A SUMMARIZATION MODEL BASED ON THE COMBINATION OF EXTRACTION AND ABSTRACTION

Osminin P. G. (osperevod@gmail.com)

South Ural State University, Chelyabinsk, Russia

We suggest a model of automatic summarization for scientific and technical texts. This model combines extractive and abstractive approaches for summarization and was developed on the basis of comparative analysis of authors' summaries and full texts of corresponding papers. The model consists of three main components: a keyword extractor, a domain and task oriented static knowledge base and a summarization algorithm. The keyword extractor is off-the shelf tool LanAKey_Ru, adapted to the application. Static knowledge includes stop lexicons, conceptual net, templates for summary content selection and rules for the generation. Stop lexicons are used for removing text segments irrelevant for the document summary. The conceptual net is used for semantic analysis of a document text helping content selection. Templates for information extraction are frame structures. Their slots are to be filled with extracted fragments of document sentences. Rules for summary generation define the grammar of summary sentences and their order. The summarization algorithm consists of four top level procedures—preprocessing, analysis, content selection and summary text generation. The model is described on the example of Russian scientific papers in mathematical modeling domain.

Keywords: automatic summarization, information extraction, knowledge base, conceptual net

1. Introduction

Automatic summarization allows quickly processing of great volumes of the scientific and technical documentation that is especially important in modern conditions when society encountered with the information overload problem.

Research in the field of automatic text summarization continue more than 55 years, but the problem of automatic creation of high-quality summaries still is not solved [29]. Automatic summarization methods can be divided into 3 groups: extractive methods, abstractive methods and hybrid methods.

Extractive methods [1, 9, 17, 21, 23, 26] create the text of the summary by extraction of the most significant text fragments (sentences, paragraphs etc.) of the source document. Fragments, as a rule, are extracted without changes and are inserted in the summary in the same sequence as in the source. The relevance of fragments can be defined by various criteria, for example by containing keywords, by fragment's location in the source text (titles, subtitles etc.), by presence of cue phrases [20], for example, “the following results are obtained”, "it is important to note" etc. The advantages of extractive methods are relative independence of subject domain, lack of necessity of the detailed linguistic analysis and creation of extensive knowledge bases. Disadvantages of these methods consist in possible incoherence of summaries because text fragments of the source document are usually extracted without any processing. Also no information generalization is performed because words are not substituted by the more general concepts [11].

Abstractive methods [12, 19] create the text of the summary, as a rule, in three stages [30]: the source text is analyzed by means of the deep linguistic analysis, then an internal formal representation of the source text meaning is created, for example, in the form of frames, a semantic net, scripts etc. [7]. At this stage ontologies and knowledge bases of subject domains can be used. The compression of the source text meaning defines the summary content. The last stage is the generation of the summary text in a natural language [24].

Advantages of abstractive methods consist in providing more quality summary, than by extractive methods. The disadvantage is complexity for practical realization, abstractive methods need considerable amount of linguistic knowledge.

Due to complexity of abstractive methods the majority of automatic summarization methods are based on extraction of text fragments, and these methods show good results for specific types of texts [8].

Hybrid methods of automatic summarization are developed to overcome disadvantages of abstractive and extractive methods. In hybrid methods, the sentences (or their parts) are extracted from the source text and processed in different ways. For example, some parts of the sentences are omitted, some sentences are merged, sentences are inserted in the abstract in the other order as in the source etc. [25, 22]. Difficulties of hybrid methods developing consists in a choice of the most suitable combination of abstraction and extraction. In comparison with purely abstractive methods hybrid methods are easier to develop, in comparison with purely extractive methods hybrid methods can provide better quality of the resulting summary.

In automatic text summarization evaluation of the results is a very challenging problem. There are many works devoted to this problem [4, 14, 15, 16]. In [2] two groups of evaluation methods are distinguished—intrinsic and extrinsic. The intrinsic evaluation is oriented toward quality of the summary itself. For example, coherence of the text, its fluency, informativeness of the summary. The intrinsic evaluation often carrying out with participation of human experts who judge the quality of the summary comparing it with the gold standard (summaries created by human) or with results of other automatic summarization systems. However the problems of agreement between judges exist, as it is possible to create various summaries for the same text [6 18]. There are automatic evaluation methods, for example ROUGE [5].

The extrinsic evaluation assumes solving some task by means of the summary—for example, understanding of the full text by its summary. The judges are given a summary of the text and are asked some questions. Answers mean the comprehension of the full text. If the judge can provide answer the summary is considered correct.

