Automatic Keyword Extraction from Bengali Text using Improved RAKE Approach

Mozammel Haque
Dept. of Computer Science and Engineering
Britannia University
Cumilla, Bangladesh
bappy.mozammel@gmail.com

***Abstract*—Keyword extraction refers to the identification of words or short phrases that concisely describe the contents of the document. Rapid Automatic Keyword Extraction (RAKE) is a well-known keyword extraction approach. But we found that RAKE fails to extract the significant Bengali keywords. In this paper, we have proposed an improved version of the pristine RAKE called RAKEB. We have also shown that RAKEB works significantly well for Bengali than the pristine RAKE.**

Keywords— RAKE, Keyword Extraction, Bengali, RAKEB, Natural Language

I. INTRODUCTION

Keyword extraction is the automatic identification of words or short phrases that concisely describe the contents of a text [1]. Rapid Automatic Keyword Extraction (RAKE) is one of the most popular and well-known keyword extraction models.

Stuart Rose et al. had developed the RAKE in 2000 [2]. This approach is now a widely used NLP technique for English and similar structure language. In 2018, S. Siddiqi and A. Sharan found some weakness of RAKE for Hindi language and suggest few modified scoring techniques [3].

The grammatical structure of Bengali is different from English as well as very complicated [4]. We found that the pristine RAKE model fails to extract keywords from Bengali because the consequential smaller keywords would unavoidably have lower scores than longer keywords. As a result, the original algorithm needs to be amended for extracting keywords from Bengali.

In this paper, we have proposed an amended version of RAKE called RAKEB (Rapid Automatic Keyword Extraction for Bengali) that is felicitous for the keyword extraction from Bengali. We have used a list of 398 Bengali stopwords for the experiment of both RAKE and RAKEB [5]. We have analyzed both approach of RAKE as well as RAKEB and show that RAKEB is a better approach than RAKE for the Bengali.

The rest of the paper is organized as follows. Section II introduces the structure of RAKE model as well as explains the motivation for this study. The proposed approach has been presented elaborately in Section III. In Section IV, We present the experimental results and its analysis. Section V includes limitation and our future plan. We conclude our proposed work in Section VI.

II. RAKE DESCRIPTION

Keywords are words or short phrases that concisely describe the contents of a document. RAKE is an automatic keyword engendering approach. The algorithm is as follows [6].

1. Split the text document into a list of words by breaking it at word delimiters (like spaces and punctuation).
2. Split the obtained list of words into sequences of contiguous words by breaking each sequence at the stopwords. Each sequence is now called a "candidate keyword".
3. Calculate the "score" of each individual word from the list of candidate keywords.
4. For each candidate keyword, add the word scores of its constituent words to calculate the candidate keyword score.
5. Take the first one-third top scoring candidates from the list of candidates as the final list of keywords.

A. Candidate keyword Scoring using RAKE

A candidate keyword of RAKE may have multiple words and summation of each word score will generate the keyword score. The score of each word is obtained by dividing the degree of the word by frequency of that word. Suppose a word w occurs in five candidate keywords, where w_1, w_2, w_3, w_4, w_5, w_6 and w are the distinct words of those keywords. The candidate keywords are listed in below:

i) w_1 w_2 w
ii) w
iii) w_5 w
iv) w_2 w_3 w w_4
v) w w_6

The entire process of scoring a word is explained in below [3], [6].

978-1-5386-9242-4/18/$31.00 ©2018 IEEE

1. Frequency of w is the occurrence of w occurs in a document. So the frequency of w with respect to above keywords is 1 + 1+1 + 1 +1= 5 i.e. Frequency (w) = 5
2. Degree of the word w measures the number of words with which a particular word w occurs in the candidate keywords. Here, need to count the number of words that occur in candidate keywords containing w, including w itself to find the Degree of word w. So, the degree of word w is 3 + 1+2 + 4 + 2 = 12 i.e. Degree (w) = 12
3. Degree(w)/frequency(w) will provide the score of word w. So, word_score (w) = 12/5= 2.4
4. For a candidate keyword of multiple words, add the word scores of its constituent words to find the candidate keyword score. For instance, score of candidate keyword "w_1 w_2 w_3" is score(w_1 w_2 w_3) = word_score(w_1) + word_score(w_2) + word_score(w_3).

