Indian Journal of Science and Technology, *Vol 8(33), DOI: 10.17485/ijst/2015/v8i33/71257, December 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

An Efficient Medical Document Summarization using Sentence Feature Extraction and Ranking

P. Gayathri* and N. Jaisankar

School of Computing Science and Engineering, VIT University, Vellore – 632014, Tamil Nadu, India;
pgayathri@vit.ac.in, njaisankar@vit.ac.in

Abstract

Background/Objectives: As all documents related to medical domain do not come with author written summary, the objective is to introduce a summarizer that exploits medical domain-specific knowledge. **Methods/Statistical Analysis**: Sentence ranking technique has been used to produce high quality summary. The features such as sentence position, length, cue-words (domain-related terms) and acronyms are extracted to assign sentence score. Sentences are ranked and arranged in the decreasing order of their normalized score. The existing summarization approaches in the literature use few or more sentence features but we have opted for few best sentence features. Pre-existing summarizers are used for performance evaluation. **Findings:** The few best features to be considered in developing medical domain-specific summarizers are sentence position, sentence length, number of cue-words and number of acronyms. Summary produced by any summarizer can be highly informative if and only if it contains dissimilar sentences. Therefore, similarity between sentences is an important feature to be considered for creating highly informative summary. The proposed summarizer is compared with the pre-existing summarizers. The evaluation is done by using traditional metrics such as precision and recall and ROUGE. Not all medical documents come with an author written abstract or summary. So, medical documents with author written abstracts are used to test the performance. Results reveals that the proposed summarizer performs better when compared with existing summarizers and attained ROUGE scores also reveals the same with respect to quality of summary produced. Thus, proposed summarizer provide highly acceptable summary to user. **Application/Improvements:** Summarization is one of the information retrieval tasks. It helps to determine whether the retrieved document is relevant for in-depth study or not.

Keywords: Feature Extraction, Medical Document Summarization, Sentence Feature, Sentence Ranking, Summarization

1. Introduction

The amount of information available on the Web in the form of pages is increasing exponentially. Users use this information to be updated with day to day activity, to gain knowledge and for learning purpose. Since so much information is available about a topic, user finds it difficult to select the best option among the available options. This, in turn, is the time consuming process. Even after selecting the best option, the user might face the difficulty in understanding the novelty and validity of the information. The same case is involved with medical information.

The process of producing a compact and concentrated representation of the matter present in a document for human use is known as Document Summarization[1].There are basically two types of summarization techniques based on the type of input provided to the process. First is single document summarization in which only one document is given as the input and the second one in which multiple related documents can be given as an input is called as multi document summarization. The other categorization is based on the nature of the text obtained after summarization, namely, abstractive and extractive summary[2,3].

Document summarization can also be categorized based on the different kinds of user: The first one is

**Author for correspondence*

generic summary, which is meant for broad readership community and the other one is user focused summary, which are made according to the requirements of a particular user or a group of users.

An informative summary is a summary that covers and provides all the important features in the document with some level of detailing. These summarization approaches helps to understand the insights of data in the document, incase if a document does not contain the author-written summary[4,5].

The summarizer proposed in this paper consists of following phases for generation of extractive informative single-document summary to exploit domain-specific knowledge. They are:

1. Pre-Processing.
2. Sentence feature extraction.
3. Sentence score computation and ranking.
4. Final summary creation by using highly dissimilar sentences.

Pre-Processing includes two major activities:

- Sentence Segmentation.
- Stop Word Removal.

The second phase in the process of summarization is sentence feature extraction. The following features have been used:

- Position of the sentence.
- Length of the sentence.
- Number of medical related terms in the sentence.
- Number of medical related acronyms in the sentence.

The third phase is sentence score computation and ranking in which sentences are arranged in the decreasing order of their sentence score. The last phase is final summary creation by using highly dissimilar sentences. The similarity between sentences is determined by using cosine similarity measure. Final summary consist of *N* number of highly dissimilar sentences.

