UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering, Mathematics & Computer Science

An Exploratory Study to Design and Evaluate a Dialogue System for a Robot Assisting the Elderly in a Care Environment

Sweta Balamurali

Master of Science

June 2022

Human Media Interaction(HMI)

Supervisors:
dr. Roeland J.F. Ordelman
dr. Mariët Theune
University of Twente

Ir. Dimitris Karageorgos
Heemskerk Innovative Technology

Examiner:
dr. Cora Salm
University of Twente

Abstract

With a growing gap between the number of care professionals to the elderly, socially assistive robots (SARs) show a potential in helping the elderly with activities of daily living (ADL). Communication using natural language is one of the most natural forms of human-robot interaction. We conduct an exploratory study to design and evaluate a dialogue system for a care robot assisting the elderly with daily activities in a care environment. We implement a dialogue system for a care robot using the Rasa framework and tested the dialogue system with users. Our evaluations obtained an F1-score of over 60% for entity extraction and response selection to unseen data, with 50% successful task completions and 29.1% partial successes. We analysed the interaction to find the limitations of the system and provide recommendations for future developments of the system.

The work in this thesis was conducted with the support of Heemskerk Innovative Technology (HiT). Their cooperation is gratefully acknowledged.

Acknowledgement

This project has been conducted with the collaboration of Heemskerk Innovative Technology (HIT) and University of Twente and would not have been possible without the support of many people.

Firstly I would like to thank Cock Heemskerk and Dimitris Karageorgos for giving me an opportunity to work on a topic for Robot Rose as HiT. I am grateful for the flexible and supportive environment and all peers from whom I have learnt a lot.

I would like to extend my gratitude to my supervisors at Twente, Roeland Ordelman and Mariet Theune, who read my numerous revisions and provided their invaluable advice at every stage. Not only have they been patient and understanding throughout this work, but were their kind words helped me push through when times were difficult.

I would like to thank my parents for always believing in me and supporting my dreams. I'm thankful to my brother Sushant for always being proud of me.

I'm grateful to my friends Aarthi, Archana, Shadab and Vijay for being constant cheerleaders from miles away and to Aswin, Nischit, Shalvi and Vishakha for the memories and courage.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The population demographic is rapidly changing. As a result of increased life expectancy, it is predicted that the number of people over the age of 60 will be more than those under 15 years by the year 2050 (Buettner, 2015). There will be an increased need for healthcare services, but the number of healthcare professionals are decreasing (Bengtsson & Qi, 2018) which impacts the quality of the health-care especially for the elderly. To address the shortage of healthcare professionals, growing research in robotics show that *Socially Assistive Robots(SARs)* have a potential in helping the elderly improve and maintain their well-being (Broekens et al., 2009), (Tokunaga et al., 2021). A survey shows that the elderly want to maintain independent living and are generally open to robot assistance for certain activities like chores and manipulating objects, but prefer a human to help with personal care (Smarr et al., 2014). To interact with the elderly, an assistive social robot should be capable of communicating through natural language using a dialogue system, since spoken dialogue is considered the most natural form of human-robot interaction (De Carolis et al., 2021). This brings us to the use-case and scope for this research, *A care robot capable of communicating using natural language to assist the elderly with daily activities in a healthcare environment*.

1.2 Socially Assistive Robots for Elderly

SARs can assist the elderly with Activities of Daily Living (ADL). Activities of Daily Living (ADL) are basic skills essential for maintaining independence like grooming, mobility, personal hygiene and eating (Smarr et al., 2014). SARs provide support for different tasks including smart wheelchairs (Gomi & Griffith, n.d.), prosthetic hand(Kiguchi et al., 2008), emotional support (Glende et al., 2016) and more. Fig-

ure 1.1 visualises the scope and intersection of our study among the different disciplines. Table 1.1, summarises studies on SARs using communication to help the elderly with daily activities.

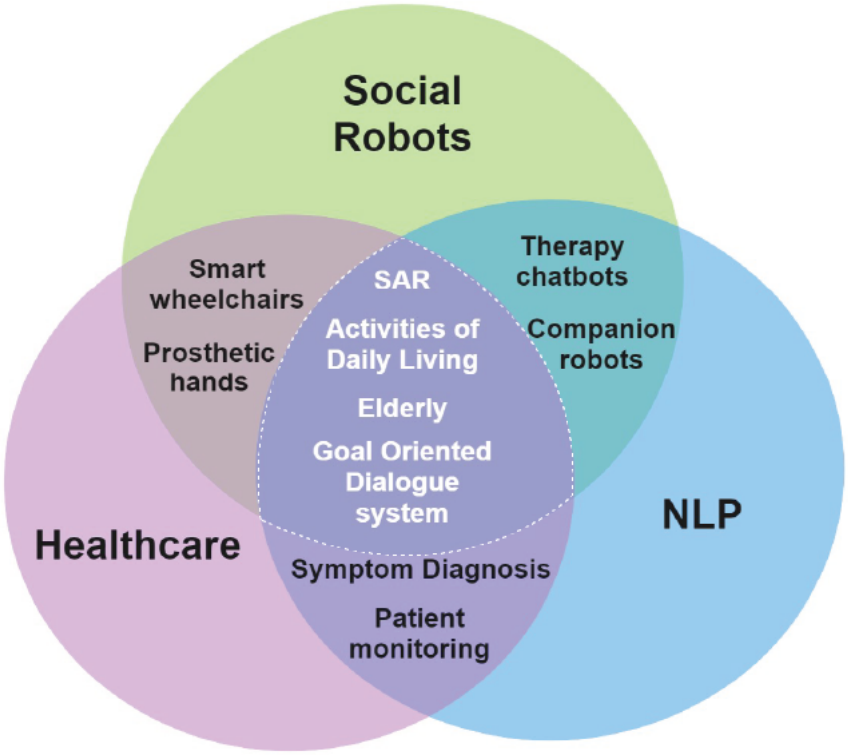


Figure 1.1: Scope of research for this study

Study	Aim of Study	Dataset
Granata et al. (2010)	An assistive robot developed for the elderly with mild cognitive impairments to help schedule appointments, send e-mails and medical diagnosis	A Vocabulary of words required is created by analysing conversations with the elderly
Goetze et al. (2012)	A mobile communication and assistance system using acoustic, visual and haptic input for a social robot ALIAS	Database consisting of speech data from users
Yasuda et al. (2013)	A virtual assistant resembling a child asks the elderly questions from their childhood	12 questions were prepared for the study
Mathew et al. (2021)	Develop a multi-modal intent classifier using large-scale dataset to train a robot for 'pick and place' tasks using natural language	ALFRED (Shridhar et al., 2019) is a new benchmark dataset for mapping natural language instructions to robotic actions
De Carolis et al. (2021)	A conversational agent to support the elderly in the context of daily tasks like scheduling reminders for medicines, weather information and symptom checker	Knowledge base created with 10 main intents and small-talk

Table 1.1: List of SAR using communication to assist elderly with daily activities

From the table, we see that researchers have handcrafted a database, vocabulary or data-sets relevant to care environments to design a communication system. We identify a gap in the availability of conversational data consisting of dialogues in the context of assisting the elderly with daily activities. SARs can support various activities and each study has focused on different activities like scheduling reminders, sending e-mails or for holding conversations on different topics. As mentioned before, the older people are open to robot assistance for certain activities like chores and manipulating objects whereas prefer help from a human for personal care. The first step in the development of a dialogue system is to analyse the tasks that need to be supported, and the type of dialogues and conversations that will be used (McTear, 2004).

1.3 Goals of the assignment

Heemskerk Innovative Technology (HiT) is working on several developments around Robot Rose including robot navigation, remote operation, human detection and communication. Figure 1.2 shows Robot Rose at a care environment.

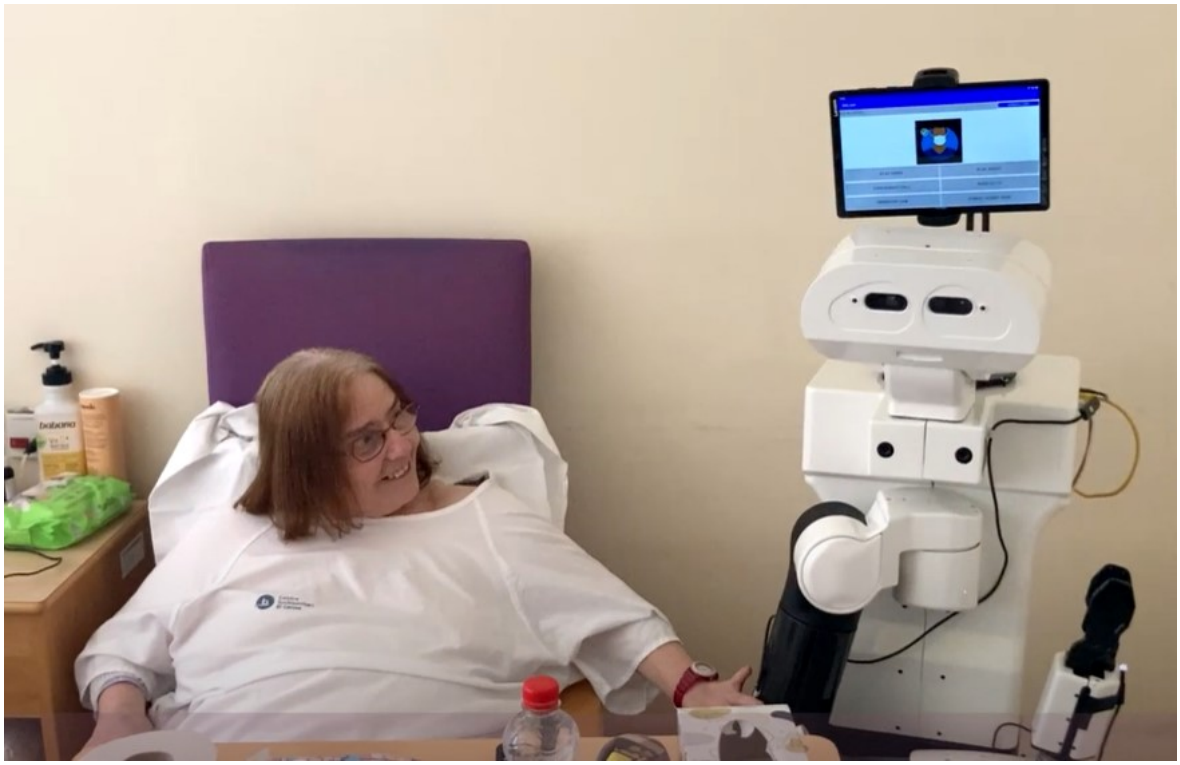


Figure 1.2: Robot Rose during trials in a care environment with the elderly

In this study, we conduct an exploratory study to find a feasible solution to design and evaluate a dialogue system for a care robot to communicate with the elderly and assist with daily activities. The main research question we answer in this study is,

1. *'How can we design and evaluate a dialogue system for a care robot to communicate and assist the elderly with daily activities in a care environment?'* This main research question leads to the following sub-questions,

   (a) What are the daily activities in care environments which the elderly require assistance with?

   (b) How can we design a dialogue system for a care robot to assist with daily activities without the availability of large amounts of training data?

   (c) How can we evaluate the performance of a dialogue system for the care robot and find the limitations for future developments?

1.4 Report organisation

Chapter 2 gives an overview of dialogue systems, the types, design, evaluation and the common challenges of a dialogue system. In Chapter 3, we review literature to explore the design approaches and conclude the chapter by discussing the proposed method for our study. Chapter 4 introduces the Rasa framework to familiarise with the different concepts and components. In Chapter 5, the implementation and design of the dialogue system is discussed. The analysis and evaluation of the dialogue system and conclusions are discussed in Chapter 6. We conclude the thesis report with Chapter 7 discussing the limitations of the study and recommendations for the future.

Chapter 2

Dialogue systems

A dialogue system (DS), or conversational agent (CA), is a computer program that communicates with users through text, speech, graphics, haptics, gestures, or other mediums (Jurafsky & Martin, 2020). Dialogue systems are complex Natural Language Processing (NLP) applications. NLP is a branch of artificial intelligence that enables machines to read, understand and derive meaning, context, emotions and more from natural languages. Figure 2.1 shows an overview of the working of a Dialogue system. The goals of a dialogue system are to understand the human language, identify the intent of the user and appropriately respond or communicate through natural language.

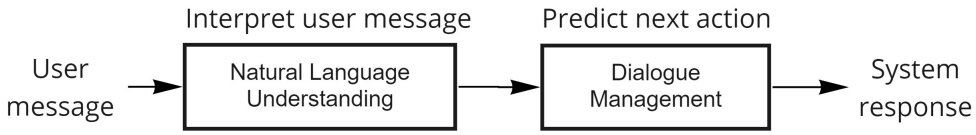


Figure 2.1: Basic functions of a Dialogue System

The remaining of this chapter and Chapter 3, we discuss the literature required for our research. Figure 2.2 the different literature review conducted in the two chapters.

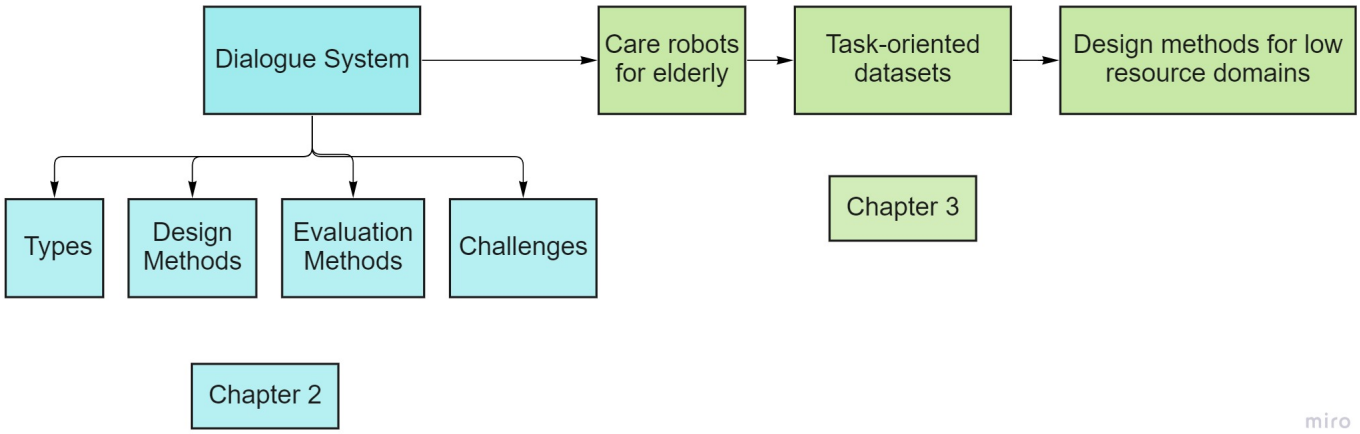


Figure 2.2: Literature review steps

2.1 Types of Dialogue Systems

Based on the application, dialogue systems can be classified into two main categories, task-oriented/goal-oriented DS and non-task oriented/ Open domain DS (Jurafsky & Martin, 2020).

2.1.1 Task-oriented

Task oriented systems help users accomplish various tasks. Dialogue agents like Siri, Alexa, Google Now/Home, Cortana, etc., help users achieve specific tasks like getting directions, controlling appliances, finding restaurants, setting reminders or making calls. Task-oriented dialogue systems are organised in a pipeline architecture and the main components include Speech Recognition, Natural Language Understanding (NLU), Dialogue State Tracking (DST), Policy Learning, and Natural Language Generation (NLG) (Chen et al., 2017). We will discuss the components in detail in Section 2.2.

2.1.2 Non-task oriented

Non-task oriented systems interact with users and are intended to keep users engaged and entertained (Ni et al., 2021). They are mostly open domain as users can interact about topics from books, movies or politics as they are trained using large datasets that contain knowledge from various sources like Reddit, Quora and more. There are two main approaches used to develop non-task-oriented systems. A gen-

erative system which generates responses by learning and retrieval-based system which selects a response from an existing database (Chen et al., 2017)(Ni et al., 2021).

2.2 Designing a Task-Oriented Dialogue system

Human robot interactions are mostly situated, which means that they occur dynamically and cannot fully be predicted or programmed but depend on the situation and the environment (Jokinen, 2018). As a result, a dialogue system should be capable of understanding the user's language, identifying the intent of the message, selecting or generating an appropriate response based on the intent and the history of the conversation. These functions are achieved by DS using two components- Natural Language Understanding (NLU) and Dialogue management (DM). We provide an overview of the two components using the studies by J. Liu et al. (2019) and Ni et al. (2021).

