Automatic Document Summarization based on Statistical Information

Aigerim Mussina[1], Sanzhar Aubakirov[1] and Paulo Trigo[2]
[1]*Department of Computer Science, Al-Farabi Kazakh National University, Almaty, Kazakhstan*
[2]*Instituto Superior de Engenharia de Lisboa, Biosystems and Integrative Sciences Institute / Agent and Systems Modeling, Lisbon, Portugal*

Keywords: Summarization, Automatic Extraction, Key-words, N-gram, TextRank.

Abstract: This paper presents a comparative perspective in the field of automatic text summarization algorithms. The main contribution is the implementation of well-known algorithms and the comparison of different summarization techniques on corpora of news articles parsed from the web. The work compares three summarization techniques based on TextRank algorithm, namely: General TextRank, BM25, LongestCommonSubstring. For experiments, we used corpora based on news articles written in Russian and Kazakh. We implemented and experimented well-known algorithms, but we evaluated them differently from previous work in summary evaluation. In this research, we propose a summary evaluation method based on keywords extracted from the corpora. We describe the application of statistical information, show results of summarization processes and provide their comparison.

1 INTRODUCTION

In this work, our goal is to make a research and comparison on summarization algorithms. Automatic summarization is the process of generating a reduced text from document, which will save the idea of original text. There are three main types of automatic summarization processes: (a) extraction-based, (b) abstraction-based, and (c) aided. In this paper we follow an extraction-based approach. It uses parts of the original text, sentences, and construct the short paragraph summary. Extraction-based approach does not make any modifications in the text. Summary construction may be influenced by several features, from syntactics to semantics, but we focus on statistical data, which is a frequency statistics of N-grams. Extraction-based summarization best suited to statistical data. Counting the similarity of text units and units importance is the popular approach in algorithms based on statistical data. Text unit could be a word, sentence or paragraph. We use sentence as a text unit. Similarity depends on the presence of keywords in the sentences. Key-words are words that indicate the topic of the text.

1.1 Related Work

The research focused on previous work related with approaches on paragraph extraction, sentence extraction, definition of position in text of main information, sentence similarity, informative sentence extraction. In the work (Jaruskulchai and Kruengkrai, 2003) presents an algorithm for extracting the most significant paragraphs from a text in Thai, where the significance of a paragraph is considered based on the local and global properties of a paragraph. The main emphasis is on the known correct distribution of paragraphs, since Thai language is very different from European languages and is more like Chinese and Japanese in terms of fuzzy division of words and sentences. In our case, we consider Russian and Kazakh languages, which have a clear sentence structure. The (Mitra et al., 2000; Fukumoto et al., 1997) works propose that each word in text can have weight and depending on this weight it is possible to denote the important part of information. However, article (Fukumoto et al., 1997) uses words weight among a paragraph and the extraction unit in this work is a paragraph. The works (Barrios et al., 2016; Yacko, 2002) mainly depict one view of summarization methods. Authors suppose that each sentence has connection with other sentences and this connection is their similarity.

In work (Barrios et al., 2016) TextRank algorithm presented with different variations of similarity functions. The main feature is denoted in construction of a graph with sentences as vertex(tops) and similarity connections as edges, where each edge has its

Mussina, A., Aubakirov, S. and Trigo, P.
Automatic Document Summarization based on Statistical Information.
DOI: 10.5220/0006888400710076
In *Proceedings of the 7th International Conference on Data Science, Technology and Applications (DATA 2018)*, pages 71-76
ISBN: 978-989-758-318-6

value calculated from similarity function. In work (Yacko, 2002) similarity of sentences defined in common words, sentence with more connections recognized as informative. The way of constructing a graph seems the most preferable since it operates with sentences, and similarity functions use statistical data as word frequency. One of the most important stage described in the work (Page et al., 1998), it is about PageRank. PageRank is an algorithm used in ranking of edges in any graph. In work (Barrios et al., 2016) author used PageRank and domain Random Walker in summary construction.

The summary evaluation process described in (Barrios et al., 2016; Sripada and Jagarlamudi, 2009), they involve usage of ROUGE. Recall-Oriented Understudy for Gisting Evaluation (Lin, 2004) is a set of metrics used in automatically generated summary evaluation and in machine translation. ROUGE evaluation compares "ideal summary" with automatically produced summary. The "ideal summary" generates by human. This research work does not assume interaction with human, therefore we can not use ROUGE. The hypothesis from work (Sripada and Jagarlamudi, 2009) stays that the summary must act as the full document, such that their probability distributions are very close to each other. Authors propose application of KL (Kullback-Leibler) divergence, the calculation of entropy of summary, in evaluation process.

