International Journal of Engineering Research
Volume No.6, Issue No.5, pp :281-286
ISSN:2319-6890 (online),2347-5013(print)
1 May 2017

# An Intelligent Chat-bot using Natural Language Processing

**Rishabh Shah, SiddhantLahoti, Prof. Lavanya. K**
Department of Computer Engineering, VIT University, Vellore-632014, Tamil Nadu, India

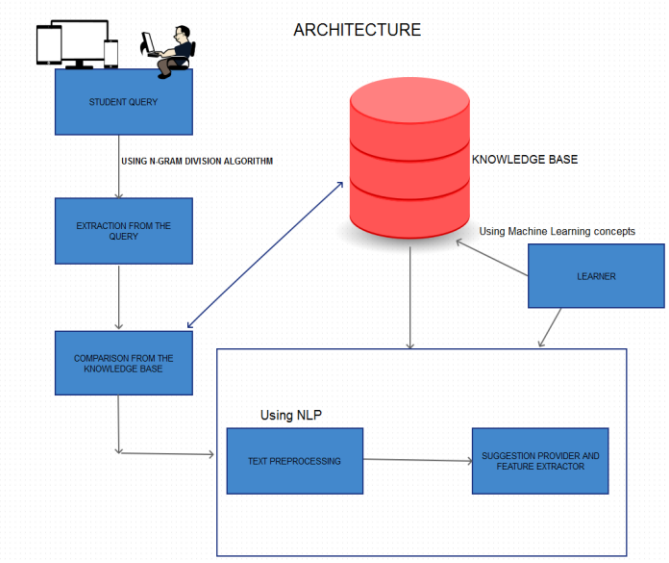
***Implementing NLP involves initiating the process of learning through the natural acquisition in the educational systems. It is based on effective approaches for providing a solution for various problems and issues in education.***
***The country where education prices are increasing day by day and the population of lower-middle/middle class is increasing at an exponential rate, we need a cheaper way for education. These people are not able to receive the level of education they deserve. Well the technology we are working on is basically for these people and also for all the learning enthusiasts (independent of their class).***

***KEYWORDS:*** chat-bot, natural language processing, skip gram, knowledge base, Unsupervised Learning

## I. INTRODUCTION

Chat-bots are computer programs that interact with users using natural languages. They use natural language processes for interaction. Chat-bot is a computer program which conducts a conversation via auditory or textual methods. Chat-bots are also used in dialog systems for various practical purposes including customer service or information acquisition. Some chat-bots use sophisticated natural language processing systems, but many simply scan for keywords within the input and pull a reply with the most matching keywords, or the most similar wording pattern, from a textual database.
The idea is a basically about creating a platform where students can learn and clear their doubts. We aim to include the concept of NLP (Natural Language Processing & machine learning concept) which will work as a tutor and a genius friend for students. This amazing feature of ours shall act as mate for students whom they can approach any time they want, can ask any question they want to and discuss with them in a more interactive way just like as we all do with our friends. All the current companies in this field have the idea of solving doubts online but the problem is that always this same set of questions are available for students which becomes kind of boring so instead our innovation called "The Personal Tutor" will create random questions and judge the student on the basis of series of questions it had asked.
The Machine Learning that we have included will continuously be learning from discussions of teachers and students. The software will then work as a perfect companion/tutor for students. These services will not only make studies at home more interesting but also children will find it as a fun activity to do.
The software will also ask the students questions that are related to their query which helps them solve their doubts and will scale down their problems, analyze them and will make sure that the student is now cleared with the concepts of that topic.

With the increase in cost of education, not everyone is able to provide their children the best education so by making this platform we are intending to provide those students the proper knowledge and a different way of learning. We want every student to get equal knowledge and a proper guidance.




### I.I PROCESS MODEL

The model which we have opted to follow for our project is the Incremental model. We have the option to follow other models but Incremental model turns out to be the best choice related to our project.
Justification:-Construct a partial implementation of a total system and then slowly add increased functionality for the developed software module.

- First we will make a database of all the common educational terms and phrases used in a particular field.
- Now we will try to make clusters of the similar words and clusters of frequently used words or phrases.
- We will use NLP to interpret the input and respond to the given question accordingly.
- Many features will be incremented step by step by given requirement and later be modified.
- Breakdown of our tasks into simpler activities like dividing and sorting words, counting the number of search for a particular words etc.

