Solution of the Answer Formation Problem in the Question-Answering System in Russian

Sergey A. Belyaev[1], Alexander S. Kuleshov[2], Ivan I. Kholod[3]
Faculty of Computer Science and Technology
Saint Petersburg Electrotechnical University "LETI"
Saint Petersburg, Russia
[1]bserge@bk.ru, [2]yol0317@gmail.com, [3]iiholod@mail.ru

***Abstract* — The question-answering systems were being investigated for several decades, but the majority of researches were carried out in English. The subject of this paper is the knowledge-based question-answering system. The unique mathematical model describes the process of answering when the question is presented in Russian as a natural language. The model is executed by mapping the question to the existing structure of the offered knowledge base. Question mapping in the natural language to any logical form is performed by using the semantic-syntactical analysis under the conditions of a limited annotated semantic corpus in Russian.**

There exist loads of approaches to execute the syntactic, semantic-syntactic, semantic analysis and approaches to create the answer, but these approaches are not fully transferrable into Russian.

The paper describes features of implementation of the semantic-syntactical analysis in Russian with using SRL algorithm, and features of answers creation.

The described rules are based on the ontology offered by A. Grasser and containing 18 categories to determine the category of a question and to retrieve relations from the text.

The paper shows the results of experiments in forming the answers for ifferent subject domains, including technical texts from the journal "Izvestiya SPbGETU «LETI»", books of Turgenev, and results of the user's requests to a search engine.

The directions of further researches in the field, which will increase quality of the model work, and supposed expansion of the available ontology of questions' categories are also described in this paper.

Keywords — question-answering system; knowledge-based question-answering; semantic role labeling; semantic-syntactic analyzer; annotated corpus

I. INTRODUCTION

Recently there is a tendency of growth of information requests to different search engines. The purpose of such requests is to find information about the event, phenomenon, date, place, reason, consequence or object. Most of the requests of this category are defined as a question, using such language constructions as "what is it", "who is it", "how is it".

To solve such problems as formations of an answer to a question, the question-answer systems are used. The main purpose of such system is to provide to the user a short answer in a natural language, instead of the set of snippets on which the user needs to collect independently information and on its basis to form the answer to the question. Among the systems, which solve the problem of this category there are: IBM Watson, START, Exactus, Yandex Object answer and Google Knowledge graph.

The series of tasks and subtasks [1] specific to the question-answer systems was developed and presented by the group of scientists in 2003. The most actual tasks for the present scientific research are given below:

- **Question classes.** The question-answer system should include taxonomy of questions covering as much as possible conceptual categories. In this system questions can be mapped by derived analytical procedure.
- **Question processing.** The system should recognize the semantically equivalent questions presented in different forms by means of different styles of the speech, syntactic dependences and phraseological units.
- **Question context.** All questions and answers have the same context. The context is used also for the explanation of question, elimination of ambiguity or tracking the progress of thought which is carried out through series of questions.
- **Data sources.** Processing of the question and evaluation of the answer largely depends on data source. If there is no answer to the question in the data source, it does not matter how well the processing of the question and formation of the answer are performed since the correct result will not be obtained.
- **Answer extraction.** Formation of the answer depends on many factors: the complexity of the question, the type of the answer, the data sources, the focus and the context of the question. Therefore in a question-answer system it is worth treating extraction of the answer carefully.
- **Answer formulation.** The result of the system operation has to be presented in the most natural form.

978-1-5090-4865-6/17/$31.00 ©2017 IEEE

In some cases, a direct answer is provided; in others – only a simple selection in a small fragment of the text.

- **Multi-lingual.** The ability of the question-answer system to support not only English is a very important task.

The question-answering systems were being investigated for several decades, but the majority of researches were carried out in English. Among the Russian systems that support Russian at the moment there are no products that can compete with foreign systems in quality. First of all, this is due to the lack of the necessary resources in Russian, contributing to the analysis of the question. This stage has a significant effect on the performance of the whole system.