The corpora meta data and dictionary extraction of the subject area are fully examined in our previous work (Mussina and Aubakirov, 2017). Therefore in this work we are focusing on the dictionary based text summary extraction

2 METHODS

The used corpora consists of news articles that were parsed by web-crawler from government and news portals. Texts are in Russian and Kazakh languages. The main themes of texts are floods, earthquakes, storms and other emergency situations. Sometimes they contain not necessary information, for example, long requisites about department or region. Summary can help people to concentrate only on necessary facts without information noise.

2.1 Summarization Techniques

Work (Barrios et al., 2016) describes TextRank automated summarization algorithm with authors modifications. In their work document represents as a graph with sentences as nodes, where edges between nodes show the similarity between sentences. Similarity calculates using different similarity functions. In our research work, we implemented three variations of similarity functions: general, BM25 and Longest common substring. The summary size is equal to the 30% of the original text size.
Formula 1 shows the similarity calculation by the general TextRank version.

$$Sim(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)} \quad (1)$$

, where S - sentence and W - word.

Algorithm 1. General TextRank

1. Extract list of sentences from text.
2. For each sentence i ∈ [0, sentence list size - 1]
3. Extract N-grams of sentence[i]
4. For each sentence j ∈ [i+1, sentence list size]
5. Extract N-grams of sentence[j]
6. Count the number of similar N-grams by formula 1
7. If similarity is greater than 0, add edge between sentences with weight equal to their similarity.

For example, consider work of summarization algorithm on message about earthquake. Since all data is in Russian and Kazakh language, in this paper we will provide machine translation of example text.

Example 1. The machine translation of original text.
"Residents of Shymkent and Taraz felt an earthquake in Afghanistan, Tengrinews.kz correspondent reports with reference to the SI "Seismological Experimental-Methodical Expedition of the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan." Underground tremors were recorded on April 10 at 16:28 on the time of Astana. The epicenter of the earthquake was located on the territory of Afghanistan, 787 kilometers to the south-west from Almaty. The energy class of the earthquake is 14.5. Magnitude - 6,8, depth of occurrence - 20 kilometers. Tremors were felt in Shymkent and Taraz - 3 points. There is no information about the injured and the destruction. Recall, April 9 earthquake of magnitude 4.9 occurred in 141-km from Almaty. Underground tremors were recorded at 23:31 on the time of Astana. The epicenter of the earthquake was 141 km south-east of Almaty on the territory of Kyrgyzstan. Energy class of tremors - 10.2, depth of occurrence - 5 kilometers."

The BM25 variation based on the below formulas:

$$BM25(R,S) = \sum_{i=1}^{n} IDF(S_i) * \frac{f(S_i,R) * (k_1+1)}{f(S_i,R) + k1 * (1 - b + b * \frac{|R|}{avgDL})} \quad (2)$$

where

- IDF inverse document frequency
- $f(S_i,R)$ occurrence frequency of a word i from sentence S in sentence R
- $|R|$ - a length of sentence R
- $avgDL$ average length of sentences in the document
- $k1$ and b are parameters

Values for last parameters we took from work (Barrios et al., 2016), $k1$ = 1.2, b = 0.75 This formula states that if a word appears in more than half of sentences it will cause negative result value. To avoid problems caused by negative value in future work of an algorithm next calculation of IDF was proposed:

$$IDF(S_i) = \begin{cases} \log(N - n(s_i) + 0.5) - \\ -\log(n(s_i) + 0.5) & ,if \quad n(s_i) > \frac{N}{2} \\ \varepsilon * avgIDF & ,if \quad n(s_i) \leq \frac{N}{2} \end{cases} \quad (3)$$

, where ε - between 0.3 and 0.5, we use 0.5

Algorithm 2. BM25

1. Extract list of sentences from text.
2. Calculate IDF for all N-grams and the average length of document sentences.
3. For each sentence i ∈ [0, sentence list size - 1]
4. Extract N-grams of sentence[i]
5. For each sentence j ∈ [i+1, sentence list size]
6. Extract N-grams of sentence[j]
7. Count the sentence similarity by formula 2
8. If similarity is greater than 0, add edge between sentences with weight equal to their similarity

The Longest common substring is the easiest in implementation algorithm, but it also can show the same results as BM25 and General TextRank. For similarity value used length of the longest common substring.

