Universität Zürich[UZH]

Bachelorarbeit
zur Erlangung des akademischen Grades
Bachelor of Arts
der Philosophischen Fakultät der Universität Zürich

Modeling dialogues for querying orofacial pain patients

Verfasserin: Salome Wildermuth
Matrikel-Nr: 10-289-544

Referentin/Referent: PD. Dr. G. Schneider

Institut für Computerlinguistik

Abgabedatum: 01.12.2022

Abstract

Software applications that talk to humans in natural language become increasingly important in a world of ubiquitous computing. So-called conversational agents already replaced humans in advisory functions, help desks, and customer services. There is increasing focus of such applications in complex and sensitive fields like virtual assistance, e-learning, car and air traffic or medical diagnostics.

The present study deals with conversational agents in the health care domain. The purpose is to analyze and engineer requirements and to implement them in form of an application that asks questions and records answers related to a certain topic in the field of orofacial complaints. The term *orofacial* refers to everything related to the anatomic face, mouth and nasal sinus. This thesis introduces into the domain of intelligent conversational agents and discusses prerequisites for a coherent and constructive conversation, putting emphasis on the empathy component. Empathy is considered as an important precondition for establishing and maintaining a familiar and trusted conversation setting. Finally, decision making processes for implementation details, respectively the employment of methods, techniques, and language models are documented, as well as the evaluation of the resulting conversations.

Zusammenfassung

Softwareanwendungen, die sich mit Menschen in natürlicher Sprache unterhalten, gewinnen an Bedeutung in einer Welt allgegenwärtiger Computer. Sogenannte Konversationsagenten haben Menschen bereits in vielen beratenden Funktionen, Informationsschaltern und Dienstleistungen ersetzt. Der Fokus der Forschung auf solche Anwendungen in komplexen und sensitiven Domänen, wie virtuellen Assistenten, elektronischem Lernen, Auto- und Flugverkehr oder medizinischer Diagnostik nimmt stetig zu.

Die vorliegenden Arbeit beschäftigt sich mit Konversationsagenten im Gesundheitswesen. Ihr Zweck ist die Analyse und Formulierung von Anforderungen und die Implementation eines Systems, das Fragen stellt und Antworten entgegennimmt bezüglich eines bestimmten Themas in der Domäne von orofazialen Beschwerden. Der Begriff *orofazial* bezieht sich auf alles im Zusammenhang mit dem anatomischen Gesicht, Mund- und Nasenraum. Die Arbeit führt in das Gebiet von intelligenten Konversationsagenten ein und diskutiert Voraussetzungen für eine kohärente und konstruktive Unterhaltung mit Schwerpunkt auf die Komponente der Empathie. Empathie wird als wichtige Voraussetzung für den Aufbau und den Unterhalt eines vertrauten und sicheren Unterhaltungsrahmens erachtet. Schliesslich sind Entscheidungsprozesse für Implementationdetails respektive der Einsatz von Methoden, Techniken und Sprachmodellen dokumentiert sowie die Evaluation der daraus resultierenden Konversationen.

Acknowledgement

***

I want to thank **Gerold Schneider**, my supervisor, who supported and guided this work with enthusiasm and inspiring inputs.

***

I want to thank my lovely partner, **Gianni Pedrazzini**, for his thorough proofreading, constructive critique on structure and content, for our interesting discussions and for his patience.

***

I want to thank my work mate - the best and most loyal one can imagine - **Yves Savoy**, for the verification of my installation guide and for testing the MDA.

***

I want to thank my little sister, **Andrea Wildermuth**, for testing the MDA with creative conversations.

***

I want to thank **Allen Hovsepian** for reviewing my use of the English language.

***

Contents

Abstract i

Acknowledgement iii

Contents iv

List of Figures vi

List of Tables vii

List of Acronyms viii

1 Introduction 1

2 Conversational Agents 3
- 2.1 Definition . . . 3
- 2.2 CA in Health Care . . . 4
- 2.3 Medical Dialogue Agent (MDA) . . . 4

3 Computational Methods 5
- 3.1 Research and state of the art . . . 5
- 3.2 Transformer and probabilistic tasks . . . 6
  - 3.2.1 Machine translation . . . 7
  - 3.2.2 Sentiment analysis . . . 8
  - 3.2.3 Topic modeling and inference . . . 8
  - 3.2.4 Named entity recognition . . . 9
  - 3.2.5 Question generation . . . 10
  - 3.2.6 Training data . . . 12
- 3.3 Rulebased tasks . . . 14
  - 3.3.1 Predefined conversation elements . . . 14
  - 3.3.2 Preprocessing . . . 15
  - 3.3.3 Runtime memory and data archiving . . . 16

4 Modeling Dialogues in Health Care 17

4.1 Research . . . 17
4.2 Realization in the MDA . . . 18
4.2.1 Empathy . . . 18
4.2.2 Common Ground . . . 19
4.2.3 Continutity and variety . . . 19
4.2.4 Topic Coverage . . . 20
4.3 Preconditions for meaningful question generation . . . 20

5 Implementation Details **22**
5.1 Data model . . . 22
5.1.1 Diagram . . . 23
5.1.2 Classes . . . 23
5.2 Configurability . . . 24
5.3 Steps per conversation turn . . . 26
5.4 Integration of Language Representation Models . . . 28
5.5 Ruleset for element selection . . . 28

6 Results and Discussion **30**
6.1 Empathy . . . 30
6.2 Testing frame- and agent-based QG . . . 31
6.2.1 Setting . . . 31
6.2.2 Test results . . . 32
6.3 Evaluation of methods and tools . . . 33

7 Conclusion **35**

Glossary **37**

References **40**

A Test Conversations **43**

B Lists of predefined content **49**
B.1 Profile questions . . . 49
B.2 Empathic phrases . . . 50
B.3 Topics . . . 52
B.4 Topic questions . . . 54
B.5 Questions for more details . . . 56

C Other **57**
C.1 Information text for test persons . . . 57

List of Figures

1	Employed language models	7
2	Bag-of-words vs. LDA	9
3	Usage of questions according to conversation phase	15
4	Conversation data model	22
5	Steps per conversation turn	27

List of Tables

1 Forbidden question words . . . 9
2 QG examples . . . 10
3 Trigger terms for question words . . . 12
4 Replacement of sensitive data . . . 13
5 Orthographic mistakes . . . 14
6 Map first person to second person rules . . . 16
7 Utterances supporting common ground . . . 19
8 Profile questions . . . 24
9 Questions asking for more details . . . 24
10 Empathic phrases . . . 25
11 Topic questions . . . 25
12 Topic model . . . 26
13 Representations of user's answer . . . 28
14 Test conversation 1 . . . 43
15 Test conversation 2 . . . 44
16 Test conversation 3 . . . 45
17 Test conversation 4 . . . 46
18 Test conversation 5 . . . 47
19 Test conversation 6 . . . 48

List of Acronyms

BART	Bidirectional Auto-Regressive Transformers
BERT	Bidirectional Encoder Representations from Transformers
CA	Conversational Agent
IR	Information Retrieval
LDA	Latent Dirichlet Allocation
MALLET	Machine learning for language toolkit
MDA	Medical Dialogue Agent
MT	Machine Translation
NER	Named Entity Recognition
NLP	Natural Language Processing
QG	Question Generation
SOLID	Principles of Single-responsibility, Open-closed, Liskov substitution, Interface segregation, Dependency inversion
SOTA	State-Of-The-Art
SQuAD	The Stanford Question Answering Dataset
WISE	Web-based interdisciplinary symptom evaluation

1 Introduction

With rapidly growing information repositories like the internet or companies' big data stores and increasing microchips performance computing machines learn and build up knowledge and experience faster and faster.

One of the growing domains is the development of intelligent conversation systems. They are able to record human language in written or spoken form and to produce an appropriate reaction in a sense of giving their human counterpart the feeling that their statements have been understood and that they are being taken seriously. They do this by using language representation models that were trained on big text corpora. There are many different shapings of conversational agents (CA), from invisible conversation partners behind an online messenger, up to very manlike robots with gestures and facial expressions. But all of them come up with competence in processing and producing language.

This thesis aims to contribute to the research of conversational systems' development. It relates to work done by Ettlin et al., who designed and constructed a "web-based interdisciplinary symptom evaluation" (WISE) for patients questionaries. It analyses how CAs can be employed to retrieve information from patients, how a constructive conversation, where the patient is willing to reveal information, can be maintained, and where the captured information is used to create a meaningful follow-up question. The fulfillment of this easy-sounding task is a topic of extensive research. The required criteria touch many different domains like information and communications technology, computational linguistics or psychology. Not to forget the field where the CA is actually implemented - in this case healthcare.

One important factor considered for the establishment of a trusted and secure conversational environment is the empathy that conversational partners show towards each other. This is one of the main research questions: Which role does empathy play in a written human-computer interaction and how can empathy be integrated in a real-world medical conversational agent? Another goal is to identify ways of information retrieval (IR) from given answers related to the domain of orofacial pain, that support the generation of meaningful follow-up questions best. Appropriate language representation models per NLP task have to be selected and a medical dialogue application that combines the different NLP steps through a well-designed data model has to be developed. Finally the application's potential is assessed, based on the resulting conversation skills. In the end the employed techniques are evaluated, and the most promising results as well as results that did not satisfy the expectations are discussed.

This medical dialogue agent (MDA)[1] is implemented in form of a hybrid system: it contains intelligent as well as rulebased components. That means, it accomplishes some tasks by applying deep learning models, namely transformers, and probabilistic models. The logic of how the different NLP tasks are constituted and combined together bases on hardcoded conversation rules. The system does not conduct a dialogue with generated content only, but provides a mix of generated and predefined sentences.

[1] https://github.com/salwil/conversational_agent_for_medical_dialogues

2 Conversational Agents

2.1 Definition

Conversational agents (CA) are systems that maintain a dialogue with humans in natural language. Their goal is to establish a conversation as close to a human conversation as possible. Alan Turing formed the famous Turing-test in the 1950s which determined whether a computer was able to behave indistinguishably from a human. The test requires a subject to hold two conversations via keyboard and screen. One with a computer and one with another human. When the subject cannot determine which of the two is the human and which one the computer after a longer conversation with them, the computer has passed the turing test. So-called task-oriented dialogue agents seem to have a higher chance to pass the turing test than open domain agents. It is easier to train agents for a conversation of good quality in one specific field, than to train an all-round genius conversation agent engaging in a discussion about various different topics.

CAs can be classified by various taxonomies. Laranjo et al. characterize them based on the following categories: Dialogue management, dialogue initiative, input and output modality and task or open domain orientation. Dialogue management refers to how the conversation is maintained. This can be either finite-state, frame-based or agent-based. Finite-state refers to a straightforward system working through a sequence of tasks, without letting the user actively participate or influence the structure. Frame-based systems also have predefined elements, but their employment depends on the course of the activities. They are more flexible than finite-state systems. Finally an agent-based system consists of intelligent subsystems (agents) executing tasks. Such a system is considered to be highly flexible and responding to the course of activities. The dialogue initiative specifies the leader of the conversation. It can be the user, the system or a mixed form. Finally, for the input and output Laranjo et al. distinguish between spoken and written modality. The written input can be categorized further into fixed user input, like in multiple choice or unrestricted input, like free text.

