ACIT5930

MASTER'S THESIS phase III

in

Applied Computer and Information Technology (ACIT)

May 2022

Cloud-based Technologies & Operations

Smart Chatbot

Shairaz Sultan

Department of Computer Science

Faculty of Technology, Art and Design

Acknowledgment

I would like pay tribute to my Supervisor Pål Halvorsen for his guidance, support, throughout my thesis. It was not easy for me to complete this degree due to some personal problems I faced last year. But it is thrilling to remember that whenever I contacted my supervisor, the response was overwhelming and motivational, that provided me the energy to stand up and continue this journey again after being away for some time. I also want to thank Oslo Metropolitan University and the Department of Computer Science Faculty of Technology, Art and Design for granted me a chance to get higher studies.

I dedicate my thesis to my parents for their support, courage, and prayers throughout my master’s degree journey.

At last, I would like to thank all my professors, teachers, friends, and well-wishers who support me and guide me at any level or to any extent. I feel proud, that I am about to complete my degree besides having quite a hiccups during this journey of master’s degree.

Table of Contents

List of Figures

List of Tables

Abstract

Chatbots or smart conversational chatting machines are being build using Artificial Intelligence and machine learning technologies to solve the existing problems in the area of natural language processing. Thanks to the recent advancements in the 'Human-computer interaction' field where text-based chatbot frameworks are the hot topic of research for the industrial sector as well as academia, initiated the development of advanced intelligent conversational chatbot systems. These days the use of chatbot frameworks is not limited to the hard-rule (domain specific) conversational services such as ticket booking, hotel reservation, customer service assistant, or as "suggestion provider" tools. A new breed of these advance intelligent chatbots can provide much more versatility and agility to handle complex business or personal tasks quickly and more efficiently.

In Last decade, Impeccable innovations in deep learning field has forced the migration of template based chatbot frameworks to the innovative trainable, lenient end to end based ML frameworks, thanks to the introduction of recurrent encoder-decoder model [56] and recurrent neural networks (RNN). To overcome the existing limitation in rule-based or retrieval-based models during dialogue generation industry favourite is modern day Sequence to Sequence based framework with encoder-decoder based architecture relying on attention mechanism for dialogue modelling including question answer-based conversations. These seq2seq encoder-decoder based models uses RNN with Long short-term memory cells to handle the vanishing & gradient problem in the network. However, we must understand that this seq2seq based approach is widely adopted for Machine learning based tasks, but it is still for from immaculate. Training of these models is very challenging that required high processing resources and computing power and finding appropriate hyper parameters for the model optimization becomes matter of extreme importance as we experienced all these obstacles during the training of such chatbot model.

In this thesis we will provide analytical overview of chatbot technologies, how it started and how much it progressed and what approaches were adopted throughout this time span. Later we will conduct some of the comparisons with sequence-to-sequence based existing chatbot models based on different neural networks and compare the results for the audience.

CHAPTER 1. INTRODUCTION

1 INTRODUCTION

Chatbots are the software's or applications with the ability to self-generate human alike responses, based on the training provide to the base chatbot model. The idea behind the development of Chatbots were to mimic humanly conversations based on specific scenarios to reduce the human interaction in obvious situation. Initially domain-specific chatbot systems were introduced in our routinely work and engaging human interaction with these technologies to address specific tasks like customer services help, product suggestion, ticket reservations, etc. However, the developments in IT sector forced a shift in the approach of being rule-based to advance AI based generative chatbot frameworks. That resulted development of multi-dimensional AI chatbot frameworks to handle multi-tasking, Sentiment based discussion and in assisting with routinely operations, technologies like "Cortana", "Siri", "Alexa" and "google Assistant" [5] are the fine examples in this context.

Before moving to the technical part first we would like to discuss the reason and motivation behind this study. Child abuse is one of the major social concern still existed in our society, the number of cases in high-earning and developed countries are shockingly higher than expected, According to the data published by WHO, In Europe the number of reported child abuse cases every year are over 118 million, [23] which bring the severe consequences in the victim's lives, such as “low self-esteem”, “Anxiety”, “depression”, “mental disorder” and in some cases “self-harm”. These social disabilities may not look life-threatening in most of the cases, but it may cause long-lasting effects on children's lives. Unfortunately, existed systems lack the efficiency level required to handle these societal issues especially during communicating with potential abused victims, who are the most important source of the information regarding the case and prosecution. To assist the current practices, this project aims to develop a Virtual avatar “Virtual Avatars for Investigative Interviews with Children” where the purpose is to come up with the solutions to conduct productive interviews with potentially abused children to overcome the existed flaws in the current system where lack of corroborative information from the victim causes weak judicial convictions. The parent project has a wide scope and covers all possible aspects such as “Interview training”, “Chatbots”, “Text-to-speech & Speech-to-text” and “Image simulation” required to build a “Virtual Child Avatar”, but for this master thesis, specifically we will conduct a study on “how to development a smart chatbot", A chat bot system that could be integrated into the parent project and ideally assist to conduct productive interviews with victims and possibly provide assistance in police training.

More about chatbots, Due to the successful inclusion of domain-specific chatbot systems in our typical work and engaging human interaction with these technologies to address specific tasks like customer services help, product suggestion, ticket reservations, or helping customers in e-commerce related solutions, kick-started the research and development of advanced Artificial Intelligence-based conversational chatbot systems to handle daily life complex problems. [17] In the second stage, the focus was on the development of multi-dimensional chatbot frameworks to handle multi-tasking and assisting, technologies like "Cortana", "Siri", "Alexa" and chatbots like "google Assistant" are noted examples of that. [5]

Recently, the focus of researchers has been shifting from simple multi-tasking machines to the development of more advanced, smart, and something human alike machines by adding the deep awareness about the context and emotional or empathy factor will allow these machines to have a more constructive and meaningful conversation. This way the machine will be able to engage humans for a long time that will allow these technologies to learn more and get better with the time being. [5, 18]

The purpose of this study is to conduct a technical survey about the existing models and frameworks appropriate to develop an advanced chatbot systems, that could be trained for question answer-based conversation to interview potentially abused children. When interviewing with children we must consider the emotion state of the traumatized child also. There is a lot of work already has been done regarding the development of advanced chatbot machines where researcher proposed different methods to create smart conversational chatbots by combining Artificial intelligence with emotional intelligence and adding empathy factor [14, 2] or using persona-based models to identify human personality [3] to make conversations more consistent and engaging. The implementation of AIML "Artificial Intelligence Mark-up Language" is surely helping to achieve a good level of trust to make conversations long and believable. After my literature review, we came to know that most of these advanced AI chatbot frameworks are using "sequence to sequence" architecture to achieve agility, and by implying LSTM "Long Short-Term Memory" approach makes the conversations more engaging and realistic. For my master thesis, we aim to build an advanced chatbot system by using "seq2seq" architecture and Google's NMT "neural machine translation" framework. To be precise, we opted modern day Sequence to Sequence based framework with encoder-decoder based architecture relying on attention mechanism for dialogue modelling including question answer-based conversations. Our seq2seq encoder-decoder based model uses RNN with Long short-term memory cells to handle the vanishing & gradient problem in the network.

