Potential of Bots for Encyclopedia

Saracevic, Mirhet; Ebner, Markus; and Ebner, Martin

Abstract: ***The wide range of applications and the capability to process and understand natural languages made chatbots very popular. Besides that, many applications chatbots are also used as information retrieval tools. Chatbots are changing the way users search for information. This document focuses on a chatbot that is used as an information retrieval tool. The chatbot enables information search in natural language in a geography domain. In case of a large number of search results, the chatbot engages users with clarification questions. It also provides support to users when uploading multimedia content to the website.***

Index Terms: ***chatbots, information retrieval tool, information search***

1. *INTRODUCTION*

Austria Forum is an online encyclopedia, an online collection, that provides Austria related information. The content of Austria Forum is divided in several categories and written in English. The category of interest for this work is the geography category. It offers information about all countries of the world. Each country page includes general information and links to category pages. Each category page stores data presented in form of text, tables or pictures. The "Community Contributions" category provides forms for uploading interesting pictures, video and audio clips.

It is known that online encyclopedias provide a large amount of information. The information search on an online encyclopedia can be illustrated in two scenarios. The first one is to navigate through the website using links between individual pages. The second, more common way, is to use the search engine integrated in the website. The search engines are mainly based on keyword matching algorithms and provide a list of results when supplied with an input. In order to find desired information, the user needs to browse the list. No matter which of the ways is chosen, the information search is time consuming. The relevance of the information retrieved is questionable. As stated in [1], finding relevant information has always been an issue since the first search engines were built.

This publication is about design, architecture and development of the chatbot prototype for Austria Forum that can be used as an information retrieval tool. The purpose of the chatbot is, on the one hand, to enable information search in natural language and, on the other, to guide users when uploading content to the website. Using natural language, users are able to express their needs better and more accurately; and using natural language processing and understanding, the chatbot is able to understand what users are searching for. Users participate actively in information search providing answers to clarification questions and thus contribute to the relevance of search results. The chatbot requires additional information in cases of ambiguous questions or a large number of search results. In terms of upload of content, the chatbot engages users with a finite number of questions in order to gather needed information.

The main focus of this work will be on improving the relevance of the search results and a faster access to information.

2. *ABOUT CHATBOTS*

Chatbots are software programs that enable users to chat, communicate, and interact with them in natural languages. They are also called dialog-based systems, virtual assistants, conversational agents or machine conversation systems depending on the area of deployment. [2]

Chatbots were primarily built to amuse and entertain the users. ELIZA was the first conversational system that was developed by Joseph Weizenbaum in 1966. The system was programmed with scripts and based on pattern matching algorithms. The goal of ELIZA was to imitate human conversation. Later in 1972 the psychiatrist Kenneth Colby developed a chatbot called PARRY, which used to simulate a paranoid individual. In 1995 a chatbot named ALICE was developed and used to entertain users. ALICE is an open source software and can be used to build customized bots. [2], [3], [4], [5]

Many chatbots developed in the last 50 years were inspired by ELIZA. Though ELIZA was limited and did not provide real understanding,

Manuscript received May 4, 2004. (specify the date on which you submitted your paper for review.) This work was supported in part by the U.S. Department of Defense under Grant DR456 T. C. Author is with the Electrical Engineering Department, University of Tsukuba, Japan (e-mail: author@nxdfrim.go.jp).

users still wanted to communicate. It gave them the feeling that they were talking to a real person. This was the reason for the increased development of chatbots. In the 90s the Leobner Prize, a contest where chatbots performed the Turing Test, was founded. The Turing Test is a method proposed by Alan M. Turing in 1950 for measuring intelligence of a computer system. The Leobner Prize had a big impact on developing chatbots and the Turing Test has increased the interest in the area of artificial intelligence (AI). [6], [7]

In 2015 the attraction to develop chatbots intensified because the use of messaging applications surpassed the use of social networks. Big companies launched platforms for bot development and integration. An increased amount of data on the internet and improvements in data processing and machine learning enhanced artificial intelligence. It is possible to develop more complex chatbots that can execute multiple tasks. Additionally, the range of chatbot applications has widened. Nowadays chatbots are used for customer service, marketing, finance, human resources, e-commerce or entertainment. [8]

The type of a chatbot depends on various parameters. The categorization can be performed based on knowledge domain, service, goal or methods for input processing and response generation [6]. The conversation length is also a possible parameter to perform the classification of chatbots. Some chatbots tend to engage users and have long conversations. These chatbots are able to recall earlier conversations and determine the context of the current conversation. Short conversations are characteristic for chatbots which provide some sort of information when supplied with a question.

