LangBot-LanguageLearning Chatbot

Madhubala S, Pattabhi RK Rao and Sobha Lalitha Devi
AU-KBC Research Centre,
MIT Campus of Anna University, Chennai, India
sobha@au-kbc.org

Abstract

Chatbots are being widely used in educational domain to revolutionize how students interact and learn along with traditional methods of learning. This paper presents our work onLangBot, a chatbot developed for learning Tamil language. LangBotdevelopedintegratesthe interactive features of chatbots with the study material of the Tamilcourses offered by Tamil Virtual Academy, Government of Tamil Nadu. LangBothelps students in enhancing their learning skills and increases their interest in learning the language.Using semi-automatic methods,we generate question and answers related to all topics in the courses. Wethen develop a generative language model and also Retrieval Augmented Generation (RAG) so that the system can incorporate new syllabus changes. We have performed manual user studies. The results obtained are encouraging. This approach offers learners an interactive tool that aligns with their syllabus. It is observed that this enriches the overall learning experience.

1 Introduction

With the emergence of conversational AI, chatbots or dialogue systems are becoming central tools in many applications such as virtual assistants helping the citizens in getting authentic information from governments, academia, health sector and industry and so on. Since user interests may change frequently over time, the AI agents may continuously see unknown (new) user intents. It is necessary and essential to understand the intent and accurately identify the intent behind the user utterance. This will help in generating the correct response for the user intent. Data created manually by manual annotation cannot catch up with the requirement and algorithms which use annotated data cannot give accurate understanding of the intent by the user.The history of Chatbots was started in 1950 as a result of the Turing Test, and ideas of that test essentially laid the foundation for the revolution of Chatbots. Then, many Chatbots have come to the stage such as Eliza (1966), Parry, Jabberwacky, DrSbaitso (1992), Alice (1995), Smarterchild (2001), IBM's Watson (2006), Siri (2010), Google Now (2012), Alexa (2015), Cortana (2015)etc .

Natural Language Processing has transformed many domains, enabling machines to understand the human and technical language helping to carry out the task more efficiently and correctly. One such domain is the educational domain, where NLP has significantly improved by assisting in automatic content generation, enabling teachers in their up skilling.

In recent years, significant efforts have been directed toward developing Large Language models such as OpenAI, Mistral2B, Med-PaLM, BioBERT, and ClinicalBERetc, These have been designed and trained on vast corpora of texts.

The rest of the paper is organized as follows. In Section 2, we review related works.In Section 3, we give our approach along with corpus description. Section 4 describes the results followed by conclusion.

2 Related Work

The journey toward developing chatbots to aid in learning has not gathered much attention both in the academics and industry oriented research. A growing body of research has explored the integration of general-purpose models with domain-specific knowledge through techniques like Retrieval-Augmented Generation (RAG). This approach involves augmenting the model's generative capabilities with relevant external knowledge retrieved from large datasets, which has been demonstrated to significantly boost the

accuracy of model predictions in a particular domain. For instance, the study presented in(Xiong et al., 2024) exemplifies how incorporating retrieval mechanisms into generative models can lead to substantial improvements in the accuracy of biomedical information processing.

There has been many studies to understand user intent from various domains, ranging from search engine questions (Hu et al., 2009) to medical queries (Zhang et al., 2016). The approaches used for intent classification include Deep learning models such as convolutional neural networks (CNN) (Xu and Sarikaya, 2013) and attention-based recurrent neural networks (RNN) (Ravuri and Stolcke, 2015; Liu and Lane, 2016).

Advancements have been made in creating toolkits and frameworks designed to facilitate the development of RAG models tailored to specific domains. The work by (Li et al., 2024) introduces a comprehensive toolkit aimed at streamlining the creation of RAG models to build reliable models for various medical tasks. Additionally, innovative applications of RAG have extended beyond traditional healthcare, as seen in (Vakayil et al., 2024), which explores the use of RAGmodels to provide psychological support to survivors of trauma. In this work RAG methodis used to develop a generative, interactive chat system.

3 Our Approach

Here are the steps that were followed in the development of the chatbot

a) Data preparation
b) Model Training
c) Retrieval Augmented Generation (RAG) and Response Generation

3.1 Data Preparation

The data/text from the text books comprising of four levels of courses viz., Certificate, Diploma, Higher Diploma and Degree level are taken in this work. This is a huge data. Table 1 shows the statistics of the data.

The text obtained from the text books is pre-processed by performing tokenization. The tokenized data is converted into vector representations such that it can be fed into the neural network. Word embedding's method of Word2Vec is used to represent as dense vectors.

Table 1: Data Statistics

Course Name	Books	Chapters	Sentences	Words
Certificate	1	18	9345	143445
Diploma	4	45	44767	564897
Higher Diploma	4	42	47879	623456
Degree	9	172	102567	1846206

3.2 Model Training

Once the data is prepared we develop custom language model. The model training is done using TensorFlow'sKeras API. We use transformers based language model architecture. The hyper parameters are defined as follows: embedding dimensions are 512, and we take a maximum sequence length of 100. The number of training epochs in our study was not predetermined, but dynamically determined based on the loss function's behavior. Training continued as long as the loss showed a consistent decreasing trend and was terminated when a significant spike in loss was observed. This approach optimized model performance while mitigating over fitting risks.