1.1 Background and Motivation

The use of Chatbots and conversational machines is increasing rapidly, we interact with these machines quite frequently. Most of these technologies we encounter are very limited in scope, as they build on hand-crafted rules to solve predefined problems. In 21st century it was the breaking point for the advance chatbot research work where not only the educational institute, but the industry and organizations started investing in it to get the potential this technology can bring to the table when it comes to customer services and reservation related operations. In 2010, the research was to make these machines smart using Advance AI and Machine learning that can help these chatbots to make decisions when needed.

All this data suggested that we could use these technologies for this project “Virtual Avatars for Investigative Interviews with Children” where the purpose is to come up with the solutions that could help to conduct productive interviews with potentially abused children to overcome the existed flaws in the current system where lack of corroborative information from the victim causes weak judicial convictions. In this project, I have worked on developing a smart chatbot system using advance seq2seq model, LSTM and recurrent neural networks that can help to build a machine that could provide a vital contribution not only in conducting interviews also in skill training for polices officers while communicating with potentially abused children. The parent project has a wide scope and covers all possible aspects such as “Interview training”, “Chatbots”, “Text-to-speech & Speech-to-text” and “Image simulation” required to build a “Virtual Child Avatar”. My motivation is that this technology has potential to contribute to the concern topic as recent studies elaborated that computer-based interactive communicative activities improve the productivity and performance of the “interrogative interviews” while interacting with children. [19]

1.2 Existing Problems

After the success of domain-specific chatbot systems and their contribution towards handling routinely operations such as customer services help, product suggestion, ticket reservations, etc led the innovation in this field and we are in the point where retrieval based and generative chatbots frameworks are being used all around the world in different aspects of the life. Last decade or so the investment and resources poured to this interesting field are uncanny. Now we have different type and genre of chatbots for example data retrieval chatbots, multi-dimensional AI chatbot frameworks to handle multi-tasking, chatting models for Sentiment based discussion and in assisting with routinely operations, technologies like "Cortana", “Replica”, “Siri", "Alexa" and "google Assistant" [5] are the few examples in this context. Now, the research is happening more about how these machines can be used as a prominent aspect of the communication, how to make these machine smart enough to the level where a person can establish a conversation with the smart AI chatbot framework with minimal assistant of the actual human and still the system can handle all the conversation productively and maintain the interest level throughout the discussion.

Right now, we have many smart chatbot models in the market to achieve the certain level of stability in the conversation with human, and the quantity of new experiments is growing day by day as the stakes are very high. Now if you research that how the smart chatbot could be built the data you find will be overwhelming and confusing. Therefore, we will conduct a detailed survey based on the research publications in last decades to provide overview to the audience that how a smart chatbot can be built and how this field has been evolving throughout last 5 decades based on limited research we conducted.

Other problem sentiment or emotion detection, that chatbot community is working to provide sentiment-based responses in the conversations as it will increase the engagement level and interest during the communication. Recent developments indicates that some level of the emotions can be embedding into the communication, but it is far from perfect. Very little improvements have been made in this aspect but as the technology is new the future looks very promising. We will provide information about one of the models that can be implemented with smart chatbot system to get batter results in the emotion-based conversation.

Another problem that exists is use of these models in interview-based communication, very little work has been done in this aspect and use of these smart chatbot systems. Our parent project aims to develop a Virtual avatar “Virtual Avatars for Investigative Interviews with Children” where the purpose is to come up with the solutions to conduct productive interviews with potentially abused children to overcome the existed flaws in the current system where lack of corroborative information from the victim causes weak judicial convictions. The parent project has a wide scope and covers all possible aspects such as “Interview training”, “Chatbots”, “Text-to-speech & Speech-to-text” and “Image simulation”. For matter of discussion, Interview training based on these sensitive issues is extra expensive as it would require professional actors and a huge amount of time and resources to train authorities which will not be a feasible solution for the long run. An alternative solution can be the use of computer-based virtual avatars to conduct interrogative interviews with potential abused victims as research indicates better performance with these technologies to gather more relevant and accurate information. [19] Unfortunately, right now there is no technological computer-based system that provides solutions to these problems, therefore in this research (parent project) is aimed to develop a "Child virtual avatar" that can be a step forward in the direction of using computerised systems in the interview conducting and training when it comes to children. Using Chatbot as a communication bridge is very important component of this research therefore this study will focus on the resources available to build such systems and we will try to find and register best frameworks and models that can help to build chatbot systems to fulfil the requirements of the parent project.

1.3 Limitations

The use of Chatbots and conversational machines is increasing rapidly, we interact with these machines quite frequently. Most of these technologies we encounter are very limited in scope, as they build on hand-crafted rules to solve predefined problems. Even though the integration of advance AI and Machine learning has made tremendous progress in the Chatbot industry, but still it is far away from being called "perfect". The new breed of these advanced chatbots is better to handle multiple tasks comparing with rule-based chatbots, but they still have many limitations when it comes to long engaging conversations, lack of consistency is a big problem we need to counter. The training process of these chatbots is a hard task especially when we lack a rich collection of datasets and making of relevant data corps requires enough resources and time. Furthermore, the use of the encoder-decoder model also comes with few consequences such as the way it handles sequences of variable length and vocabulary size, it becomes slow during handling huge vocabulary size and matching every word of the sentence out of that collection. "Emotions" always play vital role in the long engaging conversations and researchers are trying hard to achieve the level of success that might make this system reliable but still there is lot to do and one thing that we always need to remember, these are machines not humans.

1.4 Research Questions

This thesis work is a part of the main project “Virtual Child Avatar” that consists of 5 components and chatbot is one of them. The overall project has a wide domain with a wide set of research questions. For this study based on the thesis scope, we will be adopting theoretical approach to focus bellow mentioned research questions.[2]

- How chatbot frameworks evolves throughout the history?
- What are the resources (framework) available to build an advanced chatbot system?
- Can this sequence-to-sequence based chatbots models be used for question answering?

