http://doi.org/10.35784/iapgos.5628 *received: 30.11.2023 | revised: 13.02.2024 | accepted: 12.03.2024 | available online: 31.03.2024*

FORMATION OF HIGHLY SPECIALIZED CHATBOTS FOR ADVANCED SEARCH

Andrii Yarovyi, Dmytro Kudriavtsev
Vinnytsia National Technical University, Department for Computer Science, Vinnytsia, Ukraine

Abstract. *In this research, the formation of highly specialized chatbots was presented. The influence of multi-threading subject areas search was noted. The use of related subject areas in chatbot text analysing was defined. The advantages of using multiple related subject areas are noted using the example of an intelligent chatbot.*

Keywords: text-processing, intelligent data analysis, chatbot, advanced search

TWORZENIE WYSOCE WYSPECJALIZOWANYCH CHATBOTÓW DO ZAAWANSOWANEGO WYSZUKIWANIA

Streszczenie. *W tym badaniu przedstawiono tworzenie wysoce wyspecjalizowanych chatbotów. Zwrócono uwagę na wpływ wielowątkowego wyszukiwania obszarów tematycznych. Zdefiniowano wykorzystanie powiązanych obszarów tematycznych w analizie tekstu chatbota. Na przykładzie inteligentnego chatbota odnotowano zalety korzystania z wielu powiązanych obszarów tematycznych.*

Słowa kluczowe: przetwarzanie tekstu, inteligentna analiza danych, chatbot, zaawansowane wyszukiwanie

Introduction

In recent years, chatbots have gained considerable popularity among software tools using artificial intelligence technologies. Among the possibilities of their application, the most popular is the directory function, which consists in finding information upon request and forming the most correct answer. At the same time, the source of information can be both static and dynamic, according to needs and tasks [16]. During previous research in recent years, the main attention was paid to the multi-subject areas of the source, and the possibility of using the Internet as the largest source of information for searching for relevant information for the chatbot user [18]. Considering the spread of popular chatbots ChatGPT from OpenAI and Bard from Google, the research direction confirms the relevance of all previous efforts, and further development is possible due to increasing speed from a technical point of view and improving algorithms for searching and filtering data from a subject point of view [15]. Using the results of recent research on the use of machine learning algorithms and deep recurrent neural networks, it is worth noting the possibility of arbitrary horizontal scaling of the number of subject areas [6, 18]. But despite this, it is the need for expert knowledge that plays a key role in further research in the field of application of intelligent information technologies in chatbots, thanks to the levelling of the amount and volume of data operated by such chatbots as Bard and ChatGPT [5]. Applying machine learning technologies and deep neural networks as artificial intelligence technologies, it is possible to achieve a significant advantage in narrow-profile subject areas of technical, scientific, or commercial orientation. Using search engine databases as data sources, significant progress has been made, based on the current use of chatbots such as ChatGPT, Bard [2, 14]. But with an increase in the horizontal scaling of subject areas, the possibilities for vertical scaling are significantly reduced, which leads to a superficial level of information search and, accordingly, the results of information analysis and filtering [16]. Based on this statement, there is a need to research the use of a limited number of subject areas with the use of similar machine learning methods and artificial intelligence technologies to determine the feasibility of limiting the chatbot data source to achieve a more in-depth analysis of information and provide more correct search information. During further research, it is necessary to investigate the impact of the general application of intelligent search algorithms and data filtering of all knowledge bases available during the research in comparison with the selective application to a limited number of chatbot knowledge bases. As a result of the research, the qualitative impact of limiting the amount of information on the quality of its assimilation will be revealed, and the priority directions of further research in the field of commercialization and expert use of chatbots in solving industry tasks of a search nature will be determined.

