Designing Smart agent using Retrieval-Based Model

Salunke Anagha Ashok
M.Tech(ACDS)
Pune, India

Samrit Kumar Maity
CDAC
Pune, India

Manish Kumar Nirmal
CDAC
Pune, India

Prof.Mohan Nikam
Sandip University
Nashik, India

***Abstract*—A chatbot is smart when it becomes aware of user needs. A chatbot (also known as a talkbot, bot, interactive agent or Artificial conversational entity is a computer program or artificial intelligence which conducts a conversation via. auditory or textual methods. Chatbots are typically used in dialogue system for various practical purposes including customer service or information acquisition. The objective is to improve the performance of questions-answers based using retrieval-based model. Retrieval-based model use a repository of pre-defined responses. Question-answering(QA) systems have been widely developed in many domains. Genrally speaking there are two kinds of commonly used QA systems : Information-retrieval based model and generation based model . We focus on information retrieval based model question-answer systems and improving the performance of question-answering system using retrieval based model .**

Keywords—Chatbots, NLU, Retrieval-based model , neural network, Deep Learning

I. Introduction

Chatbot is an automated system designed to initiate a conversation with human users or other chatbots that communicates through text message . Chatbots have more functionalities than question and answer system and very easy to build now a day's using tools like DiaglogFlow.ai, wit.ai, LUIS, and IBM watson ,etc. Retrieval -based model which compares the queries with the message-pairs in predefined databases . People can ask any question, the QA system will find the answer from web or other sources and give the user the respective answers. The retrieval-based model can retrieve the sentence with high naturality ad fluency but is usually used in closed domain. we believe that retrieval-based method is more appropriate for our system.[6]. Glove method is to train word vector model .It use Chinese Gigaword corpus containing about 17.9 million sentences and MHMC database to train the globe model. Retrieval based model is it receives the answers or questions from a set of predefined responses and some kind of heuristic to pick an appropriate response based on the input and context.[1] Retrieval based model don’t generate any new text , they just pick response from a fixed set. It 100% works well for business problem and customer satisfaction. People can ask many questions related to a particular domain ex- Health care medicines, the questions or answers system finds answer from database. A chatbots typically have three things: Intent, Entities and action or response. Intent is the intention of the query asked by the user, named entities in query like location names, people names and dates and action is the result to throw back to the user. There are two types of chatbots:

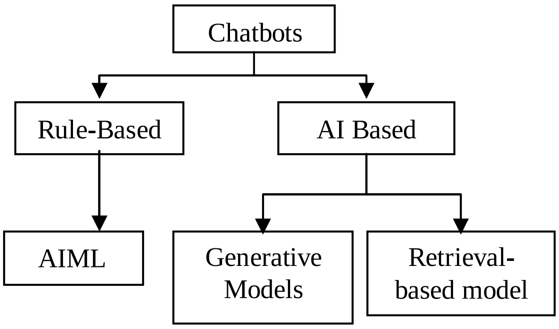


Fig1.1 Chatbots overview

A. Rule-Based :In rule-based approach, a bot answers-question based on some rules on which it is trained on. The rule can be defined simple query to complex query. The bot can handle simple query but manage to fail complex queries. Hence the bot can never pass the Turning test if based on some rule based models. One of such language is AIML(Artificial intelligence markup language) a language based on XML (Extensible markup language).

B. AI(Artificial intelligence) based: These are the bots that use some machine learning based approaches that makes them more efficient than rule based bots. There are two types bots

1. Generative model: Generative models are better than rule based models that they can generate the answers and not one replies with one of the answers from set of the answers. This makes more intelligent as they take word by word from the query and generate the answers.

