




HealFavor: Dataset and A Prototype System for Healthcare ChatBot

Abdullah Faiz Ur Rahman Khilji¹ , Sahinur Rahman Laskar¹ , Partha Pakray¹ 

Rabiah Abdul Kadir², Maya Silvi Lydia³, Sivaji Bandyopadhyay¹

¹Department of Computer Science and Engineering, National Institute of Technology Silchar, India

²Institute of IR4.0, Universiti Kebangsaan Malaysia, Selangor, Malaysia

³Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

{abdullahkhilji.nits, sahinurlaskar.nits}@gmail.com, partha@cse.nits.ac.in
rabiahivi@ukm.edu.my, maya.silvi@usu.ac.id, sivaji.ju.cse@gmail.com

Abstract—A chatbot is a software application aimed at simulating real-time conversations. This system has been designed to address a plethora of domains where they have proved themselves worthy to complement or in some areas replace human-based information acquisition. Though some domains like travel and food have advanced with the growing consumer demand, the healthcare-based system does require significant advancement to address the issue of medical accessibility. The work aims at providing a suitable dataset as well as proposes a prototype system architecture. The prototype system with the self-created dataset is then analyzed on different parameters by numerous experts.

Index Terms—Data, Healthcare, Chatbot, Classification, Sequence Prediction

I. INTRODUCTION

Healthcare based AI research draws a lot of attention citing global healthcare costs in the order of trillions [1]. Any significant advancement towards automated diagnosis contributes to the cause of medical accessibility and helps in cutting costs in today's realm of high disparity between healthcare and customer affordability. In today's age, there is a greater need for improved patient-doctor communication. Since current technology does not allow for such a time-consuming exercise, innovation is inevitable. One such innovative system is a chatbot.

A chatbot is an interactive software application to simulate natural user interactions based on AI modeling. Since such a system learns from conversations, it is capable of responding to different situations based on experience. This real-time interaction requires that the data prepared for training the system is sufficient over the use cases the user might require over the course of its interaction. The data prepared also needs to be in line with current medical trends so that the system is able to correctly answer various health-related questions and arrive at a definite conclusion. Such a system would facilitate identifying a patient's condition and provide proper medical service through conversation, thereby alleviating the pressure on medical institutions together with greatly improving medical outreach. Since such data is currently not available for the system we have proposed a dataset for the commonly occurring medical conditions together with a prototype model to provide quick assistance to the patients. Apart from the

proposed prototype model, this dataset can be used in other conversational models as well.

Initial sections of our work depict the data preparation strategy whereas the later sections explain the framework utilized for testing and surveying the reliability of the proposed work.

II. RELATED WORKS

Study on chatbot based applications have been in the limelight for quite long, one such initial attempts include ELIZA [2] which scripted hardcoded patterns so that the statement made by the bot seemed pertinent enough for the conversation to move on. Complementing ELIZA [2] with knowledge-based patterns stored in Artificial Intelligence Markup Language (AIML) files gave birth to ALICE [3]. Recent implementations like Jabberwacky¹ and Cleverbot² not only use those responses but also learn from them. One such system proposes a Finite State Graph [4] in order to achieve accurate diagnosis catering to each path a user may take to arrive at the conclusion. Recent platforms like Google Dialogflow³ and IBM's Watson⁴ let developers build their system by configuring intents and entities.

Over the years many systems have been developed to provide healthcare developmental facilities. One such prominent system includes DeepMind health AI technology [5] which enables rapid diagnosis based on information collected from patients. HealthTap⁵ startup connects patients with doctors and curates a database of similar cases for faster prognosis. Molly⁶ a virtual nurse improves 20% efficiency of medical staff by monitoring a patient until the next visit. In spite of the recent developments in the field, it is quite evident that proper guidelines are virtually nonexistent. Moreover, a specialized solution for a use case are unavailable and lack holistic details required for a particular setting [6]. Since there does not exist suitable data for our system we develop a self-created dataset along with a working

¹<http://www.jabberwacky.com/>

²<https://www.cleverbot.com/>

³<https://dialogflow.com/>

⁴<https://www.ibm.com/watson>

⁵<https://www.healthtap.com/>

⁶<https://www.sensely.com/>

prototype that is both usable on various use cases and reliable based on different health ethics.