The previous proposed models for document summarization were based on few or all of the features related to any sentence like sentence position, relevancy of sentences with the title of the document, sentence position in the document, term frequencies, standard keywords or cue phrases related to the topic discussed in the document and acronyms[6]. The final scores computed using these above mentioned features are sometime needed to generate a short and efficient summary consisting of only summary worthy sentences using the extractive summary model. The summarization systems provides rank to the sentences of a document based on the similarity to its centroid, position of the sentence in the given document, extent of similarity to the first sentence of the article which is considered to be the summary worthy sentence and finally the length of the sentence[7]. The final values obtained are then normalized between 0 and 1 which are obtained from the linear combination of the sentence features. With the help of a variation in MMR (Maximal Marginal Relevance) algorithm, the redundancy in the document has been removed[8]. This uses a recursive approach to rank the sentences. After the top sentence in the summary has been selected, the other sentences are re-ranked and weights related to the terms are reduced in order to discourage information redundancy in the summary by multiplying its probability with itself.

The summarization techniques that are used in other domains are used as it is for medical domain articles. In the medical domain, MiTAP (MITRE Text and Audio Processing) is one of the software that is used to summarize the medical domain articles by extractive approach[9]. Cluster signature of the document is used to rank the extracted sentences. The Journal of the American Medical Association's articles and abstracts as well as the full texts were used to carry out this project. Another system called TRESTLE (Text Retrieval Extraction and Summarization Technologies for Large Enterprises) is also used to generate single sentence summaries of the newsletters related to pharmaceutical field[10]. It produces summaries with the help of information extraction process to fill out the templates.

The system called HelpfulMed has been designed to help the high-level users and professionals to access the medical domain related information available on the Internet and also in the medical related databases[11]. Another query based medical information system employs an ontology approach to expand the query words and perform scores assignment to the sentences that is completely based on the number of keywords also known as the query words and expanded keywords. The latest system that is used in biomedical document summarization is Bio-Chain that uses an extractive approach and on top of that it employs domain knowledge to build the summary[12]. In comparison to above approaches, we have developed a model for medical document summarization using extractive model of summarizing, sentence feature extraction and ranking method to generate a summary.

2. Proposed System Model

In this section, we present the approach of performing document summarization. The extractive informative single document summarization approach has been used in which the important step is to identify summary worthy sentences from the source document and at the same time reducing the redundancy from the original text so that the final summary generated is relevant to the users. The proposed summarizer undergoes four phases namely, Pre-processing, Sentence feature extraction, Sentence score computation and ranking and Final summary creation by using highly dissimilar sentences. The approach incorporated in the proposed summarizer is presented in Figure 1.

2.1 Pre-Processing

In any natural language processing system, data pre-processing is one of the most important step and it should be carried out before performing any major task in the process. The methods related to data gathering are often loosely handled and this leads to out-of-range values, impossible combinations of data and missing values[13]. Thus, representing the data in a valid form and the quality of data is the first and foremost thing to be done before doing any analysis on it. There are two main activities in this step. They are:

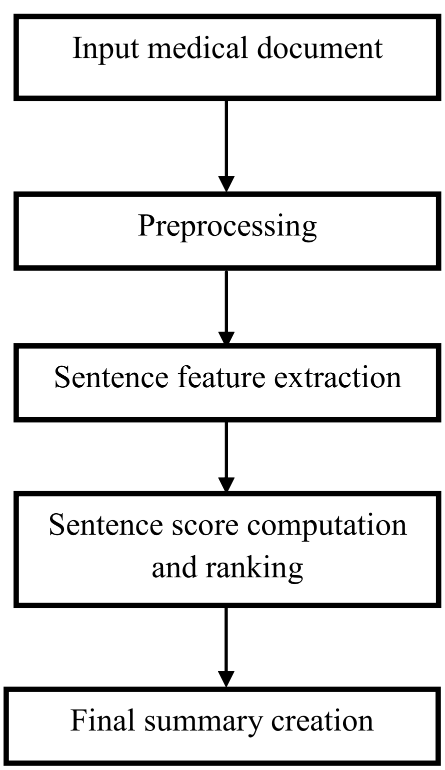

Figure 1. Proposed summarization approach.

2.1.1 Sentence Segmentation

Sentence segmentation is the process of dividing the document into parts based on the delimiter or the boundary detection. The delimiter can be a full stop. For example, consider the document given below:

The heart muscle requires a constant supply of oxygen-rich blood to nourish it. The coronary arteries provide the heart with this critical blood supply. If you have coronary artery disease, those arteries become narrow and blood cannot flow as well as they should?

The above document can be segmented as:

1. The heart muscle requires a constant supply of oxygen-rich blood to nourish it.
2. The coronary arteries provide the heart with this critical blood supply.
3. If you have coronary artery disease, those arteries become narrow and blood cannot flow as well as they should?

