UDC 519.7

Kanishcheva O.V.
National Technical University “Kharkiv Polytechnic Institute”

USING THE TEXT RELATIONSHIP MAP (TRM) FOR AUTOMATIC SUMMARIZATION

This paper is devoted to the use of statistical TFIDF, TLTF and Text Relationship Map (TRM) methods for automatic construction of abstracts for the Russian and Ukrainian languages. These methods are implemented using C++ software programming language in Borland Builder 6.0 and databases created in Microsoft Access.

Keywords – automatic processing of natural language, automatic abstracting, TFIDF, TLTF, Text Relationship Map (TRM).

У статті запропоновано використання статистичних методів TFIDF, TLTF та Text Relationship Map (TRM) для автоматичної побудови реферату для українсько- та російськомовних текстів. Ці методи програмно реалізовані за допомогою мови програмування C++ у середовищі Borland Builder 6.0 та бази даних, створеної в Microsoft Access.

Ключові слова – автоматична обробка природної мови, автоматичне реферування, TFIDF, TLTF, Text Relationship Map (TRM).

Introduction

The civilization development is a reason of steady increase of volume of the knowledge accumulated by humanity. Millions of paperbacks and manuscripts contain information on different subjects, different fields of scientific and cultural life, but they are more and more often replaced by electronic media. Now there are already electronic versions of many books, popular printed editions appear both in a print and electronic format, the amount of online documents is increasing exponentially. As a result there are many problems appearing, such as information classification, analysis, search, and their solution is related to natural language intellectual processing on a large scale.

Artificial intelligence development as a scientific field became possible only after creation of the computer (in the 1950s and the 1960s), when it gathered together mathematicians (theorists and practitioners), psychologists and experts in robotic engineering, electronics, cybernetics, for them to teach the computer to think and behave as a human (natural intellect). At the same time, another scientific field appeared that was named as computational linguistics (CL). As a scientific field, it was supposed to teach the computer to understand and process a human language (natural language texts).

Despite a lot of theoretical developments, the problem of creation of effective industrial systems within the framework of each of directions determining the modern development level of such major scientific and technical industry as informatics is exceptionally important up to the present moment.

Automated data processing involves automatic abstracting and annotating of scientific and technical texts. As there is no use keeping everything in electronic format, that it is done by a man, in fact technical descriptions become antiquated, it is enough to leave only external information about them: author, theme, that it is done. It concerns streams of information even more, it is necessary to sort them in different arrays according common themes and sources they are from, it is necessary to compress information content, formalize records, placing them in the knowledge bases where you can obtain them and give out answers for queries from. This class of problems is one of most difficult among the other tasks of automated texts processing, as it needs deep linguistic analysis that is supposed to educe the most informative and important parts of text content. And it is in the area of difficult intellectual systems.

The simplest method of text compression is an automatic extraction of those sentences, that contain one or more keywords or word-combinations that are so called "climaxes" in text semantic information distribution. These sentences displayed on the screen in the order of their following form a document machine report or annotation.

Task of reviewing as process of the automated treatment of language

An annotation and a report effectively provide a rapid exchange of new scientific and technical information, these are them that substantially abbreviate time of specialists spent on information processing. Annotating and reviewing is supposed to reduce the volume of information generator to the maximum and to substantially keep its basic content.

Language redundance and absence of univocal correspondence between the content and the form of a language piece of work are the fundamental basis for such information compression. At reviewing a report gets rid of everything second-rate and illustrative, a basic idea of what explains is kept only.

An annotation and a report are supposed to give only the most substantial information about new achievements of science and technology. If a report and an annotation interest a reader, and the information in them is not enough for them, it is always possible to find the original data source after indicated bench-mark data in them and get necessary information on a full scale. Thus, an annotation and a report perform an important function: they carry out systematization of necessary to the user information.

A report and an annotation belong to the secondary documentary scientific information sources. These are documents in which information is reported about primary documents. Information processing includes the process of study of every primary document or their complex, for example, a collection of articles, and preparation of the information that represents the most essential elements of these documents. On the basis of secondary documents use informing editions are completed, such as, abstract journals, reference materials, scientific translations etc. An annotation and report compress original sources in fundamentally different ways. While an annotation only enumerates the questions from the original source, not covering these questions content, a report not only enumerates all these questions but also focuses attention on the content of each of them. It is possible to say, that an annotation reports only what the original source is about, and a report informs about what is written in relation to each of lighted up in an original source and a report informs of that is written concerning each of the questions covered in the original source. Thus, an annotation is only a pointer for original sources selection and cannot replace them, while a report can fully replace an original source, as it reflects the material content. As it was said above, both for an annotation and for a report a certain degree of rolling up of information on the basis of its previous analysis is characteristic.

Annotating of text documents

An annotation (from lat. annotatio meaning “a remark”) is a short description of a printed work or manuscript content. It is a compressed to the maximum description of the original source. It covers the publication topic in a generalized way without the complete covering of its content. An annotation gives an answer to a question, what the primary information source is about.

According to content and intended function annotations can be referral and recommendatory. Referral annotations cover document subjects and give certain information about it, but it does not give a critical estimation. Recommendatory annotations contain a document estimation from the point of view of its serviceableness for the certain category of readers.

According to the scope of the annotated document content and reader serviceableness general and specialized annotations are distinguished. General annotations characterize a document in general and created for a wide range of readers. Specialized annotations cover expose certain document aspects that interest a domain expert. They can be quite short, consisting only a few words or small phrases, and unfolded to 20-30 lines, but also in this case, unlike a report, they give only the most substantive provisions and conclusions of a document in a condense form. Only substantial signs of the document content are specified in an annotation, id est those that allow to educe its scientific and practical value and novelty, to distinguish it from the others, similar to it on subjects and serviceableness.

