International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.03, April 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

INTELLIGENT CHAT BOT

A. Mohamed Rasvi, V.V. Sabareesh, V. Suthajebakumari

Computer Science and Engineering, Kamaraj College of Engineering and Technology, India

ABSTRACT

This paper discusses the workflow of intelligent chat bot powered by various artificial intelligence algorithms. The replies for messages in chats are trained against set of predefined questions and chat messages. These trained data sets are stored in database. Relying on one machine-learning algorithm showed inaccurate performance, so this bot is powered by four different machine-learning algorithms to make a decision. The inference engine pre-processes the received message then matches it against the trained datasets based on the AI algorithms. The AIML provides similar way of replying to a message in online chat bots using simple XML based mechanism but the method of employing AI provides accurate replies than the widely used AIML in the Internet. This Intelligent chat bot can be used to provide assistance for individual like answering simple queries to booking a ticket for a trip and also when trained properly this can be used as a replacement for a teacher which teaches a subject or even to teach programming.

Keywords : *AIML, Artificial Intelligence, Chat bot, Machine-learning, String Matching.*

I. INTRODUCTION

Social networks are attracting masses and gaining huge momentum. They allow instant messaging and sharing features. Guides and technical help desks provide on demand tech support through chat services or voice call. Queries are taken to technical support team from customer to clear their doubts. But this process needs a dedicated support team to answer user's query which is a lot of man power. Help desks are grown and almost all companies has their own support team.

Computer assistants like Siri, Cortona and Bixby are mature enough to answer the common queries raised by the users. But they are predefined to answer set of queries and not adaptively learning to answer any new queries. They can only provide support using the predefined data which are not accessible to public.

Machine learning is a computer term that gives ability to a computer to learn. The learning process is done by training the computer with set of data which are fed in to the computer as some patterns. In turn the computer utilises the data to identify certain patterns. The pattern is compared with the test data to make a decision. The comparison will result in its best match and the accuracy is given by the difference in the trained data and test data.

The proposed work is to train the data to give high accuracy for building a chat bot. The text messages are to be trained to provide relevant reply.

The rest of the research paper will unfold as follows: Section two will present related work and section three focuses on the methodologies. The section four contains system design and section five contains our future work. Section six concludes the research.

II. RELATED WORK

There are several chat bots that conducts conversations using the following algorithms . [1] Shengnan Zhang, Yan Hu and Guangrong Bian used Levenshtein distance algorithm to compare the strings. They suggested an efficient method making use of the longest common substring and the longest common subsequence. The results were universally acceptable by increasing the accuracy while comparing different strings. [2] Jun Choi Lee and Yu-N Cheah suggested sentence to sentence paraphrase detection. They proposed the method adopting the conceptual semantic relatedness in paraphrasing. This algorithm's performance is considered reasonable for semantic relatedness paraphrasing detection. [3] Salto Martinez Rodrigo and Jacques Garcia Fausto Abraham proposed the various stages involved in development of the chat bots, by comparing various algorithms for string matching . They primarily focused on reducing the time taken for comparison of string parameters stored in the knowledge base. [4] Luka Bradesko, Janez Starc, Dunja Mladenic, Marko Grobelnik and Michael Witbrock proposed the algorithm focusing on the crowd sourced knowledge acquisition approaches.

All the above proposed ways have some kind of drawbacks like either they are not accurate at all times or they do only a single part in entire semantic content extraction process. Most of the above proposed works only on particular type of inputs.

Thus to improve the efficiency of the chat bot, our propose model runs a comparison of possible results when using different algorithms and proceeds with the efficient one.

III. METHODOLOGY

We will use four different algorithms to compare the messages. Since four algorithms are better than one good algorithm to produce more relevant reply.

A. Pre-process

The first step in training any data set is to pre-process the data by removing unwanted data in it. Using cleaned data improves the accuracy. To clean the data we use Natural Language Toolkit. Which removes all the stopping words like 'a' and 'the' which makes little to no sense. In this process all the punctuations are also removed. So only the words which make sense are obtained in this step.