Replacing human services by CA is of increasing interest in many sectors such as government, insurance, tourism, banking, or health care [Van Pinxteren et al., 2020, 204].

2.2 CA in Health Care

Laranjo et al. found that the potential benefits of using conversational agents for health-related purposes are high. With increasing popularity of artificial intelligence the focus towards intelligent dialogue agents increased. However, the majority of the systems they reviewed allowed only for constrained user input (e.g. multiple-choice) without the capability to understand natural language input.

2.3 Medical Dialogue Agent (MDA)

The aim of this work is to design and implement a simple conversational agent that questions patients with orofacial complaints. The goal is to engage in a conversation that captures as much information as possible before the patients get in direct contact with the doctor. Many aspects are relevant to achieve this goal. Principally, it is a question of how the dialogue can be modeled such that users are willing to disclose information about themselves. Furthermore, the agent should be able to record the provided inputs and to classify and process information appropriately.

The MDA is a hybrid system, consisting of transformer- probability- and rulebased question generation. The dialogue management consists of finite-state, frame-based as well as agent-based components. The MDA is responsible to lead the conversation, it is a system-initiative architecture. The MDA is clearly a task-oriented dialogue agent. The implemented rules as well as the predetermined questions and sentence parts are aligned exclusively to the topic of orofacial pains.

3 Computational Methods

Question generation (QG) requires a system-initiative architecture. Implementing a system that controls a dialogue is challenging. It is hard to get out of a dead-end, for example, if the system gets stuck in some topic or even worse: in nonsense. While the system has to understand and interpret user input appropriately, it also has to be able to produce a follow-up question that keeps an informative conversation going and does not repeat itself. Some compelling principles for maintaining such a conversation are listed in chapter 4. This chapter focuses on the employment of computational tools and frameworks that can contribute to the fulfillment of crucial conversation qualities.

3.1 Research and state of the art

There is extensive research about communication in CAs in general and an increasing amount about health care domain specific CAs. The wide variety of literature is due to the countless forms and characteristics that CAs show. For the MDA implemented in the course of this thesis, research and findings in relation to frame- and agent-based conversations, system-initiative architectures and written free-text input and output are of prevalent interest.

CA applications with similar competences to the MDA typically share some common characteristics regarding their architectures and their employment of certain natural language processing (NLP) steps. Most applications are so-called hybrid systems which are built up of a mix of rule- and corpus-based components. *CORK* (*COnversational agent framewoRK*), a "modular framework to facilitate and accelerate the realization and the maintenance of intelligent Conversational Agents"[Catania et al., 2019], for example, contains modules for sentiment analysis, topic analysis, user profiling and last but not least for output creation. The output creation module works with content from a table of anticipated templates. Based on the results of the previous analysis steps the best fitting template is selected and filled with appropriate content for the next conversation turn.[Catania et al., 2019]

The approach of using templates for text generation appears also in other contexts. Fabbri et al. proposed an unsupervised template-based approach for training a QG model. A very similar technique, called masked-language modeling (MLM), is widespread in the field of unsupervised training of language models. SOTA language models that are suitable for text generation tasks, like BERT, have been pretrained with text containing masked tokens. The model learns for generative NLP tasks by learning to predict the masked tokens.

NLP tasks are often supported by neural language representation models. There are numerous open source models pretrained on datasets for executing NLP tasks, like machine translation (MT), information retrieval or extraction, sentiment analysis, or text generation. Some of them are pretrained on large datasets that contain data, tailored to a specific task (e.g. question-answer pairs for a QG model). Some years ago Google researchers presented groundbreaking results with a new approach, called Transformer architecture. Those model architectures rely entirely on attention mechanisms instead of recurrence. At that time the dominant sequence transduction models were based on complex recurrent or convolutional neural networks with encoder-decoder architectures [Vaswani et al., 2017]. Transformers outperformed SOTA translation quality while using less computation power for training than SOTA language models. [Vaswani et al., 2017]. In the meantime, NLP cannot be thought without Transformers anymore. Well-known sequence-to-sequence models, like BERT, GPT-2, or GPT-3, implement the Transformers architecture.

Other NLP steps, like topic analysis, use probabilistic models rather than neural methods. Topic models provide a way to analyze large volumes of unlabeled text. A "topic" consists of a cluster of words that frequently occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meaning. [Mimno et al., 2002] The most popular algorithm for topic modeling is the Latent Dirichlet Allocation (LDA), which has been introduced in the early 2000 by "David Blei and friends" [Graham et al., 2002]. LDA draws probability distributions of keywords in a dataset consisting of many documents. Each cluster of keywords forms a *latent* topic. Every document in the training dataset can be assigned to each topic with a certain probability. From a pretrained topic model, topics can be inferred for new unknown data.

3.2 Transformer and probabilistic tasks

Figure 1 shows the employment of the transformer and probabilistic models that the MDA uses within one conversation turn.

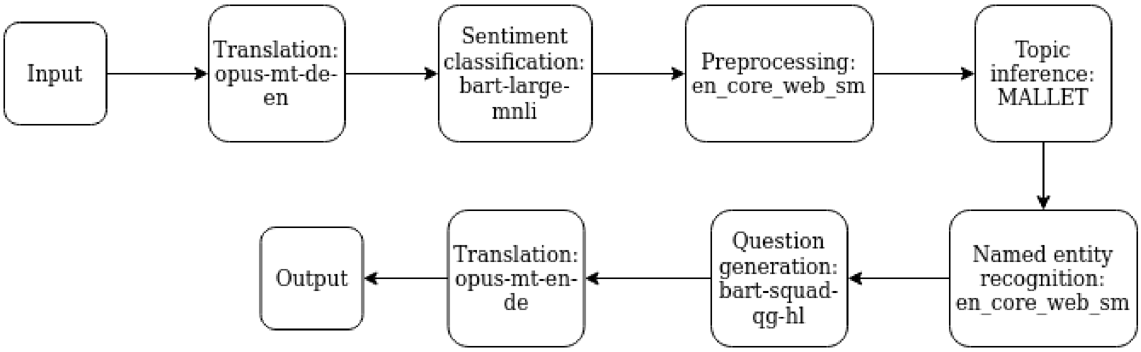

Figure 1: Employed language models

3.2.1 Machine translation

Pretrained language models for the NLP tasks that the MDA implements are available in both German and English language. But in general, English models perform better. The main reason for that is that there is more data for pretraining in English available than in German. Leading companies in the NLP area, like Google or Facebook, tend to start developing new methods on base of English data and expand to other languages later. Even though German models are often not far behind, English models are usually ahead in terms of performance in most usecases. In addition to that, there are more research papers, documentations and tutorials related to NLP tasks in English, especially in the text and question generation domain.

The MDA can talk English or German. When the user selects German as conversation language, the MDA translates every input to English before processing it. The final output is translated back to German, before displaying it to the user. Herefore, the pretrained Helsinki-NLP models, `opus-mt-de-en` and `opus-mt-en-de`, finetuned on translated WISE data, are used.

The `opus-mt-de-en` is finetuned on a German-English translation dataset which contains all the patient answers to the five main WISE questions in both German and English. The idea is to sensitize the MT model on patterns of patients answers and on domain specific terminology. The second model has to be sensitized on appropriate medical questions in context of the orofacial complaints. It is therefore finetuned on an English-German translation dataset consisting of all the survey questions. The translation dataset with the survey questions in English and their German correspondents consists of about 140 questions and 232 subquestions (belonging to multiple choice questions) in both languages.

3.2.2 Sentiment analysis

Sentiment analysis describes a technique to classify text into categories *positive*, *negative* and *neutral*. However, for getting insight into the patient's mental state, a more distinct classification of the answers would be valuable. Yin et al. proposed a method how a model pretrained for natural language inference (NLI), can be used as zero-shot text classifier. With NLI a hypothesis is classified as true (entailment), false (contradiction) or neutral. A BART model, pretrained on the Multi-GenreNLI corpus, is foreseen to measure the compatibility of two sentences (premise and hypothesis) by classifying them either as contradiction, entailment, or neutral. The model can indeed be used also for predicting a sentence (premise) and a candidate label (hypothesis) instead of a second sentence. The sentence-pair classifier is repurposed to a sequence-to-one classifier. Huggingface provides a zero-shot classification pipeline that can encapsulate this classification model. Any combination of sentences and labels can be passed to the pipeline for multi-label classification of the given text.

3.2.3 Topic modeling and inference

MALLET (machine learning for language toolkit) is a topic model package for training topic models on data with LDA and inferring topics for unknown data. It has been used for training a probabilistic model with 10 topics on the WISE datasets and the MDA uses it for topic inference on new data, i.e. on the user input. Each latent topic consists of 10 keywords that are semantically related. Figure 2 shows, how LDA topic models emerge: instead of mapping the documents to every token in the vocabulary, like in the bag-of-words approach, latent topic variables build a consolidation layer between documents and vocabulary. On the basis of the token occurrences inside a document, the topic probability of this document can be determined. In Figure 2 Doc1 and Doc2 have the higher probability for the latent topic "eating" because of the keyword distribution, while Doc3 and Doc4 have the higher probability for "healthcare". In the topic modeling for the MDA, each survey record is considered as one document. So the topic model was created from 117338 preprocessed tokens in 2840 different documents.

The result of the topic inference is a file with inferred topic distributions for the given input. The system selects the three topics with the highest weights. Unfortunately the current topic model has very unbalanced weights. The topics with the highest weights can be inferred under certain circumstances, even if the input sentence has nothing to do with the topic at all. An approach of balancing the behavior, is to

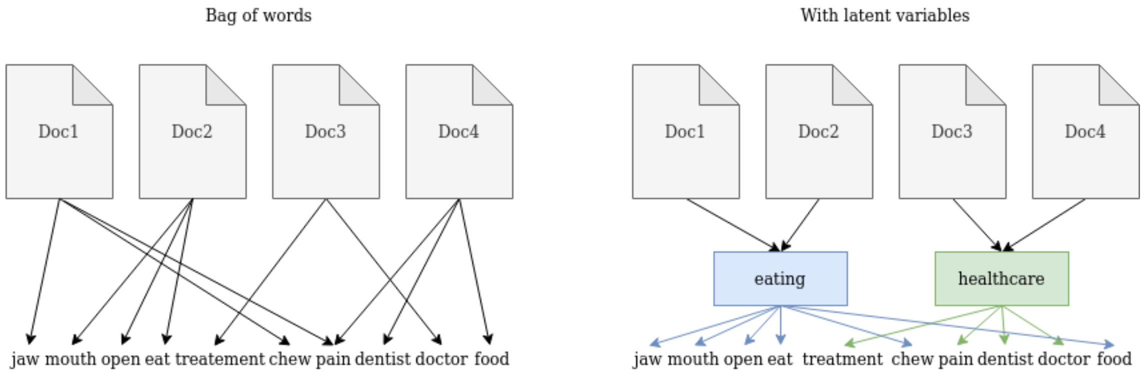

Figure 2: Bag-of-words vs. LDA (inspired by Ma [2019])

compute relative topic weights after the topic inference. That means the inferred topic distributions are divided by the respective topic weight of the topic model.

