IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol-6, Special Issue-6, July 2018 E-ISSN: 2347-2693

A Proposal of Chatbot for Malayalam

S. Sandhini[1*], R. Binu[2], R.R Rajeev[3,] M.M Reshma[4]

[1,2]Dept. Of Computer Science, Government Engineering College, Palakkad, India
[3,4]Computational Linguistics. International Centre for Open Source and Software, Trivandrum, India

[*]*Corresponding Author: sandinisukumar@gmail.com*

Available online at: www.ijcseonline.org

***Abstract*—** A chatbot is a conversational agent which interacts with humans via natural languages. Text as well as speech, is used as the input to these systems. We propose a Malayalam chatbot based on a natural language processing with machine learning techniques library on a language-independent platform. The chatbot is a retrieval based model which can converses in Malayalam. Malayalam is a Dravidian language talked over the Indian state of Kerala. Machine learning as well as NLP (Natural Language Processing) approaches are used to analyze user queries and generate responses. We experiment an AIML (Artificial Intelligence Markup Language) based chatbot and a machine learning based chatbot. Among the two bots, the machine learning chatbot performs better. We developed a domain-specific chatbot. It is a commercial product, so we can apply this to any domain.

***Keywords*—** Chatbot, Natural Language Processing, Machine Learning, Artificial Intelligence.

I. INTRODUCTION

Chatbot is a computer program which communicates with human users through natural languages [1]. It is likewise known as a chatterbot, Bot, IM bot, talkbot, Interactive agent, or Artificial Conversational Entity. The user can interact with the bot in written, oral, or mixed format of conversation. The chatbots can be used in various fields such as Business, Health, Tourism, Customer Support and so on. The architecture of a Chabot integrates language models with computational algorithms.

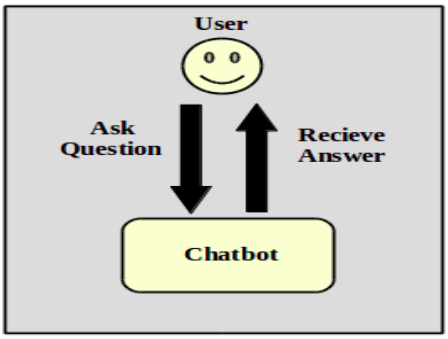

Figure 1. Chatbot Components

Many chatbots are available in English, because it is easy to build chatbots in that language. A lot of platforms are available for building chatbots in English. A lot of works are going on to develop chatbots in other languages these days. Figure 1.1 shows the simple design of the chatbot. The chatbot conversation framework is of two types: retrieval based and generative based

1. Retrieval model: The model uses a knowledge base of predefined responses and it works based on pattern matching algorithm. According to the matching the model selects an appropriate answer. The system does not generate any new responses as it outputs the known set of responses as answer to the user query.

2. Generative model: This model is a more intelligent one. It generates answer to user query according to its knowledge. The model learns from a set of known interactions and give answer by understanding the user query. Here, NLG (Natural Language Generation) process is involved.

This paper proposes a chatbot that returns a relevant answer to questions asked by user. The paper is composed as follows: Section I describes the introduction of the chatbot technology. Section II gives the details of the related works on the chatbot technology. In this, the previous works related to chatbot is described. Section III contains the design and methods used to build Malayalam chatbot. In section IV the two methods proposed are compared. Section V concludes research work with future directions.

II. RELATED WORK

The concept of chatbot was by Weizenbaum [3] who implemented the first chatbot ELIZA to mimic a psychotherapist. The ELIZA chatbot is based on the keyword matching mechanism. ALICE is another chatbot developed by Abu Shawar et.al [1] that implements various dialogues using AIML (Artificial Intelligent Markup Language). This is an extension of XML to represent patterns and templates

that are the fundamental dialogues. The categories are the basic element of AIML. The most commonly used mechanism to develop chatbot is an AIML based method. The pattern matching concept is used in this method. AIML contains a set of predefined rules.

The AIML based chatbots are the most popular because they are very easy to implement. The ALICE chatbot won Leobner Price three times. Here, each category consists of a pattern which is the input given by the user and the template which includes the output by the chatbot to the user. Here, there are three categories such as atomic, default, and recursive.