"How are you?" = {"How", "you"}

For every sentence there should be a reply. The question is pre-processed and stored in database. The reply message is stored as it is with the question in same database. So when a question is matched higher than all other questions with the user message, the reply is sent from the reply message which is stored along with the question.

The Natural Language Toolkit has text processing libraries for classification, tokenization, parsing and semantic reasoning, wrappers for industrial strength libraries. Apart from this it has Wordnet which groups set of words with similar meaning. This is used to remove stopping words.

Now the inference engine can match the user query with the database and send reply by comparing each of the question with the user query.

B. **String comparisons**

Comparison is the important process as it determines the accuracy of the reply message. There are four comparison algorithms used.

1. Levenshtein distance
2. Synset distance
3. Sentiment comparison
4. Jaccard similarity

Levenshtein distance is the number of character change needed to convert one string to another string. It is the count of character difference. This should be low for a good match.

Synset distance matches against the synonym of each word to get matching similar meaning word. This comparison is taken again word by word instead of character by character.

Sentiment comparison calculate the similarity of two statements based on the closeness of the sentiment value calculated for each statement.

Jaccard similarity is a crude word by word match that shows the percentage of words matched against the user query.

C. **Reply Prediction**

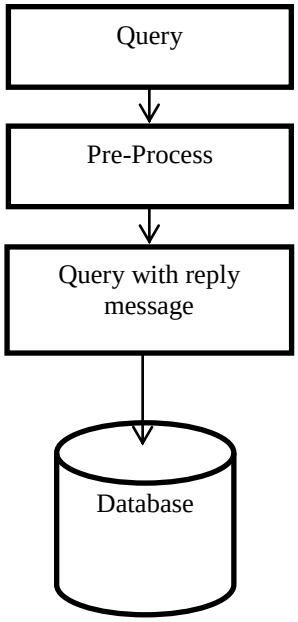
All the data are pre-processed and stored in database. When a user sends a query, the query is pre-processed in the same way all the data are done. The pre-processed data is then compared with the data one by one using the four algorithms. A confidence level is maintained for each comparison to show how much the query is relevant for that particular data in database. The confidence level ranges from 0.00 to 1.00. 1.00 declares 100% match and 0.00 declared 0% match for the user query.

The comparisons algorithms are used separately to get confidence level for same data. The average of the confidence level is used for that data. The data with most confidence level is chosen as the best match and its corresponding reply message is sent to the user.

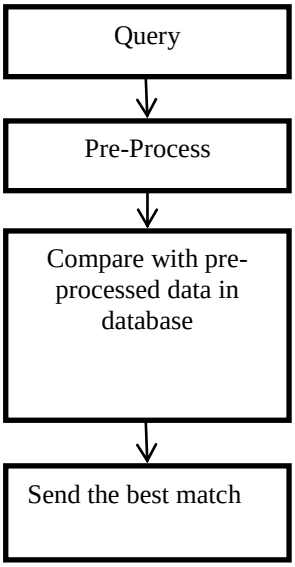
For example, for the query "How are you?" it could match with two different questions in database as "who are you?", "How are you?" with 50% {"who", "you"} match and 100% {"How", "you"} match so the 100% match is chosen as best.

