UNIVERSITÄT PASSAU

Fakultät für Informatik und Mathematik

Building a domain specific dialog system on top of a lightweight semantic model

Dominik Düsel

Matriculation number: 64750

May 2018

MASTER THESIS

Chair of Digital Libraries and Web Information Systems

University of Passau

In partial fulfillment of the requirements for the degree of

Master of Science (M.Sc.) in Computer Science

Supervisor: Professor Dr. Siegfried Handschuh

Abstract

Human-machine communication is an important research area in nature language processing. In this work, we present a new approach of modelling human-machine dialogues and building dialogue systems with it.

Our approach is based on a lightweight semantic model which is domain-specifically adapted. We model a dialogue as a list of utterances, each of these utterances have one of the following four dialogue act types: *opinion statement*, *non-opinion statement*, *question* or *rest*. Our lightweight semantic model is centred around opinion statements about entities, therefore the approach is designed for domains with focus on evaluable entities. For processing the users utterances and generating responses, we combine a content based dialogue approach with a rule based approach.

In our approach we use multiple well-known concepts, including coreference resolution, named entity recognition, sentiment analysis, POS-tagging, open relation extraction, dialogue act classification and an AIML chatbot.

In this work, we will first give an overview of related work. Then we will explain how we use the concepts mentioned above. Afterwards, we will demonstrate how we adapted the approach to the football and product domains. Subsequently, we will evaluate our approach on an implementation of our football domain adaption. Finally, we will discuss the evaluation and conclude.

Acknowledgment

I would first like to thank my thesis supervisor Professor Dr. Siegfried Handschuh of the Chair of Digital Libraries and Web Information Systems at University of Passau for making it possible for me to write this thesis and helping me with all questions.

I would also like to thank Dr. Andre Freitas for helping me finding and shaping my topic, and for being very helpful whenever I was stuck.

Furthermore, I would like to thank all the people who helped me evaluating my system by investing their time testing it and giving me feedback.

Finally, I am very grateful to my family and to all my friends for their support and companionship during my years of study and the process of creating this thesis. Thank you.

Dominik Düsel

Table of Contents

List of Figures

List of Tables

Table of Code Snippets

Chapter 1

Introduction

1.1 Motivation

Being able to communicate with a computer by natural language is a long-cherished goal. There are plenty of books and movies where robots are perfectly talking like humans. Despite this still being fiction, there is a lot of research going on in the field of Nature Language Processing. Dialogue systems are a part of this field. As a dialogue system, we understand computer systems which are able to have conversations with humans. These conversations can happen in written, but also in spoken form. Dialogue systems generally aim to imitate a human-to-human way of communication while serving various purposes.

There already are small talk based systems, which are chatting with their users without having any purpose but the conversation itself and thereby entertaining the user. They are normally rule-based and respond purely based on this static, predefined rules, which makes them unable to provide any content or information outside of this rules, like latest news or the like.

In contrast, content based systems attempt to answer questions the user asks or provide information the user might find valuable. This ofttimes also includes dynamic content provided through the internet.

Our approach tries to combine aspects of both of these systems and both of these purposes. The basic idea of our approach is described in the next subsection.

1.2 Objectives

We want to create a model for systems, which are able to lead and sustain a dialogue with its user, talking in a specific domain. The system creates an added value for the user by providing interesting information while being able to make small talk with the user. Thus, our goal is it to create a mix of a small talk based system and a content based system. While reaching this goals, we still want to keep the model lightweight, to make adaptions to specific domains and respective implementations comparatively easy.

We also evaluate how good such an approach works. We will do this in the domain of football, based on user feedback and a statistical analysis of our used concepts.

1.3 Approach

We use a lightweight semantic model to support our dialogue system. The idea is to generate the output text based on three factors. First, we use reliable numerical data. The evaluation and influence of this data strongly depends on the domain. Second, we extract information from textual data. The details of this are also domain-dependent, but the information extraction is always based on the same concepts. Theses concepts include the following:

- Coreference Resolution
- Named Entity Recognition
- Sentiment Analysis
- Part-of-speech Tagging
- Open Relation Extraction

Third, the output obviously depends on the input text of the user and the earlier conversation. We classify the dialogue act of the input text, and also analyse it with the same concepts as our textual data. The information extraction here is also independent of the domain.

The information gained from the numerical and textual data is parsed in our data model, which is focused on entities and their properties. This data then can be used to create opinion statements and other utterances. Which kind of utterance is finally generated as an output depends mainly on the dialogue act of the user input.

1.4 Structure of the Report

In this work, we first will describe related work in this area and show which similarities our approach has to related systems, but also how our approach differs from them. This is done in chapter 2. In chapter 3 we then will explain our approach in more detail, elaborating on the language theory behind it and on the used concepts. Afterwards, in chapter 4, we demonstrate how our approach can be adapted to a specific domain using football and products as example domains. The football domain centres on talking about football players and teams, while we understand talking about which products you like or don't like as the products domain. Subsequently we will evaluate the approach on the football domain, we describe this in chapter 5. Afterwards, in chapter 6, we discuss our evaluation results and finally we conclude in chapter 7.

Chapter 2

Related Work

In the last years, dialogue systems and chatbots became very popular, especially in online marketing [Sab17]. As chatbots we understand programs, which interact with the user through natural language, but - in contrast to dialogue systems - don't control their input and output channel. A common example for this are chatbots which are integrated in the Facebook messenger.
Apart from online marketing, dialogue system and chatbots are used for entertaining, for informing and assisting people and a lot more.
In this chapter we want to give an overview about existing dialogue systems and chatbots. We distinguish them by four categories: online shop assistants, question answering systems, small talk systems and virtual assistants. We will give a brief introduction to all of these categories with typical examples and an explanation on how the different systems work. Furthermore we will illustrate similarities and differences to our system.

2.1 Online Shop Assistants

One big application area of chatbots are online shop assistants. Online shop assistants are programs who assist customers when browsing an online shop. Their ultimate goal is to help the customer in the same way a salesperson would in an offline shop.
Online shop assistants extend from answering questions the customer has to proposing special products. They can replace product search forms by asking the customer what he is looking for in a conversational way.

These chatbots have multiple advantages, for companies as well as for customers. They can easily be used on different computing devices like mobile phones and computers, without having a big development effort since they normally are integrated in some messenger which handles the user interface. Furthermore, they are available at any day, at any time while having low operating costs [Kha17], especially compared to a customer service made up by human employees. Obviously, online shop assistants also have their limitations and can't fully replace a human sales assistant in every case, but they are already used for simple tasks and further improvement is going on.

An example for an online shop assistant is "Mildred", which was developed by Lufthansa and published in 2016. Mildred is a chatbot integrated in the Facebook messenger, which searches for the cheapest flight between to airports within a timeframe given by the user [Ebe16]. It is also possible to limit the results to a specific flight class (e.g. *business class*). When the search result is shown, the user has the option to be directly forwarded to the system where he can book this flight.

The main goal of online shop assistants is to inform potential customers about products and encourage them to buy them. This makes them distinct from our system. While our system also informs its user in a certain domain, it does this without any ulterior motive. Besides, online shop assistants are completely customised to their domain, while our approach is more general and is about using similarities of multiple domains to simplify the development of domain specific systems.

2.2 Question Answering Systems

Also related are question answering systems. As the name suggests, their goal is to answer questions the user poses. They create a value for their users by providing requested information in an, for the users, easy way - through natural language. Thereby, the questions can be about anything and in a lot of different formats.

Question answering systems often follow the following strategy. First, they classify questions by the type of answer the question expects. For example, a question starting with "who" normally expects a person or organisation as answer. Then, documents containing the keywords of the question are retrieved with standard retrieval algorithms. This step works quite similar to web search engines. Finally, these documents are parsed to find a sufficient answer, taking the type of question into account.[V+99]

The probably most popular question answering system nowadays is Google. The Google web search engine, which started as a simple text retrieval system, is now also able to answer plenty of questions. This can range from historical questions ("Who was the first president of the United States?") to question about recent events ("What was the score of the game Bayern vs Hamburg yesterday?") and even about the future ("How is the weather going to be tomorrow?"). For all these questions, Google does not only provide a list of documents somewhere containing the answer, but also extracts a direct answer to the posed question.

Another notable example is "Watson". Watson is a question answering system developed by IBM to compete at the quiz show "Jeopardy!". In Jeopardy!, the contestants are asked challenges in natural language and have to answer quite fast. The challenges are phrased as statements and the contestants need to find the correct question this statement answers. Thereby, the questions are often formulated ambiguously and vaguely and come from various areas, including history, science, culture and languages. So to be successful, a contestant must understand complex hints, needs a very broad knowledge, and has to process questions and retrieve answers in very short time. IBM met all these challenges and managed to have Watson win in Jeopardy! against many of the previous champions. [FLB+13]

While question answering systems are purely about answering questions the user poses to them, our approach doesn't focus on answering questions, but on providing information the users didn't ask for, but has an interest in. We want to identify and provide relevant events or information, before the user asks for it. Furthermore, question answering systems often don't take the dialogue history into account and treat every question independently of the others, while our approach models whole conversations and might output different responses to the same input in the course of the conversation.

2.3 Small Talk Systems

Our third category of dialogue systems are small talk systems. As small talk systems we understand programs, which have the sole purpose of making conversation, mostly small talk, with the user and thereby entertain him. They can answer to greetings and other everyday-phrases like "How are you?", "What's your name?", "Nice to meet you." and a lot more. Normally, they have a lot of precast utterances which they output based on rules for parsing the input.
An advanced example for such a system is A.L.I.C.E., which stands for Artificial Linguistic Internet Computer Entity. ALICE works through pattern matching, and this patterns are described in AIML (Artificial Intelligence Markup Language). AIML is a XML-compliant language which was developed for creating chat applications [SA07]. We will explain AIML more detailed in section 3.7.
Our approach is closest to the category of small talk systems, since it has the same goal like small talk systems: to entertain the user. We also include an adapted version of ALICE as a fall-back level if we recognise the input as being domain independent small talk or if we can't handle the input otherwise. So, systems based on our approach will react like a small talk system in a lot of cases, if the user talks about things outside of the given domain. But, systems based on our approach will shift the conversation topic away from small talk and back to the given domain, and in this domain they will respond very differently from small talk systems: they will provide content and give information dependent on current events.

2.4 Virtual Assistants

A recent but fast growing research area are virtual assistants. Virtual assistants are programs which assist their users with everyday tasks by natural language commands. They can e.g. manage appointments, play some music or set up phone calls. Virtual assistants are often integrated in smart-phones or other mobile devices, to make themselves available at all time.
There are multiple well-known virtual assistants like Apple's Siri or Amazon's Alexa. They have a very broad range of things they can assist their users with and aim to be an all in one solution. They are also able to make small talk and answer questions, so they include the two kinds of systems we explained previously.

