Universidade de Aveiro
2023

JOÃO ANTÓNIO ASSIS REIS

UM CONSTRUTOR DE PESQUISA CONVERSACIONAL EM BASE DE DADOS CLÍNICAS MÉDICAS

A CONVERSATIONAL QUERY BUILDER ON MEDICAL CLINICAL DATABASES

PROPOSTA DE TESE

Universidade de Aveiro
2023

JOÃO ANTÓNIO ASSIS REIS

UM CONSTRUTOR DE PESQUISA CONVERSACIONAL EM BASE DE DADOS CLÍNICAS MÉDICAS

A CONVERSATIONAL QUERY BUILDER ON MEDICAL CLINICAL DATABASES

PROPOSTA DE TESE

"*The greatest challenge to any thinker is stating the problem in a way that will allow a solution*"

— Bertrand Russell

Universidade de Aveiro
2023

JOÃO ANTÓNIO ASSIS REIS

UM CONSTRUTOR DE PESQUISA CONVERSACIONAL EM BASE DE DADOS CLÍNICAS MÉDICAS

A CONVERSATIONAL QUERY BUILDER ON MEDICAL CLINICAL DATABASES

Proposta de Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à conclusão da unidade curricular Proposta de Tese, condição necessária para obtenção do grau de Mestre em Engenharia Informática , realizada sob a orientação científica do Doutor João Rafael Almeida, Professor associado do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro, e do Doutor José Luís Oliveira (co-orientador), Professor catedrático do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro.

Dedico este trabalho à minha esposa e filho pelo incansável apoio.

o júri / the jury

agradecimentos / acknowledgements

Agradeço toda a ajuda a todos os meus colegas e companheiros.

Palavras Chave

texto livro, arquitetura, história, construção, materiais de construção, saber tradicional.

Resumo

Um resumo é um pequeno apanhado de um trabalho mais longo (como uma tese, dissertação ou trabalho de pesquisa). O resumo relata de forma concisa os objetivos e resultados da sua pesquisa, para que os leitores saibam exatamente o que se aborda no seu documento.
Embora a estrutura possa variar um pouco dependendo da sua área de estudo, o seu resumo deve descrever o propósito do seu trabalho, os métodos que você usou e as conclusões a que chegou.
Uma maneira comum de estruturar um resumo é usar a estrutura IMRaD. Isso significa:

- Introdução
- Métodos
- Resultados
- Discussão

Veja mais pormenores aqui:
`https://www.scribbr.com/dissertation/abstract/`

Keywords textbook, architecture, history, construction, construction materials, traditional knowledge.

Abstract

An abstract is a short summary of a longer work (such as a thesis, dissertation or research paper).
The abstract concisely reports the aims and outcomes of your research, so that readers know exactly what your paper is about.
Although the structure may vary slightly depending on your discipline, your abstract should describe the purpose of your work, the methods you've used, and the conclusions you've drawn.
One common way to structure your abstract is to use the IMRaD structure. This stands for:

- Introduction
- Methods
- Results
- Discussion

Check for more details here:
`https://www.scribbr.com/dissertation/abstract/`

Contents

List of Figures

List of Tables

Lista de Excertos de Código

*

CHAPTER 1

Introduction

The continuous quest for medical answers and advancements in clinical research, combined with the diversity of medical databases, has sparked complex challenges for researchers. A recurring issue is the scarcity of specific medical data that are the focus of a study, such as cases of patients with rare diseases. In this regard, a promising strategy has emerged, which consists of integrating multiple and diverse medical databases.

However, the implementation of this strategy is not free of obstacles, with the issue of heterogeneity between databases being a prominent challenge. In other words, databases contain different types, formats and/or sources, which are often not compatible with each other. The existence of these diverse data is not very effective, as they cannot be easily shared or integrated with other data.

It is in this context that the OMOP CDM (Observational Medical Outcomes Partnership Common Data Model) and the OHDSI (Observational Health Data Sciences and Informatics) initiative have emerged as good solutions to the problem of heterogeneity and interoperability among clinical medical data. Generally speaking, OMOP CDM is a common data model that establishes a universal standard for representing patient clinical information, allowing for interoperability among disparate databases. The OHDSI initiative is, in turn, an international collaboration composed of researchers and scientists committed to the mission of developing analytical, open-source solutions for an extensive network of medical databases, following systematic analysis of this heterogeneous data.

