The European Proceedings of
Social and Behavioural Sciences
EpSBS

www.europeanproceedings.com
e-ISSN: 2357-1330

DOI: 10.15405/epsbs.2020.08.49

# WUT 2020

## 10th International Conference "Word, Utterance, Text: Cognitive, Pragmatic and Cultural Aspects"

# AUTOMATIZATION OF ANSWERS TO QUESTIONS BY MATCHING SYNTACTIC GRAPHS

Aleksander Gashkov (a), Maria Eltsova (b)*
*Corresponding author

(a) Perm State Institute for Culture, Perm National Research Polytechnic University, Perm, Russia
gashkov@dom.raid.ru
(b) Perm National Research Polytechnic University, Perm, Russia maria_eltsova@mail.ru

## *Abstract*

In this paper, we are providing a solution for a relevant interdisciplinary problem of QA-automatization. However, there are very few question-answer systems for the Russian language at the moment. This study develops a model of a question-answer system that allows to give an accurate answer to a specific question while being fast-acting and easily expandable. This paper aims to develop an algorithm that automatically allows to find answers to questions to interrogated units of the text or sentence. The model of a sentence represents a syntactic tree obtained as a result of the automatic processing. The system is rule based. The rules are the probability modifiers of subordination of two words in a tree. The carried out experiments show good robustness and high quality of answers finding more than 90% in case the correct trees of question and answer were built. In case some sentences of text were analyzed incorrectly, the quality depends on percent of wrong trees and can vary from 0 to 90%. The quality of answers discerns highly for different question types. The experiments indicated another problem: a method in question requires the exact match of lexemes in the question and the text. In addition, experiments availed to identify ways for improving the system.



**Keywords:** Question-answer systems, automatization, syntactic graph, natural language processing.

https://doi.org/10.15405/epsbs.2020.08.49
Corresponding Author: Maria Eltsova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

# 1. Introduction

## 1.1. Problem Statement and Research Questions

The actual and rapidly developing field "Natural Language Processing" has formed into an independent field in research on artificial intelligence in the late 60s of the twentieth century. Within this field, several specialisms can be distinguished: the development of machine translation systems, the development of information retrieval systems, the development of question-answer systems, the development of speech communication systems, and many others. It should be noted that until recently, the main attention in the question-answer systems development was paid not so much to the possibility of their practical use as to the opportunity they could be applied to the development of systems and models for translating statements in a natural language into a formal notation and vice versa (Popova, 1990). However, in the current realities of digitalizing all spheres, this specialism has proved to be very relevant, since there is a tendency to both increase a number of search queries and demands on the quality, low latency and speed of search of information systems (Both, Diefenbach, Singh, Shekarpour, Cherix, & Lange, 2016; Dinan, Roller, Shuster, Fan, Auli, & Weston, 2019; Ha & Yaneva, 2019; Hu, Zou, Yu, Wang, & Zhao, 2017; Kunneman, Ferreira, Krahmer, & van den Bosch, 2019; Lapshin, 2012; Solovyev & Peskova, 2010; Vig & Ramea, 2019). At the same time, the number of queries asked in the form of a natural language question is growing. In addition, these systems are increasingly being introduced into everyday life: chatbots in institutions, navigators, voice assistants, information retrieval systems, etc. (Both, 2018; Galitsky, Ilvovsky, & Goncharova, 2019; Gentzkow, Kelly, & Taddy, 2019; Gu, Ling, Ruan, & Liu, 2018; Jurafsky & Martin, 2014; Trusov & Gashkov, 2019).

# 2. Problem Statement

Thus, this study is relevant because it develops a model of a question-answer system that allows to give an accurate answer to a specific question and that is fast-acting and easily expandable. Extensibility is achieved due to the fact that the construction of syntax trees and the search for an answer are two independent subsystems. Each subsystem can be improved or replaced with another independently of the other subsystem.