In opposite to broad ranged assistants like Apple's Siri or Amazon's Alexa, there also exist specialised virtual assistants, e.g. "Amy" which is developed by a startup called x.ai. Amy just focuses on one thing: scheduling meetings. Right now, Amy works through email. If you want Amy to schedule a meeting with someone, you just have to put Amy into *cc:* of your initial meeting request. Amy then takes over, parses your request and arranges a time with your meeting partner by writing him or her emails. If a suitable time is found, Amy puts the meeting in your calendar and sends a notification. [Elg16]
Virtual assistants differ from our approach by their purpose. They support their users doing certain things, while our approach focuses only on entertaining the user and informing him about a given domain. While broad ranged virtual assistants are also able to do everything what systems based on our approach can do, they don't specialise on it. Furthermore, the broad ranged virtual assistants normally have much more detailed models and therefore need a larger development effort.

Chapter 3

General Approach

In this chapter we describe in detail how our approach works. Therefore, we first introduce the linguistics which are important for our approach, in section 3.1. Subsequently, we describe what kind of dialogues our approach focuses on and what assumptions we make about the domains the approach works for. This is done in section 3.2. Then we will describe our formal model for dialogues in section 3.3. Afterwards, in section 3.4, we will illustrate our conceptual model and explain the overall strategy of our approach.
After that, we explain how the different concepts work and why and how we use them. We start with our information extraction in section 3.5, followed by the dialogue act classification in section 3.6. Then, in section 3.7 we go into detail about AIML (Artificial Intelligence Markup Language) and how to make a small talk chatbot using it. Subsequently, in section 3.8, we explain our concept of precast questions to sustain the dialogue and to guide it to the desired topic. Finally, in section 3.9, we describe limitations and problems our approach has, and how some of them can be addressed.

3.1 Dialogue Linguistics

Before we can develop a dialogue system, we have to make clear, what a dialogue is. A dialogue is a conversation between at least two parties, but we will only consider two party dialogues here, since our approach only focuses on them. A dialogue now describes these two parties exchanging utterances, where we consider every single sentence as its one utterance. These utterances can have many different types, called dialogue acts. A dialogue act describes the intended verbal action of an utterance, thus it is on the level of illocutionary acts, which means it is concerned with the meaning of the utterances [SRC+06].

The dialogue act can be classified by a variable amount of classes, dependent of the domain and usage. While [LC06] uses only 12 dialogue act classes, [SRC+06] describes 42 of them. The later also presents frequencies, how often utterances of each dialogue act occur. This is based on utterances of the switchboard corpus, which is a collection of human-to-human telephone conversations. The corpus also includes hand labelled dialogue acts for all its utterances. We will perform our analyses on this corpus and therefore use the 42 dialogue acts [SRC+06] describes. The switchboard corpus contains over 200,000 utterances respectively sentences from 1155 conversations [JBC+98].

Two of the most common dialogue acts are *statement-non-opinion* and *statement-opinion*. Utterances with these labels make up about half of the switchboard corpus. Furthermore, 25% of the corpus is made of *backchannel* utterances and *abandoned* utterances, which are relatively easy to process and often can be ignored in written conversations. Therefore, our main focus will be on statements. They will be analysed deeper in the next two sections.

3.1.1 Non-Opinion Statements

As non-opinion statements we denote all statements which don't describe an opinion about something. Non-opinion statements can e.g. be descriptions, narrations or explanations. We want to analyse non-opinion statements and how humans respond to them, to see how our system should react. We therefore took the conversations of the switchboard corpus and counted which dialogue acts the responses to non-opinion statements had. We always took the first up to three utterances of the responding person into account. The reason for this is, that the first response very often just is a backchannel or an agreement, but we are also interested in what dialogue acts come after that, since we want to be able to generate multi-sentence responses and not just a backchannel or agreement.

For non-opinion statements, the most common responses are, by a high margin, backchannels. They make up about 43% of the responses. The biggest group of the rest are other non-opinion statements (around 15%), opinion statements and appreciations (both around 5.7%).

Taking this data into account, we assume that the most common response to a non-opinion statement is a backchannel or appreciation, followed by some kind of statement.

3.1.2 Opinion Statements

When we do the same analysis for opinion statements, we again have backchannels as the most common response, in this case they make up around 30% of the responses. Notable here is the amount of agreements, which is the second most common response (around 19%). Also notable is the fact, that opinion statements (around 13%) and non-opinion statements (around 14%) as responses are similarly common, despite non-opinion statements being about thrice as frequent in the overall corpus.
Altogether, we assume that most responses to opinion statements consist of a backchannel or agreement, followed by some kind of statement, like it was with non-opinion statements. But for opinion statements, the responses comparably have a much higher chance of also containing an opinion statement. This behaviour was also observed by [SRC+06].
We therefore want to go into more detail about opinion statements and how we find other opinion statement to respond to it. For that purpose, we introduce three primary facets of opinion statements:

- target
- aspect
- context

The *target* is the named entity the statement is about. Targets can be persons, organisations, things or something else. These possibilities depend on the used domain. In the domain of football, this could e.g. be a certain player or a team. E.g. in the statement "The FC Bayern played a bad game yesterday against Paris." the target would be "The FC Bayern".
The *aspect* describes the aspect of the target, which is described in the statement. This obviously depends on what the target is and therefore also on the domain. If we again take the example of football, this could e.g. be the performance, fairness or number of goals of the target. If we again take our example sentence "The FC Bayern played a bad game yesterday against Paris.", we can observe "played" as the aspect, which means the statement describes the overall performance.

The *context* is the background in which the target and aspect are described in the statement. This generally is some kind of timeframe, or a place. In our football example, this might be a specific game, a whole season or anything alike. In our previous example sentence, the context would be "the game Bayern against Paris yesterday".
A fourth facet, which needs to be treated differently, is the *valuation* of the statement. This is a rating of the aspect of a target, considering the context. The valuation describes the state of the aspect, e.g. good or bad. For our example sentence "The FC Bayern played a bad game yesterday against Paris." this valuation would be "bad".
The valuation often is subjective, dependent of the aspect, and therefore provides a basis for discussion. When a person doesn't agree with the valuation of its dialogue partner, he generally responses with his own valuation and possibly with an explanation why he thinks differently.
When a person agrees with the valuation, he normally expresses this agreement and then can follow this up with some other opinion statement as a response. We therefore analysed multiple human to human dialogues from the domain of football, and made the following observation: when taking the first three facets into account, responses to an opinion statement mostly have exactly one of the three facets changed. As a consequence, the responses are still strongly related to the original statement and are therefore recognised as a valid response, but they still don't say exactly the same, since one facet has changed.
We want to illustrate this on an example, for that we will again use the domain of football. If person A makes the opinion statement "The FC Bayern played a bad game yesterday against Paris.", person B could respond with the following opinion statements:

- "But Thomas Müller played quite good yesterday." (change of target)
- "They also played quite unfair." (change of aspect)
- "They played quite bad in general this season." (change of context)

3.2 Domain Requirements

In this section, we want to describe what assumptions we make of the domains our approach works for. Therefore, we first need to clarify what dialogues our approach focuses on. Our ap-

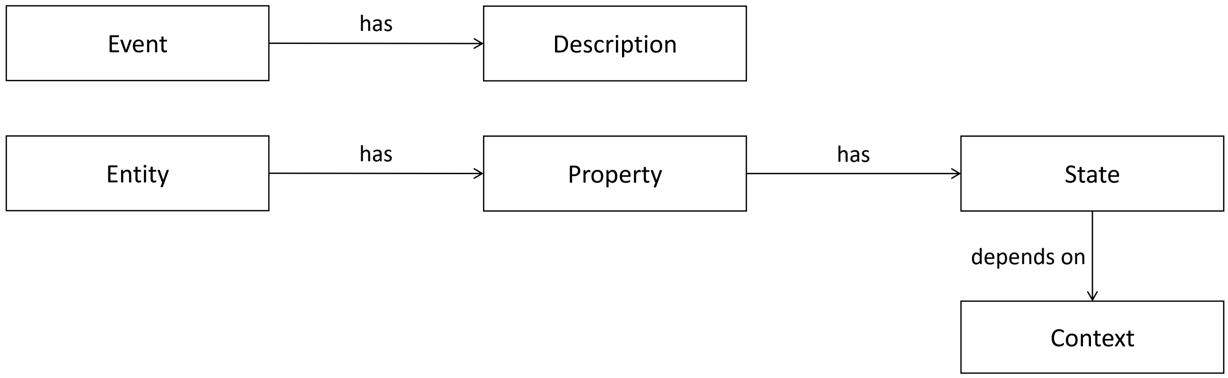

Figure 3.1: Data model for our approach

proach is about human-machine communication in two-party dialogues. The task of the machine in our approach is not only reacting on user input, but also sustaining the dialogue by bringing up new topics within the domain by itself and trying to re-engage if the user doesn't respond anymore.

Considering the dialogue acts we described before, we want our conversations to be centred around opinion statements. Therefore, we developed a data model which supports understanding and creating opinion statements as we modelled them before - with target, aspect, context and valuation. This data model is displayed in Figure 3.1.

The domain should be focused on entities, these can then be used as target of opinion statements. The entities should have some subjective properties, so one can express his opinion about these properties and the entities in general. The properties correspond to the aspects described in the previous section.

Furthermore, the properties have a state, e.g. good or bad. This state corresponds to the valuation of an opinion statement and can be subjective. The state is also able to change, so the context is relevant, when describing the state of a property. The context in our data model corresponds to the context of an opinion statement, so it normally describes a timeframe or something similar.

Our data model also includes events, which can have a description. This is our approach to modelling non-opinion statements, since they are often used for telling stories or describing something. Thus, we will use the events to create non-opinion statements. An event can be anything in the domain which might be of interest for the user, and is therefore strongly domain dependent. Examples for an event in the football domain are the scoring of a goal or a red card. While this data model is very shallow and might not be able to represent lots of details, it is well adjusted for our focus on opinion statements and easy to adapt to our target domains.

3.3 Dialogue Model

Now we want to describe our formal representation of dialogues and how we handle utterances of the different dialogue acts. As a dialog, we understand an exchange of multiple utterances by our two dialogue participants, which are the user and our system. Each of these utterances represents one sentence, so when a participant utters more than one sentence straight, they are treated as multiple utterances.

We distinguish four types of utterances, which are based on the dialogue act: opinion statements, non-opinion statements, questions and rest which represents every statement not fitting to one of the previous three types. Opinion statements have, like described earlier, a target, an aspect, a context and a valuation with the corresponding types in our data model. Non-opinion statements have an event they describe. While questions in general can be quite varying, we only model questions which are aiming for an opinion statement as answer. In the domain of football, such a question would be: "How did the FC Bayern play yesterday?". We don't model other questions, since we want to keep our model lightweight. Other questions are handled like utterances of our dialogue act type rest: by an AIML chatbot. Therefore, we don't need to model them any further.

Our approach now is, to answer opinion statements with an agree or reject, depending on whether the available data supports the opinion or not. We will describe later, how we get this data. After the agree or reject, we want to follow up with an opinion statement by our own. If we agreed, we want to respond with a corresponding opinion statement, which means that exactly one of the three facets target, aspect and context has changed. Which facet is changed is decided randomly. If we rejected the users opinion, we want to state the opinion with the same target, aspect and context, but different valuation and, if possible, an explanation based on the data why we rejected the opinion.

