Journal of Theoretical and Applied Information Technology
29th February 2020. Vol.98. No 04
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# HYBRID APPROACH TO AUTOMATIC SUMMARIZATION OF SCIENTIFIC AND TECHNICAL TEXTS

**AIGERIM M. BAKIYEVA[1], TATIANA V. BATURA[2]**

[1]Department of Information Technologies, Novosibirsk State University, Novosibirsk, Russian Federation

[2]Department of Information Technologies, Novosibirsk State University, Novosibirsk, Russian Federation

[2]Laboratory of Complex Systems Simulation, A. P. Ershov Institute of Informatics Systems (IIS) SB RAS, Novosibirsk, Russian Federation

E-mail: [1]m_aigerim0707@mail.ru, [2]tatiana.v.batura@gmail.com

**ABSTRACT**

The paper is devoted to the methods of automatic summarization, which use the representation of a text in the form of a graph. And contains an attempt at formal description of the text transformation in terms of the predicate calculus logic. The proposed method combines the use of a linguistic knowledge base, graph representation of texts and machine learning. The fragments of a text, such as words, sentences, paragraphs, are represented as graph nodes, and relations between nodes, for example, rhetorical relations, are denoted by edges. Automatic determination of rhetorical relations in the text allows you to set the location of the nucleus and satellite. To compile a brief annotation, it is necessary to transform the original text, based on the assumption that the nucleus contains the most important part of the statement. The relations between discursive markers in the text define a hierarchy that allows one to solve various problems of word processing in a natural language, including the task of automatically compiling a short abstract on a large volume of text. The summarization process created by the authors consists of six main steps: preprocessing, topic modeling, rhetorical analysis and transformation, weight evaluation, sentence selection, and smoothing. Topic modeling is used to discover key terms. First, unigram topic models, that contain only one-word terms, are constructed. These models are further expanded by adding multiword terms. The most significant fragments of the source document are determined in the process of rhetorical analysis using discursive markers.

Presentation of texts in the form of graphs helps to demonstrate the transformations with the text necessary to highlight important fragments. In assessing the importance of the text fragments are also included keywords, multiword and scientific terms, describing the scientific and technical texts. To store the marker information has created a linguistic knowledge base. The final step in the formation of the annotation is smoothing — a text conversion procedure that allows you to make the text of the abstract (annotation) received more coherent and consistent. The importance of sentences is determined using discursive markers and connectors. We used additive regularization for topic modeling (ARTM) to extract keywords and discover the topics. Our proposed BigARTM and Rake hybrid method for obtaining thematic models and the task of obtaining an abstract using RST markers, action and templates showed its effectiveness and efficiency in testing and in comparison with other methods as was shown in comparisons using the precision, recall and F- measure calculated in a way similar to [2, 10].

**Keywords:** *Automatic Text Processing, Theory of Rhetorical Structures, Discursive Marker, Text Analysis, Rhetorical Relationships, Semantics.*

## 1. INTRODUCTION

### 1.1. Relevance of the Study

Today there is a tremendous increase in the amount of information created by people and machines in a natural language. Due to a rapid increase in the bulk of textual information in the Internet, active research in the field of computer linguistics remains to be highly demanded. Meanwhile, abstracting articles on information technology is especially important, since information technology is used in almost all branches of science and technology.

The development of algorithms and the creation of systems for automatic referencing, retrieval and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

extraction of information, classification and clustering of text documents are still considered complicated issues. A continuous increase in the intensity of the flow of textual information makes the task of semantic compression of textual information more and more important. The connections between rhetorical markers, connectors and keywords in the text define a semantic hierarchy that allows you to solve various tasks of word processing in a natural language and is an important element in the auto-abstracting and determination of text topics.

### 1.2. Novelty of the Paper

This article proposes a hybrid method for automatic constructing of scientific texts' annotations in the field of information technology, which is still beyond the attention of researchers who develop abstracting systems.

## 2. RELATED WORKS

The attempts of applying discursive analysis to solving various tasks of computer linguistics can be found in current practice. A detailed review of the literature presented in the paper [1] reveals that, in most cases, discursive analysis can enhance the quality of automatic systems by 4-44%, depending on the specific task. The Rhetorical Structure Theory (henceforth: RST) approach has been widely used in many industries.

### 2.1. RST + Constraint Satisfaction Algorithm

In [2], RST was applied to identify important units in a document. The author proposed to use a constraint satisfaction algorithm to assemble all the trees that organize the input text, and then employed several heuristics to prefer one tree to the others.

### 2.2. Practical Procedure for Extracting the Rhetorical Structure of Discourse

In [4], a computational model of discourse for Japanese informative texts is proposed. This paper proposes a practical procedure for extracting the rhetorical structure of discourse. The rhetorical structure is presented in the form of a tree. The abstract compilation process is preceded by the extraction of rhetorical structures from the text of the article and their analysis. Evaluation of the results showed that the abstracts contain up to 74% of the important sentences of the original article.

### 2.3. Automated Multilingual Text Summarization System SUMMARIST

An automated multilingual text summarization system called SUMMARIST is described in [3]. This system combines symbolic concept-level world knowledge with information retrieval and statistical techniques. The algorithm consists of three steps: topic identification, interpretation, and generation. SUMMARIST produces extracted summaries in five languages: English, Japanese, Spanish, Indonesian, and Arabic.