B. Motivation to Modify RAKE

RAKE is a well-known NLP technique, but its application depends on few factors like the language in which the text is written. We modify the RAKE considering following two issues.

Firstly, a single-word keyword unavoidably has lower scores in RAKE than multi-word keyword because there is a simple addition of individual word scores for the multi-word keywords. RAKE generate candidate keywords list from a document by breaking each sequence of the word at the stopwords. But Bengali stopwords may not always found as the word while making a sentence, because it can be adjacent with another word. The following text is the topmost extracted keyword using RAKE from ভাষার-মনীষা [7].

"বিদেশীয় ভাষার সাহায্যে জ্ঞান বিজ্ঞানের চর্চার মতন সৃষ্টিছাড়া প্রথা"

In the above text, no stopword be found as word but can be found as the substring of another word. For instance, in the word "বিজ্ঞানের", the stopword "এর" is adjacent with the word "বিজ্ঞান" and create a new word "বিজ্ঞানের". As a result, RAKE cannot break the text at this stopword but generates high score because of having more words.

Secondly, RAKE ranked the multiple candidate keywords with the same score where each of them consists of same words but not in the same order. For instance, "X Y Z", "Y X Z" and "Z X Y" are the candidate keywords that found from the same text. They all are not equally important as keywords but RAKE generates the same score for all of the above candidate keywords [3].

Table I shows the top ten extracted keywords and the respective score using RAKE from Bengali article "ভাষার-মনীষা". The result shows that RAKE extracts longer keywords rather than an important one. To overcome this problem, we proposed the keyword-length normalization version of RAKE called RAKEB. RAKEB is also suitable to overcome the second limitation of RAKE as we have described above.

TABLE I. TOP TEN EXTRACTED KEYWORDS FROM "ভাষার-মনীষা" USING RAKE

Sl.	Keywords	Score
1	বিদেশীয় ভাষার সাহায্যে জ্ঞান বিজ্ঞানের চর্চার মতন সৃষ্টিছাড়া প্রথা	77.63
2	'পাকিস্তানের রাষ্ট্রভাষা সমস্যা' শীর্ষক প্রবন্ধে শহীদুল্লাহ্ লিখেছেন 'বাংলাদেশের কোর্ট	70.48
3	ফেলেছে পূর্ব পাকিস্তান সাহিত্য সম্মেলনে সভাপতির অভিভাষণে মুহম্মদ শহীদুল্লাহ্	63.49
4	সময় প্রাপ্য তথ্যের অভাবে প্রজ্ঞাবোধের সাহায্যে 'ইনফারেন্স ড্র'	62.00
5	পূর্ব পাকিস্তানের ভাষার আদর্শ অভিধান প্রকল্পের সম্পাদক হিসেবে	59.88
6	রোমান হরফের প্রবর্তনকে মুহম্মদ শহীদুল্লাহ্ অত্যন্ত পশ্চাদ্‌গামী পদক্ষেপ	58.24
7	কয়েকবার অংশ নিয়েছেন বাংলা লিপিপদ্ধতি নিয়েও ভেবেছেন	46.82
8	বাংলা বিশ্ববিদ্যালয়ের প্রধান ভাষার স্থান অধিকার করিবে	46.69
9	কাজের বর্ণনার মধ্য দিয়ে শহীদুল্লাহ্ সম্পর্কে পুরোপুরি	44.31
10	মুসলমান হওয়ার কারণে ঢাকা বিশ্ববিদ্যালয়ের শিক্ষকের চাকরিতে	44.25

III. RAKEB DESCRIPTION

The original RAKE approach fails to extract significant keywords from Bengali. In this paper, we have modified the approach of RAKE and proposed an improved version of RAKE that is called RAKEB. RAKEB is specially designed for the Bengali. The modification is actually done on the scoring measure of keywords and the rest of the steps remain same in RAKEB as like as the RAKE.

A. Candidate keyword Scoring using RAKEB

RAKEB does not rank a candidate keyword by simply adding up the word score as like as the pristine RAKE. The scoring of a candidate keyword is done using (1).

$$KS(K) = \frac{\sum_{w=1}^{N} \frac{D_w}{F_w} * O}{N} \qquad (1)$$

Here,

K = Candidate Keyword

KS = Score of the keyword "K"

D_w = Degree for each word w in "K" which measures the number of words with which a particular word w occurs in the candidate keywords.

