A

SEMINAR REPORT

ON

NATURAL LANGUAGE PROCESSING IN CHATBOT TECHNOLOGY

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE OF

BACHELOR OF ENGINEERING
INFORMATION TECHNOLOGY

BY

Pranish Warke
Roll No: 33381
Exam Seat No:

Under the guidance of
Mr.Sachin D. Shelke

Department Of Information Technology
Pune Institute of Computer Technology
Sr. No 27, Pune-Satara Road, Dhankawadi
Pune - 411 043.
AY: 2022-2023

SCTR's PUNE INSTITUTE OF COMPUTER TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY

CERTIFICATE

This is to certify that the Seminar work entitled
NATURAL LANGUAGE PROCESSING IN CHATBOT TECHNOLOGY

Submitted by

Name : Pranish Warke
Exam Seat No: .

is a bonafide work carried out under the supervision of Name of the Seminar Guide and it is submitted towards the partial fulfillment of the requirements of Savitribai Phule Pune University, Pune for the award of the degree of Bachelor of Engineering (Information Technology).

Mr.Sachin D. Shelke Seminar Guide	Dr. A. S. Ghotkar HOD IT

Dr. S. T. Gandhe
Principal

Date: 07/11/2022
Place: Pune

Acknowledgement

Purpose of acknowledgments page is to show appreciation to those who contributed in conducting this dissertation work / other tasks and duties related to the report writing. Therefore when writing acknowledgments page you should carefully consider everyone who helped during research process and show appreciation in the order of relevance. In this regard it is suitable to show appreciation in brief manner instead of using strong emotional phrases.

In this part of your work it is normal to use personal pronouns like “I, my, me” while in the rest of the report this articulation is not recommended. Even when acknowledging family members and friends make sure of using the wording of a relatively formal register. The list of the persons you should acknowledged, includes guide (main and second), head of dept, academic staff in your department, technical staff, reviewers, head of institute, companies, family and friends. You should acknowledge all sources of funding. It’s usually specific naming the person and the type of help you received. For example, an advisor who helped you conceptualize the seminar,someone who helped with the actual building or procedures used to complete the seminar,someone who helped with computer knowledge, someone who provided raw materials for the seminar, etc.

Name : Pranish Warke
Exam Seat No:

Abstract

Chatbot Technology using natural language processing is a computer program, which responds like a smart entity when conversed with through text or voice and understands one or more human languages by Natural Language Processing (NLP). Basically, a chatbot is defined as “A computer program designed to simulate conversation with human users, especially over the Internet”. Artificial intelligence (AI), natural language processing (NLP), and machine learning are chatbot underlying technologies. To make the message understandable for a chatbot, NLP follows main techniques : Semantic analysis and neural machine translation. Chatbots are also known as smart bots, interactive agents, digital assistants, or artificial conversation entities. Chatbots are capable of constant and automated refinement. Smart bots are used in spam detection, chatbots are used as digital assistants by Google, Amazon, etc. Chatbots are used to enhance customer experience by various healthcare and e-commerce websites. Customer service chatbots provide value to customers as well as businesses.

Keywords: Natural Language Processing, Machine Learning, Artificial Intelligence, neural networks, NLU, deep learning

Contents

List of Figures

List of Tables

Abbreviations

NLP	:	Natural Language Processing
AI	:	Artificial Intelligence
LSTM	:	Long Short Term Memory
GRU	:	Gated Recurrent Unit
RNN	:	Recurrent Neural Networks
DNN	:	Deep Neural Networks
CNN	:	Convolutional Neural Networks
GUI	:	Graphical User Interface
ML	:	Machine Learning

1. Introduction

1.1 Introduction

Natural Language Processing is a field in Artificial Intelligence where computer understands the human languages like English, Marathi, Hindi etc. Chatbot is an application of Natural Language Processing where user can chat with the computer like a user would have otherwise done with a human.

1.2 Motivation

Nowadays, Chatbot technology is widely used by various organizations over the Internet to provide support and service to their customers. Using Natural Language Processing, we can improve the interaction between user and chatbots by using spoken language for communication.

1.3 Objectives

Studying the implementation of natural language processing and neural networks in AI-powered chatbot technology.

1.4 Scope

Scope consists of study of general purpose in-app support chatbots that provide 24/7 customer service and text-to-text conversation using natural language processing.

2. Literature Survey

This seminar report consits of various methodologies,approaches and algorithms to implement Natural Language Processing in Chatbot Technology.

