An Information Retrieval-based Approach for Building Intuitive Chatbots for Large Knowledge Bases

Andreas Lommatzsch and Jonas Katins

TU Berlin, DAI-Labor, Ernst-Reuter-Platz 7, D-10587 Berlin, Germany

Abstract. Finding quickly the relevant information is essential in many application scenarios. In the past years, huge data collections have been created, but for most users it is still very difficult to find the information relevant for a specific, often complex problem. With the advances in automatic language processing chatbots have been developed to simplify the information search providing an intuitive user interface that gives the user the needed information in a natural dialog.
In this work we present a chatbot framework that answers questions related to services offered by the public administration. The framework enables complex dialogs and supports the user with giving hints and recommendations. Based on the framework, public chatbot services have been deployed for two major German cities designed to answer questions related to offered service, locations, and appointments. The paper discusses the architecture of the system and explains the developed algorithms. We report experiences running the systems as well as discuss the strengths and weaknesses of the developed approach.

Keywords: chatbots, information retrieval, context, human-computer interaction, natural language processing, question answering

1 Introduction

Personal assistants are getting more and more popular in a growing number of domains. These "virtual agents" act as experts providing answers to questions and supporting users in solving routine tasks. These personal bots have been developed as an additional channel to FAQs, hotlines and forums enabling a natural interactive conversation with the user. In contrast to classic search systems, chatbots should support longer dialogs (interactive search) and guide the user in finding information for complex problems. In addition, chatbots should be able to handle both keyword queries and complex "natural" sentences.

The development of chatbots leads to several challenges. One main problem is the variety of ways that can be used for describing a problem. In addition, the used vocabulary and the meaning of terms often depend on a specific domain.

This typically requires large collections of training data from the target domain. For very narrow scenarios (e.g., ordering a pizza) the training data might be created within an acceptable timeframe; for more general scenarios characterized by several thousand information objects, the generation of training data is extremely expensive and often infeasible at all. Thus, the use of information retrieval approaches extracting queries from natural user inputs seems to be a promising approach.

Another challenge is the translation of colloquial language into the often formal "official" language used in knowledge bases created by domain experts. Moreover, users do not pay much attention to spelling and correct grammar in chat interfaces. Thus, natural language analysis tools typically trained on large text corpora show only a limited precision.

In this work we present our approach for building a chatbot framework optimized for answering questions related to the services offered by public administration (e.g., applying for a new passport or registering an apartment). The framework combines Information Retrieval techniques and machine learning methods. Based on the framework we have deployed chatbot instances linked on the official web sites of two major German cities. This enables us to collect real user feedback and to evaluate the strengths and weaknesses of our approach.

The remaining work is structured as follows. In Section 2 we give a brief overview on existing approaches and discuss strengths and weaknesses. Section 3 explains our approach and describes the developed methods. Section 4 discusses the evaluation results. Finally, a summary and outlook on future work is given in Section 5.

2 Related Work

In this section, we review related approaches and analyze existing chatbot systems.

Personal Assistance Bots and Chatbot Frameworks Conversation agents ("Chatbots") have got in the focus of interest in recent years. This can be explained by ubiquity of mobile devices and installed personal assistants informing users about the weather, calendar entries and news as well as allowing users to control smart home devices [2]. These systems analyze the user input by applying rule sets. Based on the determined user intent information from the calendar or encyclopedias (e.g. Wikipedia), are queried. These systems usually are focused on providing short facts in a direct answer without supporting long dialogs.

For developing personal agents several frameworks have been developed, such as Amazon Lex ("Alexa")[1] and Google DialogFlow[2]. The core concept used by the systems is the *intent*. For each intent, several example sentences ("utterances") must be defined. These sentences typically contain slots that are matched with specific user inputs. In order to ensure that the frameworks reliably match the user input, a large number of patterns must be defined for every

[1] https://docs.aws.amazon.com/lex/latest/dg/what-is.html
[2] https://dialogflow.com/

intent. This typically leads to an explosion of complexity. Thus, the systems are developed for use cases characterized by a rather small number of intents. When defining a new virtual assistant, it is difficult to include existing knowledge bases since they are incompatible with the intent structure of the chatbot frameworks.

A similar approach for building on premise chatbots is used by the chatbot framework "rasa" [4]. The framework integrates several natural language analysis tools and a neuronal network that must be trained with annotated sample sentences.

