Comparative study of word prediction techniques in the context of ML-ChatBot

A report submitted for the rephrase Project II (CS-300)

By

Rohit Kesarwani

Roll No. 17010102

Supervised By

Dr. Navanath Saharia

THESIS
SUBMITTED TO

Indian Institute of Information Technology, Manipur

Bachelor of Technology, Computer Science and Engineering

April, 2020

Chapter 1

Introduction

“ Verizon Ventures is an active investor in the chatbot market. According to Christie Pitts, Manager – Ventures Development of Verizon Ventures, “Chatbots represent a new trend in how people access information, make decisions, and communicate. We think that chatbots are the beginning of a new form of digital access, which centers on messaging. Messaging has become a huge component of how we interact with our devices, and how we stay connected with the people, businesses and the day-to-day activities of life. Chatbots bring commerce into this part of our lives, and will open up new opportunities.”

– Christie Pitts

It all began in 1950, when Alan Turing, an English computer scientist, published a paper entitled "Computer Machinery and Intelligence.". In this paper Alan Turing he wrote "'Can machines think?'" In his article, he outlined the Turing Test, a way to measure whether one was speaking to a human or to a chatbot. In many ways, this was the beginning of AI, a test to discover the answer to his question.

In 1966, ELIZA was created by **"Joseph Weizenbaum"**, was one of the first chatbots. Although she was able to fool some users into thinking that they were actually talking to a human, she failed the Turing Test. Despite that, the principles used in ELIZA laid a foundation for the structures of chatbots, such as keywords, specific phrases, and preprogrammed responses.

In 1995, a popular online bot was A.L.I.C.E. was made which was basesd on language-processing. Although she was unable to pass the Turing Test, she did receive many other rewards for being the most advanced bot of her time.

In 2001, that is, until smarterchild came out. In many ways, it was the precursor to Apple's Siri and Samsung's S Voice.

In 2010-2015, over the next decade or so, bots became very popular among big tech companies, starting with in Siri (2010), Google Now (2012), Alexa (2015), and Cortana in (2015).

1.1 Chatbot

A chatbot is a computer program or an artificial agent which conducts a conversation via auditory or textual methods. Now a days, Chatbots are used in the various practical purposes including cutomer services, college sevices, e-commerce sites and applications, etc.

1.2 Why chatbots are important?

Chatbot applications streamline interactions between people and services, enhancing customer experience. At the same time, they offer companies new opportunities to improve the customers engagement process and operational efficiency by reducing the typical cost of customer service.

Chapter 2

Existing System Study

2.1 Introduction

Many more chatbots are implemented in various domain. In this paper, I have implemented chatbot using various techniques(TF-IDF, Word2Vec, etc) to analyze the chatbot working and what are the drawbacks of one method to the other method.
Few chatbots are listed below which are implemented in the several domains.

2.2 KrishiBot(A chatbot for farmer)

This chatbot, I have implement during my first project which is dedicated in the area of agriculture where the farmers can query the based on the keyword matching rule, the chatbot will retrieve the information. In this project, the keyword matching rule done by using the TF-IDF with cosine similarity which have some drawbacks.
Few drawbacks are listed below:

- It only works for lexical relationship between the words, there is no semantic relationship.
- The words should be in bigram or above, i.e It won't work for unigram word.
- Cosine similarity have some threshold above which it only matches, otherwise not.

2.3 FarmChat: A Convertational Agent to Answer Farmer Queries

In this the authors acknowledge two sources of knowledge that informed the development of FarmChat: 1- Farmer's information inquiries from the Kisan Call Center (KCC) 2- Findings from a formative study with local farmers and agriculture experts.

The Government of India has made all logs of calls to the KCC that has asked by the farmers from January 2015 to September 2017 which is available publicly. In total, this corpus contains data for around 8,012,856 calls. Each call log has 11 fields, including the date and time of the call, location, crop (one of the 306 crop types), query, and the answer provided by the KCC agriculture experts.

The paper authors conducted semi-structured interviews with 14 farmers (9 male, 5 female) and 2 male agriculture-experts, in September 2017. They worked closely with a local agriculture NGO, where the two agriculture experts were employed. They helped recruit the farmers and obtain their consent for participation, following their own internal ethics policies.

The farmers and agriculture experts provided the researchers with similar questions as the ones they found in the KCC dataset. Based on both the sources, they identified that the four major areas requiring information support:

1- Plant Protection: In the KCC dataset, 60.6% of the potato farming calls were related to remedies for protecting.

2- Pests and diseases: Agriculture experts stated that a majority of farmers seek suggestions on which medicine to spray for a particular crop disease. None of the farmers aware about any disease when the researchers interviewed. Usually, farmers describe crop diseases by their visible symptoms to the agri-expert with a few back-and-forth questions, the agriculture expert hypothesizes the issue and recommends medicine with dosage information that will be provided to the crops.

3- Weather: In the KCC dataset, 39.4% of the overall calls were about weather-related questions; 13.5% of potato farming questions were about weather. Farmers eagerly wanted to know weather information, as rains can wash away expensive sprayed pesticides and weather conditions determine the best time to harvest the crops.