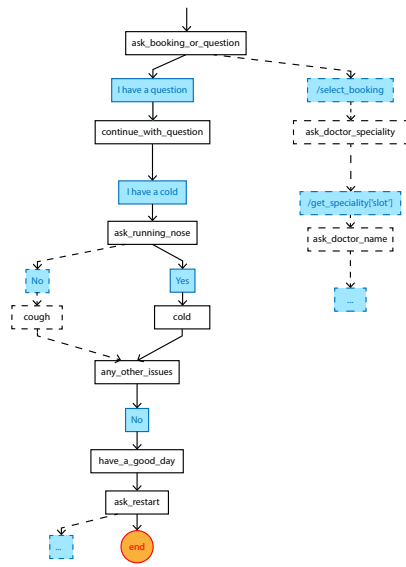


Fig. 1. Sample Data Representation

III. DATA PREPARATION

A. Sources and Observations

For creating the data we have taken the help of WebMD⁷ to acquire credible health information. Drugs.com⁸ was utilized for comprehensive drug information. For understanding the domain of patients which generally recur commonly we consulted with medical experts to gain an internal insight into those commonly occurring symptoms be it related to headache, stomach pain, skin diseases or abdominal pain. The experts also briefed us on carefully differentiating between cold and cough, diarrhea and loose-motions, etc.

B. Quality and Filtering

The prepared data for the chatbot is checked for consistency for the proper working of the system. Many diseases have common symptoms which are to be carefully taken care of. Each symptom carries certain terminology and common language associated with it. It is important that at an initial stage large classification clusters are made to encompass a large set of diseases and later narrow down based on specific easy to diagnose symptoms. Since a patient usually may have a wrong understanding of the disease symptom mapping, he may himself lead to a wrong diagnosis. Thus, the system needs to revert with a set of easy to answer questions for a proper diagnosis.

It is quite evident that such a system is quite useful for common or subtle problems that might occur and may require frequent access to the system, critical or chronic problems may also be present for such a use case it is essential for the

system to recognize these conditions and ask the patient to visit a specialist as soon as possible and recommends him/her to proceed to book an appointment using the system. These critical conditions are recognized by keywords such as chest pain, diabetes or through the questionnaire by asking whether the patient has any past medications or suffering from the symptoms for quite some time now. These critical conditions are predefined on consultation with a practicing medical doctor. The feature vectors created (as discussed in Section IV) ensures that the synonyms of these vital keywords are also taken into consideration.

C. Preprocessing

To make the data consistent for the classification task we have first preprocessed our data. Common preprocessing techniques include lowercase conversion, removal of stop words, tokenization and stemming [7]. The text is initially tokenized or segmented with the help of delimiting characters which include non-alphanumeric characters such as white space or punctuations. After tokenization, the stop words are removed which include articles conjunctions prepositions, etc. Since these words have very little dependency or contextual information removing them helps to focus on important keywords. Also, all uppercase characters are converted to their lowercase forms.

Another preprocessing technique used is lemmatization which is similar to stemming but it does not need to produce a word stem [8]. Thus, in addition to the benefits of stemming lemmatization offers to replace the word with basic word forms. Apart from the above-mentioned techniques we also explored the avenues of Phonetic-Based Microtext Normalization [9] since our data was aimed to be used for the chatbot system and was reported to improve classification accuracy by about 4%. For example, words like “gud” and “2mrw” were normalized to “good” and “tomorrow”.