2.1.2 Removal of Stop Words

Stop words are the most general words that are used frequently in a sentence and they provide very less meaning to the content of the document. Stop words are maintained in a file for checking like 'a', 'an', 'the', 'above', etc. For example, consider the following sentence:

The heart muscle requires a constant supply of oxygen-rich blood to nourish it.

The stop words present in the above sentence are: the, a, of, to, it. After removing these words, we get the following sentence:

Heart muscle requires constant supply oxygen-rich blood nourish.

2.2 Sentence Feature Extraction

In this step, all the preprocessed sentences are made to go through test that checks the features related to it. We have laid emphasis on four important features because according to the medical domain context, these features are enough to categorize a sentence into "summary worthy" or "not summary worthy" sentence. These four features are described below:

2.2.1 Sentence Position

The sentence location in the document gives the rank to all the sentences according to the equation 1.

$$P_a = \frac{1}{\sqrt{a}} \tag{1}$$

Where, 'a' indicates the position of the sentence in the document[14]. The first sentence is given the highest rank because when a = 1, P_a = 1 and as the number of the sentence increases, the positional value decreases. This feature is applied to preprocessed sentences.

The sentence position in the example document is as follows:

[a=1] heart muscle constant supply oxygen-rich blood nourish.
[a=2] coronary arteries heart critical blood.
[a=3] coronary artery disease, arteries blood flow.

2.2.2 Sentence Length

The sentences that are not summary worthy like journalist names, scientist names, venues, timelines or datelines, historical events, etc., are usually short sentences and thus are not beneficial for the summary. The length of the sentence is calculated by counting the number of words in the sentence. This feature is evaluated for the original sentence i.e. before application of stop word removal. For example, the sentences with their respective word count are shown below:

1. The heart muscle requires a constant supply of oxygen-rich blood to nourish it. **[Count = 13]**
2. The coronary arteries provide the heart with this critical blood supply.**[Count = 11]**

2.2.3 Number of Medical Related Terms

While performing domain related document summarization, the important task is to discover cue phrases and medical related terms because these words are summary worthy. For example, the medical related terms "World Health Organization", "Stem cell therapy" is summary worthy terms[15]. Hence, a sentence which contains highest number of domain specific terms should get higher score than a sentence which contains less number of such terms. In order to identify all such terms, we have maintained a dictionary of words called as domain specific vocabulary which is built up by using MeSH (Medical Subject Headings) which is NLM's (U.S. National Library of Medicine) controlled vocabulary thesaurus. So, if a sentence contains *m* domain related terms, it should get a score of *m*. This feature is applied to preprocessed document. For example, the sentences with their respective word count of domain specific terms are shown below:

1. Heart muscle constant oxygen-rich blood nourish **[Count =6]**
2. Coronary arteries heart critical blood **[Count = 5]**

2.2.4 Number of Medical Related Acronyms

The scoring of sentences is also based on the number of acronyms it contains. In medical domain, acronyms are widely used because of long and complicated words present in it and it helps to memorize them better. Hence, acronyms are one of the features to be considered. All the acronyms are stored in a file and then each of the word is compared with it. If a sentence contains *n* acronyms, then it will get a score of *n*. For example, consider the sentence given below:

Nowadays, IMRT based operations are widely used for CT scan in detection and curing of the cancer tissues in the brain.**[Count=2]**

2.3 Sentence Score Computation and Ranking

The summation of all feature values of a sentence gives its sentence score. Data normalization is the process of reducing data to its canonical form.

The method that we have used here to achieve the task is min-max normalization method. The normalized scores are calculated using equation 2.

$$y' = \frac{y - \min_B}{\max_B - \min_B}(new\_\max_B - new\_\min_B) + new\_\min_B \tag{2}$$

Where, $\min_B$ and $\max_B$ be the minimum and maximum sentence scores of document B. [new_min_B,new_max_B] be the new range, y is the original sentence score and y' is the normalized sentence score.

For example, consider the sentence given below:

The heart muscle requires a constant supply of oxygen-rich blood to nourish it.

The score generated by the algorithm for the above sentence is 20. The scores range is 10.8 to 85.5 which is to be normalized to range of [0.0, 1.0]. Then 20 is mapped as follows by using equation 2.