The information retrieval is also a field where chatbots find their application. The chatbots to question answering (QA) systems and the chatbots to frequently asked questions (FAQs) attract attention. QA systems tend to provide answers when receiving a query in contrast to search engines that deliver search results in form of lists [3], [9]. As stated in [9] a chatbot can be used as an interface to an open domain QA system. Bayan Abu Shawar presented a chatbot that is used as a natural web interface to QA system [3]. ALICE bot was retrained to be able to answer university related FAQs. The chatbot is designed in such a way that it can be used to provide answers to FAQs of any university [10]. Another example is the chatbot based on ALICE that was trained on FAQs of the School of Computing at the University of Leeds [4]. Natural language systems were already built to provide access to semi structured data of yellow pages [11].

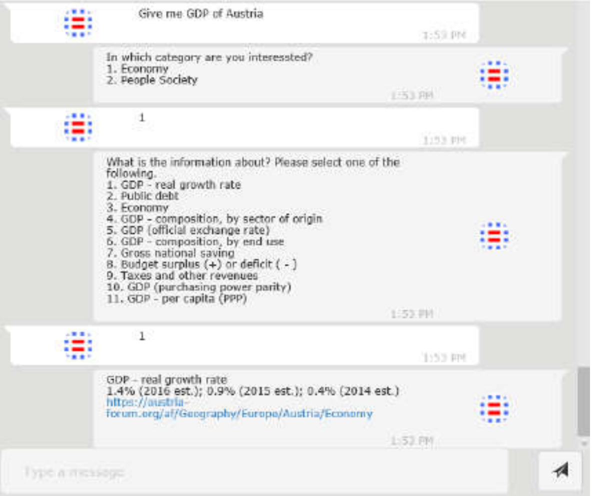
3. *Prototype Implementation*

The chatbot for Austria Forum is a standalone application developed with Java 8 technologies and runs on the Tomcat Apache server. The knowledge base of the chatbot stores information retrieved from the geography part of Austria Forum. The chatbot architecture consists of a client and a server side. The communication between the client and the server is enabled with the help of Java API for Web Sockets. The server side communicates also with a natural language understanding (NLU) platform, called Dialogflow.

The following should clarify how the chatbot basically works. When the chatbot is invoked in a browser, a single client page is displayed. The conversation starts when a user enters and sends a question. The user input is forwarded to the chatbot system running in background. The chatbot then performs input processing and searches for keywords and location information within the input. Subsequently, the user input is sent to Dialogflow for entity and intent recognition. The response from the NLU platform, which is sent in JSON string, is parsed. Having the parsed information, the chatbot knows what the user is asking for. It either creates a query based on the parsed information and performs a search, or engages users with questions and gathers information needed for the upload. In the case of a search, the search results are retrieved. If the number of results is within a defined range they are displayed to the user. Otherwise, the chatbot activates a conversation context and engages the user with clarification questions, collects information, updates the query and performs the search again. This procedure is repeated until the results are displayed to the user, or the chatbot cannot not find any information in the knowledge base. In the case of an upload, the chatbot activates a conversation context and verifies if the entered information corresponds to the requested format. If the verification is successful the chatbot uploads the information to the website. Otherwise, the chatbot notifies the user about the failed verification and is ready to process the next question.

3.1 Design

The design of the single client page was kept as simple as possible. It was developed using HTML, CSS and JavaScript. The client page, as can be seen in Figure 1, includes an input text field, a send button, and a conversation history field. The important thing is that a user does not need additional knowledge in order to communicate with the chatbot. The design of the client page is similar to commonly used conversational interfaces in messenger applications.



