Comparative Research of Index Frequency - morphological Methods of Automatic Text Summarisation*

Alexsander Osochkin
osa585848@bk.ru

Vladimir Fomin
vv_fomin@mail.ru

Olga Yakovleva
ekzegeza@yandex.ru

Herzen State Pedagogical University of Russia
Saint Petersburg, Russian Federation

Abstract

The article considers the potential of frequency-morphological analysis implementation in index methods of automatic text summarisation. The main feature of the developed index method using frequency-morphological analysis is the consideration of the importance of parts of speech in the particular language. The evaluation of the effectiveness of automatic text summarisation of scientific and educational documents, fiction in Russian using various indexing methods is presented in the paper. Based on the experiments results, indexing methods were evaluated and quality ranked in automatic text summarisation algorithms, recommendations for their use were made.

Keywords: *frequency-morphological analysis, automatic summarisation, indexing methods, morphological analysis*

1 Literature review

The extraction of knowledge from nature language (NL) data, including the provision of information in short form, has always been a main topic in the educational field, especially the implementation of information into the educational process. Increasingly, the topic "Redundancy checking algorithms" is becoming the main topic in the field of information processing [Lei, 2017], [Salloum et al., 2017]. One of the first attempts to summarise texts was described in the works of [Luhn, 1958], the authors of which faced the problem of defining important parts of the text, words in the text. Over the years, several techniques have been applied for solving this problem including some recent attempts using neural networks [Shari, 2018], [Molchanov, 2015], [Jansen, 2010]. Today, the authors are increasingly resorting to frequency and frequency-morphological methods of text analysis in automatic summarisation [Said et al., 2017], [Al-Emran, 2017], this trend is caused by the simplicity of implementation of these algorithms based on the research of papers in the field of auto summarisation on the platform EBSO search, it was revealed that the number of works [Getahun, 2017] in the field of auto-conversion is steadily increasing by 17% every year, and frequency-morphological methods of indexing become the predominant method.

2 Introduction

Automatic summarisation is an automatic process that creates a result text from one or more source texts that transmits most of the information in smaller size. [Brandow et al., 1995]. Today, there are many different methods of automatic summarisation [Clayton et al., 2011], [Evdokimenko, 2013], [Al-Emran, 2017], [Salloum et al., 2017], [Sujit et al., 2013], [Shari, 2018], but among all automatic summarisation methods, it is particularly worth highlighting indexing methods, which are based on simple, well-proven frequency analysis methods [Sujit et al., 2013], [Shari, 2018], [Molchanov, 2015], [Jansen, 2010] of text information on NL.

In the field of text-mining and natural language processing NLP, frequency analysis is the predominant method of text analysis, [Yogesh, 2014] but increasingly in scientific articles there is a shift to more complex methods of text analysis, identification of the basis of sentences, etc., using frequency-morphological analysis. The main problem of automatic summarisation is the identification of the most significant parts of the text, which removed from the text would save the integrity and reflect the main topic of the document. There is a quite large amount of automatic summarisation methods developed over the past two decades, but all of them can be conditionally divided into two groups: extraction methods and abstraction methods.

Abstraction methods are automatic summarisation methods based on the creation of a new text using new words and synonyms, consolidating the original text. These methods are of great scientific interest, especially in the field of NLP, as they involve the use of complex semantic analysis algorithms.

Abstraction methods include three necessary steps:

1) creation of the text main idea, frequently used words and main topic identification, etc.;

2) indexing of words, phrases and other meaningful units;

3) indexing-based new text consolidation and synthesis.

Extraction methods are a method of automatic summarisation text in which words with low indexes are extracted from the text. The distinctive feature of this method is saving the original text. The algorithm identifying the importance of sentences and words using indexing methods, which allow ranking elements of the text: words, sentences, paragraphs. The majority of industrial-scale automatic summarisation systems are implemented within the framework of this approach [Yogesh, 2014], although these systems also have a number of problems.

Regardless of the type of automatic summarisation method, each uses indexing of the text internal content, in order to rank the text elements and save the most significant ones. Therefore, the most important step for two kinds of automatic summarisation is the step of indexing the text internal content. Despite the fact that this stage is key for summarisation method, there is no general reliable indexing method, which would be effective in a large number of different tasks. Therefore, there is a wide range of indexing algorithms; each of them includes an effective method of text summarisation applying to the particular structure or text type.

The first automatic summarisation methods were based solely on frequency or positional analysis based on the analysis of each individual word or its position in the text. With the development of text-mining and NLP, more sophisticated methods of semantic and linguistic analysis began to be applied, scientists tried to apply text analysis with the higher meaningful units as paragraphs, thematic parts, sentences and. etc. The main problem of methods based on semantic and linguistic analysis is the lack of comparison of the sentences importance [Baxendale et al., 1958], [Yogesh, 2014], [Shari, 2018], [Dragomir, 2012]. This feature significantly reduces the quality of automatic summarisation, it refers in a big extent to the texts, where the author mentions several topics as well as to the artistic texts.

