3-бөлім Раздел 3 Section 3

Информатика Информатика Computer science

IRSTI 20.19.19

Automatic document summarization based on statistical information

Mussina A., Al-Farabi Kazakh National University,
Almaty, Republic of Kazakhstan, +77759295274, E-mail: mussina.aigerim95@gmail.com
Aubakirov S., Al-Farabi Kazakh National University,
Almaty, Republic of Kazakhstan, +77002200051, E-mail: aubakirov.sanzhar@gmail.com
Ahmed-Zaki D., Al-Farabi Kazakh National University,
Almaty, Republic of Kazakhstan, +77772469374, E-mail: darhan.ahmed-zaki@kaznu.kz
Trigo P., Instituto Superior de Engenharia de Lisboa,
Lisbon, Portugal, E-mail: ptrigo@deetc.isel.ipl.pt

Actual problem in nowadays is to efficiently process the large amount of data that pass through our mind everyday. The object of study of this paper is automatic summarization algorithms. The main goal is to implement and make comparison of different summarization techniques on corpora of news articles parsed from the web. This research work contains the description of three summarization techniques based on TextRank algorithm: General TextRank, BM25, LongestCommonSubstring. It is specially noted the languages of used corpora: Russian and Kazakh languages. The results of summarization processes and their comparison are provided. It should be emphasized that used algorithms are well-known, but the way of their evaluation on defined corpora is different from those which usually used in summary evaluation. The method of summary evaluation proposed use the special dictionary of extracted key-words on the topic of corpora. As the title implies the article describes applying statistical information. The semantic and syntactic features of text are not examined.

Key words: summarization, automatic extraction, key-words, N-gram, TextRank

Статистикалық ақпараттар негізінде текстерді автоматты түрде реферирлеу

Мусина А.Б., әл-Фараби атындағы Қазақ Ұлттық университеті,
Алматы қ., Қазақстан Республикасы, +77759295274, E-mail: mussina.aigerim95@gmail.com
Аубакиров С.С., әл-Фараби атындағы Қазақ Ұлттық университеті,
Алматы қ., Қазақстан Республикасы, +77002200051, E-mail: aubakirov.sanzhar@gmail.com
Ахмед-Заки Д.Ж., әл-Фараби атындағы Қазақ Ұлттық университеті,
Алматы қ., Қазақстан Республикасы, +77772469374, E-mail: darhan.ahmed-zaki@kaznu.kz
П. Триго, Instituto Superior de Engenharia de Lisboa,
Лиссабон, Португалия, E-mail: ptrigo@deetc.isel.ipl.pt

Біздің күнделікті ақыл-ойымыздан өтетін көптеген ақпараттарды тиімді өңдеу - бүгінгі күннің өзекті мәселесі. Автоматтандырылған реферирлеу алгоритмдері жұмыстың зерттеу объектісі болып табылады. Мақалада сипатталған мақсат интернеттен алынған жаңалықтар мақалаларының корпусында реферирлеу алгоритмдерін жүзеге асыру және салыстыру. Берілген зерттеу жұмысы TextRank алгоритміне негізделген General TextRank, BM25, LongestCommonSubstring реферирлеудің үш алгоритмдерінің сипаттамаларын қамтиды. Орыс және қазақ тілдері қолданылған қорпустың ерекше тілдері ретінде атап өтілген. Реферирлеулер мен олардың салыстыруларының нәтижесі де берілген. Қолданылатын алгоритмдер жақсы танымал екендігін атап өтуге болатынына қарамастан, зерттеу барысындағы бағалау тәсілі әдеттегі қысқаша мазмұндағы бағалаудан ерекшеленетінің айта кету керек. Ұсынылып отырылған аннотацияларды бағалаудың әдісі корпус тақырыбындағы арнайы бөліп алынған кілттік сөздерді пайдаланады. Тақырыпқа сәйкес мақалада статистикалық ақпаратты пайдалану сипатталған. Мәтіннің семантикалық және синтаксистік қасиеттері қарастырылмайды.