3.2.4 Named entity recognition

Named entity recognition (NER) is used to find out whether a sentence contains information about dates and times, events, locations, organisations or persons. This information is used for the rulebased question word enforcement. The system should never enforce question words with functions that eventually ask for an information that was already included in the last answer. Table 1 shows some examples how NER helps to prevent triggering of inappropriate question words for follow-up questions.

Last answer	**NE**	**Forbidden question words**
The headaches are particularly bad in the evening.	TIME	When
Dr. Muller told me to stay in bed and rest.	PERSON	Who
I take three aspirin per day.	QUANTITY, PRODUCT	What, how many
I have already been at the USZ for an examination.	LOC	Where

Table 1: Sentences and forbidden question words for follow-up question

NER is done with the en_core_web_sm pipeline from the Python spaCy library. This is a well performing model that is also used for other linguistic features that spaCy offers. There are indeed newer technologies - spaCy also has a NER API with Transformer models, but for this usecase the en_core_web_sm was decided to be sufficient since it is already used for lemmatization as part of the preprocessing task in the MDA. Once loaded, it can serve multiple purposes.

3.2.5 Question generation

Three criteria are relevant for the selection of the best question generation model:

1. The model should be open source.
2. The model should not occupy too much computer memory.
3. The model should be trained and perform well on question generation.
4. The model should provide an API that fits into the architecture of the implemented CA system.

Many of the freely available sequence-to-sequence models trained on question generation perform respectably in terms of quality and memory usage. However none stands out sufficiently from the others so that a selection justified by the criteria 2 or 3, could be made. When it comes to criterion 4, the model architectures compel attention: Models for QG are designed to work with different shapes of inputs depending on their architecture. One model generates questions solely based on an input text, but provides multiple possible follow-up questions, another accepts not only an input text, but also context, but returns only one follow-up question. In combination with the other features, like the topic inference and the predefined questions, it is not urgently necessary to have multiple questions to choose from. Instead, the experiments and analysis revealed a need for having a certain ability to take influence on the QG. Most of the investigated models lack on that. They offer little or no possibility to steer the model's QG.

Input:	**Generated question**
Answer: Eating food is painful. Context: I have painful hardening and cramps in the muscles of the jaw.	What is the eating pain?
I have painful hardening and cramps in the muscles of the jaw.	What is the problem with my jaw?
I have painful hardening and cramps in the muscles of the jaw.	What do I have in the muscles of the jaw?

Table 2: QG examples with various pretrained T5 models

The examples in table 2 illustrate this weakness: The possibility to take influence, for example on the question type, is small. The models tend to generate "what"-

questions and it is hard to countersteer this preference. Some models can be outwitted by adding a token to the input sentence like the word *because* at the beginning. Sometimes that makes them generate a "why"-question, but this approach is not attractive due to its experimental character and also because it is limited. For many models is not possible, for example, to provoke a "when"-question with the sample sentence, even when an expression that indicates unambiguously an adverbial of time, like *during the night*, is added.

In this regard, the BART-HLSQG model stands out from the others. HLSQG stands for highlighted sequential question generation. What makes the BART-HLSQG model interesting for the MDA, is the possibility that a part of the input can be highlighted. This happens with so-called highlighting tokens [HL] before and after the given answer span [Chan and Fan, 2019]. This highlighting is intended to "avoid possible ambiguity in specifying answers for QG" [Chan and Fan, 2019, 155], but it turns out that at the same time it provides the option to greatly influence the question type based on a given answer. The Transformer-QG-on-SQuAD project[1] trained various sequence-to-sequence models on SQuAD (The Stanford Question Answering Dataset) for QG with the HLS (highlighted sequential) architecture on top. From the evaluated HLSQG models, it reports BART-HLSQG to outperform all the others. Some domain-specific experiments with the model gave the impression of solid performance regarding the produced questions, even without any finetuning.

The BART-HLSQG model can be tricked by adding certain terms manually to the highlight-part of a given answer before feeding it to the model pipeline for the follow-up question generation. The interrogative word *why* for example can be enforced, when the given answer is enhanced with the highlight *because*. For the most important interrogative words a reliable *trigger expression* was found. Reliable in a sense, that it did not fail in a single of over 50 experiments to trigger the desired wh-question. Table 3 lists the most important question words and their corresponding trigger cues.

The triggers have to be carefully selected, namely as neutrally as possible. For example the term *My jaw* would perfectly trigger the what-question. But at the same time, it leads the model to ask something about the jaw with a certain probability, as if the patient had written something about their jaw. This may happen to be the case by accident. But rules should make the behavior of the agent more predictable and not more arbitrary as it already is.

[1] `https://github.com/p208p2002/Transformer-QG-on-SQuAD`

Question word	Trigger term
Why	Because
When	During
Where	At the hospital
What	Something
Who	Somebody
How	In form of

Table 3: Best trigger term for most important question words

3.2.6 Training data

For finetuning pretrained language models and training the topic model, data from the WISE data set is used. It contains survey questions and the answers from patients, suffering from orofacial pains, that filled this survey. The data encompasses about 600 questions and subquestions, with answers from 2840 patients. The predominant answer language is German.

For the topic model training and the German-to-English translation model finetuning, the answers on a small subset of the questions are taken into consideration:

1. Please describe your chief complaint for which you seek consultation.
2. What does your chief complaint stop you from doing?
3. What do you expect as the result of the examinations and treatments in our clinic?
4. Which factors aggravate your complaints? (e.g. chewing hard / soft food, biting, drinking, mouth opening (e.g. yawning), talking, physical / emotional stress, playing a musical instrument, ...)
5. Which factors alleviate your complaints? (e.g. distraction, rest, relaxation, keeping jaw in fixed position, ...)

Those questions require free-text answers from the patients. The reason, why the selection was reduced on such a small subset from all the questions is that usually free-text is used for finetuning language models. Data that origins in single and multiple choice questions could be useful, but would need additional formatting and restructuring to be valuable training data. This would exceed the frame of this work. And having a manageable size of data attributes, can make it easier to explain behaviors or dynamics of the MDA and analyze conversations.

For the English-to-German translation model finetuning the survey questions are used. Eventually the question generation model could be finetuned with those questions. But for this work, this was not done. The number of questions is rather small so it can be questioned how valuable a finetuning would be. Instead, some questions have been picked out and used as predefined questions (see chapter 4.2.4).

For the conversation, only English-trained language models are used. So, the data that is foreseen for the finetuning pretrained models was translated to English. Google Translate is considered as one of the best performing machine translation software that is currently available [Alsan, 2022]. It however requires that the data is uploaded to the Google cloud storage for the translation process. As the patients answers represent sensitive data in terms of privacy protection, they have to be anonymized first. With named entity recognition (NER), sensitive patient data was detected. The anonymization itself was done manually, because the NER returned a relatively high quote of false-positives, like medical (latin) terms, for example. In addition to that, there was no desire to replace every named entity by a generic term. Medicament names, for example, should be kept. Table 4 contains a list of entities that were replaced.

Sensitive object	Replace expression
Name of patients	die Patientin, der Patient
Medical staff	Doktor, Arzt, Ärztin
Places (villages, towns, districts)	XXX, YYY, ZZZ
Foreign countries	Ausland
Address	None (delete)

Table 4: Replacement of sensitive data

In the course of the anonymization, also the most problematic orthographic mistakes could be revised. The focus was on terms that have high importance for the orofacial pain domain, like the term *Schmerz*, that was sometimes spelled *Schmertz* or *Kiffer* instead of *Kiefer*. Table 5 lists the most severe orthographic mistakes that were detected and corrected.

The Google translation performs well. Even ambiguous terms (like *Kiefer*, which was translated to *pine* instead of *jaw* by lower-quality MT systems), are translated correctly by Google Translate, possibly thanks to the attention on the orofacial context.

Term wrong (case insensitive)	No of occurrences in document	No of occurrences in 5 main columns
tinitus	76	76
artzt	35	6
kiffer	17	13
schmertz	12	10
kiever	8	8
scherzhaft	3	3
tintius	2	2

Table 5: Most severe orthographic mistakes

For the training of the topic model, the data has been preprocessed furthermore with punctuation and stopword removal, tokenization, and lemmatization.

3.3 Rulebased tasks

The rulebased components of the system are responsible for a coherent and fluent conversation by orchestrating the intelligent components. Besides, the maintenance of predefined conversation elements and some preprocessing steps are rulebased. And finally, the repositories that memorize past answers and generated questions and the file archiving of the conversation follow hardcoded rules as well.

3.3.1 Predefined conversation elements

The system allows to predefine questions for user profiling, questions specific to the orofacial pain domain, questions asking for more details, and empathic phrases. First, the system runs sequentially through all the predefined user profiling questions. After the user profiling is completed, one hardcoded question with a transitional character is asked. This marks the transition from the finite-state component to the frame-based component. From this point on, the system does a probabilistic topic analysis and inference for every user input. When the computed relative weight of an inferred topic exceeds a threshold of 4.0, and there is a question defined for this topic number that has not been asked yet, this question is selected as the next question. More than one question can be predefined per topic number. If there is no topic question to ask, the system initiates the agent-based QG.

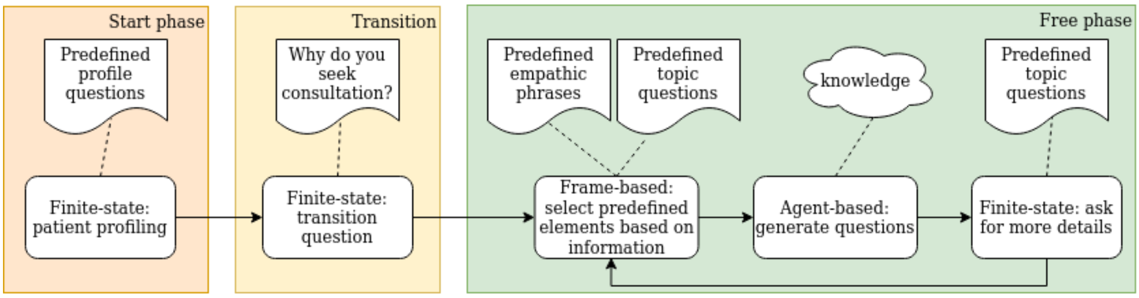

Figure 3: Usage of questions according to conversation phase

Empathic phrases are set up in a similar way like the domain specific questions. But they come together with a mental state, instead of a topic. The mental states are predefined as well and the system uses them as candidate labels in the sentiment classification step. There is no hardcoded set of candidate labels for the classifier. They are read from the respective file and are subsequently employed as candidate labels.

When the system has no topic question found and generates an answer that has already been used previously, a question asking for more details is randomly selected and outputted to the user. These more-detail questions are allowed to be used multiple times.

Details about how the predefined elements have to be set up, before the system can be used, are described in the chapter 5.2. The system is set up with some predefined default elements, but they can be changed according to the needs. Details and recommendations about manipulations on the MDA configurations can be found in chapter 5.2 as well.

3.3.2 Preprocessing

Except from the answers to the profiling questions, every answer given by the user is preprocessed. Steps, like stopword removal and punctuation removal are rulebased. The patient's answer has to be switched from first person singular to second person singular. Without that, the BART-HLSQG would generate a question in first person singular, what is obviously inappropriate. The system should not ask *When did **my** toothache start?*, but *When did **your** toothache start?*. This interference is relatively simple thanks to same verb forms for both persons, except from the verb *to be*. With the spaCy morphology feature, pronouns in first person and verbs can be detected. The below mapping table shows, how the tokens are replaced.