ISSN:2319-6890 (online),2347-5013(print)
1 May 2017

- Following the requirements asked by the client we tend to deliver the solution and suggest similar kind of words.
- Customer can respond to the module and if it doesn't meet his/her specifications it will learn from the user modifications.
- The model is highly compatible and efficient.

## II. LITERATURE SURVEY

[1] In this paper the author explains on how to analyze social media information using NLP. Automatic summarization is a three phase technique which reduces large paragraphs into small texts. Objects, name of a person and place can be identified using Named entity recognition. Speech tagging is most important technique where verbs, nouns, adjectives are classified. Sense of a whole sentence is checked through word-sense disambiguation. This basically all constitutes a web mining system which follows 4 steps i.e. data gathering, pre-processing, indexing and mining. So these techniques were combined and made an efficient use of NLP.

[2] This paper takes into consideration the rise in the number of documents present on the web and each document needs to be classified into some or the other category so that searching for them again can become easy. A number of pre-processing steps are discussed to reduce the complexities present with the document , this also means dimensionality reduction . The text being processed is tokenized so that is the point where NLP comes into play . The way the documents are treated in the numeric format is because of the feature selection part where the documents are processed in such a way that the learning algorithm can be implemented on it . The learning algorithms explained are the k-NN and k-means algorithm representing the supervised and unsupervised classification respectively .

[3] This paper focuses on the basics of making an expert system for Question Answering(QA) system . This system unlike the search engine does not show related documents instead shows how to create a perfect answer to a question with the help of natural language processing . Knowledge base is the first and the foremost important thing within an expert system since it is the place where the answer is formulated and takes into account many important aspects like document retrieval or information retrieval .The knowledge base is also the place where the query or the string inserted is tokenized and analyzed further and hence this is where NLP comes into play .

[4] This paper focuses on the techniques required for making a chat-bot. It also draws comparison between many chat-bots and the types of algorithms used to make them function properly.The important techniques required to design a chat-bot are parsing, pattern matching, AIML, chat script , relational database , Markov chain and language tricks . The paper discusses about how to create a knowledge base and how to interact with it, with the help of NLP. Speech analysis plays major role in the chat-bot since that is the process which will make the bot interactive. It has many steps and they are 1)converting speech to text 2)splitting the words and go for speech tagging 3)chunking of on the techniques and algorithms used to phrases 4)choosing a phrase 5)making a response.

## III. METHODOLOGY

The process starts from the input query. This input query will be entered by the student. It may be a single word or a phrase. If this query is in the form of a sentence then the query will be first passed through query modulation process.
The NLP taking place in the modulation part follows a particular pipeline
A pipeline in NLP is a chain of independent modules, each one taking as an input the output of the module before it.
Raw Text -> Tokenization -> Lemmatization -> POS-tagging -> Dependency parsing -> Role labelling
Tokenization, Lemmatization and POS tagging will play an important role in the bot's query modulation since there many words in a sentence which are not in its natural form and have to be classified into adjectives, nouns, conjunction and other speech of sentence.
So to bring a word into its natural form , the process of stemming and lemmatization come into play .
For e.g.: Running->Run
Since both mean the same thing and this is how the query is modulated.
query modulation also has the responsibility of removing the words that will not help the bot in the further process of document retrieval.
E.g.: What are the newton's Laws of motion?
The modulator removes such words from the query so that the search within the documents gives better results with optimum methods.

*Query Extraction*

The extraction will be based on N-gram division algorithm. Here n means the no. of words to be considered as a single entity for relating its metadata. They are set of co-occurring words within a defined window. For example, for the sentence *"The cow jumps over the moon"*. If N=2 (known as bigrams), then the n-grams would be:

- the cow
- cow jumps
- jumps over
- over the
- the moon

If X=Number of words in a given sentence K, the number of n-grams for sentence K would be:

$$Ngrams_K = X - (N - 1)$$

After the division of the query the metadata related to each of the gram is checked from the knowledge base. A knowledge base here is the data warehouse for all the possible questions with their solutions. And it consists of a learner which constantly updates it if any new query or a new solution for an old query is found.Training process involves loading example dialog into the chat bot's database. This either creates or builds upon the graph

ISSN:2319-6890 (online),2347-5013(print)
1 May 2017

data structure that represents the sets of known statements and responses. When a chat bot trainer is provided with a data set, it creates the necessary entries in the chat bot's knowledge graph so that the statement inputs and responses are correctly represented.