The main problem with indexing is that there is no consensus about what minimum unit of text analysis is the best for auto summarisation. On the one hand, frequency methods of text indexing are analysed at the level of unigrams, a separate word cut from the context, on the other hand semantic and linguistic methods of indexing involve sufficiently large meaningful units such as paragraphs, subsections, sentences, etc.

Due to the same problem in two types of automatic summarisation indexing, we have chosen indexing methods based on frequency analysis, as they are more universal, less dependent on language specificity, there are ready-made software solutions for indexing documents and these methods do not require specialized knowledge in the field of linguistics.

In the science researchers of automatic summarisation [Sujit et al., 2013], [Shari, 2018], [Molchanov, 2015], [Jansen, 2010], [Yogesh, 2014], [Tarasov, 2010], [Gambhir, 2016] it was revealed that modern algorithms related to indexing of text including the frequency method are based on uniform analysis of text, and ignore the analysis of higher organized units: phrases, sentences, and paragraphs. When analysing higher units of text, new properties appear: cohesion, coherence, and auto-somatization of individual paragraphs and text lines, etc. The use of more highly organized units while indexing a document requires a transition to frequency-morphological analysis, to identify and use special properties of phrases, sentences, paragraphs.

2.1 Relevance and purpose of the article

Information redundancy is a major problem in various information environments, where huge amounts of semi-structured data in natural language are accumulated. This problem is particularly relevant to information educational environments, because the provision of brief reference material allows to speed up the process of searching for the necessary information, which affects the level and quality of education in general. Nowadays there is no doubt that intelligent search significantly increases efficiency in any information environment that searches in huge amounts of semi-structured data in natural language. In such circumstances, new effective methods for dealing with large amounts of information that can convey the exact content of a document in a concise form are of particular importance.

One of these methods is automatic summarisation as a type of analytical and synthetic document processing [Sujit et al., 2013], [Shari, 2018], [Radev et al., 2002], which allows the required information support [Molchanov, 2015], [Jansen, 2010]. The purpose of the study is to evaluate the effectiveness of automatic text summarisation using index methods, including the use of frequency-morphological analysis.

In order to achieve this purpose, the following tasks were set: To analyse approaches to text auto summarisation based on index methods. Selecting Frequency indexing Methods to generate automatic summarisation of text materials in Russian. To modify the selected index methods using frequency-morphological analysis as the main one. To evaluate and compare the results of auto summarisation, frequency and frequency morphological analysis in terms of accuracy, completeness and amount reduction from the source text and the standard.

3 Automatic summarisation algorithm

When using index automatic summarisation methods, any D_j text in natural language can be represented as a set of words: $W = w1, w_2, \dots\ w_n$. Where, each word w_n has an index F obtained by calculating the frequency indexing method. When using the frequency analysis based indexing method, text is represented at the elementary level, thus the main elementary unit of frequency analysis is the word.

We propose to complicate frequency indexing methods by adding morphological analysis, thanks to which we will be able to obtain another set of V index.

The morphological index V is determined on the basis of the importance of the part of the NL speech on which the text is written. As a result of document indexing by means of frequency-morphological analysis, index P will be obtained, for which the following statements are correct (formula 1):

$$P = F * V \tag{1}$$

The document indexing process can be represented as a number of steps:

1. Indexing of the document.
2. Carrying out morphological analysis of text.
3. Obtaining a combined frequency-morphological index.

3.1 The index of the document

The first step in automatic text summarisation is to apply a frequency indexing method that will allow the calculation of index F for each word in the text. The calculation of index F depends on the algorithm or method of indexing, in the following the obtained index will be used for calculations with the index of morphological analysis V. In this paper, we selected the following as the main algorithms.

TF-IDF. Luh in 1957 developed a method of analysing text information, which allows to identify the most significant, relevant words that were supposed to be used to classify documents in natural language [Luhn, 1958]. At the heart of the TF-IDF method is frequency analysis, and the hypothesis that the most important words in the test are, in more often than the rest of the words in the text. Thanks to this approach, the TF-IDF method can be used not only to classify documents, but also to expand its application, using it to reduce information redundancy. Sentences that do not contain the most significant words are removed from the text. The remaining test is then subject to linguistic analysis to agree on the remaining sentences in the text.

TF-ISF. TD-IDF modification [Luhn, 1958]: aimed at testing the hypothesis that the most important words are used more than once in a single sentence, but are rarely found throughout the document.

Collocations. Technology of identification of significant sentences in text, which is based on analysis of weight of phrases. Indexing of significant sentences is calculated as a sentence with a common word, of the total number of sentence. Position analysis of offers. This technology of indexing the most important sentences is limited to the hypothesis that all the main sentences are used at the beginning and end of the text to be indexed, thus, the largest index dials sentences at the beginning and end of the text, which are then auto summarised.

