



Fachhochschul-Masterstudiengang
SOFTWARE ENGINEERING
A-4232 Hagenberg, Austria

Chatbot Prototype Development and Evaluation from a Development Perspective in a Given Domain

Masterarbeit

zur Erlangung des akademischen Grades
Master of Science in Engineering

Eingereicht von

Gerald Spenlingwimmer

Betreuer: FH-Prof. PD DI Dr. Stephan Dreiseitl

Hagenberg, Juli 2020

Declaration

I hereby declare and confirm that this thesis is entirely the result of my own original work. Where other sources of information have been used, they have been indicated as such and properly acknowledged. I further declare that this or similar work has not been submitted for credit elsewhere.

Hagenberg, 07.July.2020

Gerald Spenlingwimmer

Contents

| | |
|--|-----------|
| Abstract | iv |
| Kurzfassung | v |
| 1 Introduction | 1 |
| 1.1 Introduction | 1 |
| 1.2 Motivation | 2 |
| 1.3 Company | 3 |
| 1.4 Research Questions | 3 |
| 1.5 Prerequisites | 3 |
| 1.6 Expected Results | 3 |
| 2 Basics | 5 |
| 2.1 Chatbot | 5 |
| 2.2 Chatbot/NLU Frameworks | 6 |
| 2.3 Suitable Problems for Chatbots | 6 |
| 2.4 Intent | 7 |
| 2.5 Utterance | 8 |
| 2.6 Entity | 9 |
| 2.7 Chatbot Response | 10 |
| 2.8 Story | 10 |
| 2.9 Webhook | 11 |
| 2.10 Domain | 11 |
| 2.11 Evaluation of NLP capabilities | 12 |
| 2.12 Natural Language Processing/Understanding | 12 |
| 3 State of the Art | 13 |
| 4 Methodology | 22 |
| 4.1 Approach | 22 |
| 4.2 Comparison and Evaluation | 25 |
| 4.3 Course of Action | 28 |
| 5 Design | 30 |
| 5.1 Architecture | 30 |
| 5.2 Story/Flowchart | 31 |
| 5.3 Intents, Entities and Utterances | 32 |
| 5.4 Psychological Aspects | 33 |
| 5.5 Webhook | 34 |

| | | |
|-----------|---------------------------------|-----------|
| 5.6 | Messenger UI | 34 |
| 6 | Prototype Implementation | 36 |
| 6.1 | Dialogflow | 36 |
| 6.1.1 | Intents and Entities | 36 |
| 6.1.2 | Webhook | 37 |
| 6.2 | Watson Assistant | 38 |
| 6.2.1 | Intents and Entities | 38 |
| 6.2.2 | Dialog and Webhook | 39 |
| 6.3 | Rasa | 40 |
| 6.4 | LUIS | 44 |
| 7 | Result | 46 |
| 8 | Discussion | 64 |
| 9 | Conclusion | 67 |
| 10 | Appendix | 70 |
| | Bibliography | 79 |
| | List of Figures | 80 |
| | List of Tables | 81 |
| | List of Listings | 82 |

Abstract

The 3 Banken IT wants to use chatbots for two internal use-cases. The first use-cases are sickness notifications and the second one are vacation requests. The main questions which need to be answered are what makes a problem suitable for a chatbot solution and are the given problems suitable for a chatbot. A problem is suitable if it's repetitive, can be automated, and only requires simple back-and-forth communication Raj [29]. This is the case for the given problem which makes it suitable for a chatbot solution. For the proof of concept, prototypes were developed with Dialogflow, LUIS, Rasa, and Watson Assistant. Based on a detailed evaluation that included the entity extraction performance, the intent classification performance in English and German a technology ranking was built to determine the best technology for the given problem. In the evaluation process the similarities and differences of the frameworks were compared. All of the tested frameworks used the concepts intent, entity, and utterance. The biggest difference is that Dialogflow, LUIS, and Watson Assistant are cloud-based solutions whereas Rasa is a local solution. The evaluation also focused on the difference between the cloud frameworks and the local framework. An important factor for companies are the costs for a technology and in general the prices of the frameworks are on an equal level as expected. The entity extraction evaluation was won by Dialogflow by a large margin backed up by the highest f-score. The intent classification in English worked best with Watson Assistant closely followed by Dialogflow and LUIS. Rasa was unable to keep up with the cloud technologies in the English test setting. The result for the German test setting looked quite different since the intent classification worked measurably better and all technologies performed well. The entity extraction performance of the technology dropped significantly with the exception of LUIS which showed a performance increase. For the development of a prototype in English, all technologies can be recommended. The only technology which showed a great performance on the German tests was LUIS followed by Rasa which performed quite well. These two technologies can be recommended for a German bot. The entity extraction performance of Dialogflow was horrible and can't be recommended. Watson Assistant doesn't offer annotated entities for training sentences when German is selected as input language. This renders Watson Assistant useless for the German language until the feature is available for German and can't be recommended. The two remaining technologies for the final recommendation are LUIS followed by Rasa.

Kurzfassung

Chapter 1

Introduction

1.1 Introduction

A chatbot is a conversational agent able to communicate with users in turns using natural language [30, 31, 24, 22]. Chatbots are becoming more and more popular in recent years. The three main reasons for the growing popularity are the rise of chat systems in everyday life, the advancements in machine learning, and the advancements in NLU in recent years. [15] The original chatbot technologies required expertise in machine learning. Hence, a machine learning expert was required for the development of a chatbot and the broad public of developers was unable to implement chatbot systems. Thanks to the recent advances in machine learning no machine learning skills are required for the development of chatbots anymore. This makes chatbot systems attractive to programmers. Another reason for the increasing attention is the development effort which has been reduced greatly in the last few years. Nowadays small to no programming knowledge is needed to develop and implement a chatbot [15]. This means that people without ML and programming skills are still able to develop a chatbot. Because of these characteristics, chatbots are becoming popular in the customer support section where they are used to replace humans [20]. They are also used to increase the user acceptance of FAQ pages. The major benefits of chatbot systems are that they never complain, are cost-efficient, and are available 24 hours a day seven days a week. Chatbots can greatly reduce the costs of customer services and increase customer satisfaction [32]. A chatbot never complains, is available seven days a week 24 hours a day, and can handle large demands [32]. They are always available if e.g. assistance is required at 3 AM the chatbot will be available whereas a customer service is bound to the opening hours and might not be available at 3 AM. Large demand handling is especially interesting for companies which don't have the resources to hire enough people to match the service demands. It's possible to replace humans with chatbots as mentioned in Rahman et al. [28] or aid them by reducing the workload like in Deshpande et al. [16]. Chatbots are perfectly suited for monotonous routine work which can be used to aid people with daily work where they can focus on more important tasks. All these reasons are key factors why chatbot systems are getting popular. In this master's thesis, multiple important factors need to be considered. The knowledge-base for the company for further development needs to be built. It needs to be researched which technologies are state-of-the-art technologies to build chatbots. Lots of chatbot frameworks have been built where a developer can choose from. The big tech companies offer Dialogflow (Google), Watson Assistant (IBM), Alexa/Lex (Amazon), LUIS + Microsoft

Bot Framework as cloud services. From the found state-of-the-art technologies, some need to be chosen and compared. The chosen technologies need to be compared with domain-specific data to determine which frameworks can be recommended for further development. The chosen domain for this thesis is a combination of sickness notifications and vacation requests. There are some key requirements for the bots. The bot needs to collect information from a natural language conversation of short text messages. The extracted information is used to identify the person and is sent to a server for processing. To ensure a fair comparison in a fair environment the frameworks will be trained and tested with the same data to prevent bias. The end result is expected to look and feel similar with all technologies. The frameworks are expected to share the same or at least similar concepts. They should be easy to develop and the functionality offered is expected to be similar. The NLP performance is expected to be on an equal level like in Braun et al. [15] where the difference between Rasa and LUIS was quite small. Some papers like Braun et al. [15] compare the NLU and NLP capabilities of services. In this thesis, the frameworks are compared from the NLU and development perspective. Hence, things like setup, development difficulty, and the available features of the frameworks are compared. This information is used to give a recommendation for future development. After the evaluation of the framework, a recommendation will be provided. The recommendation provides a ranking of the frameworks based on conditions. Another important aspect is the development of chatbot systems in general. This includes the development approach, the development differences to regular development, and psychological aspects that need to be considered.

1.2 Motivation

The company currently has no knowledge of chatbot systems. To justify the use of a chatbot system knowledge about those systems needs to be collected and evaluated. Knowledge needs to be collected about chatbots, the possible use-cases, the state-of-the-art technologies, the concepts, the evaluation of chatbot systems, and psychological aspects. It needs to be determined if problems like the sickness notification are suitable for a chatbot. It's of general interest to find out which use-cases are suitable for chatbots in general to give recommendations and inspiration for further projects. With the collected knowledge the sickness notification use-case needs to be implemented using multiple frameworks to determine the best framework for the use-case. At the moment, the sickness notification process starts with a call at the company. A person receives the call, enters the information into the computer system and enables an out of office email notification for the sick person. The head of the department also needs to be notified. This repetitive task needs to be solved by a chatbot. The new process also starts with the call but this time the person receiving the call enters the retrieved information into the chatbot system and an out of office email notification is created automatically. This saves time the employees can spend on important, non-repetitive tasks. For this use-case, the chatbot needs to collect the information from natural language and submit the data to a server for processing. Chatbots can serve requests all the time and can serve a larger amount of people immediately [25]. Chatbots are accessible, efficient, 24/7 available, scalable, cheap, and user behavior can be monitored [29]. They are also a great solution when it is impossible to hire a large enough staff to handle lots of user requests [25]. Chatbots can be used to make complicated search operations easier to understand and easier to use for humans [25]. It's also a natural way of communication. Anyone who can talk to or write with a

person can work with a chatbot interface [29]. Hence, it is an easily understandable interface suitable for almost everyone. Other major points are the performance of the different frameworks, the ease-of-use, the development effort, and the supported features. Based on these and similar criteria the frameworks are evaluated to find the best suitable framework and the strengths and weaknesses of these systems. All this information together builds the knowledge-base for further research and development.

1.3 Company

The master's thesis is for the 3 Banken-IT GmbH. The managing directors are Karl Stöbich and Alexander Wiesinger [2]. The 3 Banken-IT is the general contractor for all IT-services inside the 3 Banken-Gruppe [3]. The core competences of the 3 Banken-IT are the software development and maintenance of banking applications, IT-security, IT-support, operation of the data-centers, and the operation of central and distributed IT-infrastructure [3]. The 3 Banken-Gruppe consists of three independent regional banks namely the Oberbank, Bank für Tirol und Vorarlberg and the BKS Bank [1]. The supervisor of the master thesis from the company side is Thomas Reidinger the chief of enterprise architecture.

1.4 Research Questions

- Where are the differences and similarities between the chatbot frameworks?
- Which framework is the most promising for the given problem?
- Which problems are suitable for the use of chatbots?

1.5 Prerequisites

There are some prerequisites which have an influence on the chosen technologies for this thesis. Based on this thesis further projects will be developed. The company has currently no experience with chatbot systems. Hence, the collection of information about state-of-the-art technologies is a key area for the company and an important prerequisite. The company uses a local container environment for deployment. A offline capable chatbot framework needs to be included in the tests to allow the use in the companies local systems for further projects. The company does use IBM software Hence the use of Watsons Assistant is an option. For a comparison at least three technologies need to be chosen. The domain defined by the company are sickness notifications and vacation requests. This means that the chosen chatbots need to support the collection of information in a given domain. The main purpose is to recommend a technology for further development. For this reason the chatbots need to be compared from a technical view and the development view.

1.6 Expected Results

At least three chatbot frameworks have been selected based on the findings in state-of-the-art. The given problems of sickness notification and vacation request have been analyzed

for the suitability for chatbots. The domain information has been converted to dialogs. The dialogs have been implemented with all technologies if possible. The training and test data for the chatbots have been defined. All technologies are trained and tested with the same data for a fair comparison. Evaluation criteria have been extracted from the state-of-the-art chapter. Additional evaluation criteria have been defined. A detailed comparison has been created for the different technologies showing the strengths and weaknesses from a development point of view. New use cases were found for the use of chatbot systems for further projects. Based on the prototype evaluation the best framework has been recommended for further development.

Chapter 2

Basics

2.1 Chatbot

A chatbot is a conversational agent able to communicate with a user using natural language [30, 31, 24, 22]. The term chatbot is a combination of the words chatting and robot. In literature synonyms like chatterbot and conversational agent are used. The communication with a chatbot is most often done in written form like in a chat but not restricted to it since speech is also classified as natural language. The communication between man and machine is done in turns until a conversation end is reached or the conversation is aborted [34]. A conversation end is when the underlying task has been achieved. The user input is interpreted by the chatbot and the most likely response is generated [17]. In the best case, the generated response is reasonable and intelligent [34]. In the worst case a generic message like "I didn't understand that" is generated. It's important to keep in mind that a chatbot can mimic intelligence, human dialog, and personality but is often limited to specific functionality [25, 17]. Three major chatbot types are present in the literature. There are menu-driven chatbots (Singh et al. [32]), domain-specific/goal-oriented/task-oriented chatbots (Deshpande et al. [16], Williams et al. [35], Braun et al. [15], Williams et al. [36]), and chatbots for general/open-ended/end-to-end conversations (Brandtzaeg and Følstad [14], Singh et al. [32]). Menu-driven chatbots mimic a classical interface and navigate the user through the interface using text or speech. An example of a menu-driven chatbot is a phone call at a support hotline. The caller is guided through the interface by the system and isn't connected to a person immediately. The system gives options like "press one if you'd like to request a refund, press two if you need replacement parts", navigates the caller through the process, collects information like the order number, and connects the user to the correct department. Domain-specific chatbots offer functionality that is limited to a domain. Such systems provide the functionality to achieve specific tasks and are not supposed to handle anything else. One example would be a refund desk in a shop. If the customer asks for refund specific information or wants to issue a refund the customer is served. If the customer asks where the TVs are located in the shop he is redirected to the information counter. A chatbot system in the refund desk domain offers only the functionality necessary to get and collect refund information and to issue a refund. If the bot is asked something else no useful response can be expected. Open-ended systems are made for a general conversation. Contrary to the other two approaches the user can ask open-ended systems anything and the most likely response is generated. There are no fixed rules what users may or may not ask and the conversation can develop

in any direction. These systems have a hard time to collect specific information since there is no way to tell which information is needed. Every conversation can vary greatly hence no information can be presumed. Popular examples for open-ended systems are Google Home, Siri (Apple), and Alexa (Amazon) [32]. The most limited approach is the menu-driven approach since the options are given and a user only inputs numbers to select a functionality. Domain-specific bots communicate with users in natural language and are good in achieving tasks and collecting information. Open-ended systems are perfect for general communication and are not restricted to a domain but it's hard to collect specific information or achieve a domain-specific task with such systems. General tasks are not a problem for open-ended systems. A simple exemplary chatbot conversation is shown in Figure 2.1. In general, an ML-based chatbot system consists of rules and training data. The training data is used to train natural language capabilities and should be as realistic as possible. The training data can be handcrafted for prototyping or generated from large sources like emails, phone calls, or the internet. The rules define the available functionality, types, and possible input statements. One simple rule could be that a person's name is required as input and consists of a first and last name. A chatbot has two major tasks. It needs to classify the intention of the user called intent and extract the relevant pieces of information from the message. A relevant piece of information is called an entity. An entity has a type and can be e.g. a date or the name of a person.

2.2 Chatbot/NLU Frameworks

Many chatbot and NLU frameworks were created to enable non ML experts to develop chatbot systems. The frameworks can be divided into cloud-based and local solutions. The cloud-based solutions are created, edited, and developed through a web-interface and are used in applications via a REST API endpoint. The developer is dependent on the provider and the training and analysis are done in the cloud of the provider. This needs to be kept in mind when dealing with sensitive information like bank accounts. The big cloud providers offer chatbot or NLU frameworks in their cloud environment. Microsoft has LUIS (Williams et al. [35], lui [7]), Google develops Dialogflow (dia [4]), IBM offers Watson Assistant (wat [10]), Amazon provides Lex and Alexa. The local solutions are standalone applications running on the development machine or the company server. The data is never passed to another computer or the cloud since it runs locally. A popular local open-source solution is Rasa (Bocklisch et al. [12], ras [8]). All the frameworks listed above use slot-filling. A slot is a predefined type/entity which a user needs to enter in a conversation. In Table 2.2 the slots which need to be filled for a hotel reservation could be the hotel name, start and end date, and a name for the reservation. All of these slots need to be filled before the system can process the hotel reservation automatically.

2.3 Suitable Problems for Chatbots

Chatbots seem like intelligent systems but most often they are designed for specific tasks and are unable to solve complex problems [29]. Therefore, not every problem can be solved using chatbots [29]. A problem can be solved effectively with chatbots when it is highly repetitive, can be automated, and can be done with simple back-and-forth communication. A chatbot system is meant to reduce the workload of employees or serve a large number of users. Therefore, the problem needs to be repetitive to fulfill these

requirements. It is possible to solve non-repetitive tasks with chatbots too but it is not cost-effective. The problem needs to be automated to simplify the process. If it can't be automated an employee is needed for the process and it's harder to serve a large number of users. Chatbots are based on natural language communication with text messages or voice. Hence, the problem must fit in the back-and-forth communication pattern since there are no other means of communication available for chatbots. If all the statements above can be answered with true the problem is suitable for a chatbot. A simple QnA bot for a FAQ page of a website is an example of a suitable problem [29]. The task can be automated, is highly repetitive, and the answers are predefined by the FAQ page. As mentioned in Section 2.1 there are three major types of chatbots. The menu-driven and domain-specific approach are suitable for tasks with a goal where data needs to be collected to execute an action where only a limited set of functions needs to be supported. When the user can select an option from a limited list of actions the menu-driven approach works best (support, helpdesk chatbot). When users should be able to communicate using natural language to ask questions related to a domain then the domain-specific approach works best (FAQ chatbot). In the FAQ example, the domain consists of all questions and answers of the FAQ page. It would be bothersome to navigate through a FAQ forum using numbers because if e.g. 1000 topics are available the user would need to enter a number between one and 1000 and also would need to know the correct number for the question. This would be horrible to use and no one would want to use such a system. A user just wants to ask a question to get an appropriate answer. This works best with natural language hence, the domain-specific approach is suitable. The open-ended approach is suitable when no specific data needs to be collected and a general conversation needs to be held with the users. The Google Assistant can answer general questions like "What's the weather in New York" and gives a general response. The Google Assistant can't process a request like "Tell my employee I am sick and won't come to work today" since this would require domain-specific data and actions.

2.4 Intent

Intents are the core concept of any chatbot system. An intent is the action a user wants to perform [17, 28]. In example sentence two in Table 2.2 the user wants to book a hotel. The action the user wants to perform is "book a hotel". The intent can be named "book hotel" for instance. In general, the user queries the system, the system identifies the best matching intent, collects relevant information, and generates a response. The identification of these intents is the core function of chatbots. To identify the intent the system finds the intent where the training phrases are closest to the users query or uses the fallback if no intent reached a high enough score. The score is represented by a percent value between zero and one. One means a perfect match and zero is treated as absolutely no match. A threshold defines the minimum score the system has to detect before an action is executed. All values below the threshold are handled by the fallback action e.g. a generic response like "I'm sorry but I didn't understand that". In Table 2.2 example sentences are listed with their intent, and their entities. In example sentence one of Table 2.2 the intent can be labeled as "Greet" since the user greets the bot. Table 2.1 shows a small excerpt of the information generated for the default welcome intent created by Dialogflow. The intent consists of training phrases and a response list. The training phrases are used to train the ML model. Everything which is similar to the training phrases and reaches the highest score will be classified as the intent. The system then responds

with a random pick from the response list. The random pick fakes intelligence since it would be boring for users to get the same response every time. If the same response is returned all the time users would rate the communication as robotic but the goal is to make the conversation as natural as possible.

| No. | Intent | Type | Sentences |
|-----|-----------------|------------------|--|
| 1 | Default Welcome | Training Phrases | hey howdy partner heya hello hi hey there hi there greetings howdy |
| 2 | Default Welcome | Response List | Hi! How are you doing? Hello! How can I help you? Good Day! What can I do for you today? Greetings! How can I assist? |

Table 2.1: Default Welcome Intent of Dialogflow

2.5 Utterance

An utterance is a sentence or a part of a sentence. Especially in the area of chatbots, an utterance is often a part of a sentence since chat messages are often shortened. All sentences listed in Table 2.2 are utterances for instance. The messages of users are utterances [32, 17]. The chatbot also responds to users with utterances. Table 2.1 shows some training phrases and the response list. The sentences listed as training phrases and the response lists are utterances. Table 2.1 shows that a collection of utterances belongs to an intent and builds the base for the classification of user input. The training phrases are different versions of the same question or statement. The different versions allow the system to correctly classify utterances which never were part of the training data as long as the utterance is similar. Example sentence two of Table 2.2 can be used as utterance for the book hotel intent. Utterances need to be chosen with care to prevent overfitting to unimportant features of the sentences. If all utterances contain a specific word at a specific position the model might learn that a specific word must be at that position all the time. Utterances for training purposes can come from multiple sources. They can be handcrafted for development and testing. The data can come from live systems. Real user input from e.g. e-mails, phone conversations, or online resources can be used to get a large amount of training data [32].

| No. | Sentence | Intent | Entities |
|-----|---------------------------------------|-------------|---------------------------|
| 1 | Hello | Greet | - |
| 2 | I'd like to book a hotel | Book Hotel | - |
| 3 | The Twelve Months hotel | - | Hotel Item |
| 4 | From the 27th of April | - | Date (From) |
| 5 | To the 4th of May | - | Date (To) |
| 6 | Ethan May | - | Person (Name) |
| 7 | Order one green t-shirt size M please | Place Order | T-Shirt Item, Color, Size |

Table 2.2: Examples for Concept Explanation

2.6 Entity

An entity is a piece of information that needs to be extracted from natural language. An entity is metadata which belongs to an intent and has a type. Entities can be represented as volumes, counts, and quantities for instance [29]. An intent can have multiple entities like in example sentence seven of Table 2.2. In sentence seven one entity is the color of the shirt, the second is the shirt as an item, and the third one the size. Sentences three to six of Table 2.2 show entities in a book hotel example. In a simplified example, to book a hotel the name of the hotel (Hotel Item), the check-in and check-out dates (Date), and a name (Person) are needed for the reservation. To extract entities in a context from natural language Named-Entity-Recognition (NER) is used. NER is a sub-area of NLU. Entity recognition deals with word ambiguity. Sentences five and six show a word ambiguity. In sentence five May is the month and needs to be treated as a date whereas in sentence six of Table 2.2 it's the last name and needs to be treated as a person entity. A named entity is a noun phrase relating to individuals. Classic individuals/named entities are person (Bill Gates), location (New York), and organization (Google) [19]. Frameworks sometimes provide predefined named-entities like the ones above which are needed on a regular basis. A developer can define named entities for a specific domain like in the hotel example where an entity is needed which knows all valid hotel names. Example sentences four and five of Table 2.2 show a date entity, and example sentence six shows a person entity. An example entity for food items like apples and peas is shown in Table 2.3. An entity can be a collection of items like for the fruits but it can also be a collection of synonyms. Entities are stored in the session and form the context of a conversation [32]. They are only valid for the current conversation.

| No. | Entity | Item | Synonyms |
|-----|------------|---------------------------------|---|
| 1 | Fruit Item | apple banana pea peach | apples bananas peas peaches |
| 2 | Sick Item | sick | ill, unwell, poorly, indisposed, sickly, ailing |

Table 2.3: Entity examples

2.7 Chatbot Response

The response message of a chatbot is picked from a pool of utterances as described in 2.5. The response can be picked at random or in a sequence. Alternatives are picked to keep the conversation interesting for the users especially if they are using the bot more than once. The alternatives prevent a robot-like feel in the conversation. If the same response is picked all the time users realize fast that they are talking to a bot system and it gets boring fast. An exemplary list of responses for the default welcome intent of Dialogflow is shown in Table 2.1 next to response list.

2.8 Story

A story defines the dialog flow. It's a step-by-step script for a conversation between man and machine. The bot and the user interact in turns using natural language [30, 31, 24, 22]. The user asks a question or makes a statement and the bot processes the request and delivers a response. A story can be a "happy" or "sad" path. A "happy" path is a successful interaction from start to finish without any problems. The "sad" path handles the behavior when something doesn't work out. It could be that the user input is wrong and needs to be re-entered when the validation fails. The user can also quit the conversation at any time. These "sad" paths need to be handled in separate stories. In Figure 2.2 a story for a greeting sequence with a user is shown. The user starts the conversation by entering a greeting phrase. The response of the chatbot is also a greeting phrase and then the conversation ends. An example conversation from a person with a chatbot is shown in Figure 2.1. The story for 2.1 is that the user asks for open hours, the bot respond open hours, the user asks for an item, the system validates the item and a response is generated with the item in the response text.