In the second phase of my project, we had conducted detailed research (Investigative research Survey) by studying hundreds of research papers related to advanced AI-based chatbots to get closer to finding the answers to the above-elaborated research questions. The results of the investigative survey provide enough data to understand the tools and technologies suited for the development of the smart chatbot framework. Furthermore, after developing a prototype, rich conversational data resources (dataset) will be needed to train the smart chatbot machine that eventually will be used to conduct real interviews with the police authorities. [2]

1.5 Research Methods

For this research study, First, we had conducted a technical research “survey” by studying the most relevant research work in the development of advance, multi-functional, and emotionally aware chatbot frameworks that allowed me to decide the most suitable architecture with the required tools to complete the job. But, to make the research work more consistent and rewarding it is vital to decide the research approach that increases the productivity of the research work.

Before starting the development of the advanced software systems (chatbot), it is important to gain a deep understanding of the technology and its scope, as having enough information and data about the subject is always key to productive research practices. There are many theoretical and experiment-based research methodologies and classifications [24] aimed to assist the research like ours in more coherent and consistent manners, such as guidelines suggested by the “Association of Computing machinery” in 1989, [25] the ACM classifications. These guidelines helped to maximize the scope of “Computer science” as a scientific field and disapproved of the perception that Computer science is only programming. The ACM divided the computing subject into three somehow co-related “paradigms” mentioned in the report “Computing as discipline” [25], (**i**) Theory, (**ii**) Abstraction, (**iii**) Design. [25, 26]

1.6 Contribution

The main contribution of this thesis is to provide a theoretical overview of the chatbot technology to the new users in this very field. As the data available is extremely humongous and overwhelmingly confusing. That cause lack of interest among many new audiences who want to pursue their research and future work in this field. For that purpose, we have conducted a technical survey from 1950 to 2022. Based on the limited research done in this thesis, we have provided an easy data to track the progress of chatbot technology from the beginning to this very day. We have divided the survey into two different time spans (early 1950 - 2000) and latest (2010-2022). Our findings are also providing the insight about the change in approach during that time where initially handwritten script-based models were used in 20th century and how latest deep learning encoder decoder-based framework replace those early models.

This study also provides the information for the audience that how they can build advance deep learning based chatbot framework by using sequence to sequence model

with neural networks. In the experiment section we did compare two Models based on the research findings. Model 1 with seq2seq based encoder decoder architecture with bidirectional LSTM formation and Model 2 with seq2seq based encoder decoder architecture with GRU neural networks. We compared these models on same settings and provided the findings in Chapter 4.

CHAPTER 2. BACKGROUND

2 BACKGROUND

2.1 Related Work & Literature Review

In recent times there is a lot of new work is being done for the development of realistic conversational chatbots which could be able to assist in complex tasks enriched with the ability to understand the core characteristics of human such as mood, feelings, or emotions to some extent to make the conversation more consistent and human alike. Here in this section, we would like to discuss some of the research literature we have studied in previous phases. Here we will start listing the research publications based on the relevant domain field of smart chatbots and later we will organize them in a meaningful context.

A paper published in 2015 by **O. Vinyals et al.** [1] where they proposed that conversational modelling is core to understand machine intelligence. But still, the problem is the given approaches and models have restricted the domain to a certain level such as the use of chatbot in customer service or hotel reservation or flight reservation, etc., and still, it required hand-made rules for functioning. To overcome the existing limitations Author used the new proposed seq2seq "Sequence to Sequence framework". Good things with this approach are it rarely depends on hand-crafted rules, and it converses the next conversation or response mostly from the data of previous sentences or responses from existed conversations. By using the seq2seq framework, the author experimented and concluded that it is possible to train an intelligent, realistic chatting bot engine. Without fully depending on the predefined hand-crafted rules. The results after the experiments were surprising. But this engine had limitations and lack some level of consistency with the outcomes which cause failure in the Turing test. [23]

'Sentiment' component for the success of any conversation is vital especially when we are talking about the fact where machines could communicate with human **H. Zhou et al.** [2]. Emotions play a vital role in any dialog-based conversation or discussion. Problem is that it is not studied on a big scale when it comes to conversational or chat-based systems or bots. Here in this paper author proposed an "ECM" 'Emotional chatting machine'. [2] To get this system working. Researchers used 3 new mechanisms for the development of ECM.

- Analysing the expression of emotional conversation and create categories.
- Record the fluctuation & change in an internal emotional state.
- The vocabulary which used to show explicit external emotions. (Vocabulary for emotion expression)

After experiment analysis, the results were encouraging, the 'ECM' system was able to respond to the questions more consistently and structurally, not only content-wise but also the response was showing the factor of emotion by using the wording from the vocabulary.

A human persona-based evaluation method proposed by **Jiwei Li et al.** [3] is very interesting in this regard. This model is supposed to encode the participant's persona and try to figure the characteristics of the person such as his speaking signature and person's psyche which reflect in his conversation style. Human judges were presented to oversee the results of the experiments. This persona-based model was supposed to overcome the problems which the data-driven model had such as consistency in the results. After the research and experiments, this approach proved more consistent than seq2seq based models but still the results were not extraordinary. As the scope of the paper was not handling the other characteristics of human nature just like emotions, psyche, and mood. Still, results were consistent, but the author suggested future work to make this model ideal.

'Emotionally Realistic Chatbot Framework to Enhance Its Believability with AIML and Information States' was step forward to counter the inconsistency flaws of earlier experimental approaches **Rhio Sutoyoa et al.** [4]. Where they tried to enhance the trust or believability of the current conversational chatbot system. Authors proposed a chatbot system which supposed to be emotionally realistic in conversation with the human. The purpose is to make this framework believable during interaction by using AIML "Artificial Intelligence Mark-up Language" and Information state. The author believes that adding an emotional component to the framework will allow this system to be more realistic & up to some extent will make it to understand the natural conversation of the participant in a better way that will maintain the level of consistency, which will help to make more realistic and engaging responses to the conversation. The authors experimented with 2 groups of students with different chatbot frameworks. In the end, the Authors were happy with their results they believe that student's interaction with the emotional chatbot was more engaging and realistic. The results of the participants with the "emotional model" based chatbot framework were more promising than the simple responsive model.