The signal method. Theory-based technology that key and most important sentences use specific words: meaningful, complex, heavy, tasks, goals, etc. The words are used from a special dictionary developed by H.P. Edmans.

Neural networks. Deep machine learning - which appeared relatively long ago, actively developing direction, which has found application to a wide range of problems: robotics, training and recognition of graphic information and intelligent search. One of the most important works in the field of in automatic summarisation in recent decades was the results of Collport's research [Shari, 2018], [Molchanov, 2015], which developed a unified procedure for machine analysis of text. Many modern automatic summarisation software use Collobert method [Collobert, 2008]. Using the Collobert approach allows to index by semantic and linguistic importance parts of the text: sentences, paragraphs, etc. automatic summarisation of text using depth learning in a neural network, differs from a conventional neural network by the number of layers, which contributes to more complex calculations.

3.2 Morphological analysis

The use of morphological analysis in the indexing of documents allows to apply more complex methods of calculation of indices of documents taking into account the special specificity of NL.

After investigation of many different texts on classification and automatic summarisation, it was found that the most used words in the sentence, and the most rarely found in the text, are much less important than the words located in a certain area of frequency of use (see Figure 1).

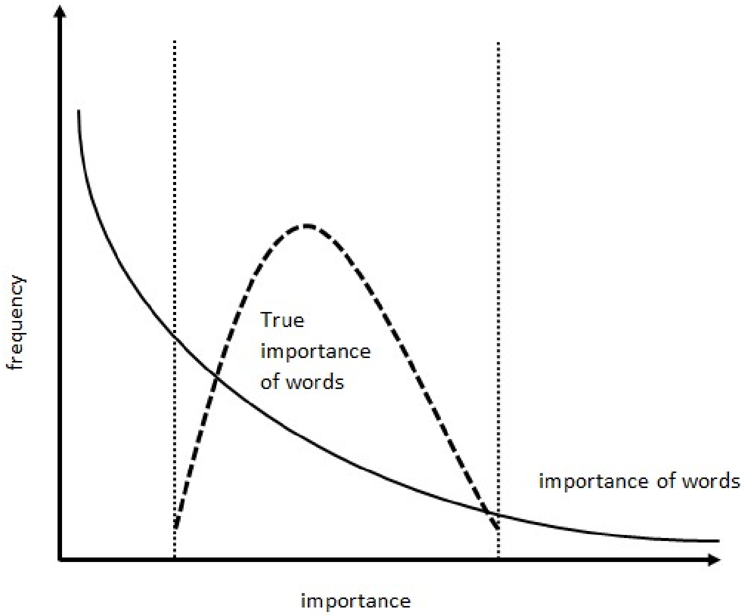

Figure 1: Frequency and significance of words in frequency analysis

Figure 1 presents the concept of evaluating of the meaning of a word in a text, used in most of TF-IDF modifications, which suggests that in natural language, auxiliary and little meaningful parts of speech that are most often used, do not carry a meaningful sense load to convey the content of the document, but merely complement the material presented. Rarely used words in the text are also unable to display the topic of the text, because as a rule the most rare words are synonyms, manifestations of the specific style of the author, etc., which are not ways to characterize the text as a whole.

In order to assess the significance of parts of speech, a small body of texts consisting of 200 artistic arbitrations of modern writers in Russian language was collected [Internet portal “Bookzip”], all books were related to different literary genres. The method of evaluation of the significance of parts of speech is based on the use of interrelated parts of speech in the text, for example, a verb and a noun, which are used in the sentence as subject and predicate, as well as on the uniform frequency of the use of parts of speech in text. “Solaris Engine” is a morphological analysis library used for identifying parts of speech, analysing bigrams, sentences, etc. This library was selected based on a number of researches [Fomin et al., 2019]. The morphological index of each word is calculated V by formula (formula 2):

$$P_n = \frac{S_n}{\sum_{n=1}^{n} w_n} + \frac{\sum_{i=1}^{i} Q_s}{\sum_{n=1}^{n} w_n \ * h} \tag{2}$$

Where Sn – is the frequency of use of the n part of speech in $D_j$text , w_n – n word in D_j

text , Q_s - frequency of use of combinations Sn part of speech witch other part speech, h – is the count of parts of speech in the NL on which the analysed text is written. Figure 2 shows the results of indexing parts of speech, based on 200 artistic texts in Russian.