With the assistance of OMOP CDM and OHDSI, the challenges of data interoperability are overcome, enabling the discovery of crucial insights for the advancement and improvement of medical studies. Sharing this data presents numerous advantages for researchers, including promoting new fields of study and a significant increase in the impact and recognition of research results.

1.1 Motivation

The search for data sources of interest for a researcher's study can be complex due to the large number of databases in the community. To face this challenge, some of these databases are grouped into database catalogues. This strategy consists of characterizing the data by aggregating data and metadata.

The EHDEN (European Health Data Evidence Network) portal is an excellent example of a platform that provides a catalogue of medical databases from across Europe. It is a centralized repository that facilitates the discovery of relevant data sources for researchers.

Despite the assistance provided by the catalogue offered by EHDEN, identifying the most suitable databases for a specific study remains a challenge. Thus, to facilitate search in the catalogue, Networkdashboards has emerged, offering statistical and aggregated information about the databases available on the EHDEN network. With this tool, researchers can filter the most suitable databases for their research needs and make more informed decisions.

Even with all this help, choosing the most appropriate databases is difficult and time-consuming, making it difficult to achieve the ideal search desired for the study. The challenge to be addressed is to assist a medical researcher in reaching the ideal search based on the protocol and parameters of their study.

1.2 Goals

How can a conversational query builder support medical researchers when defining a study protocol?

Goals:

1. study of state-of-the-art
2. developed a chat-like search engine to help discover the best databases for a study
3. enhance the engine to collect additional information to provide a query as outcome.

1.3 Dissertation outline

Que resume a estrutura do documento.

CHAPTER 2

State of Art

2.1 Information Retrieval

In computing and information science, Information Retrieval (IR) involves retrieving information from a database or multiple databases. According to **p_m_efficient_2021**, the IR system requires users to input queries, retrieving pertinent information from the database that aligns with the users' needs. Thus, this prevents the user from accessing a massive amount of information.

In conformity with **hambarde_information_2023**, conventional text retrieval systems were predominant in the initial stages of the IR field. These systems mainly depended on matching terms between queries and documents. Nevertheless, these systems based on terms have limitations, including issues like polysemy, synonymy, and linguistic gaps, which may restrict their effectiveness.

With the advancement of technology, deep learning techniques emerged, improving conventional text retrieval systems and overcoming the constraints associated with term-based retrieval methods. For this reason, the performance of these systems increased significantly, resulting in a more accurate and streamlined retrieval of information for end-users.

In turn, deep learning methods have evolved. Neural Network Architectures, transfer learning, and pre-training techniques emerged. These approaches have advanced the representation of textual data and bolstered the IR system's comprehension of natural language queries.

More recently, Transformer architectures with attention mechanisms have been implemented in IR systems to enable concentration on crucial query segments and documents for improved matching. Moreover, incorporating pre-trained language models like Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT)-2 has proven to enhance the performance of IR systems, offering an advanced understanding of the semantics and context within natural language queries and documents.

This field has many applications in the real world. **p_m_efficient_2021** highlights the following: streamlined and adaptable indexing and retrieval, information extraction, semantic matching, and multimedia retrieval. IR generally functions across three main scales: searching the web, retrieving personal information, and conducting searches for enterprises, institutions, and domain-specific contexts.

In this section, I explored the IR field and how it can be applied with Natural Language Processing (NLP).

2.1.1 Overview of Information Retrieval Systems

According to **hambarde_information_2023**, an IR system can be separated into two stages: Retrieval and Ranker. The following figure, adapted from **hambarde_information_2023**, shows us an overview of modern IR systems, highlighting the two main stages.

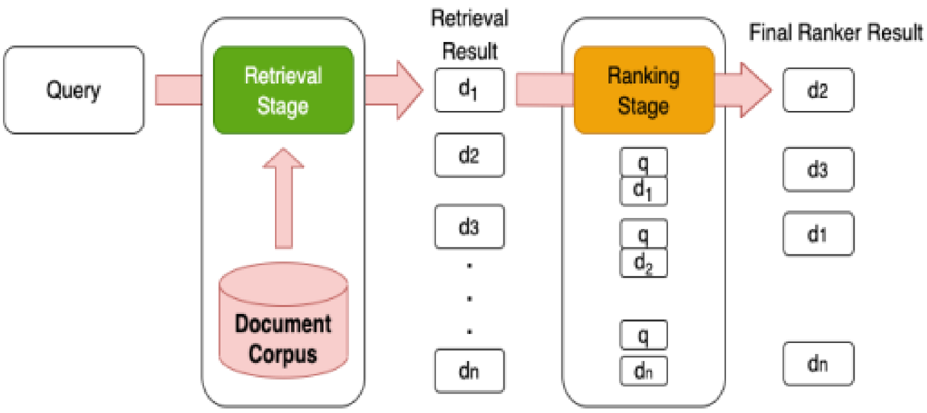


Figure 2.1: Overview of modern IR system [REFAZER IMAGEM]

