ব পাকিস্তানকে গণপ্রজাতন্ত্রী বাংলাদেশ হিসেবে ঘোষণা করছেন। ঘোষণাটি পূর্ণ পাঠ ভারতের বৈদেশিক মন্ত্রণালয় কর্তৃক “বাংলাদেশ ডকুমেন্টস” -এ প্রকাশিত হয়েছে। সেটায় (ইংরেজি ভাষায়) বলা হয়েছিল: এটা সম্ভবত আমার শেষ বার্তা। আজ থেকে বাংলাদেশ স্বাধীন। আমি বাংলাদেশের জনগণকে আহ্বান জানাই, যে স্থানেই আপনারা থাকুন, এবং যা কিছু আপনার আছে, তাই নিয়ে শেষ পর্যন্ত দখলদার বাহিনীকে প্রতিরোধ করুন। আপনাদের লড়াই চলতে থাকবে যে পর্যন্ত না পাকিস্তানের দখলদার বাহিনীর শেষ সৈন্য বাংলাদেশের মাটি থেকে বিতাড়িত হয় ও চূড়ান্ত বিজয় অর্জিত না হয়।’ (সিদ্দিক সালিক, উইটনেস টু সারেন্ডার, করাচি, ১৯৭৭)। জন্মকথা মুজিবনগর, বাংলাদেশ। ১০ এপ্রিল, ১৯৭১। একটি প্রকাশ্য অনুষ্ঠানে পাঠ করা হলো ‘স্বাধীনতার ঘোষণাপত্র’ যার অনূদিত তিনটি অনুচ্ছেদের উদ্ধৃতি দিচ্ছি: ‘ঘোষণা দিতেছি ও প্রতিষ্ঠা করিতেছি যে, বাংলাদেশ হইবে সার্বভৌম জনগণের প্রজাতন্ত্র এবং এতদ্বারা ইতঃপূর্বে বঙ্গবন্ধু শেখ মুজিবুর রহমান কর্তৃক ঘোষিত স্বাধীনতার ঘোষণাকে নিশ্চিত করিতেছি, এবং এতদ্বারা নিশ্চিত করিতেছি ও সিদ্ধান্ত লইতেছি যে, সংবিধান যে সময় পর্যন্ত প্রণীত না হয়, বঙ্গবন্ধু শেখ মুজিবুর রহমান প্রজাতন্ত্রের রাষ্ট্রপতি থাকিবেন ও সৈয়দ নজরুল ইসলাম প্রজাতন্ত্রের উপরাষ্ট্রপতি থাকিবেন, ইংরেজি ভাষায় লিখিত আমরা আরও সিদ্ধান্ত গ্রহণ করিতেছি, স্বাধীনতার এই ঘোষণাপত্র ১৯৭১ সালের ২৬ মার্চ হইতে কার্যকর হইয়াছে বলিয়া বিবেচিত হইবে।’ অতঃপর স্বাধীনতার ঘোষণাপত্রের ধারাবাহিকতায় রাষ্ট্রপতি শেখ মুজিবুর রহমান কর্তৃক স্বাক্ষরিত ‘বাংলাদেশের সাময়িক সংবিধান’ ১৯৭২ সালের ১১ জানুয়ারি জারি হয় এবং তার ৪ অনুচ্ছেদে বলা হয়, ইতিপূর্বে বাংলাদেশের এলাকা থেকে (পাকিস্তানের) জাতীয় সংসদে ও (পূর্ব পাকিস্তানের) প্রাদেশিক সংসদে নির্বাচিত সদস্যদের সমন্বয়ে গণপরিষদ গঠিত হবে। ১০ মার্চ বাংলাদেশের সংবিধানের খসড়া প্রণয়নের উদ্দেশ্যে ৩৪ জন সাংসদের একটি কমিটি গঠন করা হয়। সর্বশেষ গণপ্রজাতন্ত্রের সংবিধান প্রণয়ন ও গৃহীত হয় যার উল্লেখ সংবিধানের ‘প্রস্তাবনা’ ভাগে করা হয়: ‘এতদ্বারা আমাদের এই গণপরিষদে, অদ্য তেরশত ঊনআশী বঙ্গাব্দের কার্তিক মাসের আঠার তারিখ, মোতাবেক ঊনিশ শত বাহাত্তর খ্রীস্টাব্দের নভেম্বর মাসের চার তারিখে, আমরা এই সংবিধান রচনা ও বিধিবদ্ধ করিয়া সমবেতভাবে গ্রহণ করিলাম।’ ….. continue

Fig. 5. Snippet of the documents from which the keywords have been extracted using the proposed procedure

শেখ মুজিবুর, সরকারি, শেখ সাহেব, বাংলাদেশ, গণপ্রজাতন্ত্রী, ঘোষণা, বৈদেশিক মন্ত্রণালয়, বাংলাদেশ ডকুমেন্টস, বাংলাদেশ স্বাধীন, স্বাধীন, মাটি, মুজিবনগর, স্বাধীনতার, স্বাধীনতার ঘোষণাপত্র, প্রজাতন্ত্র, বঙ্গবন্ধু শেখ মুজিবুর রহমান, বঙ্গবন্ধু, রাষ্ট্রপতি, বাংলাদেশের সাময়িক সংবিধান, প্রস্তাবনা, সংবিধান, সংবিধান সংশোধন, বিধান, মৌলিক অধিকার, রাষ্ট্র, মূলনীতি, রাষ্ট্র, অনুচ্ছেদ, সমাজতন্ত্র, সংশোধন, আপিল বিভাগ, সুপ্রিম কোর্ট, কোর্ট, সুপ্রিম, বিভাগ, প্রজ্ঞাপন, জনগণ, সামরিক আইন প্রজ্ঞাপন, বেআইনি, বাতিল, ঘোষণাপত্র, সংবিধান, প্রজাতন্ত্রের রাষ্ট্রধর্ম ইসলাম, হাইকোর্ট, সার্বভৌম, গণপরিষদ, সংবিধানের খসড়া, অধিকার, ধ্রুবতারা, সংবিধানের ধ্রুবতারা, প্রজাতন্ত্রের সর্বোচ্চ আইন, আইন, রায়, অনৈতিক ও বিশ্বাসহন্তা যুদ্ধ, নির্বাচিত, অধিকার, ইংরেজি ভাষা