2. Retrieval based model: These bots are trained on set of questions and their possible outcomes. For every question the bot can find the most relevant answers from the set of all possible answers and then outputs the answers .The complexity can range from simple rules for a query to complex rule using some machine learning algorithms to find the most appropriate answer. The retrieval -based bots come up with a set of written responses to minimize grammatical errors, improve coherence and avoid a situation a system can be hacked to respond in less appropriate ways. Retrieval-based chatbots best suit closed domain systems. Closed domain chatbots systems are built to specifically solve simple recurring problems for instance an elevator

voice assistant. The drawback of closed domain chatbot is that the set of data they come with does not come with responses for all possible scenarios.

II. QUESTION-ANSWERING SYSTEMS

There are two types of question and answering systems:

A. Information Retrieval(IR) based question answering

IR based gets the answer from documents collected , again it does not generate the answer , it just copy-paste from the documents , if the text is not present in the documents , these model can't give the answers.

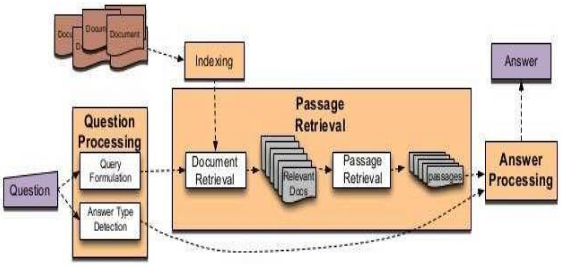


Fig. Information Retrieval

B. Knowledge based question and answering

The core idea of KBQA is convert the natural language query into structured database query. For ex., when was mady born ? it's get converted into database query and it returns the answer back to the user.[1]

III. NATURAL LANGUAGE UNDERSTANDIG

The purpose of dialogue system (DS) often also term conversational agents(CA) to converse with human to provide information , help in decision making , perform administrative services or just for the sake of entertainment.[2].

The NLU module processes the raw user input and extracts useful information and features that can be used by the dialog manager to update internal states ,send query to a knowledge base(KB) ,finds actions based on the script. NLU services is the extraction of structured, semantic information from unstructured natural language input, e.g. chat messages.

IV. TYPE OF APPLICATION PALTFORM AVAILABLE BASED ON CHATBOT TECHNOLOGY

- LUIS
- Watson Conversation
- API.ai
- Wit.ai
- Microsoft Bot framework

A. IBM Watson: IBM Watson is the best Question-answering chatbot systems. IBM Watson is the first choice as a bot platform for 61% of businesses . One of the Watson's most important part is a conversational service. It is build on neural network (one billion Wikipedia words) , understands intents, interprets, entities and dialogs, support English and Japanese languages and developer tools like Node SDK, JAVA.SDK , IOS SDK and Python SDK. IBM offers free , standards and premium plans.

B. Microsoft Bot Framework: The entire system consists of three parts: Developer portal , Bot connector and Bot directory. The Framework provides the direct line Rest API , which can be used to host a bot in an application or website. Microsoft bot framework understand user's intents. It is possible to incorporate LUIS for natural language understanding , cortana for voice and the Bing's API for search.

C. Wit.ai: Wit.ai allows using entities, intents, contexts and actions and it incorporates natural language processing(NLP). It is available for developers to use with ios android , Windows phone , python, c and it has also a javascript plugin. It supports about 50 languages and it's free.

D.Api.ai : Api.ai matches the query to the most suitable intent based on information contained in the intent and agents machine learning models. It transforms the query text into actionable data and returns output data as a response object. There are predefined knowledge packages collected over several years.

E. Chatfuel: More than 360,000 chatbots have been created using chatfuel , serving more than 17 million users globally. It consist of one or more message cards that are sent together to a bot user. Many plugins were developed: Google search, Bing search, JSON API , user input and LiveChat. It supports about 50 languages and it' s free.