### 2.4. Summarization Problem as a Reduction of Data

Some authors [5] consider the summarization problem as a reduction of data, namely, the original document is considered as a high-dimensional data, and the summarization task is to reduce the dimension of the document and keep the main content of it.

### 2.5. An RST-based Summarization System with 7 Rhetorical Categories

An RST-based summarization system for scientific articles, identifying seven rhetorical categories, is described in [6].

### 2.6. Query-based Summarization

Structural analysis formed the basis of sentence weights in [7]; the authors applied RST to create a graph representation of a document from which a query-based summarization was produced.

### 2.7. Including Information from Neighboring Documents

Building a generic and extractive summary for a single document that includes information from its neighboring documents is discussed in [8]. The produced summary has sentences extracted from the single document and makes use of the additional knowledge from its multi documents.

### 2.8. Schema-based Summarization Approach

Mithun S. describes a schema-based summarization approach for query-based blog summaries that utilizes discourse structures [9]. This approach performs four main tasks, namely: question categorization, identification of rhetorical predicates, schema selection, and summary

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

generation. The author built a system named BlogSum and evaluated its performance for question relevance and coherence. The achieved results show that the proposed approach can effectively reduce question irrelevance and discourse incoherence of automatic summaries.

### 2.9. Researches for Russian Languages Texts

Research in this field for the English language has reached a sufficiently high level, but there is not enough research for Russian. The summarization problem in scientific and technical texts in Russian using similar approaches was stated by scientists in papers [10] and [11].

#### 2.9.1. Combining nonlinear + hierarchical nature of the text

The research [10] describes the methods and algorithms that take into account the nonlinear and hierarchical nature of the text. With the help of rhetorical relations, the problem of extraction is solved. The author has developed a system based on inference rules and a highly specialized dictionary of key phrases.

#### 2.9.2. Extraction + abstraction combination approach

A hybrid approach is proposed in [11], combining extraction and abstraction methods. This approach was implemented by the author in a summarization system focused on automatic translation. The described system is constructed for texts on mathematical modeling. In our work, we solve a broader problem; we analyze any scientific texts (articles, theses, reports) on any topics. In addition, we solve also the topic detection problem and the problem of searching for keywords and multiword expressions.

#### 2.9.3. RST + discursive markers

The work [12] describes the experience of building a corpus in Russian containing discursive markers. The corpus includes the texts of different genres, such as scientific, popular-science, and news, therefore, it is publicly available. Before using the theory of rhetorical structures, it should be adapted for a specific language. This is due to grammatical features. In their paper, the authors exemplified the hierarchy of rhetorical relations, which proved to be more convenient and correct to be taken into account when summarizing texts.

## 3. PROPOSED METHODOLOGY

Our work describes an approach that allows forming brief abstracts of scientific and technical texts and determining their topics. The proposed method forms an abstract based on the most essential sentences of the original document. The importance of a sentence is determined in the process of rhetorical analysis.

In its turn, the method of additive regularization of thematic models (ARTM) is applied to determine the themes of the texts. Additive regularization of thematic models [13] allows one to solve the problems of non-uniqueness and instability by introducing additional restrictions on the required solution. The following methods of smoothing and sparsing out the distribution of terms in topics, topics in documents, and others can be utilized as regularizers.

### 3.1. Semantic Analysis and Formal Features of Rhetorical Relations

Rhetorical Structure Theory [14] postulates that a coherent text can be characterized by means of a tree structure whose leaves are the “elementary discourse units” (henceforth: EDUs). Using the RST relates to coherent method [2] of extracting text. In RST, rhetorical relations are considered as semantic relations. This theory is based on the assumption that any unit of discourse is connected with another unit of this discourse through some meaningful connection. Thus, the basic concepts of RST discourse are unit and relation.

Two types of EDUs are defined in RST: nucleus and satellite. ***The nucleus*** is considered as the most important part of the statement, while the satellites explain the nuclei and are secondary. The nucleus contains basic information, and the satellite contains additional information about the nucleus. ***A satellite*** is often incomprehensible without a nucleus. While the expressions where the satellites were deleted can be understood to a certain extent.

Successive EDUs are interconnected by rhetorical relations. These parts are the elements from which larger fragments of texts and entire texts are built. Each fragment, in relation to other fragments, performs its certain role. Textual connectivity is formed by means of those relationships that are modeled between fragments within the text [15]. Expressions where the satellites were deleted can be understood to a certain extent. Consider the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

following example:

**Text**: There may be some regularity in the target function, besides being smooth.
**Marker**: besides
**Relationship name**: Elaboration

The following notation is provided for convenience in the example below.

Suppose, x is a nucleus; y is a marker; z is a satellite;
S (x) is a predicate for EDU, which is a nucleus;
S' (x) is a predicate for EDU (which is a nucleus) beginning with a capital letter;
S (z) is a predicate for EDU, which is a satellite;
S' (z) is a predicate for EDU (which is a satellite) beginning with a capital letter;
y' is a marker beginning with a capital letter;
p ( ) is a punctuation character, the argument can be ".", ",", ":", ";".