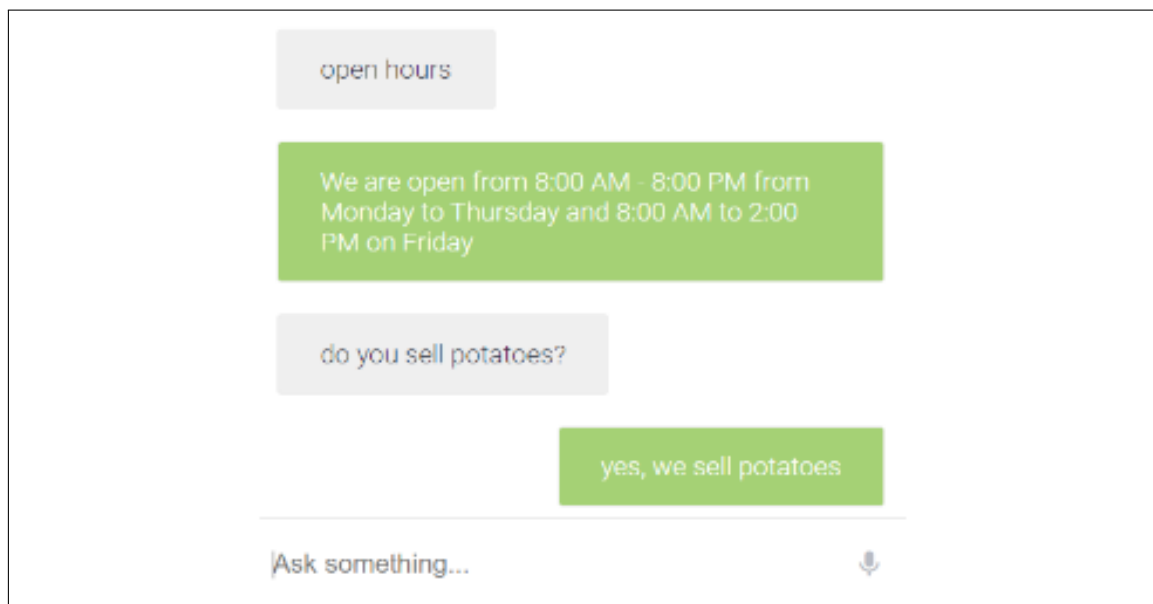


Figure 2.1: Example FAQ Conversation with Dialogflow



Figure 2.2: Greeting Story

2.9 Webhook

A webhook is a URL where the frameworks send metadata and intent data for analysis, completion, and processing. In the case of this thesis, the webhook is a simple Python REST service implemented with Flask. Each framework gets a route in the service which is `"/watson/webhook"` for the Watson Assistant requests for example. In Listing 2.1 a simplified JSON request from Dialogflow is shown for the default welcome intent. In the case of Dialogflow the incoming message can be found under `query text`, the response message under `fulfillment text` and the confidence under `intent detection confidence`.

```

1  [
2  {
3    "originalDetectIntentRequest": {
4      "payload": {}
5    },
6    "queryResult": {
7      "action": "input.welcome",
8      "allRequiredParamsPresent": true,
9      "fulfillmentMessages": [
10     {
11       "text": {
12         "text": [
13           "Good day! What can I do for you today?"
14         ]
15       }
16     }
17   ],
18   "fulfillmentText": "Good day! What can I do for you today?",
19   "intent": {
20     "displayName": "Default Welcome Intent"
21   },
22   "intentDetectionConfidence": 0.72385716,
23   "languageCode": "en",
24   "parameters": {},
25   "queryText": "Hi, how are you doing"
26 },
27 ]
28 ]

```

Listing 2.1: Dialogflow Webhook Request Example

2.10 Domain

The domain of a chatbot is equal to the functionality. A domain-specific chatbot has a limited set of operations and works with data relevant to the domain. If the chatbot

is built for a refund-desk scenario then the domain is refund for example. The bot can process and answer refund-related issues, collects refund specific information, and can issue a refund. It doesn't offer functionality which is unrelated to the refund domain-like "What's the weather in New York today" and it is not supposed to do so. Open-ended systems escape the domain limitation but have problems to execute domain-specific tasks or collect domain-specific information.

2.11 Evaluation of NLP capabilities

The scenario for this explanation is a medical test for a disease. Positive is treated as the person has the disease. A true positive (tp) is when the test is positive and the person has the disease. A false positive (fp) is when the test is positive but the person is healthy. A true negative (tn) is when the test is negative and the person is healthy. A false negative (fn) is when the test is negative but the person has the disease. In the area of chatbots, a true positive is when the input is e.g. "Hello" and the identified intent is "Greet Intent". The precision is calculated as $\frac{tp}{tp+fp}$. The recall is calculated as $\frac{tp}{tp+fn}$. The f-score is calculated as $2 * \frac{precision * recall}{precision + recall}$. The range of the f-score, precision, and recall values is zero to one where zero is the worst possible result and one is a perfect result. The precision shows how many retrieved elements are relevant. The recall shows how many relevant items have been retrieved. The f-score can be seen as the average of precision and recall. In the medical area the recall is more important than the precision because the goal is to correctly identify as many sick people as possible. It's not that important if people are false positive since further checks are always possible to verify the test result. In the area of chatbots neither false negatives nor false positives are desirable hence the f-score is the most relevant value. The precision is based on the selected elements (TP + FP) and the recall is based on the relevant elements (TP + text). Figure 10.3 from the appendix can be used as a visual aid for a better understanding.

2.12 Natural Language Processing/Understanding

The term "natural language" means input is received through voice or writing in the user's language of choice [29]. Natural language processing in the area of chatbots is necessary to analyze the text input provided by the user and figure out the intent and entities necessary to process the given information [29]. Natural language processing is seen as the brain of a chatbot since it receives the information from the user, processes the data, and retrieves the relevant parts from the input for post-processing [29]. Part-of-Speech (POS) tagging is a process where parts of speech like noun, verb, and adjective are assigned to each word/token [29]. This information is used to understand the context and is part of NLU. An important part of natural language processing is the search for the root of words which can be done with stemming and lemmatization [29]. Stemming reduces a word to the base form e.g. "saying" is reduced to "say" by removing the endings of words [29]. Lemmatization tries to return the dictionary form of a word and has better correctness and is more productive than stemming and hence the preferred method [29]. Lemmatization uses a vocabulary and morphological analysis of words to achieve its task [29]. Stop words are words that hold little to no meaning and occur in a high frequency like "the" and "a" which are removed.

Chapter 3

State of the Art

Williams et al. [36] and Bordes et al. [13] state that the two major chatbot types are goal-oriented and end-to-end systems. Kane [25] defines two different categories of chatbots. They can either be classified as a local standalone application or a web-based chatbot [25]. The types/classes of chatbots from Williams et al. [36], Bordes et al. [13] and Kane [25] don't contradict they just look at chatbots from a different perspective. From the design perspective, the conversation type of a chatbot is either goal-oriented or end-to-end. From the development perspective, the used framework is either web-based or local. Goal-oriented or end-to-end is a general statement about the kind of base problem which the chatbot needs to solve. Web-based or local has an influence on where the chatbot is developed. Rasa is one example for a local standalone application chatbot whereas LUIS, Dialogflow and Watson Assistant are cloud-based/web-based chatbot frameworks. Rasa, Dialogflow, LUIS, and Watson Assistant are all capable of goal-oriented dialog.

Deshpande et al. [16] shows the evolution of chatbots from the first chatbots to current state-of-the-art chatbots like Alexa and Siri. The first chatbots created used pattern matching to respond to user inputs [16]. Current technologies like Siri use NLP to respond to user inputs [16]. With NLP the intent is identified and the question is analyzed to detect commands and actions [16]. The basic workflow of a chatbot system starts with the user input [16]. Then the NLP engine extracts entities and detects the intent [16]. Entities are used to retrieve the relevant data [16]. After the relevant data has been retrieved a response is generated [16]. Raj [29] explains the core concepts of chatbots, NLU, and NLP necessary for the development of chatbots. A food order chatbot is built with Dialogflow and a horoscope bot is built with Rasa in Raj [29]. Chatbots are easy to use because it's as natural as talking to a human without the need for complex interfaces [29]. Before starting to implement a bot the problem the bot should solve needs to be analyzed. Chatbots can't do everything, in fact, they are most often designed for one specific task (task-oriented) [29]. Some major questions need to be answered to determine if the problem is suited for the use of a chatbot [29]. If the conversation is simple and consists of simple back-and-forth communication, and the task is highly repetitive and can be automated then the problem is suitable for a chatbot [29]. It doesn't make much sense to implement a chatbot for a non-repetitive task which can't be automated [29]. If the task is non-repetitive then the development of the chatbot would take more time than the communication between the user and an employee. A perfect example use-case for a chatbot is a FAQ web page [29]. The answers are predefined, it's repetitive, it consists solely of questions and answers, and can be automated. Some of the best chatbot and NLP frameworks are Rasa, LUIS, and Dialogflow [29]. Geyer et al. [19] is about named entity recognition (NER) in micro-posts.

NER is the core concept used by chatbot tools to extract information from text. The classic examples for named entities are people, locations, and organizations [19]. Geyer et al. [19] examines how good custom entities are recognized by the system. The frequency of entities in the training data can influence the NER performance [19]. Geyer et al. [19] checked the performance of entity recognition which can be done for chatbot frameworks in a similar fashion. Rahman et al. [28] gives a small overview of the concepts of Dialogflow and the general architecture of chatbot systems. Commonly used cloud chatbot solutions are IBM Watson, Dialogflow, and Microsoft Bot [28]. In general, the chatbot architecture consists of intent classification, entity recognition, a response generator, and a response selector [28]. The intent classification identifies the best matching intent for the users input. The entity recognition module extracts the information/data from the user's message [28]. The response generator provides response candidates and the response selector chooses the best matching response [28]. Goal-oriented chatbots are most common in the business sector and help users to achieve tasks [28]. The tech giants provide chatbot frameworks for regular developers [28]. Google provides API.ai/Dialogflow, Microsoft has LUIS, IBM develops Watson, and Amazon provides Lex [28]. The key concepts of Dialogflow are intent, entity, and utterance. Intents link the user's input to actions which are executed [28]. Deshpande et al. [16], Raj [29], Geyer et al. [19], Rahman et al. [28] focus on the basics of chatbot frameworks and chatbot related concepts. In Deshpande et al. [16], Rahman et al. [28] describe the general workflow of chatbot systems. Geyer et al. [19] is about NER and explains the core concept of entities. Additionally, Raj [29] also explains which problems are suitable for a chatbot in general.

The first major type are the end-to-end systems. Shawar and Atwell [30] focus on the application of chat systems in different languages and the evaluation of the system. The evaluation of chatbot systems needs to be adapted to the application and the user needs [30]. The best way to evaluate a chatbot is to check if the specific service or task can be achieved and the evaluation needs to be adjusted to the use-case [30]. Each component of a chatbot system can be tested individually and the whole system can be tested based on user satisfaction and acceptance [30]. The whole system is evaluated through feedback by users in a test setting [30]. This feedback is then used to improve the system [30]. In Shawar and Atwell [30] a chatbot for frequently asked questions (FAQ) has been built which competed against the classical Google search. The participants had to answer a few questions like how many results the system found and which system they prefer and why [30]. The result of the user study was that the FAQ chatbot found an answer more often than Google [30]. 47% of the users taking part in the study preferred the custom solution and 11% preferred Google [30]. The major reason why the users preferred the FAQ chat was because the chat can often give direct answers and return fewer links on average which saves browsing time [30]. The users preferred Google because it's familiar and it can give different answers when the input query is adjusted slightly [30]. This shows that custom-built solutions can outperform classical approaches in small areas. Bordes et al. [13] analyses the strengths and weaknesses of end-to-end dialog systems in goal-oriented settings. The end-to-end approaches have recently shown promising results with chit-chat/general conversation with a user [13]. However, The most useful application areas for chatbots are either goal-oriented or transactional settings [13]. A classical dialog system uses slot-filling to collect information from the conversation [13]. Slot-filling is reliable but has limited scalability [13]. The slots are predefined and hence don't scale well because they can be very different for each problem. End-to-end systems scale well because they learn directly from conversations and make no assumptions regarding the

domain or dialog structure [13]. The goal-oriented example used by Bordes et al. [13] is the classic restaurant reservation bot. The main question is if the end-to-end system performs well on the classical goal-oriented restaurant reservation task. The restaurant reservation is a domain where goal-oriented systems perform well and is the ideal candidate to determine the strengths and weaknesses of end-to-end systems in comparison to goal-oriented systems [13]. The result of Bordes et al. [13] show that end-to-end systems are not yet ready to replace goal-oriented systems and have to improve before they can perform reliably in goal-oriented settings. Bordes et al. [13] compared the chatbot types based on the number of correctly identified requests (true positives) and the number of dialogs where every request was identified correctly. Multiple conversations were used to test the system and the end-to-end system never identified all requests of the conversation correctly [13]. The request identification as such worked well [13]. No conversation went smoothly since at least one request was identified incorrectly. In other words, each user had at least one problem in the conversation. This would lead to bad user experience. In summary, the most important statement of Bordes et al. [13] is that the current end-to-end systems are not performing reliably enough in goal-oriented settings. Williams et al. [36] introduces an approach that combines end-to-end dialog with domain-specific knowledge. A task-oriented system helps the user to achieve a task using natural language [36]. A restaurant reservation is a task-oriented operation that requires domain-specific knowledge [36]. The domain-specific knowledge in a restaurant setting can be the restaurant name and the number of people for the table reservation for instance. This information needs to be collected from the conversation. End-to-end systems lack a mechanism to inject such domain-specific information [36]. There is also no mechanism for constraints in goal-oriented systems [36]. In a banking app, the user needs to be logged in (constraint) before the account information can be retrieved [36]. A programmer can code such constraints with a few lines of code but lots of training data needs to be provided to a system to learn such a mechanism on its own [36]. The hybrid-coding-networks (HCNs) introduced in Williams et al. [36] extend the open-ended approach with domain-specific knowledge. The domain-specific knowledge is provided by the programmer and requires more development effort [36]. Since the domain-specific knowledge is provided by the programmer the system doesn't need to learn these parts and less training data is required in comparison to existing end-to-end approaches [36]. A large amount of training data can be retrieved from real dialogs [36]. The experiments show that the HCN approach performs on an equal level with existing end-to-end approaches in end-to-end settings and needs less training data for the same performance [36]. The HCN system also performed better in the task-specific setting than the existing end-to-end approaches [36]. HCNs are a mixture of end-to-end and goal-oriented systems. The developer has more control than with the classical end-to-end approach but it also requires more development effort to create a chatbot with HCNs [36]. The main benefits are the usage of domain-specific information in end-to-end dialog and the reduced amount of training data required. Other references focus on end-to-end and/or goal-oriented chatbots whereas Williams et al. [36] combined them to get the best of both worlds. Williams et al. [36] does not compare the HCN approach to a goal-oriented approach like Bordes et al. [13] did for regular end-to-end systems. Hence, no statement can be made if HCNs could replace goal-oriented systems or not. In Shawar and Atwell [30] testing focus lies on the user side whereas the testing focus of Braun et al. [15] lies on the NLU capabilities. The NLU capabilities are important for the developers of the systems before a chatbot is developed and help to choose the best framework for development. The users feedback is a valuable source

of information for the actual product and is important long after the framework has been chosen. These are two very different views on the performance of a chatbot in two different phases of a project. Bordes et al. [13] focuses on end-to-end systems like Shawar and Atwell [30], Williams et al. [36] but uses a goal-oriented setting for the comparison. The important information gained by Bordes et al. [13] is that currently end-to-end systems are not reliable enough for goal-oriented tasks. This leads to the conclusion that a goal-oriented technology should be chosen for goal-oriented projects. Williams et al. [36] is trying to combine the end-to-end and goal-oriented approaches to create a new type of system which can do goal-oriented dialog better than current end-to-end approaches. The result of Williams et al. [36] shows that the HCN system is better in goal-oriented settings than current end-to-end systems but doesn't compare the system to goal-oriented technologies.

Singh et al. [32] states that the conversation type of a chatbot can be categorized as a general conversation which is about a broad generic subject or a specific conversation about one product or service. An example of a general conversation is e.g. when a customer walks into a bank and starts talking to an employee [32]. In this example, we have no idea what the person wants or who the person is [32]. Popular examples for chatbots capable of general conversation are Google Home, Siri, and Amazon Alexa [32]. An example of a specific conversation is the refund desk in a store [32]. Specific rules apply to the refund process and specific information is required from the customer [32]. The domain could be named refund and the specific rules and information are part of the domain. The customer can't get any other information than refund related information at the refund desk [32]. The outcome of the conversation can be one of the predefined outcomes, the fallback, or a conversation end [32]. A specific conversation ends as soon as the underlying task is achieved or a conversation end is reached [32]. In general, specific conversations are easier to predict and are handled with higher accuracy [32]. Chatbot systems for specific conversations communicate with information [32]. Nowadays, chatbots are AI-driven. The core component of AI-driven chatbots is the NLP engine which takes care of data extraction from natural language [32]. The extracted information is used to determine the next steps [32]. There is no difference in functionality between a regular application and a chatbot in terms of functionality [32]. The difference is that a chatbot uses conversation whereas a regular app is a self-service application [32]. Conversational training data for chatbots can be acquired from lots of sources like emails, phone calls, chats, and social media [32]. Companies use chatbots because they can save a lot of money and because they offer a good customer experience [32]. Especially in areas with personal data the usage of cloud chatbot solutions like Dialogflow, Alexa, and Watson is problematic because all the conversation data is sent to and stored at the providers servers [32]. Because of high-security requirements, an in-house/local chatbot has been built in Singh et al. [32]. The first step to create a chatbot system is to define a conversational flow [32]. The conversational flow is a decision tree which describes all events, decision, and outcomes in a conversation [32]. Singh et al. [32] also gives an overview of the basics of Microsoft Bot (LUIS), Rasa, and Google Dialogflow.

The second major type are goal-oriented systems. Braun et al. [15] states that modern conversational agents can be built without programming knowledge because of natural language understanding services (NLU) [15]. The recent advancements in machine learning (ML) and NLU and the popularity of messenger platforms has led to huge progress in recent years [15]. People are familiar with messengers because many use them on a daily basis. Hence, chatting has become a natural way of communication. A big question is how

chatbots can be evaluated. Recent publications have discussed the use of NLU service but not why one service has been chosen over another [15]. In general, The architecture of a chatbot consists of three elements [15]. The request needs to be interpreted, the response needs to be retrieved and the response message needs to be generated [15]. The purpose of these services is the extraction of information from natural language [15]. Popular NLU services are the cloud services LUIS, Watson Conversation, Dialogflow/API.ai, and RASA as an open-source alternative [15]. The cloud-based solutions have advantages when it comes to hosting and scalability and Rasa has advantages when adaptability and data control are needed [15]. All three NLU services share the same basic concepts of intents, entities, and batch import in JSON format [15]. The cloud-based services are secretive when it comes to the ML algorithms and the initial training data [15]. The exception is Rasa where the ML backend can be chosen by the developer [15]. For the evaluation of the NLU services training data from a production chatbot and two other training sets from SatckExchange were used [15]. The NLU capabilities of the frameworks were compared. The comparison included the services LUIS, Rasa, Dialogflow, and Watson Conversation [15]. The same training data was used for all services [15]. This is important to get a comparable result. The evaluation of the services NLU capability is based on the true and false positives and negatives, the recall, precision, and F-score [15]. The better the F-score is the better the NLU service has performed [15]. The evaluation of NLU services is only a snapshot since the services are continuously improving [15]. LUIS performed best on all datasets but in the area of chatbots, Rasa and LUIS performed on an equal level [15]. Rasa can be customized further than LUIS and could perform better than LUIS through customization [15]. The domain had little to no influence on the ranking of the NLU performance [15]. In other words, a good NLU system performs good independent of the domain. If the training data is sparse then there is no significant difference between the services [15]. Before using an NLU service different services should be tested with domain-specific data [15]. In Dutta [17] a chatbot is developed that assists high school students with learning general knowledge subjects and analyses the impact the chatbot had on learning. AI-based chatbots are used for banking systems, customer services, and education [17]. Popular chatbot platforms are Dialogflow, LUIS, Wit.ai and Pandorabots [17]. The developed chatbot is a web-based education solution using natural language processing (NLP) techniques to answer questions [17]. It is also able to participate in small talk [17]. The chatbot systems are evaluated based on their NLP performance and the available features of the platform [17]. All platforms are trained with the same knowledge base for comparison [17]. LUIS and Wit.ai performed slightly better in the NLP part than Dialogflow [17]. Dialogflow offers the concept of follow-up intents which is an important feature for the development of sub-tasks and was chosen for development over LUIS and Wit.hai [17]. To create the illusion of talking to a human being some small talk is implemented [17]. This can motivate learners and might increase the interest in the topic [17]. Dutta [17] uses goal-oriented technologies like Braun et al. [15] and both compare the technologies based on their NLU capabilities. Additionally, Dutta [17] also focuses on the available features of the platforms and selects Dialogflow over the other technologies because of these features. Dutta [17] also touches the psychological aspects of chatbot systems like the motivation of users through small talk. Go and Sundar [20], Brandtzaeg and Følstad [14], Følstad and Brandtzaeg [18] go deeper into the psychological aspects of chatbot systems. Gregori [22] evaluated the NLU capabilities of Wit.ai, Luis, Api.ai/Dialogflow, and Amazon Lex[26] by the confidence score. The higher the confidence the better the user request matched the training data [22]. Each tool

was trained with the same data and was tested with the same questions to ensure fair conditions [22]. LUIS, Wit.ai, and API.ai/Dialogflow performed on an equal level in terms of intent classification in Gregori [22]. Gregori [22] chose Api.ai/Dialogflow for the development of a prototype. In Ahmad et al. [11] a chatbot is built to aid customers with questions regarding medication for a pharmacy company using IBM Watson. The bot can provide medications for an illness, give information about a specific medicine, and can give information on the medication intake [11]. The bot is a domain-specific bot since it answers questions to medication only. Important performance metrics for chatbots are the conversation length and structure, the ability to provide personalized communication, and the number of conversation steps [27]. The general trend is to keep the conversations short [27]. Personalized communication is used when recommendations or tips are provided for the user based on the information related to the user [27]. Retail chatbots need a larger amount of conversation steps to provide information and recommendations and hold the user's attention [27]. The measurement criteria for chatbots vary depending on the domain of the bot. Williams et al. [35] gives an overview of LUIS the state-of-the-art language understanding service of Microsoft. Historically, language understanding could be implemented via machine learning or handcrafted rules [35]. The ML model approach is robust but requires expensive expertise [35]. In general, software developers can build language understanding without the assistance of a framework with handcrafted rules [35]. But systems with handcrafted rules don't scale well [35]. With LUIS developers don't need machine learning knowledge to build language understanding models nor do they need to write handcrafted rules [35]. LUIS allows regular developers without ML expertise to develop cloud-based, domain-specific language understanding models [35]. Developers need to understand the concepts intent, entity, and utterance to work with LUIS [35]. Developers can create custom entities and use existing entities like location, date, and time. The communication with LUIS is done through an HTTP endpoint in JSON format [35]. Dialogflow, Rasa, and Watson Assistant also use an HTTP endpoint for communication and JSON format for the messages. Bocklisch et al. [12] describes the architecture and function of Rasa. Rasa is an open-source tool for natural language understanding and dialog management for developers [12]. Rasa uses the concepts of intent and entity [12]. The cloud chatbot technologies have these two concepts in common with Rasa. The state is represented by slots and the events which led to the current state [12]. This information is stored in the tracker [12]. An action has access to the tracker and determines the next step to take [12]. An action can be something simple like an utterance or something complex like a function which is executed [12]. Rasa consists of a natural language and a dialog component that can be accessed via an HTTP API [12]. The training data for Rasa can be specified in JSON or markdown format [12]. Rasa supports live training of the system where the developer can correct the bot while communicating with it [12]. Live training is an effective way to create training data for plausible conversations [12]. With live training, there is a higher possibility to create a complete and natural conversation because the developer is forced to go through the conversation step by step. Braun et al. [15] showed that Rasa performs on an equal level with tools like LUIS. Braun et al. [15] compares some NLU frameworks regarding their natural language capabilities. The compared frameworks are suitable for task-specific problems. Shawar and Atwell [30], Bordes et al. [13], Williams et al. [36] focus on end-to-end systems, Braun et al. [15], Dutta [17], Williams et al. [35], Bocklisch et al. [12], Ahmad et al. [11], Gregori [22] focus on task-oriented systems, and Singh et al. [32] is about both. Singh et al. [32] gives a general overview of the chatbot types, gives development advice, and introduces

popular chatbot technologies and their basics. Braun et al. [15] gives an overview of the architecture of chatbot systems and evaluates the NLU capabilities of chatbot frameworks. Gregori [22] and Dutta [17] also evaluate the NLU capabilities. When frameworks are compared for their NLU capabilities the same training data needs to be used for all tested frameworks like Braun et al. [15], Gregori [22] did. Dutta [17] also introduces a framework comparison based on the available features and some psychological aspects of chatbot communication. Unlike other references which either use or evaluate frameworks Williams et al. [35] and Bocklisch et al. [12] introduce the frameworks LUIS[35] and Rasa[12]. They focus on the technology itself and don't build a chatbot with it nor do they compare frameworks. In Dutta [17], Ahmad et al. [11], Przegalinska et al. [27] actual chatbots for specific use cases have been implemented.