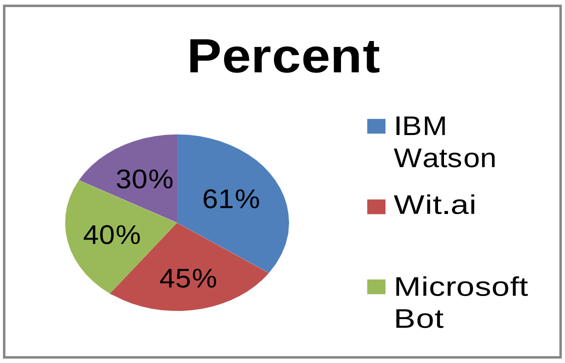


fig II. Best perform to build chatbot

V. Knowledge Base

Knowledge base(KB) are powerful tools that can be used to augment conversational models. Since knowledge bases usually entail some kind of domain specific information. These techniques usually used for task – oriented dialog systems. In a KB information related to task at hand can be stored, for example information nearby restaurants or about public transportation routes. Simple dictionaries or look-up tables can be used to match an entity with information about it.

VI. Related work

Question-answering is one of the most important application and difficult applications at the border of information retrieval. In this paper , there are two methods first method that combines information retrieval techniques optimized for question-answering with deep learning inference models for natural language inference in order to tackle the mutli-choice question answering in the science domain. The second method this two-step model outperforms the best retrieval based solver by 3% in absolute accuracy.[5]. Question-answering can be divided into closed domain and open domain . System in closed domain only focuses on answering questions in the specific field such as medicine or law and can have better performance. An open domain QA system is able to answer all the questions , not limited to a specific field . Therefore it is more difficult than a closed domain system. [6]. Bag-of-words (BOW) model is a ability for large -scale image retrieval . An image retrieval method that uses the BOW model for local feature extraction and the quantization of them to visual words. In this paper , it can improve the retrieval accuracy effectively compared to the state-of-the-art methods.[7]. Retrieval-based chatbots enjoy the advantage of informative and fluent responses, because they select a proper response for the current conversation from a repository with response selection algorithms[12]. The challenges of the task include (1) how to identify important information (words, phrases and sentences) in a context , which is crucial to selecting a proper response and leveraging relevant information in matching ; and (2) how to model relationships among the utterances in the context. For example, First , "hold a drum class" and "drum" in context are very important . Without them, one may find responses relevant to the message (i.e. fifth utterances of the context) but nonsense in the context (e.g. "what lessons do you want?"). Second the message highly depends on the second utterances in the context, and the order of the utterances matters in response selection exchanging the third utterance and fifth utterance may lead to different responses. The results show that our model can significantly outperform state-of-art methods and improvement to the best baseline mode on R @1 is over 6% [12]

VII. LSTM model

LSTM cell that process one word at a time and computes the probabilities of the possible values from next word in the sentences. The memory state of the network is initialized with a vector of zeros and gets updated after reading each word. LSTM have extra piece of information which is called memory. LSTM cell contains the following components: Forget gate, candidate layer, input gate, output gate, hidden state and memory state. Forget gate ,candidate , input gate and output gate are single layered neural networks with the sigmoid activation function except candidate layer. These gates produce vectors between 0 and 1 for sigmoid function and 1 and -1 are tanh function. If forget value is 0 then the previous memory state is completely forgotten. If forget value is 1 then previous memory state is completely passed to the cell.

VIII. Word embedding

Word embedding such as word2vec and GloVe is a popular method to improve the accuracy of model. Word2vec and GloVe carry semantic meaning-similar words have similar vectors The vector representation of the word generally has two types: one hot coding and distributed word embedding. One hot coding is used to distinguish each word in a vocabulary from any other word in a vocabulary and does not contain semantic information. So the distributed word embedding is more common in usage. Word embedding has the advantages of fixed dimensions and continuous dimensions [8]. The statistical language model and neural network language model are two classical word embedding generation methods. However, they both are complex and have many parameters which makes the computational complexity rise quickly when corpus or vocabulary reach a large level. There are many context of which the semantics are not complete and chaotic. When the new window size is small, some contexts are only part of sentence or a paragraphs and some are spans two sentences or paragraphs. The semantics and syntax of such contexts are not complete and chaotic, which causes semantic missing and confusion and directly reduce the quality of word embedding. Eg. The corpus is “Xiao-Ming like’s basketball. Xiao-Peng hates football”. With setting window size equal 3, context units may be “Xiao-Ming likes basketball”, “likes basketball Xiao-Peng”, “basketball Xiao-Peng hates”, “Xiao-Peng hates football”, in which “likes basketball Xiao-Peng”, “basketball Xiao-Peng hates” are incomplete and chaotic.