Now the provided example can be represented as a formula of the predicate calculus:

$$S'(x) \wedge p(,) \wedge y \wedge S(z) \wedge p(.) \quad (1)$$

Automatic determination of rhetorical relations in the text allows you to set the location of the nucleus and satellite. For the formation of the abstract, it is necessary to perform certain actions with the text, depending on different markers and discursive relations. Since the core contains the most important part of the statement, the proposed method can be used in systems of summarization and extracting information from texts. To assess the precision of the proposed method, there was used expert assessment. Based on the data obtained, it was decided to consider the most common relationships. Precision was evaluated for each collection using the formula:

$$Precision=TP/(TP+FP), \quad (2)$$

where *TP* is a truly positive decision; *FP* is a false positive decision.

Table 1 provides an assessment of the accuracy of determining the most common rhetorical relationships.

*Table 1: Evaluation of the Precision of Determining Rhetorical Relations*

| | Names of relations | Precision |
|---|---|---|
| 1 | Condition | 0.89 |
| 2 | Cause-Effect | 0.99 |
| 3 | Example | 0.98 |
| 4 | Restatement | 0.97 |
| 5 | Contrast | 1.00 |
| 6 | Purpose | 0.99 |

Therefore, we can conclude that for many languages, scientific and technical texts are more characterized by the following relationships: Condition, Cause-Effect, Example, Restatement, Contrast and Purpose.

To assess the precision of the proposed method, an expert assessment was used. Based on the data obtained, it was decided to consider the most common relationships.

### 3.2. Formal Description of Text Transformation

**Step 1.** The nucleus EDUs are essential to be found in the text in order to obtain a short abstract automatically.

**Step 2.** Transformation of the statements containing these nucleus EDUs, so that the text of the resulting abstract turns out to be connected. Depending on different markers and discursive relationships, these transformations will be different.

**Step 3.** Some of the considered transformations are provided. For a formal description of the actions performed by the system, it was decided to use the predicate logic of the first and second orders.

#### 3.2.1. First-order predicates

According to the notations introduced in the previous section, the actions performed by the system can be described as follows.

In the example from the previous section with the marker $y$ = "*besides*", it is necessary to delete the satellite along with the marker and leave the previous clause, which is nucleus EDU that can be illustrated as:

$$S'(x) \wedge p(,) \wedge y \wedge S(z) \wedge p(.) \rightarrow S'(x) \wedge p(.) \wedge \neg(y \wedge S(z) \wedge p(.)) \quad (3)$$

For the predicates shown above, we have introduced special actions that are performed to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

create a quasi-abstract. They depend on some verbs, nouns, markers and connectors.

**Markers (Discourse markers)** are words or phrases that do not have real lexical meaning, but instead they have an important function to form the structure of summarizing a text they use to connect, organize and manage the authors' intentions.

**Connectors** are groups of words that replace markers and characterize certain rhetorical relations. The connectors provide inter-phrase connection; they show the semantic incompleteness of the sentence. Table 1 shows the actions for markers and connectors.

*Table 2: Actions for Markers and Connectors*

| № | Rhetorical relations | Markers and Connectors | Actions |
|---|---|---|---|
| 1. | Cause-Effect | therefore | DELETE_SAVE |
| 2. | Contrast | however | SAVE_DELETE |
| 3. | Elaboration | moreover | SAVE_DELETE |
| 4. | Elaboration | for example | SAVE_DELETE |
| 5. | Elaboration | in this regard | SAVE_DELETE |
| 6. | Elaboration | at the same time | DELETE_SAVE |
| 7. | Elaboration | thereby | SAVE_SAVE |
| 8. | Evidence | in this way | DELETE_SAVE |
| 9. | Restatement | in other words | SAVE_DELETE |

During the research, we created a dictionary consisting of 121 markers and connectors, 120 nouns and 108 verbs with weights that are often found in scientific and technical texts. In total, eight actions were considered. Some actions are explained below.

DELETE_SAVE: This action removes the forthcoming clause and saves the clause with the given marker.

SAVE_DELETE: This action saves the upcoming clause and removes the clause with the given marker.

SAVE_SAVE: This action completely saves the clause with the given marker and the previous clause.

### 3.2.2. Second-order predicates

The cases of nested EDU, when lower-level EDUs are embedded in higher-level EDUs, are more convenient to describe using second-order predicates. Moreover, a separate predicate is introduced for each marker. To illustrate how the text is transformed in the cases of nested EDUs, the following example is provided.

**Example 1:**
*"Most software implementations need to support operations that can return more than one tensor.* ***For example****, if we wish to compute both the maximum value in a tensor and the index of that value, it is best to compute both in a single pass through memory,* ***so*** *it is most efficient to implement this procedure as a single operation with two outputs".*
In order to write down our example in a formal form, we add the following algorithm.

Suppose $m$ is a nucleus in a dependent clause; $n$ is a satellite in a dependent clause;
$S(m)$ is a predicate for EDU, which is a nucleus in a dependent clause;
$S'(m)$ is a predicate for EDU (which is a nucleus in a dependent clause) beginning with a capital letter;
$S(n)$ is a predicate for EDU, which is a satellite in a dependent clause;
$S'(n)$ is a predicate for EDU (which is a satellite in a dependent clause) beginning with a capital letter;
$y_i$ are markers or connectors.

$$S'(x) \wedge p(.) \wedge S'(y_1 \wedge S(n) \wedge p(,) \wedge y_2 \wedge S(m)) \wedge p(.) \rightarrow S'(x) \wedge p(.) \wedge \neg S'(y_1 \wedge S(n) \wedge p(,) \wedge y_2) \wedge S'(S(m)) \wedge p(.) \quad (4)$$

where $y_1$ = *"for example"*; $y_2$ = *"so"*.