Give me GDP of Austria

1:53 PM

In which category are you interessted?
1. Economy
2. People Society

1:53 PM

1

1:53 PM

What is the information about? Please select one of the following.
1. GDP - real growth rate
2. Public debt
3. Economy
4. GDP - composition, by sector of origin
5. GDP (official exchange rate)
6. GDP - composition, by end use
7. Gross national saving
8. Budget surplus (+) or deficit (-)
9. Taxes and other revenues
10. GDP (purchasing power parity)
11. GDP - per capita (PPP)

1:53 PM

1

1:53 PM

GDP - real growth rate
1.4% (2016 est.); 0.9% (2015 est.); 0.4% (2014 est.)
https://austria-forum.org/af/Geography/Europe/Austria/Economy

1:52 PM

Fig. 1. Single Client Page

The knowledge domain represents an essential part of the chatbot. As stated in [6] there are two kinds of chatbots when considering the knowledge base: open domain and closed domain. With an open domain chatbot, the conversation can go in any direction. Closed domain chatbots are limited and can provide answers regarding one specific topic [3]. Since the chatbot for Austria Forum uses the geography website as an information source, it can be considered as a closed domain chatbot.

The information on the geography website is accessed over API and retrieved in JSON format. The JSON objects differ in content, structure depth and nested fields. Therefore, an auxiliary tool had to be developed to structure JSON objects. Each JSON object represents a chunk of information. This step was made in order to facilitate analyzing, searching, and editing of the knowledge base.

The natural language understanding (NLU) platform called Dialogflow is used for input understanding. Previously, the platform was launched under the name api.ai. Dialogflow is an artificial intelligence platform based on machine learning. It is owned by Google and includes built-in agents, which can be seen as modules required for natural language understanding. The agent “geo search” was created in order to handle text input forwarded from the chatbot system. The agent functions based on entity and intent concepts. The entities represent a chunk of information in a user input. In addition to built-in entities (e.g. date, number, city, country), custom entities were defined (category, continent). An intent can be seen as a mapping between a user input and possible responses. It helps to understand what users are searching for. The “geo search” agent has seven intents categorized in two groups, the “search” and the “upload” group. The intent groups help the chatbot understand when users want to search and when to upload information. The platform provides user interface for agent training process. It is possible to provide conversation examples and, in this way, improve the agent understanding. Of course, the agent is able to learn with the time and become more intelligent.

3.2 *Architecture*

Each chatbot follows a defined flow that begins with a user input and ends with displaying of answers. The chatbot for Austria Forum follows the general pipeline [6]. The first stage of the pipeline is concerned with input processing, followed with input understanding where named entity and intent recognition is performed. The last two stages deal with information retrieval for the response or candidate response generation, and selection or generation of the response. The architecture of the chatbot system for Austria Forum includes four components as illustrated in Figure 2: natural language processing (NLP), natural language understanding (NLU), search, and logical component.

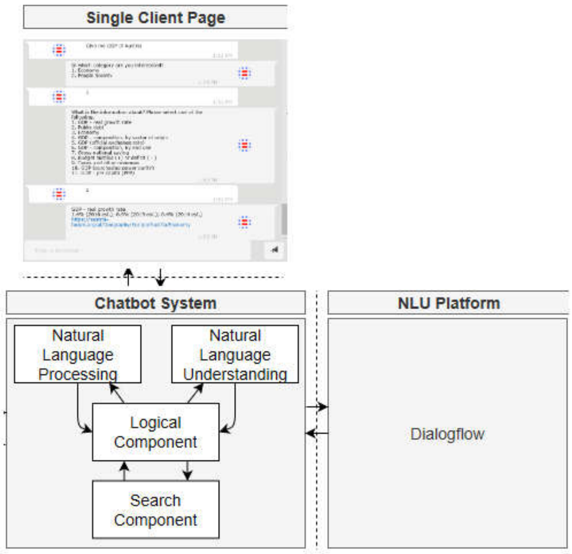


Fig. 2. Architecture of the chatbot

The Natural Language Processing Component is based on the Stanford CoreNLP software library. User inputs are passed through the pipeline of different tools that perform tokenizing, sentence splitting, part of speech (POS) tagging, parsing, named entity recognition (NER), and lemmatization [12]. The main task of the NLP component is to extract keywords and location tags. The results of the POS tagger are used for keyword extraction. This tool provides over thirty different tags. The main focus was on location tag extraction since the knowledge base includes geographical information. Location extraction was performed with the help of named entity recognizer. For English, NER recognizes person, location, and organization entities. How these two tools work is shown in an example question in Figure 3 and Figure 4.