In order to solve this problem using glove model, it can handle this variable length context. There are two improvements: first the context which is used in Glove when count the co-occurrence relation between words is no longer obtain through a fixed context window, but divided by punctuation with explicit semantics. The improve model is called Glove 1 .Second on the basis of first , the context is represented by vector and introduced into the model training. The improved model is called the Glove 2.

IX. Language models

A. Glove :

The Glove language model makes use of co-occurrence relationship between words. The Glove model is based on the fact that words with higher correlation have the higher co-occurrence count.

B. Glove 1 and Glove 2:The punctuations marks are the symbol of the auxilary language recording and part of the *written lan*guage and used to indicate the pause, tone and nature of the words. The semantic information contained in the punctuation marks is a good way to divide the natural language into units with relatively complete semantics. This paper uses the segmentation characteristics of punctuation marks to divide corpus into context units with relatively complete semantics. This paper uses sentences as contexts, so that the length of the context is uncertain. This kind of context does not apply to original Glove because of its indefinite length. In order to make Glove model suitable for this kind of context, this paper proposes two improved models based on the original Glove: Glove1 and Glove 2. [8]

- Glove 1: It is only different from original glove when they calculate co-occurance realtionship between words. The window size of glove is always fixed , but glove 1 is not. The glove 1 uses the sentence as the context, and the window size need dynamic read from the context equals the length of the context sentence .
- Glove 2: It is based on the glove 1, encodes the context into the same vector form as the word embedding, and introduces the context vector into the model to participate in the training.

X.PROPOSED WORK

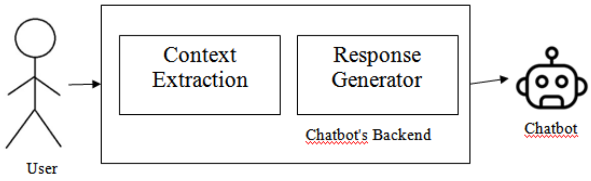


Fig. Proposed system

Generally, the chatbot will consist of two major components only 1] Context Extraction 2] Response generator. The response generator can be further broken down into above diagram fig a.

1] Encoding : This module will be filled with the entity extracted from the context extraction . This module is responsible for generating the sample action that could be potential response.

2] Retrieval: This module is responsible for generating the response from either knowledge base or on API . by checking its type.

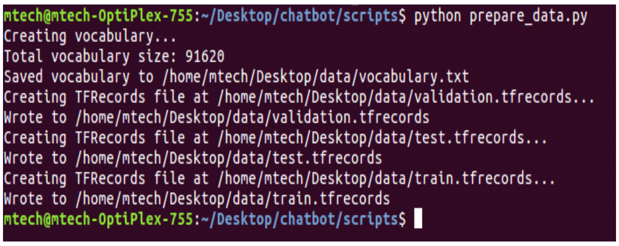
3]Candidate Selection: This module is responsible for identifying the utterance . Entity action consists of a neural networks(NN) that is encoding model fed the entity that was identified in the previous module i.e . context extraction module the entity when fed to a trained NN, it will predicted the possible response that is called the sample action. This is identified by the labeled name of it. If it is positive than the o/p is an actual response, else if it is negative then the o/p was generated.