*NLP - interpreting natural language*

This field of of study is in short concerned with task of how natu ral language can be processed in
such a way that it's semantics can be understood
or interpreted by a computer program, to then act
based on these interpretations. This is an essential part of any ch atbot since it is important to try to understand what a user wants to say in order to produce a suiting answer.
Although, this is far from a trivial problem, on the contrary it is very complex since natural language often is very abstract.
As a branch to NLP there is a field of study called NaturalLangu age Understanding (NLU).
While NLP is concerned with processing natural language to be i nterpreted,
NLU focuses on the actual interpretation. Bolinda G. Chowdhur y[5] mentionthree main problems within NLU:
The first concerns the human thought process, the second the se mantics of a given input and the third knowledge
outside the program, or common knowledge.

*Document/ Information Retrieval*:

As we have seen that the documents have to be retrieved and this comes under information retrieval part .The dataset preferred will be Wikipedia dictionary since our chat-bot focuses on the education part Wikipedia can be considered as a general platform. Since there are two possible approaches for the working of the chat-bot which are retrieval based and generative based models.Since the initial working of chat-bot could be Retrieval based but the generative models which have an open domain have been working of lately so and since education being a vast field there are many answers to various questions.
There are various steps which need to be followed so that bot can retrieve a perfect document and then a perfect answer for a particular query.
Processes:
1) Response retrieval
2) Response ranking
3) Response triggering

*1. Response retrieval*

Once the query is modulated ,the utterance of the query is compared with set of documents within the knowledge base which is considered to be set of documents . Now the sentences selected will be in a triplet format and they are :(Sprev,S,Snext)
S:This represents the optimized sentence, for which the documents are to be retrieved .
Sprev :Represents the previous statement to 'S'.
Snext: Represents the next statement to 'S'.
Well the context for the given statement has to be taken into consideration since that helps us optimize the search within the knowledge base. In the latter stages it will be clear that the context helps the LSTM network make better prediction of the word.

*2. Response Ranking*:

The ranking measure for an answer can be done through the famous Google's PageRank algorithm.
In short PageRank is a "vote", by all the other pages on the Web, about how important a page is. A link to a page counts as a vote of support. If there's no link there's no support. Similarly here in this project the vote will be counted for the best possible answer to a given query.
The PR of each page depends on the PR of the pages pointing to it. But we won't know what PR those pages have until the pages pointing to them have their PR calculated and so on…
From original google paper :
'*PageRank or PR(A) can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web'*
Calculate a page's PR without knowing the final value of the PR of the other pages. Basically, each time we run the calculation we're getting a closer estimate of the final value. So all we need to do is remember the each value we calculate and repeat the calculations lots of times until the numbers stop changing much.
After ranking is done the knowledge base finds out the cluster of answers for a particular query and search for the best rank answer from the cluster and passes out as the output.

*3. Response triggering*:

Now that the documents have been ranked a classification algorithm mostly for the unlabeled dataset so unsupervised learning comes into play. Now by unlabeled dataset we mean that the documents that have been ranked against the initial query have to classified into different category and on the basis of the number of topics which have to be covered for studying . Now this makes the process for answering to a particular query even easier

*3.1 Text clustering*

Since there are so many words present within a document and every document havemany important words so this is where textual clustering is important. For example a document of

*3.1.1 Flat clustering*

This is the technique which will be used for the text clustering since it uses the K- means algorithm which can reproduce better results on higher number of datasets.
K-means separates objects based on the attributes with the help of k-centroids,well K is user defined hence the number category of the questions will be able to help decide the number of .cluster centroids.
The k-means will basically be following this standard algorithm:

- Choose any k number of clusters which are to be determined
- Choose any number of k objects randomly as the initial cluster center
- Repeat

(Assign each object to their closest cluster
Compute new clusters, i.e. Calculate mean points. Until

ISSN:2319-6890 (online),2347-5013(print)
1 May 2017

there are no changes on cluster centers OR No object changes its cluster )

Ways to optimize results:
Now if there are problems in a document set, than it is mainly because it contains many *outliers*, now by outlier it means that, those documents that are really far away from any other documents due to which they do not fit well into any cluster. Now frequently, if a particular outlier is chosen as an initial cluster center , then no other vector would beassigned to it during fore coming iterations. Thus, it ends up as a *singleton cluster* (a cluster with only one document).

