**CAN THO UNIVERSITY**

**COLLEGE OF INFORMATION AND COMMUNICATION TECHNOLOGY**

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**🙞 🕮 🙜**

**Logo

Description automatically generated**

**GRADUATION THESIS**

**BACHELOR OF ENGINEERING IN**

**INFORMATION TECHNOLOGY**

**(HIGH-QUALITY PROGRAM)**

**MEDBOT: CHATBOT ABOUT**

**HEALTHCARE**

**Student: Nguyen Trung Tam**

**Student ID: B1910697**

**Class: 2019-2023 (K45)**

**Advisor: Dr. Lam Nhut Khang**

**Can Tho, 12/2023**

**CAN THO UNIVERSITY**

**COLLEGE OF INFORMATION AND COMMUNICATION TECHNOLOGY**

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**🙞 🕮 🙜**

**Logo

Description automatically generated**

**GRADUATION THESIS**

**BACHELOR OF ENGINEERING IN**

**INFORMATION TECHNOLOGY**

**(HIGH-QUALITY PROGRAM)**

**MEDBOT: CHATBOT ABOUT**

**HEALTHCARE**

**Student: Nguyen Trung Tam**

**Student ID: B1910697**

**Class: 2019-2023 (K45)**

**Advisor: Dr. Lam Nhut Khang**

**Can Tho, 12/2023**

EVALUATION OF ADVISOR

....................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................

Advisor

Lam Nhut Khang

ACKNOWLEDGMENTS

I wish to express my deep gratitude and sincere thanks to my professor Lam Nhut Khang– A lecturer at the College of Information and Communication Technology who gave me the golden opportunity to do this wonderful thesis on the topic “MedBot: Chatbot about healthcare”, which also helped me in doing a lot of research and I came to know about so many new things I am really thankful to them. Then I would like to thank the lecturers of Can Tho University, specifically, the lecturers of the College of Information and Communication Technology who taught me invaluable knowledge when I studied.

I am extremely grateful to my family for their love, prayers, and care for the completion of this thesis. I am very much thankful to my friends for their support when I was doing research at Can Tho University.

Sincerely,

Can Tho, 01/12/2023

Nguyen Trung Tam

TABLE OF CONTENTS

[CHAPTER 1: INTRODUCTION 1](#_Toc144840280)

[1.1. Problem 1](#_Toc144840281)

[1.2. History of problem-solving 2](#_Toc144840282)

[1.3. Purpose 3](#_Toc144840283)

[1.4. The objectives and scope 3](#_Toc144840284)

[1.5. Research Methods: 3](#_Toc144840285)

[1.6. Result 4](#_Toc144840286)

[CHAPTER 2: PROBLEM DESCRIPTION 5](#_Toc144840287)

[2.1. Detailed description of the problem 5](#_Toc144840288)

[2.2. Background 5](#_Toc144840289)

[2.2.1. Python 5](#_Toc144840290)

[2.2.2. Tensorflow 7](#_Toc144840291)

[2.2.3. Deep learning 8](#_Toc144840292)

[2.2.4. BART model 8](#_Toc144840293)

[2.2.5. Dataset 12](#_Toc144840294)

[CHAPTER 3: DESIGN AND IMPLEMENT OF THE CHATBOT 14](#_Toc144840295)

[3.1. Overview 14](#_Toc144840296)

[CHAPTER 4: CONCLUSION AND FUTURE WORK 15](#_Toc144840297)

[4.1. Conclusion 15](#_Toc144840298)

[4.2. Future work 15](#_Toc144840299)

[REFERENCES 16](#_Toc144840300)

[APPENDICES 17](#_Toc144840301)

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

**ABSTRACT**

In recent years, the integration of artificial intelligence (AI) and natural language processing (NLP) technologies has ushered in a new era in healthcare with the emergence of medical chatbots. These intelligent conversational agents hold immense promise in transforming the way healthcare services are delivered, improving patient experiences, and optimizing healthcare workflows.

This abstract provides an overview of the evolving landscape of medical chatbots, their key functionalities, and their impact on the healthcare ecosystem. Medical chatbots are designed to engage in dynamic conversations with users, offering personalized health information, symptom assessment, medication reminders, and appointment scheduling. They empower patients with on-demand access to medical guidance, reducing the burden on healthcare professionals and enhancing overall patient engagement.

Furthermore, medical chatbots demonstrate significant potential in improving healthcare outcomes through early symptom detection and continuous monitoring. By analyzing user-provided information and historical health data, these chatbots can identify potential health risks and provide timely recommendations, ultimately contributing to preventive care.

However, the widespread adoption of medical chatbots does not come without challenges. Ensuring data security, privacy compliance, and maintaining a high standard of accuracy and reliability in medical advice are paramount concerns. The need for seamless integration with electronic health records (EHRs) and healthcare information systems is another technical hurdle that must be addressed.