After analyzing the query, the retrieval stage will select an initial set of documents that are potentially pertinent to the query. Subsequently, the relevance of these documents undergoes reassessment through the similarity scores. The ranking is then refined using diverse algorithms and models, including the vector space model, Latent Semantic Indexing, Latent Dirichlet Allocation, and pre-trained models such as BERT.

This is followed by the ranking stage, in which the primary objective is to adjust the order of the initially retrieved documents based on their relevance scores. This phase prioritizes the enhancement of result effectiveness rather than efficiency. In the end, it returns a rank of documents as close as possible to the user's query criteria.

2.1.2 Traditional Methods

Some successful classical methods are ?? and ??, briefly explained next.

TF-IDF

To understand the ?? method, first, it is necessary to understand the concepts of Term Frequency (TF) and Inverse Document Frenquency (IDF). **manning_introduction_2009** explained these concepts as follows.

It is reasonable to assume that a document containing a query term more frequently is more relevant to that query. Therefore, it should be assigned a higher relevance and/or score. So, TF is the number of term occurrences in a document.

However, to evaluate the relevancy of a query, each term is regarded with equal importance, and this is the problem with the raw method explained above. **manning_introduction_2009** clarified this with the following example: the automotive industry is expected to include the term "auto" in nearly every document. The IDF calculates the rarity of a term across a set of documents. This measure is calculated as the logarithm of the inverse fraction of documents containing the term. The goal is to help prioritize some terms that are sporadic and possibly more informative.

The traditional IR method, ??, combines TF and IDF definitions, as the name suggests, and then produces a weight for each term in each document, as **manning_introduction_2009** and **chizhik_challenges_2020** mention. The weight is calculated as the product of TF and IDF values, highlighting terms that are both important within a specific document and relatively uncommon in the document collection.

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t .$$

Figure 2.2: Overview of modern IR system [REFAZER IMAGEM]

manning_introduction_2009 noted the ?? weight assigned to a term in a document is highest when the term frequently appears in a few documents, providing discriminating solid power. The weight is lower when the term occurs less frequently in a document or is widespread across many documents, indicating a less pronounced relevance signal. The weight is at its lowest when the term is present in nearly all documents.

In summary, this IR method evaluates the importance of a term within a document relative to its occurrence across a collection of documents.

BM25

??, the short form for Best Matching 25, is a ranking algorithm for IR systems, especially in the context of search engines. It builds upon the ?? model to provide more accurate and context-aware document ranking. **hambarde_information_2023** noted that ?? and other initial retrievers are employed for their effectiveness in recalling pertinent documents from an extensive pool.

The core components of ?? include TF, IDF, Document Length (DL), and tuning parameters. Recapping from the ?? section, TF is the number of occurrences that a specific term is in a document, and IDF is a measure that indicates the importance of a term in the whole document.

$$\sum_{i}^{n} IDF(q_i) \frac{f(q_i, D) * (k1 + 1)}{f(q_i, D) + k1 * (1 - b + b * \frac{fieldLen}{avgFieldLen})}$$

Figure 2.3: BM25 equation

The ?? equation is composed of the ith query term (q) and the respective IDF and TF values. Also, include a division between the DL, represented in the formula as fieldLen, and the average document length, avgFieldLen. This ratio evaluates how much the length of the document field deviates from the average length. So, the [https://www.elastic.co/blog/practical-bm25-part-2-the-bm25-algorithm-and-its-variables] explained intuitively: a document tends to receive a lower score when it contains more terms, especially those that do not match the query. The value b is a fine-tuning parameter, and it is responsible for length normalization. When b is larger, the ratio has a more significant effect on the overall score. Finally, the k1 value means term frequency saturation. It is a fine-tuning parameter that prevents the term frequency component of BM25 from having an unlimited impact on the document score.

[https://medium.com/@evertongomede/understanding-the-bm25-ranking-algorithm-19f6d45c6ce]

This algorithm is simple and effective in IR tasks, mainly search tasks. Also, it can handle vast collections. For these reasons, it is widely used and called a classic.