Thus, solving the problem of question processing, it is necessary to obtain information on syntactic and semantic representation of the question. The appropriate information can be obtained by training of classifiers at the suitable marked corpuses. The problem is that the task of determining semantic roles for Russian does not occur, unlike for English where the essential base of the annotated texts is prepared. A result of it is the absence of the full corpus with semantic annotation. The solution of this problem has to be incitement for expansion of the available annotated corpuses. First of all the quality of question-answer systems for the Russian must be improved.

II. THE SOLUTION OF THE PROBLEM

By now, there are two main modern paradigms of the question-answer systems proposed in 60`s of the last century. The **IR-based QA** and the **Knowledge-based QA** [2].

The paradigm of the IR-based QA is based on information extraction relying on the formation of answers as short segments of the text that are retrieved from large amounts of information and available in the form of texts or specialized collections of documents for specific data domains. The sequence of operations in this case is as follows: at first, the probabilistic type of the answer by mean of question processing component the answer types being named entities, such as a person, location or time. Then, the request for sending to a search system is formed. The search system returns the ranged documents divided into set of passages. At the final stage extraction of answers is carried out from the fragments and their subsequent ranging is performed.

The second paradigm, knowledge-based QA, at the initial stage creates semantic-syntactical representation of request. The value of request can be mapping on some query language (for example, SQL) or some version of predicate calculus. The predicates can be described in the form of RDF triples, where the three-tuple is presented by the predicate with two arguments expressing some simple relation. In such cases, the problem of forming the answer to the factoid question is reduced to finding the missing argument in the triplet. Factoid is a question to which there are short answers, confirmed by the obvious facts. For example, the following triplet {subject: Alexander Popov; predicate: year of birth; object: 1859} it can be used for the answer to such questions as "When was Alexander Popov born?" or "Who was born in 1859?".

For the relations representing large groups of subsets usually the extraction of rules depending on the question are manually described. For the extraction of the relation of the year of birth as in the example above it is possible to describe the template that is looking for the question word "when".

Let's carry out a short review of modern question-answer systems.

IBM Watson [2]-[4] – represents a hybrid system for formation the answer using a set of texts and the structured knowledge base for the purpose of extraction of information. The system is represented by 7 levels of question's processing. At the first stage the analysis of the question which includes classification, definition of focus, and identification of the predicate relation and determination of lexical type of the answer is performed. The second stage represents decomposition of the question into simpler parts, for this task algorithms of statistical machine learning are used. It is followed by generation of answer's hypotheses, making a search into data sources which can be structured (for example, web pages) and not structured. At the fourth stage the soft filtering of hypotheses on definitely set properties is made. It allows reducing execution time of the subsequent analysis stages of the question. Then the estimation of validity of hypothesis is carried out, the appeal to different sources of knowledge is made. At the final stage the merge of similar hypotheses in the final answer and ranging of results of merge is carried out. Assessment of reliability of each answer is made for the asked question. The answer position according to its quality estimate is defined.

START [5], [6] – is the first representative of QA system online in a natural language. The system generates the answers to questions previously distributing them into a set of categories, for example, such as: sciences, art, geography, history, culture, help information etc. For the functioning the system uses such techniques as: annotation of natural language, ternary representation, transformation rules, joint accumulation of knowledge. By accumulation of knowledge there is training of system. Expansion of the available knowledge base at the expense of the user information is supported.

Exactus [5], [7] – is the Russian development of a question-answer system, one of few supporting the interface in Russian. The user request to the system initiates its processing and redirection to such search engines as Google and Yandex. In the next step, by means of linguistic tools there occurs annotation of the received documents, at the end of the process of annotation the most relevant documents are provided to the user. The system is based on statistical algorithms of text processing, such as TF*IDF (the term frequency inverse document frequency) and linguistic methods of search hypotheses.