$((20 - 10.8) / (85.5 - 10.8)) * (1.0 - 0.0) + 0.0 = 0.123$

Like this, normalized sentence score is computed for each sentence in the document. According to normalized sentence score all sentences are arranged in the decreasing order.

2.4 Final Summary Creation

2.4.1 Sentence Similarity Computation

Finding out the similarity between sentences is considered to be one of the most important tasks which have a wide range in many text based applications. The main objective of this system is to generate a summary out of *n* sentences related to medical domain by finding a subset of sentences that are summary worthy and contain the most important information of the document. Thus, removing highly similar sentences from final summary is important in order to create more informative summary. The similarity between two sentences has been calculated using cosine similarity formula which uses the angle of vectors of the two sentences[16]. Let S_a and S_b be the two sentences, then the calculations are performed using equation 3.

$$sim_{\cos}(S_a, S_b) = \frac{\sum_{c=1}^{m} w_{ac} w_{bc}}{\sqrt{\sum_{k=1}^{m} w_{ac}^2 \cdot \sum_{k=1}^{m} w_{bc}^2}} \quad (3)$$

Where, m is the total number of terms, w_{ac} refers to the weight of the term c in the sentence S_a and w_{bc} is the weight of the term c in the sentence S_b.

Weight of term c in any sentence s can be computed by using equation 4.

$$W_{sc} = TF_{sc} * IDF_c \quad (4)$$

Where, TF_{sc} is the number of occurrences of the term c in the sentence s and IDF_c is the inverse document frequency of term c. IDF value of a term is computed using equation 5.

$$IDF_c = \log (N/sf) \quad (5)$$

Where, N is the number of sentences in the document and sf is the number of sentences that contain the term c.

For example, consider the following two sentences:

1. IMRT based operations are widely used for CT scan in detection and curing of the cancer tissues in the brain.
2. To detect and cure the cancer tissues in the brain, CT scan based on IMRT operations are used nowadays.

The above two sentences means the same but they are structured differently, this type of similar sentences can be avoided in final summary by using the cosine similarity method.

2.4.2 Summary Generation

The final summary can have *N* sentences. The value of *N* depends on the compression ration set by the user. As the length of the summary is restricted, there is a need to include highly dissimilar sentences in order to create more informative summary. We have computed similarity between sentences by using the procedure discussed in sub-section 2.4.1.

The dis-similarity between sentence x and y is computed by using equation 6.

$$Dis\_Sim(x, y) = 1 - Sim(x, y) \quad (6)$$

If dis-similarity value exceeds the configurable threshold, then sentences are determined as highly dis-similar sentences.

The steps for final summary creation are given below[27]:

Input: Ordered ranked sentences.
Output: Final summary with N number of sentences.

Step 1: Include top ranked sentence in the final summary.
Step 2: Choose next sentence from the ordered list and compute dis-similarity between chosen sentence and sentences that are already there in the final summary.
Step 3: If dis-similarity value exceeds the configurable threshold, add the chosen sentence in the final summary.
Step 4: Continue step 2 and 3 until *N* number of sentences are included in the final summary.

Likewise, final summary is created in the proposed system. In order to increase the readability of the final summary, sentences are arranged in the order in which they appear in the original document.

3. Results and Discussion

For our summarization system evaluation, 100 medical articles taken from medical information sources like Medical News Today[17], MedlinePlus[18] and PubMed[19] has been used.

For every system generated summary, reference summary is required for its evaluation because the summary generated might not be the interested one. The interestingness of the summary varies from human to human. So, the human generated summary should be considered as one of the reference summary, since it helps them to compare the generated summary to their interestingness. The following pre-existing summarizers are used for comparison purpose and human generated summary is used as reference summary:

- MEAD[20]
- Automatic Text Summarizer(Online Tool)[21]
- SMMRY(Online Tool)[22]
- Tools4Noobs(Online Tool)[23]
- MS WORD-2007

3.1 Traditional Evaluation Metrics

The traditional evaluation metrics used for experimental evaluation is precision and recall.

3.1.1 Precision

Precision is defined as the ratio of total number of system extracted sentences matched with reference summary to the total number of sentences extracted by the system.

3.1.2 Recall

Recall is defined as the ratio of total number of system extracted sentences matched with reference summary to the total number of sentences in the reference summary.

3.1.3 Experimental Results

Table 1 and Table 2 shows the average precision and recall values obtained for summary compression ratio set to 15% and 25% respectively with the human generated extractive summary as reference summary.