Ranoliya, et.al [2] implements an FAQ chatbot based on AIML and Latent Semantic Analysis (LSA) methods. They developed an interactive University related Frequently Asked Questions (FAQs) chatbot. Initially, user posts the query on the chatbot. Then, process the query that entered by the user to match the predefined format. The pattern matching is done in the query and the pattern in the knowledge base. Finally, the patterns matched answer is presented to the user. The chatbot interacts queries related to college information, admission etc. The LSA method is used to discover the similarities between words as vector representation. So that the unanswered questions by AIML will be seen as an answer by LSA.

Kalaiyarasi, T. et.al [7] proposed a Tamil chatbot POONGKUZHALI. In poongkuzhali the user can pick any current topic for discussion and ask in Tamil. Based on the context of the query, the system generates an appropriate answer to the user. The system identifies the minimal context of the input and this can be done by using a set of decomposition rules. The response is generated by utilizing a set of reassembly rules that resided in the knowledge base.

S. Chaitrali et.al [5] proposed a Bank chatbot to handle queries related to the Bank. The user interacts with the system through a web application. In this model, the user query is handled by the boot controller logic. Using NLTK libraries the query is preprocessed. Then the preprocessed query is vectorized and classification algorithms are applied to find the class it belongs to. Based on cosine similarity, the most similar answer is returned to the user as a response.

Shah et.al [6] proposed an intelligent chatbot based on Natural Language Processing (NLP) in educational systems. The database contains the topics related to the educational system. The user submits the query. The NLP processes such as Tokenization, Lemmatization, POS tagging, Dependency Parsing and Role Labeling are applied to the query. The NLP and Machine Learning (ML) are applied at their respective levels. The LSTM method is used to develop the answer from the database. It allows the model to learn the encoded part and how to create the output relevant to the question asked by the user.

III. MALAYALAM CHATBOT

Chatbots are computer program that simulates the human conversation. Here we are developing a chatbot that converses in Malayalam. The input is a Malayalam text and the output is the response to the user input which is in Malayalam. Introducing chatbot in a free order language like Malayalam is a challenging task. The attempt will be a defining moment in the Malayalam language. Regardless of whether the work begins as a domain specific, it can be brought into different domains and areas.

A. AIML based Malayalam Chatbot

AIML is the common technique used in the design of the chatbot. AIML is the extension of the XML. The AIML data objects consist of two units: Topics and Categories. The purpose of the AIML chatbot is to facilitate a better conversation modeling. The data object in the AIML language has the responsibility of modeling the conversational patterns.

The important objects in the chatbots are categories, pattern, and template. The category tag is used to specify the knowledge in the conversation. The pattern contains the input from the user to the system and the template includes the output or response to the user. The structure of the AIML categories, patterns, template are given below,

```
<category>
        <pattern> Input from the User</pattern>
                <template>
                        Response to the User
                </template>
</category>
```

The AIML chatbot is based on the pattern matching concept, the response generated is based on the mapping of keywords in each request and their patterns. With the help of the AIML Interpreter, the pattern matching between query and response is done.

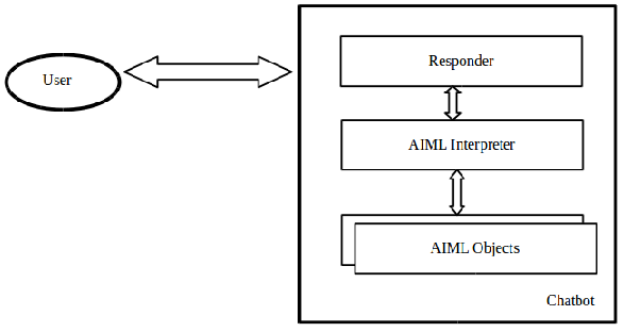

Figure 2. AIML Chatbot Architecture

The natural language input is initially preprocessed. The keywords are extracted by removing stopwords, stemming from the words. The pattern matching is achieved with the help of the tags such as <that>, <srai>, <topic> for

remembering the previous conversation. The user inputs the question to the system and the AIML Interpreter matches the keywords and generate the response then, manage the conversation. The <random> tag is used to generate random response.