Yandex object answer (in Russian *«Яндекс объектный ответ»*) [8] – at the beginning of 2015 Yandex has started a service form "object answer" which is the element of search result. The object answer represents summary of the request subject. In the knowledge base more than 110 million of named entities with several hundred millions of relations are described. The database includes such entities as famous

personalities, the movies, the music, the cities, the medicines, etc. Formation of the object answer is carried out by means of the semantic graph. It is a model describing the set of objects, their properties and relations between them. The information about the objects is formed from the set of sources. The acquired information undergoes the process of sifting of duplicates and contradicting facts. During the formation of the object answer, such method of machine learning as Matriksnet takes part [9]. The main task of algorithm is to determine which candidate will be presented as the object answer. For this purpose, the process of a comparison of contexts and search result pages is carried out.

Google Knowledge graph [10] – appeared in 2012, long before "Object answer" from Yandex. It was developed for the improvement of the quality of search engine and represents the knowledge base created from the set of sources, such as Freebase, Wikipedia, CIA World Factbook. The base contains more than half a billion objects and about 20 billion relations. As a data model the semantic network is used which allows the phenomena in the world around us described in the performed logical expressions. The relations between objects are fixed in dictionaries and thesauruses.

From the considered question-answer systems, the Russian language supports only Exactus, Yandex Object answer and Google Knowledge graph. In addition, the services from Google and Yandex are only partially capable to answer simple questions related to the well-known facts that belong to a set of named entities. The results of Exactus are presented in the form of short segment of the text containing a potential answer and differ little in their content from the SERPs. In addition, algorithms of statistical text processing which efficiency seems insufficient are put into the operation of Exactus. The two remaining systems the IBM Watson and START focus only on English. In this regard the development of question-answer system in natural Russian is urgent.

For solution of this task we will consider the approaches which are in the basis of modern paradigms of formation of question-answer systems.

IR-based QA. The paradigm offers 3 subsequent phases of the answer formation.

1. *Question processing.* The purpose of this phase is extraction of the various information which is contained in the question. The named entity type, which is determined by the type of the answer, is extracted from the question. It allows to concentrate on search of specific entity in a set of the documents prepared for formation of the answer. The classifier of the questions can be built on the basis of the hand-written rules (using regular expressions), machine learning or combination of both approaches. Then keywords are selected. The request for document retrieval which potentially may contain answers will be carried out with their help. The received request can be formulated in a different way by replacing keywords with synonyms or extending words parameters. Further focus of the question is formed. This represents some part of the question. Replacing this part with the answer, the result can be presented in the form of a full sentence.
2. *Formation of relevant fragments.* The request generated at the previous stage arrives to the input of systems of extraction of information, such as search engines or set of the indexed documents. The resulting documents pass the stage of extraction of the set of the relevant fragments which potentially contain the answer. The paragraph can act as a unit of division of the fragment, but also smaller segments, usually sentences are allowed. Fragments are ranged according to such properties as amount of named entities, the number of keywords, proximity of keywords and the greatest continuous sequence of keywords.
3. *Answer formation.* Two types of answer algorithms extracts can be applied to this task. The extraction of template as the answer and collecting the answer from N-grams. In the first algorithm, the type of the answer extracted is named entity that corresponds to it. For types which have no compliance, for example “action” or “description”, manually regular expressions are formed. In the second algorithm weight N-gram equivalent to the quantity of snippets in which it has met is assigned. Further filtering and collecting answers is carried out until there is the only candidate answer.

Knowledge-based QA. The asked questions to the system are mapped by request to the structured knowledge base. The representation of the sequence of the text in a logical form is carried out by means of the semantic-syntactical analysis. Usually, mapping in a logical form is most often represented in the form of predicate calculus or in the query language as noted earlier. The relational DB or RDF triplets can be a suitable structure for such data. In the simplest case, the task of forming the answer to the question of factoid, is search of missing argument of the triplet. This paradigm suggests creating the hand-written rules of extraction of often encountered relations in the asked questions. Two approaches of questions processing by means of the qualifier are considered.