While doing an annotation one must not retell documents content (conclusions, recommendations, actual material). It is necessary keep to a minimum the use of complicated expressions, pronouns and demonstratives.

General requirements for writing annotations are the following:

1. Sphere an annotation intended use. The scope plenitude and content of the finishing part depend on it.
2. The volume of annotation must be within 500 to 2000 printed signs.
3. Observance of structure logic that can differ from the statement order in the original.
4. Observance of language features of an annotation that includes the following :

- covering the conceptual issues of the original in a simple, clear and short way;
- avoiding repeating, including the article title;
- observing the unity of terms and reductions;
- using generally accepted reductions;
- using impersonal constructions as "to be examined..., analysed..., reported". and the passive voice;
- avoiding to use adjectives, adverbs, introductory words that do not influence the content;
- using certain summarizing words and word-combinations that provide logical connections between separate parts of expressions as "as shown", "… however", "thus,…" etc.

An annotation content:

Introduction is a bibliographic description.

Basic part is a list of basic problems mentioned in the publication.

Finishing part is a short description and estimation, intended use of the annotated work (who this publication is addressed to).

Thus, an annotation is a short generalized sketch (description) of a book or an article text. Before the annotation text bench-mark data is given (author, name, place and time of edition) in a nominative form. These data can be included and in the first part of an annotation. An annotation usually consists of three parts.

Example annotations:

Башмаков А. И., Башмаков И. А. Интеллектуальные информационные технологии / А. И. Башмаков, И. А. Башмаков. – М. : Изд-во МГТУ им. Н.Э. Баумана, 2005. – 304 с.

Интеллектуальные информационные технологии — одна из наиболее перспективных и быстро развивающихся научных и прикладных областей информатики. В учебном пособии рассматриваются ее основные направления: обработка текстов на естественном языке, моделирование знаний и базы знаний, управление знаниями, распознавание образов, нейротехнологии, интеллектуализация Internet, концептуальное программирование и др. Основное внимание уделяется математическим моделям, методам и инструментальным средствам разработки программного обеспечения интеллектуальных автоматизированных систем.

Task of reviewing of text documents

A report (from lat. "refero" meaning "report") is short exposition of maintenance of scientific work, literature on the topic, done in writing or in form public lecture, that exposes him basic maintenance in relation to all questions, that is also accompanied by an estimation and conclusions of reviewer. He must give to the reader an objective idea about character of work, to show the most material points of her maintenance.

Unlike an annotation a report doesn't not only gives an answer for a question about what the primary printed document is about but also what is said, i.e. what basic information is given in the reviewed original source. A report gives description of the primary document, informs about appearance and existence of the corresponding primary documents, also it is a source of reference data and independent means of scientific information. A report can be done in written and oral form.

An purpose of a report is give a reader relatively complete idea about the issues covered in the original source and to release a user of the necessity of complete revision of the original source.

Two basic types of reports are distinguished:

- an informing report (report-compendium);
- an indicative report (report-resume).

An informing report contains all substantive points of original, information about the research methodology, equipment use and application domain in a generalized form. An informing report is the most widespread form. One mustn't write all the points in an informing report, but only those, what closely related to the theme of reviewed document.

The reports made from one source are named monographic. The reports made from a few sources on the same theme are called correlated reports.

Among numerous types of reports it is necessary to distinguish specialized reports in which presentation is oriented to the specialists of a certain area or certain sort of activity (for example, teachers of linguistics) and takes into account their interests.

At all this variety reports have certain common features. Reasoning and historical digressions are not used in a report. Material is given in the form of a consultation or description of facts. Information is given in an exact and short form, without any content distortions and subjective estimations. Conciseness is obtained due to the use of terminological vocabulary, also tables, formulas, and illustrations. The report text must not be a brief translation or mechanical exposition of the reviewed material. It must underline what deserves some special attention from the point of view of novelty and possible use in the future productive or research work. The report text must not contain repetitions and general phrases. It is not acceptable to use direct speech and dialogues. It is necessary to add the original source author's main conclusions in text of report.

They help describe original source documents content with maximal exactness. It is reasonable to use reduction of terms in reports. The system of reductions allows to attain the considerable economy of place without any content loss. Such reductions can be generally accepted in the language (for example, adj. – adjective) and typical for this source.

It is common for the report language to use certain grammatical and stylistic means. First of all these are simple complete sentences that assist rapid report perception. It is possible to use verbal adverb phrases providing volume economy to describe different processes. The use of the impersonal sentences allows a reader to focus only on the substantial information, for example, "it is analysed, applied, examined etc".

The report volume depends on the primary document volume and report character and can be 1/8 or 10-15 % from the original source volume.

An analysis of existing methods of annotating and reviewing is for full-text documents processing

Almost simultaneously with works related to machine translation the researched of the use of computer for the aims of the automatic abstracting of scientific and technical texts began. The first machine experiment in this field was made in 1957 in the USA. Unlike machine translation, where researchers' attention, at least on the initial stage, was focused on separate sentences, as machine translation was thought as translation "phrase after a phrase", in the area of automated reviewing their attention was focused on larger text areas (mostly on paragraphs) where judgements on one theme were concentrated. In other words, the researchers' attention in this area from the beginning was focused to find out some common factors determining text semantic unity.

On the first stage of these works, the approaches based on finding out some common statistical factors of distribution of terms in the text or their mutual location in it were the most popular [1, 2]. In the future the research in area of the automated reviewing was displaced toward the use of underlying text structures and finding out that informative basis that organizes the whole text [3, 4]. The work in this direction substantially influenced the use of computer for creation of artificial texts.

The process of reviewing is done in three stages: analysis of initial text, determination of its characteristic fragments, forming of corresponding conclusion. Most modern works are concentrated round development of technology of one document reviewing.