In conclusion, medical chatbots represent a transformative technology in the healthcare domain, offering a myriad of benefits such as improved accessibility, enhanced patient engagement, and streamlined administrative processes. While challenges persist, their potential to revolutionize healthcare delivery and empower individuals to take control of their health cannot be overlooked. As research and development in AI and NLP continue to advance, the future of medical chatbots holds great promise in reshaping the landscape of healthcare for the better.

**OVERVIEW**

Medical Chatbot – a chatbot application for answering questions about medical health using Python and ML

Nowadays, the necessity for health care is becoming more and more important, especially in a technology world, because of the rapid development of IT and the need for fast and accurate information about healthcare. However, there are a lot of questions and problems that traditional healthcare is limited in providing fast and accurate information. The old manual approach needs humans to search for the info and then prepare the answer for the customers.

Medical chatbot was created to provide a useful tool to assist those who want to find and answer questions about medical health care in the most effective, fast, and convenient way. In addition, the chatbot also supports collecting and updating practical healthcare information from doctors to update the chatbot accordingly. Besides, the chatbot also allows users to search for medical information according to their desire to help them find answers and advice quickly.

# INTRODUCTION

## Problem

Nowadays, healthcare is becoming more and more important, especially in a fast-paced technology world where everyone is so busy with their phone and digital devices. Therefore, the necessity to have a medical chatbot to help people have fast and reliable information about healthcare is considerably important. However, the development of medical chatbots has faced several challenges and problems in the healthcare industry, which modern chatbots aim to address. Here are some key problems that have spurred the development of medical chatbots:

1. Healthcare Accessibility: Many individuals face challenges in accessing healthcare services, particularly in remote or underserved areas. Medical chatbots can provide instant, round-the-clock access to basic medical information and advice, bridging the gap in healthcare accessibility.

2. Appointment Scheduling: Booking appointments with healthcare providers can be cumbersome and time-consuming. Chatbots can streamline the appointment scheduling process, making it more convenient for patients and reducing the administrative burden on healthcare facilities.

3. Information Overload: The internet is filled with vast amounts of health information, making it difficult for individuals to find reliable sources. Medical chatbots offer a trusted source of medical information, helping users sift through the noise to find accurate answers to their health-related questions.

4. Health Awareness and Education: Many people lack access to comprehensive health education. Medical chatbots can serve as educational tools, providing users with valuable information about various health topics and promoting health awareness.

While medical chatbots offer innovative solutions to these problems, it's important to recognize that they are not a replacement for professional medical care. They should complement, not substitute, the expertise of healthcare providers. Additionally, ensuring the accuracy of medical information, maintaining data privacy, and addressing ethical concerns remain ongoing challenges in the development and deployment of medical chatbots.

## History of problem-solving

Related chatbots:

Woebot

- The conversational agent was built using Decision Tree and appropriate NLP algorithms and needs to be installed as software in a stand-alone computer. All conversational responses are stored in a database. The specific area of application is Cognitive Behavioural Therapy (CBT) for anxiety and mood disorders targeting young adults who are college students

- The data for the training chatbot are collected using surveys. The data processing and analysis are carried out by qualitative analysis of the responses from the surveyors

- There were several categories on which Woebot could provide therapy. Results showed that there was a significant improvement in the mean PHQ-9 score compared to controls in the mental condition of participants with 2 weeks of therapy

iHelpr

- iHelpr is a text-based interactive Chatbot intended to provide mental health support in the workplace. The iHelpr is a web-based self-assessment tool and is reported to be available for six well-being indicators viz. stress, anxiety, depression, sleep, and self-esteem

- The Chatbot is developed using the bot development framework by Microsoft’s Cognitive Services, “an Application Programming Interface (API) that can process natural language, enable a Chatbot to recognize speech, and image-processing technology”

Tess

- Tess is a web-based chatbot devised by X2AI Inc. with an access interface via SMS (on mobile) and Facebook Messenger application.

- The chatbot is developed based on machine learning algorithms integrated with psycho-educational concepts and is said to be developed in conjunction and collaboration with trained mental health professionals

## Purpose

The application "MedicalBot: Chatbot about healthcare" is used for users around the world who have healthcare problems in particular and users across the world who want to search for information about healthcare and advice for avoiding health issues. Practical healthcare news, advice, and answers are provided to users who need fast and reliable information about health issues.

## The objectives and scope

The application provides main functions to help users find, search, and explore information related to healthcare problems and advice. This thesis will be focused on the problem of searching and providing reliable answers to the medical industry. The scope of the study is: researching the problem of searching and finding health-related information online and solving the problem by implementing a chatbot with friendly UI/UX and handy features.

## Research Methods:

- Requirements analysis: study the problem related to chatbots, especially in the medical field on the network, research papers, then analyze the function, and describe the requirements to build and train the chatbot.

