Extraction of Morphological Features of Technical Systems from Russian Patent

Anatoliy Kharitonov[1], Dmitriy Korobkin[1[0000-0002-4684-1011]], Sergey Fomenkov[1[0000-0001-9907-4488]], Sergey Kolesnikov[1[0000-0002-6910-7151]]

[1] Volgograd State Technical University, Lenin av. 28, Volgograd, Russia

dkorobkin80@mail.ru

Abstract. The task of automation of the synthesis of innovative solutions in the field of technical systems and technologies is one of the most priority problems of science. The authors propose to automate the most important, initial stages of the design of new technical systems and technologies based on updated knowledge bases obtained from the world patent database, including the RosPatent patent database. According to the method of morphological analysis and synthesis, it is assumed that the main structural features (functions of technical objects) are extracted from some technical solution (patent). All these features are reduced to a morphological matrix, combined, which gives a lot of new solutions. The paper describes the developing a software for extracting the descriptions of the technical functions from Russian patents. The grammar of the presentation of technical functions descriptions according to the model "Action-Object-Condition" in the Russian-language patents was formed; algorithms for the initial processing of the patent database, the extraction of technical functions through the analysis of dependency trees, the formation of the morphological matrix was developed. The software consisting of a module of patent database processing; a module of text segmentation of the patent formula; a module of semantic text analysis; a module of extracting descriptions of technical functions; a module of presenting the results of patent database processing, was tested to solve practical problems.

Keywords. Technical functions, Natural Language processing, patents, RosPatent, Link Grammar Parser, grammar.

1 Introduction

The task of automation of the synthesis of innovative solutions [1,2] in the field of technical systems and technologies is one of the most priority problems of science. Authors of work suggest [3,4] to carry out automation of the major, initial stages of design of new technical systems and technologies based on the updated knowledge bases received from the world patent array [5,6], including the patent base of RosPatent [7].

In: P. Sosnin, V. Maklaev, E. Sosnina (eds.): Proceedings of the IS-2019 Conference, Ulyanovsk, Russia, 24-27 September 2019, published at http://ceur-ws.org

The new design decision with new properties that did not meet earlier is the result of the synthesis of an invention. Structural synthesis [1] implies the existence of a certain technical function, for which it is necessary to define some object or system that implements this technical function under certain conditions and restrictions. According to the method of morphological analysis and synthesis [1], it is assumed that the main structural features are extracted from some technical solutionы (alternative options). All these features are reduced to a morphological matrix, combined, which gives a lot of new solutions. One of the varieties of morphological analysis is the approach [8,9], where the functions of individual objects (it's elements) are taken as features. In this case, alternatives are the various implementations of these functions.

The algorithm of morphological analysis and synthesis contains the following sequential procedures:

- identification of the main features of the object;
- preparation of the form for filling the morphological matrix;
- the matrix fill the alternatives;
- choice of the best solutions;
- formation of the technical solution.

The purpose of the work is to carry out the extraction of technical functions from the RosPatent patents and present them in the form of a morphological matrix, which is the basis of the procedure for the synthesis of new design solutions.

2 The developed methods

2.1 The algorithm of patent database analysis

According to the results of the research, which consisted of the analysis of the subject area of patenting of inventions, familiarity with the structure of the patent, linguistic analysis of the patent formula, review of existing automated methods for the synthesis of new technical solutions, study of the stages of automatic text processing, the algorithm of the analysis of the patent array was proposed.

The input of the algorithm is a folder with patents granted by the Russian patent department. Technical functions are extracted from the texts of patent formulas, in which verbs determine the main action and which will unambiguously describe the essence of the invention. Then classes of actions of verbs are defined, thus forming generalized technical functions, and the functions themselves are written in the implementation variants (alternatives) of the selected generalized functions.

The output of the algorithm is a morphological matrix, presented visually on a program screen form and visually displaying the versions (alternatives) of the generalized technical function. The obtained morphological matrix can be used for the search of the fundamental idea for the synthesis of the new technical solution.