First person sentence	Second person sentence
am, 'm	are, 're
was	were
I, we	you
my, our	yours
me, us	you
mine, ours	yours

Table 6: Map first person to second person rules

3.3.3 Runtime memory and data archiving

The system is required to maintain the conversation data inventory. Before the conversation starts, all the available conversation data is loaded into dedicated repositories which are available during the entire lifecycle of the conversation. When a predefined element is used for the conversation it has to be updated with the number of its usages so far. Certain elements, like predefined topic questions are allowed to be asked only once. Empathic phrases and more-detail questions can be used multiple times. Each question appearing in the dialogue is saved in memory as are the given answers. After every conversation turn, the corresponding question and answer are physically persisted in an archive file. Any dialogue the system ever conducted is retraceable.

4 Modeling Dialogues in Health Care

4.1 Research

A dialogue between humans is perceived to be good and informative when it fulfills certain criteria. But what is valid for an interaction between humans is not necessarily one by one applicable for a human-computer interaction, found Clark et al. during their investigations about "what people value in conversation and how this should manifest in agents". A human to human conversation demands for "mutual understanding and common ground, trustworthiness, active listening and humour". However the aspect of having a common ground for example in a human-computer interaction was assessed rather negatively. "I'd be quite upset if I thought that I had a common ground with a computer.", was a statement of one of their interviewees.

The same is valid for empathy, considered as a significant factor for a building up a trustworthy environment between humans. When a computer interacts with a human, in contrary, too much empathy can even be counterproductive according to Clark et al.. And regarding empathy, a third dimension needs to be taken into account: The health care domain demands a cautious use of empathy according to Halpern. They suggest in "What is clinical Empathy?" a "neutrally empathic physician", that "does what needs to be done without feeling grief, regret, or other difficult emotions". Too sympathetic physicians risk overidentifying with patients and in general emotional responses are seen as threats to objectivity. Nevertheless, Halpern emphasizes empathy as a significant factor for enhanced therapeutic efficacy. "Engaged communication has been linked to decreasing patient anxiety, and, for a variety of illnesses, decreasing anxiety has been linked to physiologic effects and improved outcomes".

Richards et al. conducted experiments with "empathic dialogue cues". They tried to evaluate how empathic cues are percepted by users. The results on how test persons assessed the expressions are a bit sobering but also controversial. Some test persons classified statements as "stupid" while others found them "helpful". The reasons why test persons classify one and the same statement in such different ways, are often complex and multilayered.

Studies about intelligent conversation systems report some recurring challenges. Van Pinxteren et al. reviewed and summarized literature that investigates human-like communication in CAs. A common assumption is that there is a correlation between a CA's ability to behave like a human and the quality of the conversation it conducts. And a common finding is that behavior in most human-like CAs is still far from realistic. Van Pinxteren et al. categorize criteria for human-like behavior by three modalities: nonverbal behaviors, verbal behaviors and appearance characteristics. It is the nonverbal behaviors and the appearance characteristics where a system can make the big difference in terms of establishing a relation to its counterpart according to Van Pinxteren et al. The verbal behavior category is limited in contributing to an empathic dialogue. However, they identified reactions, like affecting support and social praise or politeness as written means having a positive effect on the conversation. For some techniques it depended on the human users character or even on their gender, whether the reactions of the CA had a positive impact on the conversation. The effect of humorous elements, for example, was found to depend strongly on how the user liked this type of humour.

4.2 Realization in the MDA

4.2.1 Empathy

An important aspect of empathy, mentioned by various research papers, is the perception of the mental state between dialogue partners. However, in a language-based-only CA this is a challenge. There is no posture, no tone of voice, no facial expressions supporting the detection of mental states. Empathy relies entirely on the textual content. Furthermore, classifying text into mental states is one thing, but how to react appropriately? Empathic phrases should be used with caution. A careful elaboration of appropriate formulations requires a deep understanding of the conversation domain, the conversation participants and the context. Therefore, the MDA is equipped with an mutable inventory of mental states and suitable empathic phrases. This offers the possibility of adjusting the inventory of empathic statements, by a domain expert, before the MDA is used for clinical consultation assistance. Each empathic phrase is specified together with a corresponding mental state. Details about the default inventory of the MDA and about how mental states and corresponding empathic phrases can be added, removed, or changed, by a domain expert are described in chapter 5.2.

Richards et al. designed the empathic cues, they presented to their test persons, according to the recommendations of de Rosis et al.. Two of them are relatively easy to implement: namely a friendly self-introduction and a friendly farewell. Although the farewell does not contribute anything to an informative dialogue, it (hopefully) leaves the patient with positive feelings about the conversation.

Another feature that is recommended by Van Pinxteren et al., is to use the name of the dialogue partner. A very simple template form, allows the system administrator to put placeholders for the patient's name into any statment of choice, which will be replaced at runtime by the system with the actual name of the patient.

Finally, the implementation of empathy in a system that supports solely written communication, like the MDA, is limited. The establishment of a relationship between interlocutors, which is considered as a central aspect, is hard to implement. Important non-verbal signs, like voice, gaze, mimics, or gestures, cannot be integrated into the conversation. Those missing components prevent from designing a system with real ability for elaborate empathy.

4.2.2 Common Ground

Jurafsky and Martin describe common ground as the establishment of common agreements and speakers grounding each other's utterances. The latter means acknowledging that the hearer has understood the speaker. This can be by saying a simple "OK" or by repeating what the other person says. Such communication behaviors can be implemented in a rulebased form.

Utterance examples	Type
I see.	Acknowledgment
I understand.	Acknowledgment
You mention "...".	Repetition
That makes sense.	Confirmation

Table 7: Examples of utterances supporting common ground

4.2.3 Continutity and variety

Clark et al.'s interview participants pointed out that they expect a good CA to remember information to tailor their experience. To remember what was said in previous conversation turns is crucial for two major requirements. Firstly, in order to prevent repetition, which can be considered as a conversation killer. Asking the same thing multiple times or asking for things that a patient already mentioned may ruin a conversation. Second, memorizing information also helps to prevent mental leaps respectively to stay within a context during multiple conversation turns. Based on the present questions and answers in the WISE dataset, topics can be modeled which help the system keeping a red thread during the conversation.

Variety is not only important in context of preventing repetition, but also when it comes to gathering as much information as possible from the patient. One form of variety is to alternate between different wh-questions with the goal of achieving a more informative discourse.

4.2.4 Topic Coverage

For a task-oriented system that is responsible for conducting a dialogue to gather information, it makes sense to define certain topics that must be touched at some point during the conversation. Like for the empathic statements and mental states, profound domain knowledge is required for defining appropriate topics and questions. Therefore, the system equally does not come with a fixed inventory of topic questions. It offers the possibility to adjust the inventory of topics and related questions. For each topic, any arbitrary number of questions can be added. All the details about how the entries in the inventory can be manipulated, are described in the chapter 5.2.

4.3 Preconditions for meaningful question generation

A good conversation can only be established with a proper system setup and when the user complies with certain conversation rules. First and foremost, the predefined conversation elements are expected to be meaningful and their links to either corresponding mental states and topics should make sense. Furthermore, the user should give valid answers, ideally whole sentences. The system performs significantly better when users repeat the content of the last question in their answers. For the performance of the question generation model, it makes a big difference whether the question *Have you had headaches yesterday?* is answered with a simple *Yes* or with a complete sentence *Yes, I had headache yesterday.* It is not equipped with the ability to compose the previous question with the given answer for the next question generation. For a coherent conversation it is desirable if the user adheres to common conversation rules like reading the questions thoroughly and giving adequate and informative answers. Last but not least, the patient should not use hate speech or offend the MDA. There is no guarantee that the MDA would then remain polite and not start using bad words too.

5 Implementation Details

5.1 Data model

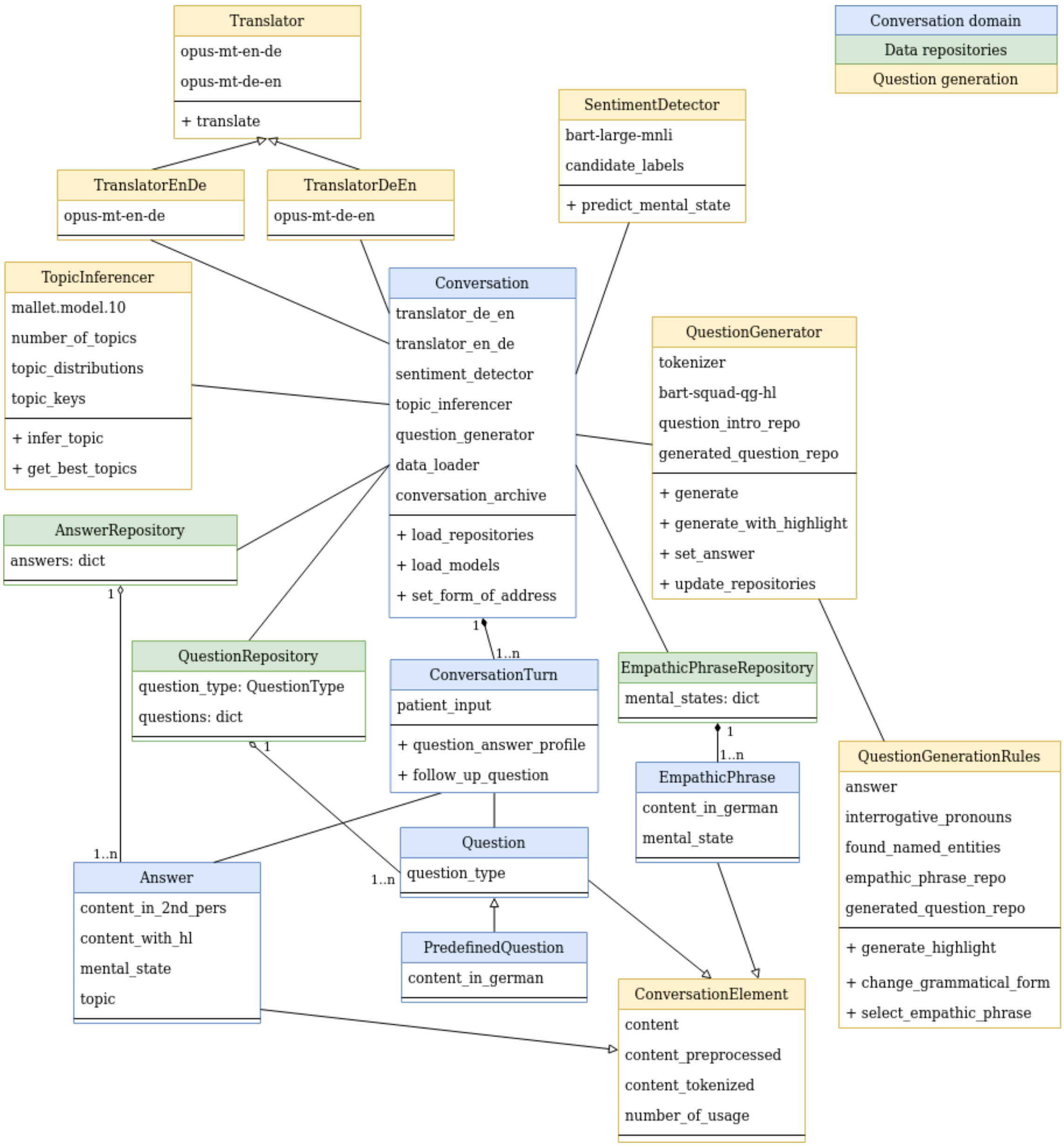

Figure 4: Conversation data model

5.1.1 Diagram

The relationships that have no cardinalities are understood as 1:1 relationships. Arrows represent inheritance, white diamonds represent aggregations and black diamonds compositions. The classes are coloured depending on their role. Blue represents the conversation domain classes, yellow all the NLP classes, green the data repository classes that are in charge of storing all the data over a conversation lifecycle. Only the public class attributes and methods are listed.