In this paper we made attempt to contribute to a solution of one of important problems of automatic information processing and we present a hybrid summarization model for scientific articles in Russian for “mathematical modelling” domain. The model is based on a combination of extractive and abstractive approaches. To our knowledge there are no hybrid automatic summarization methods for Russian. The model is created on the basis of the comparative analysis of full texts of scientific articles and corresponding authors' summaries.

2. The Comparative Analysis of Scientific Articles and Authors' Summaries

The analysis purpose consisted in detection of formal criteria of relevant fragments selection for the summary in the article’s text and choosing a technique used for summarization: extraction or abstraction, depending on what of these methods are used by the human in summary creation.

When analyzing overlaps of sentences of authors' summaries and sentences of corresponding articles the following parameters were considered:

- text presentation of the same content in the summary and article
- location of a relevant fragment for the summary in article
- lexical markers of the relevant information for the summary.

For analysis we select corpora of 107 full scientific articles (total amount of 203,729 word forms, without references) and corresponding authors' summaries (total amount of 4924 word forms), the average compression of full texts of articles was 41.4. After analysis of the material, we have found that the majority of authors’ summaries—54.3% is written by purely sentence extraction from the text. In 36.2% cases authors processed the extracted sentences from the text—paraphrased, omitted the unimportant information, added the new text, i.e. they combined extraction and abstraction. In 9.5% cases authors wrote summaries only of the new text (pure abstraction). As the majority of summaries are written by the authors from article’s sentences and/or the edited fragments of article we oriented our summarization model on a combination of extractive and abstractive approaches.

According to the requirements of the Russian state standard [13] in the summary text we can distinguish four informational parts: “Theme”—the information on a subject and an article’s theme, “Aim”—the information on the work purpose, “Method”—the information on a method or methodology of carrying out the work, “Result”—the information on results of work, a range of use of these results.

In the article’s text each of four mentioned types of the information, as a rule, is accompanied by lexical markers. To describe a theme of article markers “содержать” (to contain), “глава” (chapter), “раздел” (paragraph) etc. can be used. Markers “метод” (method), “инструмент” (tool), “находить” (to discover) etc. can be used to describe a method of research. We divided all lexical markers into the groups corresponding to specified informational parts—“Theme”, “Aim”, “Method” and “Result”. In each group markers are divided on the following semantic types: objects, attributes of objects, relations, attributes of relations. Objects describe a subject, relations describe relations between objects, attributes describe features of objects or relations.

Lexical, semantic-informational and morphosyntactic properties of markers are in a characteristic correlation. For example, the markers defining objects are expressed by nouns and pronouns; the markers defining the relations between objects are expressed by verbs and the markers defining attributes of objects and relations are expressed by adjectives, pronouns, adverbs. In the article's sentences markers can function as independent lexemes or can be a part of longer phrases. In the latter case markers according to a category are a part of noun phrases, verb phrases or adjective phrases containing terms of subject domain of mathematical modelling. For example, in a sentence "Рассмотрим следующую обратную задачу спектрального анализа", the marker "задача" (problem) is contained in a noun phrase "обратную задачу спектрального анализа".

3. A Model of Automatic Summarization

The model consists from three main components: a keyword extractor is off-the shelf tool LanAKey_Ru, the knowledge base and a summarization algorithm.

A keyword extractor LanAKey_Ru was developed by Sheremetyeva S. O. for extraction of keywords from patents in English, and then has been modified for extraction of keywords from articles on mathematical modelling in Russian [27, 28].

3.1. Knowledge Base

The knowledge base of automatic summarization model consists of the following main components: 1) stop-lexicons, 2) a conceptual net in the form of a rooted tree, 3) sets of templates for information extraction and 4) rules of the analysis of the full document, generation of the summary text, arrangement of sentences in the summary text.

The stop-lexicons are used for removing from the full text of article the irrelevant information for the summary for the purpose of facilitation of the further analysis. The stop-lexicons consist of three lists:

a) the list of the stopwords which are deleted (for example, "итак" (so), "однако" (however), "окончательно" (finally)),
b) the list of the words defining a part of a sentence for removing (for example if in a sentence the word "где" (where) is found then the sentence part after this word is deleted),
c) the list of the words defining a sentence for removing (for example, "положим" (let us assume), "пусть" (let), "обозначим" (let us define)).