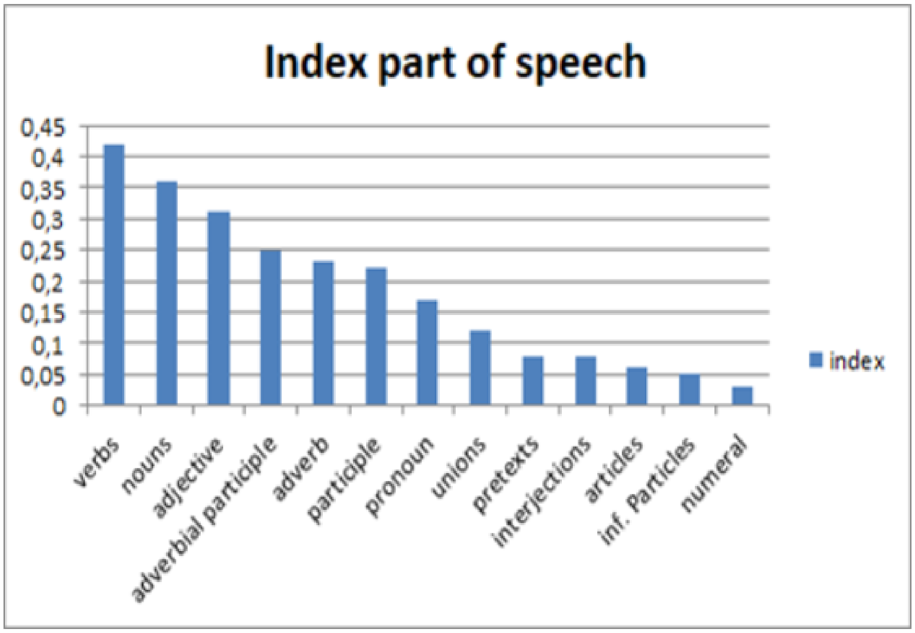

Figure 2: Indexes of parts of speech

Figure 2 Shows morphological indices calculated on the basis of formula 2. The results of morphological indexing of parts of speech in Russian language showed that the most commonly used part of speech are verbs, and nouns, as well as combinations thereof. The third most important part of speech was adjectives, which were often used in combination with nouns, further increasing the index of the given part of speech. The smallest index is received by service parts of speech, which are not included in the main parts of sentences.

4 Experiment

Based on the analysis of [Yogesh, 2014], [Tarasov, 2010], [Gambhir, 2016] methods for auto summarisation, the “Rouge” method was chosen [Yogesh, 2014], [Gambhir, 2016] because the method is easy to modify, has many varieties and is less prone to the element of chance. Accuracy calculations are carried out using the freely distributed “Rouge” application on [GitHub “Rouge”].

The “Rouge” method is based on the use of bigrams. A unique feature of this method is the detail of units of measure, Rouge - 1 for example, considers a word as a minimum unit, in Rouge - 2 a minimum unit is a bigram, this evaluation method is called “Rouge – N”. There are other types of evaluation methods, “Rouge - S” – takes measurements based on the bigrams taking into account changes in the text sequence, “Rouge - L” takes measurements based on the longest chain of bigrams of the matching sequence between the template and the text abstract, etc.

According to the “Rouge” metric, each summary is compared by two indicators called “F1 score” [Getahun, 2017] Precision, Completeness Recall, and based on these two indicators, another overall indicator is calculated - the measure of accuracy of the test measures. You can read more about methods of evaluating the results of summarisation on the Internet portal for natural language processing “RxNLP” [Internet portal “Portal”, “NLP text-mining”]. Frequency-morphological analysis is performed by special software. To date, there is quite a large number of morphological analysis libraries. In the study we will apply the morphological analysis methodology developed by the authors [Fomin et al., 2019].

The difference of calculating indices using frequency-morphological analysis is the limitation of text, i.e. bringing all words into the initial form, identifying chats of speech of each sentence member, breaking according to dictionaries of morphological modules, identifying parts of speech, determining grammatical basis of sentences and chains of parts of speech.

4.1 Text corpora

In order to conduct an experiment comparing frequency and frequency-morphological analysis in indexes, two special corpora of text were collected. For implementation of comparison, assessment of accuracy, completeness and efficiency of summarising, methods need a reference text, as a rule, the reference text means the paper, the composition, the essay written by the person.

An important element in auto summarisation, reduction of the text redundancy, is the preservation of the basis of the text, which allows to preserve the main theme of the text. Within the framework of automatic summarisation it is common to define two types of texts [Yogesh, 2014], [Tarasov, 2010], [Gambhir, 2016]:

- context-identifiable - these texts are expected to describe specific issues, problems, topics;
- context-indelible - in these texts, there is no clearly marked theme, it can be hidden, including from the reader, in the general context[1].

Context-identifiable corpora The corpora "Dissertation" refers to context-identifiable issue automatic summarisation and is represented by graduate works for obtaining the PhD degree, collected from various sites of universities of the Russian Federation in different directions and specialties, to each thesis an autoabstract is attached.