A survey based on ‘Emotionally aware Chabot’ conducted by **Endang W. Pamungkas** [5] was very important to my research field, as for the parent project of ‘Child-Avatar’ the Chabot system must be compatible with some level of emotion handling or processing. In this survey, the Author provided a systematic overview of approaches to building EAC "Emotional aware Chabot", a realistic conversation or dialogue-based framework. In this paper, the author finds that in the beginning mostly Chabot was created on pre-defined "Simple rule-based” approaches but now due to advancement in this technology and demand in the market caused a change and now the approach has been shifted to "neural-based" systems. They also predicted that in the future the work on these emotion-based bots will be increased.

The author tried to answer the existed questions regarding the development of intelligent Chatbots and later elaborated on the fact that now most of the emotion-aware bots are being developed on a neural-based approach using emotion classifiers. Modern Chabot’s are being developed using the seq2seq approach usually based on encoding-decoding architecture. [5]

As it is understood that a productive and engaged conversation require consistency and level of trust that provide a based where the parties rely on to give the opinion.

Przegalinska et al. [6] conducted a research study on chatbot technology where they suggested that ‘trust’ is a focal point in the success of realistic, persona-based conversational bots. Authors proposed a new method ‘Novel framework’ that will be able to keep track of human interaction with chatbots systems and it will be able to measure the results of the conversation and ethical performance of the chatting robots such as trust or believability. The author used neuroscientific techniques for this purpose, such as ML and deep learning, the author also believes that for the success of these chatbots it is important to have a great element of trust during the conversation between participants and the machine.

During the experiment, the author used two types of studies for the human-computer interaction, a subjective questionnaire, and secondly to get a broad perspective with psychophysiology, the objective analysis. [6]

Majority of recent developments in advance AI chatbots used NMT (Neural machine translation) model by google that provided the encoder decoder framework. Research on the course to develop ‘Intelligent Chatbot using Deep Learning’ **Anjana Tiha** [7] where the team worked to develop an intelligent chatbot system using (NMT). "Neural machine translation", NMT is a model that is developed by Google and based on the seq2seq approach. The architecture of NMT is based on the encoder-decoder framework. To track the results and performance the author used a neural attention mechanism. Hereby using a deep neural network approach, the purpose of the experiment was to generate consistent and coherent dialogs that could help in the development of intelligent and realistic conversational chatbots. To conclude, the author believes that the results of the tests could be more realistic given the rich database of a dataset, or real-life conversational data. [7] By studying this research paper we also get familiar with other approaches towards the creation of Intelligent and realistic conversational frameworks which could help to communicate with potentially abused children, secondly, the same technology could be used to train authorities for interrogative interviews.

As our Parent project is about building a virtual avatar that may help us to conduct interviews with potentially abused children and then using that data to provide training to the relevant authorities. The research work by Shriniket Yakkundi et al. [8] to develop chatbots to conduct interactive interviews. certainly, it is the work of our interest as it is near to our research work, where they used the "Natural language processing" approach and Google's "text-to-speech" technology to create advanced chatbot system that will be able to conduct interviews of the participants and based on performance and later evaluation, the chatbot framework will be able to short-list the more suited candidates. NLP enables the machine to analyse and figure out the response of the candidate during the interview and generate the dialogues that will be near to human speech. The purpose of this exercise is to build a framework that can replace the current interview system of the organizations. For an experiment, the author used participant credentials (required details) and stored in a database (knowledgebase), later, the basis of fed data chatbots will be able to conduct a human-like interview, after the completion of the process, the chatbot system will analyse the results and provide a list of most suitable candidates. The author used a "spacy framework" to extract keywords from the conversation.

The author believes that currently used interviewing methods are lengthy and expensive and not 100% transparent. By using advanced AI-based chatbots organizations certainly overcome the flaws of the existed orthodox interviewing system. More to that it will help them to save a lot of money with a more quick and transparent hiring system.

The research work on the development of chatbot systems with psychologist attributes by Miriam Romero et al. [09]. Was highly acclaimed. Thanks to the successful inclusion of chatbot technology into organizations, the research work on the development of advance and realistic chatbot systems have been increased recently. The author believes that the first-ever created conversational bot was developed on psychology fundamentals and presented as a psychotherapist. Here in this article, the author emphasized a structure to create a chatbot system with the functionalities of psychological analysis. [9] In our child-avatar project, the purpose of the chatbot framework is to assist during the interview of the potentially abused child, that means the child is in a state where the normal chatbot system will not be able to assess the possible psychological state of the children, by adding this functionality to existing 'emotional' and realistic chatbot system will be plus.

Behaviour change is common phenomena when the conversation is diverse, A case-study conducted from 25 individuals by Iwan Gulenko [10], to gather data about the common problems they face when it comes to IT-related security. "Motivational interviewing" technique was used for a question as MI allows you to educate and train people when it comes to a behaviour change. AIML (AI mark-up language) was used for the creation of chatbot systems. Chatbots are developed to handle three different sections including "concern about passwords", privacy security, & safe browsing.

Microsoft China, published a paper[14] about the development of "XiaoIce" chatbot, a project recently developed by "Microsoft China". XiaoIce is the world's most popular "Empathetic Social Chatbot" designed by combining the Artificial Intelligence technology with "Emotion Intelligence" that enabled this machine to provide a human alike emotional, sentimental, and social communicational experience. By implementing IQ and EQ (Emotion quotation) in the design of the system with MDPs "Markov Decision process" facilitated this system to analyse and process long-term communications with the decision-making ability while responding.

Chatbot technology is widely being integrated into almost every segment of the life. Education segment holds the outmost priority as the communication between students and teachers is always very important, but time constraint is always there. A research work done by Kulothunkan [78] is very interesting where the authors provided a use case of developing a chatbot with seq2seq based RNN and trained that chatbot based on seen and unseen data types to be able to provide solution when it comes to questions in a specific domain.

ConvAI or conversational AI is hot topic for many researchers in last couple of years. Even though the commercial aspect is astonishing, but the actual work done is not extensive. JD AI research paper [75] on the use of Conversational Artificial Intelligence is very interesting as they registered the work has been done and needed to be done in this aspect. They claim that using ConvAI can be used in social impact related services.

2.2 General Limitation found in Chatbots

Most chatbot technologies we encounter are very limited in scope, build on hand-crafted rules to solve predefined problems. Here are a few limitations with existed technologies.

