Chatbot for Applicants on University Admission Issues

Liudmila Shchegoleva, Grigorii Burdin
Petrozavodsk State University
Petrozavodsk, Russia
schegoleva@petrsu.ru, wow-lol01@yandex.ru

***Abstract*—The article discusses a prototype of a question-answer system for automating work with applicants during the admission campaign of Petrozavodsk State University. The main approaches to automation are the classification of questions, rules for forming a response template, and a vector model of the search engine for a collection of possible answers. The core of the question-answer system is the methods of natural language processing, including tokenization, lemmatization, morphological and syntactic analysis. The prototype of the question-answer system is implemented in the form of a chatbot for a social network.**

I. Introduction

Every year, school graduates are going to enter higher educational institutions and get acquainted with the conditions for admission to different universities. Graduates choose training areas, specifics of educational programs and curricula, evaluate their capabilities, compare them with the requirements of universities, with the results of applicants from previous years, and so on.

From year to year, the main content of the questions is repeated. The answers may change slightly, as the admission rules and the requirements for the level of knowledge of applicants change. Quantitative data on the ratio of budgetary and paid places, as well as thresholds for entrance test scores, are changing. Also, statistics on various indicators are accumulated.

During the period of submission of documents for admission to universities, the flow of questions increases significantly and the team of employees whose duties include answering incoming questions cannot cope with the volume of work. At the same time, such an employee has to repeat the same information several times during the day, which negatively affects both the result of his work (you can skip something, thinking that this information has already been said), and the emotional fatigue of employees. Such work should be automated.

One of the ways to automate the process of working with applicants is the development of a question-answer system, which, in real time, regardless of the number of questions, conducts a dialogue with each applicant (or with their parents).

The main tasks of the system are:

1. Accept a question from the user in natural language.

2. Perform preprocessing of the request text.

3. Form the answer to the user's question also in natural language.

4. Send a response to the user.

To implement the described functionality, it is necessary to develop a knowledge base and rules for forming responses based on information about the subject area.

A prototype of such a system was developed for the Petrozavodsk State University. Restrictions were defined for the system:

- Questions from users are received only in the form of text.
- Questions from users are received only in Russian.
- The subject area is limited to information related to the work of the admission committee of Petrozavodsk State University for a given year of admission. For another year of admission, it will be necessary to repeat the procedure of preliminary preparation of the knowledge base, as the quantitative data will change, as well as changes in the admission rules, changes in the list of training areas and other information are possible.

Within the subject area, the question-answer system cannot be a separate independent program. Access to it should be open to the general public. The target audience of the program is applicants who may be geographically removed from the location of the university. Therefore, the question-answer system should have a web interface or be presented in social networks, messengers, or other forms of communication. Chatbot is one of the forms of implementation of the question-answer system. Chatbot can be deployed in both social networks and messengers. Some universities and colleges are using chatbots to answer routine questions about enrollment, for example, Georgia State University [1, 2].

Further, the article will describe the principles of building a question-answer system, taking into account the peculiarities of the work of the university admission committee and the rules for admission to the university, the architecture of the system, methods of processing questions and generating answers, used tools, general algorithm, and the results of approbation.

II. METHODS

Several approaches are used to create question-answer systems. The approaches are described in sufficient detail in the reviews [3-6].

The first approach is based on information retrieval algorithms. Possible answers are treated as text collection documents. Questions are queries to a search engine. The answer is selected as the document most relevant to the request. In this approach, all the answers are pre-formulated and stored in the database. The main task is to correct pre-process the collection of answers and the development of the algorithm for determining relevance spine answers the question.

The second approach relies on the use of answer templates. The subject area in which the question-answer system will work is presented in the form of a knowledge base, thesauri, graph models, etc. The answer to the question is generated according to the specified rules.

The third approach uses neural networks that are trained on questions and answers from the subject area and generate an answer based on hidden patterns [7-10].