However, ?? can not perform a semantic analysis of the query and the documents, so getting the context and, in turn, getting better results is challenging. Another limitation is the ignorance of other crucial factors to get a better search beyond the factors relative to the TF and DL.

2.1.3 IR in Question Answering

Question Answering (QA) and IR are closely related fields, together with NLP. According to **zhong_building_2020**, QA aims to give users accurate and prompt responses to posed questions.

The traditional approach to question analysis and answering often involves mapping questions into predefined templates, such as "What-type" and "How-type". While widely utilized by existing online question-answering search engines, this template-based approach faces limitations in handling multiple questions.

So, with the advancement of technology, another approach emerged: deep learning-based question-answering. In contrast with the traditional approach, this approach employs deep learning techniques, like convolutional neural networks (CNN), to offer automatic representation and analysis of questions. These neural models, trained through end-to-end approaches, excel in extracting and understanding complex characteristics in textual documents.

Recently, deep learning approaches with attention mechanisms and transfer learning have enhanced the flexibility of representation in text classification and named entity recognition. **zhong_building_2020** highlights the tool BERT. BERT has emerged as a powerful model, utilizing contextualized representations for transfer learning. BERT-based models showcase performance in question-answering tasks, even in domains like medicine.

Natural Language Processing (NLP)

NLP is the basis for building a QA system. It is a field of Artificial Intelligence (AI) whose primary goal is to understand, interpret, and generate human language. The NLP can be divided into two major components: Natural Language Understanding (NLU) and Natural Language Generation (NLG), according to **ayanouz_smart_2020**

The **NLU** component plays a crucial role in processing and transforming unstructured data into a format the system can comprehend seamlessly. Essentially, the function of NLU is to identify topics and entities, identify the intention, and determine the structure and syntax of the sentence.

In agreement with **ngai_intelligent_2021**, for easy understanding by the chatbot, the user queries can be processed by semantic analysis, pragmatic analysis, and syntactic analysis. **ayanouz_smart_2020** explained these steps and added two more necessary steps to make it easier to understand: a lexical analysis and discourse integration.

- **Lexical Analysis**: This step involves analyzing and identifying the structure of words. It breaks down the text into chapters, sentences, phrases, and words. **chizhik_challenges_2020** defined lexical analysis as the pre-processing of the text that follows the following steps: tokenization, removal of special characters, links, and punctuation, and removal of stop-words.
- **Syntactic Analysis**: The syntactic analyzer parses the grammar and arrangement of words, making the relationships between different words more explicit. Essentially, it rejects sentences with incorrect structures. This analysis can be seen as the process of normalizing tokens.
- **Semantic Analysis**: This step ensures the text is meaningful and interprets its correct meaning by mapping syntactic constructions. It ensures that only semantically valid content is retained. The recognition of entities is part of this analysis.
- **Pragmatic Analysis and Discourse Integration**: This step analyzes the overall context to derive the conclusive interpretation of the actual message in the text. It considers factors like the true meaning of a phrase or sentence based on the broader context.

The other component is **NLG**. Encompassing text realization and planning, language generation is tasked with crafting coherent and linguistically accurate responses. Simply put, it grapples with the challenge of navigating the intricacies of natural human language.

2.2 Large Language Models

It is crucial to trace briefly the development history to understand the concept of Large Language Models (LLM). **liu_prompting_nodate** explained this simply and intuitively.

Before LLM, there were only simple Language Models (LM), which, in the initial stage, utilized a fully supervised learning approach, where task-specific models were trained on the target task dataset exclusively. Most of these were predictive models based on probabilities and Markov assumptions, also known as Statistical Language Models (SLM). This was heavily dependent on feature engineering. Afterwards, as deep learning gained prominence, an architecture designed to learn data features automatically, in other words, neural networks for NLP emerged to enhance LM's capabilities. Integrating feature learning and model training, Neural Language Models (NLM) established a comprehensive neural network framework applicable to diverse NLP tasks.

Most recently, in 2017, the launch of the self-attention mechanism revolutionized this field and the Transformer architectures have become increasingly popular. These deep-learning architectures led to the development of pre-trained models not explicitly designed for a particular task, including BERT and GPT, collectively known as Pre-trained Language Models (PLM)). PLM have shown significant performance enhancements across a diverse array of NLP tasks.

Following this, the researchers have involved the scale of model parameters, and the paradigm of "Pre-train, Prompt, Predict", like **liu_prompting_nodate** call, gained widespread acceptance. So, in terms of interaction with LM, the prompts became crucial. Researchers name these PLM with hundreds of billions of parameters as LLM. Prompts effectively allow LLM to deal with a large number of complex and diverse tasks without a lot of effort.