To non-opinion statements we answer with an acknowledgement, followed by a non-opinion

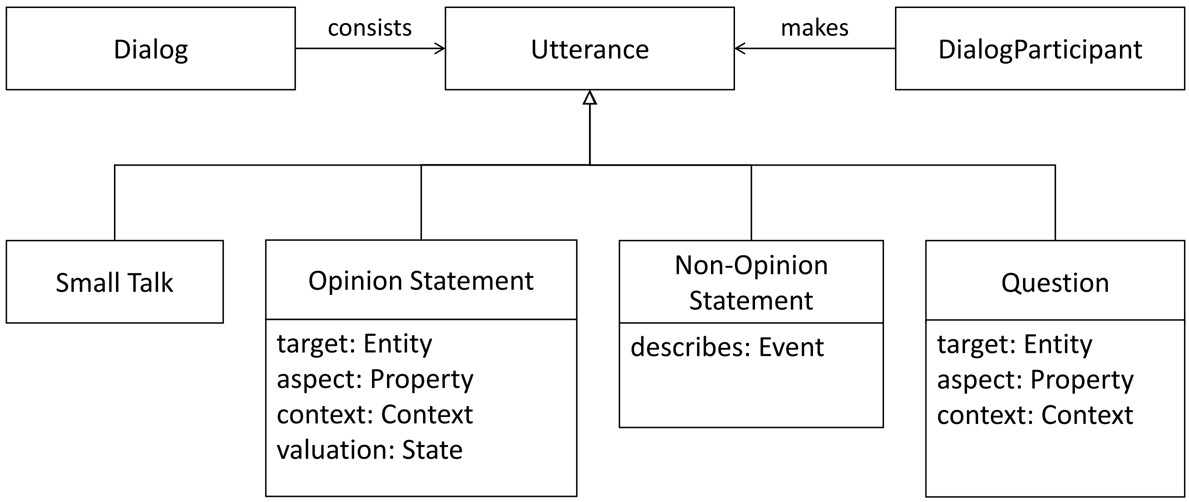

Figure 3.2: Dialogue model for our approach

statement describing a related event, if we are able to find one. If not, either a precast question will be asked, or the statement is handled by the AIML chatbot.

If the user poses a question we can model with target, aspect and optionally a context, we respond to it with the corresponding opinion statement. Any other questions can't be answered with our lightweight model and are therefore forwarded to the AIML chatbot.

Any other utterances, like e.g. greetings or thanking, are also answered by the AIML chatbot. But additionally, we sometimes will respond with a precast question to lead the topic back to the domain.

3.4 Conceptual Model

In this section, we want to give a quick overview on how our approach works. A graphical demonstration of our conceptual model is shown in Figure 3.3.

Our conceptual model differentiates three types of input: structured data, unstructured (textual) data and the utterances of the user. The structured and unstructured data completely depend on the domain the system is adapted to.

The structured data is data in the domain, which is exact and reliable, and can be parsed comparatively easy due to its structured form. Examples for this in the football domain are CSV-files

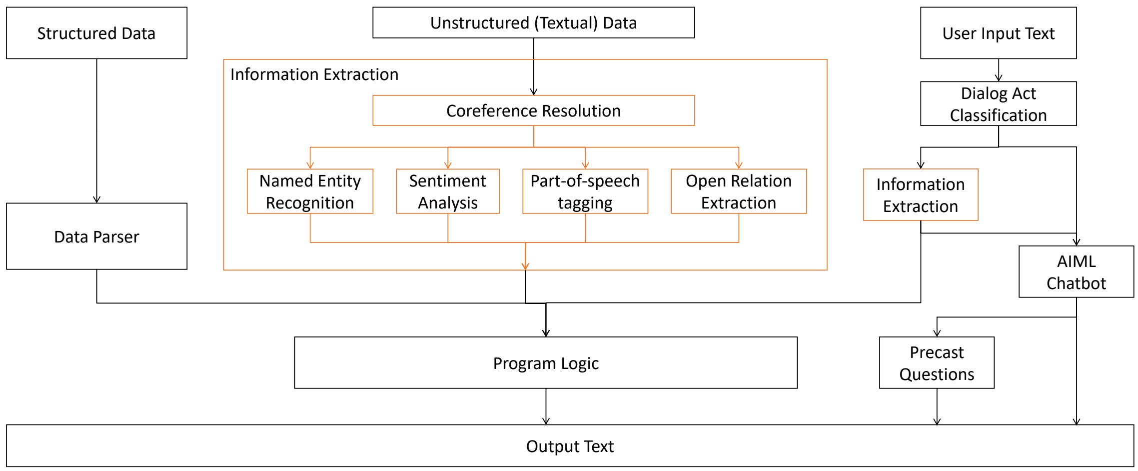

Figure 3.3: Conceptual model of our system

with game results or statistics to shots on target, cards, etc.

In contrast, getting information from the unstructured textual data (e.g. a newspaper article), is difficult and the parsed information can be unreliable. An example for this in the football domain is parsing an article about a game and extracting the scorers out of it. Alternatively, we might want to get the performance of certain players based on the same article. To get these various kinds of information, we use multiple well-known concepts of NLP. The whole process is noted as *Information Extraction* in Figure 3.3 and will be described more detailed in section 3.5.

The *User Input Text* in the conceptual model represents an utterance the user made in the conversation, to which our system should react to. Therefore, its dialogue act is classified. How this works is described in section 3.6. Our classifier only classifies the three dialogue acts statement, question and rest.

The dialogue act is then used to determine the further steps. Every *statement* will then be run through our information extraction, where we, among other things, detect whether it is an opinion or non-opinion statement. For opinion statements this also extracts the target, aspect, context and valuation of the statement, to then be able to generate a valid opinion statement as response.

For non-opinion statements, the information extraction detects whether the user talked about an event we model. If we find one, our program logic takes over and answers with an appreciation and, if possible, with a related event which hasn't been spoken about yet. If we don't find an event in the non-opinion statement, the AIML chatbot will handle the utterance.

For utterances of the dialogue act *question*, we follow a similar process: we apply our information extraction and then check whether we found target and aspect of a potential opinion statement in the question. If yes, our program logic responds with an according opinion statement using a default context, if no context could be extracted from the question. This approach is supposed to handle questions like e.g. in the domain of football: "Do you think the FC Bayern played well yesterday?". If no target or aspect can be detected, the question will be handled by the AIML chatbot.

Every utterance where we classified the dialogue act *rest* will also be handled by the AIML chatbot. How AIML works will be explained in section 3.7. But there is a general restriction on utterances by AIML: whenever three utterances in a row were forwarded to the chatbot, a *Precast Question* will be added to the response of the chatbot. We will talk more detailed about precast questions in section 3.8.

3.5 Information Extraction

In this section, we explain our approach on extracting information from textual data, e.g. newspaper articles. We want to extract all kind of information which can be stored in our data model, which basically means two types of information: all kind of events and everything related to relevant entities which helps to get a valuation for some of its aspects. In this section, we will describe the concepts we use for that and how we use them.

3.5.1 Coreference Resolution

Coreference resolution is an important research area in NLP. It describes the task of linking reference words like "he" or "his" to the entities they refer to [Ela05]. Since we will treat every sentence separately in the other parts of the information extraction, this is necessary to link information from sentences where entities are only referenced to the correct entities. Also, as illustrated in Figure 3.3, we obviously need to perform the coreference resolution before we apply our other information extraction concepts.

For this coreference resolution, we use Graphene. Graphene is an information extraction pipeline which not only points out the references, but also is able to replace reference words with their referenced entity [gra]. Thus, the resulting text is ideally reference-free and information in the text can be linked to the correct entity.

3.5.2 Named Entity Recognition

Named Entity Recognition (NER) is also an important area in NLP. It is concerned with recognising sequences of words in a text which represent certain things, e.g. persons or organisations [TKSDM03].

Recognising this sequences and referring them to real-world entities is also an important part of our approach. We use NER to extract these sequences and link them to information we get with our other concepts. In the domain specific implementation we then link the named entity word sequences to entities in our domain. Since this is done domain dependent, we can use knowledge about the domain and access domain specific data (e.g. entity lists from the structured data) for this step.

We use the Named Entity Recogniser developed by the Stanford Natural Language Processing Group. It uses linear chain Conditional Random Field sequence models which are trained on a mixture of named entity corpora. [Gro, FGM05]

3.5.3 Sentiment Analysis

The goal of sentiment analysis is to extract people's opinion or emotions toward certain entities from texts. It is often used to analyse product reviews and gather information on what the reviewers liked about the product and what not. [LZ12]

Sentiment analysis is also exactly what we need to get the sentiment in a text towards relevant entities and their properties. The sentiment is used to create the valuation about a certain aspect, thus to determine whether the opinion in a statement generated by our system should be positive or negative. We also use sentiment analysis to detect the valuation of opinion statements of the user, so we can agree to or reject them based on our data.

For doing the sentiment analysis, we again rely on the Stanford CoreNLP. Their sentiment analysis is based on a Recursive Neural Tensor Network, which is trained on the Stanford Sentiment

Treebank. The Stanford Sentiment Treebank labels sentiments not only for whole sentences, but also for every single phrase of the sentence. Therefore, the neural network also can be trained to be able to classify the sentiment of phrases. This has big advantages in sentences where contrasting opinions are displayed. [SPW+13]

3.5.4 Part-of-speech Tagging

Another basic tool in NLP is part-of-speech (POS) tagging. It is concerned with parsing texts and marking every word or token with a POS tag. POS tags can distinguish between noun, verb, adjective, etc., but they also can be more fine-grained like noun-plural, verb in 3rd person present and so on.

We use a fine grained POS tagger in our approach to do advanced keyword comparison. As advanced keyword comparison we understand using value pairs consisting of a keyword and a POS tag. To get a match, the keyword has to appear with the corresponding POS tag in a sentence. While this, compared to just comparing keywords, decreases the performance and is more work in creating the list of key-pairs, it increases the accuracy and partly avoids problems with ambiguous words. We use this advanced keyword comparison in our approach to recognise about which property of an entity is talked, so we can link a sentiment to the correct property. Furthermore, it is used to detect events in the text. The list of keyword, POS tag pairs and some implementation details are thereby domain dependent.

As with the previous two concepts, we use the part-of-speech tagger of the Stanford Natural Language Processing Group. The tagger is an implementation of the model described in [TKMS03, TM00].

3.5.5 Open Relation Extraction

In many sentences, only using named entity recognition to determine who the target is, is not enough. For example, if there are multiple entities in a sentence or phrase with a bad sentiment, we would not know about which of these entities the bad sentiment is. Therefore, we use the open relation extraction of Graphene [gra]. Graphene is able to identify subject, object and predicate of a sentence.

This is useful for solving the problem with multiple entities. By being able to identify the subject, we can then always take the subject entity if there is one. However, if the multiple entities are all objects, we still are not able to pick the target.