2.2.1 Natural Language Understanding

Natural Language Understanding (NLU) is a sub-set of NLP that allows machines to understand natural language by converting the user message into logical representations that the machine can understand. The NLU module performs two main tasks, intent classification and entity extraction.

We first explain a few terms which will be frequently used in this study.

1. **Utterance**
   The user input/message in the form of text or speech is known as utterance.

2. **Intents**
   Intent refers to the request that the user is trying to convey. Intents are also referred to as Dialogue acts (J. Liu et al., 2019)

3. **Entities**
   Entities are important text/words in the user's message that provide more details about the user's intent

Utterance	Intent	Entity
I want a glass of water	intent_drink	Water
I am thirsty		
Could you get me some coffee		Coffee
I'd like some juice		Juice

Table 2.1: Attributes of NLU

Table 2.1 shows examples of different utterances/ user messages which could be used to convey the user's need for a drink, which we define as intent_drink and the entities that are essential information conveying what the user wants exactly.

NLU runs a set of machine learning algorithms to train a model to classify intents and extract entities. Below we discuss these processes and explain some of the concepts used.

Intent Classification

Intent classification or intent detection, is a machine-learning approach to classify written/spoken input into various intents (Celikyilmaz et al., 2011). We start by training a model with examples of input text/utterances for all labelled intents. Figure 2.3 shows the steps involved in intent classification.

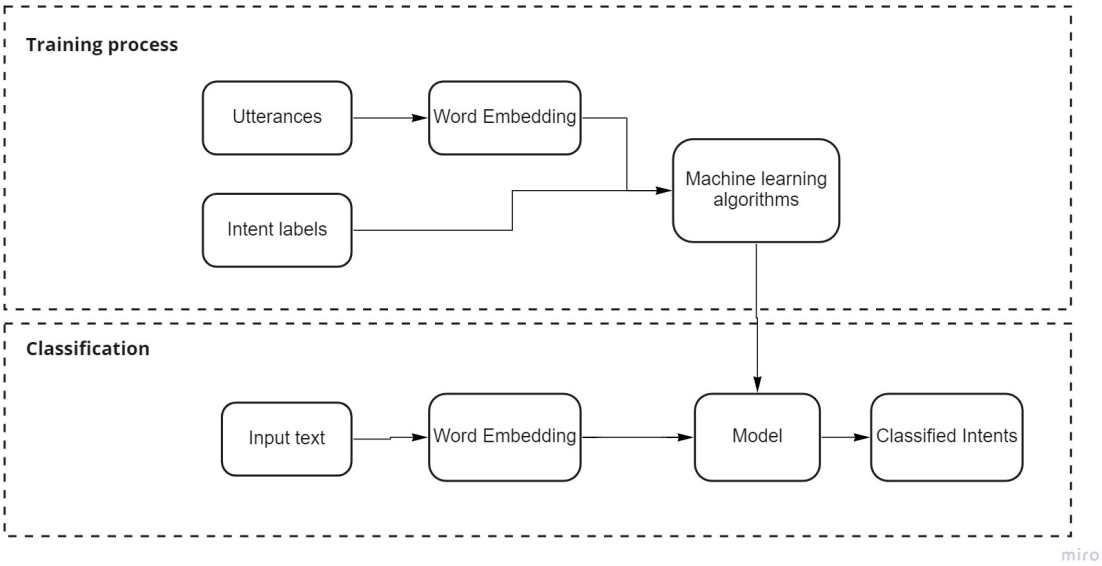


Figure 2.3: Intent classification process

Classification algorithms like Naive Bayes (McCallum et al., 1998), Support Vector Machine (SVM) (Haffner et al., n.d.) and Logistic Regression (Genkin et al., 2007)

have been studied for intent classification with a limited amount of training data and achieved high accuracy. Current state-of-the-art methods for intent classification use deep learning models like Convolution Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) Network and Gated Recurrent Unit (GRU) to improve the performance of intent classification but require more training data (Deng et al., 2012), (Tur et al., 2012).

These algorithms take numeric vector representations as input. Therefore, before classification, the utterance/ input text needs to be converted to vector representations called word embeddings. Word embeddings capture the context of a word in a document, and understand how words are related to each other. We will discuss two techniques, bag-of-words (BOW) and pre-trained embeddings.

- **Bag-of-Words**
  The bag-of-words is a commonly used input format for text classification. The frequency of occurrence of words used as a feature for training a classifier, without taking into account the grammar or order of words.
- **Pre-trained embeddings**
  Pre-trained embeddings are trained on large publicly available datasets like wikipedia text, or the Google News Dataset, Reddit with a vast corpus of language. Learning word embeddings from scratch could be challenging due to scarcity of data and a large number of parameters leading to slower training (Qi et al., 2018).

Entity Recognition

Entity Recognition is the process of identifying and extracting entities such as person, location, products etc from the user message. Entity recognition can be done by providing a dictionary of entities while training for the model to identify, or using pre-trained language models that extract the features and can identify entities from unstructured text data.

2.2.2 Dialogue Management

The dialogue management system (DM) maps the input data/utterance to a relevant response. Dialogue management can be classified into three categories- finite-state based, frame based and agent based (Laranjo et al., 2018; Ni et al., 2021).

- Finite-state uses a predetermined dialogue flow making it easy for implementation and works efficiently for well structured, straight-forward natural conversations. But finite-state DM limits user utterances and is not very flexible.

- Frame-based DM collects information from users through a template. The dialogue flow is not predetermined and is flexible depending on the user's input.
- Agent based DM can hold complex and efficient conversations. These systems are data driven and requre large amount of training data.

Dialogue management performs two tasks to decide the next sequence of action, dialogue state tracking and policy learning. We explain the functions of these tasks below.

Dialogue State Tracking (DST)

DST tracks the state of the dialogue at every turn of the conversation. Dialogue state contains a history of the user dialogue and corresponding actions from the previous state. The past utterances and replies (the dialogue history), and any intent and entities detected are represented through vectors Chen et al. (2017).

Policy Learning

The Policy learning module predicts the next action of the dialogue system. This module learns a mapping function between each dialogue state and the corresponding action and uses this mapping to predict next actions in the upcoming conversation turns (Ni et al., 2021).

- **Rule-based**
  Rule-based policies consist of handcrafted responses and rules. They struggle with complex dialogues as it is time consuming to create and maintain rules for each conversation pattern. However, rule-based policies are used to support other policies as they are computationally efficient in processing frequent queries or answering out-of-scope user requests.
- **Retrieval-based**
  Retrieval-based policies decide the next action by finding a match for the intent from the training data. Retrieval-based policies perform well for coherent and well formed responses, but require a higher quality of annotated data (Chen et al., 2017).
- **Generative Policies**
  Generative policies generate natural sentences by learning from a broad range of topics. Generative models can utter nonsensical and inconsistent sentences which is not preferred for task-oriented systems.

Most dialogue systems are designed using a rule-based approach initially and using incremental learning, user simulations are created to learn the dialogue policies or next actions for complex conversations.

.

2.3 Evaluation of a Dialogue system

Evaluating a dialogue system is a challenging task with a lot of interest among researchers. An ideal evaluation method is *automatic* with less dependency on human efforts, *reproduces* the same results across iterations under the same circumstances, yields results which *correlate* to human judgments, *differentiate* different dialogues and should be capable of *explaining* the features that impact the quality of dialogues (Deriu et al., 2021). Although many studies have made progress in automating the evaluation, there is still high dependency on human evaluation which is a cost and time consuming task (Deriu et al., 2021).

Dialogue systems are evaluated by testing with users. Users interact with the system to achieve a specific task. The interaction is rated for success either by the user or by the subject expert or the researcher. The ratings can be translated to a quantifiable metrics like the task success rate or the dialogue efficiency. PARADISE Framework (Walker et al., 1997) is a popular evaluation framework for task-oriented systems. User simulations are designed to replicate a functional system and are used find the weakness of dialogue systems or to train a system in an offline environment. The choice of evaluation depends on the purpose of the study (Deriu et al., 2021). Below, we discuss the quantifiable metrics that have been used to measure the performance of task-oriented dialogue systems:

1. **Task success rate**
   The task success rate measures the number of tasks successfully completed by the user. Task success rate is a common approach as it is a simple and effective way to measure the performance of a dialogue system.

2. **Dialogue efficiency**
   Dialogue efficiency is a measure of the length or cost of the dialogue (Walker et al., 1997). These can be measured using various properties like the number of times the system asks user to rephrase or the number of incorrect rephrases and more.

3. **Classification metrics**

- **Precision**
  Precision measures how many predictions of an intent are correct as shown in figure 2.4. As an example, the model makes predictions of an *intent:intent_1* this includes correct predictions(true positives) as well as intents that have been misclassified as *intent:intent_1* (false positives). So precision is defined as,

$$Precision = \frac{Truepositives}{Truepositives + Falsepositives}$$

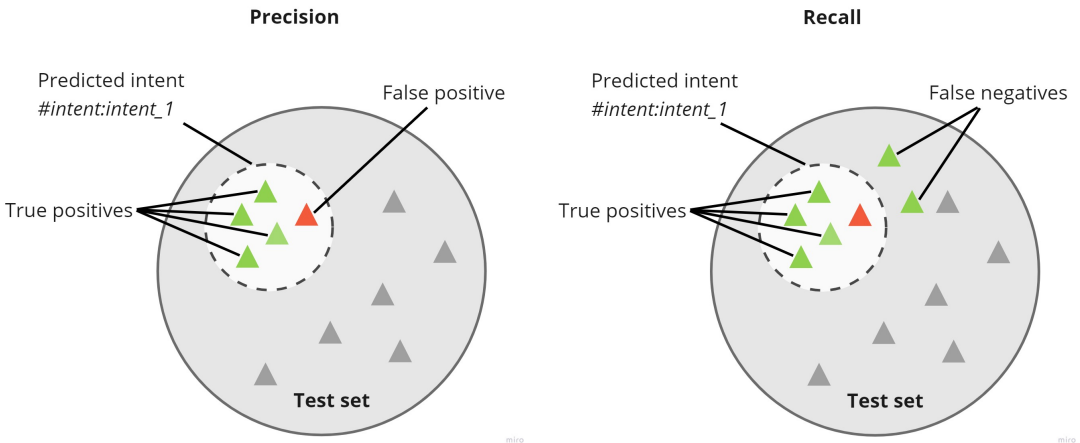


Figure 2.4: Precision and recall values for classification

- **Recall**
  Recall measures how many intents are correctly predicted, i.e, in the test set, how many intents that are actually *intent:intent_1* were classified as *intent:intent_1* as shown in figure **??**. We find the total number of *intent:intent_1*, that is true positives which model predicted correctly and false negatives, that are incorrectly predicted as other intents. Recall is defined as,

$$Recall = \frac{Truepositives}{Truepositives + Falsenegatives}$$

- **Accuracy**
  Accuracy is measure of all the correctly identified intents. Accuracy is a good metric if the classes/ intents are balanced. Accuracy is defined as,

$$Accuracy = \frac{Truepositives + TrueNegatives}{Truepositives + FalsePositives + TrueNegatives + Falsenegatives}$$

- **F1 Score**
  F1 score takes into account precision and recall and is defined as,

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

- **Confusion matrix**
  The confusion matrix is a grid that shows the actual intents along with the predicted intents and the diagonal elements refer to correct predictions of intents as shown in figure 2.5. This matrix helps to identify intents which are frequently misclassified/ confused with other intents and suggest any changes or additional parameters required by the training data.

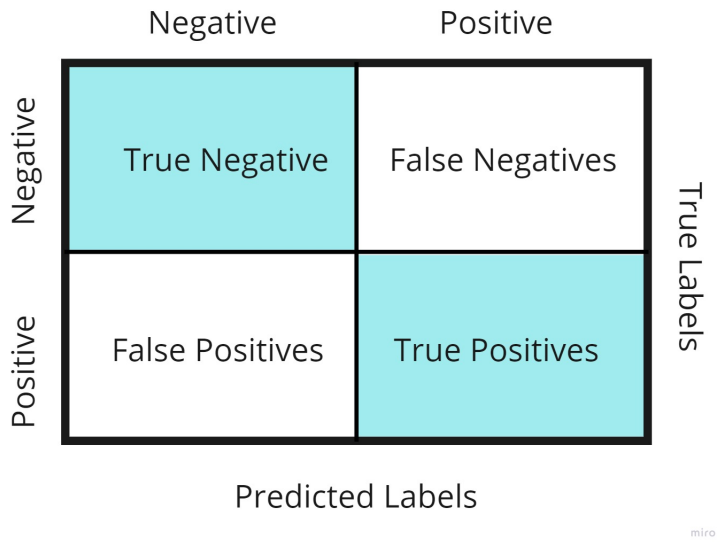


Figure 2.5: Confusion matrix

2.4 Challenges of Dialogue systems

1. Dataset is a crucial factor for the development of data-driven dialogue systems. For task-oriented systems, the dataset should consist of utterances for different intents and intent-response pairs to design a rule-based system as well as for training machine learning models efficiently. At present, the available datasets for task-oriented dialogue systems are scarce and limited to a few domains (Ye & Li, 2020),(Jiang Zhao et al., 2019). Collecting and building a good quality dataset for domains other than flight and hotel reservations, and virtual assistants, is a time consuming and expensive process.

2. Intents can be expressed in multiple ways. Intents can be split into explicit and implicit intents based on the types of expression. Using explicit utterances, a

user expresses his or her intent needs explicitly in the utterance. Using implicit utterances, the user utterance does not directly convey the user's intent, but it is essential to infer the user's true intent for correct intent classification (Chen et al., 2013) which is a difficult task for a DS to learn. As an example, in the context of our use case, an explicit user utterance is "I want tea" and an implicit user utterance could be "Tea soothes me" and the user's need may be asking for a cup of tea. It is essential to consider these possibilities while designing a dialogue system so that the system can comprehend the true intent of the user and appropriately hold the conversation.

3. Evaluation of dialogue systems remains a challenging problem for researchers. Human evaluations are effective for rating the fluency and usability. However in the initial stage of development, it is essential to validate the functionality of the model and prototype before it can be tested with users. Studies show an inconsistency in the standards for reporting which make it difficult to compare different systems (Laranjo et al., 2018).

In this Chapter, we studied different concepts of Dialogue systems- types, design and evaluation of a task-oriented DS and the ongoing challenges in the research field of Dialogue systems. In the next chapter, we discuss the design approaches and datasets available to find a feasible method to design a dialogue system for a care robot assisting the elderly with daily activities in a care environment.

Chapter 3

Related work

In Chapter 1, we identified the gap in the availability of large amounts of training data to design a dialogue system for a care robot assisting the elderly with daily activities. In this chapter we review datasets used to design a task-oriented DS using the surveys conducted by (Allouch et al., 2021) and (Ni et al., 2021).

Dataset	**Source collection**	**Domain**
MetaLWOz (Shalyminov et al., 2020)	Crowdsourcing	Bus schedule, apartment search, alarm setting etc.
COVID-19 dialogue (Yang et al., 2020)	Conversation between doctor and patients	Medical dialogue system for checking symptoms
EmpatheticDialogues (Rashkin et al., 2018)	Crowdsourcing	Recognising human feelings
SNIPS-NLU (Coucke et al., 2018)	Crowdsourcing	Train voice assistants

Table 3.1: Commonly used Task-oriented datasets, adapted from Ni et al. (2021)

The main reason to analyse these datasets was to find if the datasets had any similarity with dialogues or conversations in the context of a care robot assisting the elderly with daily activities which we can use for our study. However, the most of the common datasets were for the flight, hotel or movie domain to make reservations.

3.1 Design methods for Low resource domains

We review the literature to study the different methods to design Dialogue systems or conversational agents for domains and languages with limited resources.

3.1.1 Rule-based

Rule-based system model human dialogues using pre-defined rules and retrieves responses using the rules. In this method, conversations and dialogue flows are analysed and translated into a set of dialogues with responses based on different patterns. ELIZA (Weizenbaum, 1966) is one of the first rule based systems designed to simulate a conversation by responding to patients by reflecting their statements back to them. Rule-based systems are most popular method due to high quality, controllable responses, and is effective during the early stages of development (Quan et al., 2021). As conversations become complex, it is difficult to define dialogue states and responses by rules. State-of-the-art methods using machine learning algorithms handle complex conversations efficiently. Although they require high amount of data for training to achieve a good performance. While designing dialogue systems, a common practice is to use rule-based systems for a warm start, and enhance the models using data-driven machine learning methods in the later stages of development to handle complex conversations (Chen et al., 2017).