Түйін сөздер: реферирлеу, автоматты шығарып алу, кілтті сөздер, N-gram, TextRank

Автоматическое реферирование текстов на основе статистической информации

Мусина А.Б., Казахский национальный университет имени аль-Фараби,
г. Алматы, Республика Казахстан, +77759295274, E-mail: mussina.aigerim95@gmail.com
Аубакиров С.С., Казахский национальный университет имени аль-Фараби,
г. Алматы, Республика Казахстан, +77002200051, E-mail: aubakirov.sanzhar@gmail.com
Ахмед-Заки Д.Ж., Казахский национальный университет имени аль-Фараби,
г. Алматы, Республика Казахстан, +77772469374, E-mail: darhan.ahmed-zaki@kaznu.kz
П. Триго, Instituto Superior de Engenharia de Lisboa,
Лиссабон, Португалия, E-mail: ptrigo@deetc.isel.ipl.pt

На сегодняшний день актуальной проблемой остается эффективная обработка большого объема информации, проходящей через наше сознание каждый день. Объектами данного исследования являются алгоритмы автоматического реферирования. Описанная в статье цель заключается в реализации и сравнении алгоритмов реферирования на корпусе новостных статей, взятых из интернета. Данная исследовательская работа содержит описание трех алгоритмов реферирования основанных на алгоритме TextRank: General TextRank, BM25, LongestCommonSubstring. Особенно отмечаются языки используемого корпуса: русский и казахский. Предоставлены результаты реферирования и их сравнение. Следует подчеркнуть, что используемые алгоритмы хорошо известны, но способ их оценки на изучаемом корпусе отличается от тех что обычно используются при оценке краткого содержания. Предлагаемый метод оценки аннотаций использует специальный извлеченный словарь ключевых слов по теме корпуса. Согласно названию в статье описывается применение статистической информации. Семантические и синтаксические свойства текста не рассматриваются.

Ключевые слова: реферирование, автоматическое извлечение, ключевые слова, N-gram, TextRank

1 Introduction

In this research work, our goal is to make 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 is primarily three types of automatic summarization: extraction-based, abstraction-based and aided. In this article extraction-based approach used, it uses parts of the original text, sentences, and construct the short paragraph summary, it does not make any modifications in text. (Automatic summarization) Many text features can influence on summary, like semantic and syntactic features, but we will concentrate on statistical data, which is a frequency statistics of N-grams. Based on this we chose extraction-based summarization type. Most of the algorithms that work based on statistical data build a summary text content by counting the similarity of text units and units' importance. Text unit could be a word, sentence or paragraph, in our case as a unit was chosen a sentence. Similarity is considered the presence of key-words in the sentences. Key-words are words that indicate the topic of the text.

2 Related works

During the search on related works were used next key-words: paragraph extraction from text, sentence extraction, position in text of main information, sentence similarity, informative sentence extraction. In the work (Chuleerat 2003: 9-16) 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 (Mandar 1997: 39-46, Fukumoto 1997: 291-298) 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 1997: 291-298) uses words weight among a paragraph and the extraction unit in this work is a paragraph. The works (Federico 2016: 65-72, 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 (Federico 2016: 65-72) 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 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 1998), it is about PageRank algorithm that proposed by Google. PageRank is an algorithm used in ranking of edges in any graph. TextRank uses it when construct summary from a generated graph.

The summary evaluation process described in (Federico 2016: 65-72, Sandeep 2009: 521-529) and they involve usage of ROUGE. Recall-Oriented Understudy for Gisting Evaluation (Chin-Yew Lin 2004: 74-81) is a set of metrics used in automatically generated summary evaluation and in machine translation. This kind of evaluation does not useful for us, because it assumes comparison of automatically produced summary and human generated summary, "ideal summary". This project work does not assume interaction with human. The hypothesis from work (Sandeep 2009: 521-529) 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.

In this article we will not describe the process of dictionary extraction, since it is fully examined in our previous work (Mussina 2017:). The corpora used in previous work (Mussina 2017:) and in this work is the same.