The conceptual net is used for the semantic analysis—detection of lexical markers in the document text. The net consists of terminal and non-terminal nodes and the links realizing the relation of inclusion. The root of the tree is concept "Summary", in non-terminal nodes there are concepts, corresponding to informational parts of the summary "Theme (T)", "Aim (A)", "Method (M)", "Result (R)" and to semantic types of markers "Object (O)", "Relation (P)", "Attribute of object (AO)" and "Attribute of relation (AP)". Concepts reflect the required types of the information for summary.

Terminal nodes—lexical units—are the markers, realizing these concepts in the text. Lexicons of markers include all word forms of marker lexemes that have been selected in the course of the analysis of a sublanguage of mathematical modelling.

The fragment of conceptual net for a node "Method (M)" is showed on Fig. 1.

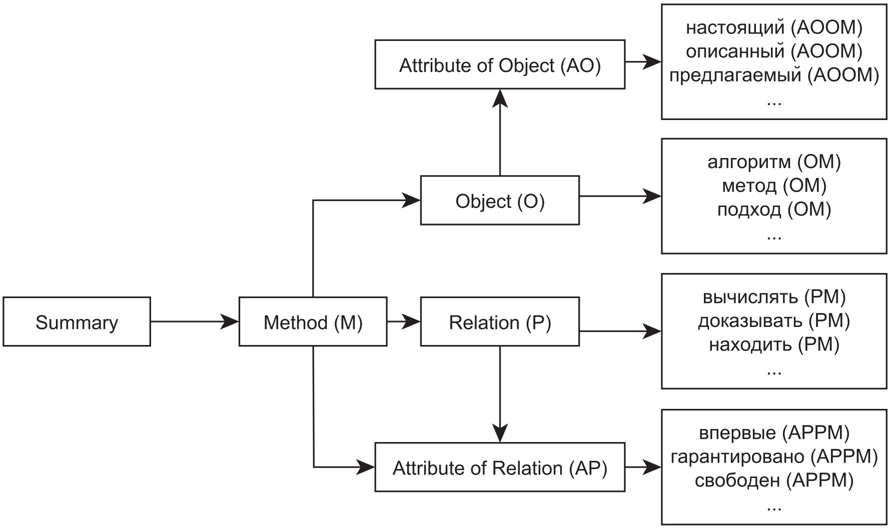

Fig. 1. A fragment "Method (M)" of a conceptual net

Templates for information extraction are divided into four categories—Theme, Aim, Method, Result, according to type of the extracted information required in accordance with the Russian state standard. Templates are frame structures of a following kind:

Template	::= (IP (structure))
IP	::= {theme, aim, method, result}
Structure	::= (X Phrase X … Phrase …X)
X	::= (word word ... word)
Phrase	::= {NP(MARKER T), VP(MARKER T), AP(MARKER T)}
Marker	::= (marker(net code))
Template number	::= (natural number)
Sentence number	::= (natural number)
Weight	::= (natural number)

where,

IP—informational part of the summary; structure—structure of an article's text fragment; X—chain of consecutive words of a fragment, can be empty; marker—a terminal node of the net; code—the net code; T—the term (can be empty); NP—noun phrase, VP—verb phrase, AP—adjective phrase, template number—a sequence number of a template which is defined at template filling, sentence number—a sequence number of the sentence used for filling a template, weight—template weight is calculated at a template weighing. If some part of a template is in square brackets it can be empty.

Rules of text generation represent the text fragments in filled slots of template in the coherent sentences. From each template, as a rule, one sentence is generated. If templates belong to the same sentence or begin with identical words ("рассматривается" (is considered), "в работе" (in the work)) then from these templates one sentence is generated. Grammatical rules include:

1) Verbs of the first person are substituted with impersonal forms (the third person with the ending "-ся"), the further noun is put in the nominative case.
2) Titles of article parts (paragraph, section, table etc.) can either be omitted, or can be substituted by expression "в статье" (in paper), "статья" (paper).
3) In templates Theme, Aim, Method the verb is used in the present. In a template Result the verb is used in the past tense (except verbs "является" (is), "смогут" (can), "позволят" (will allow)).

The order of generated sentences in the summary text is defined by the following rule:

1) Sentences follow in this order Theme, Aim, Method, Result. If in a category there are more than one sentence they are ordered by weight which is calculated from weight of the keywords and weight of the markers.

3.2. Automatic Summarization Algorithm

The algorithm consists of 4 main procedures—preprocessing of the article's text, the analysis of the article's text, content selection for a summary, generation of the summary text. The main procedures include subprocedures. Each procedure receives some information as an input and produces result which is used as the input for the following procedure.