1. Using related subject areas

During the research of the influence of the number of subject areas on the quality of information analysis when using the same software and technological capabilities, was found that the number of subject areas used during information analysis significantly depends on the size of the data source, as well as the distribution of data between subject areas [17]. For an ideal case, it is necessary that there is an even distribution of information between all subject areas included in the knowledge base of the chatbot. Thus, adding a new subject area to the chatbot knowledge base will require the addition of a significant amount of expert knowledge to perform an even distribution of data in the knowledge base. It is also worth noting that during past research it was noted that related subject areas form more relationships when using the dictionary data structure [18]. In this regard, an assumption was made that increasing the number of related subject areas for the chatbot knowledge base increases the quality component of the advanced search for relevant information due to a larger amount of data for filtering. To confirm this assumption, it was necessary to find the necessary data and configure the knowledge base of the chatbot by applying machine learning methods and recurrent neural networks.

2. Experiments

Data sets of the Kaggle platform were used as a data source. The total number of subject areas that were used for testing was 20, among which there were related subject areas and unrelated subjects, with uniform distribution of data, as well as uneven distribution during different experiments [11]. For the training sample, 60% of the entire available volume of data, which amounted to more than 10 million terms, was used. This capacity was selected as maximum that was found from Kaggle source that will be acceptable to use in the research. A total of 372 experiments were performed in which the total number of subject areas, the number of related subject areas, and the distribution of data between subject areas were varied. The tests based on average text length, number of subject areas, and using related and non-related subject areas in test groups. In average 28–34 tests per group were using all five different number of subject areas (4, 6, 10, 14, 20) and each test has his own pair with using non-related subject areas. Totally was created 12 test groups. Portion of results of tests are presented on the figures 1–5. The main characteristics that were operated by tests are text analysis

accuracy, velocity, and data distribution coefficient. Based on these result parameters can be defined influence comparison between using related and non-related subject areas. Also, from these experiments will be checked influence of number of subject areas on the accuracy of text analysis accuracy. According to the first two figures, was found that the influence of the uniformity of data distribution between the subject areas of the chatbot knowledge base decreases.

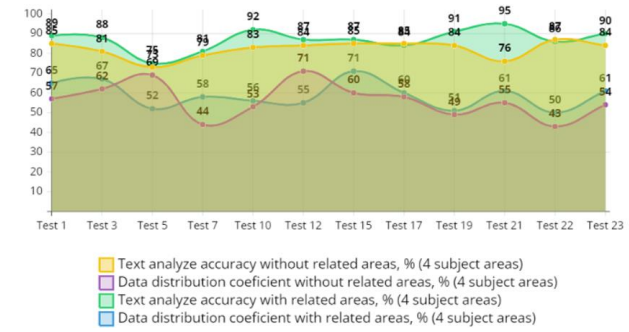


Fig. 1. Accuracy and data distribution comparison for 4 subject areas

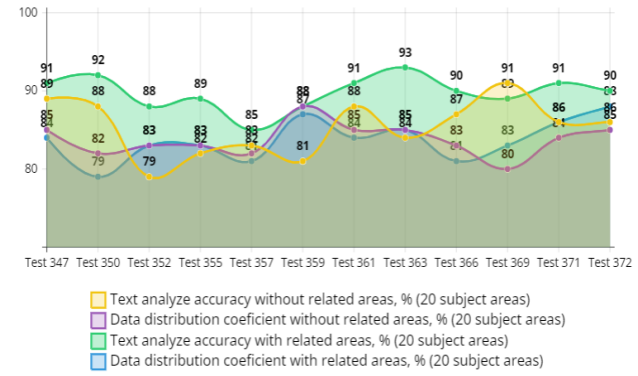


Fig. 2. Accuracy and data distribution comparison for 20 subject areas

Figure 3 corresponds to the higher average accuracy of text analyzing by keywords set up to 1–7 percent if related subject areas were used in experiments. This based mostly on bigger union kernel of terms, related to both subject areas. From technical aspects of this comparison was used improved semantic text analysis with programming module, created in previous research [18].