These frameworks are usually applied for handling a rather small set of questions. Users can train their personal chatbot so that it learns the user wording and adapts to the individual user preferences. In order to ensure a reliable detection of user intents several dozen training examples per intent are needed making this approach not applicable for scenarios characterized by a large collection of intents.

FAQ Sets and IR-based Approaches Traditionally, lists of frequently asked questions or forums are used for providing answers to questions. In order to cope with large sets of questions, information retrieval-based approaches are used. These approaches are implemented based on an inverted index, enabling the efficient matching of user questions with a huge set of texts [10]. A high result precision is reached by excluding stop words, reducing words to their stem, and by weighting terms based on a TF-IDF scheme [7]. In order to optimize the precision of the result set, term weights can be optimized based on relevance feedback [8] or query expansion techniques can be applied [3].

The advantage of the approach is that existing knowledge bases and FAQ lists can be used for creating a chatbot. The task of handling new questions could be delegated to a human expert who creates a new FAQ entry after answering the question. The disadvantage of this approach is that longer dialogs are not supported. This results often in a lack of context information that leads to a reduced answer precision. The concept of FAQ-based chatbots is used by the chatbot of the city of Vienna[3] or the bot of the Singapore administration[4].

Semantic Topic Mapping The semantic storage of information based on ontologies and semantic graphs enables a generic, language-independent knowledge representation [5, 9]. Linguistic approaches are applied, such as Named Entity Recognition and Named Entity Disambiguation, linking the natural language user inputs with the nodes in the semantic graph. In a first step, the user input is enriched using a Part-Of-Speech tagger. Detected entities are mapped to the concepts of an ontology. Semantic query languages are then used for querying the requested facts from the knowledge base.

The advantage of Semantic Topic Mapping is that the representation of facts and the processing of natural user inputs are separated, allowing for existing optimized tools to be utilized. The semantic graphs can incorporate the context resulting in an improved answer precision.

[3] https://www.wien.gv.at/bot/

[4] https://www.gov.sg/news/content/5-reasons-to-use-the-gov-sg-bot

The disadvantages are that the conversion of text-focused databases into semantic, ontology-based knowledge stores is a complex, time-consuming task. The creation of ontologies and the mapping from texts into ontologies is expensive and requires abstract knowledge about the specific domains. Furthermore, semantic query languages are usually deployed for retrieving a direct answer and do not support longer dialogs.

Discussion The existing approaches do not fulfil the requirement for building well-scalable chatbots based on large knowledge collections. The search engines are optimized for key word queries and return documents. Natural, complex dialogs are not supported.

Systems trained based on annotated question-answer pairs are not applicable for large knowledge collections since the required amount of training data does not exist. Due to the underlying probability models these approaches are often unstable. A retraining of specific questions may reduce the precision for already well-trained question-answer pairs.

The weaknesses of the existing approaches motivate us to develop a new approach for building chatbots based on existing knowledge collections. Our approach adapts existing knowledge bases in a way that information can be used in a "natural" dialog. This enables us to overcome the weaknesses of existing systems without creating knowledge bases and training data collections from scratch.

3 Approach

We develop an architecture that integrates existing databases and components for handling dialogs as well as mapping natural language user questions to knowledge base entries. The system architecture is visualized in Fig. 1. In the subsequent sections we explain the components and developed methods.

Adaptation of Data to be Suitable for a Chatbot The key components of the system are databases describing the information objects. In our concrete use case, we have to manage information about services, locations, and external topics. We split the documents into parts of reasonable size, usually on paragraph level. In order to ensure that the paragraphs are understandable without other paragraphs, references and anaphors need to be resolved (removing the dependencies from other resources). The extracted paragraphs must not be too long to make sure that the information pieces are useful, and that the agent does not answer in long monologs. The paragraphs are enriched with meta-data building the basis for the matching of user questions with answers.

Structuring services provided by the Public Administration For describing the services, ontologies and meta-data can be used. In Germany, the LeiKa catalog[5]