2.2 Extraction of descriptions of technical functions from patents

The following information is taken from an XML file of the RosPatent patent:

- <b110> … </b110> is number of the patent;
- <b220> <date> … </date> </b220> is date of issue of the patent;
- <ru-b542> … </ru-b542> is the name of the patent;
- <ru-b560> … </ru-b560> are references to other patents which are inspiration sources;
- <b721> <ru-name-text> … </ru-name-text> </b721> list of inventors;
- <claims> … </claims> is an invention formula.

Directly the text of a patent formula is put into the tags <claim-text></claim-text>. Formulas of patents are under construction on one template other than the usual sentence structure. Most often the formula represents one complex sentence in which several subordinate clauses, each of which supplements properties of an object about which it is told in the main clause. Parsing such a long sentence would be very time consuming and resource consuming. To reduce the number of mistakes in the results of the semantic analysis, this offer is divided into several parts (is segmented), and then each part of the original sentence is parsed separately.

The segmentation algorithm [10] consists of the following sequential procedures:

- the phrase “by p.” and “by paragraph” is deleted from the sentence.
- all inducement of phrases is replaced with the character of transfer of a line: "differing in the fact that"; "providing"; "including"; "which"; "whose"; also removed all prepositions relating to these words.
- are replaced by the line break character of the phrase: "a"; "but"; "where"; "and"; "thus"; "by"; "so that"; "besides"; "before"; "after".
- all signs of questions, colons, and semicolons are removed.
- the symbols denoting lists are removed, for example, "a)", "1." and so on.
- all multipurpose and unnecessary repetition of line breaks and commas clean up.

For the execution of semantic analysis of a segment of a patent formula the software of Link Grammar Parser [11], based on a link grammar is used. The parser accepts on an input sentence segments in a natural language and gives the communications found in the sentence with a marking of morphological features of words of a segment. The main idea of a link grammar is the representation of words as blocks with the connectors proceeding from them. Connectors can specify ("+") to the left or to the right ("-"). Two connectors, one of which specifies to the left and contact the connector specifying to the right form of communication. At the output, we receive not syntactic links (subject/object), but the relations between couples of words [12].

To parse the patent formula segments, the following Link Grammar Parser link types are required [11]:

- W – connects limit of the offer and the main word which is usually a noun;
- MV – connects verbs to their objects;

- SI – connects verbs to their objects answering a question that;
- E – connects an adverb to a verb;
- J – connects a pretext to a dependent noun or a pronoun;
- I – connects a verb to another a verb, dependent on it;
- A – connects a pronoun or an adjective to a noun;
- M – connects nouns to others dependent on them nouns;
- EI – connects a verb to an adverb;
- PI – connects verbal adjectives or participles to nouns;
- AXP – connects pronouns or adjectives to adjectives.

2.3 The context-dependent grammar of representation of the technical function

Description of the function of any technical object will be presented in the form of three sets represented by the formula [2]:

$$F=<D, G, H>, \quad (1)$$

where D – a set of the actions made and resulting in the desired result; G – a set of objects to which these actions are directed; H – a set of special conditions and restrictions of the performed operations.

The analysis of the set of sentences showed that the segments of patent formulas have a similar syntactic structure. This observation made it possible to construct a context-sensitive grammar (2) to extract the model components (1) from sentences of the patent formula by analyzing the links:

$$\text{Gram} = (T, N, <W>, R), \quad (2)$$

where T = {action, object, condition} is a set of terminals; N = {<MV>, <SI>, <E>, <J>, <I>, <A>, <M>} – set of non-terminals; <W> – initial non-terminal [13];