1. *The supervised learning classifier.* There is a set of training data where each instance is presented in the form of reference question and has certain logical form. Each reference question is the representative of some subset. This set is used by the system for mapping new questions to already existing logical forms. Process of mapping represents comparison of dependence tree of the investigated question and the subsequent application to a logical form.
2. *The semi-supervised learning classifier.* The problem of creation of the knowledge base to cover various forms of questions about the factoid is difficult. In this regard most of methods of mapping questions to some logical forms use the ways of text reductions. Thus, the comparison is performed of predicate and its parameters with unique artifact of some source of knowledge, for example, such as "Wikipedia".

From the available fundamental couple of ontologies, it is advisable to select hybrid system as solution. This solution combines the best qualities of each approach [5]. The general process of work of this system is presented in the figure 1.

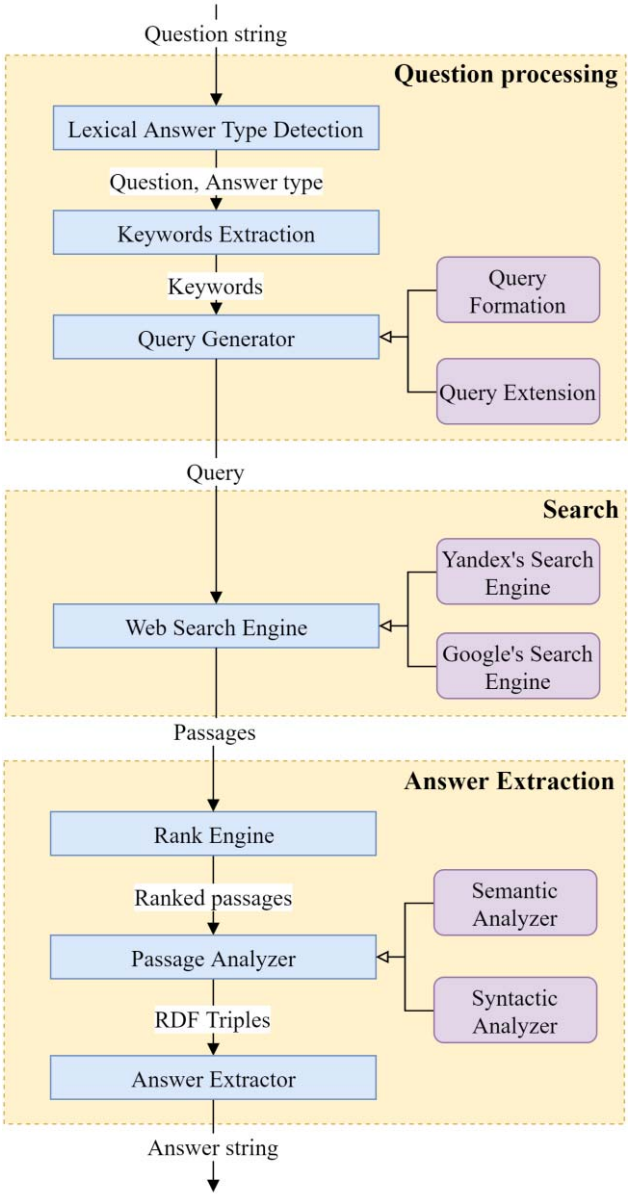

Fig. 1. Architecture of the developed system

Due to the available deficit of the existing knowledge bases in Russian as for specific knowledge domains, and in general on the sphere, processing of question should be borrowed from paradigm of the question-answer system based on extraction of information. At the initial stage the category of question and type of the expected answer, based on named entities represented in the question will be defined. For this process the ontology offered by A. Grasser [11] is used. It describes 18 categories: verification, disjunctive, concept completion, feature specification, quantification, definition, example, comparison, interpretation, causal antecedent, causal consequence, goal orientation, instrumental/procedural, enablement, expectation, judgmental, assertion, request/directive. Further selections of keywords are formed, on which the request to external system then will be created. During sampling, a set of keywords can be extended at the expense of words parameters that will allow to specify search results of target information. The selection of words parameters is carried out on the basis of the revealed named entities of the question.