In this section, I defined a LLM, [TO DO]

2.2.1 Definition

LLM are revolutionizing NLP and AI research. LLM are an AI created to comprehend, generate, and engage in human language interactions. Essentially, these advanced AI systems can mimic human intelligence. These models have a notable ability in natural language tasks, such as text generation and translation, QA, decision-making, summarization, and sentiment analysis.

These models can process and predict patterns with great accuracy due to their significant model parameters, often comprising hundreds of billions of parameters. **hadi_LLM_2023** combine sophisticated SLM and deep learning techniques to train, analyze and understand huge volumes of data, learning the patterns and relationships among the data. For this reason, according to **naveed_comprehensive_2023**, when provided with task descriptions and examples through prompts, LLM can produce textual responses to task queries. So, we can put LLM in the generative AI field.

liu_prompting_nodate say that the release of ChatGPT 1 garnered significant social attention, and research into LLM has evolved. This has led to the development of noteworthy products like PaLM, GPT-2, GPT-3 and, most recently, GPT-4, and LLaMA and LLaMa-2.

2.2.2 Architecture Overview

Transformer Architecture

The development and advancement of LLM is thankful for the introduction of Transformers by **vaswani_attention_2023** in 2017. Most LLM are built on the Transformer model, which is based on a self-attention mechanism and encoder-decoder structures. This new technology enables parallelization and efficient handling of long-range dependencies, according to **hadi_LLM_2023**, and led to the development of models that have achieved enormous results, such as GPT by OpenAI and BERT by Google.

The innovation of this model is due to the self-attention mechanism, one of the key components. It allows the model to weigh the importance of different words in a sequence when processing each word. This mechanism enables the model to focus on relevant information, capturing dependencies regardless of word order.

Another key component is the Encoder and Decoder Stacks. Essentially, the encoder processes the input sequence, and the decoder generates the output sequence. Each stack contains 6 similar layers and these layers apply the attention mechanism.

Since the model doesn't have recurrence and convolution to understand the order of the input sequence, another component, Position Encoding, provides some information about the position of the tokens in the sequence. This is crucial for capturing sequential information in the data.