As a result, we will get the following text: *"Most software implementations need to support operations that can return more than one tensor. It is most efficient to implement this procedure as a single operation with two outputs".*

**Example 2:**
In a complex sentence, the main and subordinate clause are highlighted. In this case, the lower-level EDUs are embedded in the higher-level EDUs. It is more convenient to use second-order predicates to describe actions with embedded EDEs. To illustrate how text is converted in the case of nested EDUs, we give the following example.

"***In addition****, the air that enters the freezer is already cooled to 1 ° C using a docking station refrigeration unit, which accounts for about 50% of the heat load of the incoming air.* ***Therefore****, the*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

*net effect of a refrigerated delivery dock is to reduce load infiltration. Net profit is equal to the difference between reduced infiltration of the load of the freezer and the cooling load of the shipping dock.* ***Note****, that dock refrigerators operate at significantly higher temperatures, and consume significantly less energy for the same amount of cooling.*"

Now transformations with text can be described as follows:

$S'(z) \wedge p(.) \wedge S'(x) \wedge p(.) \wedge S'(x) \wedge p(.) \wedge S'(z) \wedge p(.) \rightarrow \neg S'(El\ p(,) \wedge S(m) \wedge p(.)) \wedge \neg S'(Ev \wedge p(,)) \wedge S'(S(m) \wedge p(.)) \wedge S'(x)\ p(.) \wedge \neg S'(Cont \wedge S(n) \wedge p(.)),$

where $El$= «***In addition***»; $Ev$= «***Therefore***»; $Cont$= «***Note***».

It should be noted that the use of formalisms of first and second order logic for this purpose has not yet been sufficiently studied. In the future, it may be necessary to supplement this formalism in order to take into account the sequence of elements in the text.

It should be noted that the use of first and second order formalisms for this purpose has not yet been sufficiently investigated. In the future, it may be necessary to extend it to take into account the order of the elements in the text and the order of transformations.

### 3.3. General Description of the System

Let $T$ be the text of the article cleared after preprocessing. It consists of sentences

$$T = \bigcup_{k=1}^{P} S_k \quad (5)$$

In our understanding, the task of text summarization is to find the transformation $\Psi$ of the text $T$ into a summary $\widetilde{T}$, such that $\Psi : T \rightarrow \widetilde{T}, |T| > |\widetilde{T}| \approx 250$ words.

We developed the system "Scientific Text Summarizer". Information given on the figure 1 shows a flowchart of it.

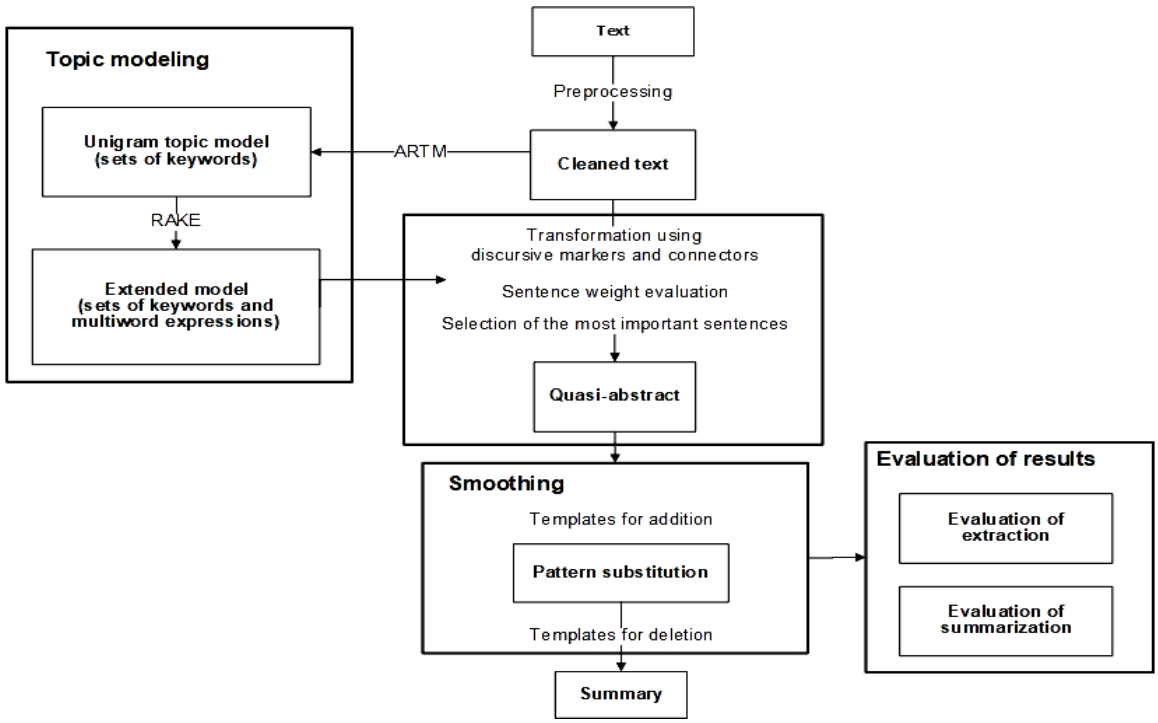


*Figure 1: Flowchart of the system "Scientific Text Summarizer"*

The main steps of our algorithm are described below.