3.2 ROUGE Evaluation

System generated summaries are also evaluated by using ROUGE (Recall-Oriented Understudy for Gisting Evaluation)[24]. In 2005, the Information Science Institute at

Table 1. Results for the compression ratio set to 15%

Summarizer	Average precision	Average recall
Proposed System	0.74	0.46
MEAD	0.63	0.35
Automatic Text Summarizer	0.62	0.41
SMMRY	0.60	0.44
Tools4Noobs	0.61	0.40
MS WORD-2007	0.52	0.30

Table 2. Results for the compression ratio set to 25%

Summarizer	Average precision	Average recall
Proposed System	0.72	0.52
MEAD	0.58	0.43
Automatic Text Summarizer	0.59	0.45
SMMRY	0.57	0.48
Tools4Noobs	0.58	0.42
MS WORD-2007	0.53	0.33

the University of Southern California developed ROUGE software. This tool helps to make comparison between system generated summary and reference summary. The older versions of this software were based on n-gram (here n-gram is a sequence of *n* consecutive words in a summary sentence) overlap between the reference summaries and automated system-produced summaries which in turn, returns separate scores for 1, 2, 3 and 4-gram matching[25].

The newer version of this software uses a recall-based measure to evaluate summaries, which requires the both summaries to be of equal length (length should be specified in terms of words or bytes) i.e. system-produced and reference summaries should be of equal length.

Equation 7 is used to compute Rouge-N score:

$$ROUGE-N = \frac{\sum_{s\in(\text{ReferenceSummaries})} \sum_{gram_n \in S}^{n} Count_{match}(gram_n)}{\sum_{s\in(\text{ReferenceSummaries})} \sum_{gram_n \in S} Count(gram_n)} \quad (7)$$

Where, ROUGE-N is an n-gram recall between a system generated summary and a set of reference summaries, n stands for the length of the n-gram, $gram_n$ and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a system generated summary and a set of reference summaries.

The usage of the previous versions of this software leads to production of the recall score only, this score did not come to as useful condition, to categorize summary as good or bad because any text which leads to high score generation need not to be the best for summary[26].

In present working version of ROUGE 1.5.5, a new precision factor is added to already existing recall factor. Inclusion of this factor helps in better judgment of the summary's acceptance level.

In a summary, the percentage of n-grams in the reference summaries that also occur in the system-generated summary is recall. The percentage of n-grams in the system generated summary that too occurred in the reference summary is precision. The average recall and precision for a system are the averages over all the summaries in the test set when system-generated summary is compared with all the available multiple reference summaries.

For automatic summary evaluation, ROUGE-1, ROUGE-2 and ROUGE-3 metrics have been widely used in the NLP (Natural Language Processing) community[27].

From the proposed summarizer, two tasks are performed for summary evaluation. One is 100 words summary generation and second is 150 words summary generation. While comparing, similar length of system generated summary and reference summary is considered.

3.2.1 Experimental Results

Table 3 and Table 4 shows the average ROUGE scores for the 100 words summary and 150 words summary with the human generated summary as reference summary.

In both the tasks, the ROUGE scores of our proposed summary is greater than that of other pre-existing summarizers. This shows the higher acceptance level of our system-generated summary.

Table 3. ROUGE scores for 100 words summary

Summarizer	ROUGE-1 Score	ROUGE-2 Score	ROUGE-3 Score
Proposed System	0.53	0.39	0.25
MEAD	0.47	0.31	0.15
Automatic Text Summarizer	0.50	0.36	0.20
SMMRY	0.47	0.31	0.21
Tools4Noobs	0.51	0.37	0.23
MS WORD-2007	0.40	0.28	0.12

Table 4. ROUGE scores for 150 words summary

Summarizer	ROUGE-1 Score	ROUGE-2 Score	ROUGE-3 Score
Proposed System	0.55	0.40	0.29
MEAD	0.49	0.34	0.18
Automatic Text Summarizer	0.52	0.37	0.26
SMMRY	0.50	0.34	0.25
Tools4Noobs	0.53	0.36	0.26
MS WORD-2007	0.42	0.29	0.15

4. Conclusion

The document summarization in medical domain on basis of sentence feature extraction has been discussed in this paper. The other summarizers in the literature used few or more sentence features for summary generation. We have opted for few best sentence features. Extractive informative single document summarization of medical document is produced by including highly dissimilar sentences. Thus, proposed summarizer provide high acceptance summary to user. By ROUGE scores, we can say that the proposed approach performs better than the pre-existing summarizers using human generated summaries as reference summaries.