The AIML is language independent. So Malayalam chatbot can be developed based on the AIML tags.

```
<?xml version = "1.0" encoding = "UTF-8"?>
<aiml version="1.0.1" encoding="UTF-8">
<!-- basic_chat.aiml -->
<category>
    <pattern>ഞാൻ *</pattern>
    <template>
      നമസ്കാരം <set name = "username"> <star/>!
</set>
    </template>
  </category>
</aiml>
```

This simple AIML file is shown above. The category part includes pattern and template tags which contains the input as well as the output. The AIML chatbot can store the name of the user using the tag <set name>. They can remember the name of the user till the conversation ends.

B. Malayalam Chatbot based on Chatterbot Python Library

Propose a retrieval based domain specific chatbot that converses in Malayalam language. The chatbot is developed based on natural language processing library with machine learning mechanism. Initially, the chatbot matches the input with the queries in the database and then calculates a confidence value for sentences that matched. Finally, selects the appropriate response based on the highest confidence value among them. Here, the response generation is retrieval based. The retrieval based method retrieve responses from the knowledge base based on the context. The input is given to a web interface. Then process the text that given by the user. Selects the sentences which match the input text.

Returns known response to the selected matches. Then, calculates the confidence value of each response. The system will return an appropriate response as the highest confidence value. Finally, the response is displayed to the user.

The chatterbot is a Python package based on Machine learning concepts. It is a language independent platform. The important module of the Chatterbot library is Chatterbot's adapters. The input is returned from the input adapter, the input is processed and stored by the logic and the storage adapters. Finally, it passed to output adapter to returns the responses to the user. Chatterbot includes training tools to make simple training process. The Chatterbot training includes loading of the dialogues into the chatterbot database. Several training classes are inbuilt in chatterbot package. Here, we train the data from the dialogue corpus using Corpus Trainer Class of Chatterbot. Initially, developed a Malayalam Tourism corpus in the predefined format in YAML file. Then, set a trainer for the Malayalam dialogue corpus. The chatterbot has built-in storage adapter that connects different databases. The Best Match logic adapters are used here to find the best response to the closest match. Here, a low confidence adapter is used and a threshold value 0.5 is set. The Best Match adapter calculates similarity between the input text to known responses. We experimented the comparison using 2 methods: Jaccard Index and Levenshtein distance. Jaccard index is the ratio of no. of items in common to total no. of items. Here are two sample sentences are:

തിരുവനന്തപുരത്തെ പ്രധാന വിനോദസഞ്ചാരകേന്ദ്രങ്ങൾ ഏതൊക്കെയാണ് , തിരുവനന്തപുരം ജില്ലയിലെ വിനോദസഞ്ചാരകേന്ദ്രങ്ങൾ .എന്തെല്ലാമാണ്

When we parse these sentences and remove the stop words, we get;

{തിരുവനന്തപുരം, പ്രധാന, വിനോദസഞ്ചാരകേന്ദ്രങ്ങൾ}{തിരുവനന്തപുരം , ജില്ല, വിനോദസഞ്ചാരകേന്ദ്രങ്ങൾ}

From the example, the intersection set is {തിരുവനന്തപുരം, വിനോദസഞ്ചാരകേന്ദ്രങ്ങൾ} which has two elements, the union set contains {തിരുവനന്തപുരം, പ്രധാന, ജില്ല, വിനോദസഞ്ചാരകേന്ദ്രങ്ങൾ}, which has count of four. The Jaccard index is two divided by four, that is 50%. So the Jaccard index is greater than the threshold value, we consider this as a match.

Levenshtein distance, a similarity measure between two sentences. The measure is calculated based on the minimum no. of single character edits that were, deletions, insertions or substitutions. From the above example the Levenshtein distance between the two sentences is 24. Accurate answers are generated when comparison is done using Jaccard index. Because it will show the response only when the correct word match is found. The bot interacts with user in a better way when Levenshtein for comparison. It finds a response even for unknown instances as it does a character level comparison. Even though the bot gives faulty responses, the conversation keeps going on.