By the time when selection of keywords is created in request, the second stage begins it is formation of the fragments of the text which are potentially containing information on the asked question. Both the search engines and the local lists of the indexed documents can be the sources providing information. Candidate lists are ranged according to the maintenance of personalized entities, number and distribution density of keywords, as well as other less-relevant properties.

The following and the final stage of the system operation is the formation of the answer. The relevant fragments prepared at the previous stage, go through a process of semantic analysis. The result of which is the predicate calculus presented in RDF-triples. For the increase of the accuracy of identification of predicates, semantic analysis is enriched with pre-stage syntactic analysis. The answer to the question is the contents of the most relevant triplet or set of triplets, contextually related.

It is necessary to pay attention to the next moments concerning restrictions of the developed question-answer system.

- Monolingual. The developed system supports requests and creates answers, only in Russian.
- The semantic representation. As data on semantic component of fragments of the text, information on lemma and part-of-speech of the word is used.
- Problem of complex structured questions. The system is not capable to recognize or give rather qualitative answer when a question has complex structure. The questions containing more than one unknown relation belong to such cases.

III. THE EXPERIMENTAL RESULTS

For the assessment of the quality of formed question-answer system experimental investigations for different data domains have been conducted. For the solution of the problem there were created selections of questions (and the answers corresponding to them) specialized in specific domains.

Questions for this data domain were formed on the base of Turgenev's works. About 100 questions according to such works as “Mumu” (in Russian «Муму»), “Asya” (in Russian «Ася»), “The first love” (in Russian «Первая любовь»), “Fathers and children” (in Russian «Отцы и дети») have been made. Some examples of the created questions are:

- What is the name of Mumu’s owner (in the story "Mumu")?
- How old is Asya (in the story "Asya")?

- What is Bazarov's profession (novel "Fathers and Children")?

The questions belonging to the technical domain were formed to the articles of the journal «Izvestiya SPbGETU «LETI». Some questions asking the system are:

- What is the hybrid cloud (Issue 9, 2015)?
- What is the process of description of verification of dynamic storage (Issue 1, 2016)?
- When the greatest packing density of topology of integrated circuits reached (Issue 7, 2014)?

For this experiment the list of the most popular questions to search engine of Yandex over the last 5 years has been made. Some user questions are:

- What movie to watch?
- When is Maslenitsa (*in Russian Масленица*)?
- How many seasons are there in series "Game of Thrones"?

It is difficult to give specifically one correct answer for the many user requests, in certain cases there are a few such answers at once. And for rhetorical questions there are no of answers, per se at all.

Summary results of experiments for different domains are presented in table 1.

TABLE THE RESULTS OF EXPERIMENTS FOR DIFFERENT DOMAINS

Domain	Art	Technical	The general
In total questions	100	100	100
The answered questions	71	53	86
Correctly answered	34	36	45
Accuracy	0.48	0.68	0.52
Responsiveness	0.34	0.36	0.45

Calculation of the Accuracy and Responsiveness parameters is taken from the paper [5]. "The answered questions" include a number of questions for which the system has given, at least, one answer. "Correctly answered" is the quantity of questions to which the correct answer, with the greatest relevance has been received.

IV. CONCLUSIONS

By the results of the experiment it is difficult not to notice, the low indicator of the row "The Answered Questions" in the technical domain. It is possible to reason it by the fact that technical texts are the completely opposite to the news or art texts, having excessive redundancy of information. In technical prevail concepts and the principles over the facts, events or objects in space and time texts [12]. Nevertheless, accuracy for the technical domain reaches 68% and is the highest indicator, among the available values. It follows due to the lack of redundancy of information in technical texts.