F_w = Frequency of each word w in "K" which simply count of the number of times w occurs in the document.

O = Occurrence of the "K" as substring in the document.

N = Number of the words in "K"

Table II shows the top ten extracted keywords and the respective score of the keywords using RAKEB from Bengali article "ভাষার-মনীষা" [7].

TABLE II. TOP TEN EXTRACTED KEYWORDS FROM "ভাষার-মনীষা" USING RAKEB

Keywords (K)	**D_w**	**F_w**	$\sum_{w=1}^{N} \frac{D_w}{F_w}$	**O**	**N**	$\frac{KS(K) = \sum_{w=1}^{N} \frac{D_w}{F_w} * O}{N}$
শহীদুল্লাহ্	77	16	4.81	18	1	86.63
ভাষা	10	5	2	26	1	52
বাঙালি	20	5	4	9	1	36
মুসলমান	12	3	4	8	1	32
মুহম্মদ শহীদুল্লাহ্	38, 77	7, 16	10.24	6	2	30.72
রাষ্ট্রভাষা	28	4	7	4	1	28
দীর্ঘ	16	2	8	2	1	16
বাংলার	15	4	3.75	4	1	15
সাহিত্য	14	3	4.67	3	1	14
মানুষ	7	2	3.5	4	1	14

B. Potency of RAKEB

RAKE produces high degree score for a multi-word keyword. The RAKEB normalizes this length issue and finds the more frequent and significant keywords using proposed (1). RAKEB solves both limitations of RAKE, as we have described in Section II.B

Firstly, RAKEB does not rank a candidate keyword by simply adding up the score of its word as like as RAKE rather it finds the average score for a multi-word keyword. Furthermore, the occurrence of the candidate keyword (O) helps to find out more frequent keywords in RAKEB, where pristine RAKE unable to do this. For instance, the keyword "বিজ্ঞান" will surely get more score for the frequency of all its different form like "বিজ্ঞানের".

Secondly, it is evident that candidate keywords "X Y Z", "Y X Z" and "Z X Y" are not equally important. But RAKE ranked them equally in this case. RAKEB solve this problem and produce the distinct score for each of the above keywords except each of them present in the document equally because RAKEB scored a keyword along with its occurrence (O).

IV. EXPERIMENTAL RESULT

The proposed approach is implemented and tested using Visual C# programming language on a laptop (Visual Studio 2012 [Windows Forms Application], 64-bit windows 7 OS, 2.4 GHz CPU, 4 GB RAM). Total four newspaper articles are used for testing. The articles are collected from the Bengali newspaper "Prothom Alo". These articles are listed in below:

1. ভাষার-মনীষা [7]
2. তিমিরবিনাশী-সংগ্রাহক [8]
3. অমৃতের পুত্র [9]
4. বাঙালি সংস্কৃতির প্রকৃত সাধক [10]

Both RAKE and RAKEB split the obtained list of words (from the document) into sequences of contiguous words by breaking each sequence at the stopwords to create candidate keywords. We have used a list of 398 Bengali stopwords for this purpose [5].

Table III shows the top ten keywords using both RAKEB and RAKE model. The result clearly shows that RAKE extracts lengthy text as the keyword because of not using the length-normalization method. Furthermore, most of them are insignificant. On the other hand, RAKEB works significantly well for Bengali. RAKEB is able to extract frequent, meaningful as well as significant keywords, because of having length normalization and improved-scoring architecture.