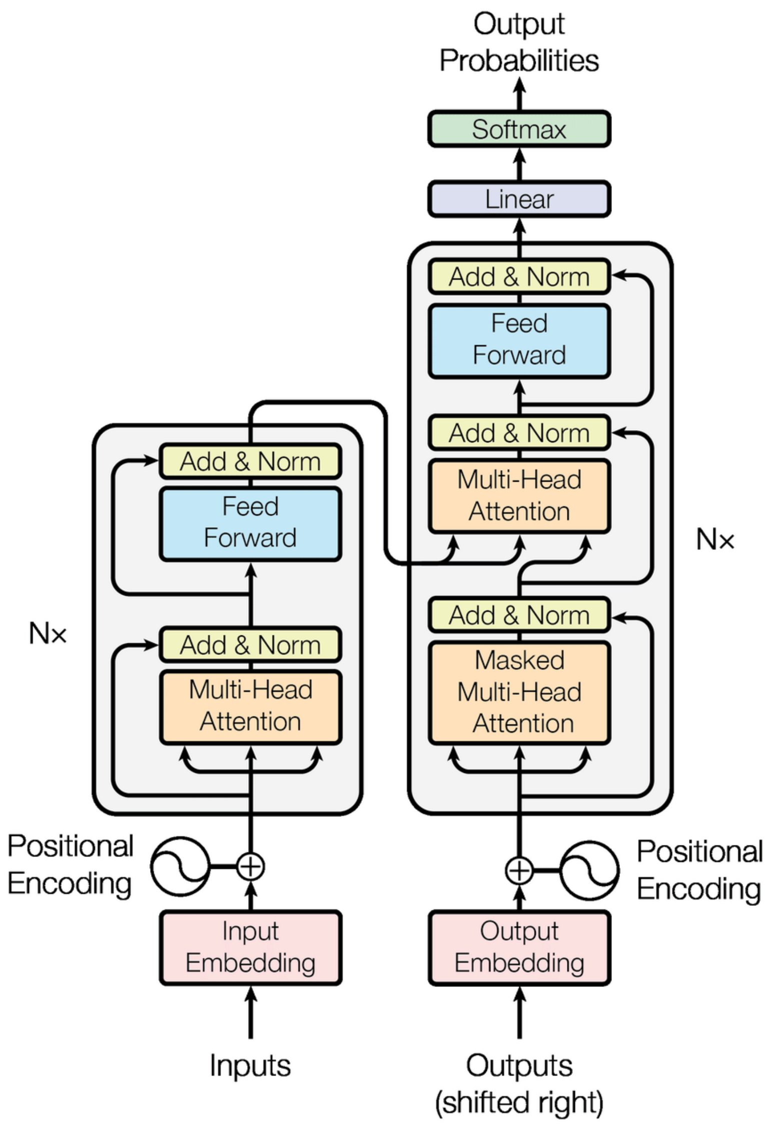


Figure 2.4: The Transformer architecture. From **vaswani_attention_2023**

Pre-training

hadi_LLM_2023

Learning the patterns and relationships among the data starts with the pre-training process. First, the LLM needs to access a vast volume of textual data from multiple sources. The goal of this phase is to predict the succeeding word in a sentence based on the context given by the previous words through unsupervised learning.

To achieve this, it is necessary to prepare and preprocess the data before the training stage. First, demand quality filtering from the training corpus. It is vital to remove unwanted, repetitive, superfluous, and potentially harmful content from the massive text data. Then, according to **hadi_LLM_2023**, duplicate data in a corpus make less diversity of LLM. So, the duplication of data will make the training process unstable, impacting the overall performance of the model. Next, it is necessary to pay attention to privacy. The data could have sensitive or personal information, so it is vital to address privacy concerns by removing this information from the pre-training corpus.

An important step, the tokenization, follows this. This step aims to divide the unprocessed text into sequences of individual tokens, which are subsequently input into LLM. Moreover, it is vital in mitigating the computational load and enhancing efficiency during the pre-training phase.

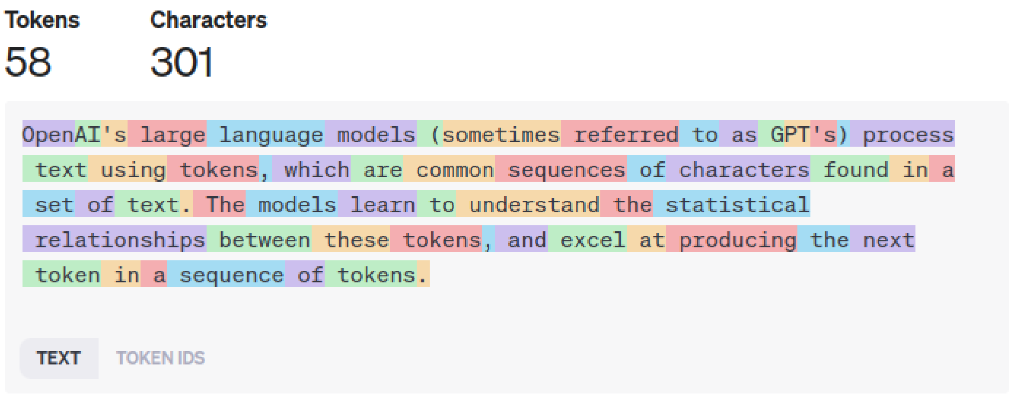


Figure 2.5: Tokenization process visually explained by OpenAI **noauthor_openai_nodate**

After the pre-training process, the LLM goes through a fine-tuning phase.

Fine-tuning

hadi_LLM_2023

During pre-training, models are generally trained with the objective of next token prediction, learning the nuances of language structure and semantics. The fine-tuning phase involves adapting a pre-trained model to specific tasks and aligning it with human preferences.

In this stage, the model is presented with labelled data to produce responses more contextually accurate for the specific task at hand. Fine-tuning enables the LLM to specialize in diverse applications, ranging from language translation and question-answering to text generation.

Parameters

The parameters in neural networks function as learnable components that include weights and biases. These values are adjusted during the training process to minimize the difference between the model's predictions and the actual target outputs.

The more parameters LLM have, the more flexibility it has in capturing patterns and relationships in data, but the risk of overfitting increases and it needs more computational power.

So ....

2.2.3 Comparation between LLM

2.2.4 Limitations

2.3 Conversational Virtual Assistant

Conversational Agents, also known as chatbots, chatterbots, or virtual assistants, have become a vital aspect of the digital landscape. These tools are generally dialogue systems that understand, interpret, and generate human language, enabling them to communicate with users to dissolve their questions.