#### 3.3.1. Step 1 — preprocessing

At the preprocessing stage, all images, tables, sentences with formulas, information about authors

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and bibliographic references were deleted from the source text. The author's abstracts were cut and saved separately so that we could evaluate the system afterwards, by means of comparing the result with the original abstract.

### 3.3.2. Step 2 — building topic models, extracting keywords and multiword expressions

Initially, a unigram model of the text is built; then the model expands with multiword expressions. To extract multiword expressions, we adapted the RAKE (Rapid Automatic Keyword Extraction) algorithm [21] for working with Russian texts.

Topic modeling consists in building a model of a collection of text documents. In such model, each topic is represented by a discrete probability distribution of words, and documents are represented by a discrete probability distribution of topics. Currently, there are different methods of topic modeling, such as PLSA (Probabilistic Latent Semantic Analysis,), LDA (Latent Dirichlet Allocation) и ARTM (Additive Regularization for Topic).

***PLSA*** is a probabilistic topic model for representing natural language text. The model is called latent, since it involves the introduction of a hidden (latent) parameter, which is the topic. First described in Thomas Hoffmann in 1999 [16].

***LDA*** is a model that allows explaining the results of observations using implicit groups, which makes it possible to identify the reasons for the similarity of some parts of the data. For example, if the observations are words collected in documents, it is argued that each document is a mixture of a small number of topics, and that the appearance of each word is associated with one of the topics of the document [17].

***ARTM*** — additive regularization of topic models, is a generalization of a large number of thematic modeling algorithms. ARTM allows you to combine regularizers, thereby combining thematic models. With this approach, the PLSA is a thematic model without regularizers, and the LDA is a thematic model in which each topic is smoothed by the same Dirichlet regularizer. The ARTM model was proposed in 2014 [18]. Currently, ARTM is becoming increasingly popular due to its versatility and flexibility in setting model parameters.

The main advantage of topic models in comparison with neural networks is that they are easy to interpret; the user understands the reasons for finding certain topics in the text and the structure of the topics themselves. In addition, it is often required that thematic models take into account heterogeneous data, identify the dynamics of topics in time, automatically separate topics into sub-topics, use not only single keywords, but also multiword terms, etc.

To select the algorithm of topic modeling, we performed a number of experiments, the results of which are presented in [19].

The choice of methods for topic modeling is due to the presence of certain features. For comparison, some of them are shown in table 3.

*Table 3: Comparison of topic modeling methods.*

| Method's name | Increasing the number of model parameters with an increase in the number of documents | Applicability to big data sets | Using multi-word terms | Uniqueness and stability of the solution |
|---|---|---|---|---|
| PLSA | yes, there is a linear dependance | no | no | no |
| LDA | no | yes | no | no |
| ARTM | no | yes | no | yes |
| ARTM + RAKE | no | yes | yes | yes |

It was decided to use the ***ARTM algorithm*** in the implementation of the Big ARTM library [20]. Due to its versatility and flexibility in parameter settings, ARTM allows you to combine regularizers, thereby combining thematic models. This method guarantees the uniqueness and stability of the solution. ARTM does not see an increase in the number of model parameters with an increase in the number of documents, so it can be applied to large sets of data. In addition, the modification we proposed allows us to use not only single-word, but also multiword expressions, which, in our opinion, increases the interpretability of the model.

### 3.3.3. Step 3 — rhetorical analysis and text transformation

At this step, we find sentences containing the discursive markers and connectors. To these sentences, certain actions are applied (see Section 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

for detailed information). As a result, we obtain a quasi-abstract. A quasi-abstract is a list of the most important sentences (or its fragments) in the text: $\Psi(T, D, V)=T'$

### 3.3.4. Step 4 — evaluation of sentence weights

The weight is assigned to each sentence of a quasi-abstract. Let us describe this procedure in more detail.

Let $S_k'$ be an arbitrary quasi-abstract sentence. Then:

$$T' = \bigcup_{k=1}^{P_1} S_k' \quad (6)$$

The weight of each sentence of the quasi-abstract is calculated depending on whether it contains keywords (or multiword terms), discourse markers and connectors, and some special lexicon that are often found in scientific and technical texts. To extract multiword expressions from the texts, the RAKE algorithm is used. It was developed for visualizing topics and finding significant n-grams in English texts [21]. In the course of our work, we adapted this algorithm for processing the texts in Russian.

As a result, the weight of each sentence is calculated by the following formula:

$$SW(s') = \frac{1}{L} \cdot \sum_{i=1}^{L} w_i + \frac{1}{M} \cdot \sum_{j=1}^{M} v_j + \frac{1}{N} \cdot \sum_{k=1}^{N} d_k \quad (7)$$

where $W = \{w_1, w_2, ..., w_L\}$ – is a set of weights of keywords and multiword expressions ($|W| = L$). The weight $w_i$ is defined as the frequency of the keyword (or the multiword expression) in the text;

$D = \{d_1, d_2, ..., d_N\}$ – is a set of weights of discursive markers and connectors ($|D| = N$). The weight $d_j$ is determined using a linguistic knowledge base.