Chatbots can also be viewed from a psychological perspective that focuses on user interaction. Følstad and Brandtzaeg [18] investigates how users are communicating on the web and shows how the recent technological advances in the area of chatbots could influence the human-computer interaction in the future. Lots of people are already using natural language as the main input method on the web through mobile messengers and social networks [18]. At the moment the natural language conversation online is from human to human through a machine interface [18]. The people are using messenger platforms (machine-interface) to communicate with other people (human to human). Some platforms already offer chatbot integration. At platforms like Twitter machine agents can already be integrated and communicate with human users [18]. Chatbots are different from classical software interfaces because they use natural language for communication instead of a classical GUI. Currently, developers are focused on designing user interfaces but for chatbots the design focus lies on the conversation itself [18]. Regular UI systems are designed by experts and improved by qualitative data which is rather sparse [18]. A big advantage of conversational interfaces is the massive amount of training data that is present all around the web [18]. Chatbots all use messenger like interfaces independent from the problem. Developers don't have to think about designing user interfaces for chatbots. The developers need to switch to a goal-oriented view where the main focus lies on understanding what the user wants and how they can be served [18]. The state-of-the-art technology of conversational interfaces is Google Assistant [18]. Brandtzaeg and Følstad [14] describes chatbots as a natural language interface using text or voice. They are used to retrieve content or access a service through natural language conversation [14]. People are used to natural language communication because they spend lots of time on messenger platforms [14]. Because of this chatbot technologies become more and more popular [14]. Person to person communication is like the open-ended chatbot approach. It is difficult to design chatbots for open-ended conversations because users can start the conversation in many ways [14]. The conversation can also develop in any direction. The communication with a chatbot should feel natural. Hence, it is important to balance the human and robot aspects of a chatbot [14]. If the chatbot is too human people might ask questions unrelated to the domain [14]. If it's too robot-like people might complain because the conversation feels unnatural [14]. Developers also need to think about how friendly a chatbot should be, how fast the bot should answer if the bot should have a gender, and how human-like the bot should be [14]. Successful chatbots inform the users what they have to expect and clarify that they are talking to a bot right from the start [14] Chatbots should inform users about what they can do and it can help to inform them what they can't do [14]. Conversational interfaces need to be improved based on the interactions with humans [14]. Three examples for chatbots are Microsofts Heston Bot

for food, cooking opportunities and fashion, H&Ms bot for shopping suggestions based on photos, and Ikeas shopping assistant bot [14]. Brandtzaeg and Følstad [14] focus on the psychological aspects of chatbot design and give recommendations independent from the type of chatbot. Go and Sundar [20] is about human-like bots and the effects they have on users and compares these bots with current chatbot systems. The main functions of current online chatbots are interaction with users, address of concerns, and question answering [20]. These chatbots often lack humanness when communicating with users [20]. The impersonal nature of the conversation can be countered with a high level of message interactivity [20]. To increase the humanness of a chatbot three approaches can be used [20]. A chatbot can be given visual cues like human figures, identity cues like a human-associated name, or conversational cues by mimicking the human language [20]. Introducing visual cues and a name to a chatbot is an easy task. Conversational cues are hard to create since they require deep understanding of human conversation and the context. These three approaches can be used to make people believe that they are talking to a human or a bot. The way a user interacts with a bot and the things a user expects change through this assumption [33, 20]. If a user expects a human agent then the chatbot is more likely evaluated as human-like and the conversation feels more natural for users compared to when they expect a bot [33]. If the bot is identified as a human then the users automatically expect more from the conversation and the bot needs to be capable of more otherwise the users will be disappointed [20]. If the bot is identified as bot the users expect less from the conversation but the conversation is more likely evaluated as robotic or unnatural [20]. If on the other hand, the bot is falsely identified as a human being by a user the expectation of the user regarding interactivity rises [20]. This needs to be kept in mind when designing chatbots since the false identification as a human can lead to a huge decrease in user satisfaction and acceptance. Go and Sundar [20] shows that the user assumption (bot or human) has a great impact on the chatbot evaluation and the feedback. Unlike the other references Følstad and Brandtzaeg [18], Brandtzaeg and Følstad [14], Go and Sundar [20] focus on the psychological aspects of chatbots and the communication with machines in general. Go and Sundar [20] is about the humanness of bots. It's important to keep in mind how human a bot should be and how the user assumption influences the feedback. If the bot appears to be human the expectations rise. If the bot is presented as bot right from the start the expectations are in general lower. Brandtzaeg and Følstad [14] also focuses on the psychological aspects and the humanness of bots. The disadvantages of too human bots are listed with real-life examples. Real chatbot conversations have shown that users ask unrelated and inappropriate questions if a chatbot is too human. Hence, it's important to balance the robot and human aspects when designing a bot [14]. Dutta [17] is about chatbot design aspects. The important message of Dutta [17] is that chatbot systems are completely different from regular systems from the design perspective. Right now, developers focus on the design of user interfaces (UIs). But chatbots all have the same UI requirements independent from the use case. They either use messengers like UIs for text communication or no UI at all through speech. This means that the design focus of a chatbot isn't the UI. It is the conversation. Hence, the main design focus needs to shift from UI design to conversation design.

ISO/IEC 25010 defines quality characteristics for the evaluation of software systems [6]. Functional suitability is split into functional completeness, correctness and appropriateness [6]. There are also performance efficiency metrics to check the time behavior of a system with the throughput time [6]. The interoperability of systems can be verified by checking the exchanged information [6]. Usability can be checked through learnability

which determines how easy it is to learn to use the product [6]. Security can be checked by checking for unauthorized access problems [6]. A part of portability is how easy it is to install a system [6]. ISO/IEC 25010 provides evaluation criteria for software systems which can be used to evaluate chatbot systems.

The researchers in the literature used for this thesis are mainly focused on introducing new approaches or the comparison of different types of chatbots. In some papers the performance of the chatbots regarding NLP and ML are compared. In most cases the actual frameworks are not compared regarding e.g. offered features, learnability, price. Most papers focus on the NLP capabilities and not on other aspects like usability, simplicity or the comparison of the functionality of the different frameworks. Nor is the ISO 25010[6] used which is a standard for the evaluation of software systems. A major point of this thesis is the recommendation of a framework for future developers and projects in the 3 Banken IT. For this reason aspects like learnability are considered because they are relevant for developers. A major factor for the use of chatbots is the price yet the prices for the frameworks are not compared in the articles and books used for research. These factors are included in the evaluation of the chatbot frameworks since they are relevant for the developers, the development of projects, and the company. Papers like Braun et al. [15] show that the differences of chatbot frameworks is small when it comes to NLU performance. Hence, it doesn't make a huge difference which framework is chosen when the focus lies solely on NLU. The NLU capabilities of the framework are important in general but to give a development recommendation other aspects like the available features and the simplicity of a framework are more important. If it's true that the NLU performance of the chatbots is roughly the same for all frameworks then a framework can't be chosen because of the NLU performance. Yet, it's the evaluation found in most papers if there is an evaluation at all.

Chapter 4

Methodology

4.1 Approach

The first step in the design process of a chatbot is the problem analysis. As mentioned in Følstad and Brandtzæg [18], the design focus of chatbots lies in the conversation. For this reason, a basic conversation has been created and analyzed to extract the required intents and entities. The example conversation of Figure 5.2 has been analyzed in Table 5.1 and was used to create the dialog flow-chart shown in Figure 5.3. The flow-chart has been used as the blueprint for the chatbot prototypes. All prototypes had to implement the same dialog structure to make a fair comparison possible. The extracted information was used to determine if the given problems are solvable through the use of a chatbot. The three criteria Singh et al. [32] mentioned have been satisfied by the problem description. The sickness notification has been classified as repetitive since the main process is about data collection for both problems. The chatbot needs to identify the type of the users message which can either be sickness notification or vacation request. Then the data needs to be extracted by the chatbot. For a sickness notification, the name of the person is a required input and the return date is an optional parameter. For the vacation request, the name of a person and a date-span are required. The process can be automated since the interaction can be done between the chatbot and the user without further interaction of a third party. The two use-cases are also solvable via simple back and forth communication because the chatbot just needs to classify a sentence and extract data. As mentioned in Chapter 2 Basics the classification of sentences and extraction of data are the two key features chatbots offer. The final dialog flow for the use-cases of Figure 5.3 confirmed that both of the use-cases are solvable via simple back and forth communication. There are three major chatbot categories and the sickness notification and vacation request problem have been classification as task-oriented and are tied to a domain. Task-oriented approaches have been described and used in Deshpande et al. [16], Williams et al. [35], Braun et al. [15], Williams et al. [36]. The chatbot has to be able to answer questions in the domain of sickness notifications and vacation requests. It's not required to answer any questions that are unrelated to the given domain. The chatbot needs to extract specific data from the sentences which is the second indicator that the problem is task-oriented. The third indicator for a task-oriented problem is that the chatbot has to collect information from the conversation to achieve a specific task. In the case of this thesis, the tasks are the submission of a sickness notification and vacation request to an endpoint. Hence, the task-oriented chatbot approach has been chosen for this thesis. The menu-driven approach has not been

chosen because the given problems require a mechanism to input dates and person names in full and partial sentences. This description doesn't fit the menu-driven approach since it uses numbers as input. The third major chatbot type found in literature is the open-ended approach found in Williams et al. [36], Bordes et al. [13], Rahman et al. [28]. The open-ended approach has no specific goal and is not limited to a domain. As mentioned before the problem is based on achieving two tasks. This disqualifies the open-ended approach since it's not meant to collect data to achieve a task. Furthermore, as Bordes et al. [13] mentioned, the open-ended approach is not yet able to perform well enough in task-oriented settings to be considered for this thesis. As Williams et al. [36] showed, it's possible to combine the open-ended and task-oriented approaches by adding state-related information. The hybrid approach works better than the classic open-ended approach but is not compared with task-oriented technologies. Since no comparison between the hybrid and task-oriented approach has been available solely task-oriented technologies have been considered for this thesis. For this thesis, the domain has been defined as sickness notifications and vacation requests. The chatbot prototypes function only inside of the given domain since further functionality hasn't been defined in the requirements section. The functionality only covers the two request types of sickness and vacation and basic structures like hello, bye, thank you, yes and no to build the conversations. No meaningful response can be expected if something else is used as input. There are a lot of chatbot frameworks available. As mentioned by Kane [25], the two major categories for chatbot frameworks are cloud-based solutions that are developed via a website and local standalone applications. The framework selection for this thesis was based on chapter 3 State-of-the-Art and Section 1.5 Prerequisites. Section 1.5 describe the requirements of the company and Chapter 3 shows which technologies are used in papers, articles, and books. To match the prerequisites at least one cloud (Braun et al. [15], Rahman et al. [28]) and one local (Braun et al. [15]) chatbot are needed and IBM Watson Assistant (Rahman et al. [28], Ahmad et al. [11], Godse et al. [21], Gregori [22]) needs to be taken because an IBM technology is on the wish-list. A common cloud technology is Dialogflow (Braun et al. [15], Dutta [17], Singh et al. [32], Raj [29], Rahman et al. [28], Godse et al. [21]) hence it is selected. A common local standalone technology is Rasa (Braun et al. [15], Singh et al. [32], Bocklisch et al. [12], Raj [29], Gregori [22]) and is the local tool of choice. The chosen chatbot technologies were Dialogflow, IBM Watson, and Rasa. The selection fulfilled all requirements by the company. Additionally, LUIS[35] was chosen for the NLU comparison since it has been evaluated as the NLU technology with the best performance by Braun et al. [15]. LUIS has been mentioned or was used by Singh et al. [32], Raj [29], Rahman et al. [28], Dutta [17], and Gregori [22]. To ensure a fair comparison the chatbot prototypes were implemented using the same dialog structure of Figure 5.3. An example conversation of the sickness notification task is shown in Figure 5.2 The NLU part of all four technologies has been trained with the same training data to ensure a fair comparison with the same initial state. The data used for training has been listed in Table 4.1, 10.2, 10.1, and 10.3. The prototypes have been implemented with the same intents and utterances. The training data for the prototypes has been defined by the author. It wasn't possible to use the same entities since they were provided by different technologies or not at all. Dialogflow and LUIS provided all required entities by default, Rasa needed external tools to provide the entities, and for Watson Assistant a person entity had to be created from scratch. For the German language, LUIS provided no person entity and it had to be created from scratch. This has been done to make the performance comparison as fair as possible regarding the NLU section. The same training data for all

| No | Utterance | Person Entity | Date Entity |
|----|---|---------------|-------------|
| 1 | I am sick | ✗ | ✗ |
| 2 | sickness notification | ✗ | ✗ |
| 3 | I am ill | ✗ | ✗ |
| 4 | [Alfred Mayer] is sick | ✓ | ✗ |
| 5 | my colleague [Maria Müller] is sick until [wednesday] | ✓ | ✓ |
| 6 | colleague [Simone Bauer] is ill | ✓ | ✗ |
| 7 | report [Franz Dorfer] sick | ✓ | ✗ |
| 8 | [Stefan Weber] is ill till [Friday] | ✓ | ✓ |
| 9 | sick [Emma Wagner] [6th of June] | ✓ | ✓ |
| 10 | [Sophia Richter] feels sick today | ✓ | ✗ |

Table 4.1: Sickness Training Utterances English

technologies approach has been used by Braun et al. [15], Dutta [17] and Gregori [22] and was adopted for this thesis to enable a fair comparison. The two key functionalities of chatbots have been tested inside of the given domain to answer which chatbot technology is the best to solve the given problem. The key functions of chatbots are the intent classification and entity extraction. The entity extraction capabilities have been tested by Geyer et al. [19]. The intent classification and entity extraction has been tested by Braun et al. [15] using precision, recall, and f-score values. The approach with precision, recall, and f-score has been adopted for this thesis since it defines an objective numeric value for the performance of a chatbot system and prevents subjective bypass for tests. The intent classification data used for all technologies has been listed in Table 7.10, 7.11, 10.4 and 10.5. The entity extraction test data has been listed in Table 7.5, 7.6, 7.7, 7.8, and Table 7.9. The test evaluation focuses solely on domain-specific information since the main purpose of this thesis is a proof of concept through prototyping. The result of Braun et al. [15] showed that the domain had small to no influence on the performance of the tested technologies. Hence, a good technology is expected to perform well in any domain and a bad technology is expected to perform measurably worse than a good technology in any setting. In Braun et al. [15] it was stated explicitly that different technologies should be tested with domain-specific data before one is selected for the actual product. The actual interesting performance of a chatbot is the domain-specific performance. If a chatbot technology shows great performance in general but performs badly in the given domain, it shouldn't be used for the domain. Hence, the general performance is not interesting for the development of a domain-specific bot. For this reason, only domain-specific training and test data were used. Another common testing approach for chatbots are user studies as mentioned in Shawar and Atwell [30]. User studies are recommended for customer-facing chatbot to increase the quality of the product and gather communication information between the chatbot and the users. User-studies are not part of this thesis since they are not required for a proof of concept through prototyping. A user study also doesn't provide any information about the development perspective which is a key part of the thesis. The psychological aspects mentioned by Følstad and Brandtzæg [18], Brandtzaeg and Følstad [14] and Go and Sundar [20] have been partially applied. The general rule that a chatbot should tell the user that he's talking to a bot was applied. The advanced rules for humanness are interesting for user studies and should be considered for future projects. They are not relevant for the proof of concept because there are no user studies in this thesis.

4.2 Comparison and Evaluation

A major part of this thesis has focused on the evaluation and comparison of the developed prototypes. The training and test data used for the comparison is domain-specific and no other data has been used. The training and test data sets were used for the evaluation of the entity extraction and intent classification capabilities of the framework. It was impossible to test the entity extraction capabilities of the frameworks unbiased since the person entity was not predefined for Watson Assistant and had to be created from scratch. An entity created by the developer might overfit or suffer from a small amount of training data. This is a bias in the test setting since the performance is dependent on the developer and not on the technology alone. With Watson Assistant it wasn't possible to mark entities in the training sentences when German is selected as the language. This rendered Watson Assistant useless for the German input language. The Rasa framework doesn't offer any predefined entities itself and relies on other optional technologies like Spacy and Duckling. The use of Spacy for date and date-spans was possible but the result date wasn't in standardized format. This led to the use of Duckling since it provided the necessary standardization for date and date-span entities. Furthermore, Dialogflow offered two entities that can handle date-spans while Watson Assistant offers none. For Watson Assistant two separate dates were used instead. Both date-span entities of Dialogflow were tested to ensure the best possible result regarding the tests. The entities which were defined by the problem analysis of the user-cases are the person, date, and date-span entity. Only those three entities have been included in the entity extraction tests. For the evaluation of the results the recall, precision, and f-score values were used as Braun et al. [15] did. To calculate the values the TP, FP, and FN values are counted for each individual test case. The comparison of the frameworks has been difficult since some entities are not available using certain technologies. Watson Assistant and Duckling had no person entity and LUIS had no person entity for the German language. In the case of Dialogflow two entities were available that could be used for the date-span extraction. The available entities were listed in Table 7.3. To make the values comparable the result was calculated for each individual entity, the combination of date and date-span, and the combined score over across all entities. The second major function of chatbots is the intent classification. The intents were trained with the same utterances for all four technologies to create a fair test setup. To measure the performance the precision, recall, and f-score values were used to provide an objective value. As mentioned in Chapter 2 Basics, the NLU services of all four frameworks provide a confidence value. The confidence value is a response that represents how well the input fits an intent. The result evaluation of this thesis has been extended with the confidence score since it is an indicator of how good the training worked for the frameworks and if additional training data would have been necessary. The additional value standard deviation has been added to compare how consistent the confidence scores of the chatbots were. The confidence scores were shown for the correct intent no matter if the intent was identified correctly or not to calculate an average confidence score across all tests. The average confidence score provides another quality metric. A high average confidence score combined with a high f-score indicates that the chatbot classifies intents reliably and vice versa. A small standard deviation in combination with high average confidence indicates that the chatbot classifies intents reliably with a high confidence score. The intent classification was tested for English and German. This was necessary since the 3 Banken IT has its locations in Austria which makes German the language of choice for potential users. The expectation for the German tests was that there is no difference to

the English tests which was not met since e.g. LUIS had no person entity for the German language. The performance expectation was that the German bot is on the same level as the English bot regarding entity extraction and intent classification. As Singh et al. [32] mentioned, the best evaluation criterion is if the underlying task can be achieved or not. This evaluation criterion has been adopted for this thesis since it's the most relevant for a recommendation of a technology. As a result value for this evaluation, a binary boolean criterion was chosen. Either it was possible to build a prototype that implemented the design or it was impossible. The evaluation didn't include an opinion or how hard the implementation was with a design. Since LUIS is an NLU technology and not a chatbot technology it was excluded from this test because it offers no dialog handling mechanism. It would have been unfair to include an NLU technology since it's not meant to handle dialogs. The Microsoft Bot Framework uses LUIS and offers dialog handling but wasn't used in this thesis and is therefore not part of this test. For future project speech recognition capabilities are on the wishlist of the 3 Banken IT. For this reason the speech-to-text and text-to-speech services offered for the chatbot technologies were compared based on their price. Chatbots which offer speech functionality can be used through voice interfaces in the future. The second evaluation criterion in the speech recognition category was if speech services are offered by the provider. It was expected in general that each framework offers speech services for their chatbot technology since these are two closely tied topics. A key factor for the use of chatbots is the price. That the price plays an important role was mentioned in Raj [29] for instance. None of the articles used for this thesis compared the prices of the available technologies. The goal of the price comparison was to find out which technologies are cheap and which are expensive. In combination with the other evaluations, the expectation was to find a cheap technology which performs well. The price of a chatbot has been calculated for different categories since the prices were intransparent and hard to compare. The first category was the price for the chatbot technologies, the second one was the price for hosting, and the third category was the price for speech-to-text and text-to-speech services. The prices for the chatbot technology for the enterprise edition of Rasa and Watson Assistant is not listed on the website since it's calculated individually. The prices for the Microsoft Bot Framework were calculated by adding the price for the Microsoft Bot Framework and LUIS. The prices for Watson Assistant are calculated per user per month while the price for Dialogflow, LUIS, and the Microsoft Bot Framework was based on the number of requests. Rasa is free by default and was excluded from the price calculation since no prices were listed on the website. The prices for the text-to-speech and speech-to-text services were comparable for Watson Assistant, Microsoft Speech Services, and Dialogflow. Rasa didn't offer text-to-speech and speech-to-text services at all. The general expectation was that the frameworks cost roughly the same but the prices were hard to compare or not comparable at all. Another evaluation criterion used for this thesis is the portability of software systems as defined in the ISO 25010[6] standard. The portability was chosen since the comparison of the cloud technologies with a local technology was interesting from a development point of view. The first factor discussed in this thesis were the necessary tools since they determine how portable the development environment of the chatbots are. The portability of the development environment of the local technology Rasa in contrast to the cloud chatbot technologies was the main focus of this evaluation. The second factor chosen was the number of providers where the chatbot can be deployed. The main focus lay on cloud technologies since the expectation was that they can only run in the environment of the provider and are not portable at all. The third portability category chosen was how

flexible the development of chatbots is regarding the devices. The expectation was that cloud technologies are developed in a web browser and can be developed from almost anywhere while the local technology requires a development environment and is comparably unflexible. Another aspect of portability is the migration from one technology to another. It was expected that all technologies save the data differently and that it's not easy to migrate between them. For the development of a chatbot, the project setup complexity plays an important role and was evaluated in this thesis. The main focus of this test was Rasa since the setup complexity was interesting for the local technology. The setup of the cloud technologies was expected to be easy since the providers take care of the setup for the customers. The deployment of chatbot systems is an important aspect when local and cloud technologies are compared. While it was expected that a cloud technology produces no deployment overhead the opposite is expected for a local technology since the deployment needs to be done by the developer. For the local technology, the Docker deployment and setup of a local Python environment were included in the evaluation. Especially the deployment using Docker has proven as interesting for future projects because of the flexibility and simplicity. The ISO 25010[6] defines learnability as sub-criterion of usability as measurement metric for software systems. The learnability is important from a development perspective since an easy to learn and use technology is the optimal case when the performance is good. The evaluation of the project setup and deployment complexity was included in the learnability since they are part of the development of a system. As mentioned before, the chatbot technologies use entity extraction to extract data from text. The entity extraction is a part of the development process of a chatbot and it takes time to define good entities. All of the chatbot technologies offer predefined entities that are ready to use. As a criterion, the number of predefined entities was defined as a measure since lots of predefined entities can save lots of development time for future projects. If a required entity is available for a given problem the developer doesn't need to create one. The second measurement criterion is the entity category if all required entities for the use-cases of this thesis were predefined by the chatbot technology. The technologies where all required entities were predefined are better suitable to solve the given problem. It was also evaluated how hard it is to create new intents, entities, training phrases, and dialogs. The expectation was that it is equally complicated for each technology. As mentioned by Følstad and Brandtzæg [18] the design of chatbots is all about the conversation. Hence, the most important aspect was the creation of dialogs. The frameworks Dialogflow and Watson Assistant offered dialog nodes to create and structure conversations. A dialog node is one step of the conversation and can be followed by another dialog node. A dialog node was e.g. the classification of the sickness notification intent followed by the collection of the name and return date to the response of the system. Rasa offered stories to represent a dialog. A story represents a whole conversation path from start to finish. The start was a greeting line and the end the goodbye response for instance. An important part of the conversations required for this thesis was the extraction of form data and has been evaluated regarding the complexity. Rasa was the only technology tested where programming skills were necessary to extract form data from the conversation. For developers, an important factor are the customization possibilities. It's an advantage for when the pipeline can be fine-tuned to increase the performance or add special custom components. The only tested framework which offered pipeline customization was Rasa. A developer can even create new custom components and include them in the pipeline. This is interesting for very specific company requirements, research, and fine tuning to increase performance. The other tested technologies didn't even display the configuration

of the pipeline. When Rasa is used for development the pipeline is adjustable to the developers needs. For the development of the prototype with Rasa the technologies Spacy and Duckling were added to the pipeline for entity extraction. This was not possible with the other frameworks since the pipeline can't be adjusted at all. The stories used by Rasa were adjustable fine granular compared to the dialog node approach of Dialogflow and Watson Assistant. A story could be used to execute multiple actions and display multiple responses in a row. Watson Assistant can't execute multiple actions but can call the webhook where the request can be processed. Dialogflow can't execute multiple actions or display multiple responses. With Rasa, the default form action was implemented using Python while no programming skills were required for form actions using Watson Assistant and Dialogflow. The advantage of a programming language for form actions is that the concepts of a programming language are available for e.g. the validation of entries and calling external services. Hence, a comparison of the programming based approach with the cloud approach was included in this thesis. As discussed in Chapter 2 Basics, the concepts used by the frameworks are interesting. The base concepts intent, entity, and utterance are present in literature and it was expected that these concepts are also present in the frameworks. Another important aspect is the communication with the chatbot services. In general, JSON is the request and response format for all tested technologies. All the technologies offered a REST interface for communication. The messages differed when the content was compared since LUIS and Dialogflow sent metadata while Watson Assistant and Rasa didn't. These are two very different approaches which were evaluated for positive and negative aspects. The meta-information can be used in the backend for future projects to create e.g. statistics. The communication can also be seen from the other side when a service sends additional information to the chatbot. It was evaluated how easy it was to process additional information in a chatbot. The expected result was that Rasa is the easiest technology for processing information since it's the only technology which uses a programming language for data processing. Another big difference between Rasa and the other technologies were the training capabilities. Live training is offered by Rasa while all the other technologies were solely trained with the training data defined by the developer. When live training is used the created stories are expected to be closer to a real-life scenario because the developer has to go through the conversation step by step. This can prevent unnatural conversations since the developer creates the conversation step by step. In the ISO 25012[6] defines the user interface aesthetics under the section usability. Hence the user interfaces of the frameworks were also compared regarding simplicity. This was evaluated from a subjective point of view. The interface of a framework was evaluated as easy to use when the overall structuring was easy to understand and the prototype was easy to implement.

4.3 Course of Action

After the problem had been analyzed the dialog flow-chart was created. Based on the flow-chart the required intents and entities were defined. The required intents were sickness notification and vacation request. For each use-case one intent was necessary. The conversation built in the flow-chart made the required entities visible. For the sickness notification the name of a person and an optional return date were required. The vacation request needed the name of the person and a date-span. The resulting entities were person, date, and date-span. For each intent, 10 training sentences were created in English and 10 were created in German. The test sentences were used to develop a German and English

prototype with each technology. For the entities and intents, test sentences were defined. The entity test sentences were evaluated as TP, FP, TN, and FN. Based on these values the precision, recall, and f-score was calculated to provide an objective evaluation. The same approach was used for the evaluation of the intent classification plus the additional properties for the confidence and standard deviation. For each test category as summary was created to compare the performance of the technology for each category. The frameworks were evaluated based on the criteria listed in Section 4.2. The results were used to rank the frameworks in different categories which show the strengths and weaknesses to make a final recommendation for further projects. Furthermore, the entity extraction and intent classification results in English and German show which technology performed best in the given domain.