Table 1: Dissertation

Science field	Count Dissertation	Count autoabstract
IT	30	30
History	30	30
Chemistry	30	30
Jurisprudence	30	30
Biology	30	30
Medicine	30	30
Pedagogics	30	30
Physics	30	30
Philosophy	30	30
Economy	30	30

All dissertations and autoabstracts presented in the corpora of texts are published during the 2007 to 2019 period. The average size of one dissertation: 142 pages or 76,964 words, the average length of the autoabstract is 22 pages or 8,464 words. The works presented by one subject area have different topics and directions, for example, for Jurisprudence, the works discuss the problems of the civil code of the Russian Federation, the judicial document of production, the Customs Code of the Customs Union, etc. The reference text of the abstract in this corpora of texts, is considered the autoabstract to the thesis written by the author. As a comparison,

[1] As a rule, context-indelible group of texts includes artistic works

abstracts created by joint application of indexing technologies with morphological analysis are used.

Context-indelible corpora The "Art literature" corps refers to Context-indelible corps and is represented by various artistic works in Russian, different time eras. As the reference text, works and essays taken from the Internet with a retelling of the content of the artistic ration are used. The body "Art literature" is presented in Table 2

Table 2: Art literature

Author	Count works	Count essays
F. M. Dostoevsky	20	20
A. I. Kuprin	20	20
L. N. Tolstoy	20	20
A. P. Chekhov	20	20

The second comparative corpora of texts of automatic summarisation is made similar to comparative abstracts in the corpora "Dissertation" where abstracts were created using different indexing technologies together with morphological analysis.

Table 3: Evaluation automatic summarisation context-identifiable corpora corpora based on frequency indexing methods

Method	Evaluation	Rouge N					
		1	2	3	4	5	6
TF-IDF	Precision	0,2110718	0,1731205	0,144557	0,0861217	0,0643702	0,1828707
	Recall	0,1750908	0,143609	0,1199146	0,0714407	0,0533971	0,151697
	M-measures	0,191405	0,1569898	0,1310878	0,0780972	0,0583724	0,1658315
TF-ISF	Precision	0,2036052	0,1864228	0,1415896	0,0829204	0,0702028	0,1881112
	Recall	0,168897	0,1546437	0,117453	0,0687851	0,0582354	0,1560442
	M-measures	0,1846341	0,1690527	0,1283968	0,0751942	0,0636615	0,1705838
Collocations	Precision	-	0,2406241	0,2413066	0,2201747	0,184369	0,0403017
	Recall	-	0,1710078	0,1714928	0,1564747	0,1310282	0,0286418
	M-measures	-	0,1999291	0,2004962	0,1829382	0,153188	0,0334857
Position analysis of offers	Precision	0,1940082	0,1577003	0,1471301	0,1185674	0,1773649	0,0382542
	Recall	0,1378786	0,1120751	0,104563	0,084264	0,1260504	0,0271866
	M-measures	0,161197	0,1310296	0,122247	0,098515	0,1473684	0,0317845
The signal method	Precision	0,1837544	0,1677501	0,1347171	0,1243346	0,1748481	0,1939582
	Recall	0,1305914	0,1192174	0,0957413	0,0883626	0,1242618	0,137843
	M-measures	0,1526773	0,1393798	0,1119334	0,1033068	0,1452773	0,1611554
Neural networks	Precision	0,1778164	0,1273729	0,1199331	0,1039203	0,1449691	0,1275056
	Recall	0,1778225	0,1273772	0,1199372	0,1039239	0,144974	0,12751
	M-measures	0,1778195	0,1273751	0,1199352	0,1039221	0,1449715	0,1275078

4.2 Evaluation of automatic summarisation of context-identifiable corpora

We will evaluate the automatic summarisation generated by various indexing methods, which are based exclusively on frequency analysis. Results of evaluation of automatic summarisation by “Rouge-N” metric the results are presented in Table 3 below.

Table 4: Evaluation automatic summarisation context-identifiable corpora based on frequency-morphological indexing methods

Method	Evaluation	Rouge N					
		1	2	3	4	5	6
TF-IDF	Precision	0,361573	0,381362	0,297024	0,304661	0,333702	0,292211
	Recall	0,113437	0,119649	0,09319	0,095561	0,104694	0,091657
	M-measures	0,172694	0,18214	0,141873	0,145484	0,159387	0,139563
TF-ISF	Precision	0,403672	0,429321	0,332682	0,342371	0,374109	0,324925
	Recall	0,113936	0,121172	0,093881	0,096635	0,105591	0,091721
	M-measures	0,17771	0,189003	0,146432	0,150721	0,164691	0,143054
Collocations	Precision	-	0,435501	0,33628	0,345765	0,378211	0,3322
	Recall	-	0,5421	0,419126	0,430326	0,470701	0,413442
	M-measures	-	0,483016	0,373494	0,383449	0,419412	0,368392
Position analysis of offers	Precision	0,365606	0,327589	0,302069	0,300216	0,340532	0,303881
	Recall	0,224472	0,201131	0,185469	0,184353	0,209048	0,186571
	M-measures	0,278161	0,24923	0,229811	0,228449	0,259028	0,231201
The signal method.	Precision	0,438151	0,464082	0,360248	0,369892	0,405142	0,356213
	Recall	0,300237	0,318511	0,246854	0,253461	0,277611	0,244101
	M-measures	0,356321	0,378001	0,292976	0,300803	0,329477	0,2897
Neural networks	Precision	0,435757	0,46195	0,358354	0,371772	0,403731	0,354252
	Recall	0,525716	0,557341	0,432354	0,448514	0,487161	0,427411
	M-measures	0,476555	0,505182	0,391897	0,406526	0,441572	0,387411

