# Question-answering system

# Question-answering system

**A A Stupina[1,2], E A Zhukov[2], S N Ezhemanskaya[2], M V Karaseva[1,2], L N Korpacheva[2]**

[1] Siberian State Aerospace University, 31 “Krasnoyarskiy Rabochiy” av., Krasnoyarsk, 660037, Russia

[2] Siberian Federal University, 3 Vuzovskii Lane, Krasnoyarsk, 660025, Russia

E-mail: saa55@rambler.ru

**Abstract**

The paper gives a brief overview of the key problems of designing question-answering systems. It offers the approach to the determination of the question-answering system of giving an answer to the question based on a modified measure of one set to another inclusion. The results of the proposed approach in the problem of automatic selection of the answer to the question of several possible options are presented. These results showed that the system is able to give the correct answer (of 4 possible ones) in more than 50% of cases with a relatively small knowledge base.

## Introduction

The information analysis given in the natural language, along with speech recognition and computer vision is one of the key directions of research in the field of artificial intelligence. Among the problems of natural language processing methods a special problem of development question-answering systems is of a great importance. According to some definitions a question-answering system is an information system that is able to get questions and answer them in natural language. In other words it is a system with a natural language interface. Conventionally, a question-answering system can be divided into highly specialized systems, i.e. those that are applied in any particular field and general question-answering systems with the capability of searching in the similar fields as well.

In the context of the conducted investigation a subclass of question-answering systems as a selection of the correct answer among the given finite set of answers, i.e., automatic test execution is considered.

## Theoretical analysis

The development of the question-answering system is a very complex process that includes the application of a variety of NLP techniques (in English it means Natural Language Processing, i.e. natural language processing) as well as the data mining algorithms such as, for example, deep and recursive neural networks. In this investigation the algorithm based on searching the relevant answers in the knowledge base will be considered.

In general case one can point out 12 key objectives in the field of the QA-systems building developed in 2002 by a group of researchers [4].

1. Questions classification. It is considered as a selection of the question type (selection of one variant from several types; to insert an empty meaning into a phrase (a direct answer to the question, etc.) as well as a class in the field of knowledge; if there was a question, for example in physics or biology. This step is necessary in order to develop a more flexible

answer algorithm to the question because each class of the question may require its own method to find an answer.

2. Stage of processing and analysis. It includes semantic analysis of the question phrases, i.e. it identifies key words, splits a complex question into a few simple patterns, detects synonyms, idioms to recognize similar ones in meaning questions but presented in the different way, etc.

3. The ability of the system to see the context of questions allows it to make clarifying questions or alternatively, to find answers to the questions posed earlier.

4. Base of knowledge generation is one of the key stages in the QA-system development. The performance of the question-answering system in addition to the effectiveness of the text analysis methods depends on the quality of the text database; if it has no answers to the questions, the QA system can't find anything. It is essential the information is accurate and presented in different forms. This helps to ensure that the QA system is more likely to find the answer.

5. Stage of the answer selection and its evaluation. It means the direct information search in the base of knowledge that contains the answer to the question. The type of the required answer depends on the type of the given question.

6. Answer wording. It implies a simple selection of the text block from the base of knowledge and answers grouping from different sources including the reformulation of the text information based on the context of the question.

7. One should pay attention to such a property of the QA-system as the answer to the question in real time regardless of the complexity of the question and volume of the base of knowledge.

8. Support of multilinguistics. It is necessary to take into account the property of the multilingual support while the system development since questions could be asked in different languages taking into account the peculiarities of the language,

9. Interactivity is necessary in the case when either the system correctly understands the question or the answer does not satisfy the user. In this case it is necessary to support dialogue with a user to clarify the information that is necessary for the question and answer.

10. Ability for consideration. The requirements for the derivation of the new knowledge on the basis of the existed belongs to the QA-systems of a higher level and considered to be one of the most promising in the in the field of this investigation.

11. User's profile. Having the information about the user (his vocabulary, speech manner) the system can increase the performance in search of answers based on the individual approach to each user.

12. And, finally, one can distinguish the stage of the combined question-answering system development to use the advantages of certain systems, eliminating the weaknesses through the strengths of other systems.

The investigation considers a simplified QA-system based on the search of answer as close as possible in accordance with the selected metric to the base of knowledge.

## Technology

First of all it is necessary to generate a base of knowledge to develop a QA-system. The base of knowledge can be local (books, scientific papers placed in local repository) and based on web technologies. The access to the base of knowledge through the Internet taking into account modern technologies does not cause any problems and it is preferable from the practical point of view. However, in cases with the large amounts of knowledge and limited Internet traffic one should pay attention to the local method of information location.