TABLE III. RESULT COMPARISON BETWEEN RAKE AND RAKEB

Article Name	**Keyword Extraction using RAKEB (Proposed Approach)**	**Keyword Extraction using original RAKE**
ভাষার-মনীষা	শহীদুল্লাহ্, ভাষা, বাঙালি, মুসলমান, মুহম্মদ শহীদুল্লাহ্, রাষ্ট্রভাষা, দীর্ঘ, বাংলার, সাহিত্য, মানুষ **[Note:** comma separates the keywords**]**	বিদেশীয় ভাষার সাহায্যে জ্ঞান বিজ্ঞানের চর্চার মতন সৃষ্টিছাড়া প্রথা, ‘পাকিস্তানের রাষ্ট্রভাষা সমস্যা’ শীর্ষক প্রবন্ধে শহীদুল্লাহ্ লিখেছেন ‘বাংলাদেশের কোর্ট, ফেলেছে পূর্ব পাকিস্তান সাহিত্য সম্মেলনে সভাপতির অভিভাষণে মুহম্মদ শহীদুল্লাহ্, সময় প্রাপ্য তথ্যের অভাবে প্রজ্ঞাবোধের সাহায্যে ‘ইনফারেন্স ড্র’, পূর্ব পাকিস্তানের ভাষার আদর্শ অভিধান প্রকল্পের সম্পাদক হিসেবে, রোমান হরফের প্রবর্তনকে মুহম্মদ শহীদুল্লাহ্ অত্যন্ত পশ্চাদ্‌গামী পদক্ষেপ, কয়েকবার অংশ নিয়েছেন বাংলা লিপিপদ্ধতি নিয়েও ভেবেছেন, বাংলা বিশ্ববিদ্যালয়ের প্রধান ভাষার স্থান অধিকার করিবে, কাজের বর্ণনার মধ্য দিয়ে শহীদুল্লাহ্ সম্পর্কে পুরোপুরি, মুসলমান হওয়ার কারণে ঢাকা বিশ্ববিদ্যালয়ের শিক্ষকের চাকরিতে **[Note:** comma separates the keywords**]**
তিমিরবিনাশী-সংগ্রাহক	ভাষা, আবদুল করিম, পুঁথির, সংস্কৃতি, চট্টগ্রাম, বাংলা সাহিত্য, মাতৃভাষা, জাতীয় ভাষা, গ্রামের, মাদ্রাসা	অসংখ্য কাহিনি কেচ্ছা গীত গাথা পালার মূল্য সাহিত্যের ইতিহাসের দিক, ‘প্রায় ৪০০ বছরের সাহিত্যিক নিদর্শন বাংলা সাহিত্যের ইতিহাসে স্থান পেয়েছে, সংস্কৃতির যথার্থ ইতিহাস রচনায় আবদুল করিম সাহিত্যবিশারদের (১৮৭১ ১৯৫৩) উত্তরাধিকার, ডেকে পাঠানোর আশ্বাস দিলে ‘ক্ষিতিমোহন গরম হয়ে গেলেন অবাকও হয়েছিলেন, দক্ষতার গুণেই দ্রুত এতকালের অপাঙ্‌ক্তেয় তামাদি সৃষ্টির পুনরুজ্জীবন ঘটল, ’ (সাহিত্যিক মাহবুব উল আলম) গ্রামের দরিদ্র গৃহস্থটি কীভাবে ব্যক্তিমাত্র, ১৯৫১ সালে চট্টগ্রামে অনুষ্ঠিত সংস্কৃতি সম্মেলনে মূল সভাপতির ভাষণে, মধ্যযুগের দেড়শতাধিক কবিকে আবিষ্কারের কৃতিত্ব তাঁর’ (ড মাহবুবুল হক), ‘১৬ই মার্চ ১৯৫১ সালের চট্টগ্রাম সংস্কৃতি সম্মেলনে প্রদত্ত সাহিত্যবিশারদ, আবদুল করিম আজীবন নিরবচ্ছিন্নভাবে হাতে লেখা পুরোনো পুঁথি সংগ্রহ
অমৃতের পুত্র	জাহিদ ভাই, মন, কথা, বল, যায়, জাহিদ ভাইয়ের, মানে, পাগল, মুনমুন আপা, বলছেন	শতছিন্ন কাপড়চোপড় গালভর্তি দাড়ি গোঁফ জট পাকানো চুল আপনমনে বিড়বিড়, বেরিয়ে একটা চাকরিতে ঢুকেছেন জাহিদ ভাই শেষ বর্ষের পরীক্ষার, জায়গায় একটা ট্রাক রং সাইড দিয়ে আসছিল দ্রুতবেগে, ছাড়া একটু কান পাতলেই শোনা যায় পাগলরা কেবল বর্তমান, পড়ে শুনিয়েছিলেন কয়েকটি লাইন ‘জীবিতের শোক মৃতরা গ্রহণ, বাড়িওয়ালার ভাড়া বাকি পড়ছে জাহিদ ভাই বাসায় তালা মেরে, বড়জন মানে শাহেদ ভাই পড়তেন প্রকৌশল বিশ্ববিদ্যালয়ে স্থাপত্যবিদ্যায়, ডাস্টবিনের পাশের ফুটপাতে দাঁড়িয়ে বক্তৃতার ভঙ্গিতে চিৎকার, নিষেধাজ্ঞার সময় সভয়ে সরে দাঁড়ানোর সময় কথা বলার সময়, বেড়াতে লাগলেন—একাই প্ল্যাকার্ড লিখে দাঁড়াতে লাগলেন প্রেসক্লাবের
বাঙালি সংস্কৃতির প্রকৃত সাধক	সংস্কৃত, বাংলার, হিন্দু, বাংলা সাহিত্যের, পূর্ববঙ্গ, বাংলা ভাষা, অনুরাগের, কুমিল্লা, ভাষার	প্রকারে প্রবেশ লাভ করিল ব্রাহ্মণগণ ইহাকে কিরূপ ঘৃণার চক্ষে দেখিতেন, গোঁড়া হিন্দু সমাজের উৎপীড়নে ইহারা স্বতঃপ্রবৃত্ত হইয়া ইস্লামের আশ্রয় গ্রহণ, সাহিত্য (১৮৯৬) রচনাকালে দীনেশচন্দ্র সেনছিলেন কুমিল্লা ভিক্টোরিয়া স্কুলের প্রধান শিক্ষক, প্রতিনিয়ত সংগ্রামমুখর তেমনি সহজ সরল জটিলতামুক্ত উদার চেতনায় স্বচ্ছ, প্রাচীন পুঁথি আবিষ্কারের কঠিন শ্রমে ব্রতী হওয়ার প্রেরণা জোগায়, উনিশ শতকের শেষার্ধে বাঙালির জাতীয় মানসে উপনিবেশবাদী চিন্তার বিপরীতে, নিম্নবর্গীয় বাঙ্গাল জীবনের সমৃদ্ধ সংস্কৃতি বাঙালি সংস্কৃতি নিয়েই গর্ব অনুভব, প্রাচীন বাঙ্গলা সাহিত্যে মুসলমানের অবদান (১৯৪০) শীর্ষক গ্রন্থ, বাংলার পূর্ব অঞ্চলে অধিক হারে নিম্নবর্গীয় অনার্য জনগোষ্ঠীর বসবাসসূত্রে, পেরেছে বাংলা সাহিত্যের সবচেয়ে ধর্মাচ্ছন্নতামুক্ত মানবিক প্রেমের আখ্যানমূলক গীতিকাসমূহ