Chapter 5

Design

5.1 Architecture

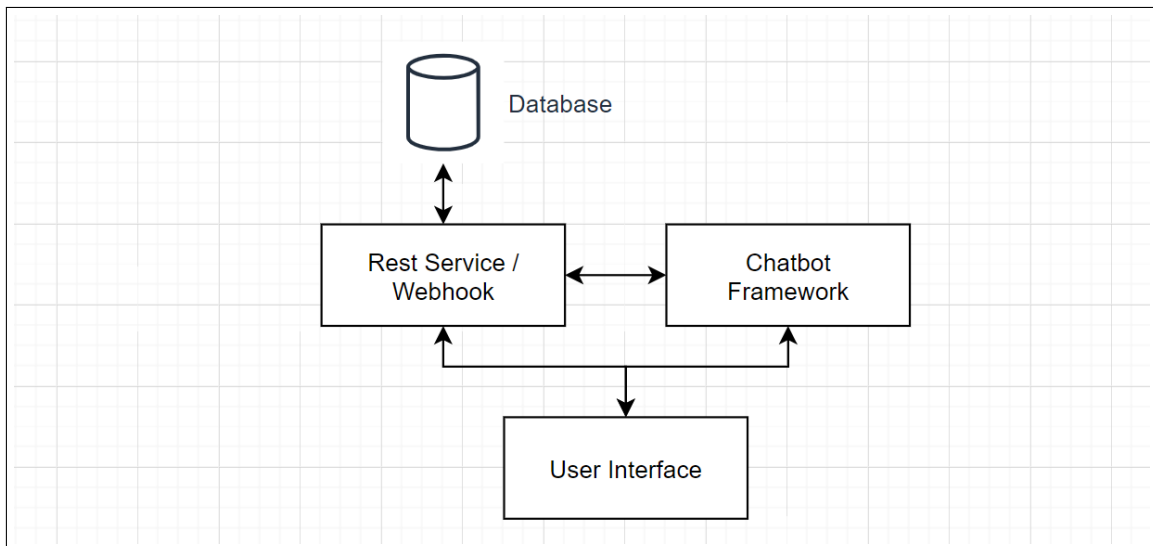


Figure 5.1: Minimalistic Chatbot Architecture

One possible architecture for a minimalistic chatbot system with database communication is shown in Figure 5.1. The chatbot framework takes care of all conversation related tasks. Conversation related tasks are the intent recognition, entity extraction, and the general flow of the dialogue for instance. The user interface of Figure 5.1 fetches the input from the user, submits it to the chatbot framework and displays the response messages. The UI can additionally communicate with the database via the REST service or webhook if necessary. The DB handles data-related tasks like the possible items and can be used for the validation of the extracted entities by the chatbot framework in combination with the webhook. The webhook separates the data logic from the application logic and is used for validation, data submission and retrieval. It's also used for other chatbot related issues like the validation of intent related data and error responses to the chatbot framework.

5.2 Story/Flowchart

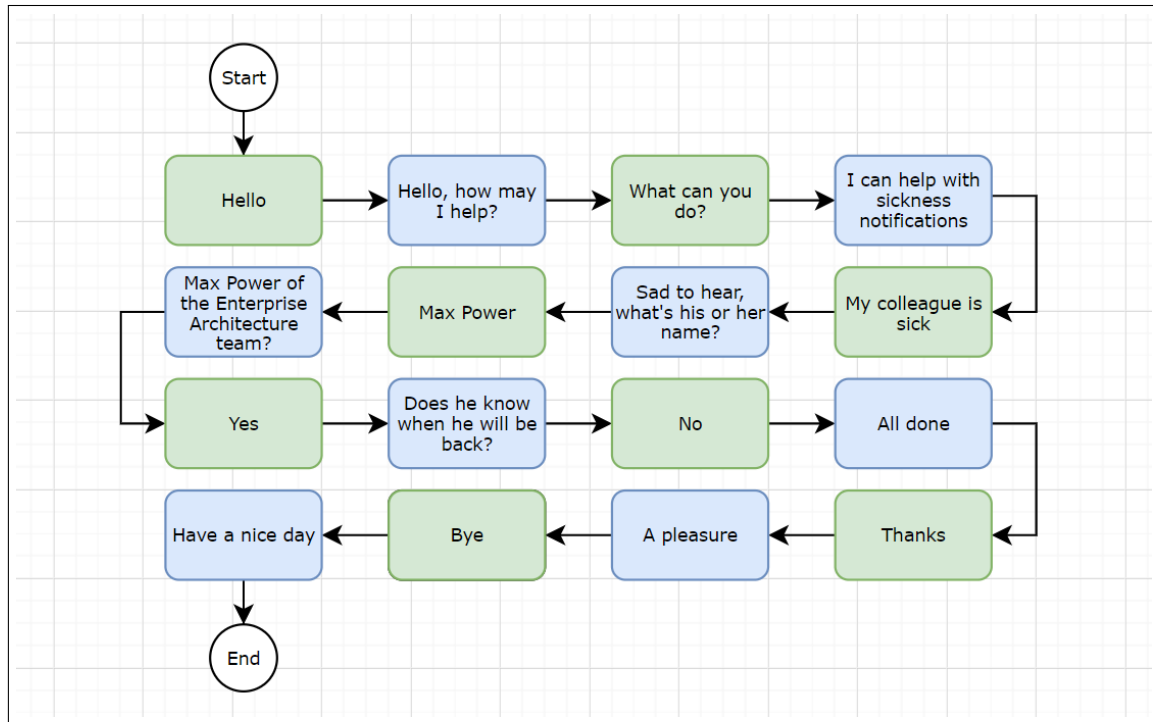


Figure 5.2: Sickness Notification Example Conversation

As mentioned in Følstad and Brandtzæg [18] the design focus for chatbot systems lies on the conversation itself. A story is an example of a dialog/conversation between user and chatbot. An example dialogue to report a colleague sick is shown in Figure 5.2 where the user input is highlighted green and the chatbot response is highlighted blue. The intent of each user input needs to be found to create the knowledge base for the domain. The intent for the user input "Hello" can be marked as greeting, the second input is the general information what the chatbot can do, the third input is the sickness notification intent. Then the necessary information is collected followed by some chit-chat and the end of the conversation. The interesting entities, in this case, are the name of the employee which is required all the time, the confirmation that the found employee is the correct one and the return date which is optional since it's not possible to tell in every case. The resulting intents, entities, and utterances for the example sentences of the conversation shown in Figure 5.2 are listed in Table 5.2. The detailed conversation structure for the use cases of this thesis is shown in Figure 5.3. The figure shows the flowcharts for the sickness notification and vacation request. User inputs are highlighted red, bot responses are green and general actions are marked purple. Both use cases start the same way because the person needs to be identified. If the entered name can't be validated the user can correct the name or enter a different name. Then the user needs to confirm that the correct person was found. If the user responds with no the name of the person can be corrected. Afterward, a date needs to be entered in both cases. The sickness intent asks for an optional return date since the yes and no path lead to the same action and the conversation ends afterward. The vacation intent requires a start and return date. The user is asked for the date until a valid date has been entered. After both dates have been entered the vacation request can be processed and the conversation end is reached.

Figure 5.3 describes every step of the conversation from start to end. Hence, it defines the functionality of the chatbot and the flow of the conversation. Table 5.1 defines the intents, name of the intents, entities, the data types of the entities, and the names of the utterances based on the information present in Figure 5.3. The combination of Figure 5.3 and Table 5.1 is the blueprint for the implementation of the chatbot. It is valid for tools using the concepts intent, entity, and utterance.

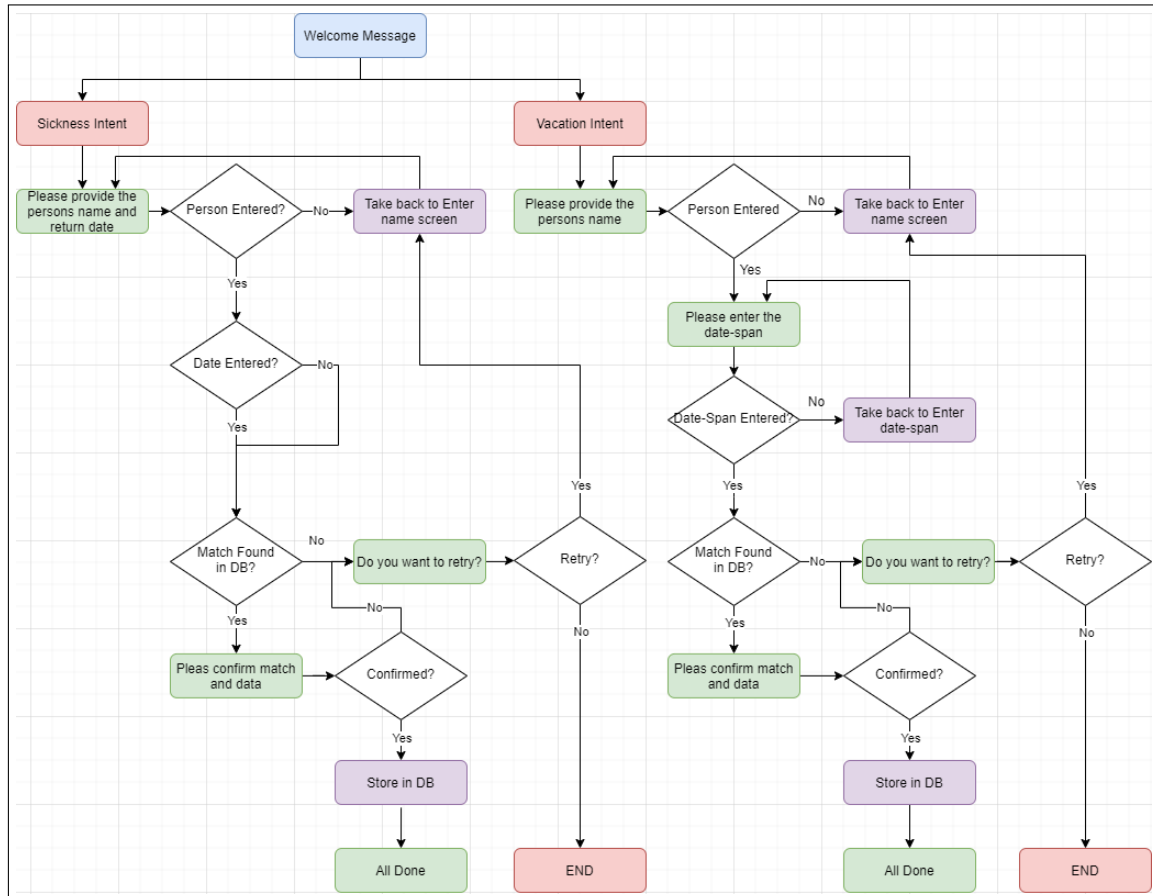


Figure 5.3: Chatbot Conversation Flowchart

5.3 Intents, Entities and Utterances

The base for this section is Figure 5.2 which shows an example conversation for a sickness notification from start to end. The identified information from the conversation is displayed in Table 5.2. In general, the greeting and goodbye intents are expected to be in almost any chatbot conversation. A thank you intent is also very likely. The user should have the option to ask for the available options. The core intent of the whole process is the sickness intent. To do the sickness notification the user needs to enter the name of a person. If the person can't be found the user needs to be notified and the operation is aborted. If the person was found the user needs to confirm that the right person was found. In production, there could be multiple employees with the same name. Because of this ambiguity, the confirmation is part of the dialog. Then the optional return-date needs to be extracted. It's optional because especially in the case of a sickness the return date is often unknown. The other case is e.g. a broken arm where the employee might return

| Intent | Reference from Figure 5.3 | Type | Utterances and Entities |
|-----------------|---------------------------|-----------|-------------------------|
| - | Welcome Message | Utterance | Utter_Welcome |
| Sickness Intent | provide sick persons name | Utterance | Utter_Enter_Sick_Name |
| | person name | Entity | Person |
| | Confirm person name | Utterance | Utter_Confirm |
| | enter return date | Utterance | Utter_Enter_Return |
| | return date | Entity | Date |
| Vacation Intent | all done | Utterance | Utter_All_Done |
| | proivde persons name | Utterance | Utter_Enter_Name |
| | person name | Entity | Person |
| | Confirm person name | Utterance | Utter_Confirm |
| | enter start date | Utterance | Utter_Enter_Start_Date |
| | start date | Entity | Date |
| | enter return date | Utterance | Utter_Enter_Return_Date |
| | return date | Entity | Date |
| | all done | Utterance | Utter_All_Done |

Table 5.1: Analysis of Figure 5.3

tomorrow. Then the sickness notification is done. In the case of this thesis, the chatbot is the important part hence the backend is mocked for simplicity. The user might thank the bot or end the conversation with a goodbye message, or the user might not respond after the all done message. These are the three possible ends of a conversation. The basic concepts extracted from Figure 5.2 are shown in Table 5.2. There are the intents discussed above, the entities and their types, and the bot actions. For this use case, the needed entities are a person or the name of a person, a confirmation of boolean type, and a date. Alternatively for the confirmation, the name of the department the person works in can be used. The department is modeled as a custom entity as described in Chapter 2: Basics in Section 2.6. The bot responds to the user with utterances. The conceptual names of the utterances are shown in Table 5.2 under Bot Action. In the simplest case, a bot action is an utterance. For the utterances, different versions need to be written to respond to the user with alternating sentences. For the vacation notification, the process remains the same. The greeting, thanks, goodbye, and list options intents remain the same. The new intent is the vacation intent. To validate a vacation request the name of the person and the department are needed again. The only change in the process is that there needs to be a mandatory start and end date. There are no new datatypes/entities needed for the second use-case.

5.4 Psychological Aspects

Some psychological aspects of Brandtzaeg and Følstad [14] have to be included. Brandtzaeg and Følstad [14] stated that the user impression depends greatly on the users mindset. If the user thinks he's talking to a bot the user expects less but the conversation is more likely evaluated as unnatural [14]. If the user assumes he's talking to a human and the bot is unable to achieve the task the user is very disappointed [14]. Since the problem is a task-oriented one the general conversation capabilities are not the primary focus. This means the bot is unable to fake a human conversation which goes beyond the expected

| Sentence | Intent | Entity | Bot Action |
|----------------------|--------------|---------------|-----------------------------|
| Hello | Greet | - | Utter Greeting |
| What can you do? | List Options | - | Utter Options |
| my colleague is sick | Sick | - | Utter Ask Name |
| Max Power | - | Person Name | Utter Confirm |
| Yes | - | Boolean | Utter Return Date or Failed |
| No | - | Optional Date | Utter Done or Failed |
| Thanks | Thanks | - | Utter You'r Welcome |
| Bye | Goodbye | - | Utter Goodbye |

Table 5.2: Sickness Notification Intents, Entities, and Actions of Figure 5.2

functionality. Hence, the user needs to be informed that he's talking to a bot. For further development steps, a gender can be introduced to make the bot feel more human-like if necessary. Another way to give the bot a more natural feel is to look at the message response times. They should not be too fast. If the response comes too fast it feels unnatural because a human being can't answer as fast as a machine.

5.5 Webhook

As explained in Chapter 2: Basic in Section 2.9 the webhook is an endpoint where services can send requests to. In this thesis three frameworks are present and each framework gets an endpoint. The service is also used to access general information about the state of the data. The Dialogflow endpoint is reachable under `"/dialogflow/webhook"`, the Watson endpoint is located at `"/watson/webhook"`, and the Rasa service submits post requests to `"/sick"` and `"/vacation"` to submit data to the REST service. For development purposes, the last request of Watson can be viewed at `"/watson/request"`, and the last request of Dialogflow at `"/dialogflow/request"`. To see all successful sickness notification entries `"/sick/all"` can be used, and under `"/vacation/all"` all successful vacation entries can be viewed. The service fakes a database and stores the request. After the information has been collected successfully by a framework a request is submitted to the service. The service can do everything with the data and is the point where the company can execute actions. All frameworks can call the webhook at the end when the data has been collected but also after each user request if necessary. This way the information can be validated and modified by the service.

5.6 Messenger UI

The design of a chatbot focuses on the conversation and not on the UI. The UI requirements are always the same for chatbots. The basic features required are writing a new message, sending a message, and displaying a message. The messenger UI shown in Figure 5.4 can enter and send a message, and displays message cards. A user can enter a new message at the bottom and click the send button to send a message. Brandtzaeg and Følstad [14] states that successful chatbot should inform users that they are talking to a bot. Therefore, the heading has been adjusted to remove ambiguities and includes the term chatbot. The messenger UI is shown in Figure 5.4.



Figure 5.4: Chatbot UI

Chapter 6

Prototype Implementation

6.1 Dialogflow

Dialogflow[4] formerly known as API.ai is the chatbot framework of Google and is hosted in the Google cloud. To create a new agent the main language needs to be selected and the name of the agent needs to be defined. The selected language is the interaction language of the users. Dialogflow offers the option to create a mega agent by combining other agents. The web interface of Dialogflow is shown in Figure 6.1 where the menu shows the intents and entities tabs. Dialogflow supports the basic concepts of intents, entities, and utterances. Hence, Table 5.1 is used for the implementation of the intents, entities, and utterances.

6.1.1 Intents and Entities

When a new intent is created some options are present. A Dialogflow intent has training phrases, actions, parameters, a response list, a context, and the fulfillment. The training phrases are a list of utterances a user would enter. Parameters are the slots that need to be filled in the conversation. The parameters for the sickness intent are the rows of Table 2.2 where the entity cell is filled with text. The response list is a list of utterances which is used to respond to the users question. The context is used to pass parameter values to other intents. In the fulfillment section, the webhook can be enabled. By default, an intent is handled by Dialogflow without communication to another service. In the fulfillment section, a webhook call can be enabled. For the sickness intent, two parameters are required. The name of the person is a non-optional parameter, the return date is an optional parameter. Dialogflow provides many predefined entities and intents. The date and person entities are predefined by Dialogflow and it's not necessary to create them. Additionally, yes and no intents are also available and can be used and adjusted for the current sickness notification use case. The final structure is a sickness intent which is followed by either the yes or the no intent. For this, the follow-up intent concept is used where follow-ups can be entered. At the end of the no intent the chatbot responds with a retry response. At the end of the yes intent, the bot responds with an all done response. This ends the conversation of the sickness use case and the yes and no intent are marked as the conversation ends. The process looks the same for the vacation use case. The vacation intent needs to be created and has a follow-up intent yes and no to confirm the user input. The entities required are a person and a date-span. Dialogflow offers a predefined entity called date-period which offers this functionality and it's not required to create one. The

date and date-period entities offer advanced features. A valid date period is for instance Monday to Friday. This is translated to the correct values if e.g. Monday is in the past the resulting date period will be from the next Monday to Friday which follows next Monday. For the example sentences, the entities can be used to create alternative ways to fill the slots. The user can be prompted slot after slot or enter them in the first request to speed up the process for more experienced users. In a sentence like "Max Power is sick and will return on Monday" all the required information is already in the sentence. The entities need to be defined in the training sentences. Then the system learns that there are multiple ways to acquire the information. With suitable training sentences, the chatbot can extract the information directly and jumps directly to the confirmation. In general, custom entities are defined as a list of entries or a list of entries with synonyms. An advanced feature of Dialogflow is small talk which can be enabled in the settings. When small talk is enabled the bot is able to respond to requests like "What are you", "Bye", and "sorry". This is another useful feature which is predefined and requires no development effort and can be enabled with one click.

6.1.2 Webhook

The Dialogflow webhook requires a HTTP POST URL as an entry point. The communication relies on JSON as a format. The entry point of the webhook is shown in Listing 6.1. A request like in Listing 2.1 can be accessed like a dictionary or an array shown in Listing 6.2 where the name of the intent gets extracted. After the intent has been extracted the parameters are extracted to process the request. In the error case, a new fulfillment message is returned. A fulfillment message as return value prevents follow up intents. In the success case, an empty response is returned and the bot continues normally. In Listing 6.3 a skeleton for intent handling is shown for a success and error scenario for an intent called "Sickness Confirmed".

```
1 @restService.route('/dialogflow/webhook', methods=['POST'])
2 def dialogflowRequestEntryPoint():
3     return asJsonResponse(dialogflowRequestHandler()), 200
```

Listing 6.1: Dialogflow Webhook Entry Point

```
1 req = request.get_json(force=True)
2 intent = req.get('queryResult').get('intent').get('displayName')
```

Listing 6.2: Dialogflow Request Parameters

```
1 if intent == 'Sickness Confirmed':
2     params = req.get('queryResult').get('outputContexts')[0].get('
        parameters')
3     employee_name = params.get('employee').get('name').lower()
4     return_date = params.get('return_date')
5     if !employeeExists(employee_name):
6         return {
7             "fulfillmentMessages": [
8                 {"text": {"text": ["Employee not found. Enter sick to retry."]}}
9             ]
10        } # RETURN ERROR MESSAGE
11    else:
12        db.store_sickness(employee_name, return_date) # PROCESS DATA
13    return {} # CONTINUE NORMALLY
```

```

14 elif intent == 'Vacation':
15     ...

```

Listing 6.3: Dialogflow Intent Handling

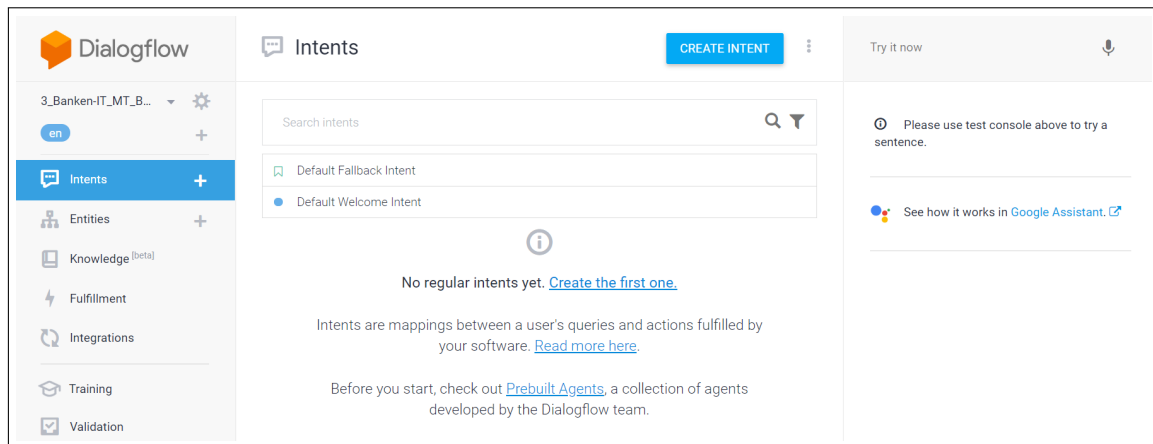


Figure 6.1: Dialogflow Web Interface

6.2 Watson Assistant

Watson Assistant [10] is the cloud chatbot technology of IBM hosted in the IBM cloud. An IBM cloud account is required to use the services. First, a new assistant needs to be created which requires a name. Then dialog skills can be created. For the sickness notification process, a sick dialog is created. The web interface of Watson Assistant is shown in Figure 6.2. The interface offers the concepts intent, entity, and dialog.

6.2.1 Intents and Entities

The required entities are person and date. Watson Assistant offers system entities for persons and dates. In the settings under options, a webhook URL can be entered. The person entity is marked deprecated and is available in English but not in German. This means a person entity needs to be defined as a replacement for the person entity using entity annotations like recommended by IBM. To do this every occurrence of person names in the test sentences need to be highlighted and marked as person. For the sickness use case, a new intent needs to be created. Then the training phrases need to be added. It is possible to create new entities by using examples or regular expressions. After the intent had been defined a new dialog needs to be created. An intent needs to be selected for a dialog. The dialog is triggered when the intent condition is met. In this case, the intent is the sickness intent. In the settings menu, the dialog can be customized. A customization is necessary to enable slot filling and the webhook call which are both necessary for the sickness use case. The required slots are person of type person and the return date of type date. The person is mandatory and the date is optional. After the information has been collected the webhook is called.