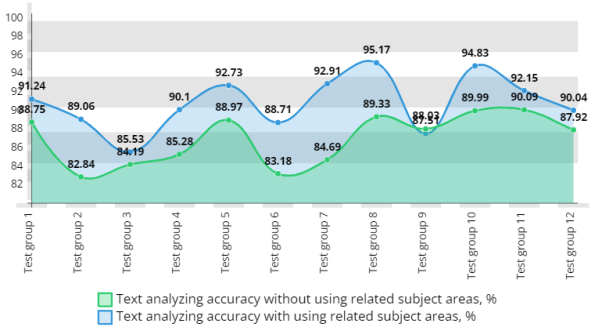


Fig. 3. Text analyzing accuracy comparison between using and non-using related subject areas

As a result of the testing, the adequacy of the test results increases (figure 3), and the speed of processing input information does not change significantly (figures 4, 5) when using most related subject areas. In each test group was selected 28–34 tests. Each test consists of 10–15 text messages from different topics, including subject areas content and context that are not related to any of existing subject area. Test group is balanced by data preparation process which includes precheck of well-known terms from subject areas in text messages from 10 up to 65 percent from whole list of words in messages. In this regard, it is expedient to state that the narrowly focused knowledge base of the chatbot increases the quality of providing information to the user, without reducing the speed of information analysis by the chatbot. As a use of these statements in further research, it is worth examining the field of commercialization, namely the formation of subject groups and criteria that will allow the use of highly specialized chatbots for the analysis of financial transactions, news, natural and man-made phenomena, etc. At the same time, the structure of the chatbot knowledge base, its organization and means of analysis and filtering of chatbot data must be clearly defined to obtain the best result.

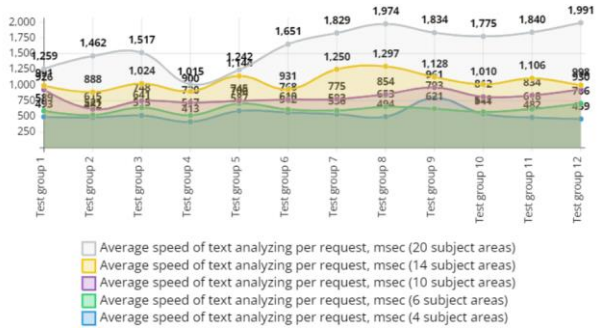


Fig. 4. Velocity of text analyzing per request with using related subject areas

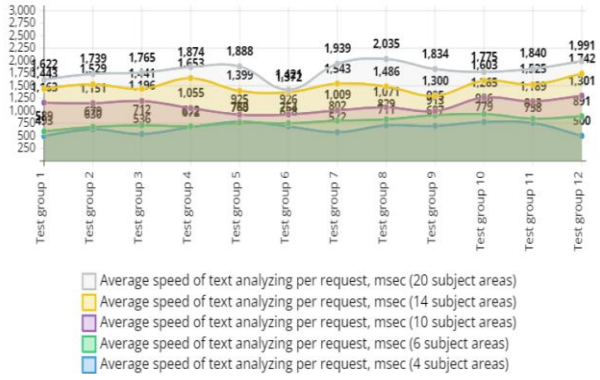


Fig. 5. Velocity of text analyzing per request without using related subject areas

3. Client-server chatbot architecture

For the actual application of a chatbot, it is necessary to develop a software architecture and use tools for its software implementation. Observing modern chatbots such as ChatGPT or Bard, a web interface using a client-server architecture is the best solution, since the key feature of use is the possibility of cross-platform support of the web interface with most modern devices and data protocols, as well as the remoteness of operational capabilities from the user interface with cloud services [10].

In addition, during previous research, the client-server architecture fully met all the requirements, namely the minimization of data transport time, the possibility of setting additional parameters, as well as the time of initial setup of the infrastructure for performing experiments [16]. In this regard, the main attention from a technological point of view is focused on the server architecture of the intelligent chatbot. Considering the latest research in the field of intelligent chatbots, let's consider the server architecture in more detail. It was based on an example from a previous research experiment with the use of additional metrics to determine the level of relatedness of subject areas, as well as an extended analysis of the distribution of data in the general and selective states [17].

Server implementation can be diverse in terms of system load, number of users, as well as the list of available hardware and software solutions. Since the key role of this article is to compare the impact of the quality of user query analysis when using related and unrelated subject areas, as well as when using data from narrow-profile subject areas, the focus is on data

and tools for their analysis. Accordingly, the server architecture is shown in figure 6.