In the context of this investigation the basis for the base of knowledge development could be books, scientific papers and the school curriculum tests from different fields of knowledge.

First of all it is necessary to allocate so-called knowledge from a source of potentially useful information in the assembled list. By knowledge here we mean a Text phrase that includes brief and specific information about any fact. Preferably if the sentence has the information about only one fact and this information is unambiguous. This kind of phrase is proposed to generate the following way:

- on the basis of tests; knowledge is a question of the test in combination with a correct answer;
- on the basis of books, papers and other text information; that means the extraction from the text separate, complete declarative sentences having at least one subject and predicate in its structure. Also a paragraph of a few sentences corresponding to the property described above can be grouped into knowledge. It is possible to apply more complex algorithms to select informative and meaningful sentences; that is one of the objects of the further investigation solutions to this problem.

Further from each resulting phrase so-called "stop words" are excluded [2] and lemmatization [2] of words is performed to bring knowledge to the universal form. Then it is necessary to form a "bag of words" on the basis of the available text information [2]. And finally, it is necessary to present each type of knowledge in the form of a binary vector [2] of length equal to the volume of a "bag of words". In this case $j$-th value of the vector is 1 if the $j$-th word from a "bag of words" is found in the knowledge at least once, and 0 otherwise. In other words, each type of knowledge is a variety of unique elements from a "bag of words".

The search for the answer to the test question is proposed to realize using the following algorithm.

1. Let we have some question $Q$ and a limited set of $n_A$ answers $A = \{a_j,\ j = \overline{1, n_A}\}$ to choose the right one.
2. For each question $Q$, $n_A$ phrases are formed: it is a combination of the question and the $j$-th answer from a set $A$ .
3. As in the case with the base of knowledge, each combination of question-answering phrase undergoes the pre-processing; it means the "stop-words" excluding, lemmatization, and representation in the form of binary vector. However, note that each phrase in combination is processed separately. Thus it is possible a binary vector will appear; it consists of only zeros for a question or answer phrase for the case these phrases includes words not found in the previously generated "bag of words". Further the combination $A$ will denote a pair of binary vectors.
4. Now it is necessary to evaluate the truth degree of each pair based on the available base of knowledge among all pairs $\langle Q, a_j \rangle$ for $j = \overline{1, n_Q}$. For this purpose it is necessary to introduce some measure of proximity of the pair to the base of knowledge. In the framework of the given investigation it is proposed to apply the measure modification of inclusion of one set into another; this modification was described by Seminum B.I. to evaluate the degree of the pair truthfulness [1].

Let we have $n_K$ knowledge and $n_A$ variants of answers for each question $Q$. The matching level (a measure of inclusion) for a pair $\langle Q, a_j \rangle$ and some knowledge $K$ is proposed to be present with the formula (1):

$$d(\langle Q, a_j \rangle, K) = [d'(Q,K) + d'(a_j, K)] \cdot (d'(Q,K) > 0) \cdot (d'(a_j, K) > 0), \quad (1)$$

where $d'(Q,K)$ and $d'(a_j, K)$ $K$ is the measure of the inclusion of question $Q$ and answer $a_j$ into the knowledge $K$.

5. The answer which is considered as correct corresponds to the maximum value of a inclusion measure among the whole knowledge from the base of knowledge (2):

$$Correct\_Answer = \arg_j \max_{j,i} d(\langle Q, a_j \rangle, K_i), \; i = \overline{1, n_A}, \; j = \overline{1, n_K} \quad (2)$$

**Experiment**

A problem of the QA system development for answering questions using school curricula tests for the 8th grade of the United States school was selected as an object for testing the proposed algorithm [8]. We have a training sample (2 500) and a test sample (21 298). The samples are a set of "question + 4 choices of answers" where the information about the correct answer for the given question is in the training sample. It is worth noting a very specific feature of the sample test questions that the question phrase is quite noisy, i.e. it contains words that may not be applied to the question. This approach to the generation of the test samples significantly reduces the opportunity to respond to the questions independently without any help of machines as the process of understanding the meaning of the question takes quite a long time. Thereby the higher requirements to the question-answering system are imposed i.e. to present the ability of the system to weed out uninformative components of the question.

The investigation was performed using three approaches to the base of knowledge generation:

1. only a training sample: a set of phrases that means a combination of a question + a correct answer;
2. a set of textbooks from different fields of knowledge (natural science, physics, chemistry, astronomy, etc.) presented on the portal ck12 [5];
3. a combination of the first and second approaches.

For the experiment let's answer all questions at random and we will receive the answers quality that is equal to about 25%. This result is expected.