6.2.2 Dialog and Webhook

The format of the webhook call is JSON. The parameters need to be specified by hand. The required parameters are the name of the intent, the person name, and the return date. The name of the intent is necessary to execute the correct action at the webhook. The name and return date are the information that needs to be stored. It can be seen that no meta-information is sent in the request of Listing 6.4. The information and format submitted to the webhook is defined by the developer. The returned information from the webhook is stored as a context variable in Watson Assistant and can be accessed like a struct in C++. The format for the webhooks response is also up to the developer. Based on the webhook response the dialog can go different ways. If the person was found and a date was entered the date needs to be displayed in the chatbot response. If the person was found but no date was entered the chatbot response needs to ignore the date. Whenever a person was found by the webhook the user needs to confirm that the entered information is correct. Whenever the person wasn't found the user needs to be prompted to retry. These cases need to be implemented with the conditional statements of Watson Assistant. Table 6.1 shows the conditions, responses, and actions that need to be implemented for the use case. Chatbot responses can be picked sequentially, at random, or as a multiline response. The multiline response displays all response lines at once. The sequential version picks them in order and the random variant picks the response at random from the available responses. The confirmation logic needs to be implemented next. No system entity or dialog for the confirmation logic is present. This can be implemented with two intents or an entity. An entity with two possible values is a logical approach. One value is true the other is false. The entity is called confirmation. For these two cases, synonyms are defined like confirm, yes, and OK for true. When the entity is detected the correct path can be selected with a condition. Similar conditional logic was used for the webhook response shown in Table 6.1. A new dialog node needs to be created for the confirmation or abort logic. There is a jump option in the settings of the dialog node. In the success cases of the sickness dialog a jump to the newly created confirm or abort dialog needs to be entered. A slot needs to be filled and a webhook call is necessary to store the collected information. These two settings also need to be enabled for the confirm or abort dialog. The slot is of type confirmation entity and is required. After the information has been gathered a request is sent to the webhook. The two possible scenarios are that the information is stored or something goes wrong. If the information has been stored successfully an all done message is displayed. Else a generic message is displayed indicating that something went wrong. This ends the sickness dialog. For the vacation use-case, a date span is needed. No predefined entity exists for date-spans but a date entity is present. It's not possible to create entities containing entities with Watson. The options are to create an intent with two dates in it or a dialog node which fetches the required data. The dialog node approach is chosen because it comes close to the behavior implemented with Dialogflow. A new dialog node named vacation is created. For the dialog node, an intent called vacation needs to be created with the example sentences. The dialog needs a mandatory person slot and two mandatory date slots. The start and return dates of the vacation. After the information has been collected the information can optionally be validated through the webhook. Then the user needs to confirm the information like in the sickness use-case. Then the final request is sent to the webhook which stores the collected information. The final all done message is displayed and the vacation use-case is finished.

| Condition | Chatbot Response | Jump To Action |
|-------------------------------|---|------------------|
| \$response.OK && \$date!=null | Please confirm \$person is sick until \$date. | Confirm or abort |
| \$response.OK | Please confirm \$person is sick. | Confirm or abort |
| \$response.ERROR | \$person not found. Enter sick to retry. | Wait for input |

Table 6.1: Conditional Responses of Watson Assistant

```

1 [{
2     "intent": "Sick",
3     "person_name": "Anna Nass",
4     "return_date": "2020-05-27"
5 }]

```

Listing 6.4: Watson Assistant Request Format

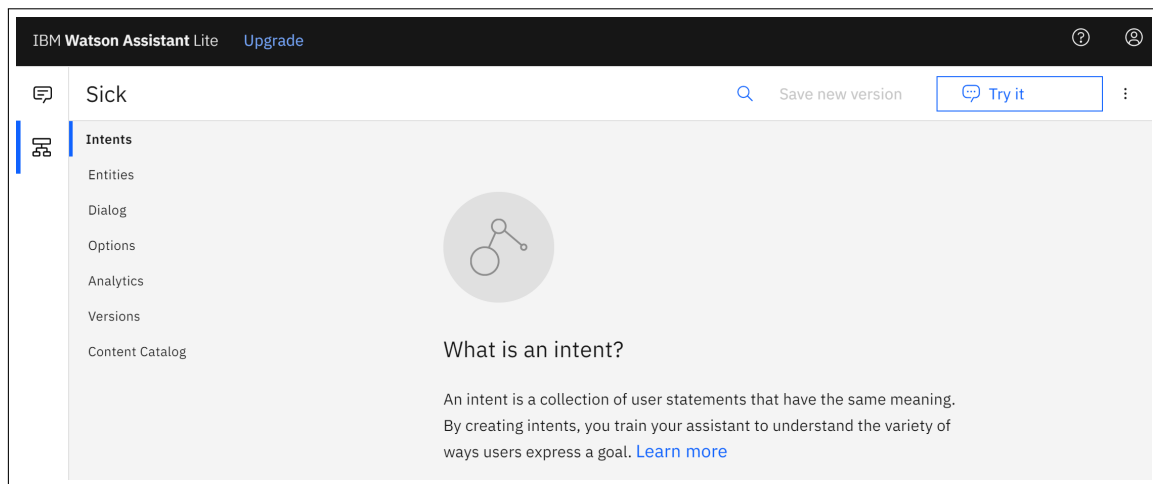


Figure 6.2: Watson Assistant Web Interface

6.3 Rasa

Rasa is an open-source framework for NLU and chatbots. Rasa[8] can either be used as a Docker container or it can be installed locally. Rasa uses Python as a development language and can be installed using pip. A Docker compose file for Rasa and the action server is shown in Listing 6.5. Rasa is split into a natural language and logic part. The natural language part takes care of everything which is related to the conversation like intents, entities, the story, and the conversation flow. The second part handles the slot filling and the actions. For each part, a Docker image can be created. The Docker compose file of 6.5 can be used in similar fashion for the Azure deployment in combination with the images. Azure also offers a Docker registry to upload and pull custom images. It's important to choose a compatible version of rasa and rasa-sdk. The Rasa image is responsible for the NLU part and the Rasa sdk runs the actions server.

```

1 version: '3.0'
2 services:

```

```

3   rasa:
4   image: rasa/rasa:1.4.6
5   ports:
6     - 80:5005
7   networks:
8     - app_net
9   volumes:
10    - ./:/app
11  command:
12    run
13    --enable-api
14    --cors *
15
16  action_server:
17  image: rasa/rasa-sdk:1.4.0
18  volumes:
19    - ./actions:/app/actions
20  expose:
21    - 5055
22  networks:
23    - app_net
24 networks:
25   app_net:

```

Listing 6.5: Rasa Docker Compose File

The information for a Rasa chatbot is split in multiple files. The `nlu.md` file stores the intents with the utterances used for training. The intent format is shown in Listing 6.6.

```

1  ## intent:sick
2    - My colleague [Max Power](PERSON) is sick until [Friday](date)
3    - somebody is sick
4    - a colleague is ill today

```

Listing 6.6: Rasa Intent Format

The `stories.md` file stores the stories. A story defines the flow of the dialog. The sickness story is shown in Listing 6.7.

```

1  ## Sickness Affirm
2  * sickness
3    - sickness_form
4    - form{"name":"sickness_form"}
5  * submit{"PERSON":"Anna Maria Mayer"}
6    - sickness_form
7    - slot{"PERSON":"Anna Maria Mayer"}
8  * submit{"time":"2020-05-30T00:00:00.000-07:00","DATE":"today"}
9    - sickness_form
10   - form{"name":null}
11   - slot{"time":"2020-05-30T00:00:00.000-07:00"}
12 * affirm
13   - action_submit_sickness_data
14   - action_restart

```

Listing 6.7: Rasa Story Format

All the actions are defined in `actions.py`. The sickness form slot filling is an action for instance. The intents, entities, and actions need to be listed in the `domain.yml` file. Otherwise, they can't be used. The defined actions can be simple utterances or something

complex like the slot filling of a form. In utterance actions, the response is chosen at random. In the endpoints.yml file, the action server needs to be entered to use the actions server for slot filling. The actions server runs separately from the NLU service at the standard port 5055. The NLU service is accessible at default port 5005. Rasa offers great freedom when it comes to configuration. This is important because Rasa does not offer predefined entities as such. But spacy does offer the required date and person entities for the use cases. In the config file shown in Listing 6.8 the pipeline includes the spacy entity extractor to enable the use of predefined entities. spa [9] lists the supported entities under named entity recognition. The entity recognition of spacy does work but the dates are stored as they are. The best case is when the date is converted in a standard format like Dialogflow and Watson Assistant do. To use a more powerful date parsing the Duckling[5] Rasa Docker image is used. Duckling converts entered dates and times to a standard format as the cloud services do. To add duckling to the project it needs to be added under pipeline in the config file. It recognizes strings like "next Monday" as dates and converts them correctly. The person entity of Spacy works as expected. The used entities need to be added under entities in the domain.yml file. People with machine learning experience can also define their own pipelines or adjust parameters. The detailed pipeline is listed in the appendix in Listing 10.1. The config file defines policies. The form policy enables the use of forms and the fallback policy enables a fallback action which is triggered when the confidence score is below the defined threshold. In the sickness use case, a generic utterance is the fallback response.

```

1 language: en
2 pipeline:
3   - name: "SpacyNLP"
4   - name: "SpacyTokenizer"
5   - name: "SpacyFeaturizer"
6   - name: "RegexFeaturizer"
7   - name: "DucklingHTTPExtractor"
8     url: "http://localhost:8000"
9   - name: "SpacyEntityExtractor"
10  - name: "EntitySynonymMapper"
11  - name: "SklearnIntentClassifier"
12 policies:
13   - name: MemoizationPolicy
14   - name: TEDPolicy
15   - name: MappingPolicy
16   - name: FormPolicy
17   - name: FallbackPolicy
18 nlu_threshold: 0.4
19 core_threshold: 0.3
20 fallback_action_name: action_default_ask_rephrase

```

Listing 6.8: Rasa Configuration

The general command required to work with Rasa are train, shell, run actions, and x. To train Rasa the command "rasa train" is used. The shell command starts an interactive shell where the bot can be tested. The action server needs to be run separately. If Rasa-X is installed the command "rasa x" starts a web UI shown in Figure 6.3. The web interface offers the discussed files in the training section with additional features like the conversation graph (right) called flow for a story (middle) shown in Figure 6.3. Rasa X offers interactive training where the chatbot is built through live interaction with the bot. Rasa stores a model every time it's trained. These models can be exchanged in

the models section. The whole history is stored which can be used for features. If the feature does not work out the model can be set back to the original version. For the

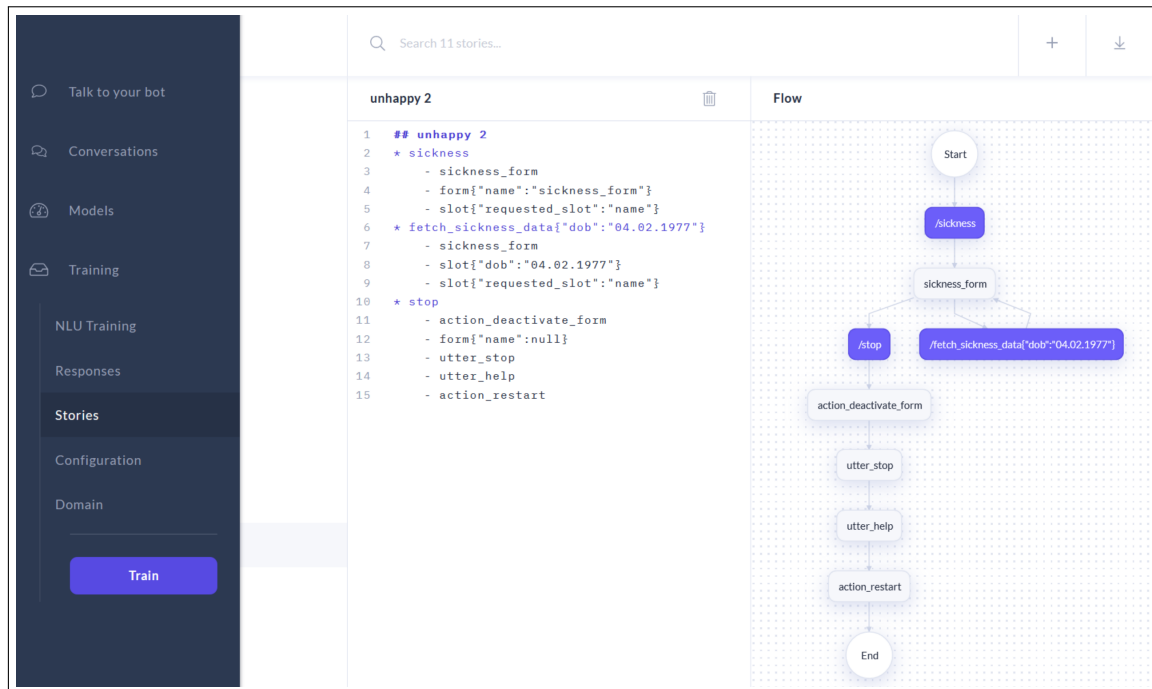


Figure 6.3: Rasa X Interface

sickness use-case, a new intent called sickness needs to be defined in the domain file. The sickness intent needs slot filling and the required entities are the person and the return date. The slots need to be defined in the slots section. Every slot needs an utterance which is named `utter_ask_{SLOTNAME}`. The utterance is displayed when the bot fills a slot. For the slot filling an action needs to be defined. An example slot fill action is shown in Listing 10.2. The action needs the name function which returns a unique name. The required slots need to be defined. If a slot is required it needs to be filled. The slot mappings function defines the type of the slots. The submit function is called as soon as the slots have been filled. In the submit function the collected information can be sent to the external endpoint for instance. Custom validation rules can be defined for each slot. This can be used to validate that the entered person name exists. Now the intent needs to be defined in the NLU file. The sickness intent and the training phrases need to be defined in the NLU file in the format of 6.6. The entities are marked with square brackets and the types with parentheses. Now the sickness story needs to be defined. When a phrase is recognized which is classified as sickness intent then the slot filling needs to be triggered. The next step is the confirmation dialog where the user confirms the entered information. A submit sickness data action is required. To classify the response the intent affirm and deny are created with some example phrases. After affirm is recognized the confirmation action needs to be triggered. In the run method of the submit action, the data can be validated and processed. If deny is recognized the slots need to be cleared and a retry message needs to be displayed. The sickness story for the successful cases is shown in Listing 6.7. The vacation use case is not different from the sickness use-case in Rasa. The Duckling time entity can process times, time-spans, dates, and date-spans. For Duckling there is no difference between a single date and a date-span. Both are stored in the time slot and are of type time. Hence, the same person and time slot of the sickness use-case

can be used for the vacation use-case. The response messages need to be adjusted. For the vacation use-case, a new intent needs to be created in the domain file. The vacation training phrases need to be listed in the NLU file. A form and a submit action need to be created for the vacation intent. The structure can be copied from the sickness use-case. Then the response messages and conditions are adjusted as needed. The actions need to be listed in the domain file to be usable in the stories. Then the new stories need to be created. With Rasa-X the stories are created with live interaction with the bot. The stories are structured like in the sickness use-case.

6.4 LUIS

With LUIS[7], the language understanding service of Microsoft, NLU services can be created. NLU is only a part of a chatbot and it's not possible to create a chatbot with LUIS alone. The chatbot features are available through the Microsoft bot services in combination with LUIS. A LUIS account is necessary to create NLU services. When a new project is created the name and a language need to be defined for the NLU application. The chosen language is the language the users will use for the communication with the LUIS application. The modification is done in the build tab. The interface shown in

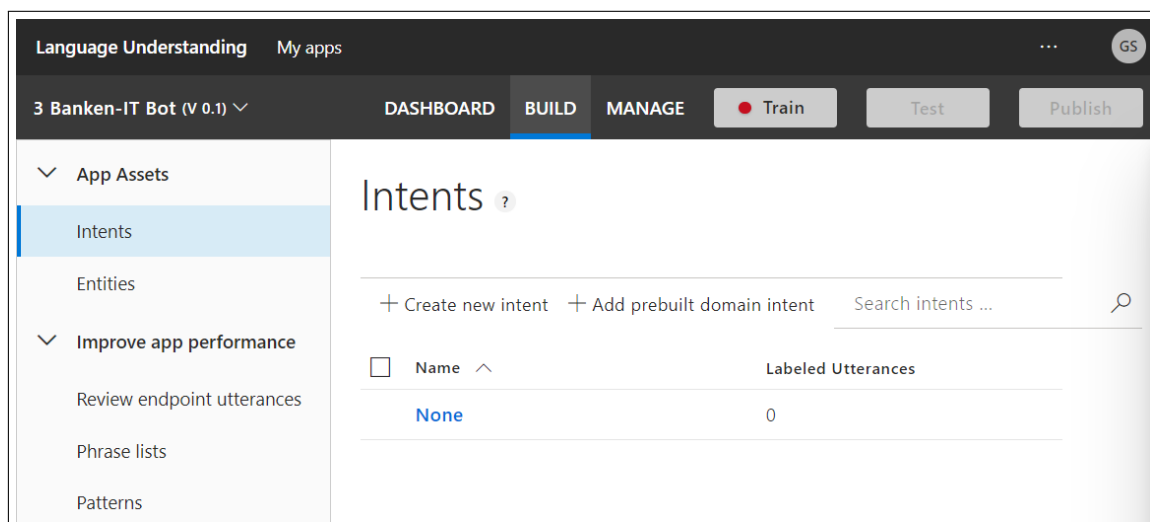


Figure 6.4: LUIS Web Interface

Figure 6.4 offers intents and entities like described in Chapter 2. To create the NLU part with LUIS Table 5.1 will be used. 5.1 defines the intents entities and response utterances that need to be created in LUIS. Figure 5.3 has no influence on LUIS because the conversation handling is not part of the NLU service. To create an intent only a name and a list of training phrases (utterances) are necessary. Multiple options are available to create different kinds of entities. There are simple entities like e.g. a city. They describe exactly one concept. There are composite entities that are built from multiple entities (parts). E.g. a ticket order entity could consist of an order number, the number of tickets, and the event. An entity list is used when the valid items are limited like in the book hotel example. There is only a limited number of valid hotel names and the list holds all valid hotels. The invalid hotels are uninteresting in such a case. Regular expressions[23] can also be used to define an entity. It is also possible to add prebuilt entities and prebuilt domain

entities. The entities required for the sickness intent are person name and date. LUIS offers the prebuilt entities date-time and person-name. This means no new entities need to be created for the sickness use-case since they are prebuilt. To test the NLU functionality two additional intents need to be created to simulate the conversation. A submit name and a submit return date intent are required. The submit name intent extracts the person name entity. The submit return date intent extracts a date-time entity. The advantage of predefined entities is the saved time and the advanced features. The date-time entity is able to match inputs like Monday or tomorrow to a valid and correct date by default. The year is also automatically added when the date e.g. first January lies in the past. This is important for the vacation use case when the vacation is requested for the next year.

Chapter 7

Result

Portability

An important factor for software systems is the portability. Portability is defined in ISO 25010 [6]. As a measurement criterion, the number of providers that support the technology is defined. A subfactor is how tied to a provider a technology is. When it comes to the chatbot cloud services (LUIS, Dialogflow, IBM Watson Assistant) the provider can't be changed. All three services run in the cloud of the provider and nowhere else. The data processing is done in the cloud and can't be done locally or on a company server. This means that the three cloud services are not portable at all and the provider can't be changed. With Rasa things are different. It is a local solution which means it can run on a local machine and a company server. The data processing is also done locally and never goes into the cloud of a provider. Docker images for Rasa are provided too. It's possible to run Docker containers in the clouds of all big providers. This means Rasa can run everywhere where Docker or a Python environment are possible and the provider can be changed easily. Another aspect of portability is the number of devices where the chatbot can be developed from. The cloud frameworks are accessible via the internet and can be used and developed from every device which has an internet connection and a browser. Rasa can only be developed on machines which have access to the source files. It is per default not accessible via the internet. However, it can be deployed in a cloud environment. Then it can be used like the cloud services but it still can't be developed easily on any device. The third aspect of portability is the migration from one technology to another. Rasa offers an import functionality (beta version) which allows The import of Dialogflow, LUIS, Wit.ai, and IBM Watson chatbots. In general, all chatbot frameworks use JSON format to store the training data. But every technology saves information differently and it's not easy to migrate an existing chatbot to another technology.

Implementation of the Design

like Singh et al. [32] said, the best evaluation criteria is if a task can be achieved with a technology or not. The goal is to implement the dialogs shown in Figure 5.3. The result of this evaluation is a binary criterion which is either true when the implementation was possible or false if it wasn't possible. The prototype implementations in Chapter 6 shown that the dialogs can be implemented with Dialogflow, Watson Assistant, and Rasa. It is not possible to implement the dialogs with LUIS. This was to be expected since LUIS is

a NLU service and not a chatbot technology and offers no dialog control. The Microsoft Bot framework is the chatbot technology which uses LUIS but is not considered in this thesis. The implementation of the dialogs of Figure 5.3 is possible with all three chatbot technologies.

Project Setup Complexity

As a measurement criterion, the required tools will be used. To set up Watson Assistant, Dialogflow, and LUIS a web-browser is needed. Those three technologies require nothing else since the development environment, files, and data are built into the websites of the technologies. This means only one tool is required. For Rasa three approaches are evaluated. Rasa can be installed locally or with Docker and Rasa-X can be used for development. Rasa-X provides a web-interface similar to the cloud technologies. A local setup requires a Python environment and a package manager like pip. A text editor or IDE is required to edit the files. This results in a count of at least three. The second way is a Docker container setup. For this approach, Docker and a text-editor are required and results in a count of two. With Rasa-x a web-interface replaces the text-editor but the score remains the same. This means the cloud technologies have a score of one and Rasa has a score of two or three depending on the approach. As second measurement criteria for setup complexity a setup time ranking will be used. To create a new project from scratch with the cloud technologies only an account is required. Then a new project can be created instantly. No tools or additional software needs to be installed to have everything development ready. For Rasa, the tools and the environment needs to be created. For the development, either Rasa needs to be installed or the Docker image needs to be pulled. The setup for Rasa takes more time because the development environment needs to be provided. This results in the first place for cloud services and the last/second place for Rasa in the setup time comparison.

Development Complexity

The easiest framework is LUIS since it is a NLU framework it doesn't offer dialog handling mechanisms. The development complexity for the NLU part is comparable to Dialogflow and Watson Assistant. They use the same concepts (intent, entity, utterance) in the UI and the mechanisms to create intents and entities are the same. The easiest chatbot framework is Dialogflow since it offers a simpler UI than Watson Assistant, more predefined entities, and the option to include features like small talk. Watson Assistant UI is close to Dialogflow but is not as transparent and additional logic can be handled inside of Watson which is a useful feature at times but makes the application more complex. Rasa is the most complex chatbot and NLU framework. The creation of intents, entities and utterances is scattered across multiple files. By default form filling actions are not supported and need to be created using Python. Luis, Dialogflow and Watson Assistant offer form filling by default and it can be enabled with a few clicks in the UI without a programming language. Rasa doesn't offer any predefined entities by default. To add predefined entities the pipeline needs to be adjusted which requires additional effort from the user. The creation of stories is also hardest with Rasa since the form filling action for instance requires additional information inside of the story, shown in Listing 10.3 where the form, person and time slots need to be set correctly. Trying to set the forms and

slots by hand is not recommended. The stories should be created using live training. It is a useful feature to create natural conversations but it's also more complex than the creation of stories with the dialog node systems of Watson Assistant and Dialogflow. The dialog node system of Dialogflow is shown in Figure 10.2. When the framework development complexity is evaluated based on the use-cases of this thesis the predefined entities are very important. The creation of new entities is equally hard with each framework. Hence, the more of the required entities are predefined the easier the development process gets. Dialogflow offers all three required entities for the German and English language as does Rasa. Watson Assistant doesn't offer a person entity in German and the one for the English language is marked deprecated. It also has no date-span entity but it can be replaced by two separate dates. This means that Watson Assistant offers two of the three entities for both required languages. LUIS offers all three entities for the English language but has no person entity for the German language.

Pricing

Dialogflow can be used by companies for free. The limits for the free tier are 180 text requests per minute. With the Essential and Plus version, 600 requests per minute are possible. Rasa is a free to use open-source tool and costs nothing. There is an enterprise edition with some benefits but the price is calculated individually. Watson Assistant offers a free tier with up to 1,000 users per month, 10,000 messages per month. The free tier of the Microsoft Bot Framework offers five requests per second and 10,000 messages per month which is the same amount of messages IBM offers. Five requests per second are 300 requests per minute which is more than the free tier of Dialogflow offers. A user is a person who interacts with the bot at least once. The Plus edition costs \$120 for 1,000 users per month with no message or user limit. The pricings and editions of the framework are shown in Table 7.1. The price of Dialogflow is based on the number of requests, IBM offers a fixed number of requests and users, and with Rasa the developer has to take care of the number of users and requests in soft- and hardware. The Microsoft Bot Framework charges per 1,000 users like IBM does. The price of the Microsoft Bot Framework is calculated by combining the prices for the framework and LUIS. The resulting cost for the Microsoft Bot Framework and LUIS is \$0.002 per request which is the same Dialogflow charges for the enterprise essential edition. The standard edition also offers 50 transactions per second which are 3,000 requests per minute which is far more than the 600 requests per second of Dialogflow. In the free tier, Dialogflow provides the best value because there is no message or user limit. In the priced tier Watson costs \$0.12 per user per month (\$120 for 1,000 users per month). Dialogflow and the Microsoft Bot Framework using LUIS charge \$0.002 per request each. Watsons cost of \$0.12 per user per month equals 60 requests per user per month with Dialogflow or the Microsoft Bot Framework plus LUIS.

| Framework | Edition | Price |
|-------------------------|------------------|---|
| Dialogflow | Standard | Free |
| | EE Essentials | \$0.002 per request |
| | EE Plus | \$0.004 per request |
| Watson Assistant | Lite | Free |
| | Plus | \$120 for 1,000 users/month |
| | Premium | Individual |
| Rasa | Rasa Open Source | Free |
| | Rasa X | Free |
| | Rasa Enterprise | Individual |
| Microsoft Bot Framework | Free | Free |
| | S1 | \$0.50 per 1,000 requests \$0.0005 per requests |
| LUIS | Free | Free (10.000 requests per month) |
| | Standard | \$1.50 per 1,000 transactions \$0.0015 per request |

Table 7.1: Pricing of Frameworks [8, 4, 10, 7]

Learnability

The easiest of the chatbot frameworks is Dialogflow. Compared to Rasa the setup is easy. The UI is minimalistic and easy to understand. It's easier to develop a chatbot with Dialogflow than with Watson Assistant or Rasa. The Watson Assistants UI is intransparent when compared to Dialogflow and a developer may need more time to orient. Rasa requires the most time to develop a chatbot and to get everything ready. As described in Section 7 the setup of cloud services is easy since the cloud providers take care of the setup. Rasa has to be installed or used as a Docker container which is harder than the cloud setup and requires more time. To use Watson and Dialogflow no programming skills are required while they are needed for Rasa which makes it harder to learn. It's also harder to define a dialog structure with Rasa since a story needs to be written in the format shown in Listing 6.7 whereas with Dialogflow and Watson dialog nodes are created which can be structured in a GUI with drag-and-drop. Furthermore, it's harder to create form filling actions in Rasa. In Rasa the form filling needs to be defined through code like in Listing 10.2 while Dialogflow and Watson support it through a simple GUI shown in Figure 10.1. When the GUI of Rasa-X is used defining new entities and intents works the same way for the three frameworks. To enable the use of predefined entities the pipeline of Rasa needs to be modified which is more complicated than with Dialogflow where they are enabled by default and Watson where the entities need to be enabled in the settings. Another measurement criteria for learnability is the number of predefined entities. If a predefined entity is available for a given problem a developer doesn't need to invest time in creating an entity or collecting training data. Table 7.2 shows the amount of predefined entities present for the tested technologies. Dialogflow provides the most predefined entities, followed by LUIS. Compared to the two above the other technologies provide a small amount of predefined entities.

| Rank | Framework | Entities |
|------|------------|-----------|
| 1 | Dialogflow | over 400 |
| 2 | LUIS | circa 150 |
| 3 | Spacy | 18 |
| 4 | Duckling | 11 |
| 5 | Watson | 7 |
| 6 | Rasa | 0 |

Table 7.2: Predefined Entities

Deployment Complexity

A major factor for the development of a chatbot is where it can be deployed and how easy the deployment process is. For the cloud technologies the deployment is done by the provider and there is no development overhead for the deployment. The deployment of Rasa needs to be done by the developer. Rasa can either run in a Docker container or it can be installed in a Python environment. The Docker container approach was chosen for the deployment because it's supported by all cloud providers. The cloud technologies share the first place in the deployment ranking since there is no overhead for the developer. The only technology tested which produces deployment overhead is Rasa. The second deployment is the local deployment. It's possible to deploy Rasa locally using Docker or a Python environment. As expected it's not possible to deploy a cloud chatbot locally or in a non native cloud environment. This ties a user to the provider and is not flexible. The only technology used in this thesis which supports a local deployment or the deployment in more than one cloud environment is Rasa.