- Data collection: Collect questions, answers, and relevant information about the medical problem to train the chatbot.

- Design: UI/UX design; Model analysis and design model architecture.

- Implementation: Using Python, and Google Colab for training the chatbot and then deploying the chatbot to Facebook messager with Python, and Flask using REST API.

## Result

- Using acquired knowledge of analysis, research, and information gathering, build an application with medical chatbot Python, and Machine Learning.

Thesis outline

**Chapter 1**: Introduction

**Chapter 2**: Background

General information about the study and main functions of the system.

**Chapter 3**: Design and implementation

Introduction of UI/UX designs, models, and implementation, describing

technologies that will be used in the study.

**Chapter 4**: Conclusion and future work

# PROBLEM DESCRIPTION

## Detailed description of the problem

Medical chatbot provides users with functions such as answering questions related to healthcare and searching for advice about health issues in an easy, fast, and reliable way. In addition, the chatbot feature is integrated into the website to help students get answers related to actual practice accurately and quickly.

Medical chatbot application consists of 2 main parts: the chatbot and the UI application.

The application part helps users interact with the chatbot and supports collecting questions and answers to improve the effectiveness of the training chatbot. In addition, the application section also supports users to find questions, answers, and news related to practical healthcare problems. The application section will also support an administrator function to maintain and update the chatbot to improve its accuracy of the model over time.

## Background

### Python

Python is a high-level, versatile programming language known for its simplicity and readability. Created by Guido van Rossum and first released in 1991, Python has gained widespread popularity in various domains, including web development, data science, artificial intelligence, scientific computing, and more.

Python is one of the most popular programming languages for machine learning and artificial intelligence. Its versatility, extensive libraries, and active community support make it an ideal choice for developing machine-learning models and conducting data analysis. Here's an overview of how Python is used in the field of machine learning:

1. Libraries and Frameworks: Python has a rich ecosystem of libraries and frameworks that simplify machine learning tasks. Some of the most commonly used ones include:

- numpy: numpy is a fundamental library for numerical computing in Python. It provides support for multi-dimensional arrays and mathematical functions, making it essential for data manipulation.

- pandas: pandas is a data manipulation library that provides data structures like DataFrames and Series. It is used for data cleaning, transformation, and analysis.

- scikit-learn: scikit-learn is a popular machine learning library that offers a wide range of machine learning algorithms for classification, regression, clustering, dimensionality reduction, and more. It also provides tools for model selection and evaluation.

- TensorFlow: Developed by Google, TensorFlow is an open-source machine learning framework that's widely used for deep learning tasks. It allows you to build neural networks for various applications, including image recognition, natural language processing, and reinforcement learning.

- PyTorch: PyTorch is another deep learning framework that has gained popularity for its flexibility and dynamic computation graph. It is known for its ease of use and is commonly used in research and development of neural network models.

- Keras: Keras is a high-level neural networks API that runs on top of other deep learning frameworks like TensorFlow and Theano. It simplifies the process of building and training neural networks.

2. Data Preprocessing: Python libraries like NumPy and pandas are instrumental for data preprocessing. You can clean and prepare your data, handle missing values, perform feature engineering, and create datasets suitable for machine learning tasks.

3. Visualization: Libraries like Matplotlib and Seaborn allow you to visualize your data, which is crucial for understanding patterns and relationships in your datasets. Visualization aids in data exploration and model evaluation.

4. Model Building and Training: Python's machine learning libraries provide a straightforward way to build, train, and evaluate machine learning models. You can experiment with various algorithms and techniques to find the best model for your specific problem.

5. Model Evaluation: Scikit-learn offers tools for evaluating machine learning models using metrics like accuracy, precision, recall, F1-score, and ROC curves. Cross-validation techniques help assess a model's generalization performance.

6. Deployment: After training a machine learning model, you can deploy it in a production environment. Python allows you to create web services, and RESTful APIs, or integrate models into applications using frameworks like Flask or Django.

7. Community and Resources: Python's machine learning community is vast, with a wealth of tutorials, documentation, and online courses available. Platforms like Kaggle provide datasets and competitions for practicing and honing your machine-learning skills.

8. Research: Python is commonly used for machine learning research, thanks to the availability of powerful libraries like TensorFlow and PyTorch. Researchers can experiment with cutting-edge techniques and algorithms.

Python's role in machine learning continues to evolve, and it remains at the forefront of AI and data science. Whether you're a beginner looking to get started or an experienced practitioner, Python's ecosystem and community support make it a valuable tool for tackling a wide range of machine-learning challenges.

### Tensorflow

TensorFlow is a free and open-source software library for machine learning and artificial intelligence. It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks.

TensorFlow can run on multiple CPUs and GPUs (with optional CUDA and SYCL extensions for general-purpose computing on graphics processing units). TensorFlow is available on 64-bit Linux, macOS, Windows, and mobile computing platforms including Android and iOS.