4- Best Practices: Information related to best practices can help increase yield in terms of the quantity or quality of potatoes. Common questions were asked by the farmers: "Till what height should I put water?" "After how many days, should I harvest?" These best practices questions comprise of 6.6% of the potato farming calls in the KCC dataset. Agriculture experts also stated that farmers consistently asked them tips to increase yield and consequently income.

Unbiased Recommendations on Products: Farmers wanted recommendations from agri-experts on products they should purchase. Questions such as "Which fertilizer to put and how many times?" and "Which seeds are the best for red potatoes?" were commonly asked. They prefer to ask these questions to agri-experts instead of local shopkeepers, believing that agri-experts would provide unbiased and trustworthy response; they feared that shopkeepers may be motivated by the profit margin of products.

2.4 College Enquiry Chatbot Using A.L.I.C.E

The author {Balbir Singh Bani, Ajay Pratap Singh} is carried on to explain the design of a chat bot specifically tailored as an application which is going to help new students to solve all the problems that they face and the questions which arises in their mind during and after the admission . In particular, the proposal investigates the implementation of ALICE chat bot system as an application named as college enquiry chat bot. A keywords-based human-computer dialog system makes it possible that the user could chat with the computer using a natural language, i.e. in English.

2.5 Online Shopping Management System Chatbot with Customer Query handling Chatbot

In the e-commerce sites, there are deal with many kind of products throughout the world. The author proposed shopping system contains different services to make user feasible in e-shopping time. When user want to buy anything from these sites, he needs guideline about product and other things in this system just like make shopping in a store. To provide this kind of things in online, they integrate an artificial chatting system with e-commerce site which gives unlimited chatting services. When user first get into the e- commerce site, he can ask queries to know in the system. E-commerce system sends customer query to the AIML Knowledge Base System to get answer by applying pattern matching algorithm. Then this answer return back to the system and then back to the user.

2.6 Chatbot for Laundry and Dry Cleaning

In this they present a chatbot for laundry and dry cleaning service. At first, they introduced a Facebook Messenger chatbot. This chatbot was used to acquaint interactions with users and to bring new customers. Based on the experience with the firstprototype that they have experienced earlier.

A customer just needs to follow these steps if he wants to interact with the chatbot:

1. make an order in the application, on the website or by phone,
2. hand clothes to Jeff at the preferred place and time,
3. take the clothes from Jeff at the arranged time and place.

Chapter 3

Methodology

In order to perform text similarity using NLP techniques some of the required steps we have to follow:

3.1 Text Preprocessing Steps

1. Collect several machine learning/AI algorithm for the dataset and save it into a text file for MLBOT(chatbot).

2. Converting whole dataset into lowercase character so that the algorithm can't treat the same word as different.

3. Tokenize the data into two parts-

   tokenize the data into sentence

   tokenize each sentence into word

4. Remove all the frequent used word called as stopwords.

5. Remove all the punctuations.

6. Lemmatize the words: It is the process of grouping together the different inflected forms of a word so they can be analysed as a single item. Eg: stems, stemming, stemmed, etc.

3.2 Various techniques to convert words into numeric

3.2.1 TF-IDF

TF-IDF stands for Term Frequency and Inverse Document Frequency which is a simple method to calculate weight of word in a document. The weight is a statistical measure which is used to evaluate how the word is important in a document.

The importance of a word increases proportionally to the number of times a word appears in the document.Tf-idf is simple and can be successfully used for stop-words filtering in various subject fields including text summarization.

TF:- Term Frequency which measures how frequently a term or a word appears in a document. Since we know that every document is different in length wise, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (the total number of terms in the document) as a way of normalization:

TF(t) = (Number of times term t appears in a document) / (Total number of terms present in the document)

IDF:- Inverse Document Frequency which measures how important a term is in the document. While computing term frequency, all terms are considered equally important. However it is known that certain terms, such as ”is”, ”of”, and ”that” (called as stop words), may appear a lot of times but have little importance. Thus we need to weight down the frequent terms while scale up the rare ones, by computing the following Inverse Document Frequency for the term:

IDF(t) = log(Total number of documents / Number of documents with term t in it)

Drawback of TF-IDF:- TF-IDF is based on the bag-of-words (BoW) model, therefore it does not capture position in text, semantics, co-occurrences in different documents, etc.

3.2.2 Skip-gram model

It is one of the type of word2vec model in which words are represented in vector space of higher dimension. Skip-gram model is used to find the most related words for a given word. It is used to predict the context for a given target word, target word is input while context words are output. First we decide what context are we looking for in terms of what will be our target words (to be predicted), source words (on bases of which we predict) and how far are we looking for context (size of window).

Eg: Considering a simple sentence, **"the quick brown fox jumps over the lazy dog"**, the task becomes to predict the context [quick, fox] given target word 'brown' or [the, brown] given target word 'quick' and so on. Thus the model tries to predict the context-window words based on the target-word.