$V = \{v_1, v_2, ..., v_M\}$ – is a set of weights of significant verbs and nouns that are often found in scientific and technical texts ($|V| = M$). The weight $v_k$ is determined using a linguistic knowledge base.

### 3.3.5. Step 5 — sentence selection

From the obtained set of sentences (see item 2), only those whose weight exceeds a predetermined threshold value (see item 3) are selected for the summary:

$$\widetilde{T} = \left[ s' \in T' : SW(s') > \beta \right]$$

where $\beta = 0.15$ is a constant defined empirically.

### 3.3.6. Step 6 — smoothing operation

The smoothing operation is a text conversion procedure that allows you to get coherent text from disparate fragments and, if necessary, further reduce it. *Smoothing* makes the resulting abstract more coherent and readable. While smoothing, some words are replaced with ones that are more suitable or deleted.

For example, let us consider the fragment *"**Indeed,** we can show how — in the case of a simple linear model with a quadratic error function and simple gradient descent—early stopping is equivalent to L2 regularization. **In order to** compare with classical L2 regularization, we examine a simple setting where the only parameters are linear weights"*. It will be replaced by *"In the case of a simple linear model with a quadratic error function and simple gradient descent — early stopping is equivalent to L2 regularization. We examine a simple setting where the only parameters are linear weights"*.

To smooth sentences, we used two types of templates: for removing fragments of sentences (in the case when the received summary is longer than 250 words) and for addition (in the case when a fragment of an unfinished sentence was included in the summary).

In cases where it is necessary to replace one fragment of a sentence with another, first substitution is applied to the template for deletion, then substitution to the template for addition. At the same time, it is important that certain conditions are met for the selection of suitable templates.

The following types of templates were used to complement:

**Introduction**;
**Novelty** (Application | Relevance | Efficiency | Feature | Perspective);
**Aim**;
**Method** (Technique | Planning | Methodology | Model | Strategy | Approach | Assessment | Definition | Formation | Analysis | Design);
**Implementation**;
**Disadvantages** (Errors | Advantages);

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

| Conclusion (Output \| Summary \| Results). |
|---|

***The template type "Introduction"*** has the form <*X*, *Yv*, *Yn*, *Z*>, where X - the added fragment. *X*∈ {"In the article", "In the work", ...}; *Yv* - the verb *V*∈ {"considered", "considering", ...}; *Yn* - part of the core, which includes a noun characteristic of scientific terms *N*∈ {"tasks", "method", "directions", "approaches", ...}; Z - is the remainder of the sentence (satellite).

***Templates of the "Target" type*** have several presentation options. For example, the most frequent of them <$X_p$, $X_w$, $Y_v$, *KW*, *Z*>, where $X_p$ - {Aim, Principal aim, Mainstream, ...}; $X_w$ - {of this work, articles, studies, models, ...}; $Y_v$ - {is, plays, takes, is considered, ...}; *KW* - keywords; *Z* - the rest of the offer (satellite with or without a marker).

***The template type "Novelty"*** has the form <*X*, $Y_n$, $Y_v$, *Z*>, where *X* is the fragment to be added. *X*∈ {"Novelty", "Novelty and Perspective", ...}; *Yn* - noun *N*∈ {"method", "algorithm", "approaches", ...}; *Yv* - part of the nucleus, which includes the verb *V*∈ {"concluded", "determined", ...}; Z is the remainder of the sentence (satellite).

## 4. RESULTS AND DISCUSSION

Our system was tested on a collection of 1200 scientific articles in the Russian language taken from the open-access journal archives "Software & Systems" for 2013–2018. There is still no generally accepted effective method for automatic evaluation of summarization systems [22].

1) Firstly, we tried to assess the quality of the received abstract with the ROUGE metric, based on counting the number of matching text elements, for example, n-grams, or sentences [23]. In this metric, the summary sentence is considered as a sequence of words. The main point is that the longer the *LCS* (the longest common subsequence) of the two summary sentences, the more similar the two summaries are. It is suggested to use the F-measure based on *LCS* to evaluate the similarity between the two sums *X* length *m* and *Y* length *n*, assuming that *X* is a reference aggregate sentence, and *Y* is the summary sentence for viewing as follows:

$$P_{lcs} = \frac{LCS(X,Y)}{n}, \quad R_{lcs} = \frac{LCS(X,Y)}{m}, \tag{8}$$

where *LCS* (*X*, *Y*) is the length of a longest common subsequence of *X* and *Y*, and $\beta = P_{lcs} / R_{lcs}$.

The following values of the ROUGE metric were obtained: precision 32.8 %, recall 59.04 %, F-measure 34.47 %. Unfortunately, in works [10, 11], which describe summarization systems of texts in Russian, ROUGE values are not given, so it is not possible to compare those results with ours. We concluded that it is incorrect to compare our results with the systems for the English language, such as, for example, [24], since such low values of ROUGE can be associated with the peculiarities of the language type. Russian is an inflected language with developed morphology.