D. Data Representation

For our chatbot system, we require two main components. One to classify patient intents as a multi-classification problem and secondly, accepting those intents as a sequence prediction task. The classification module accepts an input statement and correctly classifies them into predefined intents based on the word embeddings used. The sequence of intents is then fed into the sequence module to predict an output so that the conversation continues.

For the classification problem, we require different sentences that represent a similar intent. It is then up to the system’s sequence prediction module to ask the patient questions to reinforce the symptoms of the patient. These intents may be divided into two main constituents, one related to semantics and other related to healthcare. The latter being the one requiring much focus as extracting this information is highly unstructured, while the former has a predominantly more predictable format.

For the second of the two-component, we researched extensively on use cases for which the user might prefer the chatbot before visiting a physician. These use cases were then

⁷<https://www.webmd.com/>

⁸<https://www.drugs.com/>

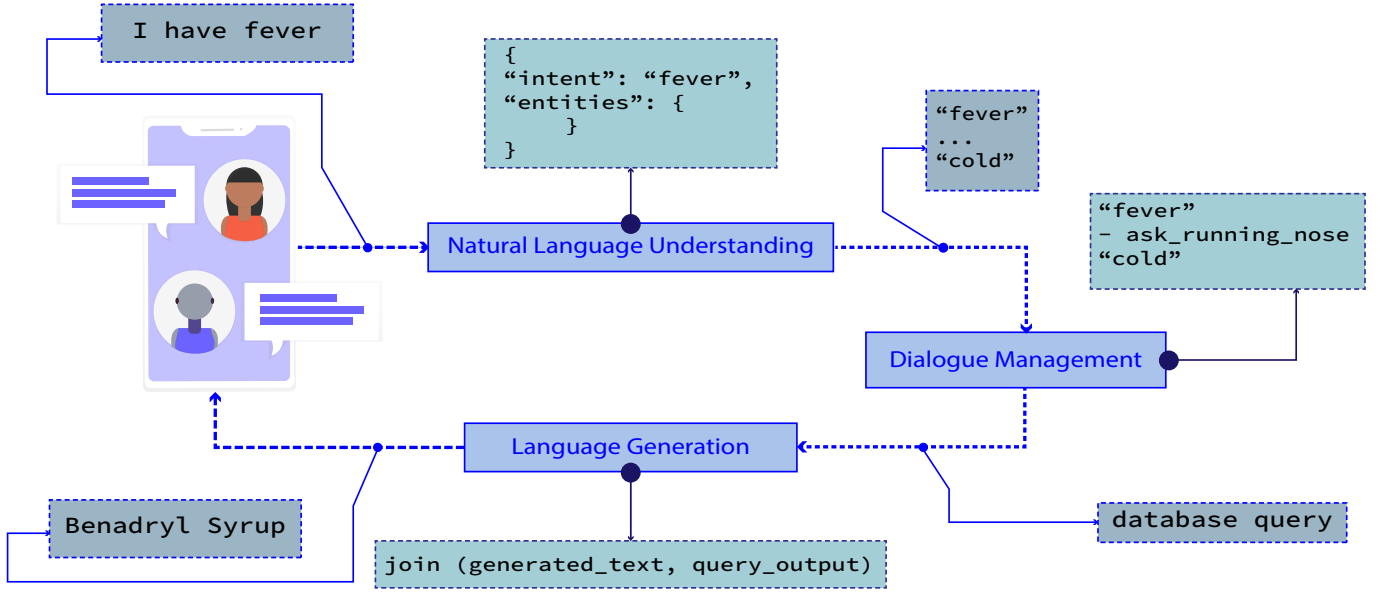


Fig. 2. Prototype System Architecture

studied for which we require a classification model to support the questionnaire. Accordingly, a list of intents was defined and common patient replies were fed into. The dataset was then constantly tested for reliability and improved in steps of reinforced manual learning to improve the confidence of the model utilized in intent classification.