3.2.3 CBOW model

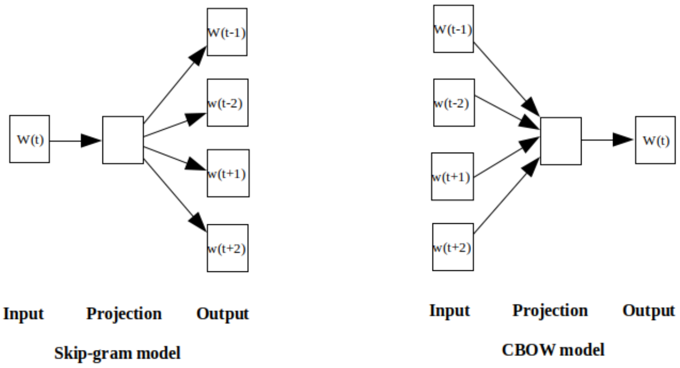
CBOW stands for Continuous Bag of Words.It is reverse of the skip-gram model. The CBOW model tries to predict the current target word (the center word) based on the source context words.

Eg: Considering a simple sentence, **"the quick brown fox jumps over the lazy dog"**, if we consider a context window of size 2, this can be pairs of **(context-window, target-word)** where we have examples like ([quick, fox], brown), ([the, brown], quick), ([the, dog], lazy) and so on. Thus the model tries to predict the target-word based on the context-window words.

3.3 Similarity techniques

3.3.1 Cosine Similarity:-

Cosine similarity gives a useful measure of how two documents are similar likely to be in terms of their subject matter. Cosine similarity measures the similarity between two document using an inner product space that measures the cosine of the angle between



the two documents. As we know that, cosine of 0° is 1, and it is less than 1 for any angle in the interval (0,pi] radians.

The cosine of two non-zero vectors can be derived by using the Euclidean dot product formula shown in equation

$$A.B = \|A\|\|B\|cos(\theta)$$

where,

A is FirstVector/document

B is Second Vector/document

θ is the angle between A and B

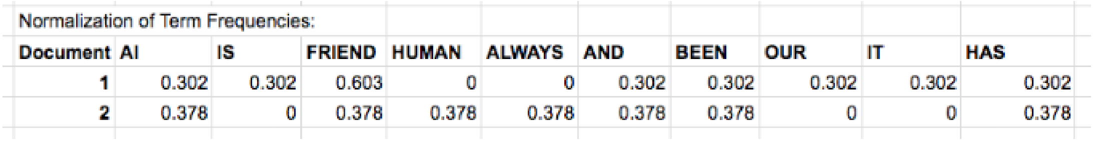
Given two vectors of attributes, A and B, the cosine similarity, cos(θ), is represented using a dot product and magnitude as:

$$cos(\theta) = \frac{A.B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i Bi}{\sum_{i=1}^{n} A_i^2 . \sum_{i=1}^{n} B_i^2}$$

Eg: Lets consider the two sentences,

S1: AI is our friend and it has been friendly.

S2: AI and humans have always been friendly.



Normalization of Term Frequencies:

Document	AI	IS	FRIEND	HUMAN	ALWAYS	AND	BEEN	OUR	IT	HAS
1	0.302	0.302	0.603	0	0	0.302	0.302	0.302	0.302	0.302
2	0.378	0	0.378	0.378	0.378	0.378	0.378	0	0	0.378

Figure 3.1: Noramlization of Term Frequencies

cos(θ) = (0.302*0.378) + (0.603*0.378) + (0.302*0.378) + (0.302*0.378) + (0.302*0.378)

cos(θ) = 0.684

3.3.2 Jaccard Similarity:-

Jaccard distance is used for comparing the two binary vectorssets. Jaccard similarity measures the similarity between two nominal attributes by taking the intersection of both and divide it by their union.

$$J(A,B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid} = \frac{\mid A \cap B \mid}{\mid A \mid + \mid B \mid + \mid A \cap B \mid}$$

If A and B are both empty, define J(A,B) = 1, where

$$0 \leq J(A,B) \leq 1$$

Eg: Lets consider the two sentences,
S1: AI is our friend and it has been friendly.
S2: AI and humans have always been friendly.

For the above two sentences, we get Jaccard similarity,
J(A,B) = 5/(5+3+2) = 0.5
which is size of intersection of the set divided by total size of set.

3.3.3 Comparison between Cosine and Jaccard similarity

1. Jaccard similarity takes only unique set of words for each sentence / document while cosine similarity takes total length of the vectors.
2. This means that if you repeat the word “frien” in S1 several times, cosine similarity changes but Jaccard similarity does not. Jaccard similarity will always be constant.
3. Jaccard similarity is good for cases where duplication does not matter, cosine similarity is good for cases where duplication matters while analyzing text similarity.

Retrieved Result:- A retrieved information is logically a subset of the representations as partitioned off by the outcome of the matching rule applied to the formal query to the chatbot.

3.4 Tools Required

- Programming Language- Python version 3.7
- Libraries- numpy, pandas, nltk, json, stopwords, csv

- Database- Plain-text/JSON format.