5. References

1. Mani I. Automatic summarization. Book of Natural language processing, Amsterdam/Philadelphia: John Benjamins Publishing Company; 2001; (3).
2. Munot N, Govilkar SS. Comparative study of text summarization methods. International Journal of Computer Applications. 2014; 102(12):33–7.
3. Mendoza M, Bonilla S, Noguera C, Cobos C, Leon E. Extractive single-document summarization based on genetic operators and guided local search. Expert Systems with Applications. 2014 Jul; 41(9):4158–69.
4. Prakash S, Chakravarthy TC, Brindha GR. Preference based quantified summarization of on-line reviews. Indian Journal of Science and Technology. 2014; *7*(11):1788–97.
5. Shri JMR, Subramaniyaswamy V. An effective approach to rank reviews based on relevance by weighting method. Indian Journal of Science and Technology. 2015 Jun; *8*(12).
6. Edmundson HP, New methods in automatic extracting. Journal of the Association for Computing Machinery. 1969; 16(2):264–85.

7. Hovy EH. Automated text summarization. In R. Mitkov (Editor). The oxford handbook of computational linguistics: Oxford: Oxford University Press; 2005. 583–98.
8. Carbonell GJ, Goldstein J. The use of MMR, diversity-based re-ranking for reordering documents and producing summaries. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: Melbourne, Australia; 1998. p. 335–6.
9. Day LD, Hirschman L, Kozierok R, Mardis S, McEntee T. Real users, real data, real problems: The MiTAP system for monitoring bio Events. In the Proceedings of the Conference on Unified Science and Technology for Reducing Biological Threats and Countering Terrorism (BTR 2002): 2002. p. 167–77.
10. Gaizauskas R, Herring P, Oakes M, Beaulieu M, Willett P, Fowkes H. Intelligent access to text: Integrating information extraction technology into text browsers. In Proceedings of the Human Language Technology Conference (HLT 2001): 2001. 189–93.
11. Chen H, Lally A, Zhu B, Chau M. HelpfulMed: Intelligent searching for medical information over the internet. Journal of American Society for Information Science and Technology (JASIST). 2003 Mar; 54(7):683–94.
12. Barzilay R, Elhadad M. Using lexical chains for text summarization. Advances in Automatic Text Summarization. 1999. 111–21.
13. Shinde RD, Routela SH, Jadhav SS, Sagare SR. Enforcing text summarization using fuzzy logic. International Journal of Computer Science and Information Technologies. 2014; 5(6):8276–9.
14. Fattah MA, Ren F. Automatic text summarization. In proceedings of World Academy of Science, Engineering and Technology: 2008 Feb 27. p. 192–5.
15. Salton G. Automatic text processing: The transformation, analysis and retrieval of information by computer: Addison-Wesley Publishing Company; 1989.
16. Achananuparp P, Hu X, Shen X. The evaluation of sentence similarity measures. Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), LNCS: 2008. 5182 p. 305–6. doi: 10.1007/978-3-540-85836-2_29.2008.
17. MediLexicon International Ltd. Available from: http://www.medicalnewstoday.com
18. Medline Plus. Available from: www.nlm.nih.gov/medlineplus/
19. National Library of Medicine PubMed. Available from: www.ncbi.nlm.nih.gov/pubmed/
20. Multi-lingual summarization and evaluation platform. Available from: http://www.summarization.com/mead/
21. AutoSummarizer. NewYork-2015. Available from: http://autosummarizer.com/
22. Elmaani LLC. Available from: http://www.smmry.com
23. Tools 4 noobs. Available from: http://www.tools4noobs.com/summarize/
24. Lin CY. ROUGE working note. Voices. 5(20).
25. Lin CY, Hovy E. Automatic evaluation of summaries uses n-gram co-occurrence. In Proceedings of Language Technology Conference (HLT-NAACL 2003): Edmoton, Canada; 2003. 1 p. 71–8.
26. Sjobergh J. Older versions of the ROUGEeval summarization evaluation system were easier to fool. Information Processing and Management. 2007 Nov; 43(6):1500–5.
27. Sarkar K, Nasipuri M, Ghose S. Using machine learning for medical document summarization. International Jounal of Database Theory and Application. 2011 Mar; 4(1):31–48.