1. A quite significant work regarding Artificial Neural Networks, and Natural Language Processing has be done in recent times.Some of these source referred are :-

i. G Krishna Vamsi , Akhtar Rasool , Gaurav Hajela A Deep Neural Network Based Human to Machine Conversation Model , IEEE 11th ICCCNT 2020 July 1-3, 2020 - IIT - Kharagpur.

ii. Ramakrishna Kumar, Maha Mahmoud Ali(2020) A Review on Chatbot Design and Implementation Techniques , National University of Science and Technology, Muscat , Oman , Feb 2020.

iii. Moneerh Aleedy, Hadil Shaiba, Marija Bezbradica (2019) Generating and Analyzing Chatbot Responses using Natural Language Processing , International Journal of Advanced Computer Science and Applications.

2. Also, work regarding implementation on Natural Language Processing using sequence-to-sequence model to create a chatbot was referred :-

i. KULOTHUNKAN PALASUNDRAM,NURFADHLINA MOHD SHAREF, KHAIRUL AZHAR KASMIRAN, AZREEN AZMAN Enhancements to the Sequence-to- Sequence-Based Natural Answer Generation Models, Intelligent Computing Research Group, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, IEEE Access, February 24, 2020.

3. Methodologies

3.1 Framework/Basic Architecture

Basic Architecture of a Chatbot using natural language processing :

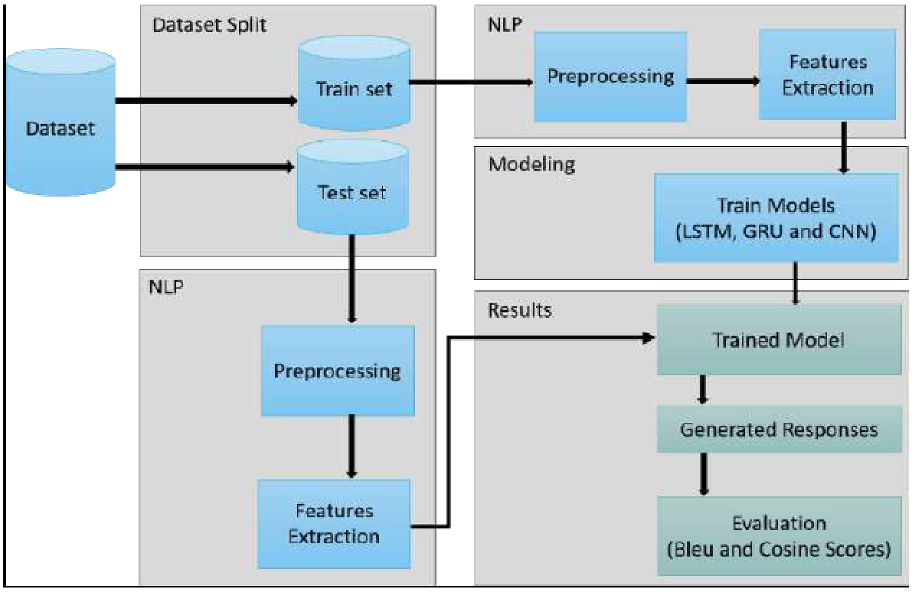

Figure 3.1: Chatbot Architecture

3.2 Different Approaches

Three approaches are studied for implementing natural language processing in chatbot technology:

1. Bidirectional Recurrent Neural Networks are the most appropriate models for processing sentences , as they have achieved substantial success in text categorization and machine translation.

2. Deep Neural Networks (DNN) can be defined as RNNs with additional depth , that is, an increased number of hidden layers between the input and the output layers.

3. Convolutional Neural Networks (CNN) are chosen mainly for its efficiency, since CNN is faster compared to other text representation and extraction methods.