Sequence prediction is a challenging part of the chatbot system and requires extensive data for different patient use cases. This can be broadly classified into the health-related sequences and booking related questions. As with the case with understanding basic semantics, the booking related questions similarly have more or less a similar definite predictable structure and thus definite sequences are presented in the dataset, the sample data representation is presented in Figure 1. The chatbot system to be proposed would comprise mainly of avenues related to healthcare and booking making it pertinent to develop a dataset catering to both these approaches. For indeterminate healthcare paths, exhaustive sets of sequences need to be prepared for unseen cases. Since the concerned domain is healthcare we have approached the problem systematically paving way for better response by the system.

The data for this healthcare system is stored in the form of text and is mainly divided into 4 main sections. One section is for the intent classification module, one contains the sequences for the sequence model. One of the other section caters to the domain of the data and the other contains the details of system actions for the user input. A sample dataset is also given at: <https://github.com/cnlp-nits/HealFavor>

IV. PROTOTYPE SYSTEM ARCHITECTURE

In this work, we propose our own system based on the conversational framework of RASA [10] employing the use of Recurrent Embedding Dialogue Policy [11] inspired by Facebook AI Research’s (FAIR) StarSpace algorithm [12] a

general-purpose neural embedding to solve a variety of tasks thereby employing cosine similarity to compare the current state with possible sets of action states.

To create feature vectors, we have used the Bag of Words (BoW) representation of the user’s message. These vectors are then passed through an embedding layer wherein embeddings are generated with the help of separate dense layers for patient input and system actions. The current user input together with the previous output from the network is then fed into the Recurrent model to calculate attention. Since the future inputs are unknown, the interpolation gate is left redundant. The input to the recurrent network is taken as the sum total of the output from the embedding layer and from the attention mechanism which is then fed into another embedding layer. The output of this layer together with the vector of system generated response is used as the final state embedding for dialogue generation. As the system needs to accommodate the skipping of a random number of steps, the LSTM states are multiplied element-wise with attention probabilities. The System Architecture is summarized in Figure 2.

For every time step we calculate the similarity given by the loss function:

$$a = \psi_- + \max_{d_-} (\text{cosm}(c, d_-)) \quad (1)$$

$$b = \psi_+ - \text{cosm}(c, d_+) \quad (2)$$

$$l(t) = \max(0, a) + \max(0, b) \quad (3)$$

The cosine similarity cosm is calculated between the target embedding d_+ and the generated embedding c . The loss is calculated by negative sampling of incorrect actions d_- ensuring similarity of incorrect actions to be low and that of correct be high. a and b are as shown in Equation 1 and 2 respectively.

A. HealFavor: Our Prototype System

For system testing our dataset on a prototype model and to check the viability of our system we have utilized the scalable framework of RASA to utilize the dataset. The RASA framework enables lucid implementation of the machine learning models and helps with integrating with the interactive front end interface together with integrating the system with Flask based pymongo architecture for Mongo database (MongoDB) access. The MongoDB enables for optimized real time storing of user chat history as well as reports and images uploaded by the user for future analysis.

V. EVALUATION

To examine the performance of the proposed prototype chatbot system, we have leveraged a user experience survey as the healthcare domain requires expert intervention for its evaluation. Thus, we have considered four evaluators who have asked a total of 162 questions to our system. The evaluators have marked the answers on a scale of 3. 1 being out of the domain and 3 correctly answered. 2 is taken as neither out of the domain or correct. Accuracy is defined as given in Equation 4. The maximum score (n_{ms}) is calculated by multiplying 3 (maximum-score) with the total number of questions which results in 486. The total score (n_{ts}) is calculated by simple summing all the scores corresponding to each question (226 in our case). This calculation results in a total accuracy of 46.50%.

$$Accuracy = \frac{n_{ts}}{n_{ms}} \times 100\% \quad (4)$$

VI. CONCLUSION AND FUTURE WORKS

In our work, we have presented a dataset suitable for training a Healthcare Chatbot. We also presented the prototype version of our system: HealFavor. Thus enabling evaluation and analysis by various domain experts to iteratively improve the dataset as well as the system. Subsequent improvements to the dataset and to the system are required which is facilitated by the thorough evaluation in order to enable scalable implementation. Since the system aims at improving medical accessibility, it is indispensable to include multilingual capabilities in our work. We also aim to include multi-modal features in the future iterations of the system.