[5] http://www.gk-leika.de/

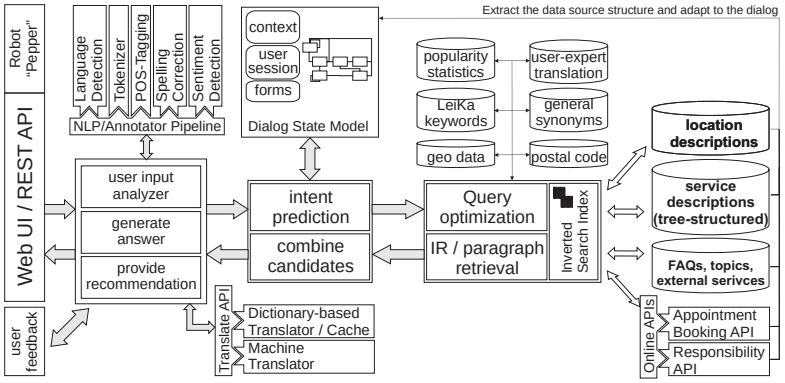

Fig. 1. System Architecture

defines a hierarchy of services offered by the public administration. In addition, keywords for the services are provided containing synonyms and colloquial terms helping to match real user language with the formal administrative language. Due to the broad range of topics in the LeiKa catalog not all ontology nodes are well covered with keywords.

Synonyms and Meta Data In a natural dialog, users often use colloquial language. This is a challenge since the used terms differ from the official language used by the administration. Thus, synonymy and colloquial terms must be learned by the system in order to "translate" the user language into the formal language. Colloquial terms such as "paperboard" (German: "Pappe") might be ambiguous and domain dependent. In addition outdated terms or region specific wording must be considered. General synonym collections are usually not enough since they do not provide enough coverage for the specific domain. This requires that these "service keyword" collections must be continuously extended and maintained based on observed dialogs and user feedback.

For supporting a location-focused search, databases of postal codes, street names and urban district names must be incorporated for handling questions such as *"What are the opening hours of the registry office Friedrichstraße."*

Providing Information in Different Levels of Detail Dialogs as well as documents usually start with general information and provide subsequently details for different aspects. This means that the information is stored in a tree-like structure. A user question does not only have to be matched with the right subtree, but with the concrete node. If a user asks "What are the fees for getting a passport", the service "apply for a passport" must be matched as well as the node "fees". Thus, the question is only precisely answered, if both criteria are matched correctly. In order to avoid giving the user incorrect answers, our system asks the user, whether the service has been identified correctly ("Are you interested in getting information about 'apply for a passport' "). This additional question

means an additional effort for the user but improves the trust in the systems since the method ensures that the user intent has been identified correctly.

Handling Dialogs – Considering the Context In chats users find the requested information interactively. This means that users often ask questions related to previous questions, typically focusing on more detailed information. Thus the chatbot must determine for every question whether the new user input should be handled as a follow-up question or whether the user wants to switch to a new topic. In our system we use a state model. If the user already selected a location or a service object, subsequent questions are handled as follow-up questions. If no information can be found based on this assumption, the systems asks the user, whether the user expects follow-up information or wants to switch to a new topic.

Recommendations Asking the right question is difficult if the domain is new to a user. To address this problem, we implemented a recommender component that suggests questions based on the behavior of other users. This is especially useful to make sure that the user becomes aware of all relevant aspects.

Detecting Off-Domain Questions Chatbot systems are typically developed for providing answers for one domain; but users often check, how the system handles questions outside the target domain, e.g. *Who is the chancellor of Germany* or *Who will win the National Soccer Championship*. Providing detailed answers for every possible question is too complex and too expensive to implement. If the system answers simple off-domain questions correctly, this motivates users to ask more complex questions. In order to ensure a high answer quality, the chatbot system should focus on one domain. We implemented classification algorithms predicting whether a question belongs to the desired domain or the question is out of the scope of the system.

Multi-language Dialogs In order to reach a larger number of users a translator component has been developed. It is based on an arbitrary third party translation service such as Microsoft Azure[6] or DeepL[7]. First, the user question is translated from any input language supported by the translation service to the system's main language, which is German in our scenario. The generated answer is then translated back to the user's original input language. Both these steps are entirely transparent to the user. Translations are highly parallelized and aggressively cached. This guarantees a good performance so that the user hardly notices any delay. The quality of the translations depends on how the input text is phrased. Since the backend of the chatbot system uses an IR-based approach, the system deals well even with translations of grammatically incorrect inputs. Due to the use integrated synonym databases the automatically translated user questions are reliably matched with the knowledge base entries.

[6] https://azure.microsoft.com/en-us/
[7] https://www.deepl.com/

Discussion The developed chatbot architecture has been designed to efficiently handle different types of user requests and to deliver detailed answers for a wide spectrum of questions based on existing data collections. The information objects are split into paragraphs and enriched with additional meta-data. A dialog handling components guides the user and ensures that the context is considered when computing answers. We have tuned the system by integrating additional meta-data collections and learning keywords and phrases. The developed system is open for new knowledge sources and can be adapted to additional domains.

