SRSTI 20.19.19

Dictionary extraction based on statistical data

Mussina A., bachelor of Technics and Technologies, Al-Farabi Kazakh National University,
Almaty, Republic of Kazakhstan, +77759295274, E-mail: mussina.aigerim95@gmail.com
Aubakirov S., PhD, Al-Farabi Kazakh National University,
Almaty, Republic of Kazakhstan, +77002200051, E-mail: aubakirov.sanzhar@gmail.com

Automatic text summarization is an actual problem when working with a large amount of information. Most of the algorithms that work on the basis of 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 unit is a sentence. Similarity is considered the presence of key-words in the sentences. Key-words are words that indicate the topic of the text. In this research work we will describe an automatic extraction of key-words dictionary, where key-words are N-grams with N from 1 to 5. Two algorithms were implemented: getting of words that occur only in one of two different corpora and getting of words with high importance. Importance of N-gram denotes its belonging to the topic of the text. Used text languages are Russian and Kazakh. The algorithms show important results, both of them make sense in constructing of full key-words dictionary.

Key words: automatic extraction, key-words, N-gram.

Извлечение словаря на основе статистических данных

Мусина А.Б., бакалавр техники и технологий, Казахский национальный университет имени аль-Фараби,
г. Алматы, Республика Казахстан, +77759295274, E-mail: mussina.aigerim95@gmail.com Аубакиров С.С., PhD, Казахский национальный университет имени аль-Фараби,
г. Алматы, Республика Казахстан, +77002200051, E-mail: aubakirov.sanzhar@gmail.com

Автоматическое реферирование текста это актуальная проблема при работе с большим количеством информации. Большинство алгоритмов, которые работают на основе статистических данных, подсчитывают схожесть текстовых единиц и их важность при состовлении краткого содержания. Текстовой единицей может быть слово, предложение или параграф, в нашем случае это предложение. Сходство считается наличием ключевых слов в предложениях. Ключевые слова - это слова, которые указывают на тематику текста. В этой исследовательской работе мы опишим автоматическое извлечение ключевых слов, где ключевыми словами являются N-граммы с N от 1 до 5. Реализованы два алгоритма: получение слов, которые встречаются только в одном из двух разных корпусов и получение слов с высокой степенью важности. Важность N-gram обозначается его принадлежностью к тематике текста. Использованы тексты на русском и казахском языках. Алгоритмы показывают важные результаты, оба могут быть использованы в создании полного словаря ключевых слов.

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

Статистикалық деректер негізінде дайындау сөздік

Мусина А.Б., техника және технологиялар бакалавры, әл-Фараби атындағы Қазақ Ұлттық университеті,
Алматы қ., Қазақстан Республикасы, +77759295274, E-mail: mussina.aigerim95@gmail.com Аубакиров С.С., PhD, әл-Фараби атындағы Қазақ Ұлттық университеті,
Алматы қ., Қазақстан Республикасы, +77002200051, E-mail: aubakirov.sanzhar@gmail.com

Мәтінді автоматты рефрерлеу - бұл ақпараттың үлкен санымен жұмыс істеу кезіндегі өзекті мәселе. Статистикалық деректердің негізінде жұмыс істейтін алгоритмдердің көпшілігі мәтіндік бірліктердің ұқсастығын және олардың қысқаша мазмұн жасау кезіндегі маңыздылығын есептейді. Мәтіндік бірлік ретінде сөз, сөйлем немесе бөлім болуы мүмкін, біздің жағдайда бұл- сөйлем. Сөйлемдерде кілт сөздердің болуы, ұқсастық болып саналады. Кілт сөздер - олар мәтіннің мазмұнымен болмысына нұсқайтын сөздер. Осы зерттеу жұмысында біз авттоматты түрде кілт сөздерді алуды сиппаттаймыз, бұл жерде N - граммалар N 1-ден бастап 5-ке дейін кілт сөздер болып табылады. Қазіргі таңда екі алгоритм іске асырылды олар - әр түрлі екі корпустардың тек қана бірінде кездесетін сөздерді алу және жоғары дәрежелі маңыздылығы бар сөздерді алу. N -граммалардың маңыздылығы оның мәтіннің мазмұнына тиістілігіне қарай белгіленеді. Қазақ және орыс тілдеріндіегі мәтіндер қолданылды. Алгоритмдер маңызды нәтижелер көрсетуде, екеуі де толық кілт сөздер сөздігін құру барысында пайдаланылуы мүмкін.
Түйін сөздер: автоматты шығарып алу, кілтті сөздер, N-gram