Its flexible architecture allows for the easy deployment of computation across a variety of platforms (CPUs, GPUs, TPUs), and from desktops to clusters of servers to mobile and edge devices.

TensorFlow computations are expressed as stateful dataflow graphs. The name TensorFlow derives from the operations that such neural networks perform on multidimensional data arrays, which are referred to as tensors.

### Deep learning

Deep learning is the branch of machine learning which is based on artificial neural network architecture. An artificial neural network or ANN uses layers of interconnected nodes called neurons that work together to process and learn from the input data.

In a fully connected Deep neural network, there is an input layer and one or more hidden layers connected one after the other. Each neuron receives input from the previous layer neurons or the input layer. The output of one neuron becomes the input to other neurons in the next layer of the network, and this process continues until the final layer produces the output of the network. The layers of the neural network transform the input data through a series of nonlinear transformations, allowing the network to learn complex representations of the input data.

Deep learning can be used for supervised, unsupervised as well as reinforcement machine learning. It uses a variety of ways to process these.

# DESIGN AND IMPLEMENT OF THE CHATBOT

## Overview

### Tokenizer

Tokenizing a text is splitting it into words or subwords, which then are converted to ids through a look-up table. Converting words or subwords to ids is straightforward, so we will focus on splitting a text into words or subwords (i.e. tokenizing a text)

Subword-based tokenization

Subword-based tokenization is a solution between word and character-based tokenization. 😎 The main idea is to solve the issues faced by word-based tokenization (very large vocabulary size, large number of OOV tokens, and different meanings of very similar words) and character-based tokenization (very long sequences and less meaningful individual tokens).

The subword-based tokenization algorithms do not split the frequently used words into smaller subwords. It rather splits the rare words into smaller meaningful subwords. For example, “boy” is not split but “boys” is split into “boy” and “s”. This helps the model learn that the word “boys” is formed using the word “boy” with slightly different meanings but the same root word.

Some of the popular subword tokenization algorithms are WordPiece, Byte-Pair Encoding (BPE), Unigram, and SentencePiece. We will go through Byte-Pair Encoding (BPE) in this article. BPE is used in language models like GPT-2, RoBERTa, XLM, FlauBERT, etc. A few of these models use space tokenization as the pre-tokenization method while a few use more advanced pre-tokenization methods provided by Moses, spaCY, ftfy. So, let’s get started. 🏃

Byte-Pair Encoding (BPE)

BPE is a simple form of data compression algorithm in which the most common pair of consecutive bytes of data is replaced with a byte that does not occur in that data. BPE relies on a pre-tokenizer that splits the training data into words. Pretokenization can be as simple as space tokenization.

After pre-tokenization, a set of unique words has been created and the frequency with which each word occurred in the training data has been determined. Next, BPE creates a base vocabulary consisting of all symbols that occur in the set of unique words and learns to merge rules to form a new symbol from two symbols of the base vocabulary. It does so until the vocabulary has attained the desired vocabulary size. Note that the desired vocabulary size is a hyperparameter to define before training the tokenizer.

Encoding and Decoding

Let us now see how we will decode our example. To decode, we have to simply concatenate all the tokens together to get the whole word. For example, the encoded sequence [“the</w>”, “high”, “est</w>”, “range</w>”, “in</w>”, “Seattle</w>”], we will be decoded as [“the”, “highest”, “range”, “in”, “Seattle”] and not as [“the”, “high”, “estrange”, “in”, “Seattle”]. Notice the presence of the “</w>” token in “est”.

For encoding the new data, the process is again simple. However, encoding in itself is computationally expensive. Suppose the sequence of words is [“the</w>”, “highest</w>”, “range</w>”, “in</w>”, “Seattle</w>”]. We will iterate through all the tokens we found in our corpus — longest to the shortest and try to replace substrings in our given sequence of words using these tokens. Eventually, we will iterate through all the tokens and our substrings will be replaced with tokens already present in our token list. If a few substrings are left (for words our model did not see in training), we will replace them with unknown tokens.

In general, the vocabulary size is big but still, there is a possibility of an unknown word. In practice, we save the pre-tokenized words in a dictionary. For unknown (new) words, we apply the above-stated encoding method to tokenize the new word and add the tokenization of the new word to our dictionary for future reference. This helps us build our vocabulary even stronger for the future.