- Lack of persistent engagement
- Lack of consistency
- Limited dialogue generation abilities
- Lack of emotional/empathy factor
- Lack of rich datasets

These days we often encounter the use of chatbot systems in our routine communication where the main purpose of these machines is to help and provide the basic services or results to existed problems. These systems mostly build on pre-defined hand-crafted rules where the scope, functionality, and domain of the work are restricted to a certain level such as the use of these machines to handle tasks like “booking and reservation”, providing solutions to “FAQs” such as auto-response generation in customer support services. These rule-based approaches are good to handle specific tasks but unable to solve more advanced and realistic problems, for example, communication with actual humans and then the response, depending on human nature like intention, emotions, psyche, or mood.

2.3 Investigative Research Survey

For a deeper understanding of chatbot technology and how this technology transformed with passing time, we decided to conduct a technical survey where we will try to gather all the respective milestones in the field of chatbot technology, starting from the first development attempt of a chatbot to recent futuristic Artificial intelligent chatting machines. It was very important aspect of this study as it will allow us to gather the technologies and methods used by the researcher to find the solution of the problem that we faced when it comes to Intelligent Chatbot conversation.

2.3.1 Initial Research Approaches

Year	Author	Topic	Purpose/ Focus	Approach
1950	A. M. Turing [23]	Turing, A. M. Computing machinery and intelligence	To analyse the ability of machines to communicate with a human being.	An experimental approach to identify, can machine think like humans? In an "imitation game", the machine must win to pass the test. [29]
1966	Joseph Weizenbaum [30]	ALIZA – A program communication between machine and human	The first chatbot used "pattern matching" to simulate the conversation.	1st proper chatbot, rule-based, used pre-written scripts to process user inputs and commands for communication.
1973	Kenneth Mark Colby [31]	Parry: The Doctor (Psychiatry) "Aliza with Attitude"	1st attempt to use a machine to counter patient with "Schizophrenia" paranoia and learn from them.	By using a model based on the behaviour of a paranoid person using the base concept of disease.
1988	Rollo Carpenter	"Jabberwacky" chatbot with humor [32]	It was designed to create Chatbot using "Natural human conversation" for entertainment purposes. One of those programs that	It was 1st attempt to create an AI framework based on HCI. It was able to mimic human interactions and then making engagement conversation.

			passed the "Turing test".	
1995	Richard S. Wallace [33]	A.L.I.C.E (Artificial Linguistic Internet Computer Entity)	Intelligent machine to have a meaningful conversation with human-based on heuristic pattern matching system.	An AI natural language bot developed based on experiments "Alan Turing" described in 1950. It uses supervised learning where the botmaster plays a vital role in an appropriate response.
2001	Active buddy Inc. [34]	Smart Child	AI-based Chabot, developed in 2000, was available on AOL & MSN messenger to have a fun textual conversation.	By using "Natural language comprehension functionality" into the AI algorithms to increase the conversation engagement for AIM & MSN messenger users.

Table 2.1: Summary of research work on smart chatbots (1950-2000)

In Table 2.1 we tried to gather all the achievements and research work in the mid of twenty century that led the idea to build a machine with the ability to behave like human beings. To start with, in 1950 the paper published by the A.M. Turing became the starting point for the discussion of building machines that can somehow relate to human thinking, the test usually called "Turing Game/test" [23], where the author suggested rules and guidelines that machine must follow and pass to be recognized as a "thinking" machine. In 1966 "Joseph W." created a program that called "ALIZA" that was the first machine code that can have a conversation with a human by using "Pattern matching" techniques where the responses were based on the pre-installed scripts and instructions, "ALIZA" was the product that indulged many Computer researchers to start working on this fascinating technology, "Kenneth Colby" was one them who was inspired by "ALIZA" program, In 1973-1975 he developed the first Chatbot that can learn from the conversation and based-on that learning it was able to behave like a "Paranoid" person of a "Schizophrenia", it was also one of the finest attempt to create a perception that machines and computers can work in the health sector. This Chatbot (Perry) [31] aka "ALIZA with Attitude" could be considered the first machine that was based on the "emotion" factor. Later, "A.L.I.C.E" [33] by "Richard S, Wallace" was the breakthrough when it won the "Loebner Prize" for human alike, talking Chatbot three times in 2000, 2001, and 2003. At the point, people started considering the potential and financial aspect of the "Chatting bots" that later stormed the e-commerce industry after decade.

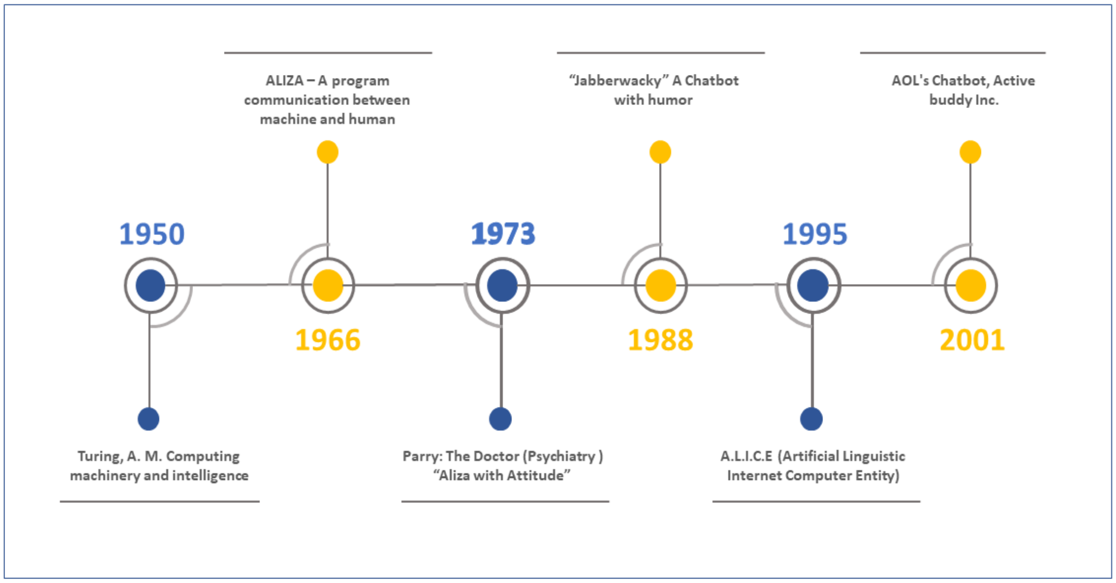

ALIZA – A program communication between machine and human
"Jabberwacky" A Chatbot with humor
AOL's Chatbot, Active buddy Inc.
1950
1973
1995
1966
1988
2001
Turing, A. M. Computing machinery and intelligence
Parry: The Doctor (Psychiatry)
"Aliza with Attitude"
A.L.I.C.E (Artificial Linguistic Internet Computer Entity)

Figure 2.1: Illustration of Table 2.2.1

The above data (Table 2.2.1, Figure 2.2.1) illustrate that it took decades of research and hard work to reach the level where the A.L.I.C.E was formed and organizations and companies felt the industrial potential of these Chatbot technologies, initially it started with the idea that human being can communicate with the computers and machines, that later evolves to the level where the conversation became more structured when machine learning techniques came and the computer started learning from the data of the communication where machines can decide the response, not the predefined rule-based replies.