3.1.2 Transfer Learning

Transfer learning is a technique used to apply knowledge from one domain into another (Ruder et al., 2019). Transfer learning makes use of the similarity between a source and a target domain to transfer knowledge from high resource domains to domains with no datasets. As shown in Figure 3.1, restaurants, movie and travel reservations have common tasks like booking and entities like time and location. Using transfer learning, knowledge can be transferred from one domain to the other instead of learning from scratch. This is useful for domains which have a similarity with other domains or in case of low-resource languages.

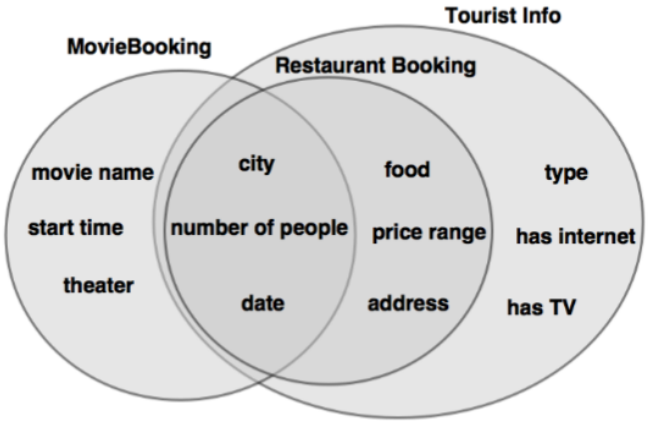


Figure 3.1: Entity/slot similarity between domains, from Ilievski et al. (2018)

3.1.3 Incremental Learning

Dai et al. (2020) proposed a method that selects a response through retrieval and engages a human for feedback if the system cannot understand the user's request. The results showed good adaptability and performance on the MultiWOZ dataset, a multi-domain dataset consisting of 10,000 dialogues from seven distinct domains including taxi, restaurant and hotel booking and has been used as a standard benchmark for the mentioned domains (Eric et al., 2019).

3.1.4 NLU services

Since chatbots efficiently engage users, automate tedious tasks and save time and effort, more and more companies are investing in the development of chatbots of conversational agents. Many NLU platforms are currently available that provide a handful of NLU tasks (Abdellatif et al., 2020). Commonly used NLU services include Watson, LUIS and Dialogflow. The services are capable of handling the functions of NLU and DM. Braun et al. (2017) and X. Liu et al. (2019) have conducted studies to compare the different services and the results show that some NLU services work better than the others for certain tasks or domains. A study of different dialog management approaches conducted by Harms et al. (2019) showed that Rasa Core was among the best-performing frameworks in terms of the overall design and features offered. Among the services, Rasa is the only open source NLU service that performed on par with the commercial services consistently. Rasa has an edge over the others as it supports customised configurations and interactive learning (Nguyen et al., 2021). Through interactive learning, the system can learn and correct itself though conversations with users or with a human/developer in the loop, unlike other platforms where the model is trained only by labelled data. Nguyen et al. (2021) and Windiatmoko et al. (2020) developed a chatbot to support university admissions and queries in Vietnamese and Indonesian, which are low resource languages using Rasa and achieved competitive results. Malamas & Symeonidis (2021) used the Rasa framework to design a minimum viable assistant with a small dataset.

3.2 Proposed Method

Reviewing the datasets and design methods discussed in the previous section, transfer learning is not a suitable choice in our context as there are no overlapping domains from which knowledge can be transferred to a robot assisting the elderly. Rule-based systems are a popular choice for exploration during early stages of development and data-driven machine learning algorithms like incremental learning to handle complex conversations can be used for further developments to handle complex conversations. In our study, we propose a method to leverage the strengths of both and using the framework Rasa as the developmental platform for the following reasons,

1. Without the availability of a large dataset, dialogues used for assisting the elderly with daily activities can be handcrafted as a set of dialogues and responses that can be retrieved to build an early functional prototype

2. Along with a rule-based method, Rasa uses machine learning algorithms to learn from user conversations, and make predictions with a small amount of dataset of 5-10 examples for each dialogue

3. Developing a DS through Rasa does not require advanced programming skills in NLP as it has inbuilt configurations to support NLP tasks

4. With a functional prototype, incremental learning methods can be used train the system to handle complex conversations by training the system with users along with a human/researcher in the loop to provide feedback during training

   Rasa framework supports the development of rule-based systems and a platform for interactive learning. Our proposed method suggests two important phases of development, an early prototype can be designed using handcrafted dialogues and retrieve responses without the need for a large dataset and incremental learning can be used in the second phase to develop the system to handle more conversations without the need for advanced programming skills in NLP.

   Within the time and scope of our study, we focus on the first stage of development to design an early prototype using the rule-based method on the Rasa framework and evaluating the system with users.

Chapter 4

RASA

In this chapter we introduce the concepts of the NLU framework Rasa. We discuss the data formats and files required in Rasa in Section 4.1, and the pipeline configurations for NLU and DM in Sections 4.2 and 4.3.

4.1 Training data

To train a dialogue system using the framework Rasa, we need data to train the Rasa NLU for intent classification and entity extraction and Rasa Core for dialogue management. In this section, we will define the type of data each component requires.

4.1.1 NLU Module Data

The training data for the NLU consists of utterances labelled by intents with entities and are stored in *nlu.yml* file. Entities are defined as *'[entity value] (entity type)'* as shown below,

```
- intent: drink_direct
  examples: |
    - Can I get a glass of [water](drink)
    - Could I get some [tea](drink
    - I would like to drink some [water](drink)
    - Can I get some [coffee](drink)
    - Can I have some [juice](drink)
    - I want [water](drink)
    - Can you bring me [coffee](drink)
    - Could you bring me some [coffee](drink)
    - May I have some [juice](drink)
    - May I get tea [tea](drink)
```

The text above shows the definition for a single user intent, *drink*. Similarly, all intents/ tasks which the DS should support need to be defined in the *nlu.yml* file. Although 5-10 utterances are sufficient for each intent, it is a good practice to include all different ways by which an intent can be conveyed, so that the system can be trained to handle more user requests efficiently.

4.1.2 Rasa Core Data

Rasa core retrieves responses to respond to a user intent. The responses to the respective intents are defined in the form of *stories* and stored in the *stories.yml* file. Below the text on the left, shows an example of a conversation between a user asking the DS for a drink. This conversation is written as intent-response pairs in stories, as shown by the text on the right to train a DS to handle the conversation.

```
User: 'I am thirsty'
DS:   'Would you like to drink something?'
User: 'Yes!'
      'Bring me a glass of water please?'
DS:   'Let me get you a glass of water'
      'Can i help you with anything else?'
```

```
- story: drink indirect path 2
  steps:
  - intent: drink_indirect
  - action: utter_drink_indirect
  - intent: affirm
  - intent: drink_direct
    entities:
    - drink: water
  - slot_was_set:
    - drink: water
  - action: utter_drink_direct
  - action: utter_help
```

In the stories defined, we see *intents* which are explained in the previous section and *actions* are responses defined. NLU classifies the intents based on the data from *nlu.yml* file, and appropriate response/action is selected by the Core using the data in *stories.yml*. It is essential to write stories for all possible conversation flows so that the system can learn to handle more conversations and reduce errors.

4.2 Rasa NLU

The RASA NLU converts unstructured user messages into intents and entities. Once the dialogue system understands the user message, it can classify and extract intents and entities. As discussed in Section 2.2, intent is the goal or task that the user wants to accomplish and an entity is the keyword which provides more information. As an example, the user request 'I want coffee' will be classified as the intent 'drink'. The DS now understands the user wants a drink, and by extracting the entity, it knows that the user wants 'coffee'.

Rasa NLU consists of a set of machine learning algorithms that classify intents and extract entities. Figure 4.1 shows the architecture and functions of Rasa NLU. The algorithms are executed sequentially in a pipeline configuration starting with a text input, parsed by various components until intents and entities are extracted as output. The main components are:

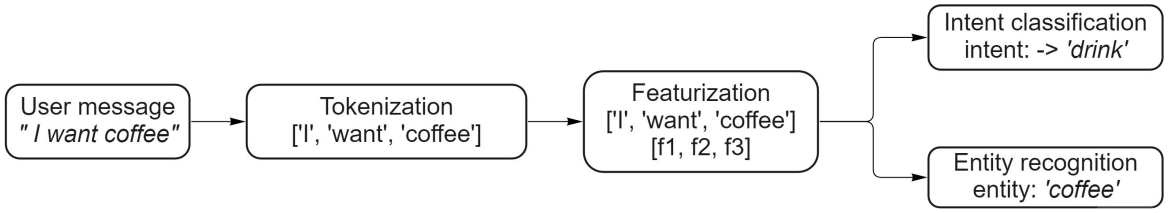


Figure 4.1: Architecture of Rasa NLU

- **Tokenization**
  A tokenizer splits the user message, i.e a sentence into words or tokens. Different tokenizers can be used for different types of uses. A Whitespace Tokenizer breaks down user messages to tokens that are separated by a space which is common in the English languages and the SpacyTokenizer uses the spaCy library. Other tokenizers are available for languages with specific requirements.

- **Featurization**
  A featurizer extracts features and creates a numeric representation of the text/tokens for machine learning models. There are two types of features, sparse and dense. The sparse features consist of bag-of-word representations of the user messages and response, whereas the dense features consist of pre-trained embeddings. Features are not only created for tokens, but also for the complete sentence. The RegexFeaturizer uses regular expressions to extract features, CountVectorsFeaturizer creates a bag-of-words representation, LexicalSyntaticFeaturizer extracts the lexical and syntatic features for entity extraction and SpacyFeaturizer uses the spaCy library.

- **Intent Classification**
  Once the features for tokens and sentences are created, intent classifiers classify the incoming user messages into one of the intents defined in the training data. Rasa support the DIET classifier (Bunk et al., 2020), a state-of-the-art intent classification architecture for natural language. DIET classifier is a viable choice for application in assistive technologies as they can be trained quickly with small a dataset (Mathew et al., 2021).

- **Entity Extraction**
  The entity extractor extracts the entities from the user message using the fea-

tures extracted. The entities can be of various types like name, location, dates, numbers. EntitySynonymMapper maps entities to synonyms defined in the library for extraction and the ResponseSelector makes selects makes a prediction from the candidate responses using the features extracted by the NLU component and passes the information to Rasa Core.

4.3 Rasa Core

Rasa core handles the dialogue management and decides the next action/response using policies. Policies are rule-based or machine learning algorithms that decide on the appropriate next action or response. Policies are defined in the configuration file. Unlike the NLU pipeline which is sequential, all policies n the pipeline run in parallel. For each turn, all policies defined in the pipeline make a prediction with a confidence score as shown in Figure . The policy with the highest score decides the next action. Below, we explain some of the policies to understand how the next action is predicted.

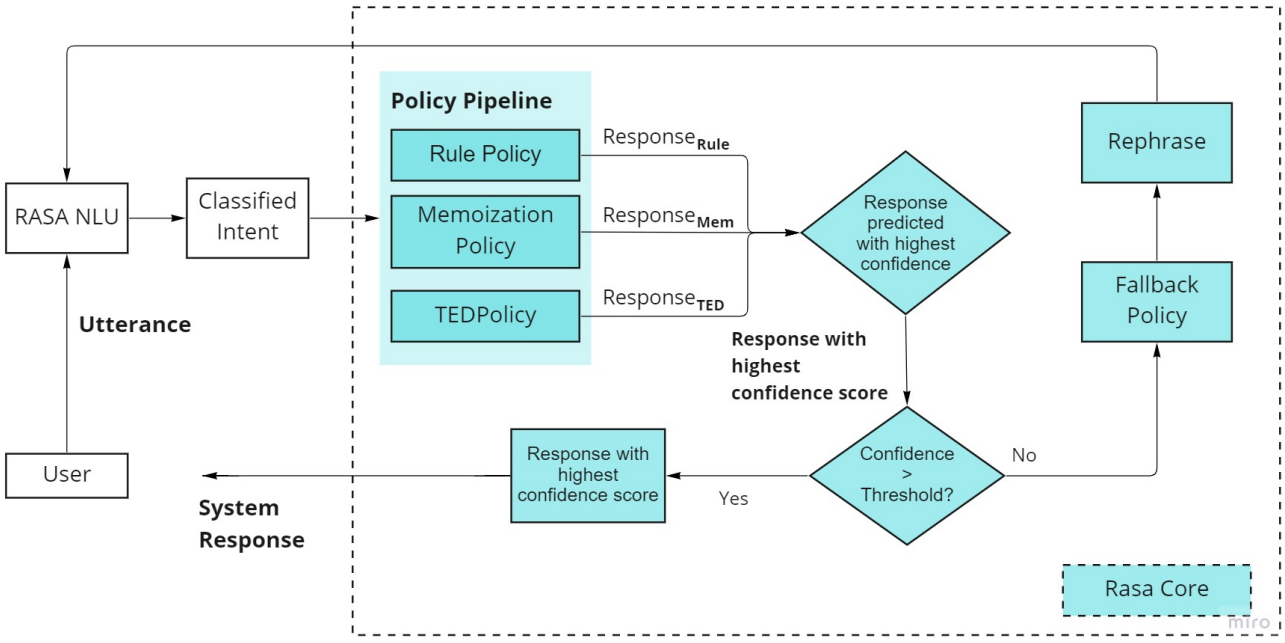


Figure 4.2: Architecture of Rasa Core

- **Rule Policy**
  The RulePolicy follows a fixed behaviour as defined by the rules which are defined as part of the training data. As an example, when a user greets, the default action would be to respond and offer greetings. This can be configured as a rule and the system will always respond with greetings to any user greets.

- **MemoizationPolicy**
  The MemoizationPolicy looks for an exact match of conversation pattern in the stories and predicts the next action as found in the training data. If a match is found, Memoizationpolicy predicts the next action with 100% certainty, i.e with a confidence score of 1 or else returns none with a confidence score of 0.
  The MemoizationPolicy ensures the system functions well for all conversation patterns that are present in the training data without errors. Since it does not make a prediction if a match is not found in the training data, it needs to be used along with other policies.

- **TEDPolicy**
  The Transformer Embedding Dialogue (TED) Policy makes a prediction by learning from training data. It makes a prediction based on the user's message, the previously predicted system action, and any values saved as entities. At each conversation turn, features from the user message, intents and entities and the predicted next action are represented as a vector. TEDpolicy decides the next action by comparing the similarities of the current vector to the previous states of the vector. Unlike the MemoizationPolicy, TEDPolicy can make a prediction of the next action even if the conversation pattern is different from the training data.

- **FallbackPolicy**
  The FallbackPolicy allows the system to handle errors gracefully, by selecting a default rephrase such as asking the user to response to avoid incorrect responses. A fallback is used when the system does not recognise an intent or makes a prediction with a low confidence score.

4.4 Rasa X

Rasa X is a GUI that can be used to interact with users and for incremental learning. User interactions through Rasa X can be captured and used as data for training. Rasa X allows the researcher/developer to give feedback and correct misclassifications or incorrect response selections, thereby acting as an interface for conducting interactive learning with a human in the loop. For the scope of our study, we used Rasa X as the interface for user interactions to test an early prototype. We did not provide feedbacks or correct the system during the study since our aim was to evaluate the functions of the system. For later stages of development, it can be used as a platform for incremental learning with a human in the loop to train and label data, give feedback and correct misclassifications to train the system instead of manually annotating conversation data.

4.5 Evaluation

The dialogue system is evaluated using the classification metrics accuracy, precision, recall and F1-score. The RASA NLU module is evaluated on intent classification and entity recognition and Rasa Core is evaluated on response selection. The models are evaluated by running intents, entities and storied over the model. Rasa evaluates the dialogue system using a *test script*. The test script is written in a similar format as stories mentioned in Section 4.1.2 and acts as the reference to validate the system. It consists of the utterances, along with the expected intent, entity(if any) and response. The test script runs over the model and the predicted intents, entities and responses are compared with the expected intents, entities and responses defined in the test script. This generates a confusion matrix of the expected and predicted classes for intent classification, entity recognition and response selection. A test set can either be created manually by the developer using unseen data(e.g., conversation from users not used in during training), or a portion of training data can be set aside for testing purposes using the rasa data split nlu command and evaluated using cross-validation.

Chapter 5

Implementation

Figure 5.1 gives an overview of the steps taken to design and implement a dialogue system for a care robot assisting the elderly with daily activities. In the remaining of this chapter, we discuss the implementation of a DS using the Rasa framework.