Algorithm 3. Longest common substring

1. Extract list of sentences from text.
2. For each sentence i ∈ [0, sentence list size - 1]
3. Extract N-grams of sentence[i]
4. For each sentence j ∈ [i+1, sentence list size]
5. Extract N-grams of sentence[j]
6. Find out longest common substring. Set its length as similarity value.
7. If similarity is greater than 0, add edge between sentences with weight equal to their similarity.

Table 1: Algorithms' results for example 1.

	General TextRank	BM25	Longest Common Substring
S_1 and S_8	0.278	1.225	13
S_2 and S_6	0.377	1.133	6
S_2 and S_8	0.326	1.225	6
S_2 and S_9	3.239	13.663	30
S_3 and S_4	0.384	1.234	13
S_3 and S_8	0.313	0.814	6
S_3 and S_{10}	1.848	6.168	22
S_4 and S_{10}	0.378	0.769	13
S_4 and S_{11}	1.211	4.11	20
S_5 and S_{11}	1.439	5.48	17
S_6 and S_9	0.403	1.048	6
S_8 and S_{10}	1.855	6.556	15

In this work we do not implement PageRank. We used the idea of symmetric summarization presented in work (Yacko, 2002). A document represents as undirected graph with sentences connected to each other. The edge weight concerns to both sentences. Sentence rank defined as a sum of weights of connected edges. Sentences with highest rank will be in summary. Consider summarization process via General TextRank for text in example 1. Generally, sentences are connected with each other. However, sometimes sentences do not have any common word. In this case we have graph presented in figure 1. Sentence with number 7 does not have any connection with other sentences. Now we have 10 sentences with connections and limit of 4 sentence in summary, therefore all sentences that have connections could not go to summary. To reduce number of sentences we define a threshold value, which is equal to the average value of the weight of all edges, see figure 2. For example 1, threshold value is equal to 1.0045. More sentences now rejected, like sentences with numbers 1, 6 and 7. The pairs that have passed through the threshold are (S2, S9) Sim = 3.239, (S3, S10) Sim = 1.848, (S4, S11) Sim = 1.211, (S5, S11) Sim = 1.439, (S8, S10) Sim = 1.855. We have reduced 7 pairs. Now we will rank each sentences with similarities that they have with other sentences, see figure 3.

1. S2 rank = 3.239 value, since it has only one link with S9
2. S3 rank = 1.848

3. S4 rank = 1.211
4. S5 rank = 1.439
5. S8 rank = 1.855
6. S9 rank = 3.239
7. S10 has two links with sentences S3 and S8, its rank is equal to 3.703
8. S11 also has two links with sentences S4 and S5, its rank is equal to 2.65

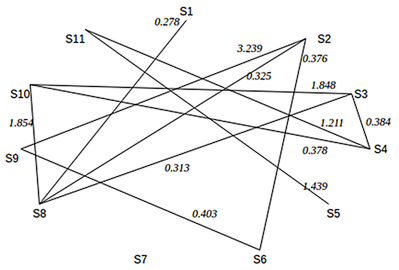

Figure 1: Graph with similarities greater than 0.

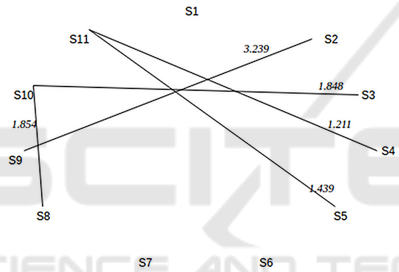

Figure 2: Graph with similarities greater than threshold.

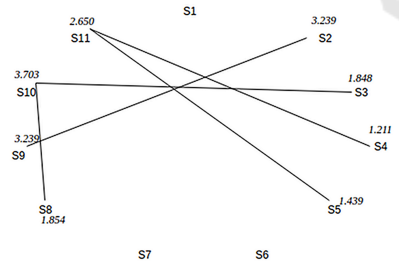

Figure 3: Graph with ranked sentences.

Sentences oredered by rank: S10, S2, S9, S11, S8, S3, S5, S4. The size of original text is 11 sentences, the 30% of 11 is 3.3, we round value up and finally get size of summary of 4 sentences. Sentences with high rank will construct the summary, we get first 4: S10, S2, S9, S11. Then we permute sentences in the order of original text and save summary. Finally, we get the summary depicted below.