Communication

All of the chatbot and NLU technologies offer Rest API endpoints for communication. There is no difference in the type of communication. Dialogflow and Watson Assistant communicate with a webhook when the extraction of information was successful. Rasa communicates with the action server which serves a similar purpose as the webhooks. The communication between the frameworks and the external service is done with JSON messages. While Dialogflow and LUIS offer metadata, shown in Listing 6.2, 10.4 and 10.5, Rasa and Watson Assistant offer no metadata by default. The required information needs to be defined by the developer. This means more logic can be handled with Rasa and Watson Assistant on the chatbot side since the format and processing of the result is up to the developer. Especially with Rasa the logic can be freely defined since the action server is written in Python and all features of programming languages are available for the result processing. The behavior of Watson Assistant and Rasa can be mimicked by shifting the logic from the chatbot to the webhook.

Entity Recognition and Extraction

Entity recognition and extraction is a key feature of chatbots. For the use-cases of this thesis a date, date-span, and person entity are required. The entity recognition and ex-

traction tests focus solely on the required entities. In Table 7.3 the supported entities are listed alongside the technologies. Rasa does not offer any entities but the pipeline can be adjusted. One part of the pipeline is Spacy which offers person and date entities. Spacy does not convert the date into a standard format. For this reason, Duckling was added to the pipeline. Duckling focuses on numeric values, dates, and times and converts dates into a standardized format. Dialogflow offers the system entities date, date-period, and person. LUIS offers a person and date-time entity. The date-time entity can also be a date-span. Hence, no entities need to be created in LUIS or Dialogflow for the use-cases of this thesis.

| Framework | Subtype | Person | Date | Date Span | Standardized Output Date |
|------------|----------|--------|------|-----------|--------------------------|
| Rasa | - | ✗ | ✗ | ✗ | ✗ |
| | Spacy | ✓ | ✓ | ✓ | ✗ |
| | Duckling | ✗ | ✓ | ✓ | ✓ |
| Dialogflow | - | ✓ | ✓ | ✓ | ✓ |
| Watson | - | ✗ | ✓ | ✗ | ✓ |
| LUIS | English | ✓ | ✓ | ✓ | ✓ |
| | German | ✗ | ✓ | ✓ | ✓ |

Table 7.3: Framework Entity Recognition and Extraction

Watson Assistant offers the system entity sys-date but no date-span entity. When Watson recognizes two sys-dates the slots are filled correctly as start and return date and the result is comparable to a date-span entity. Dialogflow extracts the date, time, and timezone in the format "2020-06-02T12:00:00+02:00". Watson and LUIS extract only the date in the format "2020-06-02". Duckling extracts dates (time entity) in the format "2020-06-07T00:00:00.000-07:00". The second date of date spans is always set to one day after the recognized day in Duckling. This means that the second dates of Duckling differ from the dates of the other technologies shown in Table 7.8 in the "Extracted Value" column. The column "Extracted Value" of Tables 7.6, 7.7, and 7.8, 7.9 show that Spacy doesn't convert the dates into standard format. The date and date-span tests focus on different date inputs. It is to be expected that a recognized date format works for all instances of the format. If Monday is recognized correctly the other weekdays will work too. The person entity has been added to Watson Assistant with custom training data. The expectation is that the person entity of Watson Assistant performs worst because it's no system entity and is built with a low amount of training data. The results of the date recognition and extraction are listed in Table 7.6 and 7.7. The expectation of the results is listed as true positive(TP) when the date should be recognized and extracted or as true negative(TN) when the entity shouldn't be recognized and extracted. The test result shows that the date entity of Dialogflow worked best with just one error followed closely by Duckling with two errors. Spacy performed worst with five errors in ten tests followed by Watson Assistant with four errors out of ten. No technology worked correctly on test case eight of Table 7.6. This is true for the english language selection. When German is selected as a language the performance on dates separated by "." increases. It also shows that weekdays and tomorrow work with all technologies. The extracted value column shows that Spacy recognizes many date inputs correctly but doesn't convert them.

The detailed result of the entity recognition comparison is shown in Table 7.4. The column name p stands for precision and r for recall. Date sum shows the framework result based on the date and date-span recognition without the person entity for the comparison with

| Framework | Type | TP | TN | FP | FN | p | r | F-Score | Tests |
|---------------------------|-----------|----|----|----|----|--------|-------|---------|-------|
| Dialogflow date-period | Date | 8 | 1 | 0 | 1 | 1.0 | 0.888 | 0.94 | 10 |
| | Date-Span | 2 | 0 | 0 | 4 | 1.0 | 0.333 | 0.5 | 6 |
| | Person | 3 | 2 | 0 | 1 | 1.0 | 0.75 | 0.857 | 6 |
| | Summary | 13 | 3 | 0 | 6 | 1.0 | 0.684 | 0.813 | 22 |
| Dialogflow date-time | Date | 8 | 1 | 0 | 1 | 1.0 | 0.888 | 0.94 | 10 |
| | Date-Span | 5 | 0 | 0 | 1 | 1.0 | 0.833 | 0.909 | 6 |
| | Person | 3 | 2 | 0 | 1 | 1.0 | 0.75 | 0.857 | 6 |
| | Summary | 16 | 3 | 0 | 3 | 1.0 | 0.842 | 0.914 | 22 |
| Watson | Date | 6 | 0 | 1 | 3 | 0.857 | 0.666 | 0.75 | 10 |
| | Date-Span | 5 | 0 | 0 | 1 | 1.0 | 0.833 | 0.909 | 6 |
| | Person | 2 | 1 | 1 | 2 | 0.666 | 0.5 | 0.571 | 6 |
| | Summary | 13 | 1 | 2 | 6 | 0.866 | 0.684 | 0.765 | 22 |
| LUIS | Date | 6 | 0 | 1 | 3 | 0.857 | 0.666 | 0.75 | 10 |
| | Date-Span | 3 | 0 | 0 | 3 | 1.0 | 0.5 | 0.666 | 6 |
| | Person | 4 | 0 | 2 | 0 | 0.666 | 1.0 | 0.8 | 6 |
| | Summary | 13 | 0 | 3 | 6 | 0.8125 | 0.684 | 0.743 | 22 |
| Spacy | Date | 4 | 1 | 0 | 5 | 1.0 | 0.444 | 0.615 | 10 |
| | Date-Span | 2 | 0 | 0 | 4 | 1.0 | 0.333 | 0.5 | 6 |
| | Person | 4 | 0 | 2 | 0 | 0.666 | 1.0 | 0.8 | 6 |
| | Summary | 10 | 1 | 2 | 9 | 0.833 | 0.526 | 0.645 | 22 |
| Duckling | Date | 7 | 1 | 0 | 2 | 1.0 | 0.777 | 0.875 | 10 |
| | Date-Span | 5 | 0 | 0 | 1 | 1.0 | 0.833 | 0.909 | 6 |
| | Summary | 12 | 1 | 0 | 3 | 1.0 | 0.8 | 0.888 | 16 |
| Dialogflow | Date Sum | 13 | 1 | 0 | 2 | 1.0 | 0.866 | 0.929 | 16 |
| Watson | Date Sum | 11 | 0 | 1 | 4 | 0.916 | 0.733 | 0.815 | 16 |
| LUIS | Date Sum | 9 | 0 | 1 | 6 | 0.9 | 0.6 | 0.72 | 16 |
| Spacy | Date Sum | 6 | 1 | 0 | 9 | 1.0 | 0.4 | 0.571 | 16 |
| Spacy + Duckling | Summary | 16 | 1 | 2 | 3 | 0.888 | 0.842 | 0.865 | 22 |

Table 7.4: Entity Recognition Result Evaluation

Duckling which has no person entity. The last row of Table 7.4 shows the result for the combination of Duckling for dates and date-spans and Spacy for persons. This makes the comparison of Rasa with the other technologies possible. The true negatives (TN) are not part of the f-score calculation. Table 7.4 shows that the technology with the best f-score across all tests (0.914) is Dialogflow followed by a combination of Duckling for dates and date-spans and Spacy for persons is used. Rasa can use Duckling and Spacy and reaches the second highest f-score. The best f-score (0.929) for the combination of dates and date-spans was achieved by Dialogflow. Dialogflow has the best date entity (f-score of 0.94) when the date-time entity is used instead of the date-period entity. The date-period entity of Dialogflow reached a low f-score of 0.5 and is not recommended for further use. Duckling, Watson, and Dialogflow were on an equal level when only date-spans are considered. All three reached an f-score of 0.909. Dialogflow also has the best f-score for the person entity. Watsons person entity performed worst as expected. With more training data the performance of Watsons custom person entity can be increased. In test case two of Table 7.8 Spacy recognized two separate dates instead of a date-span.

This is counted as a failure because in the application there is only one date slot present and the information can't be processed correctly. In test case one Dialogflows date-period entity and Luis convert Monday to a date in the past. This is now what is needed for the use-cases since it should be the next matching weekday and not the last. Test case two gives a hint that the next occurring weekday is the correct one. The result is not influenced and Dialogflows date-period entity and LUIS still take the wrong day. The functionality was retested on another day. LUIS still failed to extract the correct date but Dialogflow managed to do so. In the test result it will be counted as a failure of Dialogflows date-period entity since it doesn't deliver reliable results.

| No | Input | Expected | Technology | Recognized Correctly | Extracted Value |
|----|------------------|----------|---------------------------------------|----------------------|--|
| 1 | Franz Bauer | TP | Watson Dialogflow Spacy LUIS | ✓ ✓ ✓ ✓ | Franz Bauer Franz Bauer Franz Bauer Franz Bauer |
| 2 | franz bauer | TP | Watson Dialogflow Spacy LUIS | ✓ ✓ ✓ ✓ | franz bauer franz bauer franz bauer franz bauer |
| 3 | Anna Maria Mayer | TP | Watson Dialogflow Spacy LUIS | ✗ ✗ ✓ ✓ | Maria Mayer Maria Mayer Anna Maria Mayer Anna Maria Mayer |
| 4 | Anna Mayer-Bauer | TP | Watson Dialogflow Spacy LUIS | ✗ ✓ ✓ ✓ | Anna Mayer Anna Mayer-Bauer Anna Mayer-Bauer Anna Mayer-Bauer |
| 5 | Asdfgh | TN | Watson Dialogflow Spacy LUIS | ✓ ✓ ✗ ✗ | - - Asdfgh Asdfgh |
| 6 | Asdfgh Qwert | TN | Watson Dialogflow Spacy LUIS | ✗ ✓ ✗ ✗ | Asdfgh Qwert - Asdfgh Qwert Asdfgh Qwert |

Table 7.5: Person Entity Recogniton and Extraction

Intent Classification

The main task of a chatbot is to find the intent best matching the user input. The training utterances for the sickness intent are listed Tabel 4.1 in the Appendix. The phrases for the vacation intent can be found in Table 10.2. As an evaluation criterion, multiple parameters are used. The confidence score represents how well the input can be classified. Building upon the entity extraction the correct extraction of the entities in combined inputs is tested. A combined example is training phrase five of Table 4.1 where the entities person and date

| No | Input | Expected | Technology | Recognized Correctly | Extracted Value |
|----|----------|----------|------------|----------------------|-----------------|
| 1 | Tomorrow | TP | Watson | ✓ | 2020-06-01 |
| | | | Dialogflow | ✓ | 2020-06-01 |
| | | | Spacy | ✓ | Tomorrow |
| | | | Duckling | ✓ | 2020-06-01 |
| | | | LUIS | ✓ | 2020-06-01 |
| 2 | Friday | TP | Watson | ✓ | 2020-06-05 |
| | | | Dialogflow | ✓ | 2020-06-05 |
| | | | Spacy | ✓ | Friday |
| | | | Duckling | ✓ | 2020-06-05 |
| | | | LUIS | ✓ | 2020-06-05 |
| 3 | 6.June | TP | Watson | ✗ | 2020-06-01 |
| | | | Dialogflow | ✓ | 2020-06-06 |
| | | | Spacy | ✗ | - |
| | | | Duckling | ✗ | - |
| | | | LUIS | ✗ | 2020-06-01 |

Table 7.6: Date Entity Recogniton and Extraction

are built into the training sentence. Furthermore, the correctly identified intents are used to calculate the f-score like it was done for the entities. For the intent classification task, only date and date-spans are used which were recognized by all relevant frameworks to ensure unbiased tests. Spacy has no relevance for the application in terms of date and date-span extraction and is not considered for dates in the intent classification tests. The detailed test are listed in Table 7.10 and Table 7.11. The test evaluation is shown in Table 7.12. The tests show that Dialogflow, LUIS, and Watson have almost the same average confidence score. Dialogflow has a smaller deviation than the other three. Dialogflow always scores above 50% confidence and often reaches 100%. The scores of Watson Assistant are similar to Dialogflow with the difference that the confidence score was below 50% at times. The confidence scores of Rasa are lower than the ones of Dialogflow, LUIS, and Watson on average, shown in Table 7.12. Dialogflow and Watson managed to classify all intents correctly. LUIS classified all vacation intents correctly. The entity extraction of Dialogflow failed in test case 4 of Table 7.10 since "myself" has been extracted as a person. This means that the person slot was set incorrectly. LUIS extracted sick as person in test sentence one and didn't extract the person in test sentence six, shown in Table 7.10. The test sentences get more complex as the number increases. Dialogflow and Watson were able to correctly classify all test sentences. LUIS failed to classify test sentence seven of Table 7.10. Rasa failed to correctly classify the more complex sentences. This leads to the conclusion that Dialogflow, LUIS, and Watson work better with complex sentences when a small amount of training data is used. The result in Table 7.12 shows that Dialogflow and Watson reach a perfect f-score of 1.0. Rasa was able to reach a much higher f-score on the vacation use-case than on the sickness use-case. This implies that the vacation test sentences were better classified by the system than the sickness related sentences. The average confidence was highest with Dialogflow followed closely by LUIS and the Watson Assistant. Compared to Dialogflow, LUIS, and Watson Assistant the average confidence score of Rasa is a lot lower. The same experiment is also interesting with German as language since the potential users are from a German

speaking country. The training data is listed in the Appendix in Table 10.1 and Table 10.3. The translated test sentences are listed in Table 10.4 and Table 10.5 with a boolean value which indicates if the correct intent was identified, a boolean that indicates if the entities were extracted correctly, and the confidence score for each test case. The finalized result is listed in 7.13 and shows that the average confidence score for the German test sentences is higher than for the English test sentences for each technology individually. The f-score also increased significantly which indicates that the German test sentences are recognized better than the English test sentences. The ranking of the Technologies is also different from before since three technologies achieved perfect f-scores the ranking is based on the average confidence. The highest confidence was achieved by Watson. The second place goes to LUIS. LUIS was able to increase the average confidence from about 75% on the English tests to about 85% on the German tests. Dialogflow only reaches the third place. The intent classification is top notch for all technologies and all can be recommended from that point of view. The results in Table 10.4 and 10.5 show that the entity extraction results drop in the German setup with all technologies except LUIS. Dialogflow has problems to extract the person entity correctly in German sentences although German has been selected in the settings. The system entity date-time of Dialogflow is unable to handle German date formats although the dialog language is set to German. Watson doesn't offer entity annotation in training sentences when German is selected as language. This means that a main feature is only available for English language bots. Until the annotation of entities in training phrases is enabled for other languages Watson should only be used in English. LUIS offers a person entity when English is selected but not for the German language. The date-time entity of LUIS is able to handle German date formats. This means that LUIS is the only technology which achieved a satisfying result in the entity extraction category. All in all, the intent classification result is very good for the German language tests but the entity extraction often fails.

Training Capabilities

One major factor for chatbot technologies is the training of intents, entities, and stories/-dialogs. The development of new intents and entities works the same for all four technologies. An intent is trained with utterances. An entity is created either with examples or annotations. The training sentences for the intents can contain entities. They can be marked as entities by annotating them. Then the values are used to train the entity. This process is the same for all four technologies, hence there is no difference in the creation process of intents and entities. The creation of stories/dialogs only works for chatbot technologies. LUIS won't be part of the story evaluation. The creation of stories is done with dialog nodes in Dialogflow and Watson while Rasa uses stories. A dialog node represents a recognized intent and defines the slots, actions, and responses regarding the intent. In the UI follow up dialog nodes can be created to build a structure where one intent follows after another. Rasa uses stories which also defined the recognized intent, the slots, actions and responses. The real difference is that the stories of Rasa can be structured in any imaginable way while the dialog nodes follow a fixed structure. The second big difference is that live training should be used with Rasa since the structuring of the story files can get rather complicated. Neither Dialogflow nor Watson Assistant offer live training. With live training a developer starts a conversation with the bot and goes through the process. Everytime the bot makes a mistake the developer can correct the bot. The second option is to define the conversation step-by-step from scratch where the developer tells the bot

what to do. When the end of a story is reached it is stored and used for training. The training capabilities of Rasa exceed the ones of Dialogflow and Watson Assistant and is ranked first. There is no difference between Dialogflow and Watson Assistant when the dialog structure creation is considered.

Speech Recognition Capabilities

For the further development which is based on the results of this thesis speech recognition is an option. The result of this evaluation is true or false. Frameworks which support this feature can be used with speech input instead of text. Either the technology supports speech recognition or it doesn't. Rasa doesn't support speech functionality itself but it can be used in combination with speech-to-text and text-to-speech frameworks like Alexa. Dialogflow offers audio in- and output. The speech-to-text functionality of Dialogflow costs \$0.0065 per 15 seconds of audio. The IBM Watson Assistant also offers speech functionalities through Watson Speech to Text and Watson Text to Speech. The Watson text-to-speech service costs between \$0.01 and \$0.02 per minute which results in \$0.0025 to \$0.005 per 15 seconds. The speech-to-text service costs \$0.02 per 1,000 chars which results in \$20 per million chars. Microsoft offers the Speech Services for the speech-to-text and text-to-speech functionality. The Speech service costs \$1 per hour of speech which results in \$0.00416 per 15 seconds which is less than the costs for Dialogflow. The costs for text-to-speech are the same for the Speech Services and Dialogflow with a rate of \$4 per one million characters. The speech-to-text services are on a similar pricing level but the text-to-speech service of Watson costs five times as much as the Speech Services or Dialogflow. The only technology which doesn't support speech recognition is Rasa. The other three technologies support the speech-to-text and text-to-speech features which are required for speech processing.

Framework Concepts

The user interfaces of the cloud frameworks are similar. The interfaces are shown in Figure 6.1, 6.2, and 6.4. All three offer intents and entities. They are all structured similarly with the menu to the left. The information for the bots is always entered as text. Intents are always built with training phrases. All offer prebuilt system entities. Dialogflow has the most prebuilt entities, followed by LUIS, then comes Rasa, and Watson Assistant offers the fewest entities. For the sickness and vacation use-case the entities of Dialogflow suit best because no entity has to be defined by the developer. Watson offers a person system entity in English but the entity is marked as deprecated. It offers a date entity but no date span entity. This means with Watson Assistant two entities have to be modeled for the use-cases. LUIS offers a date-time entity which can be used for dates and date spans but it's not working as reliable as the one from Dialogflow. It did not identify "22-06 to 30-06" correctly which is no problem for Dialogflow.

Usability of the Framework

The evaluation of the usability of the frameworks is based on the subjective impressions from the prototype development section. A major part of the chatbot development process

is the creation of intents entities and utterances. The concepts are implemented in the same way for all technologies except Rasa where they need to be defined in multiple files to achieve the same result. This means that it's equally easy to implement intents, entities, and utterances with Dialogflow, LUIS, and Watson Assistant and easier than with Rasa. The structuring of the user interfaces of the cloud bots is similar. The easiest interface was the one of LUIS which was expected since it has no dialog handling mechanism. The simplest UI of the three chatbots was the one of Dialogflow since it was more transparent than the interface of Watson Assistant. Rasa-X was inbetween Dialogflow and Watson Assistant. Rasa alone had no graphical user interface and is on the last place in this category.

Customization Possibilities

A major part of chatbot frameworks is the pipeline since it defines how the training of the intents and entities is done. The cloud chatbot technologies don't show the pipeline they are using nor can you adjust it. Rasa is different in that regard since the pipeline can be adjusted freely. Furthermore, the technologies like Spacy show information about the components online. Spacy can be used with different models like small, medium, and large training sets for the predefined entities when English is the target language. The pipeline used for this thesis is shown in Figure 10.1. It's even possible to include custom components into the pipeline of Rasa. This means that Rasa is the only technology of the four where the pipeline is visible and it offers customization on top of that. The other three technologies don't offer mechanisms to adjust the pipeline. It's possible to use external services instead of adjusting the pipeline by moving the entity extraction to the webhook where necessary. Since especially Dialogflow and LUIS offer a huge amount of predefined entities the use of other entity extraction tools might not be necessary at all. It's different for Watson Assistant since it has a small amount of predefined entities.

| No | Input | Expected | Technology | Recognized Correctly | Extracted Value |
|----|-------------|----------|------------|----------------------|-----------------|
| 4 | June 6 | TP | Watson | ✓ | 2020-06-06 |
| | | | Dialogflow | ✓ | 2020-06-06 |
| | | | Spacy | ✓ | June 6 |
| | | | Duckling | ✓ | 2020-06-06 |
| | | | LUIS | ✓ | 2020-06-06 |
| 5 | 6th of June | TP | Watson | ✓ | 2020-06-06 |
| | | | Dialogflow | ✓ | 2020-06-06 |
| | | | Spacy | ✓ | 6th of June |
| | | | Duckling | ✓ | 2020-06-06 |
| | | | LUIS | ✓ | 2020-06-06 |
| 6 | 06/06 | TP | Watson | ✓ | 2020-06-06 |
| | | | Dialogflow | ✓ | 2020-06-06 |
| | | | Spacy | ✗ | - |
| | | | Duckling | ✓ | 2020-06-06 |
| | | | LUIS | ✓ | 2020-06-06 |
| 7 | 06-06 | TP | Watson | ✗ | - |
| | | | Dialogflow | ✓ | 2020-06-06 |
| | | | Spacy | ✗ | - |
| | | | Duckling | ✓ | 2020-06-06 |
| | | | LUIS | ✗ | - |
| 8 | 06.06 | TP | Watson | ✗ | - |
| | | | Dialogflow | ✗ | - |
| | | | Spacy | ✗ | - |
| | | | Duckling | ✗ | 2020-05-31 |
| | | | LUIS | ✗ | - |
| 9 | 06.06.2020 | TP | Watson | ✓ | 2020-06-06 |
| | | | Dialogflow | ✓ | 2020-06-06 |
| | | | Spacy | ✗ | - |
| | | | Duckling | ✓ | 2020-06-06 |
| | | | LUIS | ✓ | 2020-06-06 |
| 10 | 06/34/2020 | TN | Watson | ✗ | 2020-06-01 |
| | | | Dialogflow | ✓ | - |
| | | | Spacy | ✓ | - |
| | | | Duckling | ✓ | - |
| | | | LUIS | ✗ | 2020-01-01 |

Table 7.7: Date Entity Recognition and Extraction 2

| No | Input | Expected | Technology | Recognized Correctly | Extracted Value |
|----|-----------------------------------|----------|------------------------|----------------------|---------------------------|
| 1 | Monday to Friday | TP | Watson | ✓ | 2020-06-01 2020-06-05 |
| | | | Dialogflow date-period | ✗ | 2020-05-25 2020-06-01 |
| | | | Dialogflow date-time | ✓ | 2020-05-25 2020-06-05 |
| | | | Spacy | ✓ | Monday to Friday |
| | | | Duckling | ✓ | 2020-06-01 2020-06-06 |
| | | | LUIS | ✗ | 2020-05-25 2020-05-29 |
| 2 | from next tuesday to wednesday | TP | Watson | ✓ | 2020-06-02 2020-06-03 |
| | | | Dialogflow date-period | ✗ | 2020-05-26 2020-06-03 |
| | | | Dialogflow date-time | ✓ | 2020-06-02 2020-06-03 |
| | | | Spacy | ✗ | next tuesday wednesday |
| | | | Duckling | ✓ | 2020-06-02 2020-06-04 |
| | | | LUIS | ✗ | 2020-06-09 2020-06-10 |
| 3 | June 6 to June 7 | TP | Watson | ✓ | 2020-06-06 2020-06-07 |
| | | | Dialogflow date-period | ✓ | 2020-06-06 2020-06-07 |
| | | | Dialogflow date-time | ✓ | 2020-06-06 2020-06-07 |
| | | | Spacy | ✓ | June 6 to June 7 |
| | | | Duckling | ✓ | 2020-06-06 2020-06-08 |
| | | | LUIS | ✓ | 2020-06-06 2020-06-07 |

Table 7.8: Date Span Entity Recogniton and Extraction 1

| No | Input | Expected | Technology | Recognized Correctly | Extracted Value |
|----|-----------------------------|----------|------------------------|----------------------|--------------------------|
| 4 | 06/06 to 06/07 | TP | Watson | ✓ | 2020-06-06 2020-06-07 |
| | | | Dialogflow date-period | ✓ | 2020-06-06 2020-06-07 |
| | | | Dialogflow date-time | ✓ | 2020-06-06 2020-06-07 |
| | | | Spacy | ✗ | - - |
| | | | Duckling | ✓ | 2020-06-06 2020-06-08 |
| | | | LUIS | ✓ | 2020-06-06 2020-06-07 |
| 5 | 06/07 to 06/06 | TP | Watson | ✗ | 2020-06-07 2020-06-06 |
| | | | Dialogflow date-period | ✗ | 2020-06-07 2020-06-06 |
| | | | Dialogflow date-time | ✓ | 2020-06-07 2021-06-06 |
| | | | Spacy | ✗ | - - |
| | | | Duckling | ✗ | 2020-06-07 2020-06-07 |
| | | | LUIS | ✗ | 2020-06-07 2020-06-06 |
| 6 | from tomorrow to 06.06.2020 | TP | Watson | ✓ | 2020-06-02 2020-06-06 |
| | | | Dialogflow date-period | ✗ | - - |
| | | | Dialogflow date-time | ✗ | 2020-06-02 - |
| | | | Spacy | ✗ | tomorrow - |
| | | | Duckling | ✓ | 2020-06-02 2020-06-07 |
| | | | LUIS | ✓ | 2020-06-02 2020-06-06 |