Web-client will be separated project that can be scalable via using RabbitMQ or another message broker technology, Service bus is a hub of requests for chatbot which operate in parallel between user sessions. It responds to the destruction of client message to terms and then groups them into blocks with 3–7 terms in each block. By this operation, messages can be sent to all subject areas databases and results will have better accuracy to pass threshold level [17]. All services will be connected to the Service bus, including operational modules, database repository, logging service as presented in figure 6.

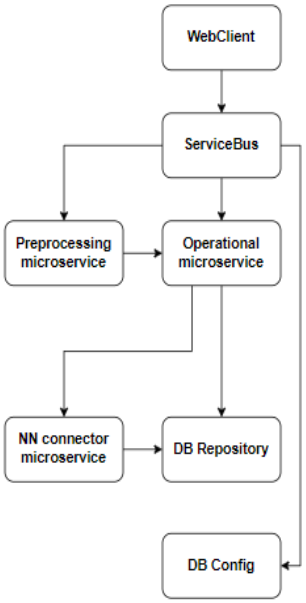


Fig. 6. Chatbot client-server architecture

As a server, ASP.NET Core Web API will be used, which has all the functionality to interact with all system components, including databases and neural network API interface. For client-server architecture will be used RESTful API.

For convenient observation of the results of the experiments, a metrics viewer program was created, which shows chronological graphs of all flows, intermediate calculations, and basic records of usage history, which can be exported to a text file or an Excel data table.

An N-number of input streams corresponding to the number of subject areas is allocated to the main-stream of user requests, and the adjacency metric of the streams is also calculated, and coefficients of affinity of the data from the user request and the streams are formed, forming a matrix of input signals with the dimension, where the additional line is the coefficients of the user's input request. Similar calculations were found in sphere of security for cryptosystems, which also use adjacency metrics for preventing side channels attacks [9]. Subsequently, this matrix is sent to a recurrent neural network, which returns a vector of values for each of the streams, which determines the level of contiguity of the subject area with the user's request [4]. As a result of the search in each of the subject areas, the streams form a list of the most relevant data in the form of a list of terms with a coefficient of adjacency to the user's request. After that, the multiplication of the term matrix of all subject areas with the vector of adjacency coefficients of the subject areas with the user's request is performed. After this operation, it remains only to screen out the terms with final adjacency coefficients to the user's request, by entering the adequacy threshold coefficient, which is described in more detail in the previous article. The final stage is the preparation of the result to a presentable appearance through the use of auxiliary syntactic structures and stylistic analysis of the user's message. In this way, it is advisable to form an improved semantic core that strengthens the context and makes.

In accordance with the chosen client-server architecture, the programming languages C# 7.0 were chosen for the server part, Typescript for the client part, which was implemented on the Angular 16 framework, YAML for combining and configuring services with each other thanks to Docker and Docker Compose, Python 3.10 for working with neural network and neural network learning [13]. Regarding technologies, only a basic list of technologies used in this client-server architecture is given: ASP.NET Core Web API, Angular 16, WPF, Entity Framework Core, SignalR, TensorFlow [7], Docker [12], MongoDB, Azure Pipelines. Visual Studio Code and PyCharm [1, 3, 8] were used as platforms for implementation.

4. Conclusion

After the program implementation, the experimental part of the solution was performed, the results of which are shown in figures 1–5, which revealed the positive impact of the use of related subject areas. Given the small amount of test data compared to the amount of data of search networks, it is difficult to clearly determine the influence of the affinity of subject areas, but more than 81% of all experiments indicate an increase in the accuracy of the response to a user's request in experiments using related subject areas by an average of 10–15% for most terms that passed the threshold than in the experiments where most of the subject areas were unrelated, as shown in comparison between table 1 (without using related subject areas) and table 2 (with using related subject areas).