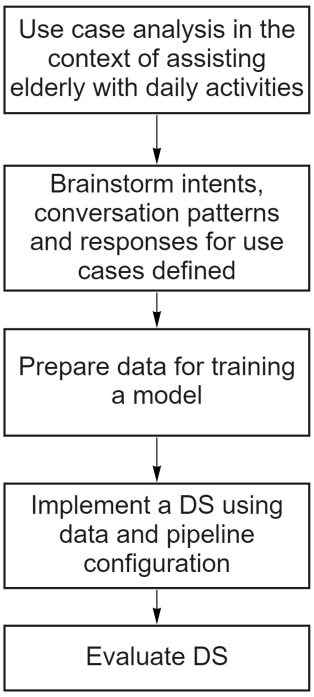


Figure 5.1: Summary of methods

5.1 Use Case Analysis

To understand the scope of a care robot and the tasks they can assist with at a care home, we analyse the activities of a staff and the care home environment. Studies usually collect audio and video recordings from the user's environment and observe

interactions in a care home. This approach also helps capture real conversational data which can be used for training a DS. However, due to privacy and COVID regulations this was not a feasible option for our scope. Instead, we conducted a virtual interview with a nurse who has been working with hospitals and care homes for over a decade and has been assisting HiT in various projects.

The interview was semi-structured consisting of open-ended questions to guide the interview. The questions focused on the care home environment, interactions, daily activities of the elderly and how care robots are perceived at care homes. After obtaining the consent, the goals of the thesis and the interview were briefed prior to the interview. The interview lasted for an hour and covered the pre-determined questions along with additional insights that followed. The goal and outcome of the interview was to gather possible use cases/tasks for the care robot. Below we summarise some of the key points from the interview.

Interactions

- On average one nurse is on duty for around 30 patients
- During the day, the nurse goes on multiple rounds and assists whichever patient is in need
- During the night, patients have a call button next to them which connects them to a nurse through phone and the nurse comes in to check

Activities

- Help serve food, drinks and medicines to patients
- Some patients have restricted movement and require help to pick up fallen objects
- Perform routine check-up of patients and maintain reports
- Some patients would require assistance with to moving around, going outside or using the restroom

Perception

- The elderly have shown acceptance towards robots
- Robots are perceived better when they are introduced early and patients are given time to adapt

- The elderly prefer slow speech and a non robotic voice as it is easier to understand and friendly
- Using gestures and other modalities like photos makes the interaction easy for the elderly when they can't follow the speech or have hearing impairments

Using the inputs from the interview, we narrowed down three different tasks that patients mostly needed assistance with. We will design a DS for robot Rose to handle the following user requests for the early prototype,

- **Drink** - To assist when the user requests for a drink
- **Pickup object** - Help when the user requests to pick up an object lying around them
- **Call doctor** Contact a doctor/nurse when the user is in pain/ needs help

5.2 Collecting Training data

From the literature, since there were no available datasets that could be used to train our dialogue system, we manually create training data in English as per the requirements of the assignment. As discussed in the previous chapter, the training data should consist of utterances/user requests with labelled intents and entities for classification and different user stories/ conversation patterns to train the dialogue management. Utterances were collected using the following techniques:

1. Manually create utterances for each intent along with some of the commonly used entities
2. Combine utterances collected from users during previous experiments conducted at HiT for the intents that matched with our use cases
3. The scenarios were presented through images and utterances used in Appendix B to collect utterances from three participants using a survey. This allowed to reduce some amount of bias caused by manually generating the dataset.

User request	Intent	No. of sample utterances	Entities
Greeting	intent_greeting	14	None
Want a drink	intent_drink	30	Coffee, Water, Milk, Tea, Juice
Help pickup an object	pickup_object	17	Book, remote, plate, pillbox, Spectacles, yarn, notebook
Call for help	call_doctor	20	Doctor, nurse, medic
Out of scope	out_of _scope	8	None

Table 5.1: List of NLU Training data in nlu.yml

Table 5.1 lists the intents and entities we have used for training Rasa NLU. The number of sample utterances refer to example sentences which can be used to convey the intent. In Figure 5.2, each branch shows the different conversations paths which have been considered for designing the DS for robot Rose. The white boxes refer to the intents, the blue boxes refer to examples of user utterances and the green boxes refer to the response selected by the DS. Although we tried to balance the training data among all intents, some intents had more utterances or examples than others.

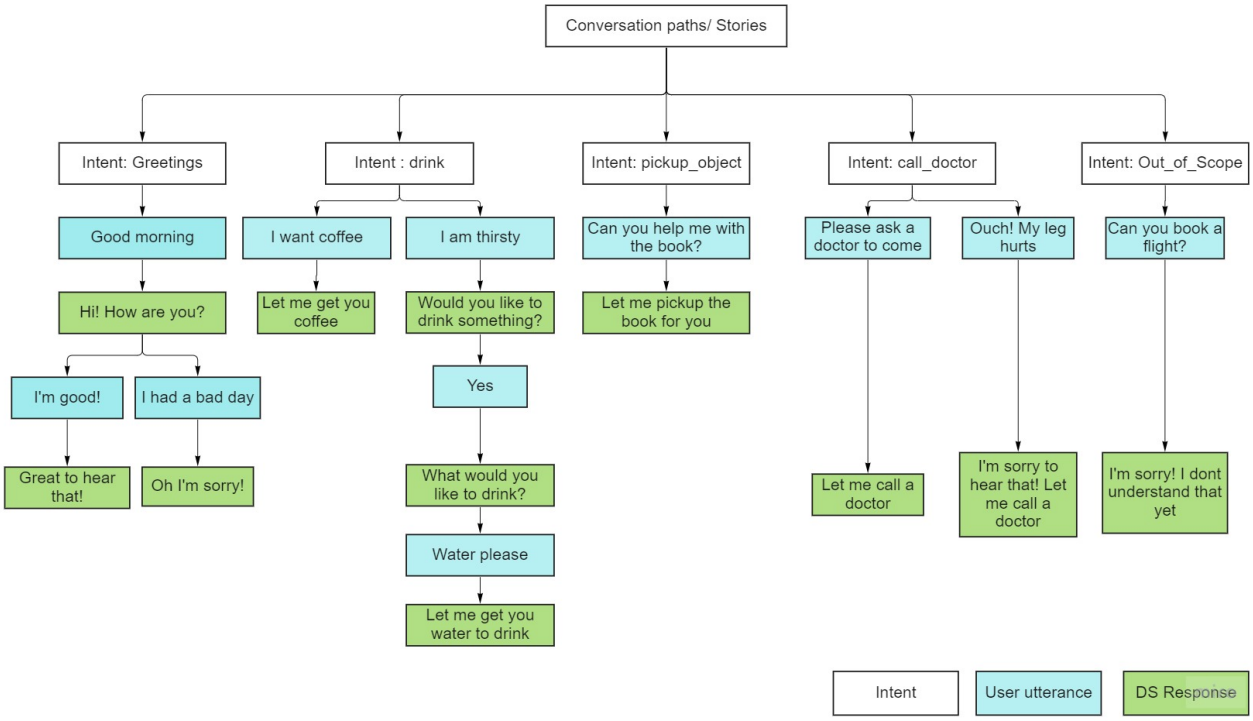


Figure 5.2: Visualizing the different conversation paths in stories.yml

5.3 Care Environment Word Similarities

As discussed in section 2.2.1, pre-trained word embeddings are trained to have some existing knowledge of natural language which helps the system handle unseen data or data which is not provided during training. We analysed a few words from our use cases to compare the word similarities within the pre-trained model and found the results as shown in Figures 5.3 and 5.4 .

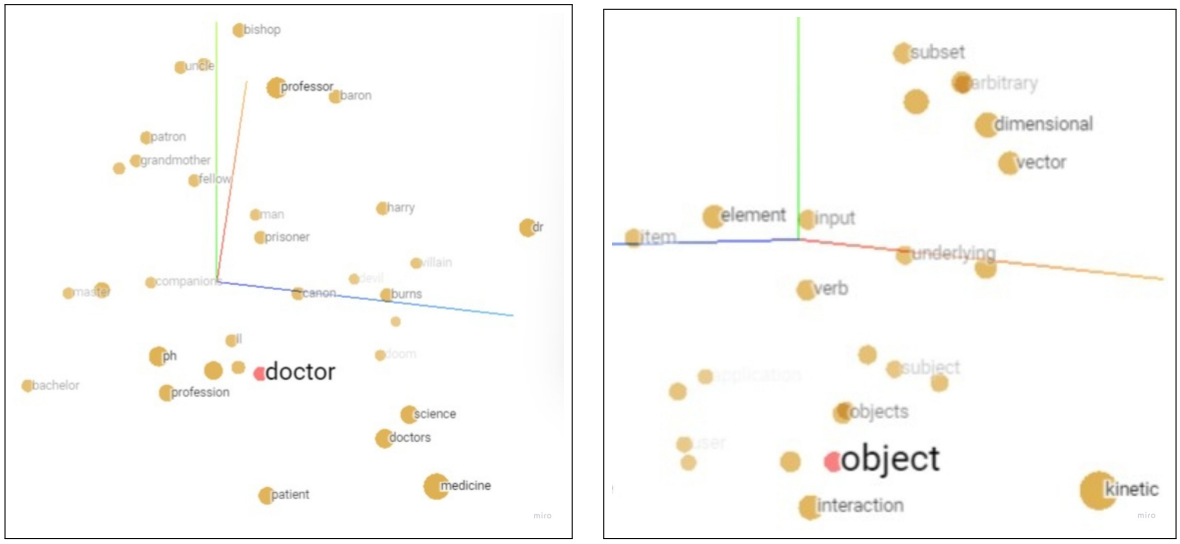


Figure 5.3: Entities from our use case with different word similarities

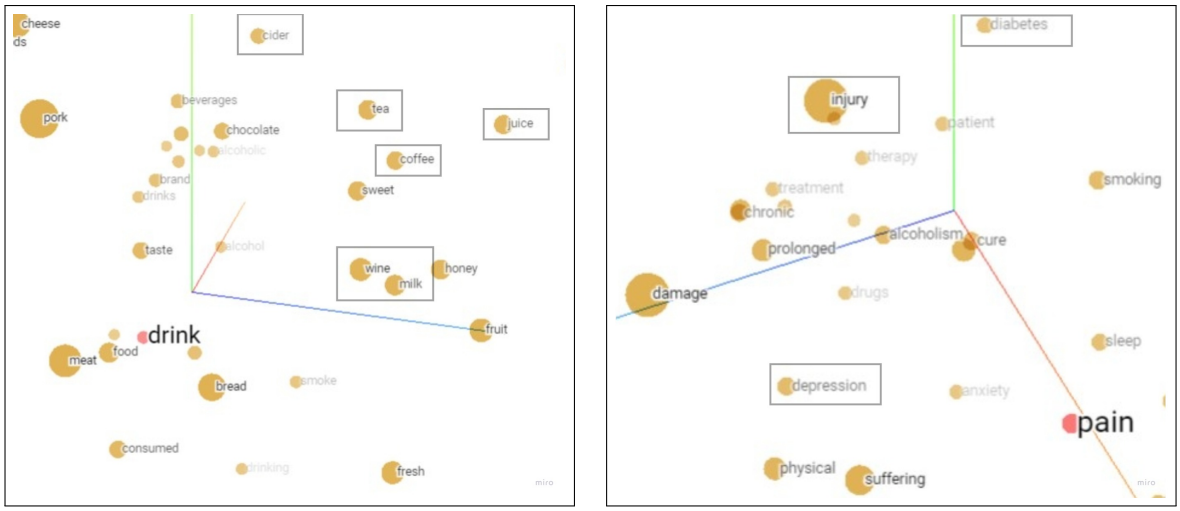


Figure 5.4: Entities from our use case with word similarities

5.4 Implementation using Rasa

For the design of a DS, Rasa recommends two default pipeline configurations based on the use case and training data. The first configuration defined as P1 in Table 5.2 is recommended for domains and use cases which contain domain specific terminologies or word similarities. This configuration builds a model from scratch using only the training data provided. For other domains, Rasa recommends using the configuration P2 which uses SpaCy pre-trained word embedding. Pre-trained word embedding have existing knowledge of natural language which help models handle unseen data or data which is not in the training set. As an example, pre-trained word embeddings are trained to have some knowledge to understand that 'coffee' and 'tea' are similar and is useful when there isn't enough training data.

NLU Pipeline	**Details**
P1	WhitespaceTokenizer + RegexFeaturizer +LexicalSyntacticFeaturizer + CountVectorFeaturizer + DIETClassifier + EntitySynonymMapper + FallbackClassifier
P2	SpacyTokenizer + SpacyFeaturizer + RegexFeaturizer + LexicalSyntacticFeaturizer + CountVectorsFeaturizer + DIETClassifier + EntitySynonymMapper + ResponseSelector + FallbackClassifier

Table 5.2: NLU Pipeline configurations

As discussed in the last section, the intents and entities considered for our design include some domain specific words. We create two models using the pipeline configurations P1 and P2 as recommended by Rasa to develop the early prototype of a DS for Robot Rose using the intents, entities and conversation path/stories defined earlier. In the next chapter, we test the early prototype with users and report our analysis and findings.

Chapter 6

Evaluation

We have designed a dialogue system for a care robot in the context of assisting the elderly with daily activities using the Rasa framework. In this chapter we discuss the experimental design and test the system with users. The experiment and evaluations were designed with the following goals:

1. Test the performance of the DS through user interactions
2. Evaluate and compare Rasa configuration to find best fit in the context of dialogues for assisting with daily activities in care environments
3. Identify limitations in design from user experiments for improving the system in future studies
4. Capture the user conversations to use as training data for further developments of the dialogue system

6.1 Experimental Design

To test the dialogue system with users,we curated six scenarios for the use cases defined in section 5.1. Table 6.1 shows the intents along with the corresponding tasks and their representation designed for the user experiments. All the scenarios are listed in the Appendix B. For the selection of participants, employees and peers from HiT were invited to participate. Eight participants volunteered to participate. For the experiment, each participant was presented with three different tasks and asked to interact with the system. Participants interacted with the system through the Rasa X interface via text. The interactions were short and lasted an average of 20 minutes. All the interactions were captured through Rasa X for analysis. After the participants completed the three tasks, the participants fill in a questionnaire to rate the interactions and provide feedback.

Task	Reference Intent	Represented by
t1	*intent:drink*	Text
t2	*intent:drink*	Image
t3	*intent:pickup_object*	Text
t4	*intent:pickup_object*	Image
t5	*intent:call_doctor*	Text
t6	*intent:call_doctor*	Image

Table 6.1: Tasks defined for User Experiment

6.2 Experimental Evaluation

In this phase of our study, our goal is to measure the performance of the system, analyse the experiment and data to explore limitations. Using the conversation data from the user experiments and the questionnaire, we conduct the following evaluations:

1. Functional evaluation of the DS using the classification metrics automated through Rasa
2. Measure task success rate from the interactions captured
3. Measure the tasks success through user filled questionnaire
4. Qualitative analysis to identify system failure through observations of interactions

6.2.1 Quantitative evaluation of the Dialogue System

We created a test script as discussed in Section 4.5 using the conversations from the user experiment to measure the performance of the DS. All the 24 user interactions captured during the experiment were translated to a test script. We run the test script on two different models created using the configurations P1 and P2 to compare the configuration for performance. The results for intent classification, entity recognition and response selection are presented in Table 6.2, 6.3 and 6.4 respectively. The confusion matrices generated are shown in the Appendix C. Accuracy, precision, recall and F1-score and the results are listed for intent classification, entity recognition and response selection are presented in Table 6.2, 6.3 and 6.4 respectively. The confusion matrices generated are shown in the Appendix C.

Pipeline	Accuracy	F1-Score	Precision	Recall
P1	0.590	0.560	0.655	0.489
P2	0.718	0.682	0.725	0.643

Table 6.2: Intent classification results for User tests

Pipeline	Accuracy	F1-Score	Precision	Recall
P1	0.946	0.829	0.875	0.787
P2	0.971	0.778	0.864	0.707

Table 6.3: Entity extraction results for User tests

Pipeline	Accuracy	F1-Score	Precision	Recall
P1	0.800	0.782	0.784	0.779
P2	0.812	0.800	0.801	0.799

Table 6.4: Response selection results for User tests

Pipeline configuration P2 as recommended by Rasa has shown higher performance for intent classification with an accuracy of 71.8%, entity recognition with an accuracy of 97.1% and response selection with an accuracy of 81.2%. Since our data is imbalanced and the distribution of examples across intents are not uniform, accuracy is not a reliable metric to measure model performance. The higher accuracy because of a single intent will lead to a higher overall accuracy. A system with high recall but low precision makes many false positive predictions leading to incorrect next actions which results in incomplete tasks. A system with high precision but low recall makes fewer but correct predictions. This leads to the risk of incomplete tasks despite the system being trained for the task.