Table IV, Table V and Table VI show the score of the top ten extracted keywords using RAKEB from Bengali text.

TABLE IV. EXTRACTED KEYWORDS FROM তিমিরবিনাশী-সংগ্রাহক

Keyword (K)	$\sum_{w=1}^{N} \frac{D_w}{F_w}$	**O**	**N**	**KS(K)**
ভাষা	4.2	22	1	92.4
আবদুল করিম	16.78	9	2	75.5
পুঁথির	7.17	6	1	43
সংস্কৃতি	6.33	6	1	38
চট্টগ্রাম	5.67	5	1	28.33
বাংলা সাহিত্য	14.25	3	2	21.38
মাতৃভাষা	3	5	1	15
জাতীয় ভাষা	7	4	2	14
গ্রামের	4.67	3	1	14
মাদ্রাসা	6.5	2	1	13

TABLE V. EXTRACTED KEYWORDS FROM অমৃতের পুত্র

Keyword (K)	$\sum_{w=1}^{N} \frac{D_w}{F_w}$	**O**	**N**	**KS(K)**
জাহিদ ভাই	8.44	36	2	151.94
মন	2.2	41	1	90.2
কথা	3.1	23	1	71.19
বল	1	62	1	62
যায়	3.5	16	1	56
জাহিদ ভাইয়ের	7.88	12	2	47.29
মানে	3.9	11	1	42.9
পাগল	3	14	1	42
মুনমুন আপা	10.57	7	2	37
বলছেন	3.25	9	1	29.25