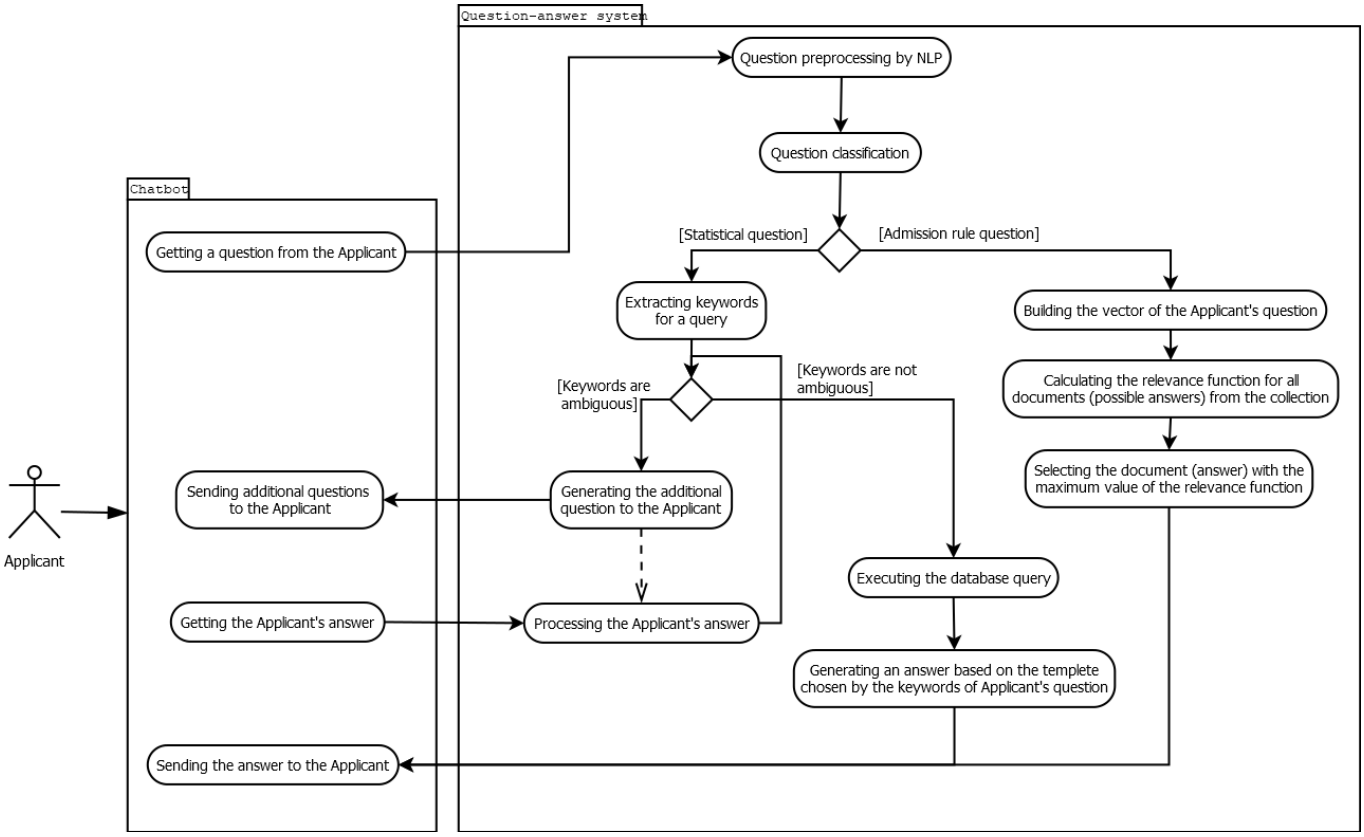

Fig. 1. General algorithm of chatbot and the question-answer system

III. THE QUESTION-ANSWER SYSTEM ARCHITECTURE

The choice of approach for creating a question-answer system should be based on the content of the subject area of questions and answers. An adequate choice of methods for solving the problem is a guarantee of the success of the information system.