Table 1. Formalized presentation of experimental results without using related subject areas

N_{SA}	4	6	10	14	20
N_{RSA}	1	2	2	3	7
N_{TE}	10	23	56	71	26
N_{TR}	1	3	5	7	10
C_{DDSA}	89%	24%	56%	72%	85%
N_{ATE}	23.39	24.16	17.328	16.24	18.25
T	0.427	0.529	0.831	0.828	0.991
C_{AA}	0.632	0.524	0.473	0.307	0.420

Table 2. Formalized presentation of experimental results with using related subject areas

N_{SA}	4	6	10	14	20
N_{RSA}	3	5	6	7	7
N_{TE}	10	23	56	71	26
N_{TR}	1	3	5	7	10
C_{DDSA}	63%	30%	53%	74%	85%
N_{ATE}	21.375	24.8	17.136	16.1	17.99
T	0.483	0.571	0.701	0.943	1.034
C_{AA}	0.794	0.678	0.572	0.415	0.515

Where columns are: N_{SA} – number of subject areas that used in experiments; N_{RSA} – number of related subject areas; N_{TE} – number of tests; N_{TR} – number of terms that used in tests; C_{DDSA} – coefficient of data distribution subject areas, which calculates as

$$AVG\left(\frac{ABS(K-N)}{K}\right) \tag{1}$$

where K – is average number of terms that used for one subject area in test, N – is number of terms that used for specific subject area; N_{ATE} – average terms in query; T – average processing time per user's request, sec; C_{AA} – avg. coefficient value of adjacency of terms which calculated as division of number of terms for primary subject area to all terms that found in user request.

Considering the results of the experiments, it is advisable to draw a conclusion about the relevance of the application of related subject areas for highly specialized industries that require the advanced search for specialized information in connection with the improvement of the quality of information search by a chatbot. As for the field of application, everything depends on the needs of the user, and the size of the knowledge

base can be arbitrary, because the main requirement is relatedness of the subject areas.

The most important and troublesome task in this case is the selection of expert data for the formation of the knowledge base of the chatbot.

For further research, it would be advisable to increase the scope of the subject areas, as well as the amount of data for analysis and training. The main sources of data used during a series of experiments were Kaggle data sets, which are quite popular among research in the field of machine learning and artificial intelligence technologies, but in the future, it is necessary to form data samples using search networks and their power, due to the lack of data limit in similar systems. One of the growth direction perspectives of the research could be cooperation with the ChatGPT development team from the OpenAI company.

During the research, an increase in the level of analysis of the user's request was noted when applying related subject areas in the chatbot knowledge base. At the same time, the dependence of the total volume of data in the knowledge base on the number of related subject areas has not been established, as well as a significant dependence on the speed of processing the user's request when the total volume of data increases, which confirms the assumption about the priority of data relatedness over uniform distribution. In the course of further research, more extensive testing will be conducted based on the use of formatted data from search networks.

References

[1] Agarwal S., Rahul P., Neetu: New Text Detection Technique Using Machine Learning Architecture. Dwivedi S., Singh S., Tiwari M., Shrivastava A. (eds): Flexible Electronics for Electric Vehicles. Lecture Notes in Electrical Engineering 863. Springer, Singapore, 2023 [https://doi.org/10.1007/978-981-19-0588-9_1].

[2] Arkoudas K.: ChatGPT is no Stochastic Parrot. But it also Claims that 1 is Greater than 1. Philos. Technol. 36(54), 2023 [https://doi.org/10.1007/s13347-023-00619-6A].

[3] Cao Y., Xu G., Gao Y., Song C.: Application of natural language processing technology based on TensorFlow framework in text mining and discovery algorithm. IET Communications 17, 2022 [https://doi.org/10.1049/cmu2.12534].

[4] Chen W. et al.: Improved Recurrent Neural Networks for Text Classification and Dynamic Sylvester Equation Solving. Neural Process Lett 55, 2023, 8755–8784 [https://doi.org/10.1007/s11063-023-11176-6Raj].

[5] Greco C. M., Tagarelli A.: Bringing order into the realm of Transformer-based language models for artificial intelligence and law. Artif Intell Law 2023) [https://doi.org/10.1007/s10506-023-09374-7].

[6] Henrickson L., Meroño-Peñuela A.: Prompting meaning: a hermeneutic approach to optimising prompt engineering with ChatGPT. AI & Soc 2023 [https://doi.org/10.1007/s00146-023-01752-8].