XI. DATASET

Ubuntu Dataset: The ubuntu dialog corpus(UDC) is one of the largest public dialog datasets available. It's based on chat logs from the ubuntu channels on a public IRC network. The training data consists of 1,000,000 examples 50% positive (label 1) and 50% negative (label 0). Each example consists of a context. , the conversation upto this point ,and an utterance, a response to the context. A positive label means that the utterance was an actual response to a context , and a negative label means that the utterance wasn't -it was picked randomly from somewhere in the corpus.

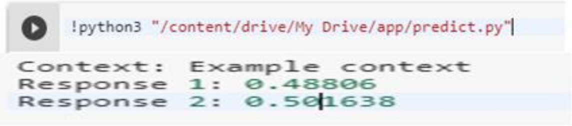
XII.EXPERIMENTS AND RESULTS

The script also replace like entities like names ,locations , organizations, URLs, and system paths with special tokens. This preprocessing isn't strictly necessary, but it's likely to improve performance by a few percent .The average context is 86 words long and the average utterance is 17 words long . The dataset comes with test and validations set. The format of these is different from that of the training data. We could work directly with CSVs . As part of the processing we also create a vocabulary.

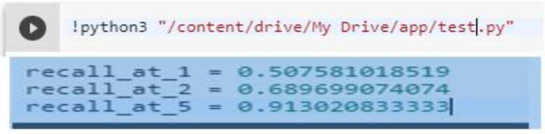


```
mtech@mtech-OptiPlex-755:~/Desktop/chatbot/scripts$ python prepare_data.py
Creating vocabulary...
Total vocabulary size: 91620
Saved vocabulary to /home/mtech/Desktop/data/vocabulary.txt
Creating TFRecords file at /home/mtech/Desktop/data/validation.tfrecords...
Wrote to /home/mtech/Desktop/data/validation.tfrecords
Creating TFRecords file at /home/mtech/Desktop/data/test.tfrecords...
Wrote to /home/mtech/Desktop/data/test.tfrecords
Creating TFRecords file at /home/mtech/Desktop/data/train.tfrecords...
Wrote to /home/mtech/Desktop/data/train.tfrecords
mtech@mtech-OptiPlex-755:~/Desktop/chatbot/scripts$
```

Evaluation of metrics: We can use the recall@k metric to evaluate our model . Tebsorflow already comes with many standard evaluation metrics that we can use , Recall@k . To use these metrics we need to create a dictionary that maps from a metric name to a function that takes the predictions and label as arguments. We use functools.partial to convert a function that takes 3 arguments to one that only takes 2 arguments . Streaming just means that the metric is accumulated over multiple batches, and sparse refers to the format of our models. The output of the predict and test data.



```
!python3 "/content/drive/My Drive/app/predict.py"
Context:  Example context
Response 1:  0.48806
Response 2:  0.501638
```



```
!python3 "/content/drive/My Drive/app/test.py"
recall_at_1 = 0.507581018519
recall_at_2 = 0.689699074074
recall_at_5 = 0.913020833333
```

XIII.CONCLUSION

We brief the introduction of various application of chatbots . and using Bag of words (BOW) have large information retrieval. IBM Watson is the best QA systems for business paltforms. Then using LSTM neural network model have improved the performance of QA systems. We Studied that word embedding vector representation technique is used for semantic information. There are language models : GloVe, GloVe1 and GloVe2 . Using SMN network, the result shows that our model on R@1 over 6% and significantaly outperforms the state-of-the-art methods. Future work, improve the performance of question-answering systems using retrieval-based model. The accuracy of the Question-answering system using deep learning technique was 90%.

REFERNCES

[1] https://medium.com/deep-math-machine-learning-ai/chapter-11-chatbots-to-question-answer-systems-e06c648ac22a.

[2] https://dzone.com/articles/understanding-architecture-models-of-chatbot-and-r

[3] Bartl, Alexander, and Gerasimos Spanakis. "A retrieval-based dialogue system utilizing utterance and context embeddings." Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on. IEEE, 2017.