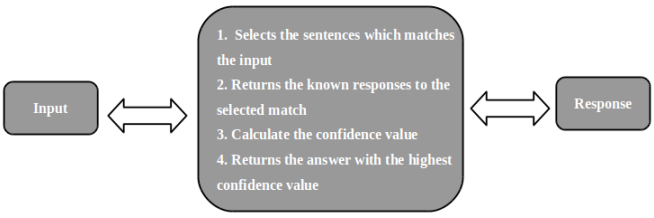

Figure 3. Flow chart of the proposed system

categories:

- കേരള ടൂറിസം
- കേരളത്തിലെ വിനോദസഞ്ചാരം

conversations:

- - കടൽത്തീരങ്ങൾ

- കോവളം, ബേക്കൽ, മുഴപ്പലിങ്ങാട്, ആലപ്പുഴ, വർക്കല, ശംഖുമുഖം, ചെറായി

We manually created Malayalam tourism corpus in a specific format shown in the figure. The data are stored in YML file. YML(*YAML Ain't Markup Language*) is a data serialization language in human-readable format. The chatbot interface is created by using Flask Library. The YML file contains a categories and conversations part, the category part describes the name or category of the data. The conversations part includes the different conversations made on the topic for training the chatbot.

The figure 4 shows the implementation of Malayalam chatbot. The principle focus of this chatbot is to produce sentences which are steady, free from linguistic errors and spelling errors. It accomplishes the objective of delivering linguistically correct Malayalam responses. Since the responses are on a par with its knowledge base so a great deal of work needs to be done to upgrade the knowledge base. However, amid building the knowledge base, must give a knowledge base free from mistakes.

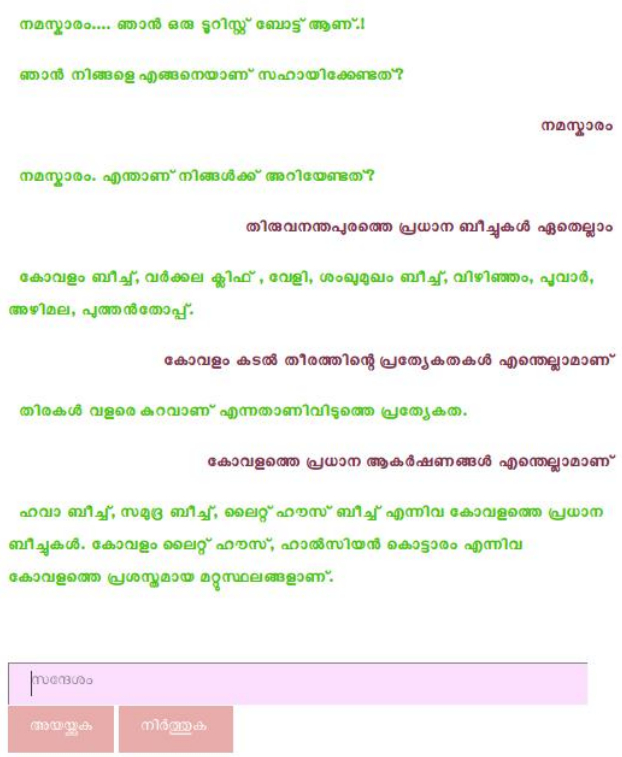

Figure 4. Implementation of Malayalam Chatbot

IV. COMPARISON

Since this is the first chatbot in Malayalam, there is no another Malayalam chatbot available for comparison. So comparison is done between the two chatbots we built, AIML based and Chatterbot based Malayalam chatbots. The AIML based chatbot is a rule-based one. The Chatterbot is integrated the Machine learning and Natural language processing concepts.

The AIML chatbot use a simple pattern-template pair to represent the user input and output. It uses simple pattern matching algorithm. The chatbot based on chatterbot starts off with no knowledge. Since there is no corpus available in Malayalam to train chatterbot, we prepared a Malayalam conversation corpus. We take care of the issue of the absence of required tools by choosing a language independent platform and picking a retrieval based model to fill the need. We examine same questions to both chatbots. The machine learning based chatbot can answer progressively like others. Since it can take include Malayalam and can give a response in Malayalam so we can state that the pattern matching algorithm is working great. The chatbot responses are free from spelling errors and any sort of semantic errors. It makes some accentuation issue which can be enhanced in future.

V. CONCLUSION AND FUTURE SCOPE

Conversational AI is a big part of the future. The users can easily enter their question in natural language and retrieve data. In this paper, propose chatbot for Malayalam language. A chatbot is an important tool for communicating with the user. General purpose chatbot must be straight forward, easy to use, must be effectively comprehended and the database must be smaller. Although some of the commercial items have developed, enhancements must be made to find a typical approach for designing a Chatbot. The Malayalam language has its own characteristic features, that make it different from other languages. This experiment is a spearheading work in the field of conversation framework in Malayalam. The principle test of this work is to make a chatbot in light of the precise learning base. Because of a need of a substantial dataset, we actualized a retrieval based closed domain chatbot which will speak with the user based on the pattern matching algorithm and will enhance its performance measure by gaining from the cooperation. Our work will give a Malayalam conversation corpus which will help in the improvement of tools for Malayalam Language Processing research.