Summary for example 1.
"Underground tremors were recorded on April 10 at 16:28 on the time of Astana. Underground tremors were recorded at 23:31 on the time of Astana. The epicenter of the earthquake was 141 km south-east of Almaty on the territory of Kyrgyzstan. Energy class of tremors - 10.2, depth of occurrence - 5 kilometers."

2.2 Summary Evaluation

The evaluation of the summary is based on the idea, proposed in work (Sripada and Jagarlamudi, 2009), that summary probability distribution model must be very close to the original document probability distribution model. Applying to our conditions we can suppose that the key-words distribution in the summary must be bigger than in the original text, because summary must reduce amount of general words and save number of key-words. In (Sripada and Jagarlamudi, 2009) work authors use uni-gram model, but we will use model from 1 up to 5 N-grams. The algorithm of calculation of key-words distribution described below.

Algorithm 4. Key-words distribution.

1. Get document from array of documents.
2. Extract N-grams from text.
3. For each N-gram check if it is in key-words dictionary. Count the sum of matches.
4. Calculate key-words distribution by dividing sum of matches by the amount of N-gram extracted from the text.
5. If there are one more document go to step 1, else calculate average key-words distribution which will describe the summary evaluation for the given TextRank variation function.

The machine translation of original text with underlined key-words:

"Residents of Shymkent and Taraz felt earthquake in Afghanistan, Tengrinews.kz correspondent reports with reference to SI "Seismological experimental-methodological expedition Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan. Underground tremors were recorded on April 10 at 16:28 on the time of Astana. Epicenter of the earthquake located on the territory of Afghanistan, in 787 kilometers to the southwest from Almaty. Energy class of the earthquake 14,5. Magnitude - 6,8, depth of occurrence - 20 kilometers. Tremors were felt in Shymkent and Taraz - 3 points. There is no information about the injured and destruction. Recall, on April 9, an

earthquake of magnitude 4.9 occurred in 141 km from Almaty. Underground tremors were fixed at 23:31 on the time of Astana. The epicenter of the earthquake was in the 141st kilometer on southeast from Almaty on the territory of Kyrgyzstan. Energetic class of underground tremors - 10,2, depth of occurrence - 5 kilometers."

The body evaluation = 0.124
The below summary is identical to all three TextRank techniques: General, BM25, LongestCommonSubstring. In this example text and summary have nearly the same key-words distribution

"*Underground tremors were recorded on April 10 at 16:28 on the time of Astana. Underground tremors were recorded at 23:31 on the time of Astana. The epicenter of the earthquake was in the 141-kilometer to southeast from Almaty on the territory of Kyrgyzstan. Energy class of underground tremors - 10.2, depth of occurrence - 5 km.* "

The summary evaluation = 0.12
The results from each TextRank variation function then compared with each other. The distribution value is normalized and it is between 0 and 1. Probably it could be not equal to 1, because document could not contain only key-words.

3 RESULTS AND DISCUSSION

Table 2 shows the amount of news articles that we have used during summary extraction tests. The average length of article presented in amount of symbols, since sentences and words could be of different length.

Table 2: Source data for summary extraction.

	Amount
Articles	74770
Average article length (in symbols)	1619

From the table 3 we can see that all summarization techniques have reduced the number of general words and the concentration of key-words increased. General TextRank stays as the best technique according to summary evaluation.

During this research work TextRank algorithm variations were tested and estimated. In the (Barrios et al., 2016) work tests show that BM25, with modification of IDF value by formula (3), was the one with better results than General TextRank and Longest common substring. Authors used the database of the 2002 Document Understanding Conference (DUC) and for evaluation used version 1.5.5 of the ROUGE package. Our implementation on corpora of news articles show another results and we have two main possible reasons for that:

Table 3: TextRank variations evaluation results.