[4] Skovajsová, L. (2017, October). Long short-term memory description and its application in text processing. In Communication and Information Technologies (KIT), 2017 (pp. 1-4). IEEE.

[5] Singh, R., Paste, M., Shinde, N., Patel, H., & Mishra, N. (2018, April). Chatbot using TensorFlow for small Businesses. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1614-1619). IEEE.

[6] Pirtoaca, George-Sebastian, Traian Rebedea, and Stefan Ruseti. "Improving Retrieval-Based Question Answering with Deep Inference Models." *arXiv preprint arXiv:1812.02971* (2018).

[7] Su, Ming-Hsiang, et al. "A chatbot using LSTM-based multi-layer embedding for elderly care." *Orange Technologies (ICOT), 2017 International Conference on*. IEEE, 2017.

[8] Pourreza, Alireza, and Kourosh Kiani. "A MapReduce-based online image retrieval system using bag-of-words model." *Knowledge-Based Engineering and Innovation (KBEI), 2015 2nd International Conference on*. IEEE, 2015.

[9] Song, Y., Yan, R., Li, C. T., Nie, J. Y., Zhang, M., & Zhao, D. (2018). An Ensemble of Retrieval-Based and Generation-Based Human-Computer Conversation Systems.

[10] Yu, J., Qiu, M., Jiang, J., Huang, J., Song, S., Chu, W., & Chen, H. (2018, February). Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 682-690). ACM

[11] https://www.tensorflow.org/tutorials/sequences/recurrent.

[12] Wu, Y., Wu, W., Xing, C., Zhou, M., & Li, Z. (2016). Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*.

[13] Hussain, S., & Athula, G. (2018, May). Extending a conventional chatbot knowledge base to external knowledge source and introducing user based sessions for diabetes education. In *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)* (pp. 698-703). IEEE.

[14] Doshi, Sarthak V., et al. "Artificial Intelligence Chatbot in Android System using Open Source Program-O." *Artificial Intelligence* 6.4 (2017).

[15] Tsutsui, S., & Fukuta, N. (2018, July). Efficient Teaching Support to Non-player Learning Agents on Multiplayer Games. In *2018 IEEE International Conference on Agents (ICA)* (pp. 30-33). IEEE.

[16] Subramaniam, S., Aggarwal, P., Dasgupta, G. B., & Paradkar, A. (2018, July). COBOTS-A Cognitive Multi-Bot Conversational Framework for Technical Support. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 597-604). International Foundation for Autonomous Agents and Multiagent Systems.

[17] Hamreras, S., & Boucheham, B. (2018, April). Adaptive content based image retrieval based on RICE algorithm selection model. In *Programming and Systems (ISPS), 2018 International Symposium on* (pp. 1-6). IEEE.

[18] Lowe, R., Pow, N., Serban, I., & Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

[19] Huang, H. K., Chiu, C. F., Kuo, C. H., Wu, Y. C., Chu, N. N., & Chang, P. C. (2016, October). Mixture of deep CNN-based ensemble model for image retrieval. In *Consumer Electronics, 2016 IEEE 5th Global Conference on* (pp. 1-2). IEEE.

[20] Lin, J., & Zhang, B. (2018, January). A music retrieval method based on hidden Markov model. In *Intelligent Transportation, Big Data & Smart City (ICITBS), 2018 International Conference on* (pp. 732- 735). IEEE.

[21] Banchs, R. E., & Kim, S. (2014, December). An empirical evaluation of an IR-based strategy for chat-oriented dialogue systems. In *APSIPA* (pp. 1-4).

[22] Ujjwal, D., Rastogi, P., & Siddhartha, S. (2016, January). Analysis of retrieval models for cross language information retrieval. In *Intelligent Systems and Control (ISCO), 2016 10th International Conference on* (pp. 1-4). IEEE.