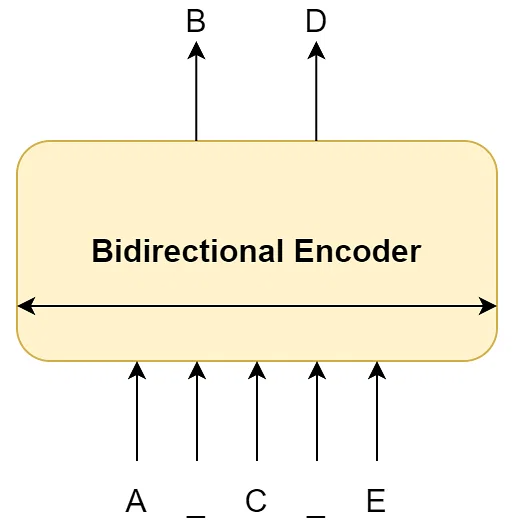
### BART model

BART is a denoising autoencoder that maps a corrupted document to the original document it was derived from. It is implemented as a sequence-to-sequence model with a bidirectional encoder over corrupted text and a left-to-right autoregressive decoder. For pre-training, we optimize the negative log-likelihood of the original document.

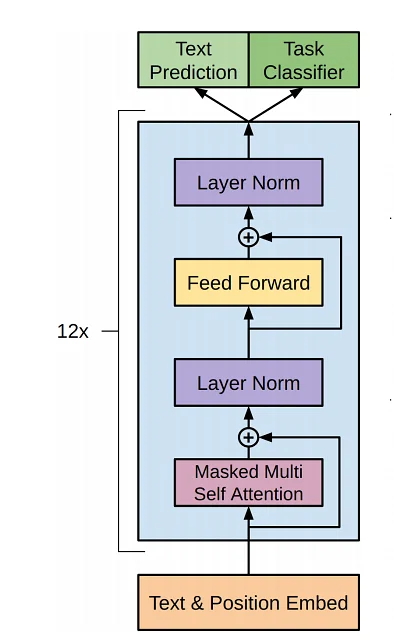
BART uses the standard sequence-to-sequence Transformer architecture from the original Transformer of Google, except, following GPT, BART modifies ReLU activation functions to GeLUs and initializes parameters from N (0, 0.02). For the base model, BART uses 6 layers in the encoder and decoder, and for the large model, BART uses 12 layers in each. The architecture is closely related to that used in BERT, with the following differences:

* Each layer of the decoder additionally performs cross-attention over the final hidden layer of the encoder (as in the transformer sequence-to-sequence model)
* BERT uses an additional feed-forward network before word prediction, which BART does not. In total, BART contains roughly 10% more parameters than the equivalently sized BERT model.

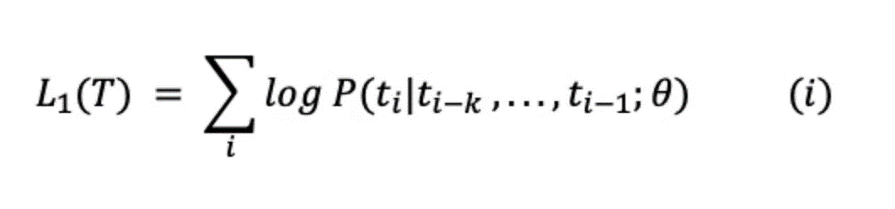
For every text sequence in its input, the BERT encoder outputs an embedding vector for each token in the sequence as well as an additional vector containing sentence-level information. In this way, the decoder can learn for both token and sentence-level tasks making it a robust starting point for any future fine-tuning tasks.



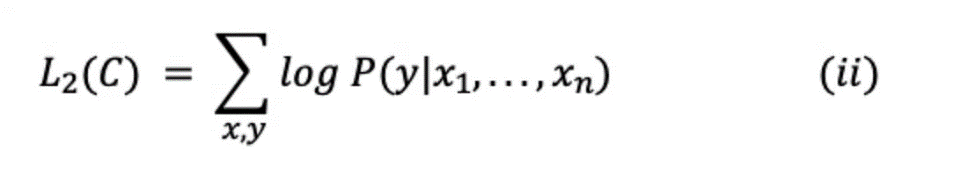
Once we get the token and sentence-level representation of an input text sequence, a decoder needs to interpret these to map with the output target. However, by using a similarly designed decoder, tasks such as next-sentence prediction or token prediction might perform poorly since the model relies on a more comprehensive input prompt. In these cases, we need model architectures that can be trained to generate the next word by only looking at the previous words in the sequence. Hence, a causal or autoregressive model that looks only at the past data to predict the future comes in handy.



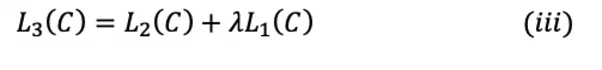
First, the model is pre-trained on tokens “t” looking back to “k” tokens in the past to compute the current token. This is done unsupervised on a vast text corpus to allow the model to “learn the language.”