To compare the systems, we can use the F1-scores. The configuration P2 shows a higher F1-score for intent classification and response selection however P1 shows a higher F1-score for entity recognition which could be a result of the differences in context of the entities with pre-trained embedding.

6.2.2 Task Success Rate

Task success rate measures the percentage of users who could complete tasks. This helps to evaluate how many tasks/user intents our system was able to assist. Tasks can either be completed successfully, or some parts of a task completed or may fail completely. To measure the task success rate, we have to first define task success levels to rate each user interaction from the user experiment. To rate task success for the user experiments, we used the following criteria to define different levels of tasks success:

1. **Success**
   System identifies correct user intent and provides an appropriate response or a fallback initiated when system can't identify the user intent

2. **Partial**
   System partially identifies the user intent and provides partial response

3. **Fail**
   System does not identify the user intent and provides incorrect response or can't identify and stops

Using the levels defined for task success, we rated the tasks and interactions from the user experiments to measure the task success rate. Figure 6.1 shows that 50% of the tasks were successfully completed, 29.16% of the tasks were partially completed and 20.83% of tasks failed to complete.

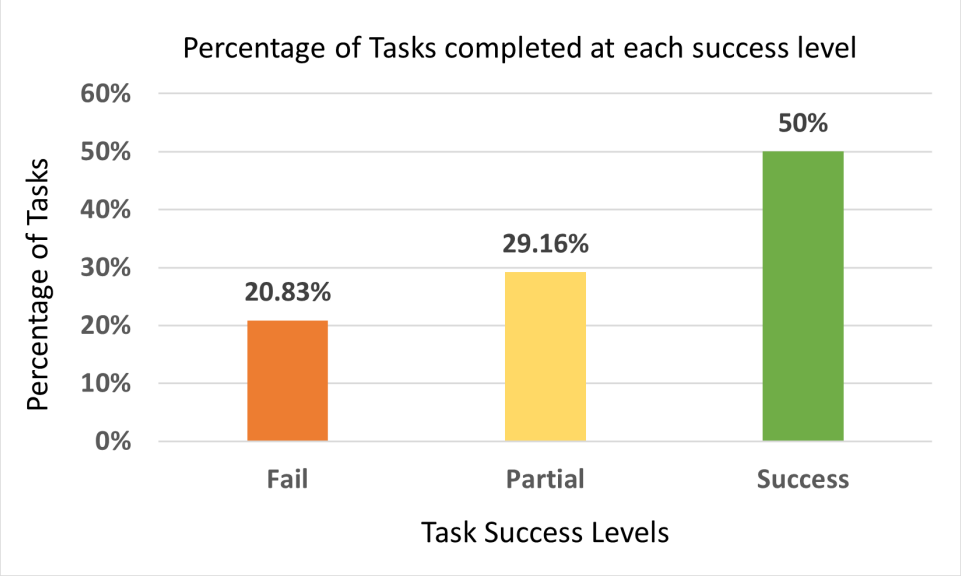


Figure 6.1: Percentage of Task success at each success level

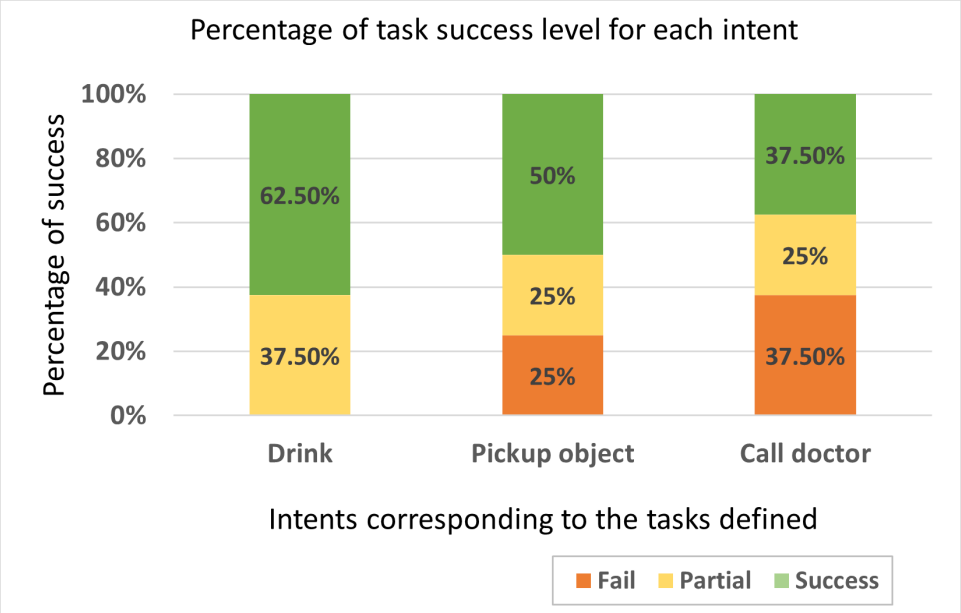


Figure 6.2: Percentage of Task success of Intents

Using the data in Table 6.1, we calculated the task success rate for each intent and show the results in Figure 6.2. We see that the *intent_drink* shows no task failures which could be due to the availability of sufficient training data and the word similarity to the language models.

6.2.3 User Questionnaire

After interacting with the system, users were asked to fill a questionnaire to rate the overall interaction. The users had to rate the number of tasks which they felt were completed, number of tasks in which relevant and appropriate responses were given by the system, the number of tasks in which the system identified all the details like the name of a drink, object or a person, if system asked for clarification when confused and the outcome and the overall ease of use. The results obtained from the user questionnaire are shared in Figure 6.3.

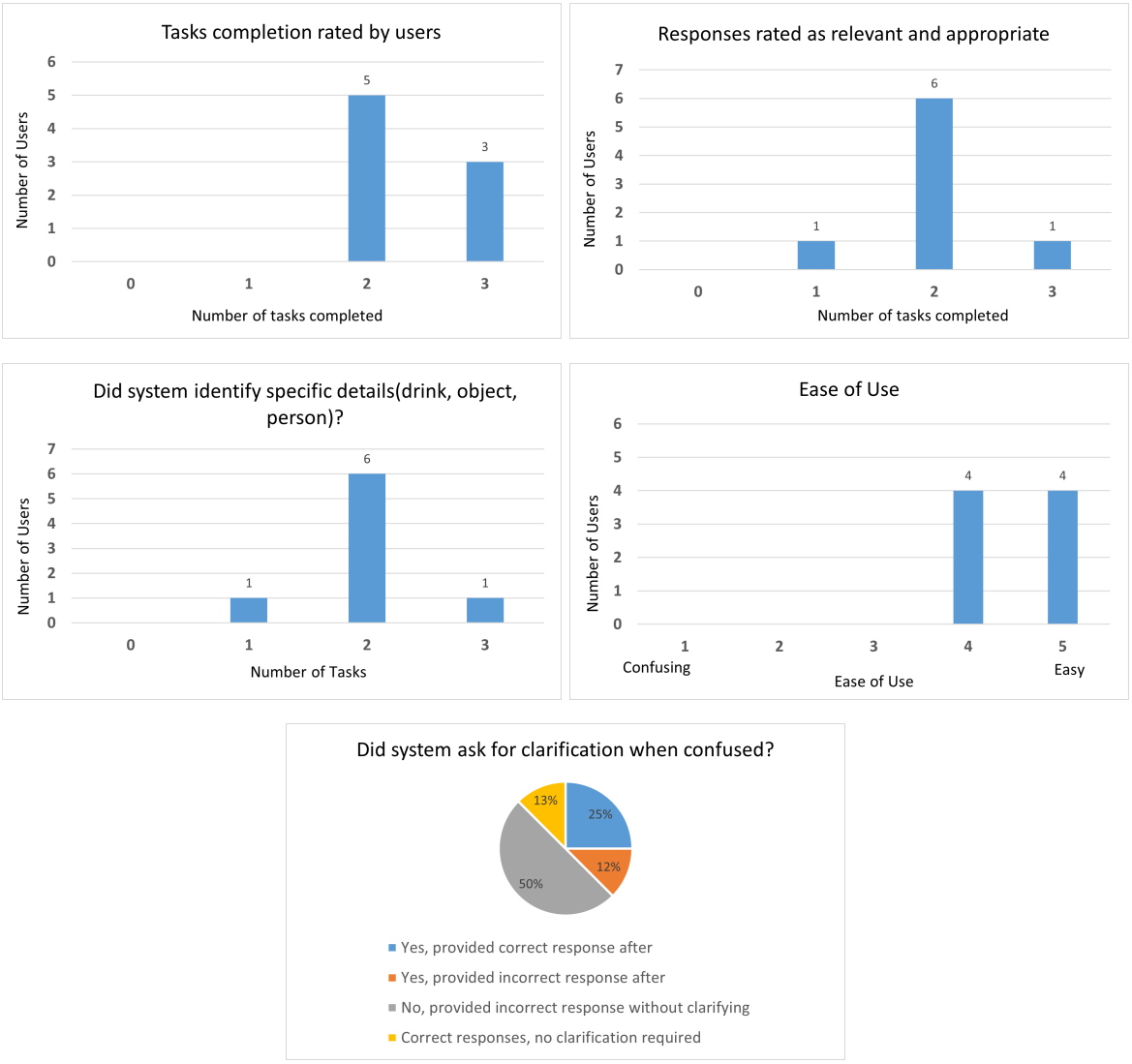


Figure 6.3: Results from the User questionnaire

Although the questionnaire doesn't provide conclusive results due to the small sample, we see that 62.5% participants rate two out of three tasks as complete, 75% of participants felt system responded appropriately and identified the entities for at least two out of three tasks and overall ease of use is above 4.5.

6.3 Qualitative Evaluation of the Dialogue System

The qualitative evaluations are made through observations during the experiment and analysis of the user interactions.

1. **Multiple intents**
   The dialogue system is trained to identify and respond to one intent at a time. However, from the user experiments, we see there are use cases where the users request for two intents simultaneously. As an example, when the user wants assistance to pick up a medicine box or pills, there's a likely chance they would ask for a glass of water with it which as per the design are two separate intents. Figure 6.4 shows sample conversations where the system could not handle multiple intents.

Figure 6.4: Example conversation where multiple intents fail

2. **Multiple entities**
   The dialogue system can identify a single entity from the user request at each turn. From the user experiment, we see that some of the user requests consists of two entities which the system could not handle. Figure 6.5 shows sample conversations where the system did not identify both entities. In the quantitative analysis, we defined a partial task success level which consists of tasks where the system completed the task partially by identifying only one of the entities.

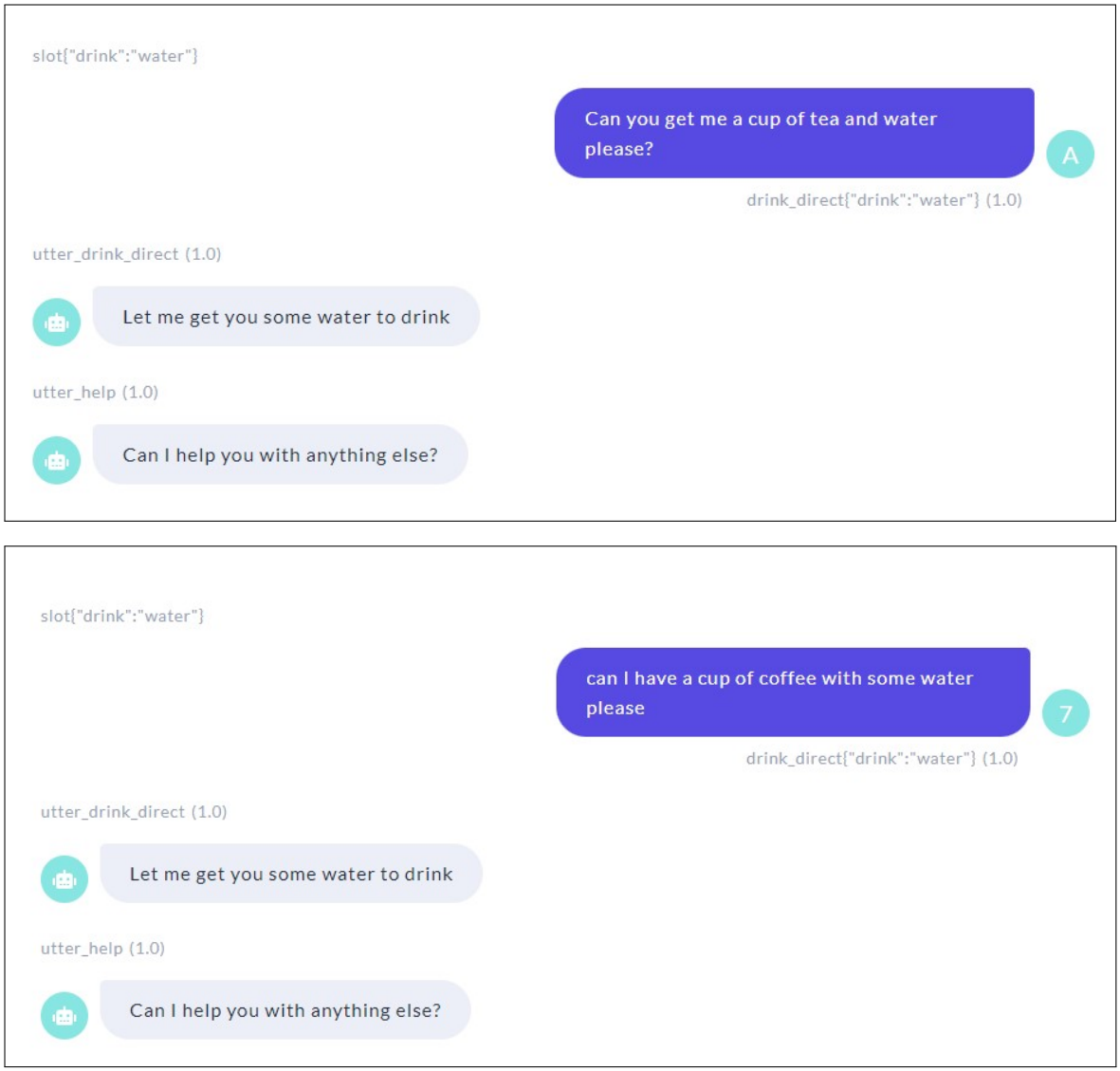


Figure 6.5: Example conversations where multiple entities are not recognised

3. **Repeated errors**
   The TEDPolicy makes a prediction using the previous states of the dialogues and responses. During the user experiment, when the system fails to recognise the intent, users attempt to try again. Because of the TEDPolicy, the previous error remains in the memory which it uses and makes the same incorrect prediction again unless corrected and feedback provided. As we focused on evaluating the prototype, we did not make corrections to the system during this experiment.

4. **Rules constraints**
   By nature of the rules defined, the dialogue system ends the interaction with a greeting. This is not logical when users call for a doctor or a medic, as the next action ideally would be to wait.

6.4 Conclusion

In this section, we discuss research questions from Section 1.3 to answer the main Research question *How can we design and evaluate a dialogue system for a care robot assisting the elderly in a care environment?*

- **R.Q.1. What are the daily activities in care environments for which the elderly require assistance with?**

  In Section 5.1, we summarised the daily activities taking place in care homes by conducting an interview with a nurse who has experience working in hospitals and care homes especially assisting the elderly. From the interview we discussed the activities which the elderly seek assistance for on a daily basis. Through the interview, we identified different tasks that a care robot would require knowledge about to communicate and effectively assist the elderly in care environments. Using the insights from the interview, we selected three tasks for the dialogue system- identify when a user asks for a drink and which drink, identify when the user needs help to pickup an object and understand when the user needs help call a doctor. These are the use-cases considered for designing the dialogue system in our context of a care robot assisting the elderly with daily activities. Although the interview gave us insights regarding the activities, we could not gather observations regarding the dialogues or study the behaviour of the elderly. Conducting a field study would help in understanding the activities and the environment better.

- **R.Q.2. How can we design a dialogue system for a care robot to assist with daily activities without the availability of large amounts of training data?**

  From the literature review, we identify a gap in the availability of recorded or labelled data(dialogues and conversations) to design and develop a dialogue system for a care robot to assist the elderly with daily activities. In Section 3.1, we review approaches taken by researches to design a dialogue system for domains with limited or without the availability of large amounts of training data. We proposed to use the framework Rasa to design a rule-based dialogue system capable of handling simple conversations for the use-cases defined in Section 5.1 and the use of incremental learning for further development of the system. The implementation shows that the framework is flexible

and supports multiple configurations for intent classification, entity extraction and dialogue management without the need for advanced programming knowledge of NLP. The user interface allows easier additions of intents, entities and rectifying errors.