It is necessary to pay attention to the results of the experiments for art and general domains. They have rather identical value of accuracy and high rate in the row "The Answered Questions". It is connected with the fact that required types represent named entities which search algorithm is based on implementation of the system.

In the paper the research of relevance and demand of the question-answer systems for Russian is made, an overview of existing approaches to solving this problem is provided, the process of formation the system is considered. Experimental researches of the quality of the system operation for different domains are conducted. The low number of the received answers to the questions for the technical domain is revealed.

During conducting, the experimental researcher there was a problem of preparation of test questions. At the moment, there are no text tracks with questions in Russian for the general and specialized domains. However, the same situation is characteristic also of other languages. For example, for an English track model of the QA TREC are specialized only for the general domain. Therefore manual selections of questions and answers in Russian for different domains have been created. Selections were suitable for experiments though they also do not show fully reliable results.

During operation of the system the local indexed documents, for example, Turgenev's works or the journal «Izvestiya SPbGETU «LETI» were not used. This task was completely delegated to the search engine, with the specifying criteria on data sources. In respect of the future development of the system, the solution of this task is a necessary condition.

In the plans of the further development of the system there should be considered expanding the semantic representation of the structure of the text. Now information on lemma and part-of-speech of the word is used. Special attention should be paid to analysis of difficult structured questions which contain more than one predicate. As one of possible solutions, decomposition of a difficult question to smaller, simple questions can be used. Work with the technical domain can be subject of a separate scientific research.

REFERENCES

[1] Burger J., Cardie C., Chaudhri V., Gaizauskas R., Harabagiu S., Israel D., Jacquimin C., Lin C., Maiorano S., Miller G., Moldovan D., Ogden B., Prager J., Riloff E., Singhal A., Shirihari R., Strzalkowski T., Voorhees E., Weishedel R. Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A), 2003, 35 p.

[2] Jurafsky D., Martin J. H. Speech and Language Processing, Ch. 28: Question Answering, New Jersey, Alan Apt Publ., 1999, 975 p.

[3] Lapshin V. A. *Informatsionnye protsessy i sistemy. Voprosno-otvetnye sistemy: razvitie i perspektivy, Nauchno-tekhnicheskaya informatsiya*, [Information processes and systems. Question-answer systems: development and perspectives, Scientific and technical information], Moscow, 2012, vol. 2, no. 6, pp. 1-9, (in Russian).

[4] IBM Watson. Available at: http://www.ibm.com/watson/ (accessed 20 November 2016).

[5] Belyaev S. A., Kuleshov A. S. Software products, systems and algorithms, *Formirovanie voprosno-otvetnoy sistemy v usloviyakh ogranichennogo ob"ema semanticheski razmechennogo korpusa,* [Formation of the question-answer system in the conditions of limited scope of semantic markup corps], 2016, no. 4, 7 p., (in Russian).

[6] START. Available at: http://start.csail.mit.edu/ (accessed 16 November 2016).

[7] Exactus. Available at: http://exactus.ru/ (accessed 17 October 2016).

[8] *Kak eto rabotaet? Ob"ektnyy otvet v poiske,* (How it works? The object answer searching). Available at: https://yandex.ru/blog/company/97446 (accessed 6 November 2016).
[9] MatrixNet. Available at: https://yandex.ru/company/technologies/matrixnet (accessed 6 November 2016).
[10] Introducing the Knowledge Graph: things, not strings. Available at: https://googleblog.blogspot.ru/2012/05/introducing-knowledge-graph-things-not.html (accessed 10 November 2016).
[11] Graesser A. C., Person N. K. American Educational Research Journal: Question Asking During Tutoring, Memphis State University, 1994, vol. 31, no. 1, pp. 104-137.
[12] Rinaldi F., Hess M., Dowdall J., Molla D., Schwitter R. Question Answering in Terminology-rich Technical Domains, 2009, 16 p.