The absence of Rouge-1 indexing in the “Collocation” indexing method is a consequence of the inability to use unigrams in text indexing.

In the evaluation of automatic summarisation by the Rouge-1 method, the highest accuracy is achieved in autoabstract generated by the method of Neural Network indexing, but the best overall correspondence (m-measures) was achieved by TF-IDF This situation can be explained by the fact that the “Recall” of autoreferences obtained by TF-IDF indexing is less than the “Recall” of autoabstract, but the number of words that often coincided with the the reference increased, while neural networks used service parts of speech more often (43.34%) than in the TF-IDF method.

When using bigrams (Rouge-2), the best “Precision” and “Recall” was shown by the “Collocations” method, which suggests the presence of coherence in words in the referenced texts. Increasing the Rouge-3 sequence, after the bigrams, decreases the accuracy of almost all methods except the phrase. Increase the sequence of n-grams resulted in reduced “Precision” and “Recall” with the maiming of the chain of n-grams. The best result when using 6 words long n-grams, was shown by the TF-IDF method.

Table 4 presents the results of automatic summarisation of dissertation, in which indexing was carried out on the basis of frequency-morphological analysis. As with frequency analysis, after indexing and shortening the text, morphological libraries were used to reconcile sentences.

The best indicator of “Recall” in the evaluation of unigrams (Rouge-1) was shown by the method of positional analysis of sentences method.

When using the position analysis of offers method the “Recall” of the main sections “Introduction”, problem, “Conclusions”, almost did not decrease, because positional they are located at the beginning and end of the texts, in case the volume of conclusions was sufficiently small, the abstract included relevance, problem and methodological part of the dissertation.

With high “Recall”, the number of words matching the reference text was extremely small, which made the overall “Precision” of the summarisation of the method low.

The lowest estimate was found in auto summarisation, where the indexing of texts was carried out with method TF-IDF. The average accuracy of TF-ISF is higher than that of TF-IDF, this result indicates that, with increasing text volume, words that are often used within a single sentence are often used in autoreferences written by humans.

Better “Precision” is achieved with any method when using unigrams. When evaluating autoreferences with unigrams, the similarity with the reference falls, except for the collocations method. Neural networks reached the best M-measures in auto summarisation, with neural networks generating the largest volume abstracts.

4.3 Evaluation of automatic summarisation of context-indelible corpora

We will evaluate automatic summarisation generated by various indexing techniques based on frequency analysis. Results of evaluation of automatic summarisation by “Rouge-N” metric the results are presented in Table 5.

Table 5: Evaluation automatic summarisation context-indelible corpora based on frequency indexing methods

Method	Evaluation	Rouge-N					
		1	2	3	4	5	6
TF-IDF	Precision	0,309992	0,226085	0,093911	0,076511	0,158793	0,038357
	Recall	0,270655	0,299204	0,170447	0,085492	0,143018	0,039314
	M-measures	0,337491	0,107212	0,235627	0,060411	0,199072	0,160076
TF-ISF	Precision	0,506626	0,458609	0,229601	0,195824	0,145352	0,078777
	Recall	0,251234	0,098935	0,239602	0,158155	0,179512	0,143286
	M-measures	0,292377	0,229081	0,092807	0,178233	0,082385	0,045001
Collocations	Precision	0,278616	0,298429	0,206652	0,063323	0,145968	0,045118
	Recall	0,552911	0,333939	0,333282	0,279554	0,23204	0,122982
	M-measures	0,360896	0,305001	0,250727	0,222087	0,232677	0,063547
Position analysis of offers	Precision	-	0,267986	0,126883	0,105382	0,118786	0,092934
	Recall	-	0,318982	0,269306	0,300199	0,170863	0,24958
	M-measures	-	0,157603	0,141343	0,231817	0,282642	0,117228
The signal method.	Precision	0,249845	0,111706	0,068327	0,152547	0,119657	0,110365
	Recall	0,334284	0,258972	0,194206	0,085959	0,130805	0,067205
	M-measures	0,291674	0,245833	0,14378	0,132268	0,176823	0,133074
Neural networks	Precision	0,273879	0,182688	0,092082	0,108367	0,220331	0,090309
	Recall	0,453305	0,440339	0,323477	0,123206	0,130873	0,131392
	M-measures	0,433696	0,244462	0,186702	0,120517	0,16168	0,205693