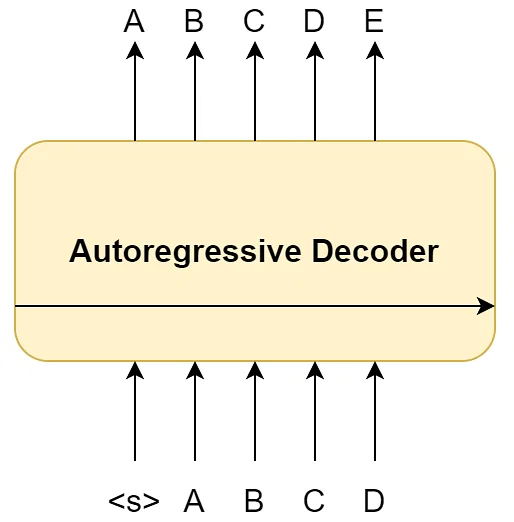
Next, to make the model robust on a specific task, it is fine-tuned in a supervised manner to maximize the likelihood of label “y” given feature vectors x1…xn.



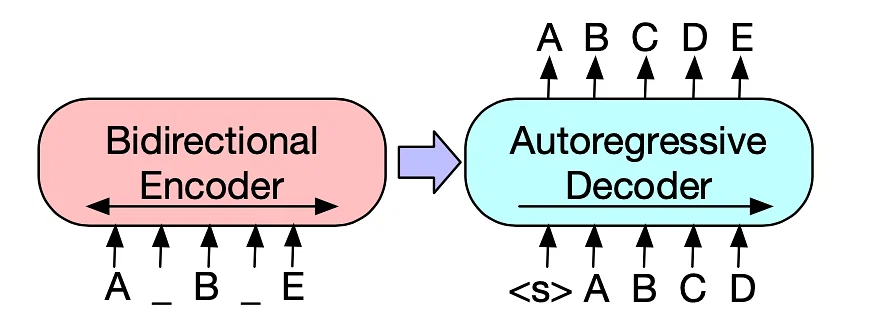
Combining 1 and 2, we get the objective in 3. Lambda represents a learned weight parameter to control the influence of language modeling.



The below image shows how the autoregressive decoder processes its input.



Although we separate the decoder from an encoder, the input to the decoder would still be a learned representation (or embedding) of the original text sequence. Thus, BART attaches the bi-directional encoder to the autoregressive decoder to create a denoising auto-encoder architecture. Based on these two components, the final BART model would look something like this:

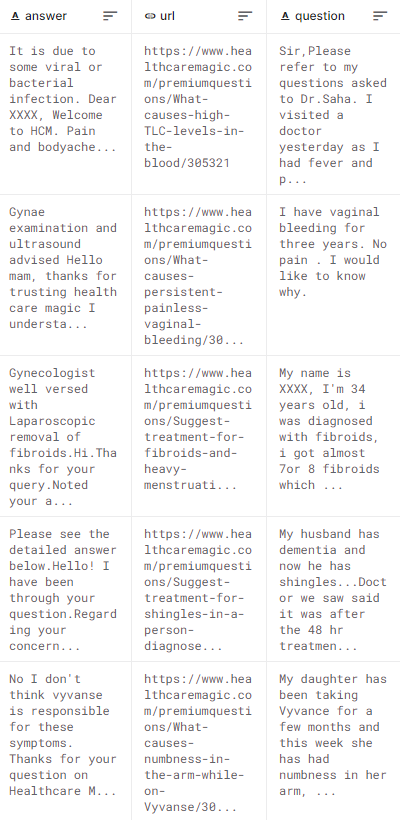


In the above figure, the input sequence is a masked (or noisy) version of [ABCDE] transformed into [A(MASK)B(MASK)E]. The encoder looks at the entire sequence and learns high-dimensional representations with bi-directional information. The decoder takes these thought vectors and regressively predicts the next token. Learning occurs by computing and optimizing the negative log-likelihood as mapped with the target [ABCDE].