Therefore, at the first stage, an analysis of questions received from applicants was carried out. Primary information was obtained in two directions. First of all, the e-mails received by the admissions committee were analyzed. As a result of manual processing of these texts, their classification was carried out to highlight the main topics to which the questions are devoted. The second source of primary information was the staff of the admissions committee. They described in more detail what applicants are most often asked about, what information parents are interested in, what questions are rare, but also important and should be presented in a question-answer system. Information about rare questions is very important. Therefore, interviewing the staff of the admissions committee had provided more information than automatically processing questions received by email.

Thus, several classes of questions were identified and the sources of information for the formation of answers were identified. The questions were divided into two large groups: questions on the rules of admission and the content of training areas and questions related to statistics of admissions in previous years. These two groups of questions are fundamentally different and separate processing approaches were used for them.

To answer questions related to the admissions committee's work procedure, admission rules, the distinctive features of the maintenance training areas, we used an approach based on a search engine with a vector model [11]. To form the index of the search engine, ten regulatory documents were taken ("Rules for admission to bachelor's, specialist's and master's programs at PetrSU", "Accounting for individual achievements of applicants to bachelor's and specialist's programs", etc.). Regulatory documents are unstructured texts containing from 300 to 2000 words (tokens). Each document contains a large amount of information that is suitable for answering various questions. Therefore, each document was divided into small fragments of 10-30 words, which could serve as a complete answer to the question. In this case, the question can be asked with an emphasis on different parts of the answer phrase. Each piece was included in the search engine's document collection as a separate document.

To build the index of the search engine, the collection of documents was processed and the frequency of occurrence of words (tokens) in the documents of the collection was calculated. The TF-IDF weights values were calculated from the frequency statistics of the words. Since the documents were short texts, the frequencies were basically equal to 1 or 2. As a result, the calculated document vectors turned out to be uninformative. After approbation of this index, the results of the question-answer system work turned out to be unsatisfactory.

Since the automated indexing of the documents collection did not give good results, we had to build the index manually. Stop words were completely excluded from the index. Keywords were selected, for which the frequency of occurrence for each document was adjusted taking into account the importance of this word in the document and its difference from other documents.

For the second group of questions related to statistical information about past admissions, a database approach was used. All statistical information was structured according to training areas and statistics indicators. To generate an answer to the question, it is necessary to select the training area, the year of admission and the name of the indicator from the text of the question. Further, a regular query to the database will return the required statistics value, which will be transmitted as an answer.

IV. Algorithm of the question-answer system

The general algorithm of the question-answer system includes the following steps (Fig. 1). The received question is processed by NLP methods to obtain tokens, their lemmas. Then, based on the rules, a binary classification of the question is carried out. If the question is assigned to the category of questions related to the admission rules, then the numerical vector of the question is constructed. Based on the cosine measure of vectors closeness, the values of the relevance function are calculated for the collection documents (possible answers to the question). The documents are ranked in descending order of the relevance function, and the document with the highest value is selected. This document is sent to the user as an answer to the question.

If the question was classified as a statistic, then a numeric value is selected from the database based on keywords of the question. Then the answer to the question is generated based on the template. This answer is sent to the user. If the extracted keywords are not enough to unambiguously determine the numerical value in the database, for example, the name of the training area is formulated inaccurately and is suitable for several training areas, and then a clarifying question is sent to the user.

To develop the system, the following tools were used:

- Python is a universal programming language with a set of modules for working with text;
- NLTK is a package of libraries for symbolic and statistical natural language processing;
- Pymorphy2 is a morphological analyzer of the Russian language;
- Mystem is a morphological analyzer of the Russian language with support for removing morphological ambiguity, developed by Yandex;
- Natasha is a set of tools for automatic processing of text documents, including sentence parsing and extensible grammars;
- SQLite is a relational database management system.