R is the set of production rules:

```
<W> → action <MV> <E> | action <MV> | action <E> | action
| ε
Action <MV> → action <MV> | action <MV> <MV>
action → action | <I> action | action <I>
<I> action → action action | <I> action action | action
<I> action
action <I> → action action | action action <I> | action
<I> action
<MV> → object <MV> | <A> object | <A> object <M> | object
<M> | object
<A> object → object object | <A> object object | object
<A> object
object <M> → object object object| object <A> object |
object object <M> | object <A> object <M>
condition <J> → condition | condition <E> | condition <J>
```

```
<A> condition → condition condition <M> | condition <A>
condition <M>
condition <M> → condition condition | condition condition
<M> | condition condition <J> | condition <A> condition
<MV> → <MV> | <SI>
<E> → <E> | <EI>
<A> → <A> | <AXP> | <PI>
```

The output of a sentence using this grammar means that a word in a sentence at the position of a terminal refers, respectively, to an action, object, or condition, depending on the type of terminal.

The links found by the parser are interpreted to represent a technical function according to the model (1). By analyzing certain types of relationships, you can discover words that are objects of action, the action itself, and the condition of the action (i.e., to define syntactic category).

The structuring of the set of possible values of the component D "Action" was made on the basis of a set of terms, the most common in the description of the functions of technical objects (for example, in the International patent classification), as well as selected through the analysis of unstructured information under the heading "Practical application" [14] available to the authors of the article of the Physical Effects (PhE) database (1200 descriptions of PhE) [15].

Action classifier (D) [2]:

- Transformation;
- Change;
- Increase;
- Decrease;
- Measurement;
- Connection;
- Separation;
- Stabilization;
- Destabilization;
- Creation;
- Destruction;
- Accumulation;
- Issuance;
- Research;
- Treatment;
- Management;
- Control;
- Passing;
- Isolation.

Using this classifier allows you to find a common class for cured actions from patent formulas. The general class will be a feature in the morphological matrix, and the actions themselves, with their objects and conditions, will be the implementation variants (alternatives) of this feature.

2.4 Example of extracting technical functions from patents

The text of the patent: "A signal converter that provides the formation of a code word of digital data according to the method of constructing the code word".

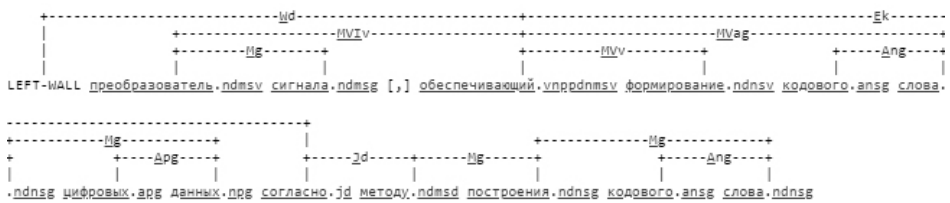

Fig. 1. Example of the tree of analysis of Link Grammar Parser

From the tree of dependencies shown in Figure 1 according to grammar (2) we receive the following:

```
<W > →
→ provides <MV> <E>
→ provides the formation <MV> <E>
→ provides the formation < a > words <M> <E>
→ provides the formation of a code word<M> <E>
→ provides the formation of a code word <a > data <E>
→ provides the formation of a code word of digital data
<E>
→ provides the formation of a code word of digital data
according to <J>
→ provides the formation of a code word of digital data
according to the method <M>
→ provides the formation of a code word of digital data
according to the method of construction <M>
→ providing the formation of a code word of digital data
according to the method of constructing <a> words
→ provides the formation of a code word of digital data
according to the method of constructing the code word.
```

The components of the technical function according to the developed model (1) recognized in the text of the patent formula will be as follows:

- Action (*D*) = “provides the formation”;
- Object (*G*) = “of a code word of digital data”;
- Condition (*H*) = “according to the method of constructing the code word.

2.5 Formation of the morphological matrix

The obtained components of the technical function representation according to the DGH (1) model, recognized in the sentence segments of patent formulas, technical functions are entered in the matrix together with the selected generalized technical function (TF) according to the action classifier (D) (Table 1).