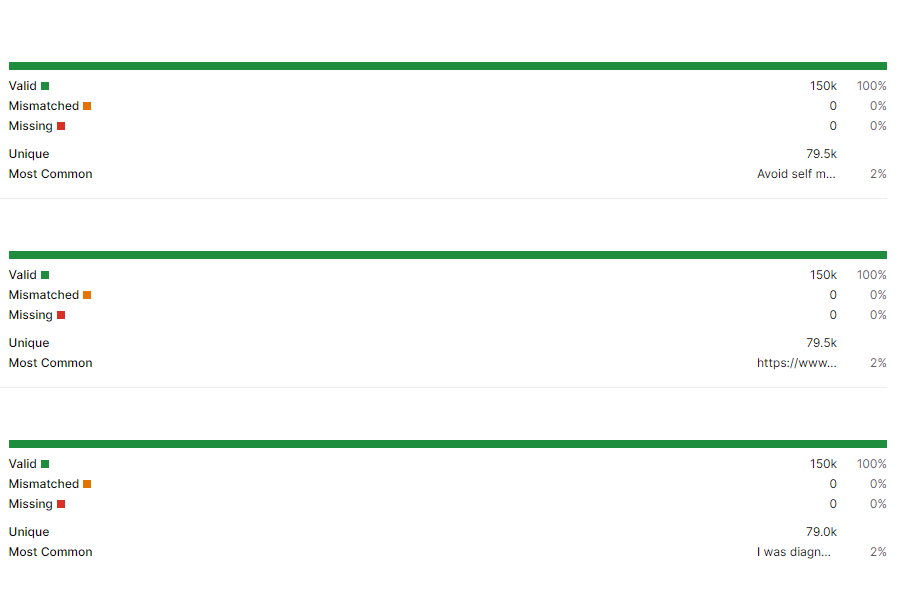
### Dataset

To train the bot to answer questions about healthcare, a list of collected questions and answers needs to be analyzed and processed. Based on the analysis, the final data to train the chatbot is organized and defined in train, test, and validation dataset files. Once the data is finalized, the chatbot needs to be fed all the data with the corresponding responses. Then the chatbot created needs to be tested and trained to fine-tune the results.

Below are some examples from the training dataset:



Also, details information about the dataset is shown below:



### Rough metric

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics and a software package specifically designed for evaluating summary, but that can be also used for machine translation as well as text generation. The metrics compare an automatically produced summary or translation against reference (high-quality and human-produced) summaries or translations.

ROUGE-N

ROUGE-N measures the number of matching between the model-generated text and a human-produced reference.

Consider the reference R and the candidate summary C:

- R: The cat is on the mat.

- C: The cat and the dog.

ROUGE-1

Using R and C, we are going to compute the precision, recall, and F1 scores of the matching n-grams. Let’s start computing ROUGE-1 by considering 1-grams only.

ROUGE-1 precision can be computed as the ratio of the number of unigrams in C that appear also in R (that are the words “the”, “cat”, and “the”), over the number of unigrams in C.

ROUGE-1 precision = 3/5 = 0.6

ROUGE-1 recall can be computed as the ratio of the number of unigrams in \_R\_ that appear also in C (that are the words “the”, “cat”, and “the”), over the number of unigrams in R.

ROUGE-1 recall = 3/6 = 0.5

Then, the ROUGE-1 F1-score can be directly obtained from the ROUGE-1 precision and recall using the standard F1-score formula.

ROUGE-1 F1-score = 2 \* (precision \* recall) / (precision + recall) = 0.54

ROUGE-2

Let’s try computing the ROUGE-2 considering 2-grams.

Remember our reference R and candidate summary C:

- R: The cat is on the mat.

- C: The cat and the dog.

ROUGE-2 precision is the ratio of the number of 2-grams in C that appear also in R (only the 2-gram “the cat”), over the number of 2-grams in C.

ROUGE-2 precision = 1/4 = 0.25

ROUGE-2 recall is the ratio of the number of 2-grams in R that appear also in C (only the 2-gram “the cat”), over the number of 2-grams in R.

ROUGE-2 recall = 1/5 = 0.20

Therefore, the F1-score is:

ROUGE-2 F1-score = 2 \* (precision \* recall) / (precision + recall) = 0.22

ROUGE-L

ROUGE-L is based on the longest common subsequence between our model output and reference, i.e. the longest sequence of words (not necessarily consecutive, but still in order) that is shared between both. A longer shared sequence should indicate more similarity between the two sequences.

We can compute ROUGE-L recall, precision, and F1-score just like we did with ROUGE-N, but this time we replace each n-gram match with the LCS.

Remember our reference R and candidate summary C:

- R: The cat is on the mat.

- C: The cat and the dog.

The LCS is the 3-gram “the cat the” (remember that the words are not necessarily consecutive), which appears in both R and C.

ROUGE-L precision is the ratio of the length of the LCS, over the number of unigrams in C.

ROUGE-L precision = 3/5 = 0.6

ROUGE-L precision is the ratio of the length of the LCS, over the number of unigrams in R.

ROUGE-L recall = 3/6 = 0.5

Therefore, the F1-score is:

ROUGE-L F1-score = 2 \* (precision \* recall) / (precision + recall) = 0.55

ROUGE-S

ROUGE-S allows us to add a degree of leniency to the n-gram matching performed with ROUGE-N and ROUGE-L. ROUGE-S is a skip-gram concurrence metric: this allows to search for consecutive words from the reference text that appear in the model output but are separated by one-or-more other words.

Consider the new reference R and candidate summary C:

- R: The cat is on the mat.

- C: The gray cat and the dog.

If we consider the 2-gram “the cat”, the ROUGE-2 metric would match it only if it appears in C exactly, but this is not the case since C contains “the gray cat”. However, using ROUGE-S with unigram skipping, “the cat” would match “the gray cat” too.

We can compute ROUGE-S precision, recall, and F1-score in the same way as the other ROUGE metrics.

### BLEU

BLEU, or the Bilingual Evaluation Understudy, is a metric for comparing a candidate translation to one or more reference translations.