5.1.2 Classes

The above class diagram shows the most important classes that are required for modeling a conversation. When a user starts a dialogue, an object of the `Conversation` class is instantiated. This class is responsible for loading the language representation models and for reading predefined dialogue data from CSV files. It basically does all the computation intensive activities once at the beginning. During the rest of the conversation it is only there for having the conversation data ready, which is updated after every conversation turn. In contrary to the actual definition of a conversation turn, which stands for a single contribution of a dialogue participant, in the MDA an instance of the `ConversationTurn` class subsumes two turns, namely an answer as well as question.

The `ConversationTurn` class is the conversation orchestrating class. One object of this class is instantiated per each question and answer pair. Via the conversation object it has access to all the loaded models and repositories that are needed for creating the follow-up question based on a given answer.

The `ConversationElement` class supports different formats of textual content, that are used by all its subclasses. In the conversation itself the `ConversationElement` class has no role, only its subclasses. The `Answer` class contains further formats of the sentence that the user entered into the system, which are only needed by the answer domain objects. For example the variable `content_in_2nd_pers`, which represents the input sentence in second person, instead of the original sentence, which is assumed to be mainly held in first person.

5.2 Configurability

The data that the MDA uses in its rulebased parts, can be modified by an administrator with domain expertise, for example, a medical professional, that installs and maintains the system for their patients. Four files can be changed. All the entries that the four files contain, has to be available in English as well as in German.

profile_questions.csv: Questions that the system should ask for the patient profiling in the beginning of the conversation have be added to this file. Those predefined questions have to be provided in English and German.

Sample profile questions
What is your name?
How old are you?
Which gender do you belong to?

Table 8: Profile questions for conversation start phase

more_detail_questions.csv: Those questions are meant to be non-topic-specific. They should encourage the user to reveal more information. This type of questions comes into play when all the previous steps failed to create a valid follow-up question.

Sample questions for more details
I'm not sure, if I fully understood your answer. Could you explain in a bit more detail, please?
Would you prefer not to talk about this topic?
Did you understand the question or should I reformulate it?

Table 9: Questions asking for more details

mental_states_with_empathic_phrases.csv: Mental states may trigger empathic phrases. Mental states can be defined together with appropriate empathic phrases in this file. When the system detects a mental state with a specified degree of certainty in a patient's answer, it will select one of the empathic phrases and output it together with the follow-up question.

The system is by default set up with emotional mental states. The list can also be customized with any other mental states than emotions. However, it is recommended to configure mental states with a more or less normal semantic distribution. This means, that every mental state should be clearly distinguishable from the others.

Mental state	Empathic phrase
Sad	I'm sorry about that.
Happy	That sounds positive.
Afraid	Don't be afraid.
Angry	That must be frustrating.

Table 10: Sample empathic phrases

The system may, for example, perform badly for the following set of emotional states: *sad*, *unhappy*, *disappointed*, *happy*. A better choice of emotional states would be: *sad*, *happy*, *afraid*, *angry*. The second distribution is inspired by the emotional states on Robert Plutchik's emotion wheel [Mohsin and Beltiukov, 2019].

questions_for_topics_x.csv This file contains predefined questions in German and English for x topics. It is possible to store an arbitrary number of questions per topic together with the corresponding topic number from the topic model. Questions should only occur once, but it is possible that a question is never asked in a conversation, in case a topic is never inferred from the answers of the patient with a sufficient high degree of certainty.

Topic number	Question in English
0	When was the first time you experienced these complaints?
0	How often do you experience these complaints?
6	Which factors aggravate your complaints? (e.g. chewing hard / soft food, biting, drinking, mouth opening (e.g. yawning), talking, physical / emotional stress, playing a musicial instrument, ...) (WISE survey question)
6	Do you remember how your complaints started resp. what triggered them?
8	In which area does it hurt the most?

Table 11: Sample topic questions for highest weighted topics

mallet_topics: The topic model that is set by default in the MDA, was trained on about 2500 WISE data records and contains 10 topics. It can be replaced by another topic model. This may be valuable from time to time, when WISE has more data available for training a better topic model or when the MDA should work with more or less than 10 topics. But whenever a topic model is replaced, also the file with the topic questions has to be revised with the new topics and when the number of topics is changed, the `TopicInferencer` object in the `Conversation` class has to be instantiated with the adjusted number of topics.

Topic number	Topic keywords
0	get pain time feel like year day sometimes problem bad since would help always anything often lot could make know
6	pain chew food hard jaw eat long keep talk still bite chewing yawn speak right mouth biting time relax joint
8	pain right left sidetooth jaw cause upper pressure ear low severe find treatment leave face area sometimes head root

Table 12: Highest weighted topics in topic model, trained on WISE

The technical details about how the files can be modified, are described in the README[1] of the project.

5.3 Steps per conversation turn

Figure 5 visualizes the most important substeps within one conversation turn. Questions the system asks can be profile questions, when the conversation is in the start phase, where the patient profiling is done. Alternatively, the questions can be derived from a topic that is inferred from a given answer. For profile questions all NLP tasks are skipped. These questions just have to be stored in the conversation file and can be outputted. The answer does not require any special treatment neither, it ends up directly in the conversation file too. Not even machine translation is required, the system is designated to read and hold all predefined elements in two languages.

In the free conversation phase, every answer from the patient goes through various NLP steps. The preprocessing consists of creating multiple representations of the same content and storing them in the `Answer` object.

In the next step, the answer is classified into mental states. The zero-shot classification pipeline from Huggingface accepts a sentence and list of candidate labels and returns the probability of each of the labels for the given sentence. The pipeline is called with the `bart-large-mnli` model. The candidate labels correspond to the different mental states.

Next, a topic inference is done for the preprocessed answer text on the base of a

[1] `https://github.com/salwil/conversational_agent_for_medical_dialogues/blob/main/README.md`

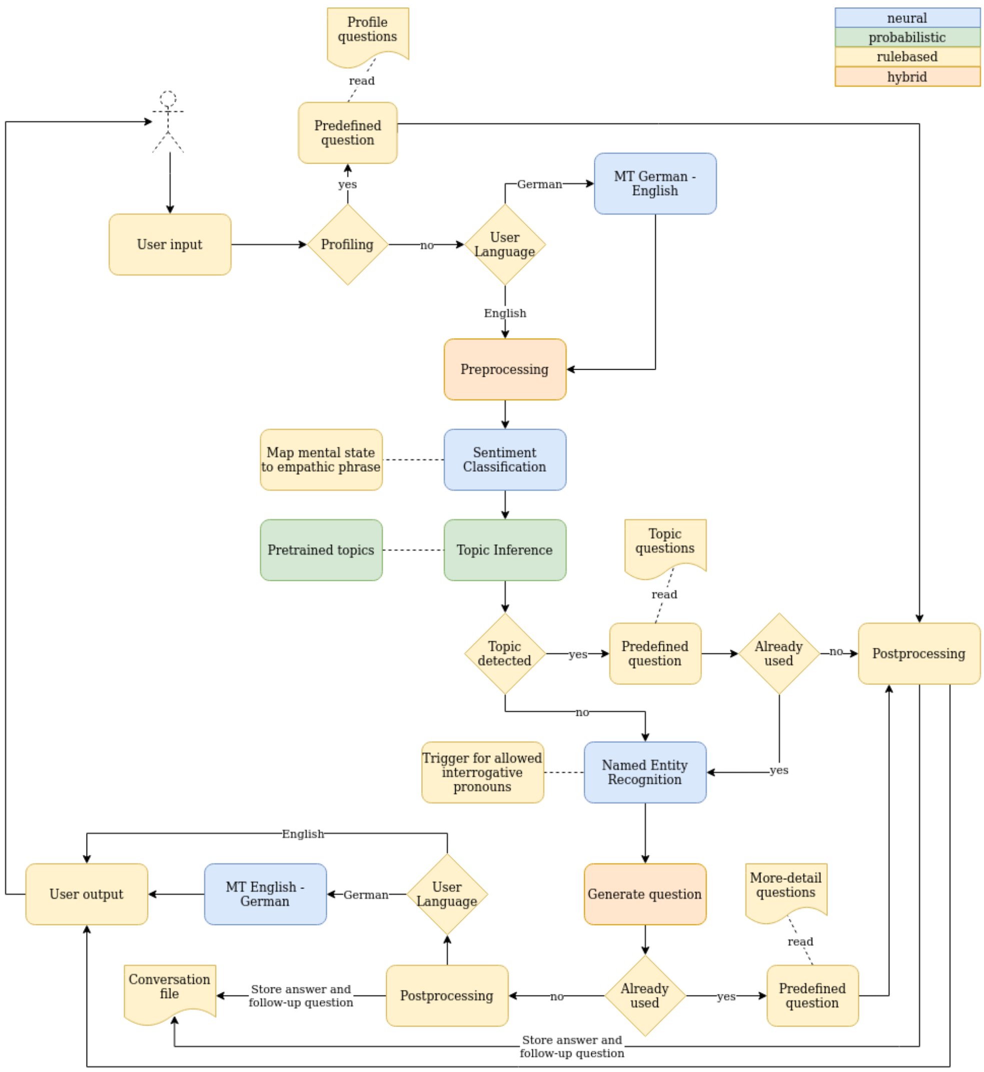

Figure 5: Steps per conversation turn

topic model pretrained on the WISE survey data. If a topic is inferred with a certain probability (the thresholds are explained in chapter 5.5), the list of associated topic questions is scanned. If there is a question that has not been asked yet, the system selects and persists it and outputs it to the user.

If no topic was inferred with the minimal required threshold, or when all the questions related to this topic have been asked already, it comes to intelligent question generation. Every third conversation turn, a named entity recognition (NER) is performed as part of the QG. This is needed for setting a highlight, that again triggers a specific interrogative word in the next question. The highlight generation is not performed in every conversation turn, because the system should not permanently oversteer the QG with the same rule. Otherwise, interesting or creative questions

Variable name	Content
\content	Original user input
\content_preprocessed	Input lemmatized, punctuation and stopwords removed (except question words)
\content_tokenized	User input tokenized
\content_in_2nd_pers	User input switched from first pers. sing. to second pers. sing.
\content_with_hl	User input extended with a highlight that enforces a certain wh-question

Table 13: Representations of user's answer

that the model creates by itself could be missed. The system generates a question based on the input answer. When the generated question passes the syntactical duplicate check, the previously selected empathic phrase and the generated question are composed into one string, and outputted. When the duplicate check fails, one of the predefined questions that ask for more details is selected at random and printed out.