3 Source and methods

The corpora, which was used, consists of news articles that were parsed by web-crawler from government and news portals. Texts are in Russian and Kazakh languages. Most of the texts about some notification situations, like floods, earthquakes and storms, also may 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.

3.1 Summarization techniques

In the work (Federico 2016: 65-72) described TextRank algorithm for automated summarization with author's modifications. It represents document as graph with sentences as nodes. Edges between nodes show the similarity between sentences. Work (Federico 2016: 65-72) compare original way of similarity calculation with different modification proposed by authors. In this project work we have implemented three variations of similarity functions: original, BM25 and Longest common substring. The summary size is equal to the 30% of the original text size.

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

Formula 1 shows the similarity calculation by the original TextRank version.
Algorithm 1. Original TextRank

1. Extract list of sentences from text. Proceed to step 2.
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.

As an example we will show the work of algorithm on "NOTIFICATION" message.
Example 1. The original text shown below.

Жители Шымкента и Тараза почувствовали землетрясение в Афганистане, сообщает корреспондент Tengrinews.kz со ссылкой на ГУ "Сейсмологическая опытно-методическая экспедиция Комитета науки Министерства образования и науки Республики Казахстан". Подземные толчки были зафиксированы 10 апреля в 16.28 по времени Астаны. Эпицентр землетрясения располагался на территории Афганистана, в 787 километрах на юго-запад от Алматы. Энергетический класс землетрясения 14,5. Магнитуда - 6,8, глубина залегания - 20 километров. Толчки ощущались в Шымкенте и Таразе - 3 балла. Сведений о пострадавших и разрушениях нет. Напомним, 9 апреля землетрясение магнитудой 4,9 произошло в 141-м километре от Алматы. Подземные толчки были зафиксированы в 23.31 по времени Астаны. Эпицентр землетрясения находился в 141-м километре на юго-восток от Алматы на территории Кыргызстана. Энергетический класс подземных толчков - 10,2, глубина залегания - 5 километров.

All N-grams are counted, for example, in sentence 9 we have phrase "*Подземные толчки были зафиксированы*" and it contains 6 N-grams: "*Подземные*", "*Подземные толчки*", "*Подземные толчки зафиксированы*", "*толчки*", "*толчки зафиксированы*", "*зафиксированы*". The word "*были*" is a stop-word.

The BM25 variation based on the below formulas:

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

where IDF – inverse document frequency, $f(S_i,R)$ – occurrence frequency of a words i from sentence S in sentence R, $|R|$ - a length of sentence R, $avgDL$ – average length of sentences in the document; k_1 and b are parameters, we used the same that author of (Federico 2016: 65-72) work used, $k_1 = 1.2$, $b = 0.75$ 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:

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

where ε - between 0.3 and 0.5, we use 0.5

Algorithm 2. BM25

1. Extract list of sentences from text. Proceed to step 2.
2. Calculate IDF for all N-grams and the average length of document sentences.
3. For each sentence i $\in$ [0, sentence list size – 1]
4. Extract N-grams of sentence[i]
5. For each sentence j $\in$ [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 sufficient results. For similarity value used length of the longest common substring. Algorithm 3. Longest common substring

1. Extract list of sentences from text. Proceed to step 2.
2. For each sentence i $\in$ [0, sentence list size – 1]
3. Extract N-grams of sentence[i]
4. For each sentence j $\in$ [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, $Sim(S_i, S_j)$	BM25, $BM25(S_i, S_j)$	Longest Common Substring, $LCS(S_i, S_j)$
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

After the graph construction we need to go to the next stage, summary construction. We have graph with sentences as nodes and edges as similarity value between sentences. The PageRank based on the assumption that the amount of connections and the source of connection play role in the "importance" of the connected object. Consider data from example 1 for general TextRank similarity function. Generally, in our graph we have sentences connected with each other. However, usually we have such a situation when some sentences do not have any common word. In this case we have graph presented in figure 1(a). We can see that sentence with number 7 does not have any connection with other sentences, this means it has no common word with others. To reduce number of edges we define a threshold, which is equal to the average value of all edges weights, figure 1(b). The average value of all edges weights will be 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, figure 1(c).