The method of extracts stowage concentrates on the selection of characteristic fragments (as a rule, sentences). For this purpose, blocks with most lexical and statistical relevancy are distinguished with the help of the method of comparison of phrase templates. The creation of the final document in this case is simply a connection of chosen fragments.

The model of linear weigher coefficients (represented in a formula 1) is used in most methods. The basis of the analytical stage in this model consists of the procedure of setting of weigher coefficients for every block of text according to such factors, as a location of this block in the original, frequency of appearance in the text, frequency of use in the key sentences, and also indexes of statistical meaningfulness. The total of individual weighers, determined as a rule after additional modification in accordance with the special parameters of setting, related to every weigher, gives gross weight of the whole block of the text U :

$$Weight(U) := Location(U) + KeyPhrase(U) + StatTerm(U) + AddTerm(U). \tag{1}$$

The weigher coefficient of location (Location) in this model depends on where this fragment appears in the whole text or in the separately taken paragraph – at the beginning, in the middle or at the end, and also whether it is used in the key parts, for example, in the introduction or in the conclusion.

Key phrases are lexical or phrase summing up structures, such as "in conclusion", "in this article", "according to the results of analysis" et cetera. The weigher coefficient of a key phrase can also depend on the term meaning in this subject domain.

Moreover, an index of statistical importance (Statterm) is taken into account in this model at determination of weigher coefficients. Statistical importance is calculated on the basis of the data, got as a result of analysis of automatic indexation, where researchers find out and estimate a number of metric data that determine the weigher coefficients of a term. These metric data allow to distinguish a document from a number of other in a certain set of documents.

One group of metric data, for example, metric data of TFIDF, characterizes balance between frequency of appearance of a term in a document and frequency of its appearance in a set of documents (as a rule, used with other metric data of frequency and facilities of normalization of length).

The basis of TFIDF methods is determination of a weigher coefficient for every element of the structure (formula 2). The degree of element influence on the distance between objects proportional to the size of the coefficient is determined according to the following formula:

$$K_{fr}(i,j) = 1 - \frac{ITF(i,j) \cdot IDF(i,j)}{N \cdot |D|}, \tag{2}$$

where N – is an amount of elements in the chain; $|D|$ – is a number of chains in the set of data; $ITF(i,j)$ – is the size that is reverse to the frequency consisting of “i” and “j” elements in the chain; $IDF(i,j)$ – is the size reverse to the frequency of chains appearence (containing this pair of elements) in the set of data.

This model gives a possibility to revise terms in a text block and determination this block weigher coefficient in accordance with the additional presence of terms (Addterm) – or they appear also in a title, in a running headline, the first paragraph and in the profile of query intended for a user. Excretions of priority terms reflecting a user’s interests most preciselyis one of the ways of a report or annotation creation for a certain person or group. On fig. 1 the generalized architecture of reviewing is brought without any support on knowledge.

On the analytical stage the model of linear weigher coefficients (Naive – bayes of Method) is used, it assumes implementation of sequence of calculations of frequency and operations of comparison of lines or templates that give out weigher coefficients of four types(Location, Cuephrase, Statterm, Addterm) for every block of initial text, according to the following formula [5]:

$$\mathrm{P}(\mathrm{s} \in \mathrm{S} \mid \mathrm{F}_1, \mathrm{F}_2, .. \mathrm{F}_\mathrm{k}) = \frac{\prod_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{P}(\mathrm{Fi} \mid \mathrm{s} \in \mathrm{S}) \cdot \mathrm{P}(\mathrm{s} \in \mathrm{S})}{\prod_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{P}(\mathrm{Fi})}, \tag{3}$$

where s is a sentence; S – is a block of initial text; $F_1, ..., F_k$ are coefficients.

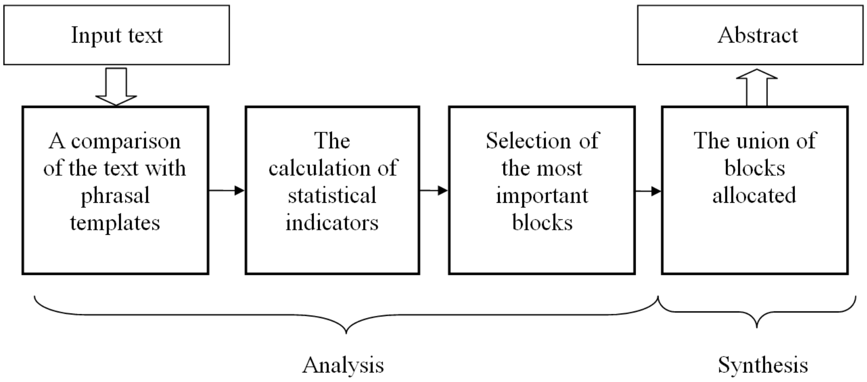

Input text
Abstract
A comparison of the text with phrasal templates
The calculation of statistical indicators
Selection of the most important blocks
The union of blocks allocated
Analysis
Synthesis

Fig. 1. Generalized architecture of reviewing without any support on knowledge

Then these coefficients are summarized for every block, whereupon n blocks are chosen they have the greatest total of coefficients (the value of n can be determined according to the degree of compression) to be used in a report.

This method was created as early as in the 1960 – 1970s but most systems that prepare such a summary on the basis of extracts, use the approach illustrated on fig. 1 until now. The analysis of comparative descriptions of different models, conducted with the aim of determination of the productivity each of them, showed that localization of blocks of text can be considered as one of the most useful functions, especially in the combination with the function of key phrases determination.

In many systems a user sets parameters depending on their needs at the moment, because descriptions can be so different for texts of different styles. Trying to automatize this process and, maybe, increase the productivity, the researchers from Xerox PARC, such as Julian Kupiach (1995) and his colleagues, invented a classifier capable to learn the rules of fragments selection. On fig. 2 it is shown how this classifier uses a set of certain reports used by a user and corresponding initial texts for the automatical determination of the criteria for an adequate fragments choice [6].