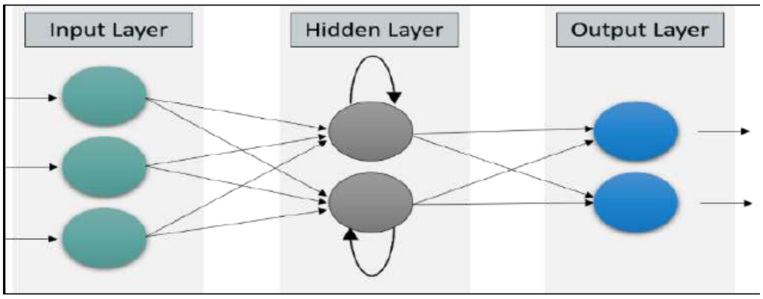

Figure 3.2: Bidirectional RNN Architecture

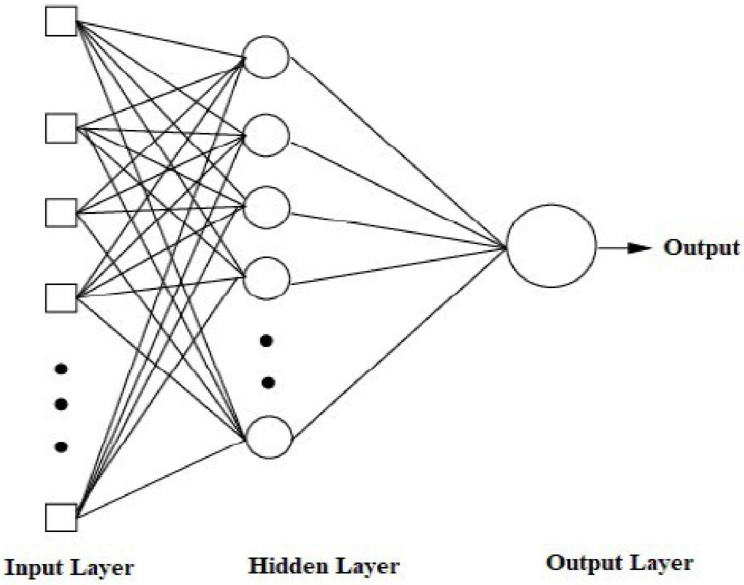

Figure 3.3: DNN Architecture

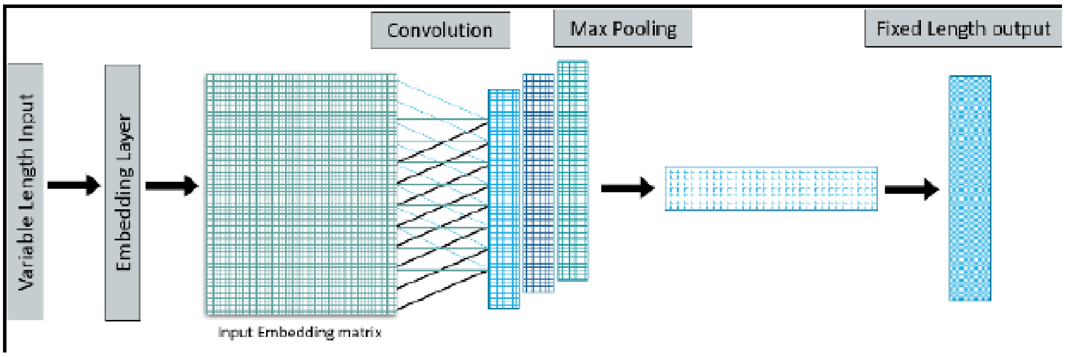

Figure 3.4: CNN Architecture

3.3 State-of-the-art Algorithms

Sequence to Sequence Model algorithm is applied to implement natural language processing in chatbot technology using Bidirectional RNNs.

Sequence-to-sequence models are used in many fields, including chat generation, text translation, speech recognition, and video captioning. The input text enters the encoder network in reverse order, then it is converted into a sequence of fixed length context vector, which is then used by the decoder to generate the output sequence