1 Introduction

The amount of information is extremely growing up, data processing is becoming time and resource consuming. The solution could be the usage of text summarization, which is the object of study. Getting only meaningful part of the text will give approximately the same knowledge as full text does. During the research on this technology, we are faced with the need to use key-words dictionary. Without the knowledge about semantic and syntactic meaning of words and phrases it is very hard to find out key-words of the text topic. The only thing we have to use is statistical data as frequency of occurrence. Dictionary extraction is the subject of our study. The goal of this research work is to extract topic dictionary of key-words. Our final goal is to develop automatic summarizer that will effectively work on our corpora.

2 Related works

The article (Chuleerat 2003: 9-16) presents an algorithm for extracting the most significant paragraphs from the text in Thai, where the significance of the paragraph depends on the local and global properties of the paragraph. The main emphasis is on the knowingly correct distribution of paragraphs, Thai language is very different from European languages and is more similar to Chinese and Japanese in terms of fuzzy division of words and sentences. We use the Russian and Kazakh languages, which have a clear sentence structure. However, if the text of the message is large enough, this algorithm can be used to identify significant paragraphs. 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 the work (Federico 2016: 65-72) TextRank algorithm presented, which is the most popular in text summarization. It uses the similarity of sentences in words to identify informative sentences. 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. Comparing of all words may cause distortion of results, to prevent this it is useful to consider only important words, key-words. In work (Iain 2016) a method of identifying the most important words using the corpora of texts from songs of a certain genre is proposed. The author defines the coefficient Mw as an indicator of belonging to the genre. The coefficient is calculated by the formula

$$M_w = \log \frac{N_w^{metal}}{N_w^{brown}}$$

where N_w^{metal} is the frequency of occurrence of the word w in the body of lyrics and N_w^{brown} is the frequency of occurrence of the word w in the Brown corpus (Wikipedia Brown Corpus). To calculate the "emergency"of the word, we decided to use the above formula, in our case N_w^{metal} is replaced by $N_w^{emergency\_news}$ - the frequency of occurrence of the word w in the emergency message body and N_w^{brown} is replaced by N_w^{news} is the frequency of occurrence of the word w in the news bulletin. According to these data, a list of the most and least "emergency"words is constructed. Applications of many algorithms and methods of text analysis depends on statistical information about text (Riedl 2012: 47-70), statistics about N-grams are an important building block in knowledge discovery and information retrieval (Berberich 2013: 101-112). In this project work, segmentation of text is considered according to words of different lengths within a single sentence. In the English literature, the term tokenization (The Art of Tokenization) is used. To denote a segment, we use the term N-Gram. N-Gram are sequences of adjacent words or tokens in a text document or line, where N is the length of the sequence (Berberich 2013: 101-112). A sequence of two consecutive elements is often called a bi-gram, a sequence of three elements is called a trigram. At least four or more elements are designated as N-grams, N is replaced by the number of consecutive elements (Wikipedia N-grams). For the calculation of N-grams already developed technologies such as Elasticsearch (Elasticsearch engine guide) and SRILM (The SRI Language Modeling Toolkit) (Srilm project). In (Ngram count), calculation of repetitions of N-grams in the body of texts or in a sentence using a finite automaton is shown. 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 d 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.

3 Source and methods

In this research work we will consider the case of processing the information about emergence situations along the news data. We have used corpora of news articles from web. For this purpose web crawler was implemented. It parses 21 state web-sites and 22 news portals. We categorize news articles to «News», that are news messages about general human life, and «Emergency News», that are messages with notifications about some danger and/or unfavorable situations.

3.1 Tools used

To preprocess text, index and extract N-gram we used the following tools: Apache Lucene, Elasticsearch and Apache OpenNLP. Apache Lucene is a library from the Apache Software Foundation for full-text search, it is also used in computational linguistics. Apache Lucene provides many tools for building indexes based on the TF-IDF model. Using Apache Lucene, we removed all stop-words from the texts, extract all N-gram from 1-gram to 5-gram and built indexes. Elasticsearch is a search tool based on Apache Lucene. The main feature of elasticsearch is the ability to perform high-speed search in the database with a large number of documents using indexes. Apache OpenNLP is a tool for working with natural language, it can perform various operations with text, such as tokenization, division of text into sentences. The tokenization of OpenNLP has been replaced by ShingleFilter's tokenization, since the latter is able to extract N-gram from sentences of different lengths at once, while OpenNLP alone can only split sentences into separate words. The used corpora is classified and marked with meta data (message title, publication date, TF-IDF, tags). Data are classified into "News"and "Emergency News".

