AI as Friend

Awais Ahmad Khan
University of Missouri Kansas City
CS5560 Knowledge Discovery and Management
Kansas City, Missorui, USA
aakpmd@mail.umkc.edu

Shah Rukh Khan
University of Missouri Kansas City
CS5560 Knowledge Discovery and Management
Kansas City, Missorui, USA
Srkhvx@mail.umkc.edu

Abstract— We are proposing a Chat bot that will act as a friend. A bot powered by NLP and a vast treasure of information stored as knowledge graphs that can help counsel user when he is angry or sad or can rejoice with user when he feels like it. Our chat bot can be named AI as a friend. Thus, our Chat bot will be a virtual friend truly inspired by the movie "HER" and "JEXI" who will have no human restraints.

I. INTRODUCTION

There are a lot of chat bots out there and almost all of them are built for a certain purpose whereas none of them Is a general-purpose chat bot. Our aim is to create something that can be a companion, that is able to connect dots like the user does. We are working on the data part, but it will most likely be an opensource project where we collect data from a crawler that goes online gets it for the project. The reason for choosing this project is to make ourselves equipped with the necessary tools used in the NLP. As this project will be focusing and using all the necessary and important concepts of NLP that is why we chose this project. Our motivation is to learn as much as we can and implement a practical implementation in the form of a project.

II. RELATED WORK

A. Converstational Modeling for chatbots
http://www.essv.de/paper.php?id=405

B. Empowering chatbots with Business Intelligence by Big Data Integration
https://www.researchgate.net/profile/Reshmi_Sankar/publication/323451431_EMPOWERING_CHATBOTS_WITH_BUSINESS_INTELLIGENCE_BY_BIG_DATA_INTEGRATION/links/5b9351b4299bf14739257a86/EMPOWERING-CHATBOTS-WITH-BUSINESS-INTELLIGENCE-BY-BIG-DATA-INTEGRATION.pdf

C. Her and Jext
Our motivation comes from two movies by the name of HER and JEXT. In this they have used a very complex and interactive AI systems.

III. PROPOSED WORK

In this project we will be focusing on two main subjects. First will be the data science part which will include all the necessary data collection and pre processing of data sets e.g. data cleaning, tokenization, extracting the necessary info. Second one will be purely AI and Machine learning part in which we will be making a neural network. The datasets which we have processed in the first phase will be used to train this model.

IV. IMPLEMENTATION AND EVALUATION

The main module of this project will be a neural network that will be trained on the pre processed data sets containing the necessary material for producing the correct results. There are two ways in which we can train the model. One will be to train different models on different datasets e.g. one model for the questions and answers and the other for suggesting somethings. The model will be able to detect the main gist of the text and will give the response accordingly.

A. Evaluation Plan

We have not narrowed down on a specific dataset yet. Since our inspiration comes from JEXI which is a very open-ended tool. It is not restricted one form of communication and most of the datasets online are very specific to a certain nature of textual data. Thus, we have a couple of datasets and we are trying to figure out how to blend them together so that our model is able to adapt to each specific style according to a consumer need. The datasets are as follows, they are open source and their references are mentioned along with them.

B. Datasets

- **Semantic Web Interest Group IRC Chat Logs**: This automatically generated IRC chat log is available in RDF, back to 2004, on a daily basis, including time stamps and nicknames.

- **Cornell Movie-Dialogs Corpus:** This corpus contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts: 220,579 conversational exchanges between 10,292 pairs of movie characters involving 9,035 characters from 617 movies.

- **The WikiQA Corpus:** A publicly available set of question and sentence pairs, collected and annotated for research on open-domain question answering. In order to reflect the true information, need of general users, they used Bing query logs as the question source. Each question is linked to a Wikipedia page that potentially has the answer.

- **Reddit:** 3.7 billion comments structured in threaded conversations.

- **OpenSubtitles:** over 400 million lines from movie and television subtitles (available in English and other languages).

- **Amazon QA:** over 3.6 million question-response pairs in the context of Amazon products.

V. Conclusion

In conclusion we can say that we have narrowed our project down into two parts. First part will be focusing on the big data part where we will be collecting the necessary data and pre processing it for the second phase for making a neural network model. The neural network model can be a single model approach or a multi model approach each focusing on the different functions of the chatbot.

VI. References

[1] https://lionbridge.ai/datasets/15-best-chatbot-datasets-for-machine-learning/

[2] https://github.com/PolyAI-LDN/conversational-datasets

[3] https://www.researchgate.net/profile/Reshmi_Sankar/publication/323451431_EMPOWERING_CHATBOTS_WITH_BUSINESS_INTELLIGENCE_BY_BIG_DATA_INTEGRATION/links/5b9351b4299bf14739257a86/EMPOWERING-CHATBOTS-WITH-BUSINESS-INTELLIGENCE-BY-BIG-DATA-INTEGRATION.pdf