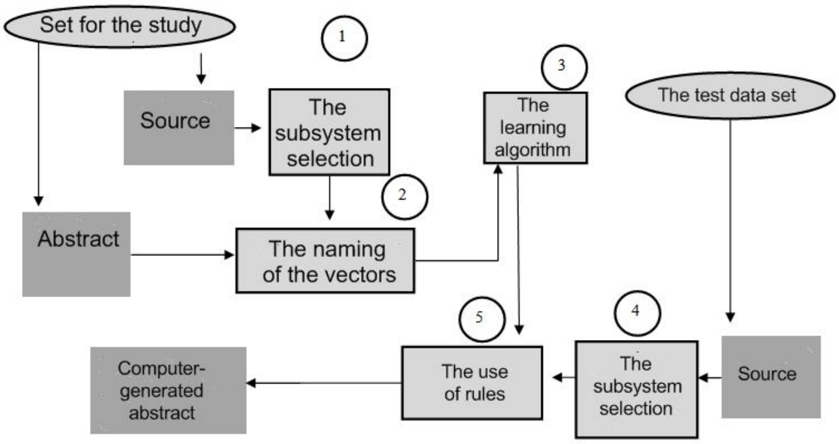


Fig. 2. Automatical determination of the criteria for an adequate fragments choice

This method, used by Inxight reviewing systems, is good for texts of different styles, but for this purpose users must have complete texts and corresponding reports for every style to do this.

The main advantage of the model of linear coefficients consists in simplicity of its realization. However, the selection of sentences (or paragraphs), that does not take intercommunications between them into reports. Some sentences can be skipped, or there can be words or word-combinations that are impossible to understand without other word or phrase. For example, if in the text there is some point explanation consisting of a few phrases and only one of them appears in the report, the meaning can be lost. The following text fragment illustrates this problem. "Bill Dickson started to work in Procter &

Gamble in 1994. In 1996 he became its vice-president". In this fragment it is possible to specify two words that are potentially unclear, these are the words "he" and "its", that does not have any meaning without the previous phrase that helps understand that "he" is Dickson, and "its" refers to Procter & Gamble company. If the first phrase is lost in the report, the text will lose its informive function. There are many researches attempting to solve this problem mainly by means of different sorts of "shreds". In a number of approaches a special window is created for the previous sentence of a report, by means of that it is possible to define the presence of semantic break or an unclear word. In other cases the sentence containing unclear words, are eliminated from a report, or attempts to find hints about the object true by means of short linguistic analysis are carried out. In this approach the degree of compression diminishes, as further information is put into the report. Moreover, when a basic report is already formed, it is difficult to restore the initial percent of compression.

This method, used by Inxight reviewing systems, is good for texts of different styles, but for this purpose users must have complete texts and corresponding reports for every style to do this. The main advantage of the model of linear coefficients consists in simplicity of its realization. However, the selection of sentences (or paragraphs), that does not take intercommunications between them into reports. Some sentences can be skipped, or there can be words or word-combinations that are impossible to understand without other word or phrase. For example, if in the text there is some point explanation consisting of a few phrases and only one of them appears in the report, the meaning can be lost. The following text fragment illustrates this problem. "Bill Dickson started to work in Procter & Gamble in 1994. In 1996 he became its vice-president". In this fragment it is possible to specify two words that are potentially unclear, these are the words "he" and "its", that does not have any meaning without the previous phrase that helps understand that "he" is Dickson, and "its" refers to Procter & Gamble company. If the first phrase is lost in the report, the text will lose its informive function. There are many researches attempting to solve this problem mainly by means of different sorts of "shreds". In a number of approaches a special window is created for the previous sentence of a report, by means of that it is possible to define the presence of semantic break or an unclear word. In other cases the sentence containing unclear words, are eliminated from a report, or attempts to find hints about the object true by means of short linguistic analysis are carried out. In this approach the degree of compression diminishes, as further information is put into the report. Moreover, when a basic report is already formed, it is difficult to restore the initial percent of compression.

Unlike the model of linear coefficients, in the methods of selection of extracts for short information presentation preparation large calculable resources are needed for the systems of human languages processing (NLP – Natural Language Processing), grammars and dictionaries for a syntactic analysis and generation of natural language structures in particular. In addition, for realization of this method it is necessary to consult ontological reference books and concepts oriented to the subject domainn to make decision during an analysis and determine major information. As shown on fig. 3, the method of short self-control forming is based on two principal approaches [7].

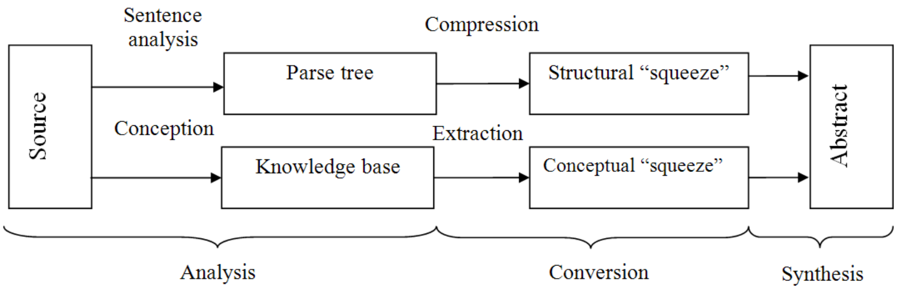


Fig. 3. Two basic approaches of presentation forming method

The first is based on the traditional linguistic method of syntactic analysis of sentences. In this method semantic information is used also for annotating of parse trees. Comparison procedures directly use trees to delete and regrouping of parts, for example, by branches reduction on the basis of some structural criteria, such as brackets or built-in conditional or subordinate clauses. After such a procedure parse trees are substantially simplified, becoming a structural "abstract" of the initial text per se.