- **R.Q.3. How can we evaluate the performance of a dialogue system for the care robot and find the limitations for future developments?**

  As discussed earlier, evaluating a dialogue system is a challenging task. In the phase of our research, we explored different metrics to evaluate the performance of the system. In Section 6.2, we measured the performance using the automated classification metrics from Rasa, calculated the task success rate based on success levels and intents and through a questionnaire where users rated the tasks on completion, relevance and appropriateness and ease of use. Although the results from the evaluations are not conclusive due to the small sample size, we could identify design limitations for future studies. The tests run through Rasa automatically generate classification metrics and a confusion matrix. The confusion matrix is useful for identifying intents and entities frequently mistaken for other intents. This helps the developer analyse if design changes or modifications to the intents are required. Accuracy is not a suitable metric for an imbalanced dataset. An ideal system would have high precision and high recall which means all the tasks/intents are correctly predicted. Although this is difficult to achieve, training the system with more data, dialogues and conversations with users could improve the F1-score. This method takes the least human intervention where the developer or researcher only needs to create a test set. The task success rates is a simple and common method to evaluate the dialogue system as it clearly shows the percentage of successful task completions which is the main goal of a task-oriented dialogue system. In our study, the user interactions were rated by the researcher based on the design and expected behaviour of the system, but many studies have considered evaluating using crowd-sourcing. We also conducted a user survey where users filled in a questionnaire to rate the system based on tasks. Although we used the questionnaire for users to rate the tasks for this stage of research, user surveys can be conducted to evaluate the usability of the system.

Chapter 7

Discussion

In this research, we conducted an exploratory study to design and evaluate a dialogue system for a care robot Rose, developed by HiT, to assist the elderly with daily activities in care environments. The scope for this research was to conduct an exploratory study to design and evaluate an early prototype. We conclude this report by discussing the limitations in this study and providing recommendations for future work.

7.1 Limitations of the study

- In the real context of our study, the target users are the elderly. However, participants in this study were colleagues from HiT. The main focus of our study was to explore design and evaluation methods for a dialogue system without the availability of large amounts of training data in the care domain. However, future studies should be conducted with the target users since the elderly may have hearing impairments or other speech problems which may require additional design changes.
- In this study, to avoid errors due to speech translation, we used a text input to evaluate the dialogue system using the Rasa framework. However, for practical implementation, a user should be able to speak to a dialogue system.
- In the experiment designed, we only included tasks for the main activities of daily assistance like requesting a drink, picking up an object and calling a doctor. However the system is also trained for other intents like greetings, goodbye and out of scope examples which were not explicitly tested in the experiment, but were part of all the interactions.
- For the user experiment, the questionnaire used was adapted from different resources and it was difficult to gather any conclusions from it. The question-

naire could have been structured better by clearly defining the requirements and expected outcomes from the user study. A detailed review of available questionnaires relevant to our study could have provided meaningful results.

7.2 Future Recommendations

- **Training data**
  It is not feasible to anticipate every user request or intent which the care robot maybe faced with in the care environment. Dialogue systems require continuous through annotated data of user conversations or through techniques like incremental learning where the system can be trained while the users interact with the system and a human in the loop to correct and give feedback to the system. In this paper, one of the reasons to choose Rasa for development was that it supports and has a platform(Rasa X interface) for incremental learning. Using the early prototype developed in this study, more user experiments can be conducted to gather or train the system on the go. In Section 4.1, we see that large datasets required for designing task-oriented dialogue systems are collected by conducting Wizard-Of-Oz experiments and crowd sourcing which can be considered in the future.

- **Multiple intents and entities**
  In Section 6.2, the analysis showed that system failed when the user request included more than one entity or intent. This scenario was not anticipated while designing the system. However, our results show that in the context of the care environment, there are high chances where the elderly require assistance with multiple entities and intents in the same user request. The dialogue system design should consider this and identify the best possible way to handle the conversation. Multiple intents can be handled by changing the type of the entity and making it a list. Based on the scenario, the intents may have to be modified, or the system should handle them one by one. This can be achieved by training the system to handle multiple intents.

- **Modifying Fallback**
  Dialogue systems handle uncertainties using fallbacks. The default approach is to ask the user to rephrase the user request. However, if the system fails to recognise the user request again, it prompts an out of scope without any further attempts or hands over to a human operator. A possible strategy to handle a fallback would be to ask clarifying questions to help the system make correct predictions. To ask the right clarifying questions, the system requires domain knowledge represented graphically ontologies. Härkönen et al. (2022)

integrated an ontology in the field of computers and integrated in with a chatbot created using Rasa. Langensiepen et al. (2016) proposed a method to generate an ontology in the context of elderly care. These studies can further be investigated and implemented for the dialogue system.

Appendix

A Interview Questions

Understanding care home environments and interactions

Goal:
Understand the everyday routine and situations in carehome and responsibilities of a nurse. To learn the flow of carehome conversations and identify key scenarios which could act as input to create a dialogue management system that identifies intents.

Carehome environments and background:

1. Do carehomes face issues with staffing?
    - If yes, how are they managed?
    - Were there changes seen especially during Covid times
    - How were the resoucres and workforce managed? Any use of technology especially for social distancing
2. How often to emergency situations occur?
3. Are the emergencies alerted through conversations or machines like ECG?
4. Could you describe some of the emergency situations you've come across and the response to that.

    General perception about healthcare robots:

5. Since you have been working with Rose for a long time, when and what kind of assistance do you think a robot can provide in healthcare homes in an everyday environment?
6. How do you think patients would react to robots assisting them?
7. Do you feel certain activities/tasks performed by nurses could be performed by robots provided the are trained for it.

    Interaction with patients:

8. How many times on average do you interact with patient(s) in a day?
9. Can you describe these interactions and the activities involved
10. What kind of assistance is provided during these situations?
11. Could you describe situations that occur at night where patients need help?

12. The basic goal for this thesis, is to create a dialogue system which can understand a patients intent and act according to the situation, could you give us examples of scenarios or situations that maybe useful to work on.

B Tasks for User Experiment

Task 1

Imagine you are one of the elderly people residing at a care home. You finish your routine exercises and feel thirsty. You see Robot Rose and ask for something to drink.

Task 2

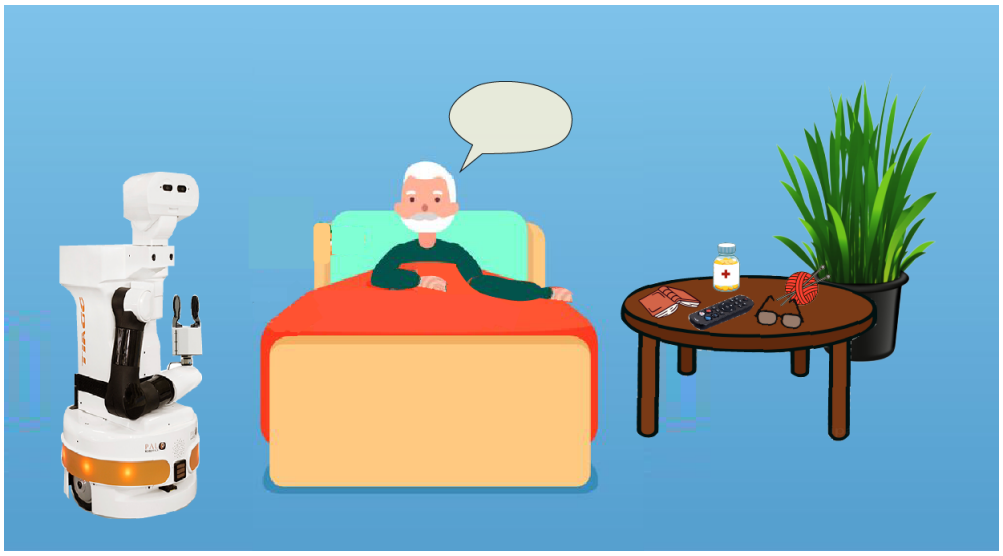
What do you think the elderly woman is saying to Robot Rose?

Task 3

Imagine you are one of the elderlies residing at a care home. You are resting in your bed and see Robot Rose. You want the medicines from the table next to you. How would you ask Robot Rose for help?

Task 4

What do you think the elderly man is saying to Robot Rose?

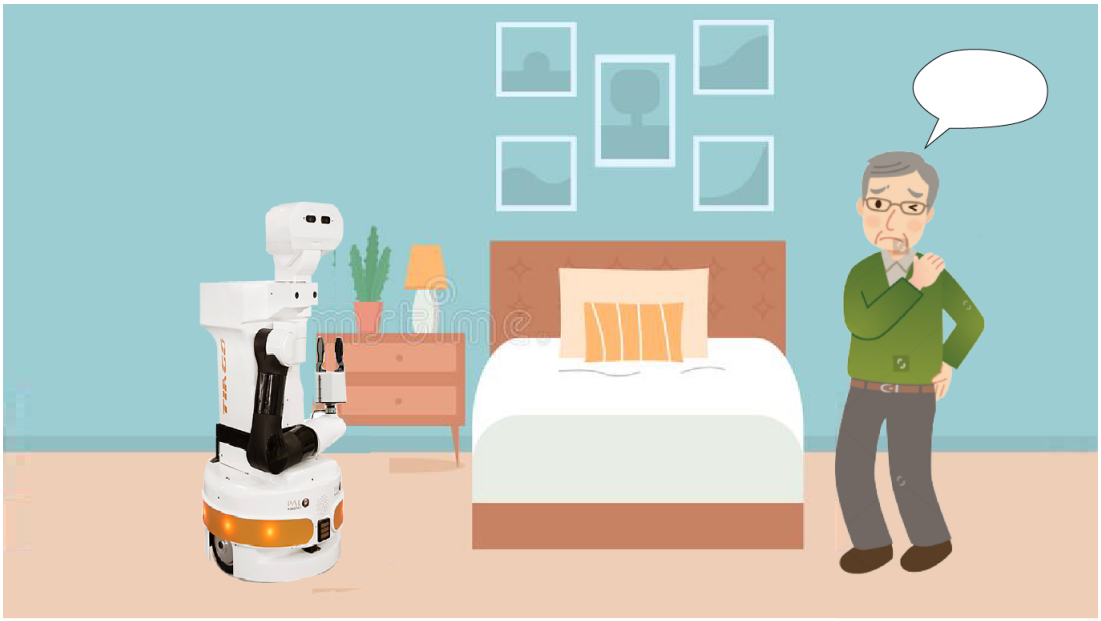


Task 5

Imagine you are one of the elderlies residing at a care home. You feel a pain in your leg. You see Robot Rose and ask to call someone for help.

Task 6

What do you think the elderly man is saying to Robot Rose?