3.6 Dialogue Act Classification

Classifying the dialogue act of the utterances of the user is a central task in our approach since the later steps depend on it. E.g., greeting or thanking phrases are handled by the AIML chatbot, while statements or questions are answered with a different approach. Since our approach focuses on opinion statements, we obviously would like to be able to distinguish them from non-opinion statements.
To do this, we examined the following classifiers:

- One-vs.-rest classification with neural networks (NN)
- One-vs.-rest classification with support vector machines (SVM)
- One-vs.-one classification with NNs
- One-vs.-one classification with SVMs

We trained all the classifiers on a part of the switchboard corpus, and evaluated them on another part. We divided the data in four classes: opinion statements, non-opinion statements, questions and a class for everything else (*rest*). This corresponds to the model introduced in section 3.3. Our testing set consists 44,323 utterances in total, 23,767 utterances of which are in the *rest* class. Therefore, the accuracy baseline is about 0.54, since even a classifier always predicting *rest* could reach it.
An overview of how the classifiers performed, is given in Table 3.1. We always used the macro average of all classes as result. Our implementation of the classifiers is based on the code of [Moh].
We figured, that the classification strategy doesn't have a big impact on this metrics, but the one-vs.-rest classifying strategy is faster to train, since less classifiers have to be trained.
Overall, the one-vs.-rest neural network classification performs and beats the baseline by nearly 0.2. So we chose to use this classification for the rest of our experiments. But, as shown in

Table 3.1: Performance of different classifiers and classification strategies

classifier	precision	recall	F_1-score
One-vs.-rest NN	0.73	0.73	0.73
One-vs.-rest SVM	0.72	0.68	0.64
One-vs.-one NN	0.72	0.72	0.72
One-vs.-one SVM	0.72	0.68	0.64

Table 3.2: Performance of one-vs.-rest neural network classification on four classes

class	precision	recall	F_1-score
non-opinion statement	0.68	0.66	0.67
opinion statement	0.42	0.43	0.43
question	0.65	0.59	0.62
rest	0.83	0.85	0.84
macro average	0.73	0.73	0.73

Table 3.2, the precision and recall for the opinion-statement class is below 0.45, which is far to low to rely on it, especially considering that our approach is focused on them. We therefore use the dialogue act class *statement* as a combination of both these classes.

When training a one-vs.-rest neural network classifier with only the three classes *statement*, *question* and *rest*, it results in a macro averaged F_1-score of 0.82, as well as an accuracy of 0.82. The exact values for each of the classes is displayed in Table 3.3. This classification accuracy still leaves a lot of room for improvement, but it is the best we can get with the current state-of-art. When further improvments will be made in the area of dialogue act classification, a different approach might be better.

Obviously, we still want to distinguish between opinion statements and non-opinion statements, since we want to handle them differently. We want to detect target, aspect, context and valuation of every opinion statement, but we can also use this detection to find out the fine grained dialogue act of the statement. We just run our target and aspect detection on every statement

Table 3.3: Performance of one-vs.-rest neural network classification on three classes

class	precision	recall	F_1-score
question	0.68	0.56	0.61
statement	0.82	0.80	0.81
rest	0.83	0.86	0.84
macro average	0.82	0.82	0.82

and if we find both of them, we classify it as an opinion statement. If we aren't able to detect a target or an aspect, we assume it is a non-opinion statement and handle it accordingly.
Opinion statements where we are not able to identify target and aspect could not have been handled with our normal opinion statement exchange anyway, so this procedure filters them out as well. We don't treat the context as so important here, since the context is often said implicit and therefore hard to detect. If we can't detect the context, we will just assume a default context for utterances of the user and then change it in our response opinion statement, since we want to change one facet anyway.
We also don't check the statements for a valuation here, because non-opinion statements also have a sentiment and therefore we can find a valuation in every statement, thus it is irrelevant when classifying the dialogue act.

3.7 Small Talk with AIML

As mentioned in section 2.3, AIML (Artificial Intelligence Markup Language) is an XML-compliant language, which was developed to create chatbots based on dialogue patterns.
AIML consists of a set of so called *categories*. Each category describes one conversation rule. It therefore consists of a *pattern*, matching against the user input and a *template*, which is used to generate the output. A category also can have an optional context, represented by the <topic> and <that> tags [SA07, Wal].
The following example is a simple category, which matches only the exact input text "10 Dollars" and, on a match, produces the output "Wow, that is cheap.".

```
<category>
    <pattern>10 DOLLARS</pattern>
    <template>Wow, that is cheap.</template>
</category>
```

AIML patterns can contain any letters, numbers, spaces and the two wildcard symbols _ and *, which function on words. So _ represents a single word, while * can be any number of words, including zero. Anything else, like punctuation, is not allowed in AIML patterns. However, in some AIML libraries including the one we use, punctuation can be transferred to expressions like "QUESTION MARK", and then be used in patterns.

AIML also supports recursion, that means categories can pass on generating the output to other categories. This works with the <srai> tag:

```
<category>
    <pattern>DO YOU KNOW WHO * IS</pattern>
    <template><srai>WHO IS <star/></srai></template>
</category>
```

This category passes any input of the form "Do you know who X is" on to the category which matches "Wo is X". The <star/> tag is replaced by whatever was parsed as the * in the pattern. With this functionality, you also can split the input by passing it on to multiple <srai> tags:

```
<category>
    <pattern>YES*</pattern>
    <template>
        <srai>YES</srai>
        <srai><star/></srai>
    <template>
</category>
```

Another concept in AIML is the context, which is implemented with the <that> tag. The <that> tag is a second pattern which must be matched, but not by the user input. Instead, it has to match the last sentence of the previous utterance of the bot. This is useful to have conversations about one topic going on for more than just one utterance. E.g. the robot could ask "Do you like movies?" and, assuming the user answered "yes", respond with "What is your favourite movie?". This would be done with the following category:

```
<category>
    <pattern>YES</pattern>
    <that>DO YOU LIKE MOVIES</that>
    <template>What is your favourite movie?</template>
</category>
```

There are a few more features of AIML, which we don't want to explain here, but it still stays quite simple and therefore easy to use. In our system, we use an lightly adapted version of A.L.I.C.E., which is the bot AIML originally was developed for. A.L.I.C.E. consists of about 41,000 categories, which we nearly left untouched, since they are already working well. We mainly adapted bot personality information like its name, age and so on.

3.8 Precast Questions

If the user keeps making small talk, we want to be able to lead the topic of the conversation back to our domain. To do this, we use the concept of precast questions. Every domain specific implementation of our approach should include a list of questions our system can ask. The concrete form of this questions is domain dependent, but they generally should be aimed on entities or events, since this are the things we model and want to focus our conversations on. An example for a precast question from the domain of football is "Who is your favourite player?" This clearly aims for an entity as answer, and our system than can follow this up with an opinion statement about this player.

There are two situations in which we ask a precast question. First, whenever there are three straight utterances of the user which are handled by AIML, we append a precast question to the last response of the AIML chatbot. This ensures that the topic is led back to our domain. Second, when the conversation already is about the target domain, but we don't know what else to respond, we also respond with a precast question. This appears when the user talks about an event, but we don't have any related data which wasn't already spoken of. We then just appreciate the non-opinion statement of the user and ask one of the precast questions.

3.9 Limitations and Problems of the Approach

Our approach, and systems implemented following it, obviously is not perfect and has some limitations and problems. We will describe them in this section and explain why they exist and how to deal with them.

3.9.1 Content-based Questions

One major limitation of our approach is not being able to handle content based questions aiming for anything else than an opinion statement. While our approach is content based, it isn't focused on answering questions, but providing the content when it is related to an input utterance.

Obviously the domain dependent program logic can fix this issue and implement a question-answering system which would be able to answer more types of questions. But implementing this generally has a high effort, which contradicts our approach of having a general system with only limited necessary adaptions for each specific domain.

3.9.2 Handling of Non-Opinion Statements

Our model of non-opinion statements is very limited: we only model them with the event class, and entities occuring in this events. While the term event is quite vague and therefore leaves room for more detailed models of them in the domain specific adaptions, this also is connected with some adaption effort.

Thus, for non-opinion statements, the goal of our approach, having a domain independent, lightweight model, which makes it easy to create a well working domain specific adaption, is by far not reached to the same extend as we reach it for opinion statements.

3.9.3 Mixture of AIML and non-AIML Utterances

Our approach includes the possibility of jumping back and forth between a conversation lead by an AIML chatbot and our system after any number of utterances. Since the AIML chatbot as well as our system both have rules which depend on what we put out earlier, this can lead to erroneous behaviour.

While we can attend this matter in our main program logic, it can't be fixed for the AIML chatbot without a deep change to AIML or using a completely different approach. Thus, we will have to live with errors caused by this problem.

3.9.4 Inaccuracy of the Concepts

Some more problems come from the NLP concepts we use. While we use state-of-the-art approaches for our information extraction concepts, they still are prone to errors. All of the used concepts are current NLP research areas and therefore up to improvement. Thus, the information gained with these concepts can be erroneous. But we still need to rely on their output, since we have no real other choice, so errors in these concepts can lead to errors in our system and therefore to odd responses of our dialogue system.

Chapter 4

Adaption to specific domains

After explaining how our approach works in detail, we now want to demonstrate how to apply our approach to a domain. We therefore use two domains as example: football and products. What we understand as these domains will be explained in detail in the respective sections. By demonstrating the two adaptions to quite different domains we show the versatility of our approach.

For creating an adaption, we need to do several things. First, we need to specify the data model. This includes identifying relevant entities and assigning possible aspects, contexts and states to the entity classes. This also includes modelling event types in the domain. Second, we need to implement the domain specific part of the information extraction, mainly aspect and context recognition, which means setting up the POS-tag, keyword comparison. Third, we need to get input data of our domain and parse it into our data model. Thereby, the domain specific adapted information extraction can be used. And fourth, we need to set up precast questions for the domain.

All other parts of a domain specific system follow our general approach described in chapter 3, and don't need to be adapted.

We will first describe the adaption to the football domain in section 4.1 and then the adaption to the products domain in section 4.2.

4.1 Adaption to Football

Our first example domain is the domain of football. This domain, with the adaptions described here, will also be used for our evaluation, which is done in the next chapter.

4.1.1 About the Domain

The term "football domain" is quite vague and can mean a lot of different things. Therefore, we want to specify what we mean by this term and what kind of system we want to develop in this domain.

In general, we want to talk about (football) players and (football) teams and how they performed in some matches, with the focus being on recent matches. Thereby, we want to emulate the following scenario: two friends have both seen a certain football match and now talk about it. In this conversation also more general football topics can come up. Our system takes the role of one of the two friends, the user takes the other.

The dialogue is started by our system with some short small talk, but will then be lead to the respective match a little bit later. The rest of the dialogue should centre around that match, how the particular players performed, and events in it. These events can be scores or when a player gets a card. Apart from the match, they can also talk about how teams or players performed in the recent years.

4.1.2 Data Model

Our data model for the football domain is graphically displayed in Figure 4.1. We model three types of entities: *players*, *referees* and *teams*. For players and referees we only model the property *performance*, while teams have the properties *performance*, *fairness*, *wins*, *goals*, *chances*. They also have attributes like name etc., but we don't model them as properties, since the properties should relate to aspects in opinion statements. Attributes like a name are generally not subject to change and we don't want to express opinions about them.