	Key-words distribution
Original documents	0.159
General TextRank	0.180
BM25	0.169
LongestCommonSubstring	0.175

1. Corpora without ideal summary
2. Not clear dictionary

The ROUGE package evaluation metric use the reference summary, or ideal summary, and has several techniques. The generation of such reference summary needs human interaction and possibly not interaction of one human, but at least three person. The professional activity of each human candidate also play role.
The alternative way of evaluation process, as was mentioned in sub-section 2.2, Summary evaluation, based on the hypothesis from (Sripada and Jagarlamudi, 2009). Authors used KL (Kullback-Leibler) Divergence which denotes the difference between two probability distributions by formula:

$$D_{KL}(P||Q) = \sum_{i \in w} P(i) \log \left(\frac{P(i)}{Q(i)} \right) \quad (4)$$

where P is probability distribution of original document and Q is a probability of summary. The basic term that used in Kullback-Leibler Divergence is entropy and information gain, but since information gain is an inverse value to entropy we will focus on entropy.
Calculation of summary "emergency" is another one evaluation method that we proposed. The idea based on emergency value of N-grams from dictionary. We suppose that distribution of key-words with high "emergency" is the key criteria of well-constructed summary. However, we have difficulties in normalizing of dustribution values.
In future we would like to continue research on completely different or hybrid summary algorithms to avoid descibed above issues. During tests it was noticed that sometimes not important N-grams repeated in several sentences, which cause that those sentences were written to the summary. The new approach will be based on dissimilarity of sentences. Algorithm is as follows:

1. Group sentences that has common N-grams.
2. Choose sentence with biggest amount of keywords among those that are in one group.
3. Generate summary from sentences that were chosen from previous step.

The used corpora contains information about emergency situations, therefore numerical data is of particular importance. The attention also will be provided to numerical data. Such information will be very helpful for emergency work specialists. The summary should contain such information and presence of it will be used in evaluation process. Finally, the main and most meaningful research should be done in synonyms. Since the basic similarity calculated by presence of common words in two sentences, it is very important to add synonyms dictionary. The sentence S_A may contain word underground tremors and sentence S_B earthquake, meaning of these N-grams mostly equal, but implemented algorithm will not recognize similarity.

4 CONCLUSION

We carried out a study of existing works in the field of the automatic summary extraction. The implemented algorithms were compared and results of this comparison show the practical meaning of this work. The results of summary evaluation mostly matched the comparison described in (Barrios et al., 2016). The General TextRank was the best one, which generates summary with a high distribution of key-words. Its average key-words distribution is equal to 0.180. The easiest in implementation algorithm LongestCommonSubstring has key-words concentration equal to 0.175. The lowest distribution 0.169 belongs to BM25.

Tests showed that the presence of identical words as a definition of the importance of sentences is not suitable for all data. Firstly, it was noticed that unimportant N-grams, repeated in several sentences, lead to summary with those sentences. Probably not all N-grams should participate in sentences similarity calculation. Secondly, synonyms are not taken into account. The sentence S_A may contain word underground tremors and sentence S_B earthquake, meaning of these N-grams mostly equal, but implemented algorithm will not recognize the similarity. However, the addition of synonyms will depend on the existence of a dictionary of synonyms and its completeness. Thirdly,numerical data does not taken into account. The used corpora contains information about emergency situations, therefore numerical data is of particular importance. In future we would like to continue research on completely different or hybrid summary algorithms to avoid descibed above issues.

More research would be done on dictionary extraction, synonyms dictionary, and summary evaluation. Dictionary extraction has more work to be done since it is very important in summary evaluation and all problems should be resolved: stop-words, stemming. In most of the cases, all three algorithms cut off useless information leaving only important part that contains topic keywords.

REFERENCES

Barrios, F., Lpez, F., Argerich, L., and Wachenchauzer, R. (2016). Variations of the similarity function of textrank for automated summarization. In *Proc. Argentine Symposium on Artificial Intelligence, ASAI*.

Fukumoto, F., Suzuki, Y., and Fukumoto, J.-i. (1997). An automatic extraction o f key paragraphs based on context dependency. pages 291–298.

Jaruskulchai, C. and Kruengkrai, C. (2003). A practical text summarizer by paragraph extraction for thai. pages 9–16.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop: Text Summarization Braches Out 2004*, page 10.

Mitra, M., Singhal, A., and Buckley, C. (2000). Automatic text summarization by paragraph extraction.

Mussina, A. and Aubakirov, S. (2017). Dictionary extraction based on statistical data. *Vestnik KazNU*, pages 72–82.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web.

Sripada, S. and Jagarlamudi, J. (2009). Summarization approaches based on document probability distributions. In *PACLIC*.

Yacko (2002). Simmetrichnoe referirovanie: teoreticheskie osnovy i metodika. pages 18–28.

COPYRIGHT FORM