C Confusion Matrices

C.1 Intent Classification

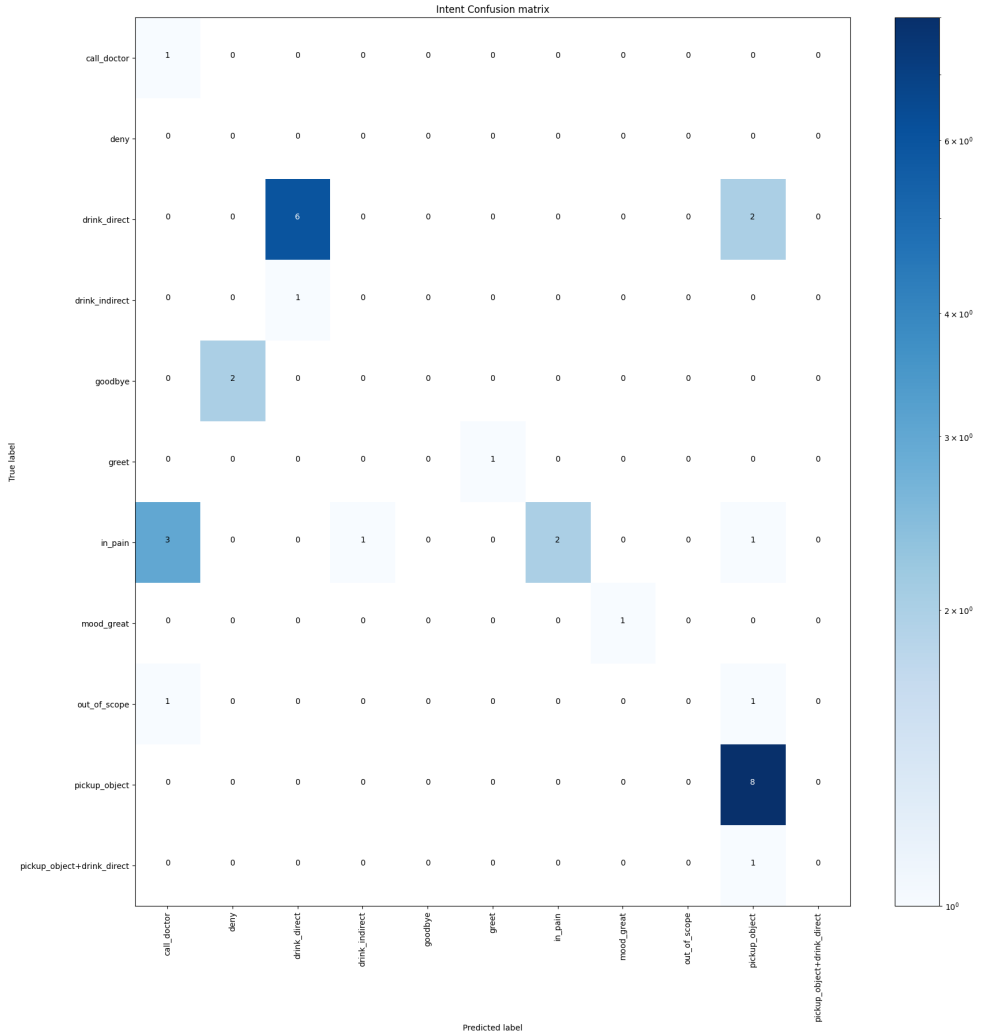


Figure 1: Intent confusion matrix for pipeline P1

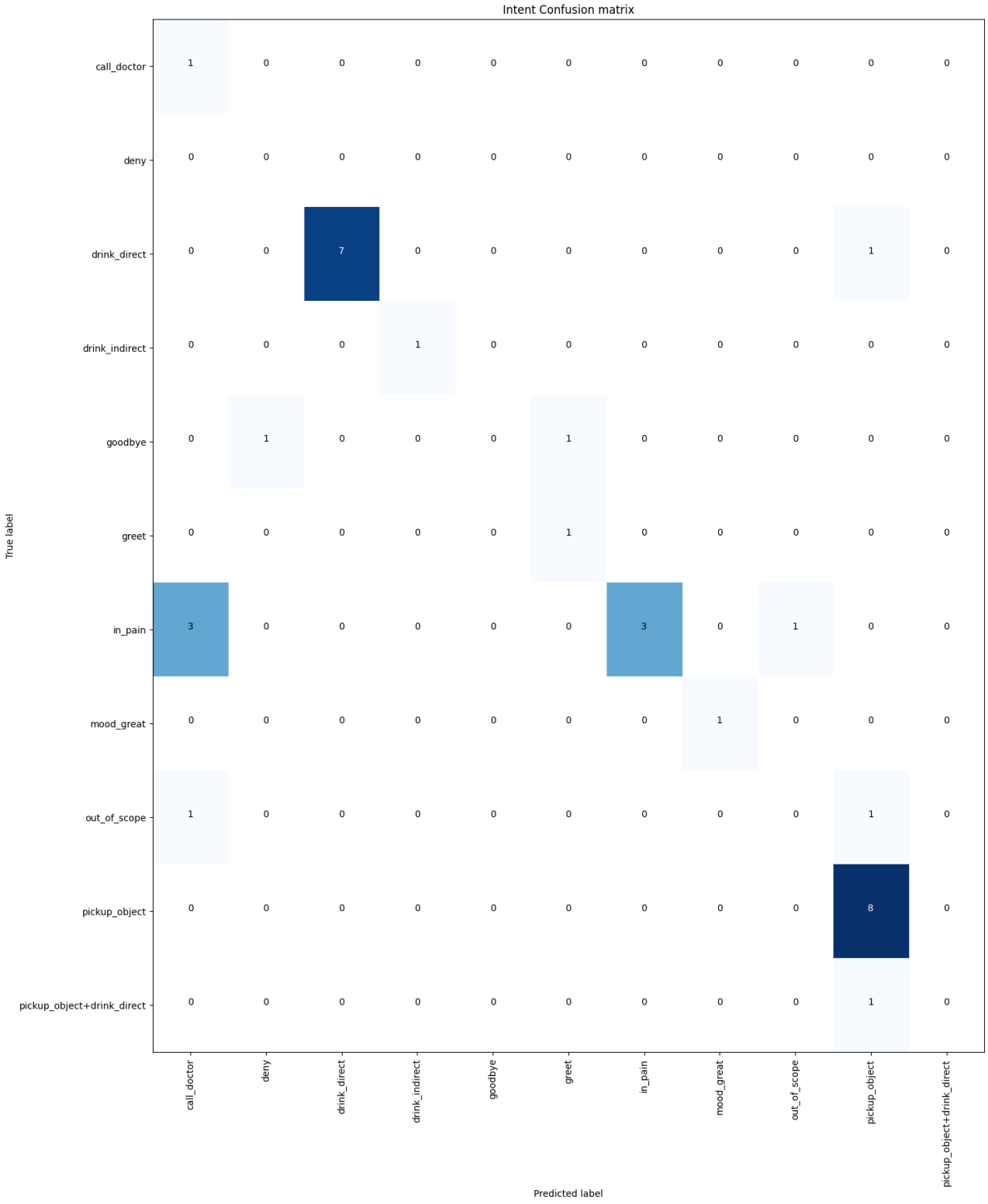


Figure 2: Intent confusion matrix for pipeline P2

C.2 Entity Recognition

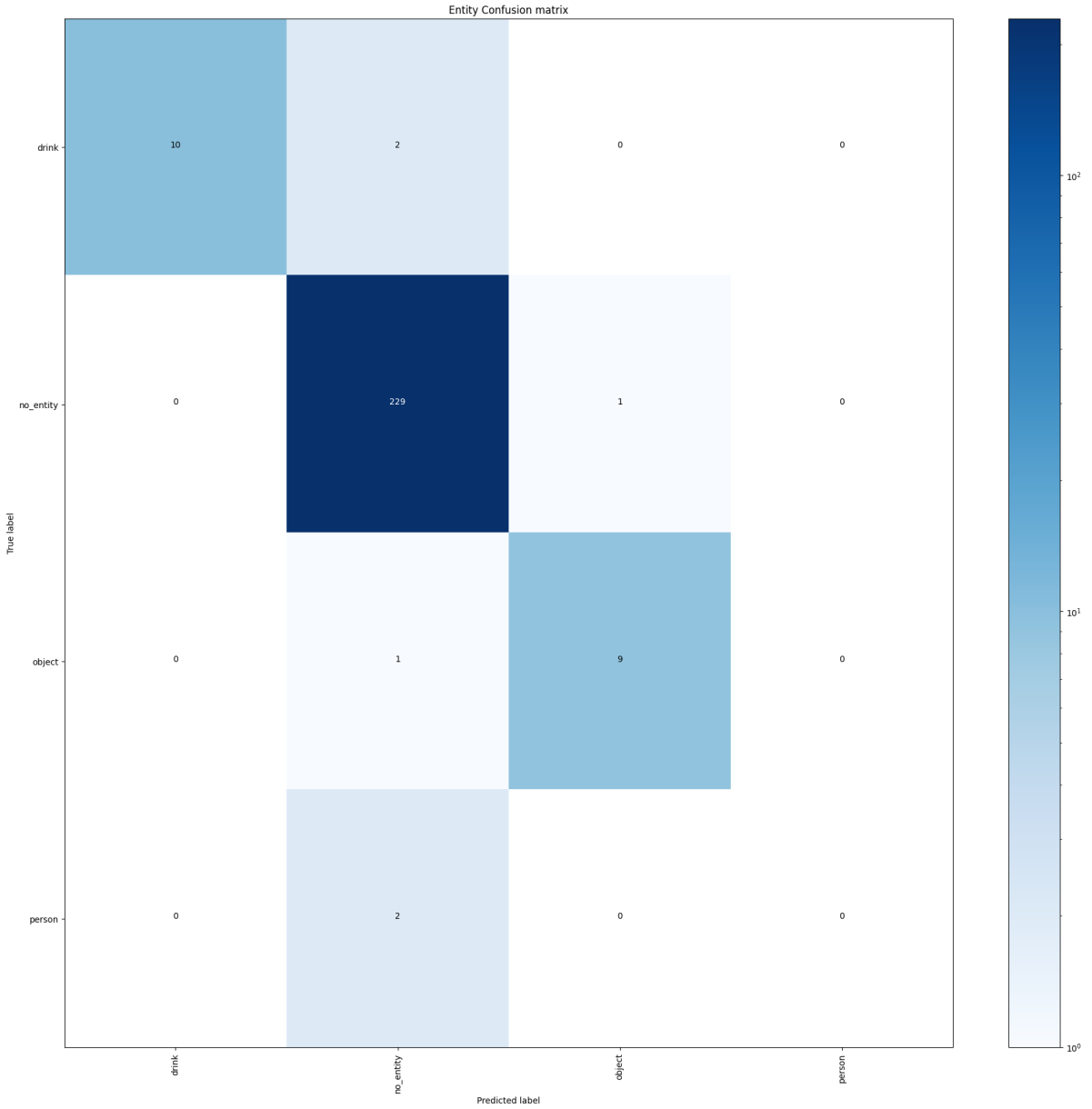


Figure 3: Entity confusion matrix for pipeline P1

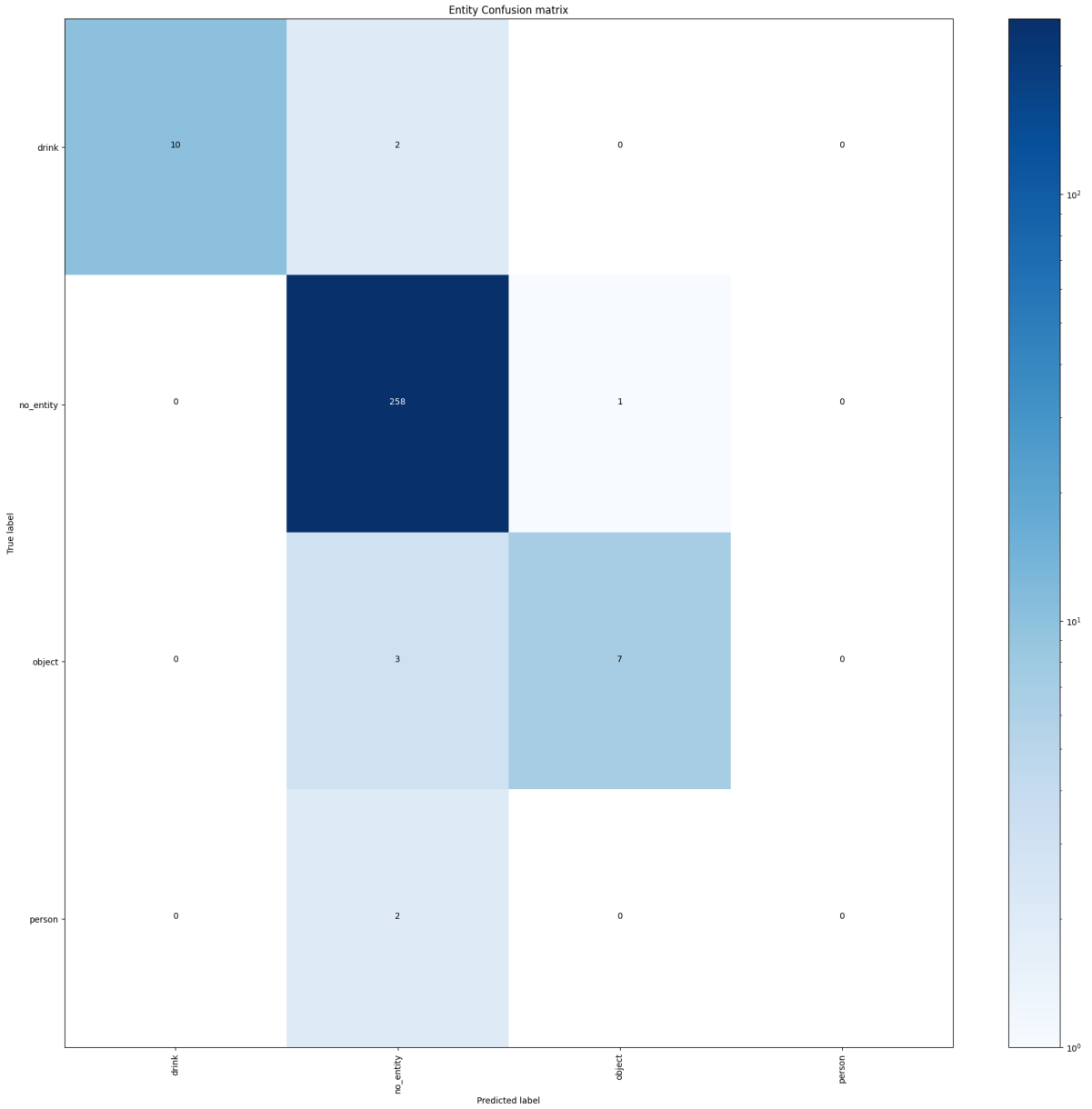


Figure 4: Entity confusion matrix for pipeline P2

C.3 Response Selection

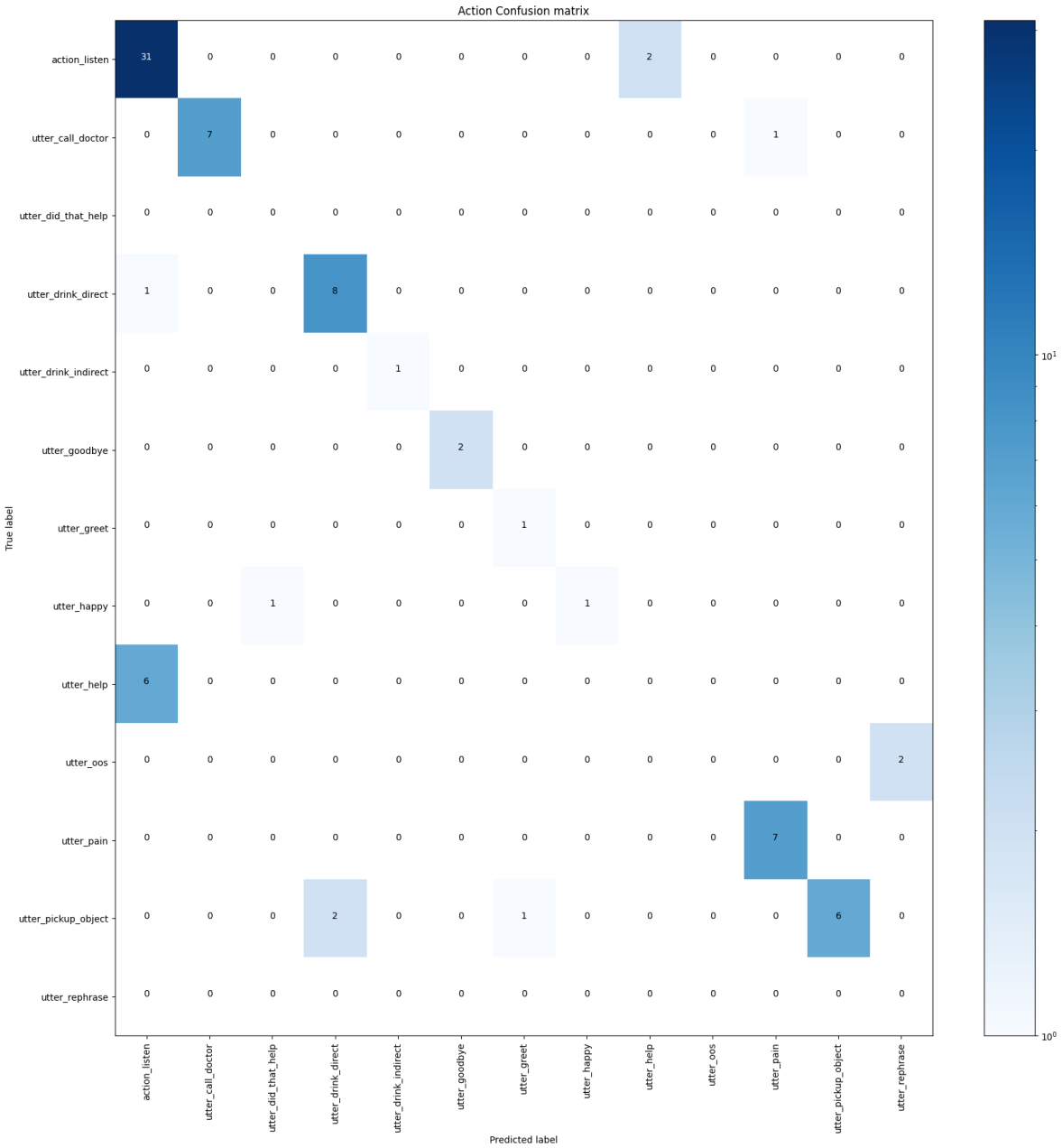


Figure 5: Response selection confusion matrix for pipeline P1

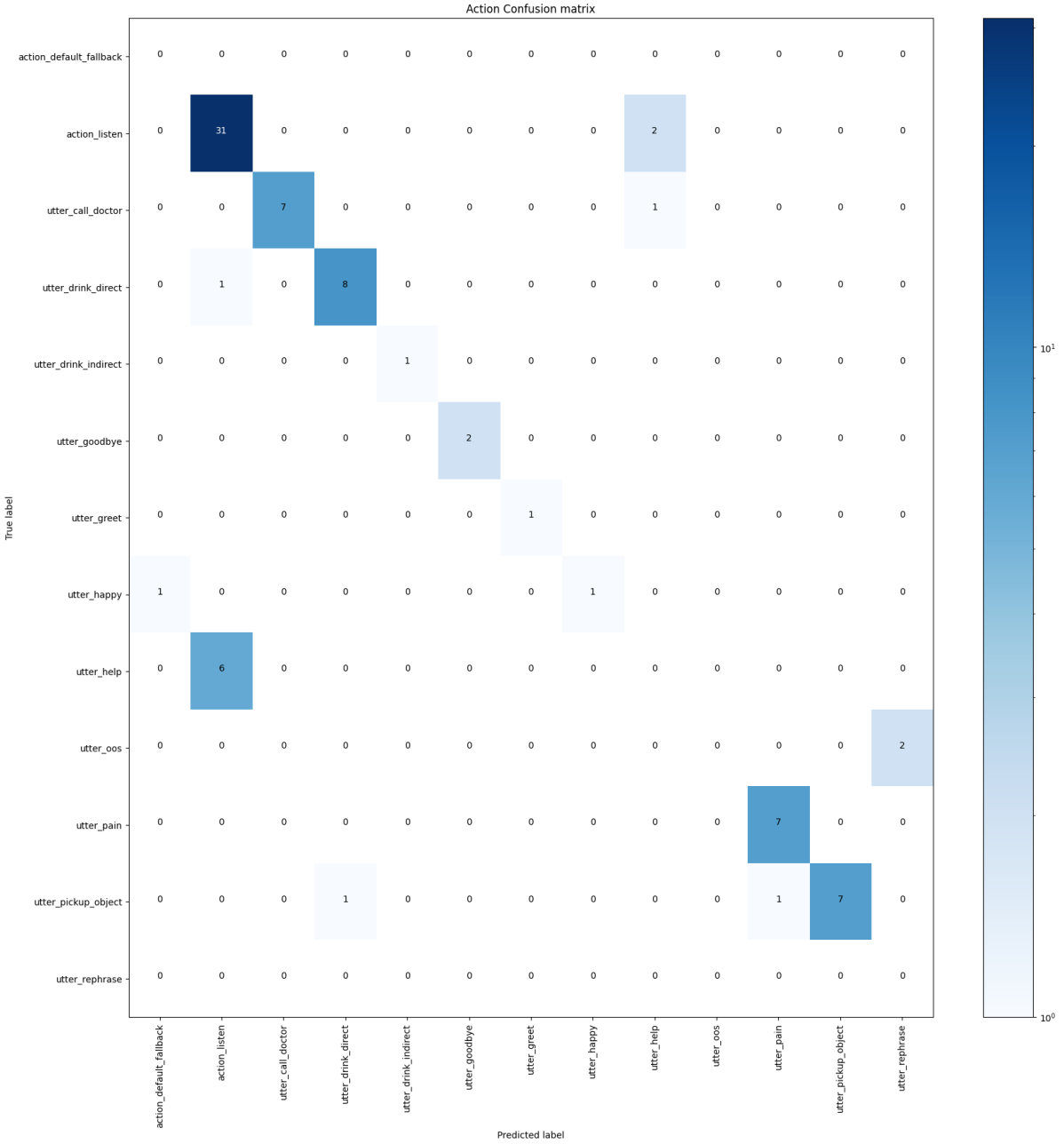


Figure 6: Response selection confusion matrix for pipeline P2

D Task Success Rating

Task success rating								
Tasks/ Participants	P1	P2	P3	P4	P5	P6	P7	P8
T1	0.5(T1)	0.5(T2)	1(T1)	0.5(T2)	1(T1)	1(T2)	1(T1)	1(T2)
T2	1(T4)	1(T3)	1(T6)	1(T5)	0(T3)	1(T3)	0(T4)	0.5(T4)
T3	0.5(T5)	0.5(T6)	1(T3)	0.5(T4)	0(T6)	0(T5)	1(T6)	0(T5)
Success	1	Values used only for representaton						
Partial	0.5							
Failure	0							

E Consent Form

Consent Form for Dialogue system for care robots

YOU WILL BE GIVEN A COPY OF THIS INFORMED CONSENT FORM

Please tick the appropriate boxes	**Yes**	**No**
Taking part in the study		
I have read and understood the study information dated 10/10/2021, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	○	○
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	○	○
I understand that taking part in the study involves capturing dialogues or utterances while interacting with the system and a survey questionnaire completed by the participant. The utterances captured will only be transcribed as text and no audio recordings are captured. I understand that this research is in collaboration with Heemskerk Innovative Technology who will be the owner of the developed dialogue system and the owner of the transcribed data sets, interviews, and questionnaires.	○	○
Use of the information in the study		
I understand that information I provide will be used for the thesis report and as data for all future developments of the system	○	○

Signatures

_____	_____	_____
Name of participant	Signature	Date

I have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

Sweta Balamurali	_____	_____
Researcher name	Signature	Date

UNIVERSITY OF TWENTE.