[7] Joseph J. F. J., Nonsiri S., Monsakul A.: Keras and TensorFlow: A Hands-On Experience. Prakash K. B., Kannan R., Alexander S., Kanagachidambaresan G. R. (eds): Advanced Deep Learning for Engineers and Scientists. EAI/Springer Innovations in Communication and Computing. Springer, Cham. 2021 [https://doi.org/10.1007/978-3-030-66519-7_4].

[8] Karchi R. P., Hatture S. M., Tushar T. S., Prathibha B. N.: AI-Enabled Sustainable Development: An Intelligent Interactive Quotes Chatbot System Utilizing IoT and ML. Whig P., Silva N., Elngar A. A., Aneja N., Sharma P. (eds): Sustainable Development through Machine Learning, AI and IoT. ICSD 2023. Communications in Computer and Information Science 1939. Springer, Cham. [https://doi.org/10.1007/978-3-031-47055-4_17].

[9] Kvyetnyy R., Ivanchuk Y., Yarovyi A., Horobets Y.: Algorithm for Increasing the Stability Level of Cryptosystems. Selected Papers of the VIII International Scientific Conference “Information Technology and Implementation" – IT&I-2021, 293–301.

[10] Meyer J. G. et al.: ChatGPT and large language models in academia: opportunities and challenges. BioData Mining 16(20), 2023 [https://doi.org/10.1186/s13040-023-00339-9].

[11] Mondal B.: Best 25 Datasets for NLP Projects. Kaggle [https://www.kaggle.com/discussions/general/150720] (avaible 13.05.2020).

[12] Pallis George, Trihinas D., Tryfonos A., Dikaiakos M.: DevOps as a Service: Pushing the Boundaries of Microservice Adoption. IEEE Internet Computing 22, 2018, 65–71 [https://doi.org/10.1109/MIC.2018.032501519].

[13] Raj A., Jasmine K.: Building Microservices with Docker Compose. The International Journal of Analytical and Experimental Modal Analysis XIII, 2021, 1215.

[14] Siad S. M.: The Promise and Perils of Google's Bard for Scientific Research. AI. 2023 [https://doi.org/10.17613/yb4n-mc79].

[15] Thapa S., Adhikari S.: ChatGPT, Bard, and Large Language Models for Biomedical Research: Opportunities and Pitfalls. Ann Biomed Eng 51, 2023, 2647–2651 [https://doi.org/10.1007/s10439-023-03284-0].

[16] Yarovyi A. et al.: Information technology in creating intelligent chatbots. Proc. SPIE 11176, 2019, 1117627 [https://doi.org/10.1117/12.2537415].

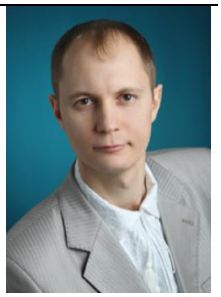
[17] Yarovyi A., Kudriavtsev D.: Dictionary data structure for a text analysis task using cross-references. IEEE 17th International Conference on Computer Sciences and Information Technologies – CSIT, 2022, 61–64 [https://doi.org/10.1109/CSIT56902.2022.10000460].

[18] Yarovyi A., Kudriavtsev D.: Method of multi-purpose text analysis based on a combination of knowledge bases for intelligent chatbot. CEUR Workshop Proceedings 2870, 2021, 1238–1248.

Prof. Andrii Yarovyi
e-mail: a.yarovyy@vntu.edu.ua

Head of Department for Computer Science of Vinnytsia National Technical University (Ukraine). Author of more than 100 technical articles (29 Scopus indexed articles), 5 monographs, 2 patents.
Scientific research related to computer science, intelligent information technologies, image processing, parallel computing.

https://orcid.org/0000-0002-6668-2425



M.Sc. Dmytro Kudriavtsev
e-mail: dmytro_k@vntu.edu.ua

Assistant at Computer Science Department from 2021. Author of more than 20 scientific publications, 7 Scopus indexed articles.
Scientific research related to artificial intelligence, chatbots, object-oriented programming, software development.

https://orcid.org/0000-0001-7116-7869