International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Advancements of Emerging Trends in Multidisciplinary Analysis and Research
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Statistical Language Models for Customer Services

Khonkulova Nilufar Ravshanovna

English Applied Translation Department, Translation Faculty
Uzbekistan State World Languages University

ABSTRACT
The article deals with modern language models for customer services. There is history, development and main components of language models are considered as well.

KEYWORDS: *language model, word sequences, software algorithms, statistical methods, neural networks, nuances of language, chatbots, text generation.*

Introduction

A language model is a statistical model that is used to predict the probabilities of word sequences in a language.

Language models (LM) are software algorithms that analyze and generate text based on the studied material. The main task of such models is to understand the structure and meaning of the text in order to be able to continue phrases, answer questions, translate texts, and perform many other tasks related to natural language processing (NLP).

The first steps in creating language models were made in the middle of the 20th century, when scientists began experimenting with various statistical methods for text processing. Early models, such as n-gram models, were based on probabilistic methods: the probability of a word's occurrence was determined by its frequency in the text. These models had significant limitations, since they could not effectively take into account the context of the material.

With the development of computing technology and the emergence of new algorithms, the situation began to change. In the 2000s, an important stage was the introduction of neural networks, which made it possible to create more complex and accurate generative language models. The advent of recurrent neural networks (RNNs) and long short-term memory (LSTMs) was a major breakthrough, as these architectures could account for long dependencies in text.

Research methodology

Modern language models, such as YandexGPT, OpenAI's GPT-4, Google's PaLM 2, and others, are complex neural network architectures consisting of tens or even hundreds of billions of parameters. They are trained on huge amounts of text data, which allows them to capture subtle nuances of language. One of the key technologies underlying modern LMs is the Transformer architecture, proposed in 2017.

Transformers use an attention mechanism that allows the model to focus on different parts of the text, assessing the importance of each for the current task. This makes them especially effective at processing long texts and understanding complex contexts.

The development of large language models (LLMs) is a joint effort between linguists and data scientists. Data scientists need to be proficient in Python and have a solid mathematical background. For example, in the course "Data Science Specialist", students learn to work with pandas, Scikit-learn, Catboost - Python libraries that specialists use every day. Big language data (hereinafter - LLM) can be defined as a type of artificial intelligence that imitates the work of human intelligence. Their work is based on the use of advanced statistical models and deep learning methods to process and understand huge volumes of text data. LLMs study complex patterns and relationships present in the data, which allows them to generate new content imitating the stylistic features of the author's linguistic personality or a given genre. Currently, this type of artificial intelligence is the basis of chatbots, which are gaining unprecedented popularity in various fields, both in the entertainment industry and in medicine, education, and financial analytics. Such successful integration of artificial intelligence into our everyday life is due not only to specific reasons, but also to a broader, philosophical view of modern science on the subject, which lies in the close interaction of two scientific disciplines - programming and linguistics. Thanks to this, we have received a form of artificial intelligence that is capable of capturing statistical patterns and linguistic nuances present in training data.

Analysis

The basic principle of large language models can generally be divided into five stages. First, in the pre-training stage, the language model learns a large amount of text from various sources, and since the training is unsupervised, it learns to predict the next word in a sentence based on the context of the previous words. This helps the model develop an understanding of grammar, syntax, and semantic relationships. Typically, the information sources that are used at the beginning of the pre-training process are divided into two categories for understanding: general data and specialized data. After collecting a huge amount of text data, it is pre-processed to create a pre-training corpus by removing low-quality, redundant, or potentially harmful material. The second stage is filtering, which removes low-quality and unwanted data from the training corpus using language, statistical, and keyword filtering. According to modern research, duplicates in the corpus reduce the diversity of the language model, which destabilizes the learning process and therefore affects the performance of the model. To avoid this, redundant data is removed. Next, issues related to the use of web data for pre-training language models must be addressed, as such data often includes user-generated content containing sensitive or personal information, which entails potential privacy violations. The data processing process is completed by tokenizing the raw text into sequences of individual segments, which are then fed to the LLM.

After pre-training, the model is trained for a specific task or work in a specific domain. At this stage, models are provided with labeled examples so that they can then generate more accurate and contextually correct answers to the target task. Fine-tuning allows the model to be used in applications that translate between foreign languages, Q&A chats, or text generation.

Thus, it is by processing and analyzing large volumes of text that language models learn to understand language at more complex levels, generating coherent responses for a specific cognitive task. At the inference stage, when interacting with the LLM, the user enters a prompt or query. The model processes the input data and generates a response based on the acquired knowledge and the available context. Therefore, it is necessary to understand the linguistic criteria for formulating a request or prompt, since the response of the language model will depend on this. Undoubtedly, due to its development potential, ChatGPT is the most popular language model in the world. Provided that its artificial intelligence develops, it will be possible to actively implement it in healthcare work. At the moment, there are already similar examples: the XrayGPT language model, which can analyze open-ended questions asked by a patient about the results of his or her X-ray, and also give answers to them.

GPT plugins can expand the capabilities of artificial intelligence to understand natural language. They are the latest way to interact with ChatGPT, significantly expanding its functionality, namely allowing developers to create their own applications that open up new features for users, with the ability to integrate them into ChatGPT. Such plugins can be used when it is necessary to access external data sources, automate tasks, and also to improve the user experience [7]. In the era of active use of ChatGPT and the emergence of various plugins, it is especially worth highlighting the OpenAI plugins, with the implementation of which ChatGPT was able to interact with third-party data sources and knowledge bases. At the time of writing, OpenAI has not yet provided all developers with access to plugin development, but several use cases are already known, for example, Expedia, FiscalNote, Instacart, KAYAK, Klarna, Milo, OpenTable, etc. Plugins have fully realized the potential of ChatGPT to compose and execute complex tasks such as sentiment analysis for any source on the web. Additionally, working with these plugins allows it to answer queries based on updated information from the web that might not have been previously included in its training data, thus increasing the reliability of its answers.

The richness and scope of developer flexibility is a notable feature of working with plugins, which is why there is increasing attention to their development. In addition to the aforementioned early contributors, OpenAI has already contributed three plugins: a web browser plugin, a Code Interpreter, and a knowledge-based search plugin. The first one exposes ChatGPT to the web to collect information that it can use to answer a query submitted by a user. This plugin allows ChatGPT to bypass the time limitation of its own training data and use the most up-to-date information on the web via the Bing Search API and a text-based web browser. Code Interpreter, in turn, allows you to run Python code directly in the chatbot interface, with the ability to use it to perform logical calculations, as well as to write code. The interpreter can understand the language model of the problem description in human languages and then use it as input for developing Python code to solve the problem.

The third search plugin with a connection to the knowledge base is also open source, which allows ChatGPT to access the data and then collect the necessary, relevant information from this data, while requests are submitted in simple human language. Technologically, this plugin can work with built-in OpenAI elements, as well as with a set of databases for indexing or searching in documents.

Discussion

Currently, new techniques for managing the behavior of large language models in order to obtain the desired result for the user without updating the models themselves are of interest - the so-called "promt engineering". Although a formal definition has not yet been formulated, this direction is promising in developing and stylizing prompts provided to large language models in order to obtain the desired answer by the user. The main criterion for increasing the efficiency of working with large language models is the correct formulation of the query - prompt. There are also some formal methods, such as explicit instructions (providing the LLM with a clear instruction to do something), system-specific instructions (asking the LLM to answer a question), formatting with an example (providing a sample question and answer to it and asking the LLM to provide the answer in the same way), control markers (using special keywords in the prompt in order to help the LLM provide an answer taking into account the provided criteria).

Promt engineers study and develop various linguistic patterns (models) of interaction between humans and artificial intelligence that help users use a chatbot more effectively. In this process, it is important to understand the role of linguistics, especially syntax, stylistics, and lexicology. Stylistics, along with an understanding of extralinguistic issues, is used at the initial stage, when it is necessary to define a role by making the first request. A correctly formulated request, for example, “act as an experienced lawyer,” will make the language model perform tasks as a person with the relevant competencies would do. Stylistic features must also be kept in mind in the case of so-called “explanatory” prompts, for example, requests of this kind: “explain the principles of probability theory as if I were seven years old.” Since the scope of application of large language models is only expanding, in the future, it is necessary to focus scientific research on improving the accuracy and performance of these models, working with their limitations, and exploring potential ways of their application. Basic Components of Language Models

- Corpus of texts: This data can include books, articles, web pages, and other sources.
- Model architecture: A set of algorithms and structures that define how the model will process and analyze text.
- Training: The process of tuning the model parameters based on the analysis of texts from the corpus. During training, the model learns to predict the next word or phrase given the context.
- Inference: Applying the trained model to various tasks such as text generation, translation, sentiment analysis, and more.

Large language models also have their limitations. They can generate text that appears meaningful but contains factual errors. They can also be susceptible to biases present in the data they are trained on. This means that if there is a bias in the source data, the model may reproduce it in its responses.

Types of models

There are several types of language models, each designed to solve specific problems in NLP. These models vary in their architectures, training methods, and applications.

Statistical language models

- N-gram models. In these models, the probability of a given word occurring depends on the previous n words. For example, in a bigram model, the probability of a word occurring is determined only by the previous word. This is a simple and effective method, but it has significant limitations as it does not take into account long-range dependencies and can quickly become computationally inefficient as n increases.
- Markov chains. These models are based on the idea that the future state (the next word) depends only on the current state. Markov chains also suffer from a lack of long-range context, but they were the basis for many early natural language processing systems.

As a conclusion, we can say where are language models used?

Machine translation:

- Automatic translation. Modern language models are capable of translating texts from one language to another with a high degree of accuracy. These models take into account the context and grammatical features of languages, which allows them to generate translations that are close to natural.
- Simultaneous translation. More advanced models are capable of performing simultaneous translation, which is especially useful at international conferences and negotiations.

Text generation:

- Content creation. Language models can be used to automatically create texts such as articles, news, or even fiction.
- Autocompletion and predictive input. In text editors and instant messengers, language models are used to suggest options for completing phrases or words, which facilitates the writing process and reduces the likelihood of errors.

Automatic text summarization:

- Extractive summarization. This method involves highlighting key sentences from the text and combining them into a summary.
- Abstractive summarization. A more sophisticated method in which models do not simply extract sentences, but also create new phrases that briefly convey the essence of the source text.

Education and training:

- Automatic assignment checking. Language models can be used to automatically check students' written assignments, including checking grammar, style, and content.
- Personalized learning. Language models can analyze each student's progress and offer personalized recommendations and learning materials based on their level of knowledge and preferences.

Virtual assistants:

- Answer generation. Assistants generate text or voice responses to queries, providing users with a convenient and interactive way to interact with the device.
- Command execution. Models allow virtual assistants to execute commands, such as setting an alarm, searching the Internet, or controlling a smart home.

Reference

[1] Dergaa, K. Chamari, P. Zmijewski, and H. B. Saad, "From human writing to artificial intelligence generated text: examining the prospects and potential threats of chatgpt in academic writing," Biology of Sport, vol. 40, no. 2, pp. 615–622, 2023.

[2] Jay Alammar. The illustrated transformer. Visualizing Machine Learning One Concept at a Time, 2018.

[3] Martín Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In ACM CCS, 2016.

[4] Samigova X., Guo T. Chjao Y. Dialogic rhetoric of English and Uzbek. Translation studies: problems, solutions and prospects, 2022. -Pp. 304–307.

[5] Bakirova H.B. Methodology of Lexical Competence Formation of Power-Engineering Students based on CBA and its Experimental Research in Teaching ESP. Pubmedia Jurnal Pendidikan Bahasa Inggris. – Indonesia. 2024. – №2,1.