



**Universität
Zürich** ^{UZH}

Bachelor Thesis

Fall Semester 2022

**Medical Chatbot:
A Hybrid Task-Oriented Dialogue System for
Collecting Dental Healthcare Data by using
T5 to Generate Questions**

Author: Kim Hai Dinh (18-735-712)

E-Mail: kimhai.dinh@uzh.ch

Supervisor: Gerold Schneider

Department of Computational Linguistics

Zurich, December 1, 2022

Table of Contents

Abstract	2
1 Introduction	2
1.1 Types of Chatbots	3
2 Related Work	4
3 Hypothesis	5
4 Chatbot Architecture	5
5 Methodology	7
5.1 Datasets	7
5.2 GPT-3 (Generative Pre-trained Transformer 3)	7
5.3 Cosine Similarity	9
5.4 Keyword Extraction	9
5.5 Machine Translation	10
5.6 Natural Language Inference (NLI)	10
5.7 Chit-Chat Bot (GPT-2)	11
5.8 Train a new T5 Model Task	13
5.8.1 Text-to-Text Transfer Transformer (T5)	13
5.8.2 Training	14
6 Results	14
7 Evaluation	17
8 Discussion and Limitations	19
9 Conclusion	21
Acknowledgements	22
References	22

Abstract

In this bachelor thesis, a variety of Natural Language Processing (NLP) methods and approaches are being used to develop a medical chatbot which can collect dental healthcare data from the patients by asking them meaningful questions. A new T5 task was trained on GPT-3 generated questions, which can generate questions about the dental healthcare domain. A chit-chat functionality was also implemented where the patient can chit-chat with the chatbot. The bachelor thesis shows that it is possible to develop a chatbot which can ask meaningful questions about the dental healthcare domain.

Keywords: Chatbot, NLP, Dental Healthcare, T5, GPT-3, Chit-Chat, Cosine Similarity, Keyword Extraction

1 Introduction

We ask questions to either gather information about other people or test their knowledge about a subject. 60%-80% of people visiting their doctor have reported that they do not tell information to their doctor that could be relevant to their health. Some of the reasons are that they have the fear of being judged or embarrassed (Kiefer, 2018). Miles et al. (2018) stated: "Chatbots could be considered as an aid for doctor-patient communication for conditions with lower perceived stigma and severity." This paper focuses on the healthcare sector, more precisely on the collection of dental healthcare data. It showcases an approach on how to develop a medical chatbot which can collect dental healthcare data from patients. The goal of the chatbot is to collect as much meaningful information as possible from the patients. For this purpose, I developed a method to create a dataset that simulates short conversations between the patient and

the dentist by using GPT-3. This synthetic dataset is then used to train a T5 model task for generating questions.

1.1 Types of Chatbots

Most of the chatbots that we see on a regular basis are task-oriented or rule-based dialogue systems, which are very limited in their conversational capabilities. However, they are very good in a specific domain e.g., making restaurant reservation, promoting movies, etc. Therefore, task-oriented, or rule-based dialogue systems cannot respond to arbitrary utterances because they are limited to a specific domain (Csaky, 2017).

The second approach is “open domain” chatbots. These chatbots are trying to imitate human conversation as well as possible. Just a few companies have achieved a level of conversation that is deceptively similar to human conversation. LaMDA by Google (Thoppilan et al., 2022) or GPT-3 by OpenAI (Brown et al., 2020) (which can also do other tasks) are already at this stage. They are trained on billions of parameters and billions of web documents/dialogues.

In chapter 5.2 I will go further into GPT-3. “For these models there isn’t a well-defined goal, but they are required to have a certain amount of world knowledge and commonsense reasoning capabilities in order to hold conversations about any topic.” (Csaky, 2017)

My approach is to combine elements of the task-oriented dialogue system and the open domain dialogue system to create a hybrid conversational agent that is robust in

a specific domain but also has the human-like ability to communicate. This approach makes a task-oriented conversational agent more approachable to the user because it can handle out of domain responses as well (Csaky, 2017).