The program is fully implemented in the Python programming language. The user's question is pre-processed using the tools Pymorphy2, Mystem, and Natasha. The search engine index was prepared using the NLTK and Mystem tools. The SQLite database stores the search engine index, collection of answers, and statistical information from the subject area.

The prototype of the question-answer system is implemented in the form of a chatbot for a social network. Currently, chatbots are actively used as assistants in various areas of human activity. And social networks are a successful platform for communication and obtaining the necessary information in the youth environment, which includes university applicants.

V. Conclusion

The subject area of the question-answer system for university applicants is very limited. Partial automation of the process of processing primary documents made it possible to reduce the amount of work in comparison with manual selection of rules and templates for generating answers. Dividing the questions into two categories and choosing the appropriate methods for generating answers gave a good result. Testing the system on questions received by e-mail showed a precision of 60%. Currently, research is being conducted to increase this metric.

References

[1] Oracle Education and Research. Artificial Intelligence and Chatbots in Higher Education, Web: https://www.oracle.com/us/industries/education-and-research/he-artificial-intelligence-br-5180773.pdf?lx=3FBMkx&f=y&topic=Finance

[2] W.E.Hefny, Y.Mansy, M.Abdallah, S.Abdennadher, "Jooka: A Bilingual Chatbot For University Admission", unpublished.
[3] B.Ojokoh, and E.Adebisi, "A Review of Question Answering Systems", *Journal of Web Engineering*, Vol. 17, Iss. 8, 2018, pp. 717–758. Doi: 10.13052/jwe1540-9589.1785.
[4] D.Diefenbach, V.Lopez, K.Singh, et al., "Core techniques of question answering systems over knowledge bases: a survey", *Knowl Inf Syst* 55, 2018, pp. 529–569. Doi: https://doi.org/10.1007/s10115-017-1100-y.
[5] A.Bouziane, D.Bouchiha, N.Doumi, M.Malki, "Question Answering Systems: Survey and Trends", *Procedia Computer Science*, Vol. 73, 2015, pp. 366-375. Doi: https://doi.org/10.1016/j.procs.2015.12.005.
[6] A.Andrenucci, and E.Sneiders, "Automated question answering: review of the main approaches", *in Proc. Third International Conference on Information Technology and Applications (ICITA'05)*, 2005, pp. 514-519, vol.1. Doi: 10.1109/ICITA.2005.78
[7] V.Yadav, V.Bharadwaj, A.Bhatt, A.Rawal, "Question–Answer System on Episodic Data Using Recurrent Neural Networks (RNN)", *in Proc. 3rd International Conference on Data Management, Analytics and Innovation (ICDMAI 2019)*, 2019, pp. 555-568.
[8] Y.Sharmaa, S.Guptaa, "Deep Learning Approaches for Question Answering System", *in Proc. International Conference on Computational Intelligence and Data Science (ICCIDS 2018)*, 2018, pp. 785–794.
[9] D.Parygin, N.Matyushin, A.Finogeev, N.Sadovnikova, T.Petrova, E.Fadeeva, "Neural Network Processing of Natural Russian Language for Building Intelligent Dialogue Systems", *Electronic Governance and Open Society: Challenges in Eurasia. EGOSE 2020. Communications in Computer and Information Science*, Vol 1349. Springer, Cham. Doi: https://doi.org/10.1007/978-3-030-67238-6_17
[10] O.Makhnytkina, A.Matveev, A.Svischev, P.Korobova, D.Zubok, N.Mamaev, and A.Tchirkovskii, "Conversational question generation in Russian", in Proc. Conference of Open Innovation Association, FRUCT, 2020, pp. 126-133. doi:10.23919/FRUCT49677.2020.9211056
[11] R.A.Baeza-Yates and B.Ribeiro-Neto, *Modern Information Retrieval*. USA: Addison-Wesley Longman Publishing Co., Inc., 1999