4 Evaluation and Preliminary Results

We have implemented a conversational bot and deployed the system on the official web portal of two major German cities. We started without official announcements as an additional channel for citizens seeking for information about the services offered by the administration. In the first month we observed a constantly increasing interest in the service. After 6 months, we served ≈ 2500 dialogs per month on one city portal giving us insights in the user preferences and user habits.

Dialog Analysis Analyzing the dialogs, we found that most users enter questions in complete sentences. In contrast to our expectations that the majority of users would only use keywords, the questions are asked in a way as if a human would be in the chat, even though the system explicitly explains that the chat is operated by a machine. Handling very comprehensive, detailed user inputs is a challenge. If there are too many questions (different intents) in one user input, it is hard for the bot to deliver an answer containing all needed information. In order to handle this problem the chatbot gives the user a hint to focus on one aspect, if the user input is too long. Very long answers from the agent we avoided by restructuring the knowledge databases (in cooperation with responsible persons of the administration) in order to make the information better understandable and suitable for a chat.

Classifying the Type of User Questions The core of our chatbot system is a large knowledge base. The implemented chatbot systems use the official service database describing all services offered by the city authorities[8]. Since the databases of services and locations are the core of the system, we expected that almost all questions should be related to these knowledge base entries. Classifying the observed dialogs, we found that 55% of the questions are related to services, 15% are related to locations, and about 5% are related to services and topics offered by external administrative offices (e.g., from the federal government). About 5% of the dialogs are general small talk, e.g., questions related to the weather or the name.

Surprisingly, about 25% of the questions relate to appointment booking, such as how to find appointments fast or what to do if the user cannot find the

[8] https://www.115.de/

details of an already booked appointment. When designing the chatbot system we were not aware of the importance of this type of questions. We trained the chatbot based on user feedback in the first weeks after going public. This meant to identify the relevant questions (that could not be answered by the chatbot system yet) and find appropriate answers in discussion with the administration. We observed that typically 250 sample sentences are needed for reliably learning these questions. This underlines that only for a limited question domain this effort is manageable. From the perspective of knowledge management, the rather big fraction of appointment related questions had been an argument for the official service website to put additional clarifications on the web portals in order to reduce the number of questions related to this topic.

Feedback and Unclear Answers A chat seems to be a natural way of getting information. This also motivates users to give directly feedback to the quality of the answers. Therefore we designed explicit feedback buttons in the system; but the users tend to give feedback in the chat in natural language sentences. This user feedback has been helpful for identifying questions not covered yet by the knowledge base and finding answers that contain confusing information. If no service and no matching FAQs are found matching the user question, we ask the user to reformulate the question. The explicit query reformulations help us to learn synonyms and to understand the wording used by users. The extracted and checked knowledge is used to extend our knowledge base.

Another important aspect is the handling of praise and complaints. If the user gives feedback within the chat, the chatbot must not answer with "I did not understand your intention", but acknowledge the feedback.

Multi-Language Support Our chatbot system integrates a machine translation service. This allows users to chat with the system in eight different languages. The user input is translated on-the-fly into German; the answers are translated back from German into the selected user language.

The multi-language support is new for the administration since almost all information related to the citizen services is only available in German. This means, that offering support in foreign languages is a value adding feature. In the live systems $\approx 7\%$ of the sessions are using the machine translation component.

A specific challenge in supporting foreign languages is the difficulty in translating domain-specific terms. The translation of user questions into German works rather well for most questions due to the IR-based approach used in the backend. Due to user inputs usually consisting of several words, the translations of the words match well with the German service descriptions. The translation of German service descriptions into a foreign language is more critical. The challenge is to ensure that the automatic translation of the administrative language is understandable by users. In order to address the problem we checked the answers given by the chatbot for the most popular services. In case of unclear translations we manually collected corrections optimized for our scenario.

Overall, providing multi-language support is a valuable feature in chatbots optimized for the public administration. Users chatting in a foreign language

gave positive feedback. Several users remarked that the 115-telephone support in English is very limited; so the chatbot should not refer to the 115-telephone support ("public authority telephone").

Robot-based User Interfaces The chatbot system has been developed for a traditional text-based chat. In order to improve the visibility of the chatbot project we have also developed a version optimized for a physical robot-based interface. As robot a 1.20 m big robot "Pepper"[6, 1] has been used.