2 Related Work

Studies about dialogue system are common in many healthcare application domains. However, these studies are more about reviewing health dialog systems for the improvement of health outcomes (Kearns et al., 2019; Vaidyam et al., 2019). In reality, there are very few studies that study the AI and NLP aspects of how a chatbot works and how it was developed. Csaky (2017) is one of the few that considers what the methods and possibilities are on creating a deep learning based chatbot model. There are only a handful of companies which are developing state-of-the-art dialogue systems. Thoppilan et al. (2022) from Google developed the Language Model LaMDA which can produce human-like conversations. There are some papers such as Jiang et al. (2019) which developed a Question and Answering system for dental healthcare. However, they are more focused on how to answer a dental healthcare question.

My bachelor thesis shows an approach that was never done before, to develop a chatbot in the dental healthcare domain which uses GPT-3 to generate synthetic data to train a T5 model for generating dental healthcare questions.

3 Hypothesis

Below are two hypotheses that were examined during the study.

1. Does the chatbot generate better questions for the dental healthcare domain when training the T5 model on another task, or does it generate better questions under a T5 model fine-tuned on SQuAD v1.1?

2. Is it possible to create a dataset with the assistance of GPT-3 to generate questions which are qualitative enough for the collection of dental healthcare data?

4 Chatbot Architecture

The chatbot is specialized in generating meaningful questions based on the patient's input. However, it has not learned any factual knowledge of the world or of the dental healthcare domain. Figure 1 shows the flowchart of the chatbot. Not everything can be illustrated, but it is a quick overview of the most important components and methods. The Architecture of the chatbot is split into two main layers: The first layer is the front end, where the patient interacts with the chatbot by typing to the system in German. The second layer is the back end, where the input of the patient is first translated into English and then processed to generate a suitable question.