Chatbots are increasingly being used in various contexts due to their many benefits. These aspects that make companies bet on the use of chatbots are the continuous availability to support and assist the customer, ensuring more consistent support, the cost-efficiency by reducing the human customer support, the time-saving both for the organization and for customers due to the immediate responses to the user queries, the ease and intuitiveness of this systems, improve service with every interaction. Because of this, the utility of the chatbots as tools is increasing as the technology advances.

The rise of conversational virtual assistants is underpinned by a convergence of technologies, including NLP, ??, and AI.

In this section, I explored the implementation of chatbots, their components, and existing tools.

2.3.1 Overview of Conversational Virtual Assistants

nuruzzaman_survey_2018 separated chatbot applications based on their main features and functionalities. There are four chatbot models: goal-based, knowledge-based, service-based, and response generated-based. Goal-based chatbots are designed for particular tasks and structured to engage in concise conversations to collect user information for task completion.

borah_survey_2019 said that the process of response generation is not the same for all chatbots and distinguished them into four models: Retrieval-Based, Generative-Based, Long and Short Conversation, and Open or Close Domain. **chizhik_challenges_2020** are in line with **borah_survey_2019**, and divided chatbots into three categories based on response generation architectures: rule-based chatbots, retrieval-based chatbots, and generative-based chatbots.

According to **chizhik_challenges_2020**, a rule-based chatbot examines fundamental features of the user's input statement and generates a response based on a predetermined set of manually crafted templates.

Conforming to **chizhik_challenges_2020** and **borah_survey_2019**, a retrieval-based chatbot picks a response from an extensive precompiled dataset. It selects the most promising reply from the top-k ranked candidates. Thus, they refrain from producing new text. It has limited flexibility regarding domain since they are usually applied to one domain, and in terms of errors, because, for example, the user cannot make grammatical mistakes.

A generative-based chatbot generates a text sequence as a response rather than choosing it from a predefined set of candidates. These chatbots are very flexible and can handle open domains because they are implemented with Machine Translation techniques. The interactions will be more identical to those of humans, as it implements a self-learning method from a large quantity of interaction data. However, this could be complex and costly to implement.

Beyond that, **borah__survey__2019** compare long and short conversations. He concluded that a more extended conversation requires saving what has been said, which makes it challenging to automate, unlike a short conversation.

Also, **borah__survey__2019** defined the differences between chatbots with opened or closed domains. In an open-domain environment, conversations can go in any direction without a predefined goal or intention. Crafting meaningful responses in such contexts demands an extensive breadth of world knowledge spanning countless topics. Conversely, closed-domain systems have more constrained inputs and outputs, focusing on specific goals. Consequently, many chatbots are inherently closed-domain, designed with a clear objective.

2.3.2 Common System Architectures

Most chatbot implementations apply standard components, such as NLP, dialogue, knowledge and data storage modules. According to **ngai__intelligent__2021** and to **dilmegani__cem__how__nodate**, the function of each module can be summarized as follows.

The dialogue module is in charge of handling the conversation flow. To effectively communicate with the user, conversation agents must understand the human language using a NLP Engine to decide what to do with the intention found.

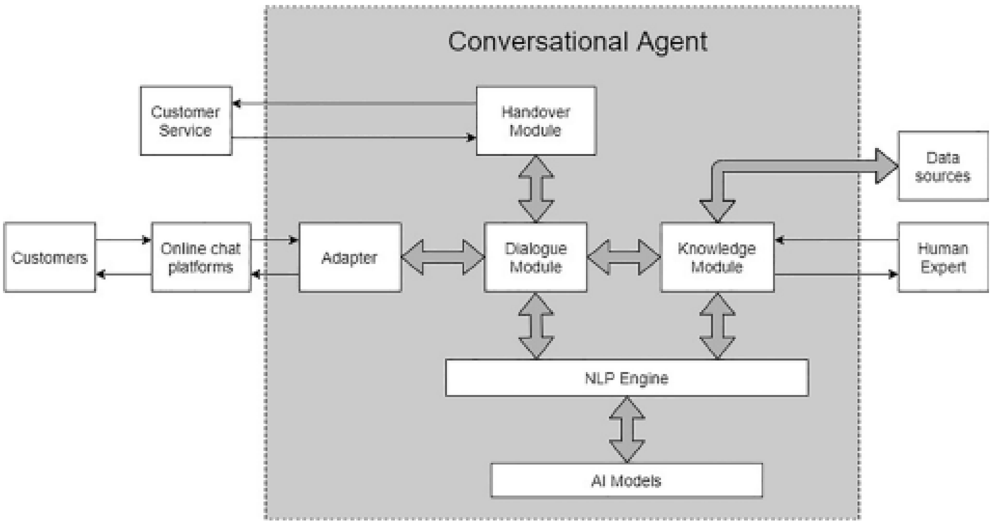


Figure 2.6: [REFAZER IMAGEM] Overview of a conversational system architecture From [An intelligent knowledge-based chatbot for customer service]

The knowledge module is the source of data and knowledge of the chatbot. After knowing the user's intention, this module will retrieve the data to respond to the user.