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

In the ascending order we will get sentences array: S10, S2, S9, S11, S8, S3, S5, S4. The size of original text is equal to 11 sentences, the 30% of this 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 from ascending array: S10, S2, S9, S11. Then we permute sentences in the order of original text and save summary. Finally, we get the summary depicted below.

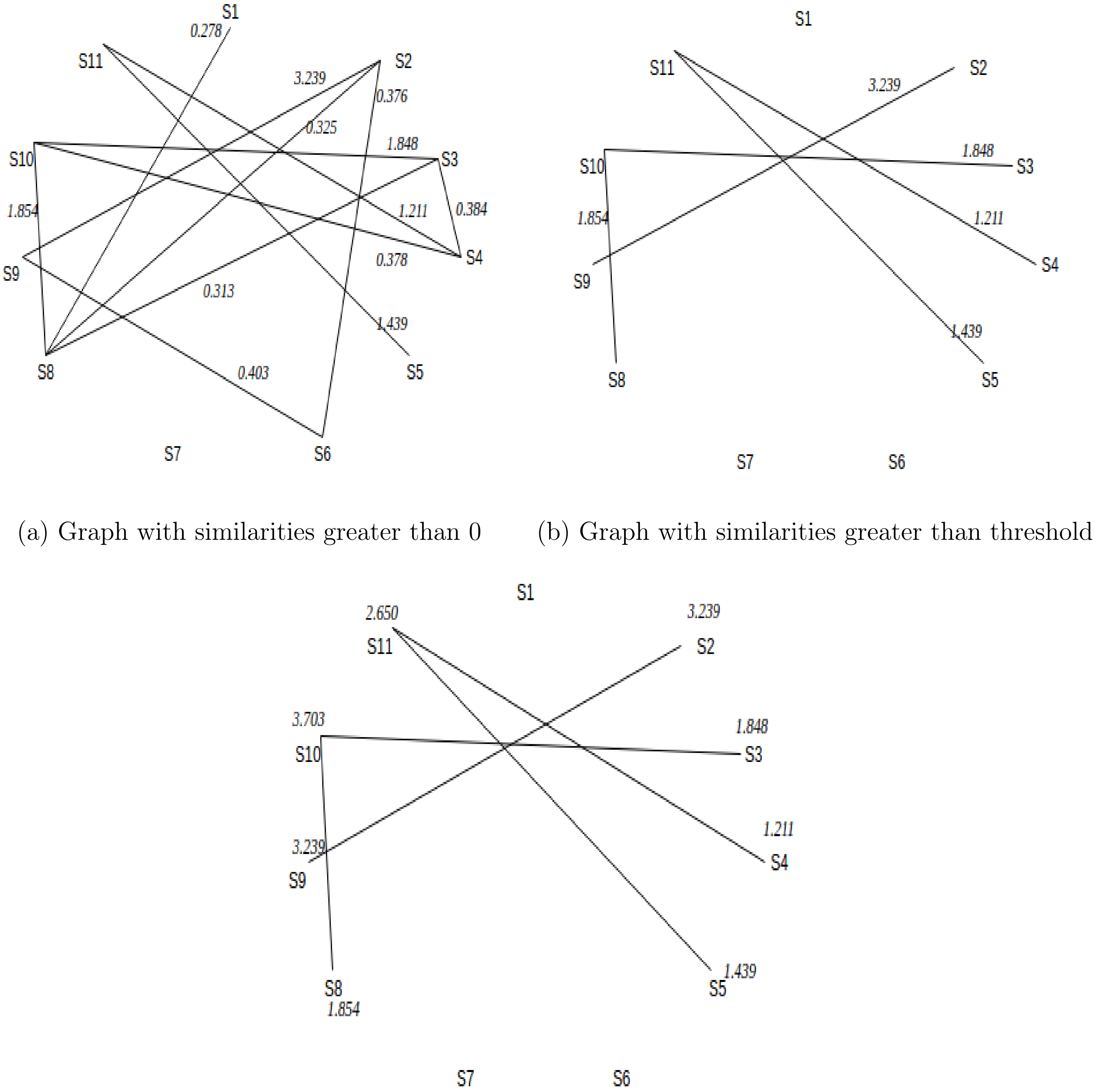
Подземные толчки были зафиксированы 10 апреля в 16.28 по времени Астаны. Подземные толчки были зафиксированы в 23.31 по времени Астаны. Эпицентр землетрясения находился в 141-м километре на юго-восток от Алматы на территории Кыргызстана. Энергетический класс подземных толчков - 10,2, глубина залегания - 5 километров.