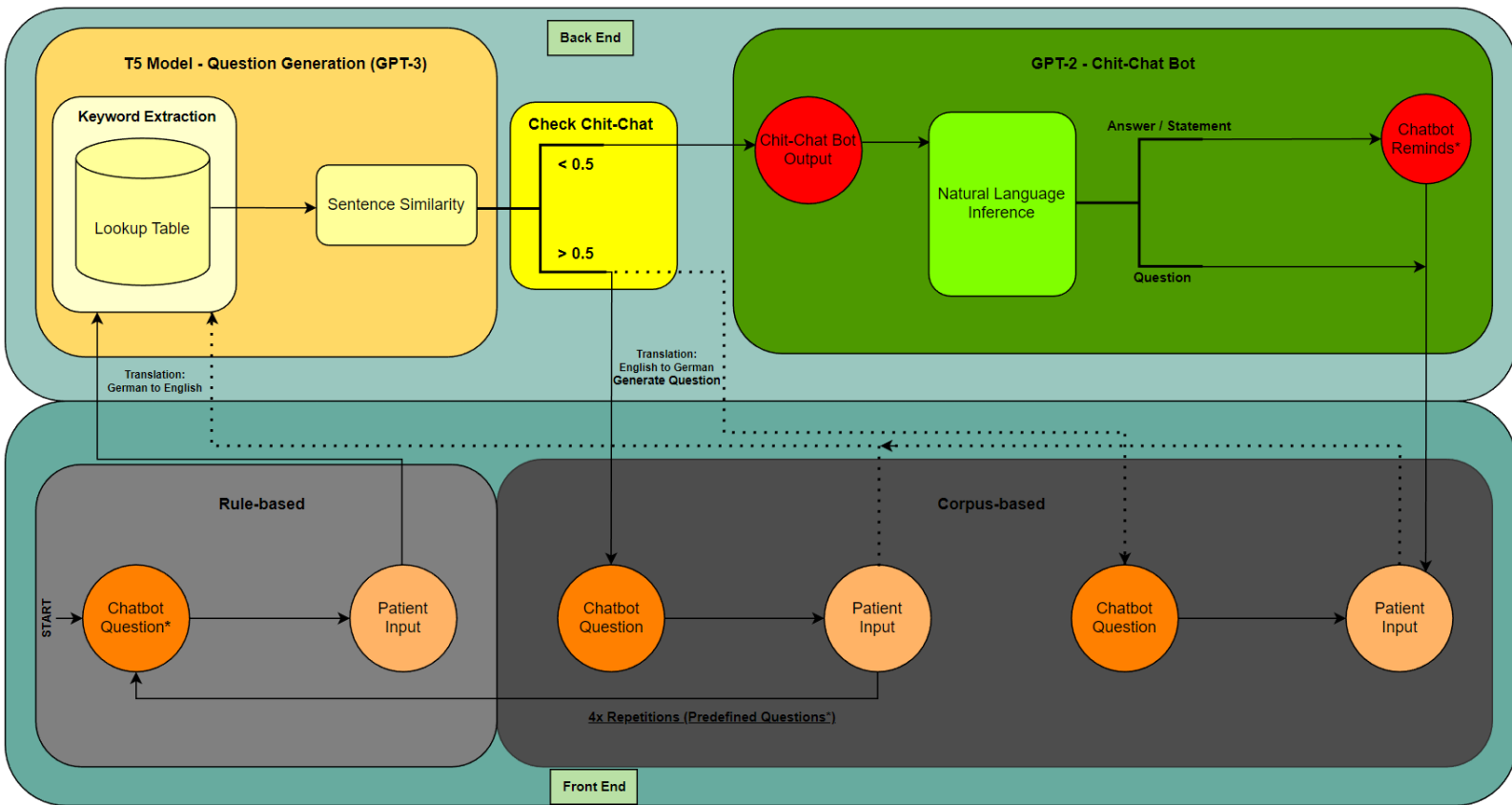


Figure 1 - Chatbot Workflow

The chatbot starts by asking a predefined question. Afterwards, the patient types in a response to the question. The patient input goes through different procedures such as Machine Translation, Keyword Extraction, comparing Cosine Similarities between sentences and potentially Natural Language Inference to generate a question with the T5 model. I go further into detail about these methods in the next chapter. The rule-based phase takes four repetitions, with each repetition having a different predefined question. After these four repetitions, the corpus-based phase begins (indicated with dotted lines) where the questions are generated solely on the patient's input. Finally, a chit-chat bot was integrated so that the patient can also do some casual small talk with the chatbot. The chit-chat "mode" will activate when the cosine similarity is under 0.5. Further details are also in the next chapter.

5 Methodology

In this chapter, I go further into the different datasets, procedures, and methods that I used to develop a chatbot. As a result, I trained a new T5 model task for generating questions.

5.1 Datasets

The first dataset is the WISE (Web-based Interdisciplinary Symptom Evaluation) dataset (Ettlin et al., 2016). A web-based questionnaire which contains, among other data, reported symptoms of dental or non-dental burdens and hopes for improvements of the patients. The data were collected from 2017 to 2021 and contains, 2236 patient reports. Furthermore, the dataset was translated from German to English and only the free texts of the dataset were used.

The second dataset is a dataset that was created with the assistance of GPT-3 and the WISE dataset. It contains 4511 questions generated by GPT-3 and the corresponding patient input from the WISE dataset.

The next section goes further into details on how the dataset was created with GPT-3.

5.2 GPT-3 (Generative Pre-trained Transformer 3)

GPT-3 by OpenAI is the largest autoregressive language model and was trained on 45 TB of compressed plaintext from books and the internet such as Wikipedia and other sources. With 175 billion parameters, it is more than 116 times larger than its predecessor GPT-2 which had 1.5 billion parameters (Brown et al., 2020). Autoregressive language model means that it can predict the most likely next

sequences of words given a set of words. Therefore, GPT-3 can perform different types of Natural Language Processing (NLP) tasks, such as Machine Translation,

Dentist asks: Which factors alleviate your complaints? Patient said: Distraction and relaxation ASK ONLY 4 QUESTIONS BASED ON WHAT THE PATIENT SAID: 1. What kinds of distractions and relaxation techniques do you find helpful? 2. Are there any specific activities or situations that make you feel better? 3. Do you have a favorite distraction or relaxation technique? 4. What advice would you give to someone who is experiencing similar symptoms?

Figure 2 – GPT-3 Question Generation example

Text Classification, Text Generation, Summarization, etc. In my case, I used GPT-3 to simulate a conversation between a dentist and the patient. Figure 2 shows an example of how I used GPT-3's ability to predict the next words to generate questions.