5.4 Integration of Language Representation Models

The language models provided by Huggingface, come with corresponding pipelines, that make the usage of the respective models quite easy by abstracting away the complex code [Huggingface, 2022]. The pipelines load models either directly from the Huggingface hub or from a local file. The finetuned MT models have to be loaded from the local machine. For topic inference with the commandline based MALLET, each shell command is encapsulated in a separate method. Some of the methods are inspired by the Little Mallet Wrapper[2] and slightly adapted to the purpose of the MDA. For the NER, morphological features and lemmatization, the `en_core_web_sm` pipeline is loaded with the spaCy top level load function.

5.5 Ruleset for element selection

Candidate mental states have to be predicted with a certain probability to trigger an empathic phrase. If this threshold is not reached, the mental state of the current conversation turn is set to *neutral*.

[2] https://github.com/maria-antoniak/little-mallet-wrapper

Predefined empathic phrases can contain `[address form]` tags. If such a tag is detected at runtime, it is replaced by the patient's name.

The topic question selection, follows a sequence of multiple steps: infer topics, pick the three highest-weighted topics, check their weights against a minimal threshold and verify, whether there is still a predefined question available to this topic, that has not been asked yet. If no topic question was found due to the failure of any of the criteria, the system generates a question.

An important rule is to have no repetition. To ensure this, a preprocessed version of the questions is hashed and stored in a dictionary data type container which can only contain unique values. The preprocessed string is lemmatized and has neither punctuation marks nor stopwords, except from the question word. Before a follow-up question is saved and displayed to the patient, uniqueness is verified by scanning the dictionary for the respective question. Duplicate questions can be detected when the hashvalues of their preprocessed strings are the same. This check is of syntactical nature, semantic ambiguity with different wording cannot be detected with it. When a duplicate error is raised, the system forgets about the generated question and picks a predefined question from the more_details_repository instead to ask for more details.

6 Results and Discussion

Measuring the performance of a conversation agent effectively and systematically is difficult, because of a missing gold standard. Both Laranjo et al. and Car et al. manifest the missing systematic evaluation of conversational agents in healthcare. Especially for CA with unconstrained natural language input capabilities, the "few published studies are mainly quasi-experimental, and rarely evaluate efficacy or safety." [Laranjo et al., 2018]. A conversation, whether it is assessed to have good or bad quality, depends on the user's perception. One user may find an utterance helpful, another may find it stupid. Nevertheless, the MDA implemented in the course of this thesis has to be evaluated against the criteria defined for this thesis. The quality assessment emphasizes basically the answer to one question: **Is the MDA able to conduct an informative conversation?** The specified research questions from chapter 1 subdivide this broad question into smaller, more measurable parts:

1. Can the MDA create a trusted and secure conversation environment by behaving in an empathic manner that encourages the human user to reveal information about them?
2. Is the MDA able to ask meaningful follow-up questions on a given answer?
3. Are the chosen tools, techniques and language models well-suited for meeting the requirements towards the MDA?

6.1 Empathy

Chapter 4 shows that empathic behavior in a written-only environment is hard to achieve. Empathy is closely linked to mimic and gestures. Nevertheless, possibilities to create a trusted and secure environment exist. Written messages can contain utterances, promoting common ground or showing empathy. The elaboration of appropriate utterances for a certain situation, requires expertise. The MDA is equipped with some incidentally (and insufficiently) designed mental states and corresponding empathic phrases. But it is implemented in a way, that they can be easily changed and enhanced as required by a domain expert.

The MDA asks meaningful follow-up questions as long as it stays in the finite-state or frame-based components. The patient profiling questions in the beginning of the conversation are asked sequentially until there are no more questions in the inventory. Under the assumption, that the preconditions, formulated in chapter 4.3, hold, every profile question is meaningful. The system simply works towards the finite state of having a complete patient profile in this phase of the conversation. After the profiling phase and the transition question, asking meaningful questions becomes more challenging for the MDA. When it is able to infer a topic and also finds a corresponding question that has not yet been asked, it is likely that this question is also meaningful. Like for empathic phrases, the quality of the topic questions depends on how much was invested in the elaboration of the questions and on the expertise of the person, doing it. The MDA is not able to notice when a topic is linked to inappropriate questions. It will just pick one of them from the list and present it to the user as follow-up question.

In the agent-based component of the system, the guarantee of a meaningful question decreases. Except from taking some influence on the question type, in every third conversation turn, to ensure a certain degree of variety, the MDA gives up all control on the question generation. The generated questions are not subject to any semantic checks. At a rather simplified syntactical level only, the MDA ensures, a question is not asked twice.

6.2 Testing frame- and agent-based QG

6.2.1 Setting

Beside thorough semantic testing during the development, two unbiased persons without any domain knowledge tested the MDA. They were told to reply to the questions like a person with orofacial complaints possibly would. They had a short introduction, based on a German text about orofacial complaints taken from Ettlin and Galli (appendix, C.1). The test persons were familiarized with the preconditions, formulated in chapter 4.3. Without any further guidance, they were left alone with the MDA for half an hour. The conversations were recorded and stored by the system. All the conversations are fictional.

6.2.2 Test results

The conversations 1 (table 14), 2 (table 15), 3 (table 16), and 6 (table 19) were hold with test persons. The conversations 4 (table 17) and 5 (table 18) were conducted by the author. Only the frame- and agent-based parts of the conversations are listed. The dialogue always starts with the hardcoded transition question: *Would you introduce me briefly why you seek consultation?* Some conversations were hold in German, but the system only records the English questions and answers.

Conversation 1 (14) shows the weakness of the system towards very short answers. The conversation 2 (15) however shows issues when the user repeats the content of the previous question. In that dialogue the system apparently ignores the negation in the user's answers about loud noises in turn 6, 7, and 8. Conversation 3 (16) shows that a meaningful conversation can be kept going for relatively many turns, when the user provides detailed answers. Due to the strong relative weights of some inferred topics, the first half of the conversation is typically characterized by the usage of the same questions. Those topics are very likely to be computed by the system, even if the input is not strongly related. Conversation 4 (17) is mainly an experiment to demonstrate how the system falls back on a more-detail question, when a generated question is duplicate. Questions for more details are asked in turn 9 and 10. In turn 11 the system manages to create another not yet asked question, despite the repeated input. From turn 12 on, the creativity of QG has finally reached its limit and the system tries to get the user to provide more details by printing one question from the more-detail question repository after the other. Conversation 5 (18) is an example of a not too bad conversation. What can be observed here in particular from turn 8 to 11, is how the system comes back on track thanks to a topic inference after having lost the orientation. The answer in turn 10 infers a new topic, which triggers a predefined topic question that gives the conversation a new content direction. Conversation 6 (19) is another example for a conversation that stays for the most part in the frame-based boundaries. The turns 2, 3, 4, 5, 6, 9, 12, 13 are based on predefined topic questions.

The questions in turn 2, 3, and 4 are very often the same. These questions are not associated to the highest weighted topic in the topic model, but they are chosen based on their high scores resulting from the computation of the relative weights of the inferred topics. The difference between the system behaviors when the real inferred weights were used or the relative weights could not be verified systematically. Often the selection of the question is accurate in the course of the conversation and prompts informative answers.

The deficient handling of empathic phrases leaps to the eye across all the conversations. For real-world usage, the empathic phrases should either be elaborated very carefully or be omitted.

6.3 Evaluation of methods and tools

For the scope of this work, the implemented tools in form of pretrained language models, frameworks, and rules, did an appropriate job. However there are many limitations. In this last section the four most and the two least promising components of the system are described.

The MALLET topic inference based on a pretrained topic model reveals promising results. Despite the very little elaborated environment of the MDA, the topic inference has a positive impact on the red thread of the conversation. With intelligent rules, eventually more topics, better elaborated topic questions and at some point also better trained topic models, the topic analysis component can be of great value in a future stage of development. It was mentioned previously that the current topic model has very unbalanced weights. The approach with the relative weight calculation does not have an obvious positive effect. For example the topic number 2 was inferred for the sentence *I make a fire in the garden* with a relative computed weight of 6.288. The threshold for a topic to be selected, lies at 4.0.

The BART-HLSQG model is probably one of the best performing QG open source models that are currently available. The outputs are solid and even in cases, where the content misses the target topic, the questions are complete sentences and rarely contain grammatical or orthographic mistakes. This was not the case in experiments with other text generation models.

The decision to construct the system in English and to add a machine translation step at the beginning and the end of every conversation turn, in case of a German conversation, was right. A negative impact of the translations on the sentence quality is limited, especially, because the generated questions are normally rather simple English. And all the predefined questions are available in both languages anyway, so there is no MT required. The advantages predominate clearly. It turned out that not only English pretrained language models are performing better, but also other tasks were easier to solve in English. First and foremost the sentence conversions from first person to second person form. In German that would have been an impossible undertaking.

The value of highlighting some parts of the input for the BART-HLSQG model in the MDA is remarkable. However, for another system architecture another model, for example, one that produces multiple possible questions for a given input, may be more suitable.

The employment of the BART-MNLI model to predict a mental state from a sentence is nice. The predictions are good, despite the a zero-shot approach. However it is questionable, how much value the knowledge about the predicted mental state really has for the MDA. The current set up with the predefined empathic phrases is very experimental. A profound analysis of mental states and appropriate reactions is essential, if empathy should be part of the conversation. But that demands a considerable effort. Given the divided views whether a conversation system needs to show empathy at all, it is debatable if further developing of the empathy feature makes sense.

The need for finetuning the English-to-German MT model can be questioned. A dataset with about 350 sentences in both languages is too sparse. In the conversations no obvious effect of the finetuning could be observed. The MT would possibly perform equally good with the employment of the pretrained model from Helsinki-NLP.

7 Conclusion

The elaboration and implementation of the MDA revealed a number of interesting insights into the development of conversational agents. First and foremost the enormous potential of the current state of the art technologies and at the same time the huge challenges to overcome when building a high-quality QG system. In view of the endless possibilities to build and shape a medical dialogue agent, the highest priority was to design a robust system that is open for change. With the dialogue data configuration that can be done without touching a single line of code and the modular design of the entire application this requirement is met. The modular structure allows to add new functionality or replace existing logic easily. Existing functionality should only little or not be affected at all, by the implementation of new features, when the SOLID principles of object-oriented programming are followed.

The test suites, consisting of nearly 50 unit and integration tests, serve as documentation of the implemented classes and methods. And they can be used as regression tests for future changes, to ensure new features don't break existing ones.

There are many extensions that could be implemented. Besides the aspects that have been mentioned already in chapter 6, there is room for improvement in the entire application. Without even touching any code, the system's performance could be improved by adding more elaborated predefined conversation elements.

User and administrator friendliness could be improved in various aspects. The system could be integrated into a browser application for example. The system could archive the conversations in a more suitable format. A format that makes questions and answers easier to use for further language models finetuning. Multithreading could be implemented in the start phase of the dialogue: meanwhile the system loads the models, it could already start the dialogue with the finite-state patient profiling. Another improvement would be, when the system was able to remember, what it asked in the previous conversation turn, so that the users do not have to repeat the content of the previous questions in their answers.