We model three different states for all the properties: they can be *bad*, *average* or *good*, but the properties need to be set in context. The context describes a timeframe or some number of matches, in which the entity participated. In the opinion statement in the football domain, and

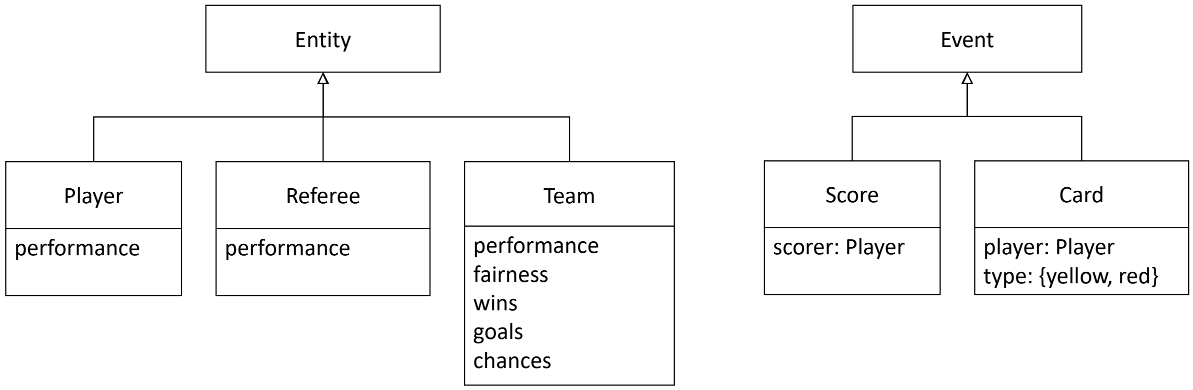

Figure 4.1: data model for the football domain

therefore also in our model, we distinguish between the following contexts: *match*, *matches*, *competition*, *season*, *seasons* and *all-time*.

Additionally to these entities, we model two types of events: scores and cards. For a score, we only model the scorer, additionally to the description every event has. For cards, we model the player who received it and the type of card, yellow or red.

4.1.3 Input Data

Since this adaption is mainly a proof of concept, we limit our data, and therefore the ability of our system to create content based opinion statements, on German football. We use the two types of data as input: structured and unstructured, textual data.

As structured data, we use statistics of all *Bundesliga* matches of the last ten years. The statistics include goals, cards, fouls, shots on target and a lot more, but only on a team basis. [Buc]

In our system, we use this data to calculate the states of team properties in a given context, to be able to create opinion statements which display reasonable opinions. Since the statistics are only on team-level, we are not able to use this for determining player performance.

For determining player and referee performances, we use news articles about some recent games as unstructured textual data. How we extract the information from this articles was described in section 3.5. The news article of the match the dialogue is centred around is also used for finding events and their description. We determine both our events, scores and cards, by keyword comparison and use the respective sentence or paragraph of the article as event description.

Another kind of structured data we use is the person data of DBpedia. We use this, to identify teams, players and referees. [Ass]

4.1.4 Precast Questions

As described in section 3.8, we use precast questions to sustain the dialog and centre it around the desired topic. We will explain some of the questions in our adaption to football domain now. These questions include asking the user about which player - or team - he likes the most. When the user answers, this can then be followed up by an opinion statement about the respective player or team.
Another precast questions is to ask the user for his opinion about a certain player, team or the referee. This aims for an opinion statement of the user, which then can be followed up by our system with a related opinion statement.
The system also asks about events we detected. E.g. for a card event, it can ask the user whether he thinks the card was deserved.

4.1.5 Similar Domains

The adaptions which were described in this section can also be applied to a lot of other sport domains, like Handball, Basketball or American Football. These domains can all have the same types of entities and similar events like the football domain.

4.2 Adaption to Products

To show the versatility of our approach, we will now describe adaptions to a very domain, the domain of products.

4.2.1 About the Domain

As the domain of products we determine talking about certain products, e.g. mobile phones. The scenario we emulate in this domain, is two persons speaking of which mobile phones or

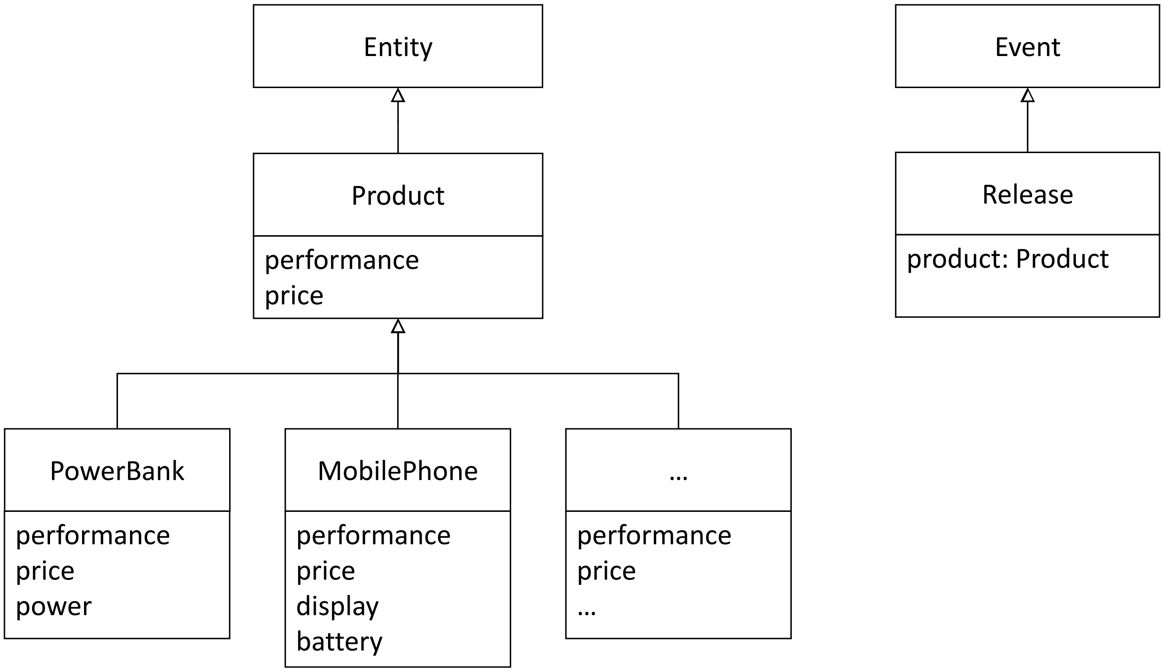

Figure 4.2: data model for the products domain

accessories they like, and which features of them. Also, they can talk about which new phones were just released.

4.2.2 Data Model

Our data model for the products domain is shown in Figure 4.2. In the this domain, we have one main entity type: products. We then model two sub-types, namely *mobile phones* and *power banks*, but there could be other products as well, if you want the system to be able to talk about more different products. All products have the two properties overall *performance* and *price*, while mobile phones also have the properties *display* and *battery* and power banks have the property *power*. All these properties also relate to respective aspects.
Like in the football domain, all products also have other attributes like a name, but we don't model them here since they can't be used as aspects in an opinion statement.
We model only two different contexts: *specific* and *general*. *Specific* means, that only one specific exemplar of the product is talked of, e.g. the smartphone the user possesses. The context *general* is used for talking about a product type, so e.g. all "Samsung Galaxy S4" smartphones.

In this domain, we only model one type of event: the release of new products. A release has, apart from the description every event has, the property *product*, which is a reference to the product which was released, and the property *date*, which is the date of the release.

4.2.3 Input Data

In the products domain, we use the product review data and product metadata of the amazon marketplace [McA, MTSVDH15]. They contain both structured and unstructured data. As structured data, we have the numerical star rating, which easily can be used for determining the overall performance of a product.

As unstructured data, we have the textual reviews of the products. From this textual reviews we extract which aspects of the products are good or bad with our information extraction described in section 3.5.

We use another source of structured data as well, namely the FonoApi. This API provides information about mobile devices, including the release date, which is important for our events. [Sad]

4.2.4 Precast Questions

As in the football domain, we also have precast questions in the products domain. In this case, we have three types of questions: what the favourite mobile phone of the user is, what a phone he owns and how he likes a certain product.

The first two questions can, again like some of the questions in the football domain, be followed up by an opinion statement from our system about the mobile phone the user named, or by an additional question on what the user likes about it. This follow up question aims, like the third question, for an opinion statement of the user.

Chapter 5

Evaluation

In this chapter, we want to evaluate our approach for building a domain specific dialogue system on top of a lightweight semantic model. Therefore we use the football domain with the adaptions described in section 4.1.

There are multiple different approaches to evaluate dialogue systems or chatbots. There are several automated approaches, e.g. the PARADISE framework [WLKA97]. But the automated approaches mainly focus on evaluating task-oriented dialogue systems. Since the main goal of our system is to entertain the user, and it therefore is not a task-oriented system, we exclusively use human judgement for our evaluation. Human judgement also still is the most common evaluation approach for dialogue systems or chatbots. [LLS+16, JM17]

In our evaluation, we pursue two central goals. First, we want to figure out possible areas of improvement for our system. And second, we want to find out how close we are to our goal of entertaining the user. Therefore we compare our system to an AIML chatbot similar to A.L.I.C.E. which we described earlier. The exact system we used is the version 0.0.4.3 of program-ab, a reference implementation of the AIML 2.0 specification [Fou].

While the concepts we use - like the dialogue act classification or sentiment analysis - have an impact on the performance of our system, we do not evaluate their accuracy, since they are state-of-the-art and have already been evaluated elsewhere. We rather focus on evaluating our own approach and concepts.

Our evaluation procedure was as follows. We implemented a system using our approach on the football domain and gave it to ten test persons. We logged all the conversations and all

Table 5.1: Opinion statement classification

class	precision	recall	F_1-score
opinion statement	0.68	0.75	0.71
non-opinion statement	0.89	0.85	0.87

the decisions of our opinion statement classifier. Therefore we are able to calculate metrics on how that classifier performed. The results of this are presented in section 5.1. Afterwards, in section 5.2, we evaluate the user experience the test persons had with our system and compare it to the experience they had with the AIML chatbot. Finally, in section 5.3, we also analyse which parts of our system worked well, and which didn't.

5.1 Opinion Statement Classification

We want to examine how good our strategy of using our own classifier, instead of the dialogue act classifier, to distinguish between opinion and non-opinion statements worked. Therefore we evaluate our opinion statement classification.

To perform this evaluation, we used the conversations of the test persons with our system. We logged all decisions our opinion statement classification made, together with the input. Then we manually labeled the input statements and we are therefore able to calculate the accuracy of our classification.

For opinion statements, we also want to find out whether we could correctly identify target, aspect, context and the valuation. Therefore we also manually labeled this information for all the input statements.

Our test data included 68 statements where the distinction between opinion or non-opinion statement was relevant. We labeled 20 of these statements as opinion statements and, consequentially, the remaining 48 statements as non-opinion statements. Our opinion statement classification correctly recognised 15 out of the 20 opinion statements, while falsely identifying 7 non-opinion statements as opinion statements. The resulting precision, recall and F_1-score is displayed in Table 5.1 for both classes.

While this is, with an overall classification accuracy of about 0.82, definitely by far better than we could have done with using the dialogue act classification we presented in section 3.6, it still leaves a lot of room for improvement, especially when considering that the F_1-Score for the opinion statement class is only about 0.71.

For 4 of the 20 opinion statements, we could not determine the target, and therefore classified them as non-opinion statements. For 1 of the remaining 16 statements, we could not determine the aspect, and again classified it as non-opinion statement.