We can imagine it as a simulated conversation between two people, the first person is the dentist, and the second person is the patient. In the example, the dentist asks the patient: "Which factors alleviate your complaints?", which the patient responds with "Distraction and relaxation". The inputs are based on the WISE dataset, where the responses of the patients were already collected for the corresponding first question. Moreover, I gave GPT-3 the instruction to "ask only 4 questions based on what the patient said". Consequently, GPT-3 will generate four follow-up questions which have high probabilities of being asked by the dentist to the patient. I did this for every WISE dataset input which had a corresponding question and was free text. As a result, I generated over 24'000 questions.

However, I manually inspected and selected for every WISE data input which of the four questions would fit the best for a follow-up question by the dentist. As a result, I manually picked, 4511 follow-up questions.

5.3 Cosine Similarity

Cosine similarity is a metric that describes how similar for example two words or sentences are to one another. In this case, I looked at sentences. I used the library “SentenceTransformer” that transforms the sentences into embeddings which means that using sentence embeddings, sentences that are close in meaning are grouped near to one another in the vector space or in other words, sentences where words that have the same meaning have similar representations (Brownlee, 2017). Finally, with these sentence embeddings I can get the cosine similarity by calculating the normalized dot product of, e.g., Sentence Embedding 1 and Sentence Embedding 2.

5.4 Keyword Extraction

The chatbot system uses the python library “KeyBERT” which uses BERT embeddings and cosine similarity to find the sub-phrases in a document that are the most similar to the document itself. (Grootendorst, 2020)

For example, the sentence: "The pain in my jaw is unbearable." has the keywords pain (0.4175), jaw (0.5388) and unbearable (0.3775). The keyword with the highest cosine similarity is jaw, with 0.5388. This means that the keyword “jaw” describes

the sentence the best. Furthermore, with this method, a lookup table was created by first extracting the two most probable keywords of every WISE free text input. For the example above, jaw (0.5388) and pain (0.4175) are taken as keywords of the sentence because they are the most probable keywords. Secondly, all the unique keywords are initialized as the keys of the dictionary (lookup table) and every sentence of the WISE data which contains the key in its sentence will be appended as the value.

5.5 Machine Translation

Machine Translation is used to translate the patient's input from German to English. English needs to be used because all methods and models work with English. For translating German into English, the models "facebook/wmt19-de-en" and "Helsinki-NLP/opus-mt-de-en" are being used. Facebook's translation model was used for the sentences because it showed better translation results from German to English than the Helsinki translation model, but occasionally it cannot translate single words, therefore the Helsinki model translates these words. Finally, "facebook/wmt19-en-de" is used for translating the generated question from English to German.

5.6 Natural Language Inference (NLI)

NLI is a very important NLP task which allows us to determine whether the given "hypothesis" logically follows from the "premise". In other words, the understanding of whether the hypothesis is true, while the premise is your only knowledge about

the subject. NLI classifies, in this case, whether a response is an answer, statement, or question (hypothesis) based on the chit-chat output of the system (premise). In the next section, I go further into details about the combination of NLI and the chit-chat functionality.

5.7 Chit-Chat Bot (GPT-2)

I implemented a chit-chat possibility which is based on GPT-2 (a smaller version of GPT-3) for the patient to use while writing to the chatbot system. Every patient input will be checked if the patient wants to chit-chat with the chatbot or not. The check is implemented as follows: Cosine similarity (sentence similarity) is used to look if the sentence of the patient is similar to the inputs of the patient in the WISE dataset via lookup table. If the cosine similarity value is smaller than 0.5 the chit-chat function activates, otherwise not. This indicates, in this case, whether a typical dental patient wants to chit-chat or not. When the chit-chat function is detected, the system gives a response with GPT-2. Moreover, it is checked if the response of GPT-2 was a question, answer or statement with Natural Language Inference covered in the previous section. The distinction between question, answer or statement is important so that the patient can give an answer if the GPT-2 output is a question.

Ultimately, an answer or a statement of the chit-chat bot is considered the same for simplicity reasons. When this case occurs, the chatbot reminds the patient of the initial dental question to pick him or her up at the right place.

```
ChatBot: Bitte beschreiben Sie Ihre Hauptbeschwerde.  
Wie geht es dir?  
ChatBot: Mir geht's gut, wie geht's dir?  
Für meine Umstände entsprechend gut.  
ChatBot: Das ist gut.  
Woher kommst du?  
ChatBot: Ich komme aus den Vereinigten Staaten.  
Wow, schön!  
  
ChatBot: Zurück zur Frage: Bitte beschreiben Sie Ihre Hauptbeschwerde.  