The language representation models, currently in place, could be exchanged by other models. The new spaCy Transformers API could be used instead of the "traditional" spaCy NER, for example. The models could be finetuned on more data. In general, this work did probably not exploit the full the potential of the WISE datasets. It restricted to the answers of five columns. A CA that has more sensibility for not only physical but also psychological aspects of the illness may be sensitised on multiple choice questions like:

- **In the past 4 weeks, how often did you feel affected by the following complaints?**
  - Feeling nervous
  - Not being able to stop or control worrying
  - Feeling down
  - Difficulty falling asleep or sleeping through the night (...).
  - ...

The integration of topic models should definitely be pursued and further developed. A systematic comparison of different topic inference approaches would help to improve the topic analysis. Topic models that are trained on more data together with well-designed questions for each topic are expected to have a positive effect on the conversation quality.

Glossary

agent-based Methods in agent-based systems are "typically statistical models trained on corpora of real human-computer dialogue" [Laranjo et al., 2018] Agent-based dialogue systems respond dynamically to the course of the conversation.

auto-regressive (language model) Autoregressive language model architectures resemble to RNNs but unlike RNNs they don't employ hidden states for previous steps, but just feed multiple inputs together (e.g. a sequence of tokens) to the model [Ho, 2019]. There is no backpropagation mechanism, i.e. they're actually feed-forward models. This type of models is convenient for text generation tasks.

chatbot Simplest kinds of dialogue systems that can carry on extended conversations with the goal of mimicking the unstructured conversations or 'chats' characteristic of informal human-human interaction [Jurafsky and Martin, 2021, 24.2].

conversational agent Jurafsky and Martin distinguish slightly between a CA and a dialogue system. A CA can "answer questions on corporate websites, interface with robots, or be used for social good." In this thesis CA and dialogue agent and dialogue systems are treated as synonyms.

corpus-based system Corpus-based systems, instead of using hand-built rules, mine conversations of human-human conversations for retrieving information from input and for generating responses. [Jurafsky and Martin, 2021]

dialogue system / agent Jurafsky and Martin distinguish slightly between a CA and a dialogue system. A "dialogue agent in digital assistants (Siri, Alexa...) give directions, control appliances, find restaurants, or make calls." In this thesis CA and dialogue agent and dialogue systems are treated as synonyms.

decoder (model) An auto-regressive model: the decoder outputs one token at a time based on an input and a preceding output token.

encoder-decoder (model) Encoder decoder models generate tokens by deviating an output (e.g. a sequence of tokens) from an input (e.g. a single token). The encoder usually operates with a recurrent neural network, while the decoder decodes states in feed-forward fashion.

finite-state In terms of conversational agent, finite-state refers to a system that takes the user "through a dialogue consisting of a sequence of pre-determined steps or states" [Laranjo et al., 2018] Such dialogues work towards a clear goal. When the goal is reached, the involved sequence of tasks and objects are considered to have reached a finite state.

frame-based A system where "the dialogue flow is not pre-determined but depends on the content of the user's input and the information that the system has to elicit". Frame-based systems address weaknesses of finite-state systems, by relinquishing on following a predefined order to fill in the required fields [Laranjo et al., 2018].

open-domain dialogue agent A CA designed to maintain a conversation about topics originating from many domains. Such systems normally do not have a clear goal or task to accomplish.

question type In this work the term question type is used as taxonomy to distinguish between the different wh-questions, namely questions that begin with *what*, *when*, *where*, *who*, *whom*, *which*, *whose*, *why*, and *how*.

rule-based system A system that processes input and creates output depending on predefined rules. Every action of the system can be explained based of the implemented ruleset.

sequence-to-sequence (model) Models with seq2seq architecture have sequences (of words) as input and produce sequences (of words) as outputs. First the network produces an internal state by encoding the inputs, and then it produces an output by decoding the input according to a given task (e.g. translation into another language)[Ibanez, 2022].

(dialogue / conversation) turn Each contribution of a dialogue participant is called a turn. In the MDA system a conversation turn, however, consists always of an answer and a question.

task-oriented dialogue agent A CA designed for a specific task that it has to fulfill.

transformer model Transformer architectures "eschew recurrence and instead rely entirely on an attention mechanism to draw global dependencies between input and output." [Vaswani et al., 2017]. That means a transformer is nothing else than an autoregressive model, as its decoder ”processes an entire input all at once” [Wikipedia, 2022], instead of processing tokens of an input sequence one after the other.

zero-shot classification In this thesis this term is used when a classification model is applied on a task it has not been explicitly trained before. For example, the MNLI sequence-pair classifier was trained to predict whether a hypothesis is true given a premise. The MDA uses it for predicting a candidate label for a given sentence.

References

M. Alsan. The best machine translation software you can try in 2022, 2022. URL `https://weglot.com/blog/machine-translation-software/`.

L. T. Car, D. A. Dhinagaran, B. M. Kyaw, T. Kowatsch, S. Joty, Y.-L. Theng, R. Atun, et al. Conversational agents in health care: scoping review and conceptual analysis. *Journal of medical Internet research*, 22(8):e17158, 2020.

F. Catania, M. Spitale, D. Fisicaro, and F. Garzotto. Cork: A conversational agent framework exploiting both rational and emotional intelligence. In *IUI Workshops*, 2019.

Y.-H. Chan and Y.-C. Fan. A recurrent bert-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162, 2019.

L. Clark, N. Pantidi, O. Cooney, P. Doyle, D. Garaialde, J. Edwards, B. Spillane, E. Gilmartin, C. Murad, C. Munteanu, et al. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

J. Davison. Zero-shot learning in modern nlp, 2020. URL `https://joeddav.github.io/blog/2020/05/29/ZSL.html`.

F. de Rosis, N. Novielli, V. Carofiglio, A. Cavalluzzi, and B. De Carolis. User modeling and adaptation in health promotion dialogs with an animated character. *Journal of biomedical informatics*, 39(5):514–531, 2006.

D. A. Ettlin and U. Galli. Orofaziale schmerzen. *dolor: Schmerztherapie in der Praxis*, 8(1):1–4, 2008.

D. A. Ettlin, I. Sommer, B. Brönnimann, S. Maffioletti, J. Scheidt, M.-Y. Hou, N. Lukic, and B. Steiger. Design, construction, and technical implementation of a web-based interdisciplinary symptom evaluation (wise)-a heuristic proposal for orofacial pain and temporomandibular disorders. *The Journal of Headache and Pain*, 17(1):1–15, 2016.

A. R. Fabbri, P. Ng, Z. Wang, R. Nallapati, and B. Xiang. Template-based question generation from retrieved sentences for improved unsupervised question answering. *arXiv preprint arXiv:2004.11892*, 2020.

S. Graham, S. Weingart, and I. Milligan. Getting started with topic modeling and mallet, 2002. URL `http://programminghistorian.org/en/lessons/topic-modeling-and-mallet`.

J. Halpern. What is clinical empathy? *Journal of general internal medicine*, 18(8): 670–674, 2003.

G. Ho. Autoregressive models in deep learning - a brief survey, 2019. URL `https://www.georgeho.org/deep-autoregressive-models/`.

Huggingface. Pipeline docs, 2022. URL `https://huggingface.co/docs/transformers/main_classes/pipelines`.

D. Ibanez. Encoder-decoder models for natural language processing, 2022. URL `https://www.baeldung.com/cs/nlp-encoder-decoder-models`.

D. Jurafsky and J. H. Martin. *Speech and Language Processing*. Prentice Hall, 3rd, draft edition, 2021.

L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. Lau, et al. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258, 2018.

X. Ma. Dimensionality-reduction with latent dirichlet allocation, 2019. URL `https://towardsdatascience.com/dimensionality-reduction-with-latent-dirichlet-allocation-8d73c586738c`.

D. Mimno, C. Sutton, G. Chandalia, and A. Hough. Mallet: A machine learning for language toolkit, 2002. URL `http://mallet.cs.umass.edu`.

M. A. Mohsin and A. Beltiukov. Summarizing emotions from text using plutchik's wheel of emotions. In *7th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2019)*, pages 291–294. Atlantis Press, 2019.

D. Richards, A. A. Bilgin, and H. Ranjbartabar. Users' perceptions of empathic dialogue cues: a data-driven approach to provide tailored empathy. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 35–42, 2018.

spaCy. Linguistic Features named entity recognition, 2022. URL https://spacy.io/usage/linguistic-features#named-entities.

M. M. Van Pinxteren, M. Pluymaekers, and J. G. Lemmink. Human-like communication in conversational agents: a literature review and research agenda. *Journal of Service Management*, 31(2):203–225, 2020.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wikipedia. Transformer (machine learning model), 2022. URL https://en.wikipedia.org/wiki/Transformer_(machine_learning_model).

W. Yin, J. Hay, and D. Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*, 2019.

A Test Conversations

1	Would you introduce me briefly, why you seek consultation?	My nose is hurting while sneezing
2	At what time of the day is the pain the worst?	In the morning
3	I see. In the morning	Yes
4	What a surprise. During what event did the question arise?	During dinner last night
5	What treatment did you have so far?	Pain killers
6	What do you expect as the result of the examinations and treatments in our clinic?	To ease the pain
7	Do you also have cold symptoms? If yes, which?	Fever
8	Ok. During what stage of the body do you get sick?	Running
9	Don‘t be afraid. It soon will get better. What is one way to run?	Fast and warm
10	Are you often exposed to loud noises? If yes, what type of noises?	yes, construction work
11	What a surprise. How often is construction work done?	Daily
12	Ok. What is the Daily	Every day
13	Alright. How does each day occur?	Loud
14	What a surprise. What is another name for Loud?	Noisy
15	Mhmm. What is another name forisy?	

Table 14: Test conversation 1

1	Would you introduce me briefly, why you seek consultation?	I'm here because my molar hurts when I'm doing sports
2	At what time of the day is the pain the worst?	The pain is worst between 10:00 and 11:00
3	What treatment did you have so far?	I haven't had any treatment so far
4	What do you expect as the result of the examinations and treatments in our clinic?	I expect from your clinic that the pain will go away when you exercise
5	Do you also have cold symptoms? If yes, which?	No, I don't have any symptoms of a cold
6	Are you often exposed to loud noises? If yes, what type of noises?	No, I'm not exposed to loud noises
7	Let‘s think about how we can solve this. Are you exposed to loud noises?	No I'm not exposed to loud noises
8	We will clarify this. How long have you been exposed to loud noises?	I am not exposed to loud noises
9	We will clarify this. Are you exposed to loud noises?	Yes
10	What a surprise. How many questions were there?	10
11	I see. What is the name of the movie?	James Bond
12	What a surprise. What is the name of James Bond?	James Bond
13	What a surprise. How often does James Bond appear?	17 times
14	Ok. How many times has this happened?	q!