The preprocessing procedure consists of three consecutive subprocedures—sentence segmentation, compression of the text, extraction of keywords.

The first subprocedure performs a sentence segmentation of the article's full text. We consider the sentences as text segments from a dot to a dot.

Then text compression subprocedure removes from the text irrelevant fragments for the summary. At this step the stop-lexicons from the knowledge base are used. After fragments' removing the sentences containing less than five words are removed (a word is a text fragment from a space to a space). Text compression is intended for facilitation of the further analysis.

After compression the extraction of keywords from the text by means of extractor LanAKey_Ru is performed. LanAKey_Ru is capable to extract nominal phrases up to four words without a preliminary annotation of the text. Keywords in our model are the most relevant noun phrases (NP) of the article. Relevance is calculated according to empirically found formula $(d / D) + U * 10$, where d—number of sentences where NP occurred at least once, D—number of sentences in the text, U—uniqueness, shows that the keyword functions individually, instead being a part of a longer phrase. This parameter is calculated as a difference between frequency of NP and the sum of frequencies of longer noun phrases containing the given NP. All specified parameters are defined in the extractor LanAKey_Ru. From the text of a full article the 10 most relevant keywords are extracted.

Analysis procedure consists of two subprocedures—partial morphosyntactic analysis and the semantic analysis. At a stage of morphosyntactic analysis detection of lexical groups and their part of speech tagging is performed. The morphosyntactic analysis is performed as follows. The phrases coinciding with keywords are tagged with noun phrase tags and are assigned the weight (relevance) automatically defined by LanAKey_Ru. Thus, LanAKey_Ru not only extracts the keywords and define their relevance but also performs an essential part of the morphological analysis, since NP—the most frequency lexical group in any text type. Then morphosyntactic analysis is finished by means of the software of Aot.ru project [3].

In such annotated article's text the semantic analysis is carried out by means of conceptual net. The lexicon from the terminal nodes of a net is compared to the annotated text. When the coincidence occurs the marker is given the network code—a path from a terminal node to the net vertex. For example, the net code for a marker "подход" (approach) is OM (Object-Method), a net code for a marker "prove" (доказывать)—PM (Relation-Method). Due to homonymy of markers—the same word forms can express various relations—during semantic analysis one marker can be given various net codes. The resolution of marker ambiguity takes place by means of templates at later stage of the analysis.

The third procedure—content selection consists of two subprocedures—scoring of sentences' weight (relevance) and filling of templates. The sentence weight (relevance) is calculated by the following formula:

$$W = 10N + M_i + K_i$$

where

W—weight of the sentence (relevance),

N—number of keywords in a sentence, the multiplier 10 was found empirically,

M_i—weight of all markers in a sentence (the weight of one marker is 10),

$К_i$—weight of all keywords in a sentence (weight is taken from Lana-key_Ru)

We define that for small texts (up to 9000 characters with spaces) the selection threshold is 10 most relevant sentences, for the bigger texts—10% of the most relevant sentences.

Filling of templates subprocedure begins with review of the selected sentences from left to right. In templates for information extraction the order of markers and other words in a sentence is set. If the sentence (or its part) satisfies template requirements then template slots are filled with sentence parts. Applying of templates on text fragments occurs between punctuation marks.

For example, the following sentence from the article's text: *В этом параграфе изложены используемые при получении основного результата факты из классической теории полугрупп операторов* satisfies the template Theme from the knowledge base *(Theme [X] [(AP(Marker(AOT) T)] (NP(Marker(OT) T)) [(AP(Marker(APT) T)] (VP(Marker(PT) T) X))*. The result of this template filling by sentence fragments is shown below:

IP	::= Theme
X	::= В
AP(Marker(AOT) T	::= этом
NP(Marker(OT) T	::= параграфе
AP(Marker(APT) T	::=
VP(Marker(PT) T	::= изложены
X	::= используемые при получении основного результаты факты из классической теории полугрупп операторов
Template number	::= 2
Sentence number	::= 7
Weight	::= 6

The order of template applying is the following Theme, then Aim, Method, Result.

The example of formally generated summary for article [10] and its author's summary is shown below.

Formally generated summary

В статье рассматривается задача Коши для абстрактного линейного эволюционного уравнения с памятью в банаховом пространстве. В статье изложены используемые при получении основного результата факты из классической теории полугрупп операторов.