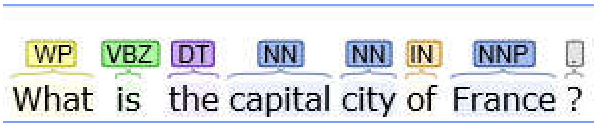


Fig. 3. Part of speech tagger

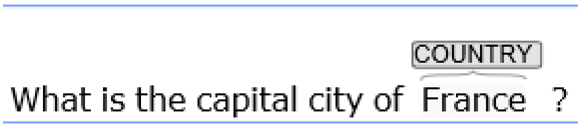


Fig. 4. Named entity recognizer

During the time of the research an evaluation of the Stanford CoreNLP and the Apache OpenNLP libraries was conducted. The number of the provided tools is almost the same. Apache OpenNLP uses different models for the POS tagger and NER tools, whereby CoreNLP needs only one model for its tools. The POS tagger of CoreNLP delivers better accuracy and takes less time than the POS tagger of OpenNLP. [13]

The Natural Language Understanding Component is used to extract meaning from the user input and makes it understandable for the chatbot. The component communicates with Dialogflow where entity and intent matching is performed. The response from Dialogflow is retrieved in JSON format containing detected entities, matched intents, actions, and parameters. Having this information in addition to keywords and location information, it is possible to create complex and efficient queries and improve relevance of the search results. Several NLU platforms were considered during the research. Almost all platforms provide capabilities for intent and entity recognition. Dialogflow and wit.ai can be used free of charge while others are for commercial use only. [6]

The Search Component provides search functionalities and is based on Lucene Core. Lucene Core is part of Apache Lucene, which is an open-source full-text search library. Apache Lucene can be used for creating search engines [14]. The library is based on an indexing and searching concept. The index of the search component consists of documents. Each document represents a mapping of a JSON object. Each key-value pair of the JSON object is stored in a text type field. The searching process begins with the query generation. Each query is created based on the information provided by NLP and NLU components. The search component implements methods for generation and update of different query types (e.g. Boolean, Term, Phrase, Range, Wildcard) and for running them on a single or multiple text fields. For example, a location query generated by a search component is a term query and is executed on the country document field. The search method executes queries on the defined index and the retrieve method retrieves them in form of documents. The search results are listed based on document score. To avoid large amounts of results and to i-prove relevance, a threshold had to be defined. Each search result is maximal three hundred characters long and includes a link to the page from where it was retrieved. An evaluation of the Solr search server was also considered during the research. Solr is built on top of Lucene and provides REST-like API for querying and retrieving of documents [15]. In contrast to Solr which is used for enterprise and content management system, the Lucene Core is suitable for programming prototypes, because it provides full control over internal processes.

The Logical Component communicates with the client side and is responsible for interaction and information flow between the components. It also manages the conversation flow and context, and forwards results to the client. The logical component performs particular actions and sets particular contexts depending on the intent matched on the NLU platform. Five different contexts can be activated. They are also grouped in “search” and “upload” contexts. In case of a large amount of search results, the logic component activates one of the “search” contexts (continent, country or category) and requests additional input from the user. The context stays active until the results are forwarded to the client. Since the chatbot also supports the user while uploading multimedia content, the logical component provides methods for acquiring and verification of the entered information. In this case one of the “up-load” contexts is activated. If the requested information is entered and its verification is successful, the information is uploaded to the website.

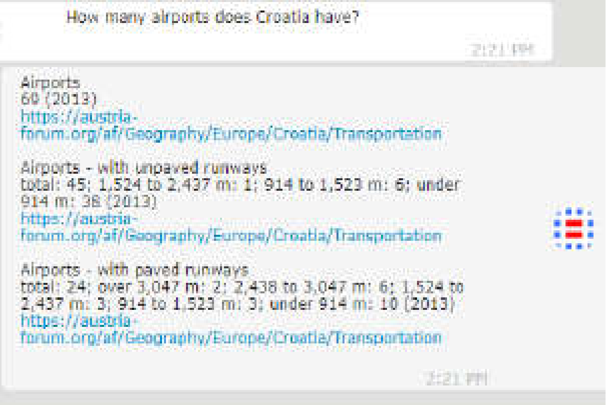
4. *Testing the Chatbot Prototype*

In order to test the chatbot, several example questions have been defined as can be seen in Table 1. The aim is to show how the chatbot behaves when it receives questions that include different information.