In the end we calculate the sum of all similarities that sentence have with other sentences. The final value we use in the summary construction. Sentence with the higher value goes to the summary.

3.2 Summary evaluation

The evaluation of the summary based on the idea, proposed in work (Sandeep 2009: 521-529), 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 reduce amount of general words and save number of key-words. In (Sandeep 2009: 521-529) work authors use uni-gram model, but we will use model from 1 up to 5 N-grams. The algorithm of summary key-words distribution calculation described below.
Algorithm 4. Key-words distribution.

1. Get document from array of documents. Proceed to step 2.
2. Extract N-grams from text. Proceed to step 3.
3. For each N-gram check if it is in key-words dictionary. Count the sum of matches. Proceed to step 4.
4. Calculate key-words distribution by dividing sum of matches by the amount of N-gram extracted from the text. Proceed to step 5.
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.



(a) Graph with similarities greater than 0

(b) Graph with similarities greater than threshold

(c) Graph with ranked sentences

Figure 1 - Graphs

The original text with underlined key-words:

“*Жители Шымкента и Тараза почувствовали землетрясение в Афганистане, сообщает корреспондент Tengrinews.kz со ссылкой на ГУ "Сейсмологическая опытно-методическая экспедиция Комитета науки Министерства образования и науки Республики Казахстан". Подземные толчки были зафиксированы 10 апреля в 16.28 по времени Астаны. Эпицентр землетрясения располагался на территории Афганистана, в 787 километрах на юго-запад от Алматы. Энергетический класс землетрясения 14,5. Магнитуда - 6,8, глубина залегания - 20 километров. Толчки ощущались в Шымкенте и Таразе - 3 балла. Сведений о пострадавших и разрушениях нет. Напомним, 9 апреля землетрясение магнитудой 4,9 произошло в 141-м километре от Алматы. Подземные толчки были зафиксированы в 23.31 по времени Астаны. Эпицентр землетрясения находился в 141-м километре на юго-восток от Алматы на территории Кыргызстана. Энергетический класс подземных толчков - 10,2, глубина залегания - 5 километров.*”

The body evaluation = 0.124

The below summary is identical to all three TextRank techniques: General, BM25, LongestCommonSubstring. As we can see the key-words distribution increased in summary.

"*Подземные толчки были зафиксированы 10 апреля в 16.28 по времени Астаны. Подземные толчки были зафиксированы в 23.31 по времени Астаны. Эпицентр землетрясения находился в 141-м километре на юго-восток от Алматы на территории Кыргызстана. Энергетический класс подземных толчков - 10,2, глубина залегания - 5 километров.*"

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. The described evaluation could be applied only to “NOTIFICATION” classified messages. Even the N-gram with negative “emergency” state that it is belonging more to “NEWS” messages, it could be not correct to say that we can use inverse “emergency” to evaluate general news articles. This assumption could be explained by the biased articles in “NEWS”, since they are about very different topics.

4 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

Table 3 - TextRank variations evaluation results

	Key-words distribution
Original documents	0.159
General TextRank	0.18
BM25	0.169
LongestCommonSubstring	0.175

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 described in the Section 3.2.