First let us investigate the algorithm using the base of knowledge based on the training samples, i.e., having only 2500 informative phrases so called a base of knowledge (Table 1):

**Table 1.** The result of answers to the questions by the system using training samples as a knowledge base

| Number of phrases | Run time | Quality of the algorithm |
|---|---|---|
| 2 500 | 4 min | 34.1% |

As you can see in Table 1 the system was able to answer nearly a third of all new tasks having only the information about the correct answers to 2500 questions.

The following investigation stage is the base of knowledge use formed on the basis of textbooks (Table 2). For the beginning we use a textbook that contains the general information but covering a wide range of topics.

**Table 2.** The result of answers to questions by a system using a textbook as a base of knowledge

| Number of phrases | Used books | Run time | Quality of the algorithm |
|---|---|---|---|
| 124 798 | Life Science concepts for middle school [7] | 17 min | 42.3% |

It is clear that only one textbook has greatly improved the quality of the system of answers to the questions. The computational time has significantly increased.

Then a training sample was included into the existing base of knowledge (Table 3):

**Table 3.** The result of answers to questions by a system using a textbook and a training sample as a base of knowledge

| Number of phrases | Used books | Run time | Quality of the algorithm |
|---|---|---|---|
| 127 298 | Life Science concepts for middle school [7] | 22 min | 42.7% |

The inclusion of the training sample into the structure of the base of knowledge has not brought a significant improvement of the algorithm quality. That means that a textbook "Life Science concepts for middle school" contains almost the whole amount of information available in the training sample as well.

Now let's add a bit more information to the base of knowledge; that is let’s include two subject-oriented textbooks: Biology [3] and Earth Science Concepts For Middle School [6] (Table 4):

**Table 4.** The result of answers to the questions by a system using three textbooks and a training sample as a base of knowledge

| Number of phrases | Used books | Run time | Quality of the algorithm |
|---|---|---|---|
| 270 535 | Life Science concepts for middle school [7],<br>Biology[3],<br>Earth Science Concepts For Middle School [6] | 47 min | 53.6% |

Again it can be noted that the use of additional training materials can significantly improve the quality of the algorithm. In this case the system could give answers probably to the subject-oriented questions the information was not included into the general textbook.

It should be noted that the use of the combined approach with only one book gives the results comparable with the use of the search engine Lucene [9] to apply it to Wikipedia [10].

## Results

Despite the fact that the proportion of the correct answers of the QA system exceeded 50% a bit, there is a reason to consider the proposed approach correct and requiring further investigation.

Among the prospects of investigation one can point out:

- including a larger number of the more subject-oriented scientific literature into the base of knowledge;
- application of vectorization methods of computation to optimize the run time speed of the algorithm;
- application of the different methods of literature parsing to identify the more specific and informative knowledge;
- development of the new metrics and the application of other well-known metrics for evaluation the conformity degree of the question-answering phrase to the knowledge.
- application of the cluster analysis algorithms and classification for pre-identification of the field of knowledge to which the system answers the question.

## References

[1] Semkin B I 2009 About the relationship between the average values of two inclusion measures and similarity measures *Bull. Botanicheskogo sada-instituta DVO RAN* Vol. **3** p. 91-101.

[2] Bird S, Klein E, Loper E 2009 Natural Language Processing with Python. O'Reilly Media 497 p.

[3] Brainard J Biology 2016 CK-12 Foundation

[4] Burger J, Issues, Tasks and Program Structures to Roadmap Researchin Question Answering (QA) / C. Cardie, V. Chaudhri, R. Gaizauskas, S. Harabagiu, D. Israel, C. Jacquemin, C-Y LinS. Maiorano, G. Miller, D. Moldovan, B. Ogden, J. Prager, E. Riloff, A. Singhal, R. Shrihari, T. Strzalkowski, E. Voorhees, R. Weishedel. / Q&A Roadmap Paper, 2003.

[5] CK-12. – CK-12 Foundation, 2016 Access: http://www.ck12.org/, free. – A title screen.

[6] Desonie D. Earth Science Concepts For Middle School CK-12 Foundation, 2016.

[7] Harwood J., Wilkin D. Life Science concepts for middle school / J. Harwood, D. Wilkin, CK-12 Foundation, 2015

[8] Kaggle: The Home of Data Science (The Allen AI Science Challenge’s page). – Kaggle inc, 2016 -. -Access: https://www.kaggle.com/c/the-allen-ai-science-challenge/, free. – A title screen.

[9] Lucene: Ultra-fast Search Library and Server. - The Apache Software Foundation, 2016-. – Access: https://lucene.apache.org/core/, free. – A title screen.

[10] Wikipedia: The Free Encyclopedia. – Wikimedia Foundation, 2016 -. – Access: https://en.wikipedia.org/, access. – A title screen.