Table 7.9: Date Span Entity Recogniton and Extraction 2

| No | Input | Technology | Intent Match | Entity Match | Confid. [%] |
|----|--|------------|--------------|--------------|-------------|
| 1 | sick | Watson | ✓ | - | 41.0 |
| | | Dialogflow | | | 84.2 |
| | | Rasa | | | 78.5 |
| | | LUIS | | ✗ | 99.2 |
| 2 | Emilia Schwarz is ill | Watson | ✓ | ✓ | 94.0 |
| | | Dialogflow | | | 78.2 |
| | | Rasa | | | 63.1 |
| | | LUIS | | | 83.4 |
| 3 | Lea Schmitt is sick until wednesday | Watson | ✓ | ✓ | 92.0 |
| | | Dialogflow | | | 73.9 |
| | | Rasa | | | 66.2 |
| | | LUIS | | | 83.3 |
| 4 | I need to report myself as sick | Watson | ✓ | - | 38.0 |
| | | Dialogflow | | ✗ | 63.3 |
| | | Rasa | ✗ | - | 9.7 |
| | | LUIS | ✓ | | 87.2 |
| 5 | Paul Armstrong is sick and will be back on Monday | Watson | ✓ | ✓ | 89.0 |
| | | Dialogflow | | | 69.6 |
| | | Rasa | ✗ | | 23.8 |
| | | LUIS | ✓ | | 73.1 |
| 6 | sick Anna Maier 06/10 | Watson | ✓ | ✓ | 90.0 |
| | | Dialogflow | | | 100.0 |
| | | Rasa | | | 80.5 |
| | | LUIS | | ✗ | 67.6 |
| 7 | I feel sickly today and won't come to work | Watson | ✓ | ✓ | 43.0 |
| | | Dialogflow | | | 51.6 |
| | | Rasa | ✗ | | 4.3 |
| | | LUIS | | | 17.9 |
| 8 | Anna Lehner is feeling sick. She will be back on Thursday | Watson | ✓ | ✓ | 91.0 |
| | | Dialogflow | | | 72.5 |
| | | Rasa | ✗ | | 26.2 |
| | | LUIS | ✓ | | 68.1 |

Table 7.10: Sickness Intent Classification

| No | Input | Technology | Intent Match | Entity Match | Confid. [%] |
|----|--|------------|--------------|--------------|-------------|
| 1 | vacation | Watson | ✓ | - | 100.0 |
| | | Dialogflow | | | 100.0 |
| | | Rasa | | | 91.7 |
| | | LUIS | | | 97.1 |
| 2 | my colleague Franz Bauer wants to go on vacation | Watson | ✓ | ✓ | 91.0 |
| | | Dialogflow | | | 64.9 |
| | | Rasa | | | 33.7 |
| | | LUIS | | | 64.9 |
| 3 | vacation Martin Huber 07/12 to July 15th | Watson | ✓ | ✓ | 44.0 |
| | | Dialogflow | | | 100.0 |
| | | Rasa | | | 57.4 |
| | | LUIS | | | 84.6 |
| 4 | holiday from 07/12 to July 15th | Watson | ✓ | ✓ | 94.0 |
| | | Dialogflow | | | 64.7 |
| | | Rasa | | | 67.4 |
| | | LUIS | | | 87.9 |
| 5 | vacation from 28th September to October 4 | Watson | ✓ | ✓ | 95.0 |
| | | Dialogflow | | | 100.0 |
| | | Rasa | | | 55.5 |
| | | LUIS | | | 88.1 |
| 6 | I really need a long vacation on a distant island | Watson | ✓ | - | 44.0 |
| | | Dialogflow | | | 79.0 |
| | | Rasa | | | 35.0 |
| | | LUIS | | | 92.3 |
| 7 | colleague Helmut Kerschbaum requested holidays | Watson | ✓ | ✓ | 93.0 |
| | | Dialogflow | | | 61.7 |
| | | Rasa | | | 46.6 |
| | | LUIS | | | 41.9 |
| 8 | Colleague Helmut Kerschbaum requested holidays. He will be unavailable from 09/10 to 09/24 | Watson | ✓ | ✓ | 45.0 |
| | | Dialogflow | ✓ | | 70.2 |
| | | Rasa | ✗ | | 21.6 |
| | | LUIS | ✓ | | 70.8 |

Table 7.11: Vacation Intent Classification

| Framework | Tested | TP | FN | p | r | f-score | Confid. avg. [%] | Confid. Median | std. dev. |
|------------|----------|----|----|-----|------|---------|---------------------|-------------------|--------------|
| Watson | sickness | 8 | 0 | 1.0 | 1.0 | 1.0 | 72.3 | 89.5 | 26.2 |
| | vacation | 8 | 0 | 1.0 | 1.0 | 1.0 | 75.8 | 92.0 | 26.1 |
| | sum | 16 | 0 | 1.0 | 1.0 | 1.0 | 74.0 | 90.5 | 25.4 |
| Dialogflow | sickness | 8 | 0 | 1.0 | 1.0 | 1.0 | 74.2 | 73.2 | 14.3 |
| | vacation | 8 | 0 | 1.0 | 1.0 | 1.0 | 80.1 | 74.6 | 17.3 |
| | sum | 16 | 0 | 1.0 | 1.0 | 1.0 | 77.1 | 73.2 | 15.6 |
| Rasa | sickness | 4 | 4 | 1.0 | 0.5 | 0.67 | 44.0 | 44.7 | 31.3 |
| | vacation | 7 | 1 | 1.0 | 0.88 | 0.93 | 51.1 | 51.1 | 22.1 |
| | sum | 11 | 5 | 1.0 | 0.69 | 0.82 | 47.6 | 51.1 | 26.4 |
| LUIS | sickness | 7 | 1 | 1.0 | 0.88 | 0.93 | 72.5 | 78.2 | 24.5 |
| | vacation | 8 | 0 | 1.0 | 1.0 | 1.0 | 78.5 | 86.3 | 18.3 |
| | sum | 15 | 1 | 1.0 | 0.94 | 0.97 | 75.5 | 83.4 | 21.1 |

Table 7.12: Intent Classification Result

| Framework | Tested | TP | FN | p | r | f-score | Confid. avg. [%] | Confid. Median | std. dev. |
|------------|----------|----|----|-----|------|---------|---------------------|-------------------|--------------|
| Watson | sickness | 8 | 0 | 1.0 | 1.0 | 1.0 | 94.1 | 94.0 | 1.88 |
| | vacation | 8 | 0 | 1.0 | 1.0 | 1.0 | 96.4 | 96.5 | 1.99 |
| | sum | 16 | 0 | 1.0 | 1.0 | 1.0 | 95.3 | 95.5 | 2.21 |
| Dialogflow | sickness | 8 | 0 | 1.0 | 1.0 | 1.0 | 80.8 | 78.6 | 19.1 |
| | vacation | 8 | 0 | 1.0 | 1.0 | 1.0 | 89.5 | 100 | 16.5 |
| | sum | 16 | 0 | 1.0 | 1.0 | 1.0 | 85.1 | 90.3 | 17.8 |
| Rasa | sickness | 7 | 1 | 1.0 | 0.88 | 0.93 | 54.0 | 44.2 | 20.2 |
| | vacation | 8 | 0 | 1.0 | 1.0 | 1.0 | 58.3 | 57.0 | 10.9 |
| | sum | 15 | 1 | 1.0 | 0.94 | 0.97 | 56.2 | 53.6 | 15.9 |
| LUIS | sickness | 8 | 0 | 1.0 | 1.0 | 1.0 | 79.6 | 77.5 | 15.1 |
| | vacation | 8 | 0 | 1.0 | 1.0 | 1.0 | 91.0 | 90.9 | 5.01 |
| | sum | 16 | 0 | 1.0 | 1.0 | 1.0 | 85.3 | 89.7 | 12.3 |

Table 7.13: Intent Classification Result German

Chapter 8

Discussion

The price comparison of Table 7.1 shows that all tested frameworks can be used for free. The non-free tiers of Dialogflow and LUIS in combination with the Microsoft Bot framework are charged based on the number of requests while the price for Watson Assistant is calculated based on the number of users. The price for the enterprise edition of Rasa and the premium edition of Watson Assistant is not stated on the websites. The prices are hard to compare but all are offered in a free tier and the bots of Google and Microsoft cost roughly the same per request. The disadvantage of Rasa is the deployment since the costs for the infrastructure are unclear for a similar workload. Since Watson Assistant calculated the price based on the number of users it offers a better value when the users communicate with the bot frequently or have long conversations. The prices for the speech functionality are comparable for Dialogflow and the Microsoft speech-services. The IBM text-to-speech service is more expensive than the one of Dialogflow and Microsoft. Rasa doesn't offer any speech services. This means that the Microsoft speech-service and Dialogflow are the best when it comes to the price. Watson is second and Rasa doesn't show any prices at all. A big disadvantage of cloud-based solutions is the dependency on the company. It's not easy to change the provider and take the current chatbot with you if it is a cloud-based chatbot. Rasa differs at that point since it can run in any cloud environment inside a docker container. Every cloud provider offers the option to host docker containers in their cloud. A major factor is security since the users enter sensitive information in the chatbot communication and the data is processed in the cloud of the provider. The cloud provider hence has access to the sensitive information which can be problematic with bank account information for instance. Rasa has no such problem since the data is processed inside of the docker container or in a local environment. The entity extraction section shows that Dialogflow and LUIS define the required entities while Watson lacks a person entity. Rasa doesn't offer any entities but can use the entities of Spacy and Duckling. If both are used then no entities need to be defined in Rasa. The result in Table 7.4 shows that Dialogflow has the best person, date, and date-span entity and dominated the entity extraction tests. The best overall performance was achieved by Dialogflow followed by Rasa. No technology performed exceptionally bad in this category and the date-span entity of Duckling, Dialogflow, and Watson Assistant reached the same f-score which means they performed on an equal level. The intent classification result of Table 7.12 shows that the cloud services performed on an equal level when it comes to precision, recall, the f-score, and the average confidence. Rasa was unable to keep up with the cloud technologies and especially the average confidence score is low. Watson Assistant and Dialogflow were able to reach a perfect f-score of 1.0. The standard devia-

tion is lowest for Watson which puts Watson Assistant in the top spot closely followed by Dialogflow and LUIS. Dialogflow extracted an entity incorrectly in test case four of Table 7.10. Rasa showed a good performance when the f-score is used as a measure but has a low average confidence score of 46%. The confidence scores of Dialogflow, LUIS, and Watson are far better with 77%, 75%, and 74%. In this category, Dialogflow, LUIS, and Watson performed good enough to be recommended. The confidence scores of Rasa can be improved through more training data. The German language result of Table 7.13 shows some interesting differences to the English tests of Table 7.12. The average confidence increased for all four technologies and the standard deviation decreased significantly for four technologies. For the German language, the f-score values also increased. All in all the German language result looks very good on the surface when only the intent classification is considered. The entity extraction performance dropped significantly on the test sentences. The German tests are shown in the appendix in Table 10.4 and 10.5. There were a total of 21 failed extractions for the German test sentences and a total of three failed extractions for the English test sentences on a total of 64 test sentences each. This means in German the entity extraction failed on every third test sentence. The only technology which delivered a satisfying entity extraction result on the German test sentences with zero mistakes total is LUIS. The performance of Rasa was worse than the performance of LUIS but it was good enough for a recommendation. The date entity of Dialogflow and Watson Assistant are unable to cope with the German date format. The person entity of Dialogflow did not work reliably in German although the names aren't different from the ones used in the English test sentences. Braun et al. [15] stated that if the training data is sparse there is almost no performance difference between the chatbot frameworks regarding NLU. This can be confirmed for the cloud frameworks since there is no big difference regarding the f-score, average confidence, and entity extraction, shown in in Tables 7.4, 7.12, 7.13. The local solution Rasa was measurably worse. The communication of the frameworks works the same way for all the technologies. There is no big difference no matter what is used. LUIS and Dialogflow offer metadata by default while Rasa and Watson Assistant don't. All four offer REST endpoints and use JSON format for messages. The easiest technology is LUIS because it's just an NLU framework and doesn't handle the dialog structure. Dialogflow is the easiest chatbot framework because it has the simplest UI, the most predefined entities, predefined intents for common tasks like a yes/no path and small talk can be enabled in the menu which saves development time. Watson is comparable to Dialogflow when it comes to the creation of entities and intents but the UI is less intuitive, there is no small talk feature, there are almost no predefined entities, and offer no predefined intents for common tasks. The most complex is Rasa simply for the facts that the base version has no UI and programming skills are required. To add predefined entities the pipeline needs to be adjusted, it has no predefined intents, and a small talk feature is not available. The creation of dialogs with Rasa is far more complicated as is the collection of form data. Compared to the cloud frameworks Rasa is incredibly complicated and doesn't offer a lot of features by default. The setup of Rasa is complicated compared to the cloud chatbots which was expected. To use a cloud chatbot only a browser is needed, for Rasa either Docker is required or a Python environment. It's harder to set up Rasa because the cloud technologies run already in the cloud and require no additional deployment. The comparably complex setup has advantages when it comes to deployment. Rasa can be deployed locally and in every cloud environment while the cloud services only run in the cloud of the provider. The deployment evaluation met the expectations. Rasa is flexible and the cloud chatbots are simple

when it comes to deployment. Since Rasa can be deployed anywhere, doesn't rely on the cloud, and can run without an internet connection it's the only real choice of the four for security-critical applications. Rasa is more complicated but it was possible to implement the design with all chatbot technologies. This means that all three technologies can be recommended when only the implementation of the design is considered. The basic concepts for the implementation are the same for all technologies, there are intents, entities, and utterances. Dialogflow and Watson Assistant use dialog nodes for the conversation structure while Rasa uses stories. But in the end, there is no big difference when it comes to the concepts. The only technology which allows customization is Rasa. If something very specific is needed or tuning is required then Rasa should be used.

Chapter 9

Conclusion

The domain consisting of the sickness notification and the vacation request can be solved with chatbots. It was possible to implement the design with all three chatbot technologies namely Dialogflow, Rasa, and Watson Assistant. Based on Chapter 6 the implementation is easiest with Dialogflow then comes Watson Assistant followed by Rasa which is the most complicated by far. It's easiest with Dialogflow because it already has prebuilt entities, it offers the utility intents yes and no, and small talk can be included with a single mouse click. Watson Assistant is second because it takes more time to get accustomed to the UI, the person entity needs to be defined by the developer, and the yes/no logic needs to be implemented. Rasa is the most complex one since it doesn't provide any entities itself. To enable the entities of other technologies like Spacy and Duckling the pipeline needs to be adjusted. To connect to Duckling and Additional Docker container needs to be started. The chatbot developer also needs to implement the form filling with Python when using Rasa. The other technologies don't require any programming skills to model the dialog. The implementation with Rasa took by far the longest. When only the implementation section is used for the ranking Dialogflow wins in front of Watson followed by Rasa. Table 7.2 shows the number of predefined entities. When only the domain of this thesis is considered then the only technology where an entity is missing would be Watson Assistant since it provides no person entity. When the number of entities is considered for a general case then the top technology by a large margin is Dialogflow. The number of predefined entities is an important number for future use-cases since lots of development time can be saved when the entities are already predefined. They are also trained well because they use large training sets. Table 7.4 ranks the frameworks based on their entity extraction capabilities. The best person, date and date-span entities are present at Dialogflow because they achieved the highest f-score values. In the date-span entity category, Duckling and Watson achieved the same f-score as Dialogflow. The best overall entity extraction performance is achieved by Dialogflow. From the entity extraction capabilities alone either Rasa or Dialogflow needs to be recommended. The other frameworks performed worse in the entity extraction department. The ranking shows that some technologies performed better than others. It also shows that no framework performed badly. The intent classification is dominated by Watson and Dialogflow since both were able to classify all test cases correctly. LUIS is a close second. All three cloud chatbots performed on an equal level. The detailed result is shown in Table 7.12. Rasa was unable to keep up with Dialogflow, LUIS, and Watson when the average confidence is compared. This is an indication that Rasa needs more training data to reach the same performance level as Dialogflow, LUIS, and Watson. An average confidence value below

50% is not satisfactory. Based on the entity extraction and intent recognition results the recommended technology in the top spot has to be Dialogflow since the performance in both categories was great. Additionally, Dialogflow has the best average confidence value. It's also possible to use LUIS and Watson Assistant since the performance was close to Dialogflow. It's possible to improve the performance and confidence scores of Rasa with more training data. Rasa needs to improve to be recommended. The intent classification was also done for the German language and the results differ greatly from the English results. The average confidence increased for each technology. The f-score of LUIS and Rasa increased significantly. This puts LUIS on par with Dialogflow and Watson with a perfect f-score value of 1.0. This leads to the conclusion that the intent classification works better for the German language in general since all technologies performed better. Contrary to the intent classification, the entity-extraction performance decreased significantly. In the English test sentences of Tables 7.10 and 7.11 the entity extraction didn't work correctly in two cases and worked in 46 cases. For the German language, test shown in Tables 10.4 and 10.5, a total of 21 wrong and 26 correct extractions are counted. The only technology that achieved a satisfying entity extraction result on the German test sentences is LUIS. Rasa is the only portable technology. It can run in a Docker container and Docker containers are supported by every cloud provider. It can also run locally on a development machine and on an internal server. The cloud technologies are tied to the cloud provider who offers the bot. Rasa wins the portability comparison. The migration from one chatbot technology to another is not easily possible with any of the technologies tested. Hence, there is no evaluation possible and no ranking can be done. The pricings of the frameworks shown in Table 7.1 are hard to compare and intransparent for the premium version of Watson Assistant and the enterprise edition for Rasa. All three offer a free tier which can be used by companies. Because of the lack of transparency, the comparison can't be used for a recommendation or a ranking. The real difference is that Watson Assistant is measured by the number of users per month and Dialogflow and LUIS + Microsoft Bot framework are based on the number of requests. This makes Dialogflow and LUIS cheaper for small amounts of requests and lots of users with short conversations. Watson is cheaper for returning users with lots of conversation or very long conversations. The prices for the speech capabilities of Dialogflow and the Microsoft Speech Services are comparable. The same functionality is more expensive when the services of IBM are used. Worst in this category is Rasa since it offers no speech services but it's possible to use the speech services of the cloud providers. The winners of the speech capability evaluation are Dialogflow and the Microsoft Speech Services. When security is an important factor then a local technology needs to be used. The cloud technologies all process the requests in the cloud and this has to be kept in mind. If it's impossible to process information in the cloud because there are security restrictions then the recommendation is to use Rasa. When all the factors discussed above are combined the clear winner for English bots is Dialogflow and the winner for German bots is LUIS. The design could be implemented fastest with Dialogflow, it's the easiest of the chatbot technologies, it offers the most predefined entities, provides utility intents, had the best person and date entity in the entity extraction tests has the highest average confidence score, and performed great in the intent classification tests. The only downside of Dialogflow is the entity extraction performance on the German test sentences. The best overall technology is LUIS and is the recommended framework for future development. The only negative aspect of Rasa is the low average confidence score but it can be improved with more training data and is a good local alternative to LUIS. When looking at the differences and similarities of the chatbot

framework the conclusion is that they are very similar and differ only in small aspects. Especially the cloud frameworks are almost identical in functionality, the provided services, and the costs. All tested frameworks use the concepts intent, entity, and utterance. The chatbot frameworks use a mechanism to structure the dialog with the user. Dialogflow and Watson Assistant use dialog nodes and Rasa uses stories. The development effort and complexity between the technologies vary. Cloud chatbots require less knowledge and are easier to use while the local technology (Rasa) offers more customization possibilities. The pipeline of Rasa can be adjusted to the specific needs of the developer or company which is not offered by any cloud technology. The third research question is about the suitability of problems for the use of chatbots. The three major criteria mentioned by Singh et al. [32] can be confirmed. A problem needs to be solvable via simple back and forth communication because it's the only thing a chatbot can do. The problem is not required to be repetitive but it's only cost-efficient if it is a repetitive problem. It should also be automatable to be efficient and to make the chatbot work independent. In general, all problems which satisfy those three criteria are suitable for the use of chatbots. Possible future use cases for the 3 Banken IT are the use of a chatbot for a banking application like mentioned in Singh et al. [32]. The banking app of [32] is used for transactions and querying account information. A popular use case for chatbots are FAQs (Shawar and Atwell [30], Raj [29], Huang et al. [24], Go and Sundar [20]). When resources like [] and [] are taken into account the psychological aspects of chatbot development have to be considered for future work. Like Go and Sundar [20] said, there are multiple ways to make users believe that they are talking to a bot. For an actual chatbot, there needs to be decided if the users should know that they are talking to a bot or not. If they believe to talk to a bot they expect less but the conversation is generally treated more machine-like [20]. When users assume to talk to a human the conversation is judged as more natural but if the bot performs badly the experience is also evaluated as more negative [20]. Like Shawar and Atwell [31] said, a chatbot should not imitate humans they should be used to reduce their workload and help them. For prototyping the generation of training data by the developer is appropriate. It's not appropriate for the training of the actual product since the performance requirements are much higher. The training data for the future work on the actual product can be collected through various real-life conversation sources like emails and conversations between customers and experts. With real-life data, the training data is as realistic as possible and the quality of the chatbot increases. User-studies are recommended to increase the quality of a chatbot further since the feedback of actual users is the most valuable to improve the actual systems.

Chapter 10

Appendix

```
1 language: "en"
2 pipeline:
3   - name: "SpacyNLP"
4   - name: "SpacyTokenizer"
5   - name: "SpacyFeaturizer"
6   - name: "RegexFeaturizer"
7   - name: "DucklingHTTPExtractor"
8     url: "http://localhost:8000"
9   - name: "SpacyEntityExtractor"
10  - name: "EntitySynonymMapper"
11  - name: "SklearnIntentClassifier"
```

Listing 10.1: Detailed Pipeline Configuration for Rasa

```
1 class SicknessForm(FormAction):
2
3     def name(self) -> Text:
4         return "sickness_form"
5
6     @staticmethod
7     def required_slots(tracker: Tracker) -> List[Text]:
8         return ["PERSON", "return_date"]
9
10    def slot_mappings(self) -> Dict[Text, Union[Dict, List[Dict]]]:
11        return {
12            "PERSON": [self.from_entity(entity="PERSON"),
13                      self.from_text()],
14            "return_date": [self.from_entity(entity="DATE"),
15                          self.from_text()]
16        }
17
18    def submit(
19        self,
20        dispatcher: CollectingDispatcher,
21        tracker: Tracker,
22        domain: Dict[Text, Any],
23    ) -> List[Dict]:
24
25        return []
```

Listing 10.2: Rasa Slot Filling Action

```

1 ## Sickness Affirm
2 * sickness
3   - sickness_form
4   - form{"name":"sickness_form"}
5 * submit{"PERSON":"Anna Maria Mayer"}
6   - sickness_form
7   - slot{"PERSON":"Anna Maria Mayer"}
8 * submit{"time":"2020-05-30T00:00:00.000-07:00","DATE":"today"}
9   - sickness_form
10  - form{"name":null}
11  - slot{"time":"2020-05-30T00:00:00.000-07:00"}
12 * affirm
13   - action_submit_sickness_data
14   - action_restart

```

Listing 10.3: Rasa Sickness Story

| No | Utterance | Person Entity | Date Entity |
|----|--|---------------|-------------|
| 1 | Ich bin krank | X | X |
| 2 | Krankmeldung | X | X |
| 3 | krank melden | X | X |
| 4 | [Alfred Mayer] ist krank | ✓ | X |
| 5 | meine Kollegin [Maria Müller] ist bis [Mittwoch] krank | ✓ | ✓ |
| 6 | Kollegin [Simone Bauer] ist krank | ✓ | X |
| 7 | melde [Franz Dorfer] krank | ✓ | X |
| 8 | [Stefan Weber] ist krank bis [Freitag] | ✓ | ✓ |
| 9 | krank [Emma Wagner] [6.Juni] | ✓ | ✓ |
| 10 | [Sophia Richter] ist heute krank | ✓ | X |

Table 10.1: Sickness Training Utterances German

| No | Utterance | Person Entity | Date-Span Entity |
|----|---|---------------|------------------|
| 1 | vacation | X | X |
| 2 | holiday | X | X |
| 3 | I need a vacation | X | X |
| 4 | colleague [Johan Schmidt] requested a vacation | ✓ | X |
| 5 | I'm on holiday [from Monday to Friday] | X | ✓ |
| 6 | vacation [Mia Koch] [06/20 to 06/30] | ✓ | ✓ |
| 7 | [Adam Schneider] is on holiday [from June 6 to June 14] | ✓ | ✓ |
| 8 | holiday request | X | X |
| 9 | vacation [from 12th of August to 30th of August] | X | ✓ |
| 10 | [Lea Klein] requested a vacation | ✓ | X |

Table 10.2: Vacation Training Utterances English

| No | Utterance | Person Entity | Date-Span Entity |
|----|---|---------------|------------------|
| 1 | Urlaub | ✗ | ✗ |
| 2 | Urlaubsantrag | ✗ | ✗ |
| 3 | Ich brauche Urlaub | ✗ | ✗ |
| 4 | Kollege [Johan Schmidt] hat Urlaub beantragt | ✓ | ✗ |
| 5 | Ich bin von [Montag bis Freitag] im Urlaub | ✗ | ✓ |
| 6 | urlaub beantragen [Mia Koch] [20/06 bis 30/06] | ✓ | ✓ |
| 7 | [Adam Schneider] ist [vom 6.Juni bis 14.Juni] im Urlaub | ✓ | ✓ |
| 8 | Ich beantrage Urlaub | ✗ | ✗ |
| 9 | Urlaub [vom 12.August bis zum 30.August] | ✗ | ✓ |
| 10 | [Lea Klein] hat einen Urlaubsantrag eingereicht | ✓ | ✗ |