During this research work TextRank algorithm variations were tested and estimated. In the (Federico 2016: 65-72) 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:

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 persons summary, from which will be chosen one ideal. The professional activity of each human candidate also play role.

The alternative way of evaluation process, as was mentioned in sub-section 3.2, Summary evaluation, based on the hypothesis from (Sandeep 2009: 521-529). Authors used KL (Kullback-Leibler) Divergence which denotes the difference between two probability distributions by formula:

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

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 discuss on entropy.

Another one alternative way of summary evaluation which was proposed by us is calculation "emergency" of summary. The idea based on "emergency" value of N-grams from dictionary. Since we have N-grams with, "emergency" < 0, we suppose that the summary with more not important, not from dictionary, words will have low "emergency", when summary with less not important words and more key-words will have high "emergency". The main problem with such method is in uncertainty of comparison of values, because we cannot properly normalize values with certain upper and lower bounds.

In future we would like to continue research on completely different summary generation. 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. We propose the possible idea to construct summary by sentences non similarity. The new algorithm is as follows:

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

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 "подземные толчки" and sentence S_B "землетрясение", meaning of these N-grams mostly equal, but implemented algorithm will not recognize similarity.

5 Conclusion

The research on already existent works about object of our study was made. 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 (Federico 2016: 65-72). The General TextRank was the best one, which generates summary with high distribution of key-words. Its average key-words distribution is equal to 0.18. The LongestCommonSubstring, which is the easiest algorithm to implement has key-words concentration equal to 0.17589730440957158. The lowest distribution 0.169 belongs to BM25. In the (Federico 2016: 65-72) work authors make changes in parameters values, such that the results of BM25 became better. In future work we will also make changes in parameters and IDF function definition. 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. The summaries constructed by all three algorithms in most of cases cut off not important information and leave the important part with key-words. The examples of algorithm work and results of their tests presented in this research work.

This work was partially supported by the grant of the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan (project 3350/GF4 MON RK).

References

[1] Chin-Yew Lin, "ROUGE: A Package For Automatic Evaluation Of Summaries,"*ACL Anthology Network* (2004): 74–81, accessed October 20, 2016

[2] Chuleerat Jaruskulchai and Canasai Kruengkrai, "A Practical Text Summarizer by Paragraph Extraction for Thai,"(paper presented at the Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages, Sappro, Japan, July 7, 2003)

[3] Federico Barrios, Federico Lopez, Luis Argerich, Rosita Wachenchauzer, "Variations of the Similarity Function of TextRank for Automated Summarization,"*Cornell University Library* (2016): 65-72, accessed November 14, 2016, arXiv:1602.03606.

[4] Fukumoto F., Suzuki Y., Fukumoto J., "An Automatic Extraction of Key Paragraphs Based on Context Dependency,"*Natural language processing* Vol. 4 (1997): 89-109, DOI:10.5715/jnlp.4.2_89.

[5] Mandar Mitrat, Amit Singhal, Chris Buckleytt, "Automatic Text Summarization by paragraph Extraction,"*Intelligent Scalable Text Summarization* (1997):39-46.

[6] Nagwani N.K., "Summarizing large text collection using topic modeling and clustering based on MapReduce framework,"*Journal of Big Data* (2015): 18.

[7] Page L., Brin S., Motwani R., Winograd T., "The pagerank citation ranking: Bringing order to the web,"(paper presented at the Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, 1998)

[8] Sandeep S. and Jagadeesh J., "Summarization Approaches Based on Document Probability Distributions,"(paper presented at Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Hong Kong, China, December 3-5, 2009).

[9] Wikipedia. "Automatic summarization."Accessed November 25, 2016,
https://en.wikipedia.org/wiki/Automatic_summarization.

[10] Wikipedia. "Stop words."Accessed June 30, 2015,
https://en.wikipedia.org/wiki/Stop_words.

[11] Yacko V.A., "Simmetrichnoe referirovanie: teoreticheskie osnovy i metodika,"*Nauchno-tehnicheskaya informaciya* Ser.2 (2002): 18-28