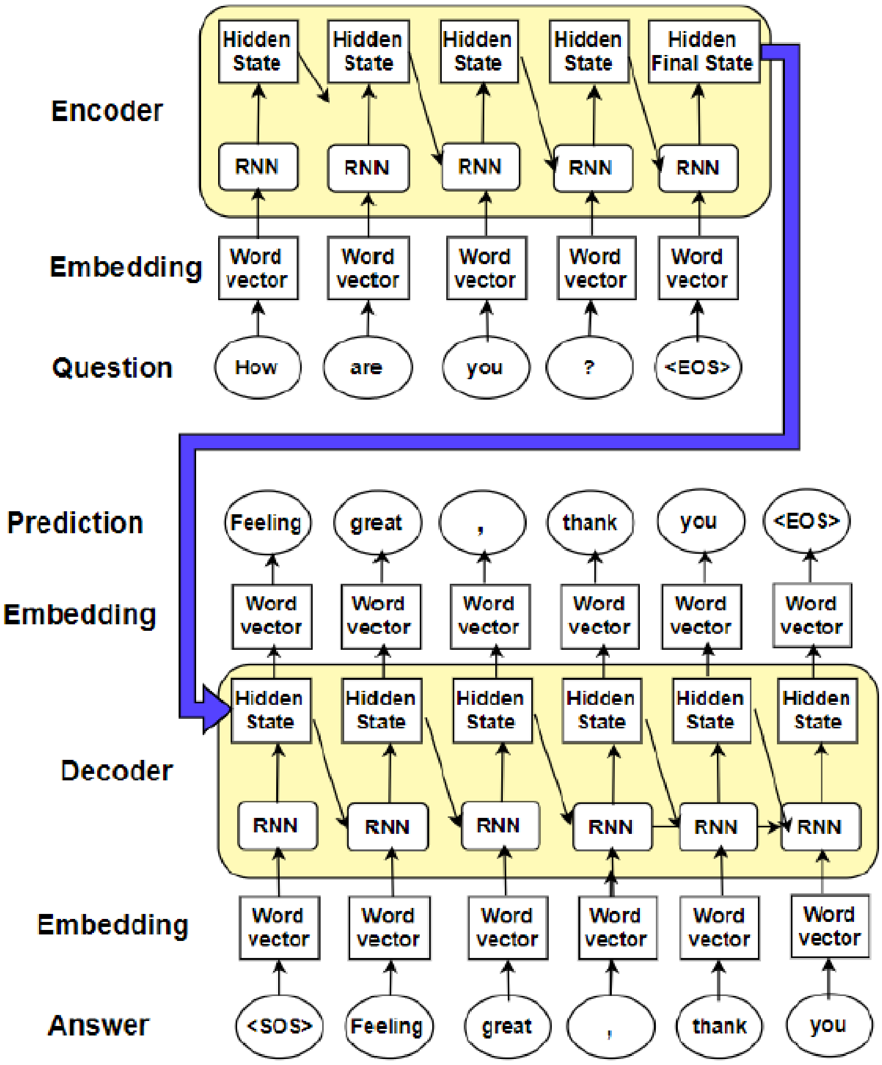

Figure 3.5: Sequence to Sequence Model

The following activity happens during the encoding phase:-

1) The question string ("How are you ?") is first tokenized into a list of tokens. A token can be a word or a sub-word of one (1) or more characters and symbols.

2) Each token is then associated with a vector which is a subset of input vocabulary.

Algorithms

Algorithm 1 : Training a Sequence to Sequence Model With Attention Mechanism for Answer Generation

Input: Question (X)-Answer(Y) pairs, Maximum Answer Tokens to Generate (T), Number of Epochs

Steps :

```
For epoch 1 to Number of Epochs
For the batch of question-answer pairs, X and Y Do
1. Encoder: Perform question encoding, generate the hidden states for each timestep
2. Decoder:
2.1 Generate tokens (with the highest probability) one by one by feeding weighted hidden states from Encoder until maximum answer length is reached, or end of sequence token is generated
2.2 Join all tokens to generate an answer (Y')
3. Calculate the cross-entropy loss (the difference between Y and Y')
4. Update the model parameters
End For
End For
```

Output: Trained Seq2Seq Model

Algorithm 2 : Sequence to Sequence Model Prediction (Beam Search)

Input: Trained Seq2Seq Model, Question (Q), Beam Size (k)

Steps :

```
1. Encoder: Perform question encoding, generate the hidden states for each timestep
2. Decoder:
2.1 Generate tokens (with probability scores) by feeding weighted hidden states from Encoder
2.2 Sort and get the k top tokens based on probability score (this is the hypothesis with 1 word)
2.3 Add to beam search list
2.4 For each of the hypothesis in the beam search list
2.4.1 Generate tokens (with probability scores) by feeding weighted hidden states from Encoder
2.4.2 Sort and get the k top tokens
2.4.3 Add to the hypothesis in the beam search list
2.5 Calculate probabilities for each of the k2 hypothesis, sort and keep only k top hypothesis
2.6 Repeat steps 2.4 to 2.5 until maximum answer length is reached or end of sequence token
```

is generated

2.7 Sort beam search list and return the hypothesis with the highest probability

Output: Answer to Question(Q)

3.4 Discussion

Bidirectional Recurrent Neural Networks and Deep Neural Networks are more suitable to implement natural language processing in text-to-text or speech-to-text chat bot technology as compared to Convolutional Neural Networks.