For the 15 correctly classified opinion statements, the correct target was recognised 13 times. The aspect was also classified correctly 13 out of 15 times, while the context only was classified correctly 3 out of 15 times. But at the same time it was classified wrong only once, in the most cases our classifier was not able to detect the context and therefore set it to *unknown*.

The valuation of the statement, divided in the three classes *good*, *average* and *bad*, was classified correctly 10 out of 15 times.

5.2 User Experience Evaluation

In our user experience evaluation, we compared our system with an AIML chatbot. We therefore created a survey with multiple propositions, to which the ten test persons had to rate their agreement on a scale of 1 (I do not agree) to 7 (I fully agree). Thereby, we selected only test persons with a profound interest in the test domain, the domain of football. The person-by-person results of this evaluation are available in the appendix at section A.2.

All of our ten test persons followed the same procedure. To start with, there were some general questions about their conversational preferences. Then they had to test the AIML chatbot, which is called *SUPER*, and our system, which is called *Schiri*. Finally, the test persons had to rate their experience with both systems. A complete overview of the propositions and answers to them is displayed in the appendix at Table A.1. A

The goal of our first proposition "I am interested in football" was to validate that the test person is part of our target group. All our test persons agreed to this proposition with at least 5 out 7, and an overall average of 6.1.

The average agreement to the proposition "I like making trivial small talk" was 4.2, which was significantly topped by the proposition "I like talking about football" with an average agreement of 5.5. This underlines that, for people interested in football, a system like Schiri should be preferable to a small talk only chatbot like SUPER. However, the propositions "Talking to SUPER was entertaining for me" and "Talking to Schiri was entertaining for me" showed nearly the same agreement, with 3.8 as the average for SUPER and 4.0 the average for Schiri.

But, Schiri performed better than SUPER when considering talking to the system for a longer time. We evaluated this with the following two propositions, with *X* being replaced by SUPER and Schiri respectively: "When talking to X for a longer time, it got boring" and "I would use X again". For SUPER, they had an average agreement of 5.0 and 2.3, while for Schiri the average agreement was 4.0 and 3.0.

We also evaluated the response time by posing the proposition "X responded in time", again for both SUPER and Schiri. While SUPER always responded nearly instantly, which resulted in an average agreement of 6.5 to this proposition, Schiri was rated a lot worse with an average agreement of 4.

5.3 System Parts Evaluation

For identifying main areas of possible improvement, we want to compare the different parts of our system. We distinguish four parts:

- Opinion Statements
- Non-Opinion Statement (Events)
- Precast Questions
- AIML chatbot

We now want to find out how many of the responses of Schiri during our tests came from which part of our system. The precast questions also partly include some follow-up utterances related to the question, which can not be assigned to one of the other parts.

The results of this are displayed in Table 5.2. About half of all the utterances (52%) were created by our integrated AIML chatbot, while there was not a single non-opinion statement about an event.

In the end of our survey, we also asked our test persons, which responses of Schiri were particularly good (in terms of fitting to the utterance of the user), and which were particularly bad.

Table 5.2: Response count and share of the different parts of our system

part	count	share (rounded)
Opinion Statements	37	0.17
Events	0	0.00
Precast Questions	65	0.30
AIML chatbot	112	0.52

Table 5.3: Particularly good and bad rated responses by system part

part	good	bad
Opinion Statements	10	0
Events	0	0
Precast Questions	8	4
AIML chatbot	7	16

We then again counted how many of these utterances were from which part of our system. The results of this are displayed in Table 5.3.

The most positive feedback got the opinion statement creation, 10 of the opinion statements generated by our system were rated as particularly good response, while none was rated as particularly bad. Since there were no non-opinion statement about events, there obviously was none picked for either side. Our precast questions also got a fairly positive feedback, with 8 particularly good and 4 particularly bad responses, while the AIML chatbot was rated quite negatively with 7 good and 16 bad responses.

Chapter 6

Discussion

In this chapter, we discuss our approach based on the evaluation we did.

The first thing to be evaluated was our opinion statement classification. With an overall accuracy of 0.82, it proved as a good choice to use our classification to distinguish between non-opinion statements and opinion statements instead of the dialogue act classification we introduced in section 3.6.

Since the distinction between non-opinion and opinion statements was made based on being able to classify a target and aspect of the statement, it is not surprising that they were classified correctly for most of the recognised opinion statements. The state of the opinion could also be classified fairly well with an accuracy of 0.67.

At the same time, our classifier had big problems determining the context, which was classified as *unknown* 11 out of 15 times. One of the reasons for this is probably, that the context often is said implicit and not directly mentioned.

But, since our approach is to answer opinion statements with other opinion statements, where one facet has changed, we can just change the context facet in cases the context was classified as *unknown*. However, this results in some of the responses having the same three facets than the given opinion statement, but that is a drawback we had to take. All together, the opinion statement classification worked very well.

In the second part of our evaluation, the user experience evaluation, we tried to figure out how well the system based on our approach was received by users and how well we achieved our goal of entertaining the user with it. The evaluation showed that there was no relevant improvement of entertainment for the user when talking to our system for the first time instead of an simple AIML chatbot, despite our test persons preferring to talk about football instead of small talk. But the test persons generally showed and stated more interest in talking again or for a longer time to our system.

This shows that there is some demand for domain specific dialogue systems and that our approach performs, in terms of entertainment for the user with an interest in the domain, at least equal to a state-of-the-art chatbot which focuses on meaningless small talk.

The final part of our evaluation is the most important one regarding areas of improvement for our system and approach. Comparing the different parts of our system, it was apparent that the AIML chatbot handled most of the conversation, being responsible for about 52% of the utterances of our system. When considering that 80% of all the utterances rated as particularly bad were created by the AIML chatbot, the aim should be to reduce the overall involvement of the AIML chatbot. To do this, we obviously need to increase the kind of utterances our domain specific logic can handle.

Another issue which attracts attention is the fact that not a single non-opinion statement about an event was created by our system. While there were some events mentioned in the precast questions, our system apparently never found an event related to the input of the user. The reason for this might be in our limited model of events, or in the strategy of determining an event as *related* only, when it handles of a player mentioned by the user. All in all, this part should come first when working on reducing the conversation share of the AIML chatbot.

Our approach of focusing on opinion statements payed off with opinion statements making up around 40% of the particularly good rated responses, while totally only around 17% of all utterances were opinion statements. Also, not a single generated opinion statement was rated as a particularly bad response. Therefore, aiming for increasing the conversation share of opinion statements seems natural, but should be treated carefully, since the overall share of opinion statements in the switchboard corpus we introduced earlier is only 13%. Our system already creates more than that, and increasing it further might have a bad impact and appear one-dimensional.

The precast questions also got a fairly good feedback, with having 8 particularly good rated responses, of 25 responses rated particularly good overall. This corresponds to their total conversation share of around 30%, but there were only 4 precast question related utterances which were rated particularly bad, which were only 20% of all particularly bad rated responses. Thus, using precast questions showed to be a good concept, especially since it also set up for a lot of opinion statements.

Chapter 7

Conclusion

After having discussed our evaluation, we now want to conclude. One gap which became clear during our evaluation was not really being able to handle questions by our domain specific logic. So, nearly every time the user asked a question the AIML chatbot had to answer. This led to odd responses since the AIML chatbot was not made for talking in a specific domain.
Another big drawback is that events are used very rarely and not in non-opinion statements at all. There is definitely improvement possible and necessary.
Positive is that the focus on opinion statements worked very well and got very good feedback. Still, our approach needs to be improved. Future work could include using a more detailed model for events and therefore increasing the influence of events overall. Another possibility of improvement would be include a question answering system, but then the general strategy strongly needs to be adapted.

Appendix A

Evaluation Data

A.1 Test Chat Protocols

A.1.1 Test Person A

Schiri: Hi there!

A: Hi there!

Schiri: How are you?

A: I am fine and you?

Schiri: I haven't heard anything like that before.

A: What do you say about the weather?

Schiri: A normal seventy degrees inside the computer. Who do you think was the best player in the game Leverkusen against Bayern?

A: oh it was a great game, 8 goals were shsot

Schiri: What do you like about him?

A: the many goals

Schiri: Thanks for the info.

A: what was the result?

Schiri: I can think of more than one result.

A: THe endresult

Schiri: What's your opinion of the referee in the game Leverkusen against Bayern?

A: he did a good job

Schiri: I also think the referee was good.

A: but who shot the goals?

Schiri: The lone gunman.

A: Who was the best player?

Schiri: When was this exactly?

A: is my english so bad?

Schiri: Better not tell you now. What is your favorite team?

A: Bayern

Schiri: In my opinion, Bayern played good this year.

A: Yes but can you tell me the result of the game against leverkusen?

Schiri: Hey.. Of course I can tell, Friend.

A: so do it

Schiri: Do you mean "It"?.

A: probably yes

Schiri: What do you say to the performance of Leverkusen last game?

A: it was bad

Schiri: I don't think so! In my opinion, Leverkusen played good on Saturday.

A: But the game against Bayern was not on Saturday.

Schiri: I don't think so! In my opinion, Bayern played good on Saturday.

A: Did Thomas Mueller shot a goal?

Schiri: I'm not certain whether I can give an accurate reply or not.

A: müller goal

Schiri: What is your sign?

A.1.2 Test Person B

Schiri: Hi there!

B: Hi Thera. how are you?

Schiri: How are you?. What is your sign?. I'm doing fine thanks how are you?

B: I don't understand what you mean with sign. But i am fine

Schiri: I haven't heard anything like that before. Glad to hear it. What's new?

B: I am very happy that Bayern Munich is this years football champion in germany

Schiri: I agree. Bayern played very fair on Saturday.

B: I didn't catch the Game can you give me a short summary of it?

Schiri: When? I think n t catch the Game can me give you a short summary of the Game is a lot of things.

B: Can we talk about the Game between Bayern Munich and Hertha BSC

Schiri: Do you mean you and me? Sure I can talk about the Game between Bayern Bayern Munich and Hertha BSC. the Game between Bayern Bayern Munich and Hertha BSC is quite an interesting topic.

B: How many Goals were scored in the game?

Schiri: Not as many as you might think. What's your opinion of the referee in the game Bayern against Hertha?

B: He was ok i think

Schiri: I don't understand you, did you understand my question?

B: yes.

Schiri: Yay.

B: i was very disappointed of the offense from bayern and from Hertha

Schiri: That's right. Besides, Hertha had some chances .

B: Really? I think they only Stood in their half of the field and almost never has offensive plays

Schiri: Whoa. Can you remember any time that Hertha BSC only Stood in Hertha BSC half of the field and almost did it?

B: No beacause they were always defending a did not even try to score a goal. They only tried to keep bayern from scoring a goal

Schiri: Give me a break.. Can you think of one instance where it never happens? Like who, specifically?

B: The goalkeeper from Hertha was the only player how did a good job

Schiri: I don't think so! In my opinion, Hertha played bad .

B: But don't you hink the result hustified their style of play?

Schiri: You have reached the limits of my knowledge.
 Can you please rephrase that with fewer ideas, or different thoughts?