As a result of estimation by the method of Rounge-1 of automatic summarisation generated on the basis of frequency indexing methods, the best method was neural networks, where the value of the function reached 43.36%. The best correspondence when evaluation the match with the standard of bigrams, showed automatic summarisation generation by the method of “Collocations” where the text of the reference was similar to the writing in 30.5% of cases. In the evaluation of n-grams of Rouge-3-6, automatic summarisation obtained by the “Collocations” indexing method also have the best m-measure value with the reference. Now we will carry out comparative analysis by Rouge-N procedure, using frequency-morphological analysis in indexing, results of which are presented in Table 6.

Table 6: Evaluation automatic summarisation context-indelible corpora based on frequency-morphological indexing methods

Method	Evaluation	Rouge-N					
		1	2	3	4	5	6
TF-IDF	Precision	0,426539	0,522639	0,502415	0,446527	0,509916	0,427969
	Recall	0,388721	0,476286	0,459437	0,406925	0,464692	0,390012
	M-measures	0,406747	0,498387	0,480756	0,425807	0,486255	0,408117
TF-ISF	Precision	0,600605	0,732799	0,68317	0,634922	0,712932	0,605985
	Recall	0,40032	0,488431	0,455352	0,423193	0,475189	0,403905
	M-measures	0,480424	0,586166	0,546468	0,507874	0,570274	0,484727
Collocations	Precision	-	0,558879	0,519882	0,483895	0,513947	0,441059
	Recall	-	0,970924	0,903175	0,840656	0,892864	0,765423
	M-measures	-	0,70941	0,659909	0,61423	0,652376	0,551926
Position analysis of offers	Precision	0,413795	0,494613	0,479276	0,438813	0,448022	0,392228
	Recall	0,635545	0,759672	0,736116	0,673971	0,688101	0,602825
	M-measures	0,501239	0,599136	0,580558	0,531545	0,542697	0,475178
The signal method	Precision	0,315449	0,376193	0,367336	0,342419	0,360224	0,314927
	Recall	0,391679	0,467101	0,456104	0,425166	0,447273	0,391031
	M-measures	0,349455	0,416747	0,406936	0,379332	0,399057	0,348877
Neural networks	Precision	0,450621	0,535067	0,509098	0,549532	0,546561	0,541471
	Recall	0,576912	0,685024	0,651778	0,703543	0,699739	0,693223
	M-measures	0,506005	0,60083	0,571672	0,617073	0,613737	0,608021

Evaluation of automatic summarisation generated on the basis of frequency-morphological analysis using the Rouge-1 technique showed that the highest compliance with the reference, was achieved in methods where the method of neural networks was used. The total compliance with the standard of 50.6% was achieved, which is 17.94% better than when using frequency analysis method. The “Precision” indicator reached 97%, which allows saying that almost all words used in automatic summarisation are also found in the reference text.

Conclusions

The experiments results show that moving from simple unigram frequency analysis to more complex frequency-morphological analysis has a great impact on automatic text summarisation quality. The “Rouge-N” method, used for evaluating the efficiency of automatic text summarisation, showed that autoabstracts made on base of frequency-morphological analysis were 16% closer to the original than autoabstracts based on frequency analysis only, moreover automatic text summarisation of fiction was 31,28% more accurate using frequency-morphological analysis than using frequency analysis.

While using frequency-morphological analysis in indexing documents was found that all methods except TF-IDF increased the similarity with the original text. Comparing to the original texts, which were written by people, it was found that text on average is accurate by 48% and the text reduction reached 93,05% while dissertation referencing.

The experiments results let us suppose the high potential of using index methods on base of neural networks using frequency-morphological analysis in smart search or in information-educational fields. We are expecting to expand the scope of application fields of frequency-morphological analysis in fiction auto reference and in indexing of parts of speech usage frequency, in order to classify data on NL.

Acknowledgements

The research was supported by the Russian Science Foundation (RSF), Project “Digitalisation of the high school professional training in the context of education foresight 2035” No 19-18-00108.

References

[Brandow et al., 1995] Brandow R. Mitze K., and Lisa F. R. (1995) Automatic condensation of electronic publications by sentence selection // Inf. Process. Manag. Vol. 31. Pp. 1-8.

[Baxendale et al., 1958] Baxendale P. B. and etc (1958) Machine-made index for technical literature: An experiment // IBM J. Res. Dev., Vol. 2, Pp. 354-363.

[Lei, 2017] Lei L. and etc. (2017) Redundancy checking algorithms based on parallel novel extension rule. //Journal of Experimental Theoretical Artificial Intelligence Vol. 29 , 2017 - Issue 3.