# 3. Research Questions

An analysis of existing works suggests that at present there are two main modern paradigms of question-answer systems proposed back in the 60s of the last century: 1) a question-answer system based on information extraction (IR-based QA) and 2) a question-answer system based on knowledge (Knowledge-based QA) (Jurafsky & Martin, 2014; Popova, 1990). The main elements of a question-answer system are a question processing block, an answer search block, and an answer generation block (Popova, 1990); within each block there are various procedures partially presented in each system (for example, morphological analysis, syntactic analysis, semantic synthesis, syntactic synthesis, etc.). Some procedures, for example, validation of answers, are presented not in each system (Solovyev & Peskova, 2010).

https://doi.org/10.15405/epsbs.2020.08.49
Corresponding Author: Maria Eltsova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

## 4. Purpose of the Study

This study aims to develop an algorithm that automatically allows to find answers to questions to interrogated units of the text or sentence (under 'interrogated unit' we understand a member of sentence to that they can ask a question according to its meaning), for example, accurately answer the following questions: “Who did something?”, “Where did the action take place?”, “What did you do?” etc. The structure of paper corresponds its aim. Firstly, we represent the model on which basis the question-answer system is built, secondly, we describe the experiment, and then we correlate the experimental results.

## 5. Research Methods

### 5.1. Model

The model of a sentence is a syntactic tree obtained as a result of the automatic processing. The automatic analysis is divided into two parts: 1) morphological analysis and 2) syntactic analysis. The morphological analysis consists of morphological tagging of all sentence tokens. If any token is ambiguous, then different variants are generated for all such tokens with their own tagging.

The system is rule based. The rules are the probability modifiers of subordination of two words in a tree. Each rule can take in consideration more words if needed. There are some samples of the rules for Russian:

1. A noun in the nominative may be a subject or a predicate nominative.
2. A noun in the dative may be an object.
3. A noun without a preposition in the accusative may be an object.
4. A personal or interrogative pronoun in the nominative may be a subject or a a predicate nominative.
5. A possessive pronoun can be an attribute.
6. A possessive pronoun does not agree with a word segregated by a preposition.
7. An indefinite pronoun can be an attribute.
8. The demonstrative pronoun agreed with the noun following it is an attribute.
9. The full form of the adjective is an attribute.
10. The short form of the adjective is a predicate.

The probability modifier is depending on two factors: 1) grammatical attributes of two words (and, possible, their neighbours) and 2) distance between two words in sentence. Before the analysis, it is presumed that any two words can be linked, thus effectively creating full weighted graph of sentence. After applying all rules, the weights of edges are changed. It allows building an optimal spanning tree of a graph resulting in the most probable syntactic tree. The tree is restricted in a number of ways: 1) multiply subordination is prohibited, 2) all words of the sentence must be in the resulting graph (except particles, pronouns and interjections), 3) the root of the tree is predicate (or subject, if there is no predicate in sentence). The restrictions imposed make it possible to use a highly efficient algorithm for constructing a spanning tree with a time complexity of $O(n^2)$, where n is the number of words in a sentence. The algorithm finds the tree with the highest weight. It is impossible to resolve ambiguity at the moment of tree building, so we must preserve ambiguous trees for later analysis. It is impossible to store all trees because their

https://doi.org/10.15405/epsbs.2020.08.49
Corresponding Author: Maria Eltsova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

number is too large, exactly $n^{n-2}$ where n is number of words in sentence. We select the best variants by creating *k* optimal trees with Eppstein algorithm (Ng, Jordan, & Weiss, 2001).

The next step of analysis is the co-reference resolution and adjusting sentences' weight with a context. If sentences contain any homonym, then all variants of combinations are used to calculate weight adjustment. As a result, trees of each sentence can be reordered and a new one takes the place of the best variant. The process could be repeated many times. As a result, we obtain list of *k* trees for each sentence, sorted by score. The score correlates with probability of sentence's tree being correct. The *k* can be any reasonable number.

The resulting syntax trees allow a number of formal operations on sentences:

- Comparison of sentences (with different surface form)
- Partial comparison
- Establishing order.

Comparison of sentences allows establishing the equality between sentences with the same syntax tree but with a different record, for example:

*Мама любит дочку, Дочку мама любит* и т. д.

A partial comparison does not establish the equality between lexemes but their grammatical properties, for example, it allows assigning the sentence to a certain class:

[ predicate [subject [attribute]]] [ predicate [adverbal modifier] [subject]].