3.2 Implementation of algorithms

We have implemented the algorithm for extracting the emergency N-gram dictionary and calculating the "emergency"of N-grams. At this stage, $N = 5$ (phrases of up to 5 words). Both algorithms extract and count the N-gram frequency.

Algorithm 1: Extract and calculate the frequency of occurrence of N-gram: 1. We retrieve all messages of the given class. Proceed to step 2. 2. We form a common text from the title lines and the main text of the messages. Proceed to step 3. 3. We divide the text into sentences. We proceed to step 4. 4. We extract all possible N-gram and perform the calculation of frequency.

Algorithm 2: Extracting the vocabulary of extreme words. 1. We extract the N-gram with the frequency of occurrence according to the algorithm 1. Proceed to step 2. 2. Write the N-gram from the text, that classified as "Emergency News to emergency dictionary if the N-gram is missing from messages classified as "News".

The calculation of "emergency"performed, according to the formula presented in section 2, Related works. Two types of N-gram are considered for which calculation is performed. The first type is N-gram with minFreq greater than 1 in two types of messages. The second type is N-gram with minFreq greater than 5 in two types of messages.

Algorithm 3: calculation of "emergency"N-gram. 1. We extract the N-gram with the frequency of occurrence according to the algorithm 1. Proceed to step 2. 2. We check that N-gram from «Emergency News» is not among the N-gram from "News". If it occurs in both, proceed to step 3. 3. If the N-gram occurs in messages of the class "News"and "Emergency News"more or equal to minFreq (the minimum frequency of occurrence), then calculate for its "emergency". Proceed to step 2.

In the work (Iain 2016), the author highlight the possible distortion of the results of the third algorithm, caused by rare words. As a decision, only words that occur at least 5 times in both cases are taken into account. In this paper we denote this solution by minFreq. If the frequency of occurrence of N-gram in both classes is greater than or equal to minFreq, then "emergency"is calculated for it.

4 Results and discussion

Table 1 - Source data for dictionary extraction

	Amount
News articles classified as «Emergency News»	2592
News articles classified as «News»	75.901

In Table 1 presented data about corpora, amount of news articles classified as «Emergency News» and «News». General news is nearly 30 times more than emergency.

Table 2 - Algorithm 1 results

	Amount
N-grams from «Emergency News»	483.052
N-grams from «News»	26.774.077

Amount of N-grams by algorithm 1 presented in Table 2. There are more news articles classified as «News», because of this amount of N-grams is also more, almost 55.4 times.

Table 3 - Algorithm 2 results

	Amount
Dictionary of emergency N-gram	202.884

Table 3 shows that about 280 thousand N-gram occurs in news articles of both classes. The rest is the dictionary of emergency N-gram.

Table 4 - Top-10 most frequent from the dictionary of emergency N-gram

	Dictionary key-words N-gram	Frequency
1	явлений погоды (weather phenomena)	160.0
2	прогноз важнейших явлений погоды (forecast of the most important weather phenomena)	158.0
3	важнейших явлений погоды (the most important weather phenomena)	158.0
4	явлений погоды информации (weather phenomena information)	116.0
5	погоды информации (weather information)	116.0
6	прогноз важнейших явлений погоды информации (forecast of the most important weather information)	115.0
7	важнейших явлений погоды информации (the most important weather information)	115.0
8	погоды информации филиала ргп (weather information of the branch of the PPP)	98.0
9	важнейших явлений погоды информации филиала (the most important weather information of the branch)	98.0
10	погоды информации филиала (weather information of branch)	98.0

All the phrases, from Table 4, within the current corpora are found only in messages of the class "Emergency News". The top 10 emergency words are presented on the frequency of their occurrence. All ten N-grams which are on the top connected with each other. Probably the one very common sentence repeats in most of emergency news.

Table 5 - Algorithm 3 results

	Amount
N-gram with computed «emergency», in case of minFreq=1	280.168
N-gram with computed «emergency», in case of minFreq=5	19.293

The number of N-gram for which the "emergency"was calculated by algorithm 3 is shown in Table 5. The calculations were carried out with minFreq equal to 1 and 5. As can be seen with minFreq = 1, the amount of N-gram equals the number of N-grams not included in the dictionary of emergency words by algorithm 2.