ACKNOWLEDGMENT

We would like to thank Department of Computer Science and Engineering and Center for Natural Language Processing (CNLP) at National Institute of Technology Silchar for providing the requisite support and infrastructure to execute this work. The work presented here falls under the Research Project Grant No. CRD/2018/000041 and supported by the Department of Science & Technology (DST) and Science and Engineering Research Board (SERB), Govt. of India.

REFERENCES

- [1] D. Dranove, C. Forman, A. Goldfarb, and S. Greenstein, "The Trillion Dollar Conundrum: Complementarities and Health Information Technology," *American Economic Journal: Economic Policy*, vol. 6, no. 4, pp. 239–270, Nov. 2014. [Online]. Available: <https://www.aeaweb.org/articles?id=10.1257/pol.6.4.239>
- [2] J. Weizenbaum, "ELIZA—A Computer Program For the Study of Natural Language Communication Between Man And Machine," *Communications of the ACM*, 1983.
- [3] R. Wallace, "The elements of AIML style," *Alice AI Foundation*, vol. 139, 2003.
- [4] S. Divya, V. Indumathi, S. Ishwarya, M. Priyasankari, and S. Kalpana Devi, "A Self-Diagnosis Medical Chatbot Using Artificial Intelligence," *Journal of Web Development and Web Designing*, 2018.
- [5] J. Powles and H. Hodson, "Google DeepMind and healthcare in an age of algorithms," *Health and Technology*, vol. 7, no. 4, pp. 351–367, 2017, publisher: Springer Verlag.
- [6] R. Pryss, R. Kraft, H. Baumeister, J. Winkler, T. Probst, M. Reichert, B. Langguth, M. Spiliopoulou, and W. Schlee, "Using Chatbots to Support Medical and Psychological Treatment Procedures: Challenges, Opportunities, Technologies, Reference Architecture," in *Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics*, ser. Studies in Neuroscience, Psychology and Behavioral Economics, H. Baumeister and C. Montag, Eds. Cham: Springer International Publishing, 2019, pp. 249–260. [Online]. Available: https://doi.org/10.1007/978-3-030-31620-4_16
- [7] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Information Processing & Management*, vol. 50, no. 1, pp. 104–112, Jan. 2014. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0306457313000964>
- [8] J. Plisson, N. Lavrac, and D. Mladenec, "A Rule based Approach to Word Lemmatization," p. 4.
- [9] R. Satapathy, C. Guerreiro, I. Chaturvedi, and E. Cambria, "Phonetic-Based Microtext Normalization for Twitter Sentiment Analysis," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, Nov. 2017, pp. 407–413, iSSN: 2375-9259.
- [10] T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol, "Rasa: Open Source Language Understanding and Dialogue Management," *arXiv:1712.05181 [cs]*, Dec. 2017, arXiv: 1712.05181. [Online]. Available: <http://arxiv.org/abs/1712.05181>
- [11] V. Vlasov, A. Drissner-Schmid, and A. Nichol, "Few-Shot Generalization Across Dialogue Tasks," *arXiv:1811.11707 [cs]*, Nov. 2018, arXiv: 1811.11707. [Online]. Available: <http://arxiv.org/abs/1811.11707>
- [12] L. Y. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, and J. Weston, "StarSpace: Embed All The Things!" in *Thirty-Second AAAI Conference on Artificial Intelligence*, Apr. 2018. [Online]. Available: <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16998>