The second approach to the stowage of short presentation is based on the approaches of artificial intelligence and based on the understanding of human language [3]. A syntactic analysis is also a component of such a method of analysis, but parse trees are not generated in this case. On the contrary, the conceptual representative structures of all initial information (accumulated in the text base of knowledge) are formed. Formulas of quantificational logic, semantic network, or set of frames can be used as structures.

A template of bank transactions (a forehand certain event) with involved organizations and people, relative date, transferred money volume, type of transaction etc mentioned in them, can be an example here. The transformations of the conceptual presentation presented on fig. 3 take a few changes. Excessive information that does not have a direct relation to the text is removed by deleting superficial judgments or chopping off of conceptual section columns. Then the information undergoes further aggregating by confluence of columns (or templates) or generalization of information, for example, by means of taxonomical hierarchies of relations of subclasses. To implement these transformations the following methods on the basis of conclusions are offered such as macro rules manipulating with logical suppositions, or operators distinguishing basic templates in the text base of knowledge [8].

As a result of transformation the conceptual representative structure of report is formed, that are conceptual "abstract" of the text per se. The presence of these formal representative layers (structural and conceptual "abstracts") distinguishes an approach on the basis of knowledge from an approach that is not on the basis knowledge. As shown fig. 3, the stage of synthesis is identical for both approaches: a text generator converts structural or conceptual presentation into a natural language report. Some systems give a user the possibility to manage received "abstracts" and do not have a stage of generation, on condition that initial texts are given together with their short presentation. This type of reviewing is based on certain structures of knowledge, that specify to the reviewing system beforehand what conception to choose more characteristic, or what conceptual characteristics (roles or fields) belong to this or that conception. This reviewing method fully represents semantic information as connections between knots in a conceptual column, as taxonomical (subclass or copy) or metonymy (part) relations. In this case, it also sets direction and criteria for choice for procedure of search or forming of conclusions. The rules of conclusion on the basis of reports or general charts of conclusion (such as a terminological classification) use this information to represent the text content in the most precise way.

The methods of abstracts creation are easily built to process large amounts of information. As their activity is limited to the choice of fragments, sentences or phrases, as a result a report text is incoherent. On the other hand, the method of forming of short abstracts produces more difficult annotations, that quite often contain information that complements initial text. As they are based on formal presentation of the informative content of a document, they can be built with a very high degree of compression, for example, those that are necessary for distribution of reports on the device of PDA (Personal Digital Assistant). The methods of templates filling are good only for the texts created after certain templates, although facilities of reviewing can use statistical technologies on the stage of analysis.

The methods that are based on knowledge need the full-scale sources of knowledge as a rule. This requirement is an obstacle for their wide distribution. The last tendencies in the area of systems of NLP on the basis of sets of texts promise perfection of syntax analyzers, that will embrace the wide range of knowledge, creation of fundamental dictionaries (such as WorldNet) and ontological reference books (such as Penman Upper Model) in the future. In addition, for the educational systems of NLP the large volume of texts is developed, in particular a set of text files, such as The Wall Street Journal, or grammatically annotated sets, such as Penn Treebank to the consortium of Linguistic Data.

Aim of work

The aim of this work is an analysis of existing models and methods that are used for creation of an automatic report; a selection of the most perspective among the considered models and methods; study of

TFIDF, TLTF and Text Relationship Map (TRM) methods for automatic construction of a report for texts by the Ukrainian and Russian languages; programmatic realization of an offered method.

A extraction of keywords for an annotation (report) construction with TFIDF model

This approach is based on the ideas offered in the works by Loon as early as in 1950s. they are based on the principles of statistical linguistics such as the Zipf law, that describes frequency of words distribution in a document.

Let’s take words that meet in a text and sort out them according to the frequency of their appearence. A word position on this list is called the grade of word. According to the Zipf law, a product of grade of word on his frequency is a permanent value. He received this result while studying English texts, however, it was further confirmed for other languages [9]. Frequency of a word appearance in a TF document is the most widespread method of words weighing in a document. The frequency is calculated as s correlation of number of appearance of a word in the general amount of words of document. This estimation is very popular and it is the basis for such a widespread method of calculation of estimation of measure of relevance as TFIDF.

To reduce meaningfulness of words, that meet in many sentences, an inverse frequency of a term IDF (inverse document frequency) is entered, it is a logarithm of relation of the number of all sentences $|S|$ to the sentences that contain a certain word t . The more frequent a word is in the documents of database, the less is the value of this parameter. Thus, for the words occur in a large number of documents, IDF is be near to zero (if a word occurs in all the documents, IDF equals zero) it helps distinguish important words (formula 4).

$$IDF = \log \frac{|S|}{s_i \ni t} . \tag{4}$$

A parameter of TF (term frequency) is a relation of frequency of a word appearance in a document to length of a document (formula 5). Normalization long document is done to make the TF parameter independ of the length of a document [10].

$$TF = \frac{|D|}{n_t} . \tag{5}$$

The coefficient of TFIDF equals the product of TF and IDF. Then it is possible to accept the coefficients of TFIDF of words the gravimetric parameters of vector model of a document which it contains. For scales to be in an interval (0, 1), and the vectors of documents to have the same length, the values of TFIDF are usually normalized according to a cosine. This formula estimates meaningfulness of a term only on the basis of frequency rate in a document, not taking into account appearance of terms in a document and their syntactic role; in other words, semantics of a document is interpreted as lexical semantics of terms that it contains, and composition semantics is not discussed.

The words with most weight are the most important here. Words with small weight can be even ignored.

Let’s show it in a simple example. Say, a document consists of three sentences.

Mother washed Masha with soap.

Mother washed a frame with soap.

Mother bought some soap in a shop.

The type of a dictionary is shown table 1.