In the previous section, we have tried to figure out how the idea of human-computer interaction started in the first place, the Turing test (Turing Paper) in 1950, [23], and later development of Perry "the doctor" [31] in 1975 kick-started the idea about the development of emotionally aware machines. We already discussed how technology, ideas, and perceptions regarding

Human-computer interaction & machine intelligence evolved in that time duration. After studying the fascinating research work that started in the previous century (1950 - 2000), now fast forward to recent times, in "next section" we will analyse the recent developments in the field of Human-Computer Interactions and conversation machine frameworks, especially due to the integration of advanced AI and Machine learning that has stormed the IT industry and now has become the hard topic of discussion.

Unlike the initial stage of research and development about Emotionally aware machine that was mainly using the rule-based architectures, now the paradigm has been shifted to more advanced neural-based approaches, this change in approach has been highly beneficial in the development of more interactive and engaging conversational chatting frameworks. In table 2, we focused on more advanced and versatile chatting frameworks that are being developed and used in recent years. According to my findings, from 2000 to onwards 2010, 2011 more of the research resources were used to build interactive multifunctional Chatbot machines mainly focused to assist in industrial use such as “Customer services”, “Customer care”, “Booking & scheduling” etc. After 2015 the research work from academia and industry has been tilted towards the integration of human emotions into the development of interactive Chatbot frameworks. Inclusion of human characters such as emotions, feelings, personality certainly can change the way we investigate these technologies, it can stretch the existed boundaries and limitations and provide new technological dimensions to the human-computer interaction domain. The research about integration of “emotion factor” into the Chatbots stared in the ’70s, but now with technological advancement in recent years provides a new meaning to it, especially in 2017 when “Minlie Huang” and co, published the first shared task about “Emotion generation challenge” [48], where 20 teams work on this challenge and 10 of them submitted the end-results and evaluation. After that, the research on “Emotional Chatbot” frameworks are a topic of interest for high-end organization and academia, “XiaoIce” [5] from Microsoft is a fine example in this regard.

2.3.2 Modern Research Approaches

Year	Author	Topic	Purpose/ Focus	Approach
2010	Mavridis, Nikolaos et. al. [35]	Human-Like memory for social Agents	Using chatbots socially would require a set of desiderata for the memory systems. Importance of shareable memory framework to understand the scenarios, feelings, situations, and natural language for common use.	A case study of the memory handling subsystems of two famous Chabot frameworks of that time "Ripley Chabot" [36] and "FaceBot" [37]. How these memory subsystems can be improved for a more meaningful relationship btw human & Bots.
2011	Ana Paiva [38]	Empathy in Social Agents	The author conducted a study on the empathy factor of social bots, and how this should be inspired by human social behaviors in society, further discussed non-verbal behaviors, emotions, etc.	Testing of 4 social bots under an empathy-based framework that handles the empathy factor as a dynamic process. This Framework consists of 2 stages "Emphatic Appraising" & "Emphatic Response" to make it effective. [39]

2014	Li Zhou, et. al. [2]	An emotional Chatting Machine	Authors worked on the development of empathy factor for Chabot's that could be used for conducting human alike social instructions.	In this paper author proposed the seq-2-seq model with GRU "Gated recurrent Unit" & "Recurrent Neural Network". They also took "Emotional Quotient" & IQ as a vital factor for empathy building.
2015	Oriol Vinyals et. al. [1]	A Neural Conversation Model	Conversational models are core to understand the level of machine intelligence but existed approaches have limitations due to its dependency on hand-crafted rules.	Seq to Seq framework allowed the versatility where a machine can learn by its previous engagement with humans and relying on the data, the machine can conduct an engaging conversation with other users.
2016	Jiwei Li et. al. [3]	A persona-based neural conversation model	A method based on human personality. This technique is supposed to encode a person's personality & characteristics, psyche during the conversation. Aim to overcome the flaw of the data-driven model.	To address the consistency flaws that come with the data-driven model like the seq2seq model where breakout in conversation was the norm. Proposed a persona-based conversational model on the baseline seq2seq model.

2017	R. Zhang, Z. Wang, et. al. [40]	Building emotional conversation systems using multi-task Seq-2-Seq framework	The author focused on the development of the emotional conversation system that could be able to generate "Viable responses" against Every type of emotion based on post sequences.	Creation of responses on every possible Emotional category was done by using Multiple task sequence to sequence framework along with GRU networks, Emotion factor was entertained as an Emotional Classifier.
2018	Endang Wahyu Pamungkas [5]	Emotionally Aware Chabot: A Survey	A systematic overview of approaches to building "Emotional aware catboats". The author discussed rule-based approaches and the "neural approach" to develop an interactive emotional machine.	A survey where the author researched and find that majority of emotionally aware systems are being developed on "Neural based" approaches such as the seq2seq model having "Emotion" as a classifier.
2018	Anjana Tiha [7]	Intelligent Chabot using Deep Learning	Using deep learning practices to develop an Intelligent machine and later to evaluate the performance of the machine using a performance tracking model.	The author worked on Google's "Neural machine translation" based on the seq2seq approach. It uses RNN with bi-LSTM cells. To track the results and performance the author used a neural attention mechanism.