Table 1. Extracted technical functions

No	Action (*D*)	Object (*G*)	Condition (*H*)	Generalized TF
1	provides the formation	of a code word of digital data	according to the method of constructing the code word	Creation
2	performs	sandwiched	on the postoperative data	Treatment
3	the result of generating	a modulating signal	from the IFFT calculator	Creation
4	forms	pressure	in an empty cylindrical form	Creation

Based on the generalized technical functions, a morphological matrix is constructed in which technical functions in the form of "object-condition-action" tuples of DGH (1) are used as alternatives to the execution of generalized technical functions. When building a morphological matrix (Table 2), all found actions (D) are viewed and those actions that have the same generalized technical function are recorded in the same column of the morphological matrix with their object and conditions.

Table 2. Morphological matrix

Creation	Treatment	…
D – provides the formation G – of a code word of digital data H – according to the method of constructing the code word	D – performs G – the inverse fast Fourier transform H – on the postoperative data	…
D – the result of generating G – a modulating signal H – from the IFFT calculator	…	…
D – forms G – pressure H – in an empty cylindrical form	…	…

The designer chooses the sign, i.e. the generalized technical function interesting him from the received morphological matrix, and in a column with the corresponding name looks through all possible alternatives of its execution. The review of the possible ways of realization of some function concentrated in one place can prompt the user the idea of improvement of the available design or creation of the absolutely new technical solution.

3 Results

The extracted technical functions and the morphological table are stored in the RDD (Resilient Distributed Dataset) scheme, which is used for the most rapid processing of patent data using MapReduce technology [16]. For extraction of technical functions from patent documents, the software consisting of 5 blocks is created:

- block of processing of the input patent document;
- block of segmentation of the text of the patent formula;
- block of semantic text analysis;

- block of extraction descriptions of technical functions;
- block for presenting the results of the processing of the patent databases.

4 Conclusion

The theoretical value of this work is: in the developed grammar of the description of technical functions in the texts of Russian patents; in the algorithms of the primary processing of the patent array, the extraction of technical functions through the analysis of dependency trees, the formation of the morphological matrix.

The functionality of the software has been tested on a set of test tasks.

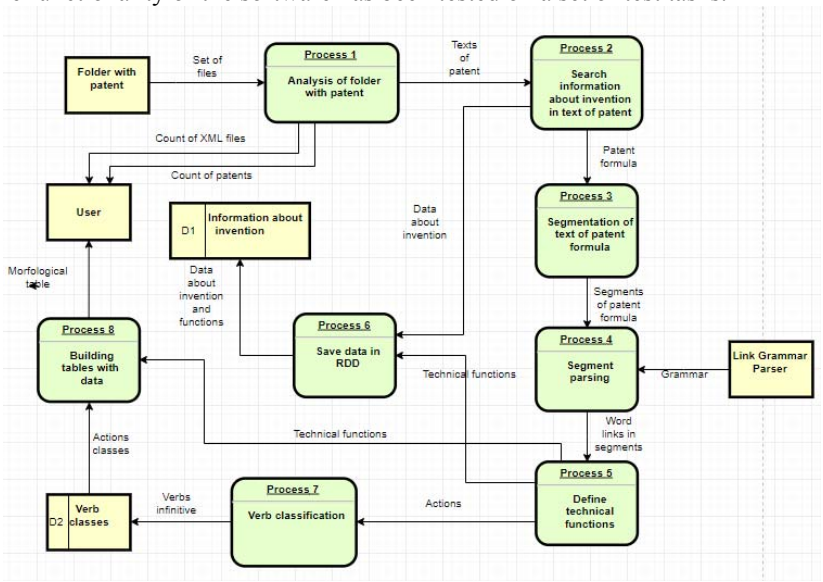

Fig. 2. Data Flow Diagram

Acknowledgment

The reported study was funded by RFBR (research project 18-07-01086), RFBR and Administration of the Volgograd region (projects 19-47-340007, 19-41-340016).

References

1. Polovinkin, A.I., 1991. Theory of New Technique Design: Laws of Technical Systems and Their Applications, Informelektro, Moscow
2. Kamaev, V.A., Fomenkov, S.A., Petrukhin, A.V., Davydov, D.A., 1999. Architecture of the automated system of conceptual design, SOFIE. Software Products and Systems (2), pp. 30– 34.