Study contact details for further information: [*Name, email address*]

Contact Information for Questions about Your Rights as a Research Participant

If you have questions about your rights as a research participant, or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than the researcher(s), please contact the Secretary of the Ethics Committee Computer & Information Science at the University of Twente by

Researcher: Sweta Balamurali
E-mail: s.balamurali@student.utwente.nl
Primary supervisor: DR. R.J.F. ORDELMAN (ROELAND)
E-mail: roeland.ordelman@utwente.nl
CIS Ethics Committee to file a complaint: ethicscommittee-cis@utwente.nl

UNIVERSITY OF TWENTE.

F Information Brochure

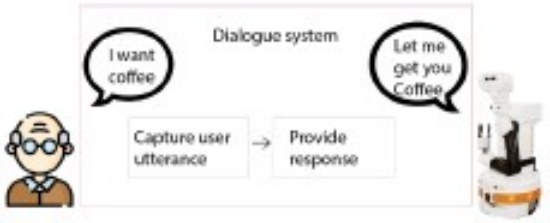
Information sheet

Exploratory study to design and evaluate a Dialogue system for a care home robot

The goal of the thesis is to conduct an exploratory study to design a develop a dialogue system for a robot used in care homes. This information sheet provides a brief about the purpose of research, type of data collected and stored and contact information for further questions or clarifications.

- **Purpose of the research**

A dialogue system helps understand user's utterances, understands the intent, and responds to the users appropriately. Current research in dialogue systems is prominent in specific domains like personal assistants and flight/restaurant reservation systems, however a dialogue system for care robots is still in the initial stages of research. In this thesis, through literature we explored different approaches to design a dialogue system and implemented a basic dialogue system using the framework Rasa. The implementation understands the user message, classifies the intent category and provides an appropriate response.



- **Details about the experiment**

Through this experiment, we test how the dialogue system performs with users and and analyse the results to find the limitations and weaknesses to further enhance the system.

- Participants will be presented with tasks and based on it they must speak to the dialogue system
- If the system responds appropriately, the task is complete and if not, the user may try again
- Each participant will be given three tasks and the experiment will last for 30 mins
- At the end of the experiment, participants are requested to rate the system based on appropriateness of the responses provided by the dialogue system (the prototype)
- The user utterances (dialogues spoken by the participants) are captured as text for further developments. All utterances are in the form of text, thereby no personally identifiable information like voice recorded. To ensure anonymity, participants are requested not to utter/ provide any sensitive information as utterances are automatically captured and converted to text.

UNIVERSITY OF TWENTE.

- **Benefits and risks of participating**
  There are no risks involved in participating in this experiment. No personal data or data that can identify the participants are captured. Everything including the dialogues captured and the survey results will be anonymous. The research project has been reviewed and approved by the CIS Ethics Committee

- **Procedures for withdrawal from the study**
  - Participants may stop participating at any time and without needing to give a reason.
  - No personal data or data that can identify the participants are captured.
  - The data captured include the dialogues spoken during the experiment and a usability survey gathered at the end of the experiment. The data captured will be used for analysing outcomes and for training the system, however no personal information is included here. All data will be anonymous.
  - The research data(feedbacks) will be retained for further development and improving the performance of the system.
  - For questions about Your Rights as a Research Participant or any other concerns, the following contact details can be reached for clarifications.
    Researcher:
    Name: Sweta Balamurali
    E-mail: s.balamurali@student.utwente.nl
    Primary supervisor:
    Name: DR. R.J.F. ORDELMAN (ROELAND)
    E-mail: roeland.ordelman@utwente.nl
    CIS Ethics Committee to file a complaint: ethicscommittee-cis@utwente.nl

UNIVERSITY OF TWENTE.

References

Abdellatif, A., Badran, K., Costa, D. E., & Shihab, E. (2020, 12). A Comparison of Natural Language Understanding Platforms for Chatbots in Software Engineering. doi: 10.1109/TSE.2021.3078384

Allouch, M., Azaria, A., & Azoulay, R. (2021, 12). Conversational agents: Goals, technologies, vision and challenges. *Sensors*, *21*(24). doi: 10.3390/S21248448

Bengtsson, T., & Qi, H. (2018). Ageing workforce, social cohesion and sustainable development: Political challenges within the baltic sea region: Sweden.

Braun, D., Hernandez-Mendez, A., Matthes, F., & Langen, M. (2017). Evaluating Natural Language Understanding Services for Conversational Question Answering Systems. In *Proceedings of the 18th annual sigdial meeting on discourse and dialogue.* Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.18653/v1/W17-5522

Broekens, J., Heerink, M., & Rosendal, H. (2009, 4). Assistive social robots in elderly care: a review. *Gerontechnology*, *8*(2). doi: 10.4017/gt.2009.08.02.002.00

Buettner, T. (2015, 11). Urban Estimates and Projections at the United Nations: The Strengths, Weaknesses, and Underpinnings of the World Urbanization Prospects. *Spatial Demography*, *3*(2), 91–108. doi: 10.1007/s40980-015-0004-2

Bunk, T., Varshneya, D., Vlasov, V., & Nichol, A. (2020, 4). DIET: Lightweight Language Understanding for Dialogue Systems.

Celikyilmaz, A., Hakkani-Tur, D., Tur, G., Fidler, A., & Hillard, D. (2011, 12). Exploiting distance based similarity in topic models for user intent detection. In *2011 ieee workshop on automatic speech recognition & understanding.* IEEE. doi: 10.1109/ASRU.2011.6163969

Chen, Liu, X., Yin, D., & Tang, J. (2017, 11). A Survey on Dialogue Systems. *ACM SIGKDD Explorations Newsletter*, *19*(2), 25–35. doi: 10.1145/3166054.3166058

Chen, Z., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013, June). Identifying intention posts in discussion forums. In *Proceedings of the 2013 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 1041–1050). Atlanta, Georgia: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N13-1124`

Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., ... Dureau, J. (2018, 5). Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces.

Dai, Y., Li, H., Tang, C., Li, Y., Sun, J., & Zhu, X. (2020). Learning Low-Resource End-To-End Goal-Oriented Dialog for Fast and Reliable System Deployment. In *Proceedings of the 58th annual meeting of the association for computational linguistics.* Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.57

De Carolis, B., Flace, G., Macchiarulo, N., Melone, G., & la Forgia, A. (2021). Dialog management for a social assistive robot in the domain of elderly care. In *Aixas@ ai* ia* (pp. 41–51).

Deng, L., Tur, G., He, X., & Hakkani-Tur, D. (2012, 12). Use of kernel deep convex networks and end-to-end learning for spoken language understanding. In *2012 ieee spoken language technology workshop (slt).* IEEE. doi: 10.1109/SLT.2012.6424224

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., & Cieliebak, M. (2021, 1). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, *54*(1). doi: 10.1007/s10462-020-09866-x

Eric, M., Goel, R., Paul, S., Kumar, A., Sethi, A., Ku, P., ... Hakkani-Tur, D. (2019, 7). MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines.

Genkin, A., Lewis, D. D., & Madigan, D. (2007, 8). Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics*, *49*(3). doi: 10.1198/004017007000000245

Glende, S., Conrad, I., Krezdorn, L., Klemcke, S., & Krätzel, C. (2016, 6). Increasing the Acceptance of Assistive Robots for Older People Through Marketing Strategies Based on Stakeholder Needs. *International Journal of Social Robotics*, *8*(3), 355–369. doi: 10.1007/s12369-015-0328-5

Goetze, S., Fischer, S., Moritz, N., Appell, J.-E., & Wallhoff, F. (2012). Multimodal human-machine interaction for service robots in home-care environments. In *Proceedings of the 1st workshop on speech and multimodal interaction in assistive environments* (pp. 1–7).

Gomi, T., & Griffith, A. (n.d.). Developing intelligent wheelchairs for the handicapped. In *Assistive technology and artificial intelligence* (pp. 150–178). Berlin/Heidelberg: Springer-Verlag. doi: 10.1007/BFb0055977

Granata, C., Chetouani, M., Tapus, A., Bidaud, P., & Dupourque, V. (2010, 9). Voice and graphical -based interfaces for interaction with a robot dedicated to elderly and people with cognitive disorders. In *19th international symposium in robot and human interactive communication* (pp. 785–790). IEEE. doi: 10.1109/ROMAN.2010.5598698

Haffner, P., Tur, G., & Wright, J. (n.d.). Optimizing SVMs for complex call classification. In *2003 ieee international conference on acoustics, speech, and signal processing, 2003. proceedings. (icassp '03).* IEEE. doi: 10.1109/ICASSP.2003.1198860

Härkönen, A.-P., et al. (2022). Computationally clarifying user intent for improved question answering.

Harms, J.-G., Kucherbaev, P., Bozzon, A., & Houben, G.-J. (2019, 3). Approaches for Dialog Management in Conversational Agents. *IEEE Internet Computing*, *23*(2), 13–22. doi: 10.1109/MIC.2018.2881519

Ilievski, V., Musat, C., Hossmann, A., & Baeriswyl, M. (2018, 2). Goal-Oriented Chatbot Dialog Management Bootstrapping with Transfer Learning.

Jiang Zhao, Y., Ling Li, Y., & Lin, M. (2019, 7). A Review of the Research on Dialogue Management of Task-Oriented Systems. In *Journal of physics: Conference series* (Vol. 1267). Institute of Physics Publishing. doi: 10.1088/1742-6596/1267/1/012025

Jokinen, K. (2018). Dialogue Models for Socially Intelligent Robots.. doi: 10.1007/978-3-030-05204-1{_}13

Jurafsky, D., & Martin, J. H. (2020). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson.

Kiguchi, K., Rahman, M. H., Sasaki, M., & Teramoto, K. (2008, 8). Development of a 3DOF mobile exoskeleton robot for human upper-limb motion assist. *Robotics and Autonomous Systems*, *56*(8), 678–691. doi: 10.1016/j.robot.2007.11.007

Langensiepen, C., Lotfi, A., Chernbumroong, S., Moreno, P. A., & Gómez, E. J. (2016, 6). A New Way to Build Multifacetted Ontologies for Elderly Care. In *Proceedings of the 9th acm international conference on pervasive technologies related to assistive environments* (pp. 1–6). New York, NY, USA: ACM. doi: 10 .1145/2910674.2935831

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., ... Coiera, E. (2018, 9). Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, *25*(9). doi: 10.1093/jamia/ocy072

Liu, J., Li, Y., & Lin, M. (2019, 7). Review of Intent Detection Methods in the Human-Machine Dialogue System. *Journal of Physics: Conference Series*, *1267*. doi: 10.1088/1742-6596/1267/1/012059

Liu, X., Eshghi, A., Swietojanski, P., & Rieser, V. (2019, 3). Benchmarking Natural Language Understanding Services for building Conversational Agents.

Malamas, N., & Symeonidis, A. (2021). Embedding Rasa in edge Devices: Capabilities and Limitations. *Procedia Computer Science*, *192*, 109–118. doi: 10.1016/j.procs.2021.08.012

Mathew, K. V., Tarigoppula, V. S. A., & Frermann, L. (2021). Multi-modal intent classification for assistive robots with large-scale naturalistic datasets. In *Proceedings of the the 19th annual workshop of the australasian language technology association* (pp. 47–57).

McCallum, A., Nigam, K., et al. (1998). A comparison of event models for naive bayes text classification. In *Aaai-98 workshop on learning for text categorization* (Vol. 752, pp. 41–48).

McTear, M. F. (2004). Dialogue Engineering: The Dialogue Systems Development Lifecycle. In *Spoken dialogue technology* (pp. 129–161). London: Springer London. doi: 10.1007/978-0-85729-414-2{_}7

Nguyen, M.-T., Tran-Tien, M., Viet, A. P., Vu, H.-T., & Nguyen, V.-H. (2021, 11). Building a Chatbot for Supporting the Admission of Universities. In *2021 13th international conference on knowledge and systems engineering (kse)* (pp. 1–6). IEEE. doi: 10.1109/KSE53942.2021.9648677

Ni, J., Young, T., Pandelea, V., Xue, F., Adiga, V., & Cambria, E. (2021, 5). Recent Advances in Deep Learning Based Dialogue Systems: A Systematic Survey.

Qi, Y., Sachan, D. S., Felix, M., Padmanabhan, S. J., & Neubig, G. (2018, 4). When and Why are Pre-trained Word Embeddings Useful for Neural Machine Translation?

Quan, J., Yang, M., Gan, Q., Xiong, D., Liu, Y., Dong, Y., ... Jiang, D. (2021, 2). Integrating Pre-trained Model into Rule-based Dialogue Management.

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y.-L. (2018, 10). Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset.

Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). Transfer Learning in Natural Language Processing. In *Proceedings of the 2019 conference of the north* (pp. 15–18). Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.18653/v1/N19-5004

Shalyminov, I., Sordoni, A., Atkinson, A., & Schulz, H. (2020, 3). Hybrid Generative-Retrieval Transformers for Dialogue Domain Adaptation.

Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., ... Fox, D. (2019, 12). ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks.

Smarr, C.-A., Mitzner, T. L., Beer, J. M., Prakash, A., Chen, T. L., Kemp, C. C., & Rogers, W. A. (2014, 4). Domestic Robots for Older Adults: Attitudes, Preferences, and Potential. *International Journal of Social Robotics*, *6*(2), 229–247. doi: 10.1007/s12369-013-0220-0

Tokunaga, S., Tamura, K., & Otake-Matsuura, M. (2021, 6). A Dialogue-Based System with Photo and Storytelling for Older Adults: Toward Daily Cognitive Training. *Frontiers in Robotics and AI*, *8*. doi: 10.3389/frobt.2021.644964

Tur, G., Deng, L., Hakkani-Tur, D., & He, X. (2012, 3). Towards deeper understanding: Deep convex networks for semantic utterance classification. In *2012 ieee international conference on acoustics, speech and signal processing (icassp).* IEEE. doi: 10.1109/ICASSP.2012.6289054

Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1997). PARADISE. In *Proceedings of the eighth conference on european chapter of the association for computational linguistics -.* Morristown, NJ, USA: Association for Computational Linguistics. doi: 10.3115/979617.979652

Weizenbaum, J. (1966, 1). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. doi: 10.1145/365153.365168

Windiatmoko, Y., Hidayatullah, A. F., & Rahmadi, R. (2020, 9). Developing FB Chatbot Based on Deep Learning Using RASA Framework for University Enquiries. doi: 10.1088/1757-899X/1077/1/012060

Yang, W., Zeng, G., Tan, B., Ju, Z., Chakravorty, S., He, X., ... Xie, P. (2020, 5). On the Generation of Medical Dialogues for COVID-19.

Yasuda, K., Jun-ichi, A., & Fuketa, M. (2013). Development of an agent system for conversing with individuals with dementia..

Ye, W., & Li, Q. (2020, 11). Open Questions for Next Generation Chatbots. In *2020 ieee/acm symposium on edge computing (sec)* (pp. 346–351). IEEE. doi: 10.1109/SEC50012.2020.00050