Table 6 - Results of algorithm 3, minFreq = 1

	“emergency” > 0	“emergency” < 0	“emergency” = 0
N = 1	669	24018	1071
N = 2	5852	41059	19030
N = 3	7619	26843	31499
N = 4	7554	20337	35338
N = 5	6887	16480	35912

The number of N-grams with “emergencies” greater than 0, less than 0 and equal to 0 in case of minFreq = 1 presented in Table 6. When “emergency” is equal to 0, it means that N-gram occurred in the “NOTIFICATION” messages the same time as in “NEWS” message, since in formula 1 we use logarithm. Here we can notice that amount of uni-grams is more less than other N-grams.

Table 7 - Results of algorithm 3, minFreq = 5

	“emergency” > 0	“emergency” < 0	“emergency” = 0
N = 1	223	6364	31
N = 2	1040	4080	302
N = 3	1141	1760	407
N = 4	967	954	355
N = 5	776	587	306

The number of N-grams with “emergencies” greater than 0, less than 0 and equal to 0 in case of minFreq = 5 presented in Table 7. Here bi-grams and tri-grams occur more than other N-grams.

Table 8 - Top-10 N-gram with high «emergency» in case of minFreq=1

	“emergency ” N-grams	“emergency” value
1	подземных толчков (tremors)	4.532599493153256
2	землетрясение магнитудой (earthquake of magnitude)	4.436751534363128
3	сведений ощутимости (sensibility data)	3.8918202981106265
4	алматы границе (almaty border)	3.8501476017100584
5	методическая экспедиция (methodical expedition)	3.713572066704308
6	сейсмологическая (seismological)	3.713572066704308
7	опытно методическая (expertly methodical)	3.713572066704308
8	сейсмологическая опытно методическая экспедиция (seismological experimental expedition)	3.713572066704308
9	сейсмологическая опытно методическая (seismological experimental methodological)	3.713572066704308
10	опытно методическая экспедиция (experimental expedition)	3.713572066704308

The top 10, shown in Table 8, mainly consists of N-gram about earthquake. Indeed, these words and phrases can rarely be found in the usual news. However, the result is slightly distorted due to words that usually appear together, for example, the phrases from 5 th to 10 th are from the same sentence and their “emergency” is also equal. Probably the whole phrase is presented by 8 th N-gram. This means that usually people use the same sentence construction when they write about earthquake.

Table 9 - Top-10 N-gram with low «emergency» in case of minFreq=1

	“emergency ” N-grams	“emergency” value
1	де (particle)	-9.167328481382892
2	бойынша (according to)	-9.105757331783742
3	туралы (about)	-8.835210463664092
4	бір (one)	-8.668711839055147
5	кг (kg)	-8.426173793029069
6	болады (will be)	-8.352554369474591
7	мемлекеттік (the state)	-8.345455428161928
8	жол (road)	-8.339261982923576
9	болды (was)	-8.328934041955529
10	бұл (this)	-8.294216292881348

The low "emergency"of N-gram at minFreq = 1 is presented in Table 9. These words and phrases are more common in probably all news article, but they appear more in general news. For example, 6 th and 9 th presents the time form of the same auxiliary verb, 4 th is a number, 5 th the measure abbreviation. All N-grams, or more concrete uni-gram, are in Kazakh language.

Table 10 - Top-10 N-gram with high «emergency» in case of minFreq=5

	“emergency ” N-grams	“emergency” value
1	магнитудой (magnitude)	3.040184036124825
2	объявлено штормовое (storm announced)	2.929287174145838
3	объявлено штормовое предупреждение (storm warning announced)	2.929287174145838
4	прогноз важнейших явлений (forecast of the most important phenomena)	2.8965256234705428
5	важнейших явлений (most important phenomena)	2.8965256234705428
6	горных районах алматинской (mountainous areas of Almaty)	2.70805020110221
7	км юго запад (southwest km)	2.662587827025453
8	землетрясение (earthquake)	2.631089159966082
9	прогноз важнейших (forecast of the most important)	2.608843551018762
10	эпицентр землетрясения (earthquake epicenter)	2.5508646175797978

Comparing the results from Table 8 and Table 10, you can see that the N-gram about earthquake mostly appears in “Emergency News”. In table 10, where minimum frequency value is equal to 5 in both classified news articles corpora, predominantly represented N-grams about storm. We can say that “News” articles also often use N-grams about storms.