3. Davydov, D.A., Fomenkov, S.A., 2002. The automated design of linear structures of the physical principles of action of technical systems. Mechanician (2), pp. 33–35
4. Fomenkov, S.A., Kolesnikov, S.G., Korobkin, D.M., Kamaev, V.A., Orlova, Y.A., 2014. The information filling of the database by physical effects. Journal of Engineering and Applied Sciences, vol. 9, no. 10, pp. 422-426.
5. Korobkin, D.M., Fomenkov, S.S., Kravets, A.G., Kolesnikov, S.G., 2017. Prior art candidate search on base of statistical and semantic patent analysis. Proceedings of the International Conferences on Computer Graphics, Visualization, Computer Vision and Image Processing 2017 and Big Data Analytics, Data Mining and Computational Intelligence 2017 - Part of the Multi Conference on Computer Science and Information Systems 2017 11, pp. 231-238.
6. Fomenkov, S.A., Korobkin, D.M., Kolesnikov, S.G., Dvoryankin, A.M., Kamaev, V.A., 2014. Procedure of integration of the systems of representation and application of the structured physical knowledge. Research Journal of Applied Sciences, vol. 9, no. 10, pp. 700-703.
7. FIPS Rospatent. 2018. – http://www1.fips.ru/wps/portal/IPS_Ru.
8. Korobkin, D.M., Fomenkov, S.A., Golovanchikov, A.B., 2018. Method of identification of patent trends based on descriptions of technical functions. Journal of Physics: Conference Series, vol. 1015, s. 032065.
9. Korobkin, D.M., Fomenkov, S.A., Kolesnikov, S.G., Golovanchikov, A.B., 2016. Technical function discovery in patent databases for generating innovative solutions. Proceedings of the International Conferences on ICT, Society, and Human Beings 2016, Web Based Communities and Social Media 2016, Big Data Analytics, Data Mining and Computational Intelligence 2016 and Theory and Practice in Modern Computing 2016 - Part of the Multi Conference on Computer Science and Information Systems 2016, pp. 241-245.
10. Korobkin, D., Fomenkov, S., Kravets, A., Kolesnikov, S., Dykov M., 2015. Three-steps methodology for patents prior-art retrieval and structured physical knowledge extracting. Communications in Computer and Information Science, vol. 535, pp. 124-136.
11. Link Grammar Parser. 1998. –http://www.abisource.com/projects/linkgrammar.
12. Korobkin, D.M., Fomenkov, S.A., Kolesnikov, S.G., 2016. A function-based patent analysis for support of technical solutions synthesis. In Proceedings of 2nd International Conference on Industrial Engineering, Applications and Manufacturing, ICIEAM 2016 – Proceedings 2, s. 7911581.
13. Chomsky, N., Miller, G.A., 2018. Introduction to the formal analysis of natural languages. R.D. Luce, R.R. Bush, E. Galanter (Eds.), Handbook of Mathematical Psychology, vol. 2, Wiley, Amsterdam, pp. 269-321
14. Korobkin, D.M., Fomenkovб S.A., Kravets, A.G., 2018. Extraction of physical effects practical applications from patent database. Proceedings of 8th International Conference on Information, Intelligence, Systems and Applications, IISA 2017 8, pp. 1-5.
15. Korobkin, D.M., Fomenkov, S.A., Kolesnikov, S.G., 2014. Ontology-Based extraction of Physical Effect description from Russian text. In Proceedings of the European Conference on Data Mining 2014 and International Conferences on Intelligent Systems and Agents 2014 and Theory and Practice in Modern Computing 2014 - Part of the Multi Conference on Computer Science and Information Systems, MCCSIS 2014, pp. 260-262.
16. Karau, H., Konwinski, A., Wendell, P., Zaharia, M. Learning Spark: Lightning-Fast Big Data Analysis. ISBN 10: 1449358624. ISBN 13: 9781449358624.