To calculate the key word in the sentence, let’s identify every sentence with a self-weighted vector $s_i = (w_{i1}, w_{i2}, ..., w_{in})$ of the words appearing in the document, where n is an amount of words in the document d.

Table 1.

Keywords and their weight

Word	Weight	Frequency in a sentence	IDF
Mother	3	3	0
wash	3	2	0,18
soap	2	2	0,47
Masha	1	1	0,47
frame	1	1	0,47
shop	1	1	0,47
buy	1	1	0,47

Weight w_{ij} of a word j depends on frequency of its appearance in a certain sentence i and in the whole set of sentences (in a document), it is determined according to formula 6.

$$w_{ij} = f_{ij} \log_2 (\frac{m}{m_j}) ,\ i = 1,...,m ,\ j = 1,...,n , \tag{6}$$

where m_j is an quantity of sentences with a word j.

The function f_{ij} of frequency of appearance of a word j in a sentence i, is calculated as follows:

$$f_{ij} = \frac{n_{ij}}{n \cdot len(s_i)} , \tag{7}$$

where n_{ij} is an amount of appearance of a word j in a sentence i.

In order to avoid a change caused by length (by the amount of words) of a sentence, a function f_{ij} is normalized in relation to length of a sentence, $len(s_i)$ is length of a sentence.

To determine closeness d_{ip} between sentences s_i and s_p Evklid distance (formula 8) is mostly used.

$$d_{ip} = \sqrt{\sum_{r=1}^{n} (w_{ir} - w_{kr})^2} . \tag{8}$$

We calculated the frequency of appearance of every word in the document. However, it is obvious that meaningfulness of frequently appearing words must go down, as these are usually function words (prepositions) etc, to take this feature into account the algorithm of calculation of weight is modified as follows.

The list of the so-called "noise" or "stop" words is entered. This list, as a rule, is formed statically for this collection or language. Then the words are brought to the normal form. Some researchers (Baker, Mccallum) mark the decline of efficiency in the use of morphological processing, although a lot of researchers used it as it help reduce the space dimension.

Another method of reduction of a dictionary is a possibility to take synonyms into account so that the synonyms are marked with the same term of a dictionary.

The place of appearance of a word is taken into account in a text. This description is explained by intuitional reasoning, i.e. from the author’s point of view the major words take place in the title of a document or its parts, or at the beginning of the text.

Certainly, at such an approach casual terms rare words, proper names and other "noise" can occur among key words. It is therefore necessary to process texts with the use of algorithms, that improve quality selection.

A method of TFIDF is the most popular. At relative simplicity this description provides quite good quality of search. A disadvantage in this case is, opposite, long documents are underestimated, as there are more words and average frequency of words is lower in the text in them. To correct this effect, complemented normalized frequency is used, that calculated in the following way $0.5 + 0.5 \cdot (TF / ATF)$, where ATF is an average frequency of appearance of a term in a document [11].

The Zipf laws describe any text on the basis of frequency analysis of including of words to a text. However, it is not enough for the estimation of a document in a collection. The model of TFIDF allows to pass to the mathematical, vector model of text, distinguish the list of keywords.

The advantages of the method are high productivity, flexibility in relation to data. However, this method has a substantial disadvantage, at the construction of vector the order of words and the context are not taken into account, id est the semantic constituent of text is important. But by means of additional algorithms, such as a list of "stop" words, the calculation of Evklid distance, exposure of synonyms can increase the productivity and efficiency of the method.

Determining the weight of words by the method of TLTF

The idea TLTF method is based on the face that the most frequent words are short. Such words do not describe the basic theme of document, id est they are "stop words". And vice versa, words that the least frequent words are long. The advantage of the use TLTF method for weighing of words, is that this method requires no external resources, and uses only information within the limits of a document [12].

In a model for the calculation of clusters of words it is not frequency of appearance of terms in text (as in many methods), that is used but more difficult rules. Let's imagine the sequence of words in a sentence as: $\beta = \{w_u, \dots, w_v\}$, words join a cluster, if the following conditions are provided:

- the first w_u and the last w_v are meaningful words in a sentence;
- meaningful words are divided by of number of insignificant words determined beforehand.

For example, we can divide the sequence of words in a sentence as follows:

$$w_1 \cdot [w_2 \cdot w_3 \cdot w_4] \cdot w_5 \cdot w_6 \cdot w_7 \cdot w_8 \cdot [w_9 \cdot w_{10} \cdot w_{11} \cdot w_{12}]. \tag{9}$$

In this case a sentence consists of 12 words. Meaningful words are the words $w_2, w_4, w_9, w_{11}, w_{12}$. The clusters are in square brackets. They are formed in accordance with the following condition: meaningful words must not be divided by not more than three insignificant words. It is necessary to mention that in a sentence there can be a few clusters, as in our example. Meaningfulness of sentence determines the most value of cluster. The value of cluster in a sentence s_i will be calculated according to formula 10.

$$L_{s_i} = \arg\max_{\beta} \frac{ns(\beta, s_i)^2}{n(\beta, s_i)}, \tag{10}$$

where $ns(\beta, s_i)$ is an amount of meaningful words in a cluster; $n(\beta, s_i)$ is a common amount of words in a cluster.

Thus, using this model, according to a number of clusters in a sentence, extraction of major and most meaningful sentences for a report is done.

Method of documents reviewing based on the use of maps of test relations (TRM)

The method of writing a report based on the use of a map of text relations (TRM – Text Relationship Map), is used in this work. The idea of method consists in the presentation of a text as a graph [13]:

$$G = (P, E), \tag{11}$$

where $P = \{p_1, p_2, \dots, p_k, \dots, p_n\}$ are weighted vectors of words that correspond the fragments of a document.