```

Figure 3 - Chit-Chat example

Figure 3 shows an example chit-chat conversation between the chatbot and the patient. The patient asked in German: “Wie geht es dir?” (“How are you?”) and the chatbot responds with “Mir geht’s gut, wie geht’s dir?” (“I’m fine, how are you?”). The system detects that the chatbot asked a question and lets the patient answer the question. Additionally, the chit-chat bot remembers the chit-chat history and can answer appropriately to previous statements, for example the patient answered to the chatbot’s question: “Für meine Umstände entsprechend gut.” (“For my circumstances, accordingly well.”) and the chatbot responds with “Das ist gut.” (“That’s good.”).

5.8 Train a new T5 Model Task

In this section, the functionality of the T5 model is explained and how it was trained for the new question generation task.

5.8.1 Text-to-Text Transfer Transformer (T5)

Transfer learning has led to a new wave of state-of-the-art results in natural language processing (NLP). The Text-To-Text Transfer Transformer (T5) was developed by Google and is a new model that is pre-trained on a large dataset and achieves state-of-the-art results on many NLP benchmarks. The T5 model is flexible enough to be fine-tuned to a variety of important downstream tasks (Raffel et al., 2019).

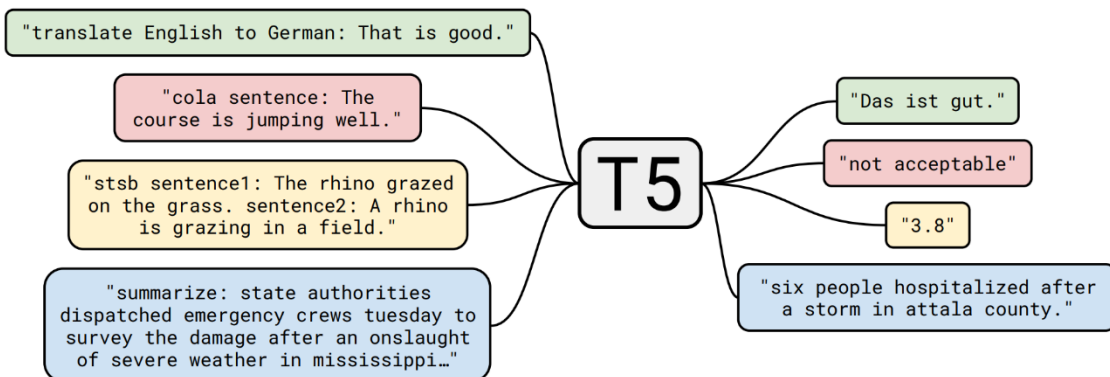


Figure 4 – Different text-to-text frameworks of T5

Figure 4 shows the text-to-text framework of T5. T5 needs a text as input and trains it to generate a target text. Moreover, it can learn different NLP tasks when adding a prefix before the input. Raffel et al. (2019) explains: “As an example, to ask the model to translate the sentence “That is good.” from English to German, the model would be fed

the sequence “translate English to German: That is good.” and would be trained to output “Das ist gut.” For the development of the chatbot, the t5-base model was used to train a new T5 task for generating questions. Finally, the next section goes more into detail how the new T5 task was trained.

5.8.2 Training

A new T5 task was trained by using the same method described in the previous section. The prefix “generate_question:” was concatenated with the input text sequence to indicate the task to perform, e.g., “generate_question: Chew, eat”. This was trained to output: “What foods or drinks do you consume that make them worse?”. Therefore, the input to a T5 model has the following pattern: “<prefix>: <input_text> </s>” and the output sequence has the following pattern: “<output_sequence> </s>”.

The model was trained with SimpleT5 with a batch size of 2, Adam Optimizer and with 21 epochs. However, the model which was trained on 5 epochs was chosen to generate the questions because the validation loss began to increase drastically after this point.

6 Results

Figure 5 shows how an example conversation between the chatbot, and a patient can look like. The chatbot takes first note of the patient’s personal information: name, age, weight etc. Afterwards, a quick introduction and remark from the chatbot. When the hybrid-based phase starts, a predefined question initiates the conversation between patient and chatbot. The chatbot stops asking questions after the 10th question.