The new "physical" user interface required several adaptations. We have integrated a voice-to-text service converting the captured speech into text. This introduces a new source of errors since the speech-to-text component sometimes ignores parts of the user input. Thus the backend must be able to handle incomplete sentences. Our IR-based approach for finding relevant answers works in most cases robustly since the question-answer matching is done based on keywords. For small talk for which the question-FAQ mapping (trained on complete sentences) often additional training examples have been needed for the speech-to-text component. Moreover, navigational elements such as lists or radio buttons must be optimized for a voice-controlled interface. For instance, users can say "select suggestion two" or "service two" for selecting the second element from a list of suggestions.

For handling the answers provided by the robot specific optimizations are needed. Very long answers should be provided in smaller parts. It should be possible to interrupt the robot when answering questions in long monologues. Moreover the users may ask the robot to repeat an answer. In text-based chats repeating an answer does not make sense since users could simply scroll up; in a conversation with a robot the query "could you repeat the answer" is rather common.

Discussion Overall the developed system shows a good performance providing reliable, highly accurate answers related to the wide spectrum of topics. The conversion of existing databases into a format optimized for chats enables "natural" dialogs for a wide spectrum of topics. Nevertheless, we had to add knowledge required for small talk and for answering questions indirectly related to the information objects.

5 Conclusion and Future Work

In this work we presented our chatbot framework optimized for answering questions related to services offered by the public administration. Compared with most existing chatbot systems our approach scales well with the size of the chatbot's knowledge. This ensures that the chatbot covers the complete range of entries from the knowledge base. Furthermore our approach allows setting up a new chatbot instance without explicit training data by mapping existing databases to the ontology used by the chatbot. For building good queries from the user question we combine statistic methods (e.g., popularity-based methods) and natural language analysis models (e.g., syn-sets and Part-Of-Speech

tagging). The deployed chatbot systems are incrementally optimized by analyzing the user feedback. We extract synonyms by applying Word-2-Vec models and analyzing the input from users when reformulating a query ("could you please rephrase your request"). If there are several answers matching a user request, the relevant answers can be figured out interactively in a dialog.

The evaluation shows that the deployed systems provide a good answer quality. The retrieval-based approach works robustly with the variety of natural language formulations. The training of the query optimization matching component with real user feedback improves the matching for the most popular services. A big challenge is the handling of off-topic questions. Since the system is optimized for one domain, it is difficult to detect whether a question is off-topic or uses a very uncommon wording. In order to address this problem we plan to build a dataset for learning a classifier. Furthermore, we are working on more sophisticated machine learning methods for extracting relevant patterns from the collected user feedback in order to improve the structure of the dialogs and to optimize the matching precision for fuzzy questions.

References

1. I. Aaltonen, A. Arvola, P. Heikkilä, and H. Lammi. Hello pepper, may i tickle you?: Children's and adults' responses to an entertainment robot at a shopping mall. In *Procs. of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, pages 53–54, New York, USA, 2017. ACM.
2. S. A. Abdul-Kader and D. J. Woods. Survey on chatbot design techniques in speech conversation systems. *International Journal of Advanced Computer Science and Applications*, 6(7):72–80, 2015. http://dx.doi.org/10.14569/IJACSA.2015.060712.
3. S. Albayrak, S. Wollny, A. Lommatzsch, and D. Milosevic. Agent Technology for Personalized Information Filtering: The PIA-System. *Scalable Computing: Practice and Experience*, 8:29–40, 2007.
4. T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*, 2017.
5. A. Lommatzsch, B. Kille, and S. Albayrak. An agent-based movie recommender system combining the results computed based on heterogeneous semantic datasets. In *Proc. of the 13th GI International Conference on Innovative Internet Community Systems and the Workshop on Autonomous Systems*, I2CS '13, Düsseldorf, Germany, 2013. VDI-Verlag.
6. J. Markowitz, editor. *Robots that Talk and Listen: Technology and Social Impact.* Walter de Gruyter, 2014. page 41, ISBN 9781614514404.
7. A. Mishra and S. K. Jain. A survey on question answering systems with classification. *J. King Saud Univ. Comput. Inf. Sci.*, 28(3):345–361, July 2016.
8. S. Robertson, H. Zaragoza, et al. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.
9. K. Toutanova, V. Lin, W.-t. Yih, H. Poon, and C. Quirk. Compositional learning of embeddings for relation paths in knowledge base and text. In *Procs. of the 54th Meeting of the Assoc for Comp. Linguistics*, volume 1, pages 1434–1444, 2016.
10. C. Zhai and S. Massung. *Text Data Management and Analysis: A Practical Introduction to IR and Text Mining.* ACM and Morgan; Claypool, NY, USA, 2016.