A vector includes the scales of words that belong to it. For example, k fragment will be presented by a vector:

$$\{\omega_{k,1}, \omega_{k,2}, ..., \omega_{k,i}, ... \omega_{k,m}\},\tag{12}$$

where $\omega_{k,1}$ is weight of a word in the sentence i of fragment k; E is a great number of arcs between the knots of the graph $E = \{(p_k, p_b,), p_k, p_b \in V\}:$.

On fig.4 there is an example of such a map. Every knot on the map corresponds to a fragment of the text (sentence) and is a weighted vector of terms.

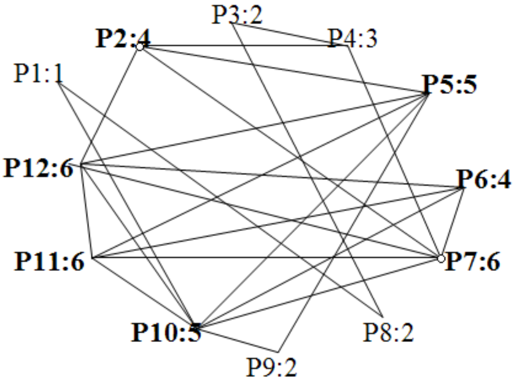


Fig. 4. Example of a map of text relations

Connections are created between two knots, if they have a high measure of similarity between the fragments of a text, that is usually calculated as a scalar product of vectors that present these fragments. If there is a connection between two knots, then it is said that corresponding fragments are "semantically similar". The number of arcs that is included in a certain knot corresponds to the importance of a fragment.

$$sim(p_i, p_j) = \frac{\sum_{k=1}^{m} p_{i,k} \cdot p_{j,k}}{\sqrt{\sum_{k=1}^{|m|} p^2{}_{i,k}} \cdot \sqrt{\sum_{k=1}^{m} p^2{}_{j,k}}}.\tag{13}$$

For example, on fig. 4 amounts of entrance arcs of knot equal 7, as its arcs are from knots. It is the maximal value. Thus, a knot can cover fragments corresponding the knots related to it, the maintenance, and it must included in a report.

The basic disadvantage of this approach is that only one aspect of importance of a fragment is taken into account, it is its relation to the other fragments of a document. Informative function of words contained in a separate fragment is not examined here. As a result, closely constrained with other fragments can get to the report, but they do not characterize the subjects of the document (id est do not contain keywords). For correct this defect it is suggested to use the concept of local and global properties of a fragment. Thus, local properties are examined as clusters of words in a sentence, which weight is calculated by TLTF method. And the relationship of this sentence with all others in a text comes forward, as global property, that is determined by the method of TRM. Combining both properties, this method determines the degree of meaningfulness of a sentence and necessity of its use in a report.

Described local and global properties determine different aspects of meaningfulness of a sentence. A local property determines a part of information in a sentence, and a global one pays attention to the structural aspect of a document estimating informative function of the whole sentence. To increase other efficiency it is suggested to examine both objects in totality, uniting them in the estimation of informative function of a sentence, that can be used for decision-making, whether to use this sentence in a report or not. To calculate the combined estimation the following formula is used:

$$F(s_i) = \lambda G' + (1-\lambda)L',\tag{14}$$

where G' – a normalized global unity of a sentence is calculated according to the formula:

$$G' = \frac{d_{s_i}}{d_{\max}}, \tag{15}$$

where $d_{\max}$ is the maximal amount of ribs for one knot on the map of relations in a text; d_{s_i} is an amount of ribs for the knot of the corresponding sentence s_i; L' is a normalized value of local clasters of a sentence s_i, it is calculated according to the following formula:

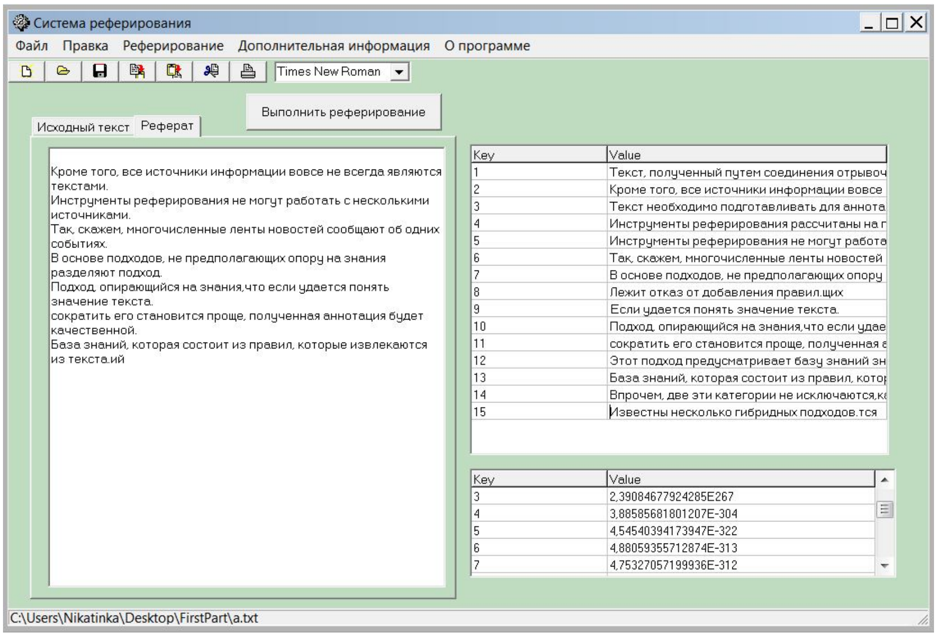
$$L' = \frac{L_{s_i}}{L_{\max}}, \tag{16}$$

where $L_{\max}$ is a maximal local clusterization in the whole text; λ is a parameter that changes according to the importance of the constituents G' or G'.

Thus, taking into account all the described methods, the integrated estimation for all sentences is created, according to which results it is possible to choose a sentence for a report or annotation.