Establishing an order allows determining whether one sentence is a part of another one.

The combination of the partial comparison and order gives a formal definition of the answer to the question to the sentence member: if the question in which the question word (or phrase) is replaced with a pseudo-word with the required grammatical characteristics corresponding to the answer is partially ordered with the sentence (partial coincidence with the sentence by the interrogative word and complete with the rest), then this sentence contains the answer at the place where the top of the syntax tree coincides with the question pseudo-word.

The questions are restricted in the following way: 1) a question must be a full sentence. 2) the question is asked of an interrogated unit (logical conclusion is impossible).

The question is analyzed in the same way. The differences are as follows:

- question word or phrase is substituted with a pseudo-word
- the pseudo-word possesses grammatical attributes of a possible answer
- possible answers are selected from a predefined list.

The answering process means the selection of attributes graph pattern of question in graphs of all text sentences (Tong, Faloutsos, Gallagher, & Eliassi-Rad, 2007). The question pseudo-word could be corresponded to many grammar patterns and any their coincidence produces a possible answer. As a result, a list of answer candidates is obtained and sorted.

### 5.2. Experiment

To carry out an experiment, 215 sentences were selected. Sentences satisfy restrictions mentioned above. There were at least 4 questions given to each sentence. The operation of the system is demonstrated

https://doi.org/10.15405/epsbs.2020.08.49
Corresponding Author: Maria Eltsova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

on the example sentence *Мальчик увидел восемь летящих крокодилов* (*A boy saw eight flying crocodiles*).

Possible questions words and phrases are as follows: *what, who, where, where from, where to, when, which, how, why, how much, do what, what for*. The question must be addressed to a particular lexeme of a sentence. For sample sentence we can ask *Кто увидел? Что сделал мальчик?* (*Who saw? What did a boy do?*) etc.

Any questions can be formed for an interrogated unit of the sentence analyzed including interrogative words or phrases: *what, who, where, where, when, which, how, whose, why, how much, what to do / do* (in personal form) and lexemes matching those of the sentence. In the example given, we can ask the following questions: *How many crocodiles did a boy see? Whom did a boy see?* and etc. Figure 01 demonstrates the system's operating results.

```
<sentence>мальчик увидел восемь летящих крокодилов</sentence>
<question text="кто увидел">
    <answer>мальчик</answer>
</question>
<question text="что сделал мальчик">
    <answer>увидел</answer>
</question>
<question text="что увидел мальчик">
    <answer>восемь крокодилов</answer>
</question>
<question text="восемь кого увидел мальчик">
    <answer>восемь крокодилов</answer>
</question>
<question text="сколько крокодилов увидел мальчик">
    <answer>восемь</answer>
</question>
```

**Figure 01.** Question-answer system operating results

As a result of the analysis according to the model described above, the system gives an answer: an interrogated unit of a sentence, an answer to an asked question.

## 6. Findings

The experiments show good robustness and high quality of answers finding more than 90% in case the correct trees of question and answer were built. If a question was analyzed wrongly, then correct answer can be found only randomly. In case some sentences of text were analyzed incorrectly, the quality depends on percent of wrong trees and can vary from 0 to 90%. Consequently, we can presume that the quality of syntactic analysis is the main factor of successful question answering.

The quality discerns highly for different question types. The best results (about 99%) were shown for questions *"who", "what"* and *"which"*. The worst quality (less than 50%) was obtained for the question "*what for*".

The experiment indicated another problem: a method in question requires the exact match of lexemes in the question and the text. There exist a number of methods to overcome this problem. One is to use

synonymous lists, another is to define semantic distance between lexemes. Both methods allow searching for fuzzy matches – synonyms and synonymous phrase.

## 7. Conclusion

The work resulted in the model of question-answer system that precisely answers the question while being fast operating and easily extensible. The algorithm is based on the original model that builds trees of the sentences analyzed and then matches syntactic graphs according to a probabilistic assessment method. Carrying out the experiments, we recognized that the model reveals the high quality of answers finding more than 90% in case if correct trees of question and answer were built. In addition, it has been experimentally proved that the quality of an automatic analysis is a decisive factor in the search for an answer. Moreover, experiments established accuracy for different types of questions: the system demonstrates the least accuracy for questions to the adverbial modifier of purpose. Also, experiments availed to identify ways for improving the system.