Fig. 6. Keywords extracted manually

বিভাগ, রাষ্ট্র, হইবে, স্বৈরশাসক, সম্মতিদান, ঘোষণাপত্র, শেখ, ক্ষমতা, গৃহীত, সামরিক, বিধান, বাংলাদেশ, অনুচ্ছেদ, সংবিধানের প্রস্তাবনা, রাষ্ট্রপতির নিকট, মূলনীতি, ধারাবাহিকতায়, অনুচ্ছেদ সংশোধন, আপিল বিভাগ, অতঃপর, সুপ্রিম, পাঠক, সংশোধন, বলিয়া, বঙ্গবন্ধু, সূত্র, প্রিয়, বিল, কোর্ট, দ্বারা, প্রজাতন্ত্র, তাহা, আইন, রায়, অনুরূপ, রাষ্ট্রপতি, সংবিধান, স্বাধীনতার ঘোষণাপত্র, গ্রহণ, স্বাধীনতার, সামরিক আইন, মুজিবুর, প্রস্তাবনা, প্রজ্ঞাপন,

Fig. 7. Keywords extracted using proposed procedure

শেখ মুজিবুর, সরকারি, শেখ সাহেব, গণপ্রজাতন্ত্রী, বৈদেশিক মন্ত্রণালয়, বাংলাদেশ ডকুমেন্টস, বাংলাদেশ স্বাধীন, মুজিবনগর, বঙ্গবন্ধু শেখ মুজিবুর রহমান, বাংলাদেশের সাময়িক সংবিধান, মৌলিক অধিকার, সমাজতন্ত্র, সুপ্রিম কোর্ট, সামরিক আইন প্রজ্ঞাপন, বেআইনি, বাতিল, প্রজাতন্ত্রের রাষ্ট্রধর্ম ইসলাম, হাইকোর্ট, সার্বভৌম, গণপরিষদ, সংবিধানের খসড়া, অধিকার, ধ্রুবতারা, সংবিধানের ধ্রুবতারা, প্রজাতন্ত্রের সর্বোচ্চ আইন, অনৈতিক ও বিশ্বাসহন্তা যুদ্ধ, নির্বাচিত, অধিকার, ইংরেজি ভাষা,

Fig. 8. Keywords not detected by the proposed procedure

REFERENCES

[1] G. Salton,“Automatic Text Processing,” Addison-Wesley, 1988.

[2] 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.

[3] A. Hulth, “Improved automatic keyword extraction given more linguistic knowledge,” In Proceedings of EMNLP, 2003, pages 216223.

[4] M. Rada and P. Tarau, “Textrank: Bringing order into texts,” In Proceedings of EMNLP 2004 (ed.Lin Dand WuD), pp.404411., Association for Computational Linguistics, Barcelona, Spain.

[5] S. Beliga, A. Metrovi, and S. Martini-Ipi, “An overview of graph-based keyword extraction methods and approaches,” Journal of information and organizational sciences 39, no. 1 (2015): 1-20.

[6] C. Zahang, “Automatic keyword extraction from documents using conditional random fields,” Journal of Computational Information Systems 4, no. 3 (2008): 1169-1180.

[7] T. Tomokiyo and M. Hurst, “A language model approach to keyphrase extraction,” Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18. Association for Computational Linguistics, 2003.

[8] Y. Matsuo and M. Ishizuka, “Keyword extraction from a single document using word co-occurrence statistical information,” International Journal on Artificial Intelligence Tools, 2004, 13(1): 157-170.

[9] Y. Tseng, “Multilingual keyword extraction for term suggestion,” ACM New York, NY, USA, 1998, 377-378.

[10] “Term frequency and inverse document frequency,” Online: http://en.wikipedia.org/wiki/Tf*idf, Accessed on Augest 31, 2018.

[11] “Chi-square distribution,” Online: http://en.wikipedia.org/wiki/Chi-squared_distribution, Accessed on Augest 31, 2018.

[12] M. Mahfuzur Rahman and M. Ruhul Amin, “Language Independent Statistical Approach for Extracting Keywords,” Proceedings of the 4th International Conference on Advances in Electrical Engineering, IEEE, 2017.

[13] “Keyword Extraction demo of Alchemy API,” Online: https://github.com/AlchemyAPI, Accessed on Augest 31, 2018.

[14] A. Ahmad and M. Ruhul Amin, “Bengali word embeddings and it’s application in solving document classification problem,” Proceedings of the 19th International Conference on Computer and Information Technology, IEEE, 2016.

[15] A. Ahmad, M. Rub Talha, M. Ruhul Amin and F. Chowdhury, “Pipilika N-gram Viewer: An Efficient Large Scale N-gram Model for Bengali,” Proceedings of the International Conference on Bangla Speech and Language Processing(ICBSLP), IEEE, 2018.

[16] A. Ahmad, M. Rub Talha, M. Ruhul Amin and F. Chowdhury, “Bengali Document Clustering using Word Movers Distance,” Proceedings of the International Conference on Bangla Speech and Language Processing(ICBSLP), IEEE, 2018.