Programmatic realization of TRM and LSA (Latent Semantic Analysis) for full-text document reviewing

The developed system of reviewing allows to work with a text, create a report, choose major sentences in a text, to analyze frequency of appearance of words. On fig. 5 a program interface is presented.



Кроме того, все источники информации вовсе не всегда являются текстами.
Инструменты реферирования не могут работать с несколькими источниками.
Так, скажем, многочисленные ленты новостей сообщают об одних событиях.
В основе подходов, не предполагающих опору на знания разделяют подход.
Подход, опирающийся на знания,что если удается понять значение текста.
сократить его становится проще, полученная аннотация будет качественной.
База знаний, которая состоит из правил, которые извлекаются из текста.ий

Key	Value
1	Текст, полученный путем соединения отрывоч
2	Кроме того, все источники информации вовсе
3	Текст необходимо подготавливать для аннота
4	Инструменты реферирования рассчитаны на г
5	Инструменты реферирования не могут работа
6	Так, скажем, многочисленные ленты новостей
7	В основе подходов, не предполагающих опору
8	Лежит отказ от добавления правил.щих
9	Если удается понять значение текста.
10	Подход, опирающийся на знания,что если удае
11	сократить его становится проще, полученная а
12	Этот подход предусматривает базу знаний зн
13	База знаний, которая состоит из правил, котор
14	Впрочем, две эти категории не исключаются,ка
15	Известны несколько гибридных подходов.тся

Key	Value
3	2,39084677924285E267
4	3,88585681801207E-304
5	4,54540394173947E-322
6	4,88059355712874E-313
7	4,75327057199936E-312

C:\Users\Nikatinka\Desktop\FirstPart\a.txt

Fig. 5 Result of the program implementation

The main menu is represented by a few points: File, Correction, Reviewing, Additional information. The program is intended to work with text files with the expansion .txt. In order to begin work with the program it is necessary to open a text by means of File of the main menu, that it is necessary to create a report for. Analogical actions can be taken by means of the buttons on the bar of tools.

The result of implementation is presented on figure 5.

At the file opening the text of the document is represented on an inset Original text. For further document processing it is necessary to push the button Execute reviewing or to choose Reviewing on the main menu. After the button is pressed a report is created and it is represented on an inset Report.

Thus, a program user will get a short document content, the most important sentences in the text, that help understand the text main ideas. After implementation of the program a user has the opportunity to read information about a text, to get the numerical value of degree of sentences similarity and understand what sentences the program put into the report and why.

Conclusions and received results analysis

The task of the automatic abstracting is a task of extraction of a text content. There are many software instruments for creation a document report. However, they do not always give necessary for the user result, that is why automatic abstracting remains one of priority tasks of artificial intelligence.

The review of literature, analysis of methods and approaches to the problem of the automatic abstracting, shows that to solve this task it is important to select keywords, word-combinations, informatively-saturated sentences of a text, artificially built sentences characterizing the basic text content.

Existing methods and models for creation of an automatic report were analyzed. A review of the existing industrial systems carrying out the automatic abstracting functions is done. Thus, undertaken studies allowed to develop a mathematical model of an automatic report creation for Russian and Ukrainian full-text documents, that is based on the use of latently-semantic analysis methods (LSA – Latent Semantic Analysis), a map of text relations (TRM – Text Relationship Map) and TFIDF metrics (Term Frequency Inverse Document Frequency) to extract keywords from a text.

On the basis of this model an built algorithm was developed, it was realized by means of C++ language programming system in the environment of Borland Builder 6.0 and a database created in Microsoft Access.

1. Михайлов А. И. Основы информатики / А. И. Михайлов, А. И. Черный, Р. С. Гиляревский. – М.: Наука, 1968. 2. Леонов Б. П. О методах автоматического реферирования (США 1958-1974 гг.) / Б. П. Леонов // Научно-техническая информация, сер.2. – 1975. – №6. – С. 16-20. 3. Пащенко Н. А. Проблемы автоматизации индексирования и реферирования / Н. А. Пащенко, Л. В. Кнорина, Т. В. Молчанова и др. // Итоги науки и техники. Сер. Информатика. – М.: ВИНИТИ, 1983. – Т.7. – С. 7-164. 4. Севбо И. П. Структура связного текста и автоматизация реферирования/ И П. Севбо. – М.: Наука, 1969. – 135 с. 5. Белоногов Г. Г. Компьютерная лингвистика и перспективные информационные технологи / Г. Г. Белоногов, Ю. П. Калинин, А. А. Хорошилов. – М: Русский мир, 2004. – 248 с. 6. Башмаков А. И. Интеллектуальные информационные технологии: учеб. пособие / А. И. Башмаков, И. А. Башмаков. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2005. – 304 с. 7. Borko H. Abstracting Concepts and Methods / H. Borko, C. L. Bernier. – Academic Press, New York, 1975. 8. Iatsko V. Linguistic Aspects of Summarization / V. Iatsko // Philologie in Netz. – 2001. – № 18. – pp. 33-46. 9. Скороходько Э. Ф. Семантические сети и автоматическая обработка текста / Э. Ф. Скороходько. – Киев: Наукова думка, 1983. – 219 с. 10. Чугреев В. JI. Модель структурного представления текстовой информации и метод ее тематического анализа на основе частотно-контекстной классификации: Автореф. канд. техн. Наук / В. Л. Чугреев. – С-Пб., 2003. – 24 с. 11. Hahn U., Mani I. The Challenges of Automatic Summarization / U. Hahn, I. Mani // IEEE Computer Cociety. – 2000. – vol. 33, no. 11. – pp. 29-36. 12. Барсегян А. А. Технология анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А. А. Барсегян, М. С. Куприянов, В В. Степаненко, И. И. Холод. – СПб.: БХВ-Петербург, 2007. – 384 с.