TABLE VI. EXTRACTED KEYWORDS FROM বাঙালি সংস্কৃতির প্রকৃত সাধক

Keyword (K)	$\sum_{w=1}^{N} \frac{D_w}{F_w}$	**O**	**N**	**KS(K)**
সংস্কৃত	4.5	11	1	49.5
বাংলার	5.13	8	1	41
হিন্দু	5.67	4	1	22.67
বাংলা সাহিত্যের	10.45	4	2	20.9
পূর্ববঙ্গ	3.33	6	1	20
বাংলা ভাষা	8.7	4	2	17.4
অনুরাগের	7.5	2	1	15
কুমিল্লা	5.5	2	1	11
ভাষার	2.33	4	1	9.33

V. LIMITATIONS AND FUTURE WORK

Our proposed RAKEB is designed to extract the keyword from Bengali Text. We do not find this is very problematic. This Improved approach works well in the large document rather than in a very small document, but accuracy in the small document is much better than the original RAKE.

In (1), "O" is a substring of the given document. In this case, few non-significant keywords get more score than significant keywords, but the amount is negligible as this frequency helps to find out informative and significant keywords in the top list.

RAKEB is the keyword scoring-measure modification of RAKE which is specially designed for the Bengali. The original RAKE does not work well for Bengali. We have a plan to modify the approach of RAKE so that we can extract the keyword from a large amount of language using a unique RAKE model.

VI. CONCLUSION

RAKE is an automatic keyword extracting approach. The original RAKE algorithm fails to extract the keyword from Bengali as the structure of Bengali is very complicated and much different from English. In this paper, we have proposed a modified version of the RAKE called RAKEB which is specially designed for the Bengali.

We have used four Bengali articles for experiments and shown the experimental results for both of RAKE and RAKEB. It shows that RAKEB works significantly well for Bengali than the original RAKE. We hope that RAKEB will be useful in the various field of computational linguistics.

REFERENCES

[1] S. Beliga, M. Ana, and S. Martinčić-Ipšić, "An Overview of Graph-Based Keyword Extraction Methods and Approaches," *J. Inf. Organ. Sci.*, vol. 39, no. 1, pp. 1–20, 2015.

[2] S. Rose, D. Engel, and N. Cramer, "Automatic Keyword Extraction from Individual Documents," in *Text Mining: Applications and Theory*, 2010, pp. 1–20.

[3] S. Siddiqi and A. Sharan, "Improved RAKE Models to Extract Keywords from Hindi Documents," *Inf. Syst. Des. Intell. Appl. Adv. Intell. Syst. Comput.*, pp. 472–483, 2018.

[4] M. Haque and M. N. Huda, "Relation between Subject and Verb in Bangla Language : A Semantic Analysis," *Int. Conf. Informatics, Electron. Vis.*, pp. 41–44, 2016.

[5] "Bengali stopwords collection." [Online]. Available: https://github.com/stopwords-iso/stopwords-bn/blob/master/stopwords-bn.txt. [Accessed: 22-Jun-2018].

[6] "Keyword Extraction using RAKE." [Online]. Available: https://codelingo.wordpress.com/2017/05/26/keyword-extraction-using-rake/. [Accessed: 21-Jul-2018].

[7] "ভাষার মনীষা" [Online]. Available: http://www.prothomalo.com/special-supplement/article/1470321/ভাষার-মনীষা. [Accessed: 24-Jun- 2018].

[8] "তিমিরবিনাশী সংগ্রাহক", [Online]. Available: http://www.prothomalo.com/special-supplement/article/1470331/তিমিরবিনাশী-সংগ্রাহক. [Accessed: 24- Jun- 2018].
[9] "অমৃতের পুত্র", [Online]. Available: http://www.prothomalo.com/special-supplement/article/1470296/অমৃতের-পুত্র. [Accessed: 24- Jun- 2018].
[10] "বাঙালি সংস্কৃতির প্রকৃত সাধক", [Online]. Available: http://www.prothomalo.com/special-supplement/article/1470301/বাঙালি-সংস্কৃতির-প্রকৃত-সাধক. [Accessed: 24- Jun- 2018].