2018	Xiao Sun et al. [41]	Emotional Human-Machine Conversation Generation Based on SeqGAN	How can a robot understand human emotions for productive conversation? The authors used “emotional tags” to the dataset that would allow the machine to understand the emotion factor & dimension from the user’s conversation.	The author used 3 types of inputs on a model based on Seq2Seq, encoder-decoder Arc & LSTM. 1: Input sequence with no Emotion factor. 2: A sequence-based on emotion category for input activity, 3: Input sequence with emotion factor for Output generation.
2019	Nurul Lubis et. al. [42]	Positive Emotion Elicitation in Chat-Based Dialogue Systems	An effective dialogue generation framework with the ability to mimic the emotion-based interactions with the users can overcome the current deficit (emotional benefit) in the existing conversational frameworks.	A response retrieval approach that only entertains the +ve emotion segment, using the seq2seq framework. To maintain the emotion-based approach they used an emotional encoder that was trained on the entire dataset but targeting only positive emotional response.
2021	Jovanovic M et al. [74]	Chatbots as Conversational Healthcare Services	Using Chatbot tech to advance Health and care system such as diagnostics and treatment.	Adoption of

2022	Peng Qi et. al. [75]	Conversational AI Systems for Social Good	ConvAI technique for Social Good: Opportunities and Challenges	ConvAI for the social communication using its user interface LUI and auto speech recognition and speech synthesis.

Table 2.2: Summary of research work in smart chatbot development (2010-2022)

Relying on the data from Table 2.2, we can say that most of the research work on emotionally aware Chatbot is based on neural-based approaches, where the advanced chatting machines are being developed on the sequence-to-sequence framework. This encoding-decoding-based framework can learn from a high amount of data and can generate meaningful and appropriate responses as compared to rule-based practices. Seq-2-Seq model is more like an end-to-end sequence-based model where encoder-decoder architecture relying on RNN "Recurrent Neural Network", can handle tasks such as image captioning, human alike dialogue creation, machine translation, interrogative conversation, and machine translation. Sequence to Sequence model used the "Long Short-Term Memory" (LSTM) networks that use RNN with the ability to learn long-term dependencies and provides the ability to generate responses based on those dependencies. [7, 12]

The other thing we understand from this technical data overview, most of the researchers are working on an "emotion classifier". We usually train our machine learning models with huge datasets to get better results, normally these datasets contain different types of data that has a unique meaning but adding an emotion classifier to the model can provide luxury to assess the dataset and recognize the text based on emotions, this technique can improve the quality of meaningful response generation.

As we have found enough evidence that sequence to sequence model is a way to go if we want to build an intelligent Chatbot with emotion factor, but there is more to the equation, many researchers created different models and tools that can be used on base "sequence to sequence" model to meet their desired state and improve the intelligence level of these emotionally aware machines. Techniques such as persona-based seq-2-seq framework, or GRU "Gated recurrent Unit" based on the seq-2-seq framework are the examples we studied in table 2 where authors anticipated and developed suitable models to meet their requirements based on seq-2-seq architecture.

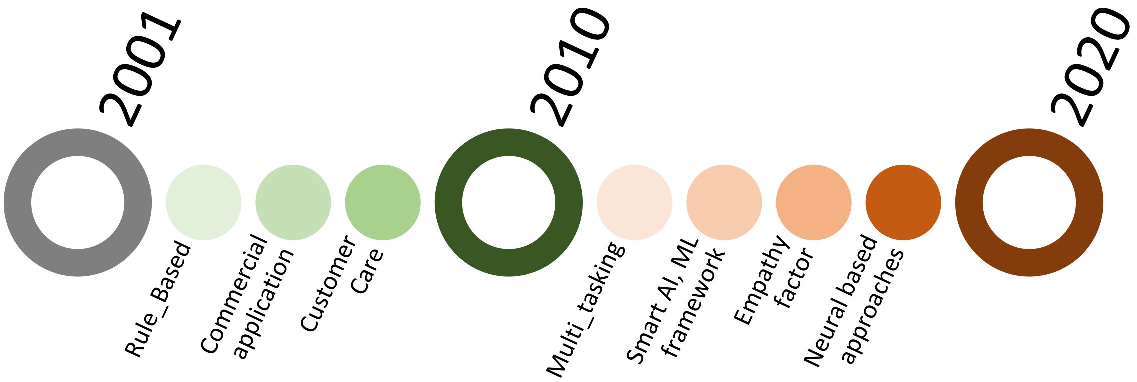


Figure 2.2: Illustration of Table 2.1

Figure 2.2: Illustration of Table 2. is a visual illustration of Chatbot technology in the last two decades, it provides a simple idea that how technology evolved and how to research topics and priorities for industry and academia changed from simple domain-based chatbots to advance neural architecture based intelligent empathetic chatbots.

By analysing all these resources from above tables, we are in state to understand that what technologies and resources (Including models and frameworks) were used by different researchers to achieve better results when it comes to development of advance chatting machines. In below we will try to write briefly about some of these techniques to provide the basic idea to our users.

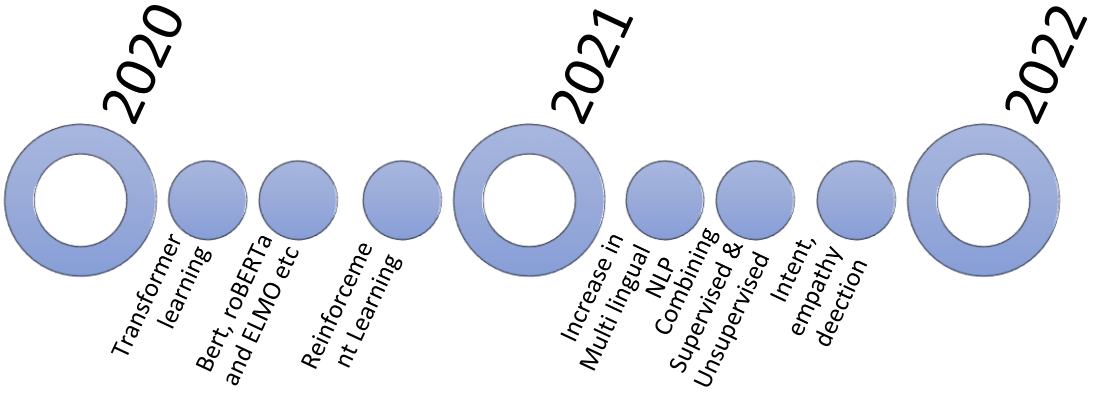

2020
Transformer learning
Bert, roBERTa and ELMO etc
Reinforcement Learning
2021
Increase in Multi lingual NLP
Combining Supervised & Unsupervised
Intent, empathy deection
2022

Figure 2.3: Latest NLP trend

Above figure is illustration of the latest NLP trends in last couple of years, as we all know how the advancements in NLP brought new exciting era of machine learning field. Based on the data we collected from our technical research survey we can easily draw the perception that how the NLP is being perceived in recent years.