IV. SYSTEM DESIGN

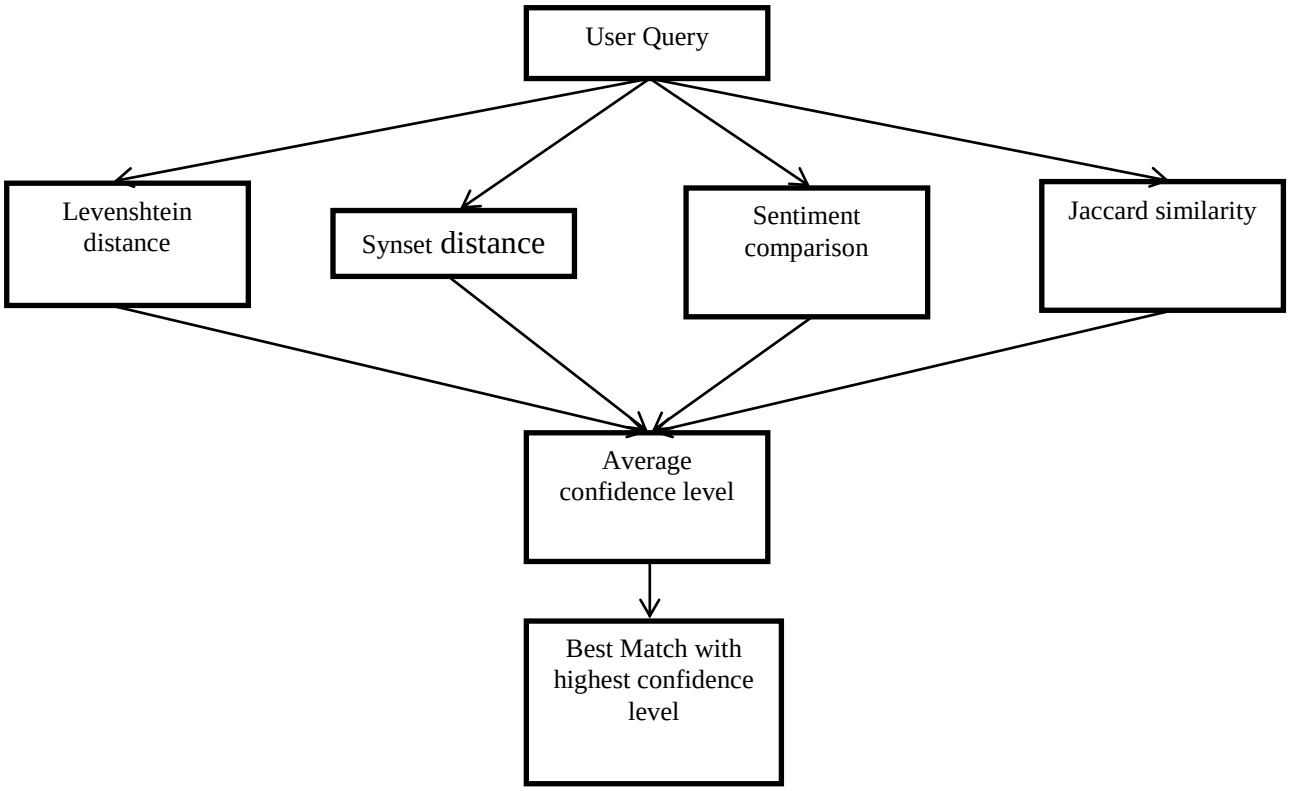
Training Data Set



Request – Response Flow



String Comparison



V. FUTURE WORKS

As of now, the chat bot performs regular chit-chat conversations in addition to answering question in C. We aim to teach more programming languages using our chat bot.

The next focus is to add more number of regional languages.

The chat bot will be enhanced to act as the personal assistant, integrating with the multiple devices.

We will be upgrading to more efficient algorithm to increase the accuracy of results.

VI. CONCLUSION

Handling growing set of objects and achieving output in lesser time is the key to semantic content extraction. Through our approach, we proved that there is no need to train the system with all objects at once and retrain again when new objects are added. Experimental results confirm the approach by saving the computation time upto 30% in overall process.

REFERENCES

[1] Shengnan Zhang, Yan Hu, Guangrong Bian (2017), "Research on string similarity algorithm based on Levenshtein Distance", in Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2017.

[2] Jun Choi Lee, Yu-N Cheah (2016), "Paraphrase detection using semantic relatedness based on Synset Shortest Path in WordNet", in International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), 2016

[3] Salto Martinez Rodrigo, Jacques Garcia Fausto Abraham(2015), "Development and implementation of a Chat Bot in a Social Network" , 9th International Science and information conference 2017

[4] Luka Bradesko, Janez Starc, Dunja Mladenic, Marko Grobelnik , Michael Witbrock (2016) "Curious cat conversation crowd based and context aware knowledge acquisition chat bot". 2016 IEEE 8th International Conference on Intelligent Systems (IS).