С помощью принципа сжимающих отображений доказана однозначная локальная разрешимость этой задачи в смысле классических решений.

Результаты работы позволят с одной стороны перейти к рассмотрению полулинейных эволюционных уравнений с памятью. Полученный результат использован при исследовании начально-краевой задачи для параболического интегро-дифференциального уравнения с памятью. Основным результатом данной работы является доказательство однозначной локальной разрешимости этой задачи.

The author's summary

Доказана локальная однозначная разрешимость задачи Коши для линейного эволюционного уравнения с секториальным оператором и с интегральным оператором памяти в банаховом пространстве. Результат работы проиллюстрирован на примере начально-краевой задачи для интегро-дифференциального уравнения с частными производными.

Formally generated summary reflects a content presented in the author's summary and satisfies the requirements of the Russian state standard in a greater degree. The evaluation of the results was performed manually and consisted in comparison of formal summaries with authors' summaries in order to detect the informational parts required in accordance with the Russian state standard. In case of considerable contradiction between the formal and author's summary we resorted to help of the experts. As a result we discover that about 70% of formal summaries was generated correct. Comparison with automatic summarization systems was not performed, as we did not have a possibility to get the data of such systems for comparison.

4. Conclusion

In this paper we presented the model of automatic summarization, developed on the basis of comparative analysis of authors' summaries and full texts of corresponding articles.

In the presented model extractive and abstractive approaches are realized. We described the model parts: 1) keyword extractor, 2) the knowledge base and 3) algorithm of automatic summarization. Steps of automatic summarization algorithm are developed. Formally generated summaries, as a rule, coincide with authors' summaries by content and satisfy the requirements of the Russian state standard in a greater degree.

References

1. *Abuobieda A., Salim N., Albaham A. T., Osman A. H., Kumar Y. J.* (2012), Text summarization features selection method using pseudo genetic-based model, International conference on information retrieval knowledge management, Kuala Lumpur, pp. 193–197.
2. *Afantenos Stergos, Karkaletsis Vangelis, Stamatopoulos Panagiotis* (2005), Summarization from medical documents: a survey, Artificial Intelligence in Medicine. Vol. 33 Issue 2, pp. 157–177.
3. Automatic text processing [Avtomaticheskaja Obrabotka Teksta], available at: http://aot.ru/
4. *Chin-Yew Lin, Eduard Hovy* (2002), Manual and Automatic Evaluation of Summaries, Association for Computational Linguistics. Proceedings of the Workshop on Automatic Summarization (including DUC 2002), Philadelphia, pp. 45–51.
5. *Chin-Yew Lin* (2004), ROUGE: A Package for Automatic Evaluation of Summaries, Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain, pp. 74–81.
6. *Das D., Martins Andre F. T.* (2007), A Survey on automatic text summarization. Unpublished manuscript, Literature survey for Language and Statistics II, Carnegie Mellon University, 31 p.
7. *DeJong G. F.* (1982), An Overview of the FRUMP System, in Strategies for Natural Language Processing, Lawrence Erlbaum, Hillsdale, New Jersey, pp. 149–176.
8. *Dubinina E. Ju.* (2013), Scientific text compression: methods and models [Kompressija nauchnogo teksta: metody i modeli], Avtoreferat dis. … kandidata filologicheskih nauk: 10.02.21, Rossijskij gosudarstvennyj pedagogicheskij universitet im. A. I. Gertsena.