2) Secondly, we used expert evaluation. The precision of the obtained summaries estimated by experts was significantly higher. An expert evaluation showed that 86.43 % of the generated abstracts coincided with the author's abstracts or to some extent differed in meaning from the author's one (which in fact does not always indicate a low quality of the abstract) and 13.57 % were incorrectly selected fragments of the texts. It should be noted that the expert evaluation we obtained is higher than 71.6% in [11] and 80.84% in [10].

We have noticed that authors often use synonyms, paraphrase and change sentences in places. The expert evaluation confirms that the order of sentences in the abstract, as a rule, does not affect its general meaning. However, the ROUGE value does not consider this. In addition, sometimes automatically generated summary is longer than we would like to have (about 500 words instead of 250). This is due to the large number of meaningful sentences in the text.

3) Thirdly, we examined the precision, recall and F-measure calculated in a way similar to [2, 10]. Let us explain in more detail. Suppose that the automatically generated summary contains a set $W_1$ of keywords and multiword terms, a set $V_1$ of special words that are often found in scientific texts, and a set $D_1$ of discursive markers and connectors. The union of these sets is denoted by $N_1 : N_1 = W_1 \cup V_1 \cup D_1$. Similar sets can be defined for the author's summary: $N_2 : N_2 = W_2 \cup V_2 \cup D_2$. Then the precision (P),

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

recall (R) and F-measure will be calculated by the following formulas:

$$P = \frac{|N_1 \cap N_2|}{|N_1|}, \quad R = \frac{|N_1 \cap N_2|}{|N_2|}, \qquad (9)$$
$$F\text{–}measure = \frac{2PR}{P+R}.$$

A comparative evaluation of the results is given in table 4.

*Table 4: Evaluation of summarization*

| System | Precision | Recall | F-measure |
|---|---|---|---|
| Marcu (1998) | 73.53 % | 67.57 % | 70.42 % |
| Trevgoda (2009) | 67.03 % | 64.81 % | 66.03 % |
| Open Text Summarizer (2016) | 12.0 % | 24.20 % | 38.50 % |
| Scientific Text Summarizer (2018) | 75.23 % | 68.21 % | 71.55 % |

The advantage of the proposed formulas is that they allow us to evaluate the contribution of each characteristic and various their combinations to the overall evaluation of the result. For example, you can evaluate the contribution of only markers and connectors, or only special scientific terms, or both but without key words and expressions, etc. In the future, we plan to conduct a similar study of this issue. The following values were obtained: accuracy of 75.23%, completeness of 68.21% and F-measure of 71.55%, which confirms the effectiveness of the proposed methods.

Possible improvement of the algorithm proposed in this article, in our opinion, is to take into account the cases of anaphora [25] and part-of-speech homonymy [26], and fill up the linguistic knowledge base with markers and connectors.

## 5. A COMPREHENSIVE SECTION OF OPEN RESEARCH ISSUES IDENTIFIED

To improve the estimates obtained, the number of templates for smoothing will be increased, the lists of markers and connectors will be supplemented. Experiments will be conducted with texts from various scientific fields. And also it was decided to add both inflectional (English, )and agglutinative languages as (Turkish, Kazakh, Korean).

## 6. CONCLUSION

In this paper, we described an approach to automatic summarization of scientific and technical texts in Russian. We extract most significant sentences based on discursive markers and connectors. Keywords, multiword terms, and some special words that are often present in scientific and technical texts are also taken into account.

The system is implemented in Python3, a tool for working with Postgre SQL databases is also used. The external libraries Scikitlearn, Gensim, Tensor Flow, NLTK, Big ARTM, Flask and some others were used. For a formal description of text transformations, predicate calculus formulas are used. To build unigrammatic thematic models, the ARTM algorithm is used in the implementation of the Big ARTM library.

The unigram model is expanded with multiword terms by means of a modification of the RAKE algorithm, which was adapted for working with texts in Russian. Experiments have shown the high quality of the proposed algorithm.

However, it should be noted that in the case of a large number of formulas, drawings and graphs in the source text, the method works worse. Among the shortcomings, it should be noted the need to manually configure the knowledge base. The proposed approach can be used in information retrieval and automatic summarization systems. In the future, we plan to conduct experiments with texts from various scientific fields in other languages.

The main results of the research:

1. A hybrid method has been developed that allows you to receive abstracts (annotations) of high quality and determine the topics of texts in the form of a set of key terms. The proposed method is based on the use of a linguistic knowledge base, graphical representation of texts and machine learning.
2. Formally describes the methodology for detecting important elements in the text, based on the concepts of the theory of rhetorical structures. A linguistic database based on the analysis of the sublanguage of abstracts was created, used to evaluate the weights of the sentences of a quasi-abstract.
3. An algorithm for constructing extended thematic models of collections of text documents is proposed.
4. The procedure for smoothing sentences is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

described, which allows to make the text of the abstract (abstracts) more coherent and consistent.

5. The proposed models, methods and algorithms are implemented in the form of a system that allows you to automatically generate annotations of scientific and technical articles.

6. A collection of texts of scientific articles in Russian (about 1200 texts) for conducting experiments has been compiled. Computational experiments have been carried out confirming the effectiveness of the proposed methods and algorithms.