Table 15: Test conversation 2

1	Would you introduce me briefly, why you seek consultation?	I'm here because I've had acute jaw pain in the last few months.
2	At what time of the day is the pain the worst?	The pain is worst in the morning from around 4.30 a.m. The pain goes away in the morning.
3	What treatment did you have so far?	I once went to see my family doctor for treatment, who then sent me to you.
4	What do you expect as the result of the examinations and treatments in our clinic?	I expect my pain to be treated and gone.
5	Do you also have cold symptoms? If yes, which?	No, I don't have any symptoms of a cold.
6	Are you often exposed to loud noises? If yes, what type of noises?	I'm only exposed to loud noises when my son plays electric guitar.
7	What a surprise. When do you hear loud noises?	I hear loud noises twice or three times a week when my son plays electric guitar. Mostly in the afternoon.
8	Which factors alleviate your complaints? (e.g. distraction, rest, relaxation, keeping jaw in fixed position, ...)	A walk in the forest relieves my discomfort.
9	Do you often feel stress? What makes you feel stressed?	I have a lot to do at home and at work from Monday to Friday, so I often feel stressed. The main reason is a lot of work.
10	Let's think about how we can solve this. Who can cause you to feel stressed?	When my boss gives me a lot of work and we have time pressure, I feel stressed.
11	We will clarify this. What can cause you to feel stressed?	The many work causes me to feel stressed.
12	I understand your confusion. What causes you to feel stressed?	Sometimes I sleep badly.
13	I see. How often do you sleep badly?	I sleep badly during the week.
14	Ok. How long has it been since you've slept badly?	Last week I slept badly and the previous weeks too.
15	Ok. Did you sleep badly the previous week?	q!

Table 16: Test conversation 3

1	Would you introduce me briefly, why you seek consultation?	I'm sick.
2	Ok. Where do you go when you're sick?	I'm sick.
3	And... Why are you sick?	I'm sick.
4	And... What happens when you're sick?	I'm sick.
5	Ok. How long have you been sick?	I'm sick.
6	Ok. Why are you sick?	I'm sick.
7	And... How often do you get sick?	I'm sick.
8	And... How do you know you're sick?	I'm sick.
9	I am not sure, if I have understood correctly.	I'm sick.
10	Could you explain that a bit more detailed?	I'm sick.
11	And... Where do you go when you're sick?	I'm sick.
12	Could you explain that a bit more detailed?	I'm sick.
13	I am not sure, if I have understood correctly.	I'm sick.
14	Could you explain that a bit more detailed?	I'm sick.
15	How do you normally deal with that?	q!

Table 17: Test conversation 4

1	Would you introduce me briefly, why you seek consultation?	I've had severe headaches for a few days.
2	At what time of the day is the pain the worst?	The worst is the headache in the evening.
3	What treatment did you have so far?	So far I have not taken any treatment.
4	What do you expect as the result of the examinations and treatments in our clinic?	I hope we find out why I have a headache.
5	Do you also have cold symptoms? If yes, which?	No, I don't have any symptoms of a cold.
6	Are you often exposed to loud noises? If yes, what type of noises?	Sometimes there are workers who make loud noises on the upper floor.
7	Let's think about how we can solve this. How do workers sometimes make loud noises on the upper floor?	They work with loud machines. Then I notice that my headache gets worse.
8	We will clarify this. How long does it take for your headache to get worse?	Most of the time, the pain gets worse immediately.
9	I see. Most of the time, the pain gets worse immediately	Yes and sometimes I also get earache.
10	I understand your confusion. When do you get an earache?	WHEN the workers are loud, then I get earache.
11	Do you often feel stress? What makes you feel stressed?	No, I don't feel very stressed. I actually have a relaxed everyday life.
12	We will clarify this. When do you feel relaxed?	Actually, I always feel relaxed when there is no noise around.
13	Let's think about how we can solve this. How do you feel relaxed when no noise is around?	Well, when it's noisy, I don't feel so relaxed anymore.
14	Let's think about how we can solve this. What do you do when it's noisy?	Sometimes I put earplugs in my ears.
15	When did you relax last time?	This morning, I was very relaxed.
16	I'm happy that you feel better. Where were you in the morning?	In the morning I was at home and it was quiet.
17	We will clarify this. When were you at home?	I was at home in the morning.
18	I understand your confusion. How long has it been since you were at home?	I'm not home now. I'm at work now.
19	I understand your confusion. How many of you are home now?	I don't think there's anyone at home now.
20	We will clarify this. Do you think there's anyone at home now?	Nope.
21	What a surprise. What was the name of the person who said no?	q!

Table 18: Test conversation 5

1	Would you introduce me briefly, why you seek consultation?	I'm here because it whistles in my ear
2	At what time of the day is the pain the worst?	The pain is worst in the evening
3	What treatment did you have so far?	I've already had treatment with the family doctor
4	What do you expect as the result of the examinations and treatments in our clinic?	I expect the whistling in my ear to go away
5	Do you also have cold symptoms? If yes, which?	I have no symptoms of a cold
6	Are you often exposed to loud noises? If yes, what type of noises?	I am often exposed to loud flying noise
7	Let's think about how we can solve this. Where are you exposed to loud flying noise?	I am exposed to loud flying noises at the airport
8	I see. Where are you exposed to flying noises?	The flight noises are exposed at the airport
9	Do you often feel stress? What makes you feel stressed?	No I don't feel stressed often
10	We will clarify this. How many times do you feel stressed?	I never feel stressed
11	That sounds positive. When do you never feel stressed?	I don't feel stressed every day
12	When did you relax last time?	I last relaxed a week ago
13	What does your chief complaint stop you from doing?	I can't go about my work because of my main complaint
14	Ok. How often do you go about your work?	I do my job for 5 days a week
15	We will clarify this. How long do you work?	q!

Table 19: Test conversation 6

B Lists of predefined content

B.1 Profile questions

The below list the default defined questions for patient profiling.

- What is your first name?
- What is your surname?
- What is your gender?
- What is your birthday?
- What is your height in cm?
- What is your weight in kg?
- What is your medical care provider?

B.2 Empathic phrases

The below list the default defined empathic phrases for default defined mental states with a placeholder for a personal form of address.

- sad
  - I understand, [address form].
  - I'm sorry about that.
  - I'm sure, you will overcome this.
  - This must be hard.
  - Don‘t worry, it soon will get better.
- neutral
  - I see.
  - Ok.
  - Alright.
  - Mhmm.
  - I agree with you, [address form].
  - Ok.
  - And...
  - Do you have an idea, [address form],
  - So
  - Hmm.
  - Right.
  - [address form],
- surprised
  - What a surprise.
- afraid
  - Don‘t be afraid. It soon will get better.

- confused
  - Let‘s think about how we can solve this.
  - We will clarify this.
- disgusted
  - Mhmm.
- angry
  - I understand your anger, [address form].
  - That must be frustrating.
  - I’m sorry about that, [address form].
- happy
  - That sounds positive.
  - Good.
- ill
  - Ok.
  - And...

B.3 Topics

The below list contains the ten latent topics with their weights and keywords from the topic model, trained on the WISE dataset.

- Topic: 0
  - weight: 0,61678
  - keys: get, pain, time, feel, like, year, day, sometimes, problem, bad, since, would, help, always, anything, often, lot, could, make, know
- Topic: 1
  - weight: 0,14792
  - keys: doctor, tongue, treatment, burn, tooth, mouth, since, implant, dental, dentist, eat, area, throat, take, see, xxx, remove, gum, lip, complaint
- Topic: 2
  - weight: 0,20515
  - keys: tooth, splint, night, bite, grind, wear, sleep, jaw, clench, position, stress, low, dentist, treatment, day, front, use, michigan, crunch, dental
- Topic: 3
  - weight: 0,27253
  - keys: sleep, work, headache, life, long, pain, stress, physical, sport, everyday, cause, time, lie, quality, day, possible, well, concentrate, activity, head
- Topic: 4
  - weight: 0,09859
  - keys: cold, tooth, drink, pain, eat, food, hot, warm, brush, heat, trigeminal, water, neuralgia, gum, upper, fruit, ice, sensitive, sweet, temperature
- Topic: 5
  - weight: 0,42317
  - keys: jaw, mouth, open, joint, chew, yawn, crack, hard, wide, eat, opening, long, bite, left, right, food, keep, without, side, move

- Topic: 6
  - weight: 0,55341
  - keys: pain, chew, food, hard, jaw, eat, long, keep, talk, still, bite, chewing, yawn, speak, right, mouth, biting, time, relax, joint
- Topic: 7
  - weight: 0,37167
  - keys: jaw, tension, muscle, neck, stress, pain, headache, relaxation, relax, area, massage, mental, sleep, exercise, physical, shoulder, tense, rest, migraine, night
- Topic: 8
  - weight: 0,4903
  - keys: pain, right, left, side, tooth, jaw, cause, upper, pressure, ear, low, severe, find, treatment, leave, face, area, sometimes, head, root
- Topic: 9
  - weight: 0,10893
  - keys: ear, tinnitus, jaw, pressure, noise, sound, hear, asleep, right, increase, loud, left, fall, hearing, music, change, since, high, possible, connection

B.4 Topic questions

The below list contains the default defined questions per topic.

- Topic 0:
  - When was the first time you experienced these complaints?
  - How often do you experience these complaints?
- Topic: 1
  - What do you expect as the result of the examinations and treatments in our clinic?
  - Do you also have cold symptoms? If yes, which?
- Topic: 2
  - At what time of the day is the pain the worst?
  - What treatment did you have so far?
- Topic: 3
  - Do you often feel stress? What makes you feel stressed?
  - When did you relax last time?
  - What does your chief complaint stop you from doing?
- Topic: 4
  - When you eat or drink, what is most problematic?
- Topic: 5
  - Do you have other complaints as consequences of the pain? (e.g. tension)
- Topic: 6
  - Which factors aggravate your complaints? (e.g. chewing hard / soft food, biting, drinking, mouth opening (e.g. yawning), talking, physical / emotional stress, playing a musicial instrument, ...)
  - Do you remember how your complaints started resp. what triggered them?
- Topic: 7

  - Which factors alleviate your complaints? (e.g. distraction, rest, relaxation, keeping jaw in fixed position, ...)

- Topic: 8
  - In which area does it hurt the most?

- Topic: 9
  - Are you often exposed to loud noises? If yes, what type of noises?

B.5 Questions for more details

The below list contains the default defined more-detail questions.

- Could you explain that a bit more detailed?
- I am not sure, if I have understood correctly.
- How do you normally deal with that?

C Other

C.1 Information text for test persons

Orofaziale Schmerzen sind schmerzhafte Beschwerden im Zahn-, Mund-, Kiefer- und Gesichtsbereich, oft unter Einbezug benachbarter Regionen, wie Ohr, Nacken/Hals und Kopf. Wird frühzeitig eine korrekte Diagnose erstellt, so können akute Schmerzen meist einfach behandelt werden. Chronische Beschwerden fordern eine interdisziplinäre Abklärung und Behandlung unter Einbezug von Zahnärzten, Ärzten verschiedener Fachrichtungen, Psychologen und Physiotherapeuten.

Orofaziale Schmerzen können vielfältige Ursachen und fachübergreifende Manifestationsformen haben. Diagnostische Verzögerungen führen zu zahlreichen Arzt- bzw. Zahnarztbesuchen und können in einer alle Lebensbereiche beeinträchtigenden Schmerzkrankheit resultieren. Eine frühzeitige Erkennung einer sich entwickelnden Chronifizierung und eine Behandlung in einem interdisziplinären Team sind daher wichtig. Ursächliche und modulierende Schmerzfaktoren sowie Patientenwünsche bestimmen die therapeutische Strategie, die häufig aus einer Kombination von nicht medikamentösen und medikamentösen Ansätzen besteht.

[Ettlin and Galli, 2008]