9. *Edmundson H. P.* (1969), New methods in automatic extracting, Journal of the ACM, vol. 16 Issue 2, pp. 264–285.
10. *Fedorov V. E., Staheeva O. A.* (2008), On Local Solvability of Linear Evolutionary Equations with Memory [O lokal'noj razreshimosti linejnyh èvoljutsionnyh uravnenij s pamjat'ju], Bulletin of the South Ural State University. Series Mathematical Modelling, Programming & Computer Software [Vestnik Juzhno-Ural'skogo gosudarstvennogo universiteta. Serija: Matematicheskoe modelirovanie i programmirovanie] no. 27 (127), pp. 104–109
11. *Genest Pierre-Etienne, Lapalme Guy, Yousfi-Monod Mehdi* (2009), Hextac: the creation of a manual extractive run, Proceedings of the Second Text Analysis Conference, Gaithersburg.
12. *Genest Pierre-Etienne, Lapalme Guy* (2012), Fully Abstractive Approach to Guided Summarization, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, pp. 354–358.
13. GOST 7.9-95 (1995), System of standards on information, librarianship and publishing. Informative abstract and indicative abstract. General requirements [GOST 7.9-95. Sistema standartov po informatsii, bibliotechnomu i izdatel'skomu delu. Referat i annotatsija. Obshchie trebovanija] Izdatel'stvo standartov, Moscow.
14. *Hassel Martin* (2004), Evaluation of Automatic Text Summarization. A practical implementation Licentiate Thesis Stockholm, Sweden, Royal Institute of Technology (KTH) 69 p.
15. *Hirao Tsutomu, Okumura Manabu, Yasuda Norihito, Isozaki Hideki* (2007), Supervised automatic evaluation for summarization with voted regression model. Vol. 43 No 6, pp. 1521–1535.
16. *Hobson Stacy President, Dorr Bonnie J., Monz Christof, Schwartz Richard* (2007), Task-based Evaluation of Text Summarization Using Relevance Prediction, Information Processing and Management. Vol. 43 No 6, pp. 1482–1499.
17. *Jatsko V. A.* (2002), Symmetrical Summarization: Theoretical Foundations and Methods [Simmetrichnoe referirovanie: teoreticheskie osnovy i metodika], NTI. Ser. 2 no. 5. pp. 18-28.
18. *Karen Sparck Jones* (2007), Automatic summarising: The state of the art, Information Processing and Management, vol. 43 issue 6, pp. 1449-1481.
19. *Korhova O. V.* (2001), Method for mathematic formalization of the Russian language in automatic text summarization aspect [Metod matematicheskoj formalizatsii russkogo jazyka v zadache avtomaticheskogo referirovanija tekstov], dis. kand. fiz-mat. nauk: 01.01.09.
20. *Leonov V. P.* (1986), Summarization of scientific and technical literatre [Referirovanie i annotirovanie nauchno-tehnicheskoj literatury], Nauka, Novosibirsk.
21. *Luhn H. P.* (1958), The Automatic Creation of Literature Abstracts, IBM Journal of Research and Development, vol. 2, no. 2, pp. 159–165.
22. *Polanyi Livia, van den Berg Martin H., Lorenzo Thione Giovanni, Crouch Richard S., Culy Christopher D., Ahn David D.* (2009), Systems and methods for hybrid text summarization, United States Patent US 7,610,190 B2

23. *Prihod'ko S. M., Skorohod'ko È. F.* (1982), Automatic summarization based on inter phrase links [Avtomaticheskoe referirovanie na osnove analiza mezhfrazovyh svjazej], NTI. Ser. 2 no. 1. pp. 27–32.
24. *Radev Dragomir R., McKeown Kathleen R.* (1998) Generating Natural Language Summaries from Multiple On-Line, Computational Linguistics—Special issue on natural language generation, vol. 24, no. 3, pp. 470–500.
25. *Saggion Horacio, Lapalme Guy* (2002), Generating indicative–informative summaries with SumUM, Computational Linguistics. vol. 28, no. 4, pp. 497–526
26. *Sevbo I. P.* (1969), Structure of coherent text and automatization of summarization [Struktura sviaznogo teksta i avtomatizatciia referirovaniia], Nauka, Moscow.
27. *Sheremetyeva S.* (2009), An efficient patent keyword extractor as translation resource, MT Summit XII: Third Workshop on Patent Translation, Ottawa, pp. 25–32.
28. *Sheremetyeva S.* (2012), Automatic Extraction of Linguistic Resources in Multiple Languages, Proceedings of NLPCS 2012, 9th International Workshop on Natural Language Processing and Cognitive Science in conjunction with ICEIS 2012, Wroclaw, pp. 44–52.
29. *Sheremetyeva S. O.* (2013), On interactive summarization oriented to machine translation [Interaktivnoe referirovanie, orientirovannoe na mashinnyj perevod], Bulletin of the South Ural State University Series Linguistics [Vestnik Juzhno-Ural'skogo gosudarstvennogo universiteta. Serija: Lingvistika], vol. 10, no. 1, pp. 89–92.
30. *Tarasov S. D.* (2010), Modern methods for automatic summarization [Sovremennye metody avtomaticheskogo referirovaniia], St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunications and Control Systems [Nauchno-tehnicheskie vedomosti Sankt-Peterburgskogo gosudarstvennogo politehnicheskogo universiteta. Informatika. Telekommunikatsii. Upravlenie] no. 6 (113) pp. 59–74.