For future work, we can implement a voice-enabled chatbot system and add pictorial representation for better understanding for people.

REFERENCES

[1] Shawar, Bayan Abu, and Eric Atwell.”Chatbots: are they really useful?." *Ldv forum*. Vol. 22. No. 1. 2007.

[2] Ranoliya, Bhavika R., Nidhi Raghuwanshi, and Sanjay Singh. "Chatbot for University Related FAQs.", 2017.

[3] Weizenbaum, Joseph. "ELIZA—a computer program for the study of natural language communication between man and machine." *Communications of the ACM* 9.1, 36-45, 1966, .

[4] Marietto, Maria das Gracas Bruno, et al.”Artificial intelligence markup language: A brief tutorial.”*arXiv preprint arXiv:1307.3091*, 2013.

[5] S. Chaitrali, Kulkarni, U. Amruta, Bhavsar, Savita Chaitrali S Pingale. "BANK CHATBOT – An Intelligent Assistant System Using NLP and Machine Learning", International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 05, 2017.
[6] Shah, Rishabh, Siddhant Lahoti, and K. Lavanya. "An intelligent chat-bot using natural language processing." *International Journal of Engineering Research* 6.5 : 281-286. 2017.
[7] Kalaiyarasi, T., Ranjani Parthasarathi, and T. V. Geetha. "Poongkuzhali-an intelligent tamil chatterbot." *SIXTH TAMIL INTERNET 2003 CONFERENCE*. Vol. 1. sn, 2003.
[8] Abdul-Kader, Sameera A., and John Woods. "Survey on chatbot design techniques in speech conversation systems." *International Journal of Advanced Computer Science and Applications* 6.7 : 72-80.2015.
[9] E. Loper, and S. Bird, "NLTK: The natural language toolkit." pp. 63-70, 2002.
[10] S. Bird, "NLTK: the natural language toolkit." pp. 69-72, 2006.
[11] A. S. Lokman, and J. M. Zain, "An architectural design of Virtual Dietitian (ViDi) for diabetic patients." pp. 408-411, 2009.
[12] A. M. Galvao, F. A. Barros, A. M. Neves, and G. L. Ramalho, "Persona aiml: An architecture developing chatterbots with personality." pp. 266-1267, 2004.
[13] Mujeeb, Sana, Muhammad Hafeez Javed, and Tayyaba Arshad. "Aquabot: A Diagnostic Chatbot for Achluophobia and Autism." *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS* 8.9 : 209-216, 2017.
[14] Bani, Balbir Singh, and Ajay Pratap Singh. "College Enquiry Chatbot Using ALICE."
[15] S. J. du Preez, M. Lall and S. Sinha, "An intelligent web-based voice chat bot," EUROCON 2009, EUROCON '09. IEEE, St. Petersburg, 2009.
[16] Wailthare, Sumit, et al. "Artificial Intelligence Based Chat-Bot." *Artificial Intelligence* 5.03 (2018).
[17] TIWARI, AMEY, RAHUL TALEKAR, and SM PATIL. "College Information Chat Bot System."
[18] Hatwar, Nikita, Ashwini Patil, and Diksha Gondane. "Ai based chatbot." *International Journal of Emerging Trends in Engineering and Basic Sciences* 3.2 85-87. 2016.
[19] Sarthak V. Doshi "Artificial Intelligence Chabot in Android System using Open Source" Program International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 6, Issue 4, April 2017.
[20] Bayu Setiaji " Chatbot Using A Knowledge in Database" International Conference on Intelligent System, Modling and Simulation 2016.

Authors Profile

Ms. Sandhini S pursued Master of Technlogogy in the field of Computer Science and engineering with specialization in Computational Linguistics from GEC, Palakkad, A P J Abdul Kalam Technological University, Kerala, India in year 2018. Her area of interests are Natural Language Processing, Machine Learning and Computational linguistics.