3.3 Retrieval Augmented Generation (RAG) and Response Generation

After completing model training, setting up the retriever component is crucial. Since this system is developed for the Tamil text books and it was necessary that the system also considers the new syllabus which is continuously evolving. The new syllabus does not have elaborate text books as it is still evolving. So for this purpose we use Retrieval Augmented Generation (RAG). For developing the RAG we take the new syllabus topics statements. The topic statements are matched to the old text books contents using similarity techniques. The chunks of text similar to the new syllabus are retrieved from the text.

Response Generation: The original query from the user is given to the language model and output is obtained. Based on this output context is set to retrieve the relevant chunks from the RAG and the final output response is generated.

4 Results and Evaluation

This chatbot has been tested using human evaluators in a practical environment. At first, people have been selected with different levels of their educational qualification levels and they were divided into four groups of 10 each such as Certificate level, Diploma level, Higher Diploma level and Graduate level respectively, As the next step, the system was freely distributed among them and asked them to chat with the system. After that they were asked to score for based on five parameters:

a) Clarity of User Interface
b) Ease of Use
c) Validity of responses
d) Usefulness
e) Efficiency and Reliability

Table 2 shows the scores given by the human evaluators.

As can be observed from the score table the system has overall satisfactory level of 85%.Majority of the participants have expressed satisfaction with the system. The system needs to improve in the area of generation of valid responses which means the answers need to be more specific and to the point.

Table 2: Human Evaluation Scores – Each participant scored the system on each parameter

Parameter	Very Good	Good	Moderate	Poor	Very Poor
The clarity of the User Interface	25	15	0	0	0
Ease of Use	20	17	3	0	0
Validity of responses	14	20	5	1	0
Usefulness	20	19	1	0	0
Efficiency &Reliability	25	10	5	0	0

Conclusion

In this paper we have described our work on the development of Tamil language learning chatbot. Here we base our work on RAG which helps developing language model specific to the Tamil syllabus text books. The system was tested with human evaluation and it is observed that it is having overall 85% satisfactory level.

References

Anjishnu Kumar, Pavankumar Reddy Muddireddy, Markus Dreyer, and Bj¨ornHoffmeister.2017. Zero shot learning across heterogeneous overlapping domains. In Annual Conference of the International Speech Communication Association (INTER-SPEECH), pages 2914–2918.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. In Annual Conference of the International Speech Communication Association (INTER SPEECH), pages 685–689.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S. Yu. 2018. Zero-shot user intent detection via capsule neural networks. In Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3090–3099

Emmanuel Ferreira, BassamJabaian, and FabriceLef`evre. 2015a. Onlineadaptative zero-shot learning spoken language understanding using wordembedding. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5321–5325.

Emmanuel Ferreira, BassamJabaian, and FabriceLef`evre. 2015b. Zero-shot semantic parser for spoken language understanding. In Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 1403–1407.

Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. 2011. Transforming auto-encoders. In ICANN, pages 44–51.

Jian Hu, Gang Wang, Frederick H. Lochovsky, JianTao Sun, and Zheng Chen. 2009. Understanding user's query intent with wikipedia. In International Conference on World Wide Web (WWW), pages 471–480.

Jinseok Nam, EneldoLozaMenc'ia, and Johannes F¨urnkranz. 2016. All-in text: Learning document, label, and word representations jointly. In AAAI Conference on Artificial Intelligence (AAAI), pages 1948–1954.

KiyoakiShirai and Tomotaka Fukuoka. 2018. JAIST Annotated Corpus of Free Conversation.

In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

PrateekVeerannaSappadla, Jinseok Nam, EneldoLozaMenc'ia, and Johannes F¨urnkranz. 2016. Using semantic similarity for multi-label zero-shot classification of text documents. In European Symposium on Artificial Neural Networks (ESANN).

SumanRavuri and Andreas Stolcke. 2015. Recurrent neural network and lstm models for lexical utterance classification. In Annual Conference of the International Speech Communication Association (INTER-SPEECH), pages 135–139.

Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic routing between capsules. In Advances in Neural Information Processing Systems (NIPS), pages 3859–3869.

SeppHochreiter and J¨urgenSchmidhuber. 1997. Long short-term memory. Neural computation, 9(8):1735–1780.

PuyangXu and RuhiSarikaya. 2013. Convolutional neural network based triangular CRF for joint intent detection and slot filling. In IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU Workshop), pages 78–83.

Yun-Nung Chen, Dilek Z. Hakkani-T¨ur, and Xiaodong He. 2016. Zero-shot learning of intent embeddings for expansion by convolutional deep structured semantic models. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6045–6049.