Question	Text
Q1	Can you provide some information about Nigeria?
Q2	Give me some information about energy in India
Q3	How many airports does Croatia have?
Q4	I need information about population
Q5	What are the most common natural hazards?
Q6	Can you provide information about energy?
Q7	I would like to upload a video

Table 1. Example questions

Since the knowledge base consists of geography information, the main focus was on location tags within a question. If the location is found and the number of search results is within a defined range, the chatbot displays the search results as illustrated in Figure 5. The location information is also present in question 1 and question 2. Because of the large number of results, the chatbot would pose clarification questions. In this way it would gather additional information and perform the search again. The procedure is repeated until the number of results is within the defined range.



How many airports does Croatia have?

2:21 PM

Airports
69 (2013)
https://austria-forum.org/af/Geography/Europe/Croatia/Transportation

Airports - with unpaved runways
total: 45; 1,524 to 2,437 m: 1; 914 to 1,523 m: 6; under 914 m: 38 (2013)
https://austria-forum.org/af/Geography/Europe/Croatia/Transportation

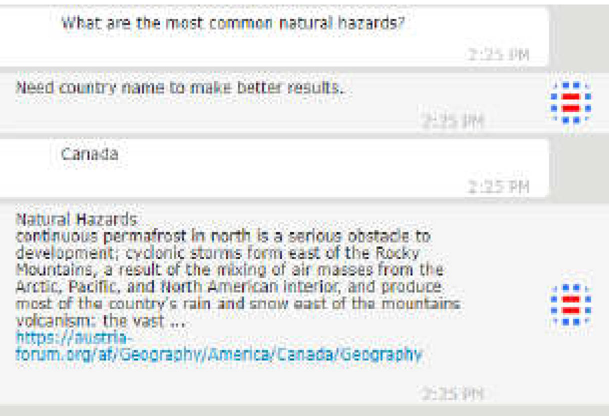
Airports - with paved runways
total: 24; over 3,047 m: 2; 2,438 to 3,047 m: 6; 1,524 to 2,437 m: 3; 914 to 1,523 m: 3; under 914 m: 10 (2013)
https://austria-forum.org/af/Geography/Europe/Croatia/Transportation

2:21 PM

Fig. 5. Search results for Q3

Questions 4, 5 and 6 do not provide any information about the location. The chatbot would perform a search and in case of a large number of results, require a location information (country name). Once the chatbot receives a location tag, it would either proceed with follow up questions or display the search results. Figure 6 shows the conversation flow for question 5.

The NER tool of the used NLP library showed shortcomings. Some of the continents were recognized as a country or as a city. This caused the chatbot to fail to answer questions containing the misinterpreted information.



What are the most common natural hazards?

2:25 PM

Need country name to make better results.

2:25 PM

Canada

2:25 PM

Natural Hazards
continuous permafrost in north is a serious obstacle to development; cyclonic storms form east of the Rocky Mountains, a result of the mixing of air masses from the Arctic, Pacific, and North American interior, and produce most of the country's rain and snow east of the mountains volcanism: the vast ...
https://austria-forum.org/af/Geography/America/Canada/Geography

2:25 PM

Fig. 6. Search results for Q5

As already mentioned, the chatbot is able to support users when uploading pictures, video or audio clips. This scenario occurs if a user poses questions similar to question 7. The chatbot would activate upload context and gather information from user, verify it and in case of successful verification upload the information to the website.

Since the chatbot is a closed domain chatbot, it does not provide answers to every user question. The knowledge base determines the intelligence of the chatbot. If a user searches for information which is not related to geography and does not exist in the knowledge base, the chatbot will answer with one of the pre-defined answers. The chatbot does not generate new answers but retrieves information and makes it available to users.

In addition to the search agent that was created, the built-in Small Talk agent was activated on Dialogflow. The Small Talk agent is able to match Small Talk intents and extends the number of user inputs that can be handled. This capability contributes to the improvement of user experience.