Comparison between RNN and CNN -

RNN	CNN
The conversation is a sequence of words (Can handle arbitrary input and output lengths)	The conversation is a fixed size (Cannot handle sequential data)
Considers the previously received inputs along with the current input. The LSTM cells used in RNN model allow RNN to memorize previous inputs.	Considers only the current input, and it cannot remember the previous input.
Uses time-series information, hence it is the best suitable model for systems that take the conversation context in its consideration.	Uses connectivity pattern between its neurons, hence the neurons are arranged in such a way that enables CNN to respond to overlapping regions tiling the visual field.
Used to create a combination of subcomponents (e. g. text generation, language translation)	Used to break a component (e. g. image) into subcomponents (e. g. object in an image)
It is ideal for text and speech generation.	It is ideal for images, videos processing and ranking candidate sentences.

Table 3.1: RNN vs CNN

4. Implementation

4.1 Algorithm/Methodologies

Implementation of NLP using Sequence to Sequence Model Algorithm.

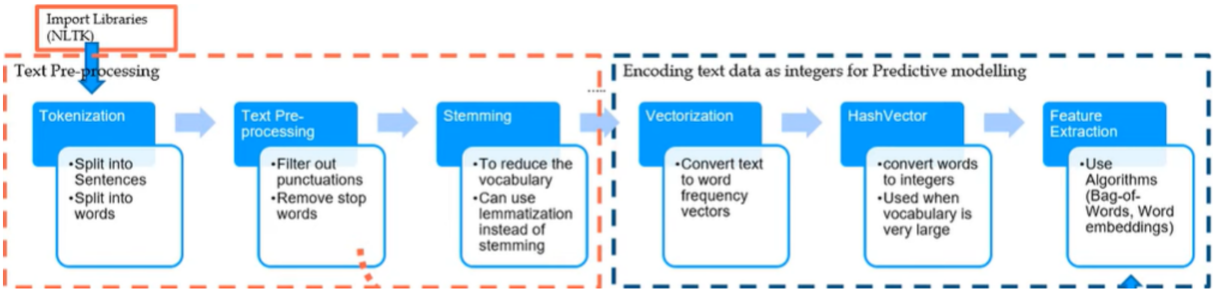
Import Libraries (NLTK)
Text Pre-processing
Encoding text data as integers for Predictive modelling
Tokenization
•Split into Sentences
•Split into words
Text Pre-processing
•Filter out punctuations
•Remove stop words
Stemming
•To reduce the vocabulary
•Can use lemmatization instead of stemming
Vectorization
•Convert text to word frequency vectors
HashVector
•convert words to integers
•Used when vocabulary is very large
Feature Extraction
•Use Algorithms (Bag-of-Words, Word embeddings)

Figure 4.1: NLP using Sequence to Sequence Model

4.2 Proposed Solution

Deep Recurrent Neural Networks are the most suitable to implement NLP in chatbot technology. Python code for steps involving Implementation of NLP using sequence to sequence model to create a general purpose chatbot is as follows -

Tokenization

```
sent_tokens = nltk.sent_tokenize(raw)# converts to list of sentences
word_tokens = nltk.word_tokenize(raw)# converts to list of words
```

Lemmatization

```
# LemTokens will take as input the tokens and return normalized tokens.
def LemTokens(tokens):
   return [lemmer.lemmatize(token) for token in tokens]
```

Pre-Processing

```
remove_punct_dict = dict((ord(punct), None) for punct in string.punctuation
                                    )
def LemNormalize(text):
   return LemTokens(nltk.word_tokenize(text.lower().translate(
                                         remove_punct_dict)))
```

Vectorization

```
 TfidfVec = TfidfVectorizer(tokenizer=LemNormalize, stop_words='english')
```

4.3 Software Requirement Specification

4.3.1 Constraints and Assumptions

Dataset used is small, and training it over deep neural networks model leads to over fitting the model, to find the optimal learning rate appropriate precautions must be taken while building the model.

Attribute Name	Description
Tags	Tags are keywords assigned to the patterns and responses for training the chatbot .e.g., ”greeting”, ”bye”.
Patterns	These are the types of queries asked by the user to the chatbot.
Responses	Responses are the answers generated by the chatbot for the respective queries.
Context	Context is given for which the queries require a search and find operation.