Table 10.3: Vacation Training Utterances German

```

1 "query": "sick Anna Maier 06/10",
2 "prediction": {
3   "topIntent": "sick_",
4   "intents": {
5     "sick_": {
6       "score": 0.6760214
7     },
8     "vacation_": {
9       "score": 0.224106133
10    },
11    "None": {
12      "score": 0.006858599
13    }
14  }}

```

Listing 10.4: LUIS Intent Response

```

1 "entities": {
2   "datetimeV2": [{
3     "type": "date",
4     "values": [{
5       "timex": "XXXX-06-10",
6       "resolution": [
7         { "value": "2020-06-10" },
8         { "value": "2021-06-10" } ] ] } ]
9   },
10  "$instance": {
11    "datetimeV2": [{
12      "type": "builtin.datetimeV2.date",
13      "text": "06/10",
14      "startIndex": 16,
15      "length": 5,
16      "modelTypeId": 2,
17      "modelType": "Prebuilt Entity Extractor",
18      "recognitionSources": [ "model" ] } ]
19  }
20 }

```

Listing 10.5: LUIS Entity Response

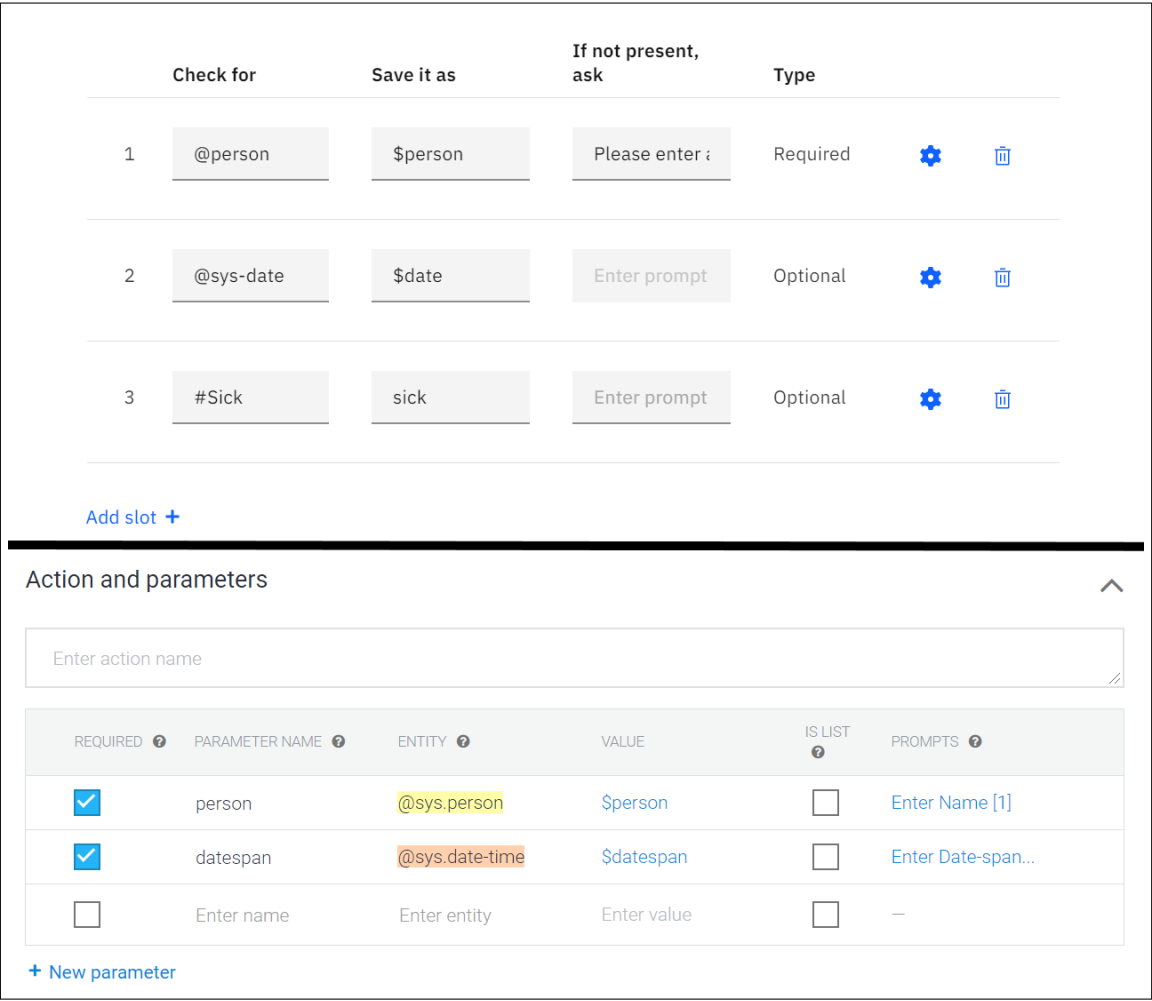


Figure 10.1: From Filling in Watson Assistant (top) and Dialogflow (bottom)

| No | Input | Technology | Intent Match | Entity Match | Confid. [%] |
|----|---|------------|--------------|--------------|-------------|
| 1 | krank | Watson | ✓ | - | 91.0 |
| | | Dialogflow | | | 77.1 |
| | | Rasa | | | 93.6 |
| | | LUIS | | | 99.8 |
| 2 | Emilia Schwarz ist krank | Watson | ✓ | ✗ | 94.0 |
| | | Dialogflow | | ✓ | 100 |
| | | Rasa | | | 72.7 |
| | | LUIS | | | 92.7 |
| 3 | Lea Schmitt bis Mittwoch krank gemeldet | Watson | ✓ | ✗ | 95.0 |
| | | Dialogflow | | ✓ | 80.0 |
| | | Rasa | | ✗ | 43.1 |
| | | LUIS | | ✓ | 79.1 |
| 4 | Ich muss mich krank melden | Watson | ✓ | - | 97.0 |
| | | Dialogflow | | | 70.5 |
| | | Rasa | | | 45.3 |
| | | LUIS | | | 95.0 |
| 5 | Paul Armstrong ist krank und kommt am Montag wieder | Watson | ✓ | ✗ | 93.0 |
| | | Dialogflow | | ✓ | 73.5 |
| | | Rasa | | | 41.6 |
| | | LUIS | | | 64.8 |
| 6 | krank Anna Maier 06/10 | Watson | ✓ | ✗ | 93.0 |
| | | Dialogflow | | ✓ | 100.0 |
| | | Rasa | | ✗ | 60.2 |
| | | LUIS | | ✓ | 72.1 |
| 7 | Ich komme heute nicht zur Arbeit weil ich krank bin | Watson | ✓ | - | 96.0 |
| | | Dialogflow | ✓ | | 45.2 |
| | | Rasa | ✗ | ✗ | 33.4 |
| | | LUIS | ✓ | | 75.9 |
| 8 | Anna Lehner fühlt sich krank. Sie wird Dienstag wieder da sein. | Watson | ✓ | ✗ | 94.0 |
| | | Dialogflow | | ✓ | 100 |
| | | Rasa | | | 42.4 |
| | | LUIS | | | 57.6 |

Table 10.4: Sickness Intent Classification German

| No | Input | Technology | Intent Match | Entity Match | Confid. [%] |
|----|---|------------|--------------|--------------|-------------|
| 1 | Urlaub | Watson | ✓ | - | 100.0 |
| | | Dialogflow | | | 100.0 |
| | | Rasa | | | 59.8 |
| | | LUIS | | | 98 |
| 2 | mein Kollege Franz Bauer möchte Urlaub beantragen | Watson | ✓ | ✓ | 93.0 |
| | | Dialogflow | | | 100.0 |
| | | Rasa | | | 45 |
| | | LUIS | | | 88.8 |
| 3 | Urlaub Martin Huber 12.Juli bis 13.07.2020 | Watson | ✓ | ✗ | 95.0 |
| | | Dialogflow | | ✗ | 100.0 |
| | | Rasa | | ✗ | 44.5 |
| | | LUIS | | ✓ | 91.4 |
| 4 | Urlaubsantrag 12/07 bis 13.Juli | Watson | ✓ | ✓ | 100.0 |
| | | Dialogflow | | ✗ | 100.0 |
| | | Rasa | | ✗ | 91.7 |
| | | LUIS | | ✓ | 98.3 |
| 5 | Urlaub beantragen für den Zeitraum 28.September - 2.Oktober | Watson | ✓ | ✗ | 97.0 |
| | | Dialogflow | | ✗ | 54.9 |
| | | Rasa | | ✓ | 69.1 |
| | | LUIS | | ✓ | 96.4 |
| 6 | Ich brauche unbedingt Urlaub auf einer einsamen Insel | Watson | ✓ | - | 96.0 |
| | | Dialogflow | | ✗ | 80.5 |
| | | Rasa | | - | 54.2 |
| | | LUIS | | - | 83.2 |
| 7 | Kollege Helmut Kerschbaum hat Urlaub beantragt | Watson | ✓ | ✗ | 97.0 |
| | | Dialogflow | | ✓ | 100.0 |
| | | Rasa | | ✓ | 66.7 |
| | | LUIS | | ✓ | 90.5 |
| 8 | Kollege Helmut Kerschbaum will Urlaub. Er wird von 10/09 bis 24/09 nicht erreichbar sein. | Watson | ✓ | ✗ | 96.0 |
| | | Dialogflow | | ✗ | 100.0 |
| | | Rasa | | ✗ | 53.0 |
| | | LUIS | | ✓ | 93.8 |

Table 10.5: Vacation Intent Classification German

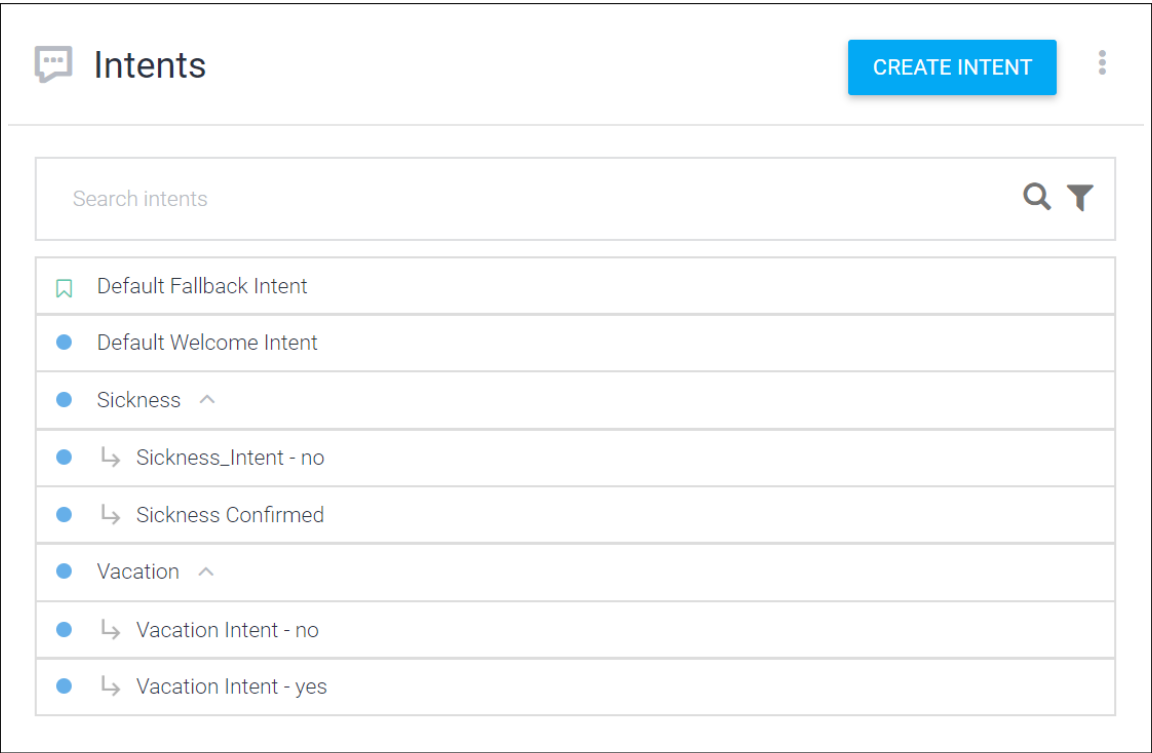


Figure 10.2: Dialog Nodes in Dialogflow

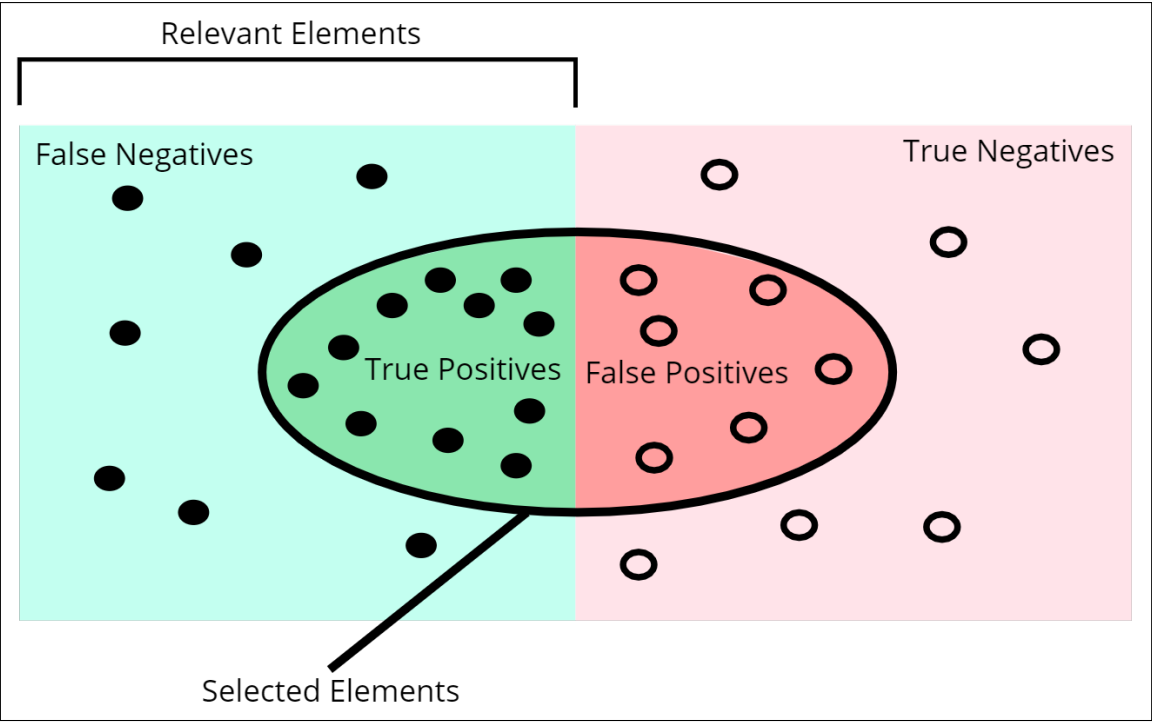


Figure 10.3: True and False Positives and Negatives

Bibliography

- [1] 3 banken it company. <https://www.3bankenit.at/company/company.xhtml>, . Accessed: 2020-03-07. (Cited on page 3.)
- [2] 3 banken it organigram. <https://www.3bankenit.at/company/organigram.xhtml>, . Accessed: 2020-03-07. (Cited on page 3.)
- [3] 3 banken it services. <https://www.3bankenit.at/competence/services.xhtml>, . Accessed: 2020-03-07. (Cited on page 3.)
- [4] Dialogflow. <https://cloud.google.com/dialogflow>. Accessed: 2020-05-09. (Cited on pages 6, 36, 49, and 81.)
- [5] Duckling. <https://duckling.wit.ai/>. Accessed: 2019-05-29. (Cited on page 42.)
- [6] Iso/iec 25010. <https://iso25000.com/index.php/en/iso-25000-standards/iso-25010>. Accessed: 2020-05-09. (Cited on pages 20, 21, 26, 27, 28, and 46.)
- [7] Language understanding (luiz) documentation. <https://docs.microsoft.com/en-us/azure/cognitive-services/luiz/>. Accessed: 2019-12-26. (Cited on pages 6, 44, 49, and 81.)
- [8] Build contextual assistants that really help customers. <https://rasa.com/>. Accessed: 2020-05-09. (Cited on pages 6, 40, 49, and 81.)
- [9] Spacy api. <https://spacy.io/api/annotation>. Accessed: 2019-05-28. (Cited on page 42.)
- [10] Watson assistant. <https://www.ibm.com/cloud/watson-assistant/>. Accessed: 2020-05-09. (Cited on pages 6, 38, 49, and 81.)
- [11] N. S. Ahmad, M. H. Sanusi, M. H. Abd Wahab, A. Mustapha, Z. A. Sayadi, and M. Z. Saringat. Conversational bot for pharmacy: A natural language approach. In *2018 IEEE Conference on Open Systems (ICOS)*, pages 76–79, Nov 2018. doi: 10.1109/ICOS.2018.8632700. (Cited on pages 18, 19, and 23.)
- [12] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*, 2017. (Cited on pages 6, 18, 19, and 23.)

- [13] Antoine Bordes, Y-Lan Boureau, and Jason Weston. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*, 2016. (Cited on pages 13, 14, 15, 16, 18, and 23.)
- [14] Petter Bae Brandtzaeg and Asbjørn Følstad. Chatbots: changing user needs and motivations. *interactions*, 25(5):38–43, 2018. (Cited on pages 5, 17, 19, 20, 24, 33, and 34.)
- [15] Daniel Braun, Adrian Hernandez Mendez, Florian Matthes, and Manfred Langen. Evaluating natural language understanding services for conversational question answering systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 174–185, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5522. (Cited on pages 1, 2, 5, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, and 65.)
- [16] Aditya Deshpande, Alisha Shahane, Darshana Gadre, Mrunmayi Deshpande, and Prachi M Joshi. A survey of various chatbot implementation techniques. *International Journal of Computer Engineering and Applications*, 11, 2017. (Cited on pages 1, 5, 13, 14, and 22.)
- [17] Debasatwa Dutta. Developing an intelligent chat-bot tool to assist high school students for learning general knowledge subjects. Technical report, Georgia Institute of Technology, 2017. (Cited on pages 5, 7, 8, 17, 18, 19, 20, 23, and 24.)
- [18] Asbjørn Følstad and Petter Bae Brandtzaeg. Chatbots and the new world of hci. *interactions*, 24(4):38–42, 2017. (Cited on pages 17, 19, 20, 22, 24, 27, and 31.)
- [19] Kelly Geyer, Kara Greenfield, Alyssa Mensch, and Olga Simek. Named entity recognition in 140 characters or less. In *#Microposts*, pages 78–79, 2016. (Cited on pages 9, 13, 14, and 24.)
- [20] Eun Go and S. Shyam Sundar. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, 97:304 – 316, 2019. ISSN 0747-5632. doi: <https://doi.org/10.1016/j.chb.2019.01.020>. URL <http://www.sciencedirect.com/science/article/pii/S0747563219300329>. (Cited on pages 1, 17, 20, 24, and 69.)
- [21] N. A. Godse, S. Deodhar, S. Raut, and P. Jagdale. Implementation of chatbot for itsm application using ibm watson. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–5, Aug 2018. doi: 10.1109/ICCUBEA.2018.8697411. (Cited on page 23.)
- [22] Eric Gregori. Evaluation of modern tools for an omscs advisor chatbot. 2017. (Cited on pages 1, 5, 10, 17, 18, 19, 23, and 24.)
- [23] Haruo Hosoya and Benjamin Pierce. Regular expression pattern matching for xml. In *Proceedings of the 28th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 67–80, 2001. (Cited on page 44.)
- [24] Jizhou Huang, Ming Zhou, and Dan Yang. Extracting chatbot knowledge from online discussion forums. In *IJCAI*, volume 7, pages 423–428, 2007. (Cited on pages 1, 5, 10, and 69.)

- [25] Danielle A Kane. The role of chatbots in teaching and learning. *E-Learning and the Academic Library: Essays on Innovative Initiatives*, 131, 2016. (Cited on pages 2, 5, 13, and 23.)
- [26] Amazon Lex. Conversational interfaces for your applications. *Powered by the same deep learning technologies as Alexa*. URL <https://aws.amazon.com/lex>. (Cited on page 17.)
- [27] Aleksandra Przegalinska, Leon Ciechanowski, Anna Stroz, Peter Gloor, and Grzegorz Mazurek. In bot we trust: A new methodology of chatbot performance measures. *Business Horizons*, 62(6):785 – 797, 2019. ISSN 0007-6813. doi: <https://doi.org/10.1016/j.bushor.2019.08.005>. URL <http://www.sciencedirect.com/science/article/pii/S000768131930117X>. (Cited on pages 18 and 19.)
- [28] AM Rahman, Abdullah Al Mamun, and Alma Islam. Programming challenges of chatbot: Current and future prospective. In *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pages 75–78. IEEE, 2017. (Cited on pages 1, 7, 14, and 23.)
- [29] Sumit Raj. *Building chatbots with Python: using natural language processing and machine learning*. Springer, 2019. (Cited on pages iv, 2, 3, 6, 7, 9, 12, 13, 14, 23, 26, and 69.)
- [30] Bayan Abu Shawar and Eric Atwell. Different measurements metrics to evaluate a chatbot system. In *Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies*, pages 89–96. Association for Computational Linguistics, 2007. (Cited on pages 1, 5, 10, 14, 15, 16, 18, 24, and 69.)
- [31] Bayan Abu Shawar and Eric Atwell. Chatbots: are they really useful? In *Ldv forum*, volume 22, pages 29–49, 2007. (Cited on pages 1, 5, 10, and 69.)
- [32] Abhishek Singh, Karthik Ramasubramanian, and Shrey Shivam. Building an enterprise chatbot. (Cited on pages 1, 5, 6, 8, 9, 16, 18, 22, 23, 26, 46, and 69.)
- [33] S Shyam Sundar, Saraswathi Bellur, Jeeyun Oh, Haiyan Jia, and Hyang-Sook Kim. Theoretical importance of contingency in human-computer interaction: effects of message interactivity on user engagement. *Communication Research*, 43(5):595–625, 2016. (Cited on page 20.)
- [34] Dana Vrajitoru, Jacob Ratkiewicz, et al. Evolutionary sentence combination for chatterbots. In *International Conference on Artificial Intelligence and Applications (AIA 2004)*, pages 287–292, 2004. (Cited on page 5.)
- [35] Jason D Williams, Eslam Kamal, Mokhtar Ashour, Hani Amr, Jessica Miller, and Geoff Zweig. Fast and easy language understanding for dialog systems with microsoft language understanding intelligent service (luís). In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 159–161, 2015. (Cited on pages 5, 6, 18, 19, 22, and 23.)
- [36] Jason D Williams, Kavosh Asadi, and Geoffrey Zweig. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. *arXiv preprint arXiv:1702.03274*, 2017. (Cited on pages 5, 13, 15, 16, 18, 22, and 23.)

List of Figures

| | | |
|------|--|----|
| 2.1 | Example FAQ Conversation with Dialogflow | 10 |
| 2.2 | Greeting Story | 11 |
| 5.1 | Minimalistic Chatbot Architecture | 30 |
| 5.2 | Sickness Notification Example Conversation | 31 |
| 5.3 | Chatbot Conversation Flowchart | 32 |
| 5.4 | Chatbot UI | 35 |
| 6.1 | Dialogflow Web Interface | 38 |
| 6.2 | Watson Assistant Web Interface | 40 |
| 6.3 | Rasa X Interface | 43 |
| 6.4 | LUIS Web Interface | 44 |
| 10.1 | From Filling in Watson Assistant (top) and Dialogflow (bottom) | 73 |
| 10.2 | Dialog Nodes in Dialogflow | 76 |
| 10.3 | True and False Positives and Negatives | 76 |

List of Tables

| | | |
|------|--|----|
| 2.1 | Default Welcome Intent of Dialogflow | 8 |
| 2.2 | Examples for Concept Explanation | 9 |
| 2.3 | Entity examples | 9 |
| 4.1 | Sickness Training Utterances English | 24 |
| 5.1 | Analysis of Figure 5.3 | 33 |
| 5.2 | Sickness Notification Intents, Entities, and Actions of Figure 5.2 | 34 |
| 6.1 | Conditional Responses of Watson Assistant | 40 |
| 7.1 | Pricing of Frameworks [8, 4, 10, 7] | 49 |
| 7.2 | Predefined Entities | 50 |
| 7.3 | Framework Entity Recognition and Extraction | 51 |
| 7.4 | Entity Recognition Result Evaluation | 52 |
| 7.5 | Person Entity Recogniton and Extraction | 53 |
| 7.6 | Date Entity Recogniton and Extraction | 54 |
| 7.7 | Date Entity Recogniton and Extraction 2 | 58 |
| 7.8 | Date Span Entity Recogniton and Extraction 1 | 59 |
| 7.9 | Date Span Entity Recogniton and Extraction 2 | 60 |
| 7.10 | Sickness Intent Classification | 61 |
| 7.11 | Vacation Intent Classification | 62 |
| 7.12 | Intent Classification Result | 63 |
| 7.13 | Intent Classification Result German | 63 |
| 10.1 | Sickness Training Utterances German | 71 |
| 10.2 | Vacation Training Utterances English | 71 |
| 10.3 | Vacation Training Utterances German | 72 |
| 10.4 | Sickness Intent Classification German | 74 |
| 10.5 | Vacation Intent Classification German | 75 |

List of Listings

| | | |
|------|--|----|
| 2.1 | Dialogflow Webhook Request Example | 11 |
| 6.1 | Dialogflow Webhook Entry Point | 37 |
| 6.2 | Dialogflow Request Parameters | 37 |
| 6.3 | Dialogflow Intent Handling | 37 |
| 6.4 | Watson Assistant Request Format | 40 |
| 6.5 | Rasa Docker Compose File | 40 |
| 6.6 | Rasa Intent Format | 41 |
| 6.7 | Rasa Story Format | 41 |
| 6.8 | Rasa Configuration | 42 |
| 10.1 | Detailed Pipeline Configuration for Rasa | 70 |
| 10.2 | Rasa Slot Filling Action | 70 |
| 10.3 | Rasa Sickness Story | 71 |
| 10.4 | LUIS Intent Response | 72 |
| 10.5 | LUIS Entity Response | 72 |