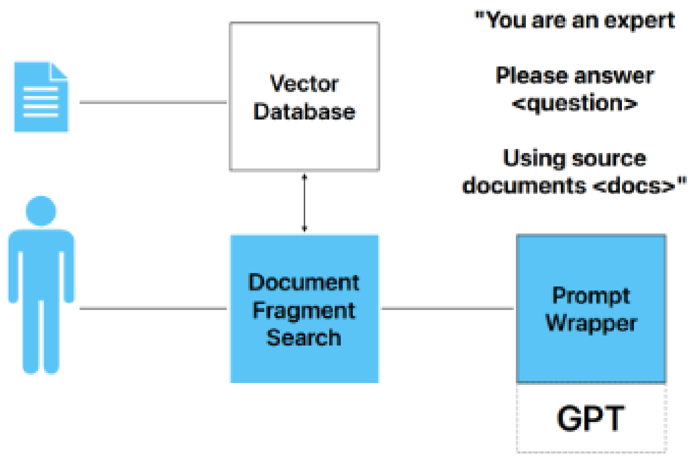
Many chatbots use a knowledge base for the knowledge module. Conforming to **pereira__querying__2023**, this choice is justified because it is advantageous to have the data organized semantically. This establishes a coherent structure for structured and unstructured data, simplifying the deduction of new knowledge.

Additionally, some chatbots are integrated with web scrapers to pull data from online resources and display it to users

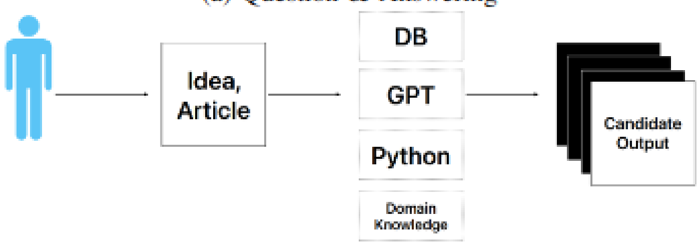
2.3.3 Generative-Based Chatbot

2.3.4 Prompt Engineering

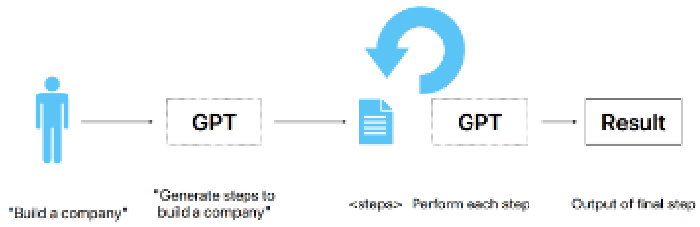
2.3.5 Retrieval-Augmentation Generation



(a) Question & Answering



(b) Creative Applications



(c) Multi-step Planning

Figure 2.7: [REFAZER IMAGEM] Templates for LLM-based application development. GPT is taken as an example scenario representing LLMs From **hadi_LLM_2023**

2.4 Interactive Query Builder

2.5 Insights/Summary

Não há muito trabalho feito para o desenvolvimento de Conversational Queries Builder.

» limitations:

* não existe query builder para dados médicos

* muitos chatbots com o seguinte flow: NLP -> query struct -> DB, que não é o caso a ser implementado

* há algumas limitações dos chatbots a ter em conta, destaco os erros gramaticais, erros ortográficos e a ambiguidade de textos

* Recommender System: Existing chatbots do not ask questions, explain, or advise the user topic. They collect information and provide responses from the knowledge base. The chatbot should be able to write questions based on previous answers. [referencia: Chatbot Implementation in Customer Service Industry through Deep Neural Networks]

* alucianação dos LLM

» oportunidades:

* a união entre as áreas de NLP e de IR pode mostrar-se muito promissora

* vários modelos de AI em NLP sobre várias sub-áreas: Query understanding, Text Classification e Text Generation