In contrast to search engines, the chatbot accepts queries in form of full sentences in natural language as well as keywords. In most cases the choice of the question form affects the search results. The search results contain a chunk of text and a link to the page containing the relevant information. The number of search results to be displayed, as well as the content length of each result, can be configured.

5. Conclusion

The goal of this work was to develop a chatbot standalone application that can be used as an information retrieval tool in a geography domain. The chatbot should be used for information search, as well for upload of information.

At the beginning of this paper, its context and the problem of the relevance were introduced. Definitions of chatbots and drivers behind increased interest in development were discussed. Types and areas of application were mentioned with the focus on information retrieval field.

The knowledge domain represents the brain of chatbots. It was shown how to design and structure a semi-structured data. In the future this step should be considered to create a relational database and use the chatbot as a natural language interface. The setting up of an agent on an NLU platform was described.

The architecture of the chatbot system, its components and libraries used were discussed. It was shown how a natural language understanding component in an information retrieval chatbot system can enable the generation of qualitative and complex queries and in this way improve the relevance of search results.

At the end of this work, the chatbot prototype was tested on several questions. The results showed that the chatbot provides satisfactory answers. The chatbot has potential as an information retrieval tool and could be used as an alternative to an integrated search engine in a closed domain.

References

[1] W. Bruce, Croft & Donald, Metzler & Trevor, Strohman. (2010). Search Engines: Information Retrieval in Practice. Pearson. ISBN: 978-0136072249

[2] Shawar, Bayan & Atwell, Eric. (2007). Chatbots: Are they Really Useful?. LDV Forum. 22. 29-49.

[3] Shawar, Bayan. (2011). A Chatbot as a Natural Web Interface to Arabic Web QA. International Journal of Emerging Technologies in Learning. 6. 10.3991/ijet.v6i1.1502.

[4] Shawar, Bayan & Atwell, Eric & Roberts, Andrew. (2005). FAQchat as an Information Retrieval System.

[5] Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. Commun. ACM 9, 1 (January 1966), 36-45. DOI=http://dx.doi.org/10.1145/365153.365168

[6] Nimavat, Ketakee & Champaneria, Tushar. (2017). Chatbots: An overview. Types, Architecture, Tools and Future Possibilities.

[7] A. M. Turing, I.—Computing Machinery and Intelligence, Mind, Volume LIX, Issue 236, October 1950, Pages 433–460, https://doi.org/10.1093/mind/LIX.236.433

[8] DALE, R. (2016). The return of the chatbots. Natural Language Engineering, 22(5), 811-817. doi:10.1017/S1351324916000243

[9] Quarteroni, S., & Manandhar, S. (2007). A Chatbot-based Interactive Question Answering System.

[10] Ranoliya, Bhavika & Raghuwanshi, Nidhi & Singh, Sanjay. (2017). Chatbot for university related FAQs. 1525-1530. 10.1109/ICACCI.2017.8126057.

[11] Kruschwitz, Udo & De Roeck, Anne & Scott, Paul & Steel, Sam & Turner, Raymond & Webb, Nick. (1999). Natural Language Access to Yellow Pages. 34-37. 10.1109/KES.1999.820113.

[12] Manning, Christoper & Surdeanu, Mihai & Bauer, John & Finkel, Jenny & Bethard, Steven & McClosky, David. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. Proceedings of 52Nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 10.3115/v1/P14-5010.

[13] Nanavati, Jay & Ghodasara, Yogesh. (2015). A comparative study of Stanford NLP and Apache Open NLP in the view of POS tagging. International Journal of Soft Computing and Engineering (IJSCE).

[14] Balipa, Mamatha & Ramasamy, Balasubramani. (2015). Search Engine using Apache Lucene. International Journal of Computer Applications. 127. 27-30. 10.5120/ijca2015906476.

[15] Kumar, Vikash & Brawal, P.N.. (2016). Implementation of highly optimized search engine using Solr. International Journal of Innovative Research in Science, Engineering and Technology.

Mirhet Saracevic works as a software developer at B&R Industrial Automation in Graz.

Markus Ebner is currently working as a Junior Researcher at the Department Educational Technology at the Graz University of Technology.

Martin Ebner is head of the Department Educational Technology at Graz University of Technology. He also works as a Senior Researcher at the Institute of Interactive Systems and Data Science.