Although developed for translation, it can be used to evaluate text generated for different natural language processing tasks, such as paraphrasing and text summarization.

The BLEU score is not perfect, but it’s quick and inexpensive to calculate, language-independent, and, above all, correlates highly with human evaluation.

How to compute the BLEU score

Consider the two reference answer R1 and R2 produced by human experts, and the candidate answer C1 produced by our chatbot system.

- R1: The cat is on the mat.

- R2: There is a cat on the mat.

- C1: The cat and the dog.

Computing unigrams precision

To express the quality of our translation with a metric, we may count how many words in the candidate translation C1 are present in the reference translations R1 and R2, and divide the result by the number of words in C1 to get a percentage. Therefore, a perfect score is 1.0, whereas the worst score is 0.0. Let’s call this metric BLEU.

In C1 there are three words (“the”, “cat”, “the”) that appear on the reference translations, thus:

BLEU\*(C1) = 3/5 = 0.6

The candidate translation is far from perfect, indeed it receives a score of 0.6. Everything looks fine.

The problem with repeating unigrams

Let’s compute the BLEU\* score of the new candidate translation C2:

- R1: The cat is on the mat.

- R2: There is a cat on the mat.

- C2: The The The The The.

This time our chatbot system is not very good, unfortunately.

Every word in C2 is present in at least one between R1 and R2, thus:

BLEU\*(C2) = 5/5 = 1

We achieved a perfect score with a non-sense translation, there’s something we need to correct on our metric.

It doesn’t make sense to consider the word “The” five times in the numerator, as it appears at most twice on each reference translation. We can try counting the word “The” only for the times it appears at most on each reference translation, that is two. Let’s call this new metric BLEU.

BLEU\*\*(C2) = 2/5 = 0.4

Now the score makes more sense, as we are accounting for the fact that a good translated word appears too many times on our candidate translation.

Considering n-grams

Let’s try computing the BLEU score on two other candidate translations C3 and C4 to check if everything looks fine.

- R1: The cat is on the mat.

- R2: There is a cat on the mat.

- C3: There is a cat on the mat.

- C4: Mat the cat is on a there.

The BLEU scores are the following:

BLEU\*\*(C3) = 7/7 = 1.0

BLEU\*\*(C4) = 7/7 = 1.0

Both candidate translations contain words that are present in the reference translations, therefore they both achieve the maximum score. However, C4 is not a well-formed English sentence.

A quick way to get higher scores for well-formed sentences is to consider matching 2-grams or 3-grams instead of 1-grams only. Let’s call BLEU\*\*₁ the score that considers only 1-grams and BLEU\*\*₂ the score that considers only 2-grams.

C3 has six 2-grams and they all appear on the reference translation R2, thus:

BLEU\*\*₁(C3) = 7/7 = 1.0

BLEU\*\*₂(C3) = 6/6 = 1.0

Instead, in C4 all the 2-grams don’t appear in any reference translation, thus:

BLEU\*\*₁(C4) = 7/7 = 1.0

BLEU\*\*₂(C4) = 0/6 = 0.0

It is generally said that the BLEU\*\*ₙ score for n-grams focuses on the sentence meaning for low n, and focuses on well-formed sentences for high n.

It has been found that the geometric mean of the BLEU\*\*ₙ scores with n between one and four has the best correlation with human evaluation, therefore it’s the score more commonly adopted. Let’s call it MEAN\_BLEU.

Penalizing short candidate translations

Let’s try now computing the BLEU\*\*₁ and BLEU\*\*₂ scores of the candidate translation C5:

- R1: The cat is on the mat.

- R2: There is a cat on the mat.

- C5: There is a cat.

The scores are:

BLEU\*\*₁(C5) = 4/4 = 1.0

BLEU\*\*₂(C5) = 3/3= 1.0

Looks like C5 achieves a perfect BLEU\*\*ₙ score for each n, even though the candidate translation is missing a piece of text with respect to the reference translations.

This can be avoided by adding a penalty for candidate translations whose length is less than the ones of the reference translations. We call it Brevity Penalty (BP).

The final BLEU score is:

BLEU = BP \* MEAN\_BLEU

That is, BLEU is the product of the Brevity Penalty BP (which penalizes short translations that don’t contain relevant text from the reference translations) and the geometric mean of the BLEU\*\*ₙ scores for n between one and four (which takes into account small n-grams, to capture the sentence meaning, and large n-grams, to get well formed sentences).

What is the value of BP?

If the length of the candidate solution is bigger than the length of the reference translation with the most similar length, then we shouldn’t penalize and therefore BP equals one. Otherwise, BP is a decaying exponential which is lower when the length difference between the candidate and the reference translations is greater. The BLEU suggests computing the brevity penalty over the entire corpus rather than over single translations to smoothen the penalties for short translations.

# CONCLUSION AND FUTURE WORK

## Conclusion

## Future work

# REFERENCES

# APPENDICES