1. Not including outliers in the initial set.
2. Going with multiple starting points and choosing the clustering with the lowest cost

*3.2 Document representation*:
Now the question remains how is text going to be clustered and on what basis ?Well the documents can converted into vector format and this is done with the help of vector space model. Since the documents are in subjective format and hence they need to be numeric format to be clustered or even be compared .Hence this is where TF-IDF comes into play.

TF-IDF
This stands for term frequency-Inverse Document Frequency, to reflect importance of every word within the documents the concept of using weights derived by TF-IDF . It is one of the most common weighing method to infer a particular document within the Vector Space Model. It has to be used when the document retrieval process is taking place .
TF: Term frequency, used to represent the frequency with which a particular words repeats within that document.

IDF: Inverse Document frequency ,this factor comes into play to determine a particular weight to a particular words since some meaning less words repeat a lot hence they are weighed less and the once that occur rarely are weighed more .

Hence a combination of both of them makes it lethal since it help in selection of the perfect document.

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \log\left(\frac{\mathrm{N}}{\mathrm{df}_i}\right)$$

$\mathrm{tf}_{i,j}$ = total number of occurences of i in j
$\mathrm{df}_i$ = total number of documents (speeches) containing i
N = total number of documents (speeches)

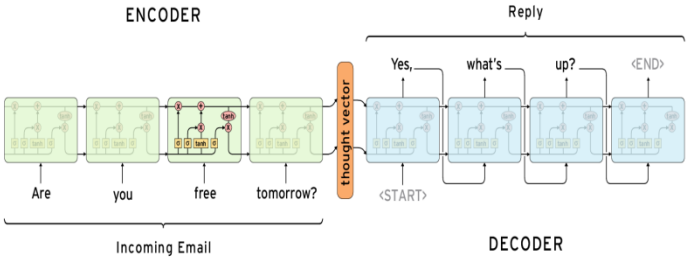
*3.2.1 Finding Similarity Score*
Use cosine similarity to identify the similarity score of a document. The method FindCosineSimilarity takes two arguments which can be considered as two vectors which contain two documents and hence it gives the similarity index between the two documents and this index is between 0 to 1.

3.3 *Clustering*
Now based on the above functions and steps the documents are clustered,well the documents sharing the maximum similarity index tend to be within the similar cluster .This is how K-means comes into play.

*ANSWERING THE QUERY*:

*1. Word2vec*
It has two parts which are continuous bag of words and skip gram. Since the continuous bag of words is not our concern so the focus shall be maintained on Skip-gram algorithm.
As already discussed for text clustering we need to convert the subjective part into numeric form hence while the bot is coming up with an answer to a relevant query , it will need to vectorize the words from the relevant document and hence to create a proper answer a large corpus would be required hence skip gram method .

*1.1 Skip Gram*
Data sparsity is a large problem in natural language processing and hence skip gram has been used to increase the number of training sentences and hence helps in increasing the size of training corpus. Skip-grams reported for a certain skip distance k allow a total of k or less skips to construct the n-gram. As such, "4-skip-n-gram" results include 4 skips, 3 skips, 2 skips, 1 skip, and 0 skips
For e.g.:"Insurgents killed in ongoing fighting."
2-skip-bi-grams = {insurgents killed, insurgents in, insurgents ongoing, killed in, killed ongoing, killed fighting, in ongoing, in fighting, ongoing fighting}
*2. Sequence to sequence model:*
We will be following the sequence to sequence model. Well a sequence to sequence model consist of a pair of RNN(recurrent neural network), the first one works as an encoder which processes the documents and the second one is a decoder which is going to generate an output. The figure is given below:



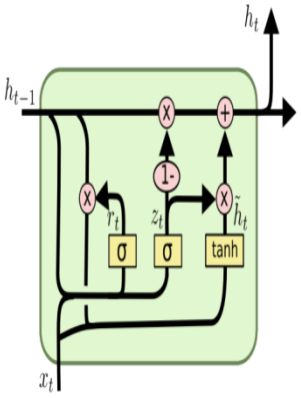

The weights are shared between the encoder and the decoder while they use different set of parameters.
The sequence to sequence model takes the input as the vector and hence the skip gram model comes into play.
Now the input to the LSTM network shall be from the finalized document, the one that has been ranked the highest and the sentences are selected from it with the contextual reference.

*3. LSTM:*
Long Short Term Memory networks – usually just called "LSTMs" – are a unique type of RNN, which have the capability of learning long-term dependencies.. LSTMs are especially designed to avoid the long-term dependency problem. Remembering information for

ISSN:2319-6890 (online),2347-5013(print)
1 May 2017

long periods of time is practically their default behavior. All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single "tanh" layer.
E.g.:" the birds are flying in the _____"
Now the LSTM with the help of skip gram model creates a large corpus and hence creates an answer "SKY"
Hence LSTM helps in creating the sentence " the clouds are in the sky".

*3.1 LSTM Architecture*
LSTM cells/ blocks always contain three to four "gates" which can be further used to control the information that has been continuously flowing into or out of their memory. These gates have been implemented using the sigmoid function(logistic function) to get a final a value in between 0 and 1. Multiplication operation has been applied to this value to partially allow or deny the flow of information into or out of the memory. The various gates used,
"input gate" : The one that controls the level up to which a new value flows into the memory.
"forget gate" :controls the extent to which a value will remain within the memory.
"output gate": This controls the extent up to which the value in memory is used for computing the outputactivation of the block




$$z_t = \sigma\left(W_z \cdot \left[h_{t-1}, x_t\right]\right)$$

$$r_t = \sigma\left(W_r \cdot \left[h_{t-1}, x_t\right]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot \left[r_t * h_{t-1}, x_t\right]\right)$$

$$h_t = \left(1 - z_t\right) * h_{t-1} + z_t * \tilde{h}_t$$

*3.2 Working:*
The LSTM has many steps before we create the output:
Step1: Deciding which part of the query has to be passed through the cell state, hence forget gate comes into play which has a sigmoid gate .
step2: decide what new information we're going to store in the cell state.
2a) a sigmoid layer called the "input gate layer" decides which values we'll update.
2b) a tanh layer creates a vector of new candidate values, C(t), that could be added to the state.
step3)Combine the previous two steps.
step4) update the old cell state, C(t-1), into the new cell state C(t). The previous steps already decided what to do, we just need to actually do it.
step5) Finally, we need to decide what we're going to output. This output will be based on our given cell state, but will be the one with a filtered version. First, we run within the sigmoid layer which decides what parts of the cell state we're going to output.It is then, we put the cell state through tanh (to push the values to be between −1 and 1) and multiply it by the output of the sigmoid gate, so that we only output the parts we decided to.

The above given algorithm works as the final step since it is the one which works on the final documents taken into consideration and LSTM then takes the sentences from those documents to finally produce the desired output/answer to the query registered within the bot .