Vorname: Max
Nachname: Mustermann
Geschlecht: männlich
Alter: 34
Körpergröße in cm: 183
Gewicht in kg: 76

ChatBot: Guten Tag Herr Mustermann, ich bin Ihr persönlicher Chatbot!
ChatBot: Ich werde Ihnen einige Fragen über Ihre Beschwerden stellen.
ChatBot: Bitte beantworten Sie die Fragen so ehrlich und kurz wie möglich.
ChatBot: Ihre Antworten werden nur im Rahmen Ihrer Behandlung verwendet.

ChatBot: Bitte beschreiben Sie Ihre Hauptbeschwerde.

Ich habe Tinnitus im rechten Ohr.

ChatBot: Was ist die Ursache Ihres Tinnitus?

Die Ursache sind akute Schmerzen im Kiefer.

ChatBot: Was können Sie wegen Ihrer Hauptbeschwerde nicht mehr machen?

Ich höre nicht mehr gut im rechten Ohr.

ChatBot: Wie hat sich Ihre Lebensqualität verändert, seit Sie diese Symptome erfahren haben?

Meine Lebensqualität hat sich sehr verschlechtert. Ich kann mich nicht mehr auf die Arbeit konzentrieren.

ChatBot: Welche Faktoren verschlimmern Ihre Beschwerden? (z. B. Kauen von harter/weicher Nahrung, Beissen, Trinken, Mundöffnung (z. B. Gähnen), Sprechen, körperlicher/emotionaler Stress, Spielen eines Musikinstruments,...)

Wenn ich spreche oder einfach den Mund öffne.

ChatBot: Gibt es andere Symptome, die Sie erleben, wenn Sie den Mund aufmachen?

Ja, ein knackendes Geräusch beim Öffnen des Kiefers.

ChatBot: Welche Faktoren lindern Ihre Beschwerden? (z. B. Ablenkung, Ruhe, Entspannung, Halten des Kiefers in einer festen Position,...)

Ich versuche zu entspannen und vermeide Stress und laute Hintergrundgeräusche.

ChatBot: Was tun Sie, um Stress und laute Hintergrundgeräusche zu vermeiden?

Ich versuche zu schlafen oder gehe nicht an Orten, wo es viele Menschen hat.

ChatBot: Gibt es noch andere Faktoren, die zu Ihren Beschwerden beitragen könnten?

Ja, ich bekomme auch andauernd Kopfschmerzen.

ChatBot: Danke für das Beantworten der Fragen!

ChatBot: Ich wünsche Ihnen noch einen schönen Tag.

Figure 5 - Chatbot conversation example

40 conversations with the chatbot have been made. I took the place of multiple potential patients to determine whether the chatbot can generate meaningful questions which also fit the conversation flow accordingly. First, I did 20 conversations with the new trained T5 chatbot to generate question which are trained on the questions generated by GPT-3. Afterwards, I did 20 conversations with the T5 model fine-tuned on the SQuAD v1.1 that stands for “Stanford Question Answering Dataset”, a large dataset which consists of 100’000+ sets of question-answer pairs (Rajpurkar, 2016). This is a model which comes the nearest to my model and is freely available.

Returning to the conversations, after each conversation I categorized the conversation between one of the three categories: Smooth Connections, Usable Connections, Useless Connections.

	Smooth Connections	Usable Connections	Useless Connections
Chatbot with new T5 Task	4	9	7
Chatbot with fine-tuned T5	0	4	16

Table 1 – Results of 20 Conversations with the Chatbot on the new T5 Task and 20 Conversations with the fine-tuned T5

Useless connections are conversations that had two or more questions that either didn’t make sense or were not suitable for the conversation. Moreover, Usable Connections have only one question that didn’t make sense or were not suitable for the conversation. Smooth connections are conversation that all questions make sense and are suitable for the conversation. An example of a Smooth Conversation is Figure 5 above.