## ACKNOWLEDGMENTS

This study was funded by RFBR according to the research project N 19-07-01134.

## REFERENCES:

[1] M. I. Ananyeva and M. V. Kobozeva, "Development of the corpus of Russian texts with markup based on the Rhetorical Structure Theory", in *Proceedings of the International Conference "Dialog"*, 2016. Retrieved from: www.dialog-21.ru/media/3460/ananyeva.pdf

[2] D. Marcu, "Improving summarization through rhetorical parsing tuning", in *Sixth Workshop on Very Large Corpora*, 1998, pp. 206-215.

[3] E. Hovy and Ch. Y. Lin, "Automated text summarization and the SUMMARIST system", in *Proceedings of the TIPSTER Text Program*, 1998, pp. 197-214.

[4] K. Ono, K. Sumita, S. Miike, "Abstract generation based on rhetorical structure extraction", in *Proceedings of Coling 94*, 1994, pp. 344-348.

[5] F. Andonov, V. Slavova, G. Petrov, "On the open text summarizer", in *International Journal "Information Content and Processing"*, vol. 3, no. 3, 2016. Retrieved from:http://www.foibg.com/ijicp/vol03/ijicp03-03-p05.pdf

[6] S. Teufel and M. Moens, "Summarizing scientific articles: experiments with relevance and rhetorical status", in *Computational Linguistics*, vol. 28, no. 4, 2002, pp. 409-445.

[7] W. Bosma, "Query-Based summarization using rhetorical structure theory", in *Proceedings of 15th Meeting of CLIN*, 2005, pp. 29-44.

[8] S. H. Huspi, "Improving single document summarization in a multi-document environment", in PhD thesis, RMIT University, Melbourne, Australia, 2017.

[9] S. Mithun, "Exploiting rhetorical relations in blog summarization", in PhD thesis, Concordia University, Montreal, Canada, 2012.

[10] S. A. Trevgoda, "Methods and algorithms of automatic text summarization based on the analysis of functional relations", in Abstract of PhD Thesis. St. Peterburg, Russia, 2009.

[11] P. G. Osminin, "Construction of a model for abstracting and annotating scientific and technical texts focused on automatic translation", in PhD thesis. Chelyabinsk, Russia, 2016.

[12] D. Pisarevskaya, M. Ananyeva, M. Kobozeva, A. Nasedkin, S. Nikiforova, I. Pavlova, A. Shelepov, "Towards building a discourse-annotated corpus of Russian", in *Computational Linguistics and Intellectual Technologies*, vol. 16 (23), no. 1, 2017, pp. 194-204.

[13] K. Vorontsov, O. Frei, M. Apishev, P. Romov, M. Dudarenko, "BigARTM: open source library for regularized multimodal topic modeling of large collections", in *Proceedings of the International Conference on Analysis of Images, Social Networks and Texts (AIST)*, 2015, pp. 370-384.

[14] W. Mann and C. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization", in *Text-Interdisciplinary Journal for the Study of Discourse*, vol. 8, no. 3, pp. 1988, 243-281.

[15] M. Louwerse, "An analytic and cognitive parameterization of coherence relations", in *Cognitive Linguistics*, vol. 12, no. 3, 2001, pp. 291-315.

[16] T. Hofmann, "Probabilistic latent semantic indexing", in *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99)*, 1999. pp. 289-296.

[17] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet allocation", in *Journal of Machine Learning Research*, vol. 3, 2003, pp. 993-1022.

[18] K. V. Vorontsov and A. A. Potapenko, "Regularization of probabilistic thematic models to increase interpretability and determine the number of topics", in *Proceedings of the International Conference "Dialog"*, 2014, pp. 676-687.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[19] T. V. Batura and S.E. Strekalova, "An approach to building extended topic models of russian texts", in *Vestnik NSU. Series: Information Technologies*, vol. 16, no. 2, 2018, pp. 5-18.

[20] K. Vorontsov, "Welcome to BigARTM's documentation", 2015, Retrieved from: http://bigartm.readthedocs.io/en/stable/

[21] S. Rose, D. Engel, N. Cramer, W. Cowley, "Automatic keyword extraction from individual documents", in *Text Mining: Applications and Theory*, vol. 12, 2010, pp. 3-20.

[22] D. Das and A. Martins, "A survey on automatic text summarization literature", *Survey for the Language and Statistics II course at CMU*, vol. 4, 2007, pp. 192-195.

[23] Ch. Y. Lin, "ROUGE: A package for automatic evaluation of summaries", in *Workshop on Text Summarization Branches Out*, 2004, pp. 74-81.

[24] J. J. Zhang, H. Y. Chan, P. Fung, "Improving lecture speech summarization using rhetorical information", in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2007, pp. 195-200.

[25] A. Kozlova, O. Gureenkova, A. Svischev, T. Batura, "A hybrid approach for anaphora resolution in the Russian language", in *Siberian Symposium on Data Science and Engineering (SSDSE),* 2017, pp. 36-40.

[26] T. Batura and E. Bruches, "Combined approach to problem of part-of-speech homonymy resolution in Russian texts", in *International Russian Automation Conference "RusAutoCon 2018"*, 2018, pp. 4-9.