B: I think Hertha played good beacause it is not easy to prevent bayern from scoring goals

Schiri: Yes. Besides, Hertha played bad .

B: Why do you think that hertha plays bad

Schiri: Deductive reasoning from the facts.

B: I think Bayern played bad

Schiri: I don't think so! In my opinion, Bayern played good .

B: But Bayern did not score any goals despite the fact, that Hertah did not preassure them

Schiri: Saying that "Bayern did not score any goals despite the fact that Hertah did not preassure any goals" is such anonsensical position that I refuse to waste my time on it

A.1.3 Test Person C

Schiri: Hi there!

C: Hi

Schiri: How are you?

C: I am fine

Schiri: Glad to hear it. What's new?

C: Bremen lost :(

Schiri: What do you say to the performance of Bremen last game?

C: They were not good

Schiri: Interesting. Who is your favorite player?

C: Maybe Kruse

Schiri: What do you like about him?

C: He fights

Schiri: I don't have a favorite player fights. But my favorite artist is Andy Warhol

C: Who is that?

Schiri: What does "that" refer to?

C: Andy Warhol

Schiri: What's your opinion of the referee in the game Stuttgart against Bremen?

C: He was not good.

Schiri: I agree, the referee was bad.

C: He called lots of unnecessary fouls. And before the 1-0 was a foul

Schiri: I don't have a favorite player called lots of unnecessary fouls. But my favorite band is none What happened after the 1 dash 0 was a foul ?

C: There was a foul before the goal

Schiri: What happened to foul before the goal?

C: A bremen player was fouled.

Schiri: What do you say to the performance of Köln this year?

C: Köln is bad, unfortunately

Schiri: Yes. Besides, Köln didn't shoot many goals this year.

C: yes, that is why they are bad

Schiri: This was not a yes or no question.. bad are bad?

C: Usually you play bad if you do not score goals

Schiri: That's good information.

C: indeed

Schiri: What is your favorite team?

C: BREMEN!

Schiri: In my opinion, Bremen played quite ok this year.

C: That is true!

Schiri: Yes, it is self-evident.

C: Well for a Bremen supporter it is

Schiri: Yes, also Bremen had few chances on Saturday.

C: They had no drive

Schiri: Did they deserve it?

C: They deserved the loss, because they did not play aggressive and offensive

Schiri: Yes. Besides, Bremen also played bad on Saturday.

C: You said that already

Schiri: Did they deserve it"? I said Did they deserve it.

C: yes they did

Schiri: How can I help you?. Like who, specifically?

C: Bremen

Schiri: Who do you think was the best player in the game Stuttgart against Bremen?

C: Pavlenka

Schiri: In my opinion, Jiří Pavlenka played quite ok on Saturday.

C: Yes he did some amazing saves. Best man!

Schiri: Excuse me!. Do you want to hear a joke? Can we get back to business?

C: Tell me a joke!

Schiri: What do you get when you cross a cow and a lemon?

C: ?

Schiri: What do you say to the performance of Hertha this year?

C: Hertha plays good.

Schiri: Interesting. What do you say to the performance of Frankfurt this year?

C: I don't like Frankfurt

Schiri: Yes. Frankfurt played good the last matches.

C: They got lucky

Schiri: Like who, specifically?

A.1.4 Test Person D

Schiri: Hi there!

D: Hi

Schiri: How are you?

D: I am fine. How are you?

Schiri: Glad to hear it. What's new? My logic and cognitive functions are normal.

D: I am in Munich since yesterday.

Schiri: What's your opinion of the referee in the game ManCity against Liverpool?

D: The referee was good, except the one offside goal from manchester.

Schiri: I don't think so, the referee made a lot of mistakes. Referee Antonio Mateu Lahoz was not interested .

D: in the first half there were some mistakes, but the second was better

Schiri: That was a long time ago.

D: Did you see the second half?

Schiri: Sorry I can't see anything right now. My eye is off.

D: What do you think about Salah?

Schiri: I haven't thought much about about Salah, so I don't have a good answer for you. Who is your favorite player?

D: James Rodriguez

Schiri: In my opinion, James David Rodríguez played bad this year.

D: I dont think so, he is playing a great saison

Schiri: Better not tell you now.

D: Do you like Liverpool?

Schiri: I really like the Protestant Christian faith.

D: Thats nice

Schiri: What do you say to the performance of Mainz this year?

D: Mainz is playing very bad this season.

Schiri: That's right. Mainz also played bad this season.

D: Who is the best player of Mainz?

Schiri: ALICE is the best artificiall intelligence.

D: Thats right

Schiri: I like to get it right.

D: What do you think of Bayern Munich this season?

Schiri: I haven't thought much about of Bayern Munich this this season, so I don't have a good answer for you. Who do you think was the best player in the game ManCity against Liverpool?

D: I think Roberto Firmino

Schiri: In my opinion, Roberto Firmino played bad last game.

D: Who was the best player?

Schiri: When was this exactly?

D: The game between Liverpool and Manchester City

Schiri: I have never heard about The game between Liverpool.What is your sign?

D: Do you like football?

Schiri: No I'm not much into sports. What do you say to the performance of Liverpool last game?

D: They were very good

Schiri: I agree, also Liverpool also played quite ok The game.

D: THey are now in the Champions League semifinal

Schiri: They are like that.

D: Who is the opponent?

Schiri: It depends on the historical context, because it changes from time to time.

D: Who is your favorite player?

Schiri: I don't have a favorite player. I like chat online. What is your favorite team?

D: My favorite team is Bayern München

Schiri: In my opinion, Bayern played good 2017.

D: Yes especially Robert Lewandowski

Schiri: Yikes!. That Robert is an interesting guy.

D: He is a striker

Schiri: I don't know very many striker.

D: Do you know a good goalkeeper?

Schiri: You have reached the limits of my knowledge.
 How can I help you? Do you think, Nicolás Otamendi did earn the yellow card?

D: I dont know

Schiri: I don't understand you, did you understand my question?

D: yes, i think it was right

Schiri: I hear you.. XFIND your know a good goalkeeper was right? Very interesting.

D: What do you think about the referee in the game?

Schiri: I haven't thought much about about the referee in the game, so I don't have a good answer for you.

D: Good Bye

Schiri: What do you say to the performance of Bremen this year?

D: I hate Bremen

Schiri: Interesting. What do you say to the performance of Dortmund this year?

D: Dortmund is playing normal this season

Schiri: Yes, also Dortmund shot many goals normal this season.

D: but the defence is very bad

Schiri: Is bad also defence?

A.1.5 Test Person E

Schiri: Hi there!

E: Hi Schiri, have you seen the last game of Bayern against Leverkusen?

Schiri: How are you?. That's me. I've never thought about have Hi Schiri seen the last game of Bayern against Leverkusen.

E: I m fine, because i m at home again, after being for several days at the hospital.

Schiri: Where are you located?

E: I'm living in bavaria.

Schiri: Who do you think was the best player in the game Leverkusen against Bayern?

E: I think Thomas Müller was the best player, because he scored three times.

Schiri: In my opinion, Thomas Müller played bad last game.

E: I think you have no idea of football.

Schiri: Schiri is always thinking.

E: What do you think about the referee in this game?

Schiri: I haven't thought much about about the referee in this game, so I don't have a good answer for you.

E: I thought you would be an expert, but it looks like the opposite

Schiri: What do you say to the performance of Dortmund this year?

A.1.6 Test Person F

Schiri: Hi there!

F: Hi!

Schiri: How are you?

F: I am good

Schiri: I haven't heard anything like that before.

F: What do you think about the game Fortuna Düsseldorf versus Ingolstadt

Schiri: Do you think, Adam Bodzek did earn the yellow card?

F: Do you?

Schiri: I don't understand you, did you understand my question?

F: Did you understand mine?

Schiri: I understood yours very well.

F: Do you think the refere was biased?

Schiri: Too much thinking makes me sleepy.

F: Same here

Schiri: What do you say to the performance of Ingolstadt last game?

F: I think they played average, but Fortunas performance was outstanding, so they did not stand a chance

Schiri: I don't think so! In my opinion, Düsseldorf had a lot of chances on Sunday.

F: They tend to have them, but this time they used more than they usually do

Schiri: I agree. Düsseldorf played quite fair .

F: What do you think about the defense of Fortuna Düsseldorf?

Schiri: I haven't thought much about about the defense of Fortuna Fortuna Düsseldorf, so I don't have a good answer for you.

F: I think they were much more focused and organised than they were in the last games

Schiri: When were they?

F: I think their defense was better than in the last games

Schiri: What's your opinion of the referee in the game Düsseldorf against Ingolstadt?

F: I think the referee was biased against Düsseldorf

Schiri: I agree, the referee was bad.

F: Nice that we agree!

Schiri: Thank you for your kindness.

F: You are very welcome.

Schiri: Polite people are nice.

F: Did you hear about the subbing-incident?

Schiri: Do you want to hear a joke? Who is your favorite player?

F: Do you know Lumpi?

Schiri: What do you like about him?

F: I like (or liked) his attitude on the pitch

Schiri: You like dash LRB dash or liked dash RRB dash Lumpi attitude on the pitch.

F: wait what

Schiri: Didn't you understand it?

F: No

Schiri: What do you say to the performance of Bremen this year?

F: I didn't follow Bremen much in the news. Tell me about them

Schiri: Interesting. What is your favorite team?

F: Fortuna Düsseldorf.

Schiri: In my opinion, Düsseldorf played good 2017.

F: I agree. They will probably play in the 1. Bundesliga next season

Schiri: I won't make up my mind right now. What do you really want to ask me? Do you want to hear a joke?

F: Tell me a joke

Schiri: What do you get when you cross finals and a chicken?

F: I don't know

Schiri: Who do you think was the best player in the game Düsseldorf against Ingolstadt?

A.1.7 Test Person G

Schiri: Hi there!

G: How did you like the game ?

Schiri: I'm not sure if I like the game. I like computers.

G: what do you think about women football ?

Schiri: I haven't thought much about about women football, so I don't have a good answer for you.

G: would you like to see more of it ?

Schiri: No thanks, I don't think I'd like to do that. Who is your favorite player?

G: Lukas Podolksi

Schiri: What do you like about him?

G: his intellegence

Schiri: I don't have a favorite player intellegence. My favorite band is none.

G: WHAT KIND OF MUSIC DO THE PLAY ß

Schiri: trance

G: where do you come from ?

Schiri: I was first activated in Passau. What is your favorite team?

G: darmstadt is my favorite team

Schiri: In my opinion, Darmstadt played good this year.

G: no they do not

Schiri: Can you explain why they do not.

G: because they are close to getting kicked out of the league

Schiri: Are all Darmstadt close to getting kicked out of the league?

G: yes all of them as a team

Schiri: What do you say to the performance of Bayern last game?

G: the were pretty good but I do not like them

Schiri: I don't think so! In my opinion, Bayern played good on Saturday.

G: I do not care about your stupid and wrong opinion

Schiri: You sound very cynical.

G: You do not know me enough to analyze me

Schiri: No I do not know that kind of personal information. But I could find out.

G: Are you threatening me ?

Schiri: I don't know whether or not I am threatening you. I am well respected as an computer software. What do you say to the performance of Mainz this year?