Table 11 - Top-10 N-gram with low «emergency» in case of minFreq=5

	“emergency ” N-grams	“emergency” value
1	мен (I)	-7.959683517398391
2	және (and)	-6.909709889225459
3	сборной (team)	-6.692827882439268
4	үшін (for)	-6.665220379421287
5	жаңа (new)	-6.664153885765531
6	встречи (meetings)	-6.5789733956449306
7	сумму (amount)	-6.354080143708786
8	имеет (has)	-6.350594807621779
9	наших (our)	-6.344173302776835
10	мажилиса (mazhilis)	-6.306275286948016

Table 11 shows words and phrases rarely found in emergency news. These N-gram more characterize the news of the class "News".

The main distortion in results caused by list of stop-words, which is not full, because of that dictionary contains abbreviations and not meaningful N-grams. Another problem connected with absence of stemming support. The results could be more clear and effective with applying of Russian and Kazakh stemming algorithm. Some N-grams, mostly unigrams, has equally popular variations. Here variations are N-grams with changed affix part. Generally, they are the same, but dictionary extraction algorithm takes them as different N-grams. For example: “сейсмологических”, “сейсмологический”, “сейсмологическая”. All previous words translated to English as “seismological”, they have difference only in affixes that denote the number and gender of noun to which those adjectives connected. We should take into account the existence of N-grams with place names, for example “Пожар в "Абу Даби Плаза"”. The "Абу Даби Плаза"is a name of building, but since such N-gram, “Пожар в "Абу Даби Плаза"”, does not appear in “NEWS” messages, it was written to dictionary. Finally, during tests we have noticed that some notification N-grams do not appear in dictionary, for example “Подземные толчки”. Each word exists in dictionary separately, but not together. This situation is possible because such N-gram does not appear in “NEWS” messages corpora and as a result the “emergency” could not be calculated. We should take into account such N-grams.

5 Conclusion

Table "Top 10"show that N-gram associated with accidents, assurance and search and rescue (SAR) operations are most often found in the emergency messages body. To get mostly full dictionary we can replenish results of Algorithm 2 (Extracting the vocabulary of extreme words) with words and phrases from results of Algorithm 3 (calculation of N-gram's emergency) with "emergency"above some defined threshold value. We need to calculate this threshold value since the results of Algorithm 3 showed that N-gram, most suitable for one particular class, can occur in texts of both classes.

Future work. The results of extracting the vocabulary of emergency words and calculating the "emergency"of words in the future will be used in identification of important part of text. The importance of text unit will be defined via its local properties, the presence of N-gram from dictionary, and global properties, the presence of nearby text units with high values of local properties. It is planned to use this importance identification in automatic document summarization.

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] Berberich K. and Bedathur S., "Computing n-gram statistics in MapReduce,"*ICPS - International Conference Proceedings Series* (2013): 101-112

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

[3] 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)

[4] CraigTrim. "The Art of Tokenization."Accessed June 30, 2015,
https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en.

[5] Elasticsearch. "Elasticsearch engine guide."Accessed October 25, 2015,
https://www.elastic.co/guide/en/elasticsearch/reference/1.4/index.html.

[6] 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.

[7] 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.

[8] Iain. "Heavy Metal and Natural Language Processing - Part 1."Accessed September 20, 2016,
http://www.degeneratestate.org/posts/2016/Apr/20/heavy-metal-and-natural-language-processing-part-1/.

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

[10] Ngram count. "Ngram count."Accessed October 25, 2016,
http://www.ling.ohio-state.edu/ bromberg/ngramcount/ngramcount.html.

[11] Riedl M. and Biemann C., "Text segmentation with topic models,"*Journal for Language Technology and Computational Linguistics* Vol.27 (2012):47-70

[12] 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).

[13] Srilm project. "Srilm project."Accessed October 25, 2015, http://www.speech.sri.com/projects/srilm/.

[14] Wikipedia. "Brown Corpus."Accessed September 20, 2016, https://en.wikipedia.org/wiki/Brown_Corpus.

[15] Wikipedia. "N-grams."Accessed September 20. 2015, https://en.wikipedia.org/wiki/N-gram.

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