Table 4.1: Dataset Description

4.3.2 Platform for Implementation and its Specifications

- Latest Python Version 3.9.0 installed on your PC(Windows/Linux).
- Jupyter Notebook installed
- A software GUI library called ’Tkinter’ for Python is used to create a simple user interface.
- Libraries Imported :-

```
import nltk # Natural Language ToolKit used for NLP
import warnings
import numpy as np
import random
import string
from sklearn.feature_extraction.text import TfidfVectorizer
```

4.4 Result

Rules-Based Chatbots vs. NLP based Chatbots

A rules-based solution “allows brands to deliver experiences to specific segments of people based on the manual creation and manipulation of business rules.” For instance, a brand may set up a chatbot rule that states “If a person mentions the word ‘return,’ have the chatbot reply with our QnA page on how to return a product.” NLP-powered chatbots, on the other hand, analyze past conversation data to “present the most relevant content or experience for each and every visitor in real-time.” It works like this: If a customer has a question about a return because something doesn’t fit, the machine won’t provide a link to a return page because of a rule that has been set up. Instead, the chatbot may provide a specific answer based on past conversations in which similar customers specifically asked about “returns” and “fit.” In short, the main difference comes down to a model’s ability to understand what someone is saying and respond accordingly without help from a human. As specific as a rules-based chatbots can be, they are only as effective as a company’s ability to anticipate every user question and comment.

Why NLP based Chatbot is better?

Rules-based software can be effective if you want to achieve something you’ve seen work before on a basic level. For example, in the past, anyone who asked us X question expected Y response. ML applications require more data to train, and as a result, can make more informed decisions about what types of behavior may lead to purchase or what types of chatbot answers are most likely to resonate with a customer in real-time.NLP based chatbots make interaction between human and interface more lively.

5. Applications

5.1 State-of-the-art Applications

- Customer Service - The implementation of chatbots in customer service also involves the use of cases of collecting customer feedback.
- Media Publishing Applications - Media publishers recognize chatbots as a promising approach for engaging with the audience alongside monitoring the engagement of the audience. As a result, chatbots can help media publishers in obtaining credible and helpful insights regarding audience interests.
- Food Ordering - The commonly visible application of chatbots is evident in the case of food delivery. Notable names such as Pizza Hut and KFC use chatbots for allowing customers to place orders through a conversation.
- HealthCare Applications - The chatbot examples in the healthcare sector also showcase the breadth of the reach of chatbots. Chatbots, such as Super Izzy has been helping medical professionals in providing quick medical diagnosis and answers to health-related questions.

5.2 Challenges

- Understanding the User Intent - For any Chatbot, the biggest challenge is to understand the user intent and to decode the intent hidden inside the queries and messages.
- Building an affordable chatbot solution - Many small and medium enterprises simply skip Chatbots because of the higher cost factor.
- Chatbot Testing - As natural language processing (NLP) capability is increasingly getting better, Chatbots are now frequently updated. This is why the testing mechanism should always be used for every update to check the effects of each value addition.
- Building a more human chatbot - The last but not the least of all these challenges is building a Chatbot that makes an impression of human conversation to the user on the other end.

6. Conclusion and Future Scope

This study involves the application of Natural Language Processing to create chatbot technology using Artificial Neural Networks.To explore Most suitable approach of implementing Natural Language Processing in chatbots. According to the results, Deep Recurrent Neural Networks are most suitable for creating a chatbot based on Natural Language Processing using Sequence-to-Sequence Model Algorithm.

Implementation of Natural Language Processing in chatbot was planned by implementing Natural Language Processing steps involved ,to import Natural Language ToolKit in Python and implement sequence-to-sequence algorithm for creating a chatbot and its working Tkinter GUI.A very basic chatbot was built using a small Kaggle dataset , all the Natural Language Processing steps are implemented by importing Natural Language ToolKit. To check whether the chatbot is working efficiently using Machine Learning model on Tkinter GUI is not implemented.

Bibliography

[1] G Krishna Vamsi , Akhtar Rasool , Gaurav Hajela A Deep Neural Network Based Human to Machine Conversation Model , IEEE 11th ICCCNT 2020 July 1-3, 2020 - IIT - Kharagpur

[2] Ramakrishna Kumar, Maha Mahmoud Ali(2020) A Review on Chatbot Design and Implementation Techniques, National University of Science and Technology, Muscat, Oman , Feb 2020.

[3] Moneerh Aleedy, Hadil Shaiba, Marija Bezbradica (2019) Generating and Analyzing Chatbot Responses using Natural Language Processing,International Journal of Advanced Computer Science and Applications

[4] KULOTHUNKAN PALASUNDRAM,NURFADHLINA MOHD SHAREF, KHAIRUL AZHAR KASMIRAN,AND AZREEN AZMAN Enhancements to the Sequence-to-Sequence-Based Natural Answer Generation Models,Intelligent Computing Research Group, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, IEEE Access,February 24, 2020.