I categorized in Table 1, 20 conversations with the new T5 task chatbot which resulted in 4 conversations labeled as Smooth Connections, 9 conversations have Usable

Connections, and 7 conversations have Useless Connections. Whereas the chatbot which T5 is fine-tuned on the SQuAD v1.1 has zero conversations categorized as Smooth Connections, 4 conversations as Usable Connections and 16 conversations as Useless Conversations. In the next chapter, I go into the evaluation of the conversations with the chatbot.

7 Evaluation

Looking at the result of conversations I conducted with the chatbot, there is clearly a better model. The chatbot with the new T5 task trained on GPT-3 generated questions has 13 conversations which were Usable Connections or better, out of the 20 conversations conducted with this model. In contrast, the T5 fine-tuned on the SQuAD v1.1 dataset performed very badly, with only 4 out of 20 conversations which are Usable Connections.

Some of the problems that occur in both but more in the first model were for example that the given user input had no keyword in the input which is needed for keyword extraction function to work, e.g., the input “Mehr oder weniger” (More or less) will not extract a keyword. Additionally, even when there are keywords in the sentence, not all extracted keywords from the sentence are in the lookup table. Moreover, the cosine similarity between the user input sentence and the sentences in the lookup table were sometimes under the value of 0.5 which means it will activate the chit-chat mode which is only desirable when the patient truly wants to chit-chat e.g., the sentence: “Ich bin nicht mehr so glücklich wie früher.” (I'm not as happy as I used to be.) scores a value of 0.41 which is under 0.5. These were the main problems that occur, and which decided

whether a conversation was categorized either into Usable Connections or Useless Connections.

On the contrary, the second model generated questions which made no sense for the conversation flow, e.g., “Was ist ein häufiges Symptom von Zahnschmerzen?” (“What is a common symptom of toothache?”) which is a general question about toothache.

Furthermore, questions were generated that only a previous, repeated answer can be given.

I know this evaluation is very subjective and not a very standardized method to evaluate a chatbot, but it can give a first glimpse into how these two models can perform in a real-world use case. The new trained T5 model performed well with a 65% chance of getting a performance better than Useless Connections while the other model only has a 20% chance to perform the same. The worse performance of the T5 finetuned with SQuAD v1.1 dataset is due to the fact that it was not trained specifically on the dental healthcare domain, which the new trained T5 model was.

A better evaluation will be conducted in the future, where the chatbot will be tested and evaluated in a real-world scenario. The chatbot will then be tested in a dental practice, where real dental patients will interact with the chatbot and give their feedback and opinions.

Finally, the question whether the new T5 model can generate meaningful dental healthcare questions is debatable. In my time interacting with the chatbot, I saw that the chatbot can generate meaningful questions about dental healthcare. But of course, it has still lots of room for improvement and how this improvement can look like and what other limitations are, will be discussed in the next chapter.

8 Discussion and Limitations

One major aspect at the beginning of the project was whether the chatbot could be implemented with only German models. This means that the front end and the back end would only process with German sentences. I saw immediately that this would be a very difficult task to process everything in German instead of English models. The reason was either that the English models performed better, or there were no German models for the problems that I wanted to solve. For instance, GPT-3 can generate German sentences, but the performance with English sentences were still better. Because GPT-3 is mainly trained on large texts which are in English. It is the case that normally machine learning models perform better when trained on large datasets, but of course not in every situation. Ansaldo (2022) stated: “More training data improves AI performance up to a certain point but can compromise performance beyond it. The quality of the data used to train AI is just as important as the quantity. Poor data quality leads to poor AI results.”

As a result, I chose to process everything in the back end layer first in English and then output a German sentence for the patient. Consequently, this sometimes leads to wrong translations and therefore inevitable noise, which can compound during the processing phase for generating a question. For example, the sentence “What foods do you have a hard time eating?” to “Welche Lebensmittel haben Sie schwer zu essen?” instead to “Welche Lebensmittel können Sie nur schwer essen?”. As a result, the patient might not take the chatbot as serious as he or she would if there were no grammatical errors by the system. Nevertheless, the chatbot produces good translations in most cases.

Furthermore, the free texts of the patients in the WISE questionnaire were mostly written as key points rather than full sentences. This can lead to the decline in quality when searching for the most similar sentence in the lookup table because I assume that the patients will write full sentences when communicating with the chatbot.