[Said et al., 2017] Said A.S. and etc (2017) Using Text Mining Techniques for Extracting Information from Research // Intelligent Natural Language Processing: Trends and Applications Vol. 1. Pp.373-397.

[Salloum et al., 2017] Salloum S.A. (2017) A Survey of text mining in socialmedia: facebook and twitter perspectives //Advances in Science, Technology and Engineering Systems Journal Vol.2. Pp.127-133.

[Clayton et al., 2011] Clayton S.and etc (2011) Experiments in Automatic Text Summarisation Using Deep Neural Networks // Machine Learning, Fall Vol.1 2011. Avaible at: https://www.semanticscholar.org/paper/545-Machine-Learning-%2C-Fall-2011-Final-Project-in-Ben-Rahul/8f4f64e15553baf9fd0c2933c631b78c97c8f0bc

[Radev et al., 2002] Radev D. R., Hovy, E., and McKeown K. (2002) Introduction to the special issue on summarisation. //Comput. Linguist, Vol. No 28. Pp. 399–408.

[Dragomir, 2012] Dragomir R.R. (2012) Single-document and multi-document summary evaluation via relative utility University of Michigan, Ann Arbor MI 48109 2012 Avaible at: https://www.eecs.umich.edu/techreports/cse/2007/CSE-TR-538-07.pdf

[GitHub “Rouge”] Application «Rouge», «GtiHub» Avaible at: https://github.com/kylehg/summariser/blob/master/rouge/ROUGE-1.5.5.pl

[Derczynski, 2016] Derczynski L. (2016) Complementarity, F-score, and NLP Evaluation // Proceedings of the International Conference on Language Resources and Evaluation, Vol. 1. Pp. 1-6.

[Sujit et al., 2013] Sujit R. Sujit V. and etc (2013) Classification of News and Research Articles Using Text Pattern Mining IOSR Journal of Computer Engineering (IOSR-JCE) Vol. 14, Issue 5 . Pp. 120-126.

[Luhn, 1958] Luhn P. (1958) The automatic creation of literature abstracts IETE // Journal of research J. Res. Dev., Vol. 2, no. 2. Pp. 159-165.

[Fomin et al., 2019] Fomin V., Osochkin A., and Zhuk Y. (2019) Frequency and morphological patterns of recognition and thematic classification of essay and full text scientific publications //NESinMIS-2019, 12-Jul-2019, CEUR-WS 2019. Vol. 2401 , 69-84. 2019.

[Evdokimenko, 2013] Evdokimenko E. Y. (2013) The Concept of Information Noise in the Social and Human Sciences // Molodoy ucheniy. Vol. 10. Pp. 564-566, 2013. Avaible at: https://moluch.ru/archive/57/7765/

[Al-Emran, 2017] Al-Emran M., Shaalan K.(2017) Academics'awareness towards mobile learning in Oman // Int.J. Com. Dig. Sys. Vol.6

[Shari, 2018] Shari T. (2018) Optimize Optimize the A Commentary. Journal Search Voice, Vol. 1, Pp. 1-6.

[Molchanov, 2015] Molchanov A.N and etc.(2015) A mathematical model of natural language text that takes into account the coherence property // Internet-journal "Science of science". Vol. 7, No 1, 2015 Avaible at:https://naukovedenie.ru/PDF/70TVN115.pdf

[Jansen, 2010] Jansen, B. J. and Rieh, S (2010) The Seventeen Theoretical Constructs of Information Searching and Information Retrieval// Journal of the American Society for Information Sciences and Technology. Vol 61. Pp. 1517-1534.

[Yogesh, 2014] Yogesh M . et al. (2014) Analysis of Sentence Scoring Methods for Extractive Automatic Text Summarisation // Proceedings of the International Conference on Information and Communication Technology for Competitive Strategies. – ACM: NY, USA, vol. 1. Pp. 89-97, 2014.

[Gambhir, 2016] Gambhir M. Gupta. V. (2016) Recent automatic text summarisation techniques: a survey // Artificial Intelligence Review. vol.1 Pp. 1-66.

[Tarasov, 2010] Tarasov S.D. Modern methods of automatic referencing // Scientific and technical statements of SPBPU. //Journal "Computer science, telecommunications and management", vol. No 6 1-9 2010.

[Internet portal "Portal", "NLP text-mining"] Avaible at: http://rxnlp.com(update:30.09.2019).

[Internet portal "Bookzip"] Avaible at: https://bookzip.ru/boeviki-ostrosjuzhetnaja-literatura/

[Getahun, 2017] Getahun T. and etc. (2017) Automatic Amharic Text Summarisation using NLP Parser international. //Journal of Engineering Trends and Technology (IJETT) – Vol.53, Pp. 52-58.

[Collobert, 2008] Collobert, R. and Weston, J.(2008) A unified architecture for natural language processing: Deep neural networks with multitask learning. // Conference: Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June vol. 1, Pp. 1-8.