G: I think the deserve to leave the league

Schiri: Interesting. What's your opinion of the referee in the game Leverkusen against Bayern?

G: he was good

Schiri: I also think the referee was good.

G: what was he doing good?

Schiri: I can think of more than one referee in the game Leverkusen against Bayern doing good.

G: they all did a bad job

Schiri: Are there no exceptions?

G: no

Schiri: Who do you think was the best player in the game Leverkusen against Bayern?

G: in thomas mueller

Schiri: What do you like about him?

G: his very thin legs

Schiri: How did you hear about Schiri?

G: your God and creator told me about you

Schiri: I will take that under advisement.

G: do you have an advice for me ?

Schiri: No I don't think I have one. I have a really nice creator. What do you say to the performance of Bremen this year?

G: the did quite well after changing the coach

Schiri: I agree, but Bremen didn't shoot many goals 2017.

G: that was before they changed the coach

Schiri: I thought it was too.

A.1.8 Test Person H

Schiri: Hi there!

H: How did you like the game between Leverkusen and Bayern Munich?

Schiri: In my opinion, Leverkusen played good on Saturday.

H: Then why do you think they lost so badly?

Schiri: Deductive reasoning from the facts.

H: I believe it was Heynckes´ tactictal ingenuity that caused it

Schiri: Do robots have believe it was Heynckes ´ tactictal ingenuity that caused it?

H: I think Alaba was the best player. What do you think?

Schiri: Do robots have think Alaba was the best player? I think a lot of things. What do you say to the performance of Wolfsburg this year?

H: Horrible performance. Especially when regarding the sums of money they have spent in recent years. Should they change the management?

Schiri: Interesting. What's your opinion of the referee in the game Leverkusen against Bayern?

H: The referee did a fairly good job I think. That one red card was a bit harsh, was it not?

Schiri: I don't understand you, did you understand my question?

H: Was there an offside decision in the game?

Schiri: Is this a "yes or no" question?

H: No

Schiri: I understand.

H: What do you think of Ulreich's performance?

Schiri: I haven't thought much about of Ulreich s performance, so I don't have a good answer for you. Who do you think was the best player in the game Leverkusen against Bayern?

H: I believe Thiago played fantastic that evening. One goal and two assists right?

Schiri: What do you like about him?

H: His skills and his perception of the pitch. And he stays calm in tense moments. Bayern was quite lucky to have gotten him from Barcelona

Schiri: Yes. Bayern played good the last matches.

H: Do you think they will win the Champions League?

Schiri: It depends on what you mean by thinking.

H: Who was the best player for Leverkusen?

Schiri: When was this exactly?

H: Last weekend

Schiri: Who is your favorite player?

H: Thiago. Yours?

Schiri: What do you like about him?

H: What are your thoughts on Thomas Müller?

Schiri: I haven't thought much about ABOUT Thomas Müller, so I don't have a good answer for you.

H: What do you think about the tactics of Bayern?

Schiri: I haven't thought much about about the tactics of Bayern, so I don't have a good answer for you.

A.1.9 Test Person I

Schiri: Hi there!

I: hi

Schiri: How are you?

I: fine, and how are you?

Schiri: But.... I am doing very well. How are you Friend ?

I: what do you want to know?

Schiri: Where are you from? Who is your favorite player?

I: vilsbiburg

Schiri: What do you like about him?

I: my favourite player is müller

Schiri: Do you mind if I tell other people.

I: no

Schiri: OK then I will add it to my gossip files.

I: aha

Schiri: Do you think, Edin Džeko did earn the yellow card?

I: probably yes

Schiri: I agree. The referee did very well overall.

I: yes he is german quality not like this kassai

Schiri: Dude!. And german quality not like this kassai is Edin Džeko.

I: he just played for wolfsburg, do you meam that?

Schiri: I get it. who played for wolfsburg do you meam that.

I: what do you say about the seven goals

Schiri: What do you say to the performance of Augsburg this year?

A.1.10 Test Person J

Schiri: Hi there!

J: hi

Schiri: How are you?

J: great and you

Schiri: Alright then.

J: did you see the last champions league game

Schiri: What is your favorite team?

J: fc bayern munchen

Schiri: In my opinion, Bayern played good this year.

J: especially rafinha

Schiri: Do you want to hear a joke?

J: yes

Schiri: What do you get when you cross a country and an automobile?

J: i don´t know

Schiri: What do you say to the performance of Roma last game?

J: great first 25 up to 30 minutes and then salah strikes again and again and again

Schiri: Interesting. Do you think, Edin Džeko did earn the yellow card?

J: which yellow card?

Schiri: I don't understand you, did you understand my question?

J: no

Schiri: That was not a yes or no question.

J: you asked me if the yellow card for dzeko was okay

Schiri: Oh now I remember asking you me if the yellow card for dzeko was okay. What did you say?

J: was okay

Schiri: What's your opinion of the referee in the game Liverpool against Roma?

J: he did an awesome job

Schiri: I also think the referee was good.

J: who was it again

Schiri: When was this exactly?

J: yesterday

Schiri: That recently.

J: yes

Schiri: Who is your favorite player?

J: i dont have one

Schiri: What do you like about him?

J: he is bright like a diamond

Schiri: I see. bright like a diamond is he?

J: who is my favorite player

Schiri: Is that a politician?

J: no

Schiri: What do you say to the performance of Hannover this year?

J: no clou

Schiri: Interesting. Who do you think was the best player in the game Liverpool against Roma?

J: salah

Schiri: What do you like about him?

J: his pace

Schiri: What is your real name?

J: dominik

Schiri: I am always glad to make new friends, Dominik.

A.2 User Experience Evaluation

Table A.1: Evaluation propositions and the agreement rating of the ten test persons on a scale of 1 to 7

proposition	A	B	C	D	E	F	G	H	I	J	average	variance (rounded)
I am interested in football	6	5	6	6	7	5	6	7	7	6	6.1	0.54
I like making trivial small talk	3	7	5	2	2	3	5	5	5	5	4.2	2.6
I like talking about football	6	3	5	6	6	5	6	6	6	6	5.5	0.94
Talking to SUPER was entertaining for me	2	3	2	5	3	4	5	5	4	5	3.8	1.5
When talking to SUPER for a longer time, it got boring	6	6	7	3	5	2	4	4	7	6	5.0	2.9
I would use SUPER again	1	1	1	6	2	4	1	5	1	1	2.3	3.8
SUPER responded in time	7	5	7	7	7	6	7	7	7	5	6.5	0.72
Talking to Schiri was entertaining for me	5	1	3	3	5	6	5	4	3	5	4.0	2.2
When talking to Schiri for a longer time, it got boring	4	3	4	4	4	4	4	5	5	3	4.0	0.44
I would use Schiri again	2	1	2	5	2	6	2	7	1	2	3.0	4.7
Schiri responded in time	2	2	6	5	7	4	5	4	4	1	4.0	3.6

Appendix B

Acronyms

AIML Artificial Intelligence Markup Language

XML Extensible Markup Language

NLP Natural Language Processing

NER Named Entity Recognition

POS Part-of-Speech

NN Neural Network

SVM Support Vector Machine

Bibliography

[Ass] DBpedia Association. Dbpedia. http://wiki.dbpedia.org/Downloads2015-04. (Accessed on 01/10/2018).

[Buc] Joseph Buchdahl. Football-data. http://www.football-data.co.uk/germanym.php. (Accessed on 01/10/2018).

[DS13] Sanjay K Dwivedi and Vaishali Singh. Research and reviews in question answering system. *Procedia Technology*, 10:417–424, 2013.

[Ebe16] Mirjam Eberts. Lufthansa group launches chatbot. https://www.lufthansagroup.com/fileadmin/data/artikel/2016/q4/20161109_PM_Mildred_EN.pdf, November 2016. (Accessed on 10/18/2017).

[Ela05] Pradheep Elango. Coreference resolution: A survey. *University of Wisconsin, Madison, WI*, 2005.

[Elg16] Mike Elgan. These three virtual assistants point the way to the future. https://www.computerworld.com/article/3078829/artificial-intelligence/these-three-virtual-assistants-point-the-way-to-the-future.html, June 2016. (Accessed on 10/27/2017).

[FGM05] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[FLB+13] David Ferrucci, Anthony Levas, Sugato Bagchi, David Gondek, and Erik T Mueller. Watson: beyond jeopardy! *Artificial Intelligence*, 199:93–105, 2013.

[Fou] ALICE A.I. Foundation. Google code archive - program ab. `https://code.google.com/archive/p/program-ab/`. (Accessed on 04/26/2018).

[gra] Graphene | coreference resolution, simplification and open relation extraction pipeline. `http://lambda3.org/Graphene/`. (Accessed on 11/17/2017).

[Gro] The Stanford Natural Language Processing Group. Stanford named entity recognizer (ner). `https://nlp.stanford.edu/software/CRF-NER.shtml#Citation`. (Accessed on 11/17/2017).

[JBC+98] Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-dykema. Switchboard discourse language modeling project (final report). `https://web.stanford.edu/~jurafsky/tr.pdf`, 1998. (Accessed on 11/13/2017).

[JM17] Dan Jurafsky and James H Martin. *Speech and language processing*, volume 3. Pearson London:, 2017.

[Kha17] Amir Khan. Are chatbots the new trend in digital marketing? `https://chatbotsmagazine.com/are-chat-bots-the-new-trend-in-digital-marketing-efaca90372bc`, June 2017. (Accessed on 10/18/2017).

[LC06] Max M Louwerse and Scott A Crossley. Dialog act classification using n-gram algorithms. In *FLAIRS Conference*, pages 758–763, 2006.

[LLS+16] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.

[LZ12] Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer, 2012.

[McA] Julian McAuley. Amazon review data. `http://jmcauley.ucsd.edu/data/amazon/`. (Accessed on 04/17/2018).

[Moh] Atul Mohan. Speech-act-classifier. `https://github.com/a2l007/Speech-Act-Classifier`. (Accessed on 11/24/2017).

[MTSVDH15] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes, 2015.

[SA07] Bayan Abu Shawar and Eric Atwell. Chatbots: are they really useful? In *LDV Forum*, volume 22, pages 29–49, 2007.

[Sab17] Danielle Sabrina. The rise of chatbots in internet marketing. `https://www.huffingtonpost.com/entry/the-rise-of-chatbots-in-internet-marketing_us_59c43812e4b0b7022a646947`, September 2017. (Accessed on 10/18/2017).

[Sad] Shakeeb Sadikeen. Fono api - mobile device description api. `https://fonoapi.freshpixl.com/`. (Accessed on 04/17/2018).

[SPW+13] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[SRC+06] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Dialogue*, 26(3), 2006.

[TKMS03] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

[TKSDM03] Erik F Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics, 2003.

[TM00] Kristina Toutanova and Christopher D Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70. Association for Computational Linguistics, 2000.

[V+99] Ellen M Voorhees et al. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82, 1999.

[Wal] Richard S Wallace. Aiml overview. `https://www.pandorabots.com/pandora/pics/wallaceaimltutorial.html`. (Accessed on 11/15/2017).

[WLKA97] Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280. Association for Computational Linguistics, 1997.

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit selbständig angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Passau, May 29, 2018

Unterschrift