This brings me also to another point, on how I generated the dataset to train the T5 Model task. I generated four questions per WISE data input and manually chose which one of the four questions might be the best follow-up question for the chatbot. The problem with this approach is that as a layperson in the dental healthcare domain, I have not the same judgement as a dentist. The more scientific way would be either to have at least two people “annotate” which question fit the best and calculate the Inter-Annotator Agreement e.g., Cohen’s Kappa for two annotators or Fleiss’ Kappa for more than two annotators to determine the annotation decisions and therefore the quality between annotators. Another way would be to let an expert such as a dentist choose the follow-up question because a dentist has an in-depth knowledge and experience of what a follow-up question should look like.

Another improvement is to get more dental healthcare data to generate more questions from GPT-3 and train the T5 model task further. 4511 questions that I used for training were still very scarce for training a new T5 model task, but the chatbot still performed well.

9 Conclusion

In summary, a medical chatbot was developed to collect dental healthcare data from patients in, e.g., dental practice. It was developed by using different NLP methods and was trained on a new T5 task which can generate dental healthcare questions. The training data for the T5 model were created by using the functionality of GPT-3 to generate the most probable following words in a sequence. With this approach, conversations between the dentist and the patient were simulated to generate follow-up questions which a dentist could ask a patient. Furthermore, a chat functionality was also built-in for the patient to interact.

The chatbot's new T5 model task was compared to a T5 model fine-tuned on the SQuAD v1.1 dataset and outperformed it with 13 out of 20 conversations with "Usable Connections" whereas the latter had only 4 out of 20 conversations with "Usable Connections".

Finally, the chatbot can generate meaningful dental healthcare questions, but still has room for improvement to perform better.

The method that I created to generate synthetic data from GPT-3 to train a T5 model or any other model will only be the beginning on how AI Engineers will train their models in the future. This is only the first step into something much bigger.

Acknowledgements

A special thanks to my supervisor PD Dr. Gerold Schneider and Dr. Tilia Ellendorff for giving me a helping hand during my bachelor thesis. In addition, a special thanks to PD Dr. Dr. Dominik Ettl and Dr. phil. Markus Wolf for providing their WISE dataset from the WISE-Project.

References

Kiefer, J. (2018). WHY PATIENTS LIE TO THEIR DOCTORS.

Retrieved November 10, 2022 from

<https://healthcare.utah.edu/publicaffairs/news/2018/11/lying-patients.php>

Miles, O., West, R., & Nadarzynski, T. (2021). Health chatbots acceptability moderated by perceived stigma and severity: A cross-sectional survey. *Digital health*, 7, 20552076211063012. <https://doi.org/10.1177/20552076211063012>

Csaky, R. (2019). Deep learning based chatbot models. arXiv preprint

arXiv:1908.08835.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T., ... &

Le, Q. (2022). Lamda: Language models for dialog applications. arXiv preprint

arXiv:2201.08239.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.

Kearns, W. R., Chi, N. C., Choi, Y. K., Lin, S. Y., Thompson, H., & Demiris, G. (2019). A systematic review of health dialog systems. *Methods of Information in Medicine*, 58(06), 179-193.

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7), 456-464.

Jiang, Y., Xu, Y., Guo, J., Liu, Y., & Li, R. (2019, October). An intelligent question and answering system for dental healthcare. In *International conference on broadband communications, networks and systems* (pp. 201-215). Springer, Cham.

Ettlin, D. A., Sommer, I., Brönnimann, B., Maffioletti, S., Scheidt, J., Hou, M. Y., ... & Steiger, B. (2016). 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.

Brownlee, J. (2017). What Are Word Embeddings for Text?.

Retrieved November 13, 2022 from

<https://machinelearningmastery.com/what-are-word-embeddings>

Grootendorst, M. (2020). Keyword Extraction with BERT.

Retrieved November 14, 2022 from

<https://www.maartengrootendorst.com/blog/keybert/>

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J.

(2020). Exploring the limits of transfer learning with a unified text-to-text

transformer. *J. Mach. Learn. Res.*, 21(140), 1-67.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Ansaldo, M. (2022). When training AI models, is a bigger dataset better?.

Retrieved November 22, 2022 from

<https://www.hpe.com/us/en/insights/articles/when-training-ai-models-is-a-bigger-dataset-better-2207.html>