2.4 Machine Learning

Machine Learning is field of computer science that has revolutionize the industry as well as our routinely aspects in a way that we all get benefited. Machine learning is about developing the data driven application and systems that are capable of learning things by getting information's from the previously available data or data based forecast mechanisms. Machine learning has evolved a lot in these last years to the level where almost every technical field has some part of it, fields such as natural language processing. Biology, computational biology, physics, computer vision, robotics all today heavily rely on the advances of the machine learning.

The initial concept was originated in early 1950, where the first time the idea was established that the machine could be taught, and they can learn [53] by depending on the network system called neural network. In recent times, the technology has been at the level where complex, yet efficient algorithms can be developed that not only learn the basic but also can learn, understand, communicate and address emotion or sentiment-based communications. The latest example of the use of advance machine learning application is auto-driven vehicles where machine learning and its other component has provided enough accuracy to build trust among developer and users.

Furthermore, Machine learning has two main categories, Supervised learning and unsupervised learning. Supervise learning is the method where we rely on labelled data. The purpose is to feed properly labelled data that will be used for learning based on the provided labelling. In supervised learning the goal is to provide the best appropriate guess between provided input and output. [54] Un-supervised Learning is opposite to the supervised learning; we do not have labelled data to feed out algorithm instead it highly relies on the formation of given data like understanding the patterns in data that could be related or finding groups with similar attributes that can be a part of same category. [54] there are also other categories available such as reinforcement learning or semi-supervised learning etc., but those are considered out of scope in this context.

2.5 Natural Machine Translation

Initially the NMT model or The Neural Machine Translation was developed to mitigate the gaps and vulnerabilities of the existed phrase-based translation approaches. IT was effective in execution and results, but it was a very expensive system for deployment as it required more resources such as finance and very large datasets for training and then building interfaces for translation. NMT model uses the vector representation in the process of word-based translation, therefore this model was better than the previous phrase-based translation model and it has single representation model for both phrase-based translation as well as recording model. During the translation it uses the bidirectional recurrent neural network also called encoder that perform encoding of the given input where the second RNN used to perform decoding. Later another NMT model was introduced by google "Google's Neural Machine Translation system" based on deep "Long Short-Term Memory" LSTM along with 8 encoders and decoders. GNMT model is aimed to provide much-needed robustness and versatility where it requires less training time, and due to the use of the "Low-Precision "approach while computation the translation speed will be robust. [7, 13, 63]

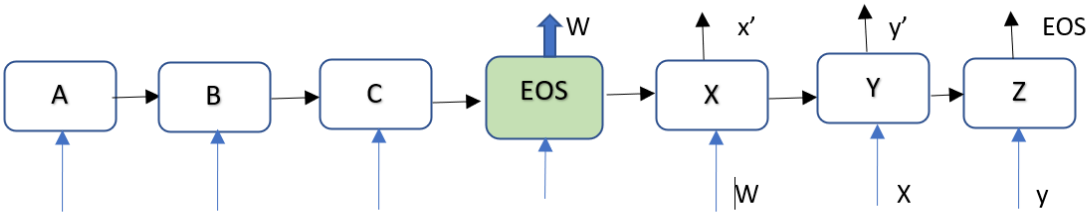


Figure 2.4: NMT model with input sequence. [63]

Above figure is the illustration of the NMT model by (Sutskever et al.,) [63] where the input sentence A, B, C fed to the simple NMT model, that produces the output sentence

"WXYZ", but one thing to notice is that the prediction process will continue until it reaches to the End of the state "EOS". To reach to the "EOS" state, LSTM will read the input in the reverse order that will help to identify the short-term dependencies in the process that eventually helps in optimization.

2.6 Neural Networks

Neural Networks form the base of a deep learning technology where the developed machine learning algorithms are relied on the structure of how human brain functions. Neural networks take the given data and base on the information on the data these networks train themselves to be able to recognize the formation and patterns in the data and predict the outcomes. These networks also referred as "feedforward" networks that based on supervised learning, uses non-linear activation technique to handle complex pattern to make decision. [55]

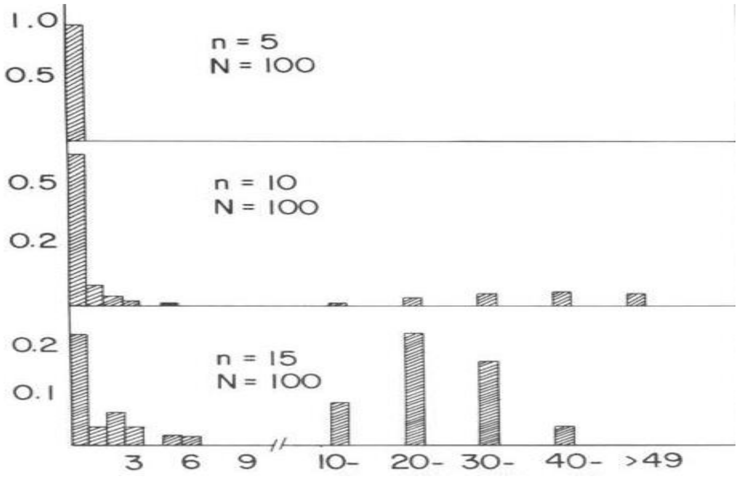


Figure 2.5: The occurrence of errors based on the probability in stable state.

Source: https://www.pnas.org/doi/epdf/10.1073/pnas.79.8.2554 [55]

In the above figure by J. Hopfield [55], the x-axis is the representation of the actual number of errors, and the y-axis is the probability. In simple words it is representing the occurrence of errors based on the probability in stable state from previous memories.

2.7 Encoder Decoder Framework

Other thing we studied during the technical survey phase is more about the adoption of encoder-decoder model in recent times. As many researchers opted for encoder and decoder technique for the purpose of making the sentence generation more constructive and coherent. The biggest difference that has the impact to change old rule-based and simple neural network-based approaches to towards the development of advance AI based smart chatbot models is the introduction of advance algorithms that are smart enough to learn from the previous utilization. Models such as Encoder-decoder has the ability to provide the futuristic results in experiments, an encoder recurrent neural network in the model provides the vector representation of the input sentence, and decoder Recurrent neural network generates the required targeted sentence formation.

Sequence to Sequence model is more like an end-to-end sequence-based model using encoder-decoder architecture relying on RNN “Recurrent Neural Network”, specifically used to handle tasks such as image captioning, human alike dialogue creation, machine translation, interrogative conversation, and machine translation. [7] Sequence to Sequence model used the “Long Short-Term Memory” (LSTM) system to make a fixed input sequence in a vector where the vector’s dimensionality is