## IV. RESULTS AND DISCUSSION

A chat-bot with such important functions should have a particular pipeline since it helps in deciding which part or the component would be requiring the highest level of precision.The NLP and ML have been applied at their respective levels and hence it would be perfect to say that they work hand in hand with each other along with IR .
When it comes to such chat-bots which follow a generative model which is open is an open ended field ,the questions / query presented to the chat-bot does belong to particular field and it is not necessary to have a previous data set for such models and hence the unsupervised learning comes into play.
Intensive usage of unsupervised learning algorithm comes into play since the data or query being modulated does not necessarily have some stored documents ,so the IR part would not work and hence the learning algorithm will come into play.
So we have used K-means clustering algorithm which is most commonly used algorithm since it work s best with the huge datasets.
It would be best to build the chat-bot in AIML (Artificial Intelligence Markup Language).Since it can work with the common questions directly which are stored within the database with pattern matching. Since this works on fixed template s it is going to provide better results until and unless machine learning algorithms come into play. Since they work on providing or deciding the rules hence creates a simple rule based within the complex model.
Now coming to the responsiveness of the chat-bot which is the answer generation and extraction ,the extraction process is including in the IR process where many algorithms are taking place to decide which documents and sentences to be referred.
Now comes into play the LSTM which is used to develop the answer from a particular documents and sentence . For creating a perfect answer a large dialog corpus will be required hence the n-skip gram method has been implemented to produce optimum results. This Word2Vec procedure takes into consideration skip gram which turns out to be procedure to create word embedding which will be further used to connect a particular vocab to its dense vector. The dense vectors created for the vocab will be passing throughan intermediate layer(LSTM) which will be further used in prediction of the next word of the statement or the answer generated.
LSTM needs to function properly every step in the algorithm should verified . Since LSTM is our final stage the errors within this stage should be prevented .The query and the answers should be padded or the process of bucketing should be

ISSN:2319-6890 (online),2347-5013(print)
1 May 2017

implemented to solve the issues of variable length or large sentences.

Now since the part with LSTM is completed we need to include softmax layer which basically the output layer of the neural network and this layer will actually determine the next word since it takes into consideration probability and the word with the highest probability shall take place next hence this is how so far the chat bot with a generative , seq2seq model shall work over many data sets and shall be implemented and cross verified.

We realize that such chat-bots will play an important part in a learner's life because the fact that it is different for each user the way it learns is the way it can reciprocate and help the user. The user's performance and the question pattern can be analyzed and the level at which the user is how much more he/she has to work can be predicted with the help of other relevant data from the other users.

## V. CONCLUSION

It is hard to draw any conclusions at all considering that the amount of user tests conducted was small. With better datasets and knowledge base ,better results will be showcased by the bot. But since there are some major components where there have been dynamic changes.The IR (information retrieval) process tends to change with time. Different algorithms have been used and with technology evolving the retrieval process including the triggering part is getting faster.

Not just IR but also the answering part with the word embedding technique coming into play makes it easier to relate or connect to different words. Such word embedding makes LSTM network more efficient.

The tendencies of the results are promising in that a NLP and self-learning techniqueswould be a good addition to a chat-bot as to make it more human-like. Although,further research would be required to make any real conclusions.

## VI. FUTURE SCOPE

This chat-bot here we aim to make is using the most trending technology of natural language processing(NLP) but what we further aim to add a suggestion provider which will suggest user the other queries similar to their question. Until now the bot has been working towards the answering of a query now the bot will post a query and will verify the answer and hence does analysis of that particular user. So our platform with the chat-bot analyses the errors and helps the user in a better way. To get the reverse part started the LSTM neural network should perform in a better way and the attention mechanism shall embellish it.

Implementing an attention mechanism with in the chat-bot will make it work in an efficient way. Now the attention mechanism will no longer encode the full input within LSTM , It allows the model to learn which part to be encoded and how the output will be created .

A suggestion for future research on the same topic would be to l ook more on which categories that is most relevant to evaluate re garding human-like behavior.

## REFERENCES

*i. John Selvadurai -A Natural Language Processing based Web Mining System for Social Media Analysis - International Journal of Scientific and Research Publications, Volume 3, Issue 1, January 2013*

*ii. Aurangzeb Khan, BaharumBaharudin, Lam Hong Lee*, Khairullah khan- A Review of Machine Learning Algorithms for Text-Documents Classification journal of advances in Information Technology, Vol. 1, No. 1, February 2010.*

*iii. Shweta C. Dharmadhikari#1, Maya Ingle, Parag KulkarniEmpirical Studies on Machine Learning Based Text Classification Algorithms Advanced Computing: An International Journal ( ACIJ ), Vol.2, No.6, November 2011*

*iv. Dr. John Woods- Survey on Chatbot Design Techniques in Speech Conversation Systems(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 7, 2015*

*v. Gobinda G. Chowdhury (2003). "Natural Language Processing" Annual Review of Information Science and Technology, 37 . pp. 51-89*