## References

Both, A. (2018). Towards Component-based, Domain-specific, Efficient Question Answering Systems. *WWW (Companion Volume),* 1047.

Both, A., Diefenbach, D., Singh, K., Shekarpour, S., Cherix, D., & Lange, C. (2016). Qanary – a methodology for vocabulary-driven open question answering systems. In: Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S.P., Lange, C. (eds.) *ESWC 2016. LNCS, 9678*, (pp.625–641). Springer: Cham. https://doi.org/10.1007/978-3-319-34129-3_38

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., & Weston, J. (2019). Wizard of Wikipedia: Knowledge-powered conversational agents. *In Proceedings of the International Conference on Learning Representations (ICLR).*

Galitsky, B., Ilvovsky, D., & Goncharova, E. (2019). On a Chatbot Providing Virtual Dialogues. *Proceedings of Recent Advances in Natural Language Processing,* 382–387.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature, 57*(3), 535-74. https://doi.org/10.1257/jel.2018102

Gu, J.-C., Ling, Z.-H., Ruan, Y.-P., & Liu, Q. (2018). Building sequential inference models for end-to-end response selection. Retrieved from CoRR, abs/1812.00686

Ha, L. A., & Yaneva, V. (2019). Automatic Question Answering for Medical MCQs: Can It Go Furtherthan Information Retrieval? *Proceedings of Recent Advances in Natural Language Processing,* 418–422.

Hu, S., Zou, L., Yu, J. X., Wang, H., & Zhao, D. (2017). Answering natural language questions by subgraph matching over knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, *30*(5), 824-837. https://doi.org/10.1109/TKDE.2017.2766634

Jurafsky, D., & Martin, J. (2014). *Speech and Language Processing. Chapter 28, Question Answering*. NJ: Pearson Education.

Kunneman, F., Ferreira, T. C., Krahmer, E., & van den Bosch, A. (2019). Question Similarity in Community Question Answering: A Systematic Exploration of Preprocessing Methods and Models. *Proceedings of Recent Advances in Natural Language Processing,* 593–601.

Lapshin, V. A. (2012). Voprosno-otvetnye sistemy: razvitie i perspektivy [QA-systems: actual stand and perspectives]. Nauchno-tehnicheskaja informacija (NTI) [Scientific and Technical Information]. Ser. 2. Informacionnye processy i sistemy [Information Processes and Systems], *6,* 1–9 [In Russian]

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On Spectral Clustering: Analysis and an algorithm . NIPS. Retrieved from https://www.semanticscholar.org/paper/On-Spectral-Clustering%3A-Analysis-and-an-algorithm-Ng-Jordan/c02dfd94b11933093c797c362e2f8f6a3b9b8012

Popova, Je. V. (Ed.). (1990). Iskusstvennyj intellekt [Artificial Intellegence. Sistemy obshhenija i jekspertnye sistemy: Spravochnik [Communication Systems and Expert Systems: A Reference Book]. M.: Radio i svjaz'. [In Russian]

Solovyev, A. A., & Peskova, O. V. (2010). Postroenie voprosno-otvetnoj sistemy dlja russkogo jazyka: modul' analiza voprosov [Creating a QA-system for the Russian language: a Module for Analyzing Questions]. Novye informacionnye tekhnologii v avtomatizirovannykh sistemakh [New Information Technologies in Automated Systems]: Proceedings of 13 Workshop. – Mosk. gos. in-t elektroniki i matematiki [Moscow State Institute for Electronics and Mathematics], 41–49. [In Russian]

Tong, H., Faloutsos, C., Gallagher, B., & Eliassi-Rad, T. (2007). Fast best-effort pattern matching in large attributed graphs. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 737-746). https://doi.org/10.1145/1281192.1281271

Trusov, A., & Gashkov, A. (2019). Iterative Procedural Internet Search. *Journal of Physics: Conference Series,* 1415. Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/1415/1/012019.

Vig, J., & Ramea, K. (2019). Comparison of transfer-learning approaches for response selection in multi-turn conversations. Workshop on DSTC7.