Universitat de Girona
Escola Politècnica Superior

Treball Final de Màster

Estudi: Màster en Ciència de Dades

Títol: Ajustament d'un model generatiu de llenguatge per a la creació de xatbots personalitzats per administracions públiques

Document: Memòria

Alumne: Martí Mas Fullana

Tutor: Josep Suy Franch
Tutor: Miquel Tarragona Margarit

Departament: Departament d'Informàtica, Matemàtica Aplicada i Estadística
Àrea: Intel·ligència Artificial

Convocatòria (mes/any): Setembre 2024

Universitat de Girona
Escola Politècnica Superior

TREBALL FINAL DE MÀSTER

Ajustament d'un model generatiu de llenguatge per a la creació de xatbots personalitzats per administracions públiques

Autor:
Martí MAS FULLANA

Setembre 2024

Màster en Ciència de Dades

Tutors:
Josep SUY FRANCH
Miquel TARRAGONA MARGARIT

Resum

Agraïments

Per començar vull agrair molt especialment a ...

Índex

Índex de figures

Índex de taules

CAPÍTOL 1

Introducció

1.1 Antecedents

1.1.1 Introduction to Dialogue Systems

Dialogue systems, also known as chatbots, have experienced a significant step-change in the last few years. Initially these systems were based on predefined rules and decision trees [1, 2], limiting their capacity for understanding and answering user queries in a natural and flexible manner. These rudimentary systems, commonly referenced as rule-based chatbots, might have been enough for simple tasks, but could not have managed the full complexity and variability of natural language.

1.1.2 Towards Language Models

As the first machine learning-based language models appeared, such as the Sequence-to-Sequence (Seq2Seq) model [3], and more recently the transformer-based models such as GPT (Generative Pre-trained Transformer) [4, 5], the capacity of chatbots to understand and generate natural language has improved significantly. These models are trained on large datasets of text, learning the complex patterns and structures of language, and are able to generate text that is coherent and contextually relevant.

1.1.3 GPT and its Contribution

The GPT model [5], developed by OpenAI, has been one of the most notable advances in this field. GPT uses the transformer architecture [4], which is a type of neural network that is particularly well-suited for processing sequences of data, such as text. Its capacity for generating coherent and contextually relevant responses has been leveraged in a wide range of applications, from virtual assistance to automated content generation.

1.1.4 Retrieval Augmented Generation (RAG)

One of the most recent advances in the integration of language models has been the use of retrieval augmented generation (RAG) [6]. RAG combines the strengths of information retrieval from databases with the generative capacity of language models. In this context, when a user query is received, the system first retrieves relevant information from a database, and then the language model generates a coherent and precise response based on this information. This approach has been shown to improve the accuracy and relevance of the responses generated by chatbots [6].

1.1.5 Applications and Benefits of RAG-based Chatbots

RAG-based chatbots offer a variety of benefits compared with more traditional systems. They are able to generate responses that are more coherent and contextually relevant. In this way users are both less frustrated and more satisfied. These systems also allow the chatbots to have access to newer, more up to date information than the data the model was originally trained on, as the data provided to the information retrieval component can be updated by simply adding new entries to the database. This makes the chatbot more adaptable and flexible, and allows it to provide more accurate and relevant information to users, reducing the necessity of performing full or partial retraining of the model, which can be prohibitively expensive.

CAPÍTOL 2

Estat de l'art

2.1 Secció

2.1.1 Subsecció

Capítol 3

Preliminars

Capítol 4

Planificació i Metodologia

Capítol 5

Contribució Metodològica

Capítol 6

Resultats

Capítol 7

Conclusions i treball futur

Bibliografia

[1] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, p. 36–45, jan 1966. (Cited on page 1.)

[2] B. Abushawar and E. Atwell, "Alice chatbot: Trials and outputs," *Computación y Sistemas*, vol. 19, 12 2015. (Cited on page 1.)

[3] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," 2014. (Cited on page 1.)

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. (Cited on page 1.)

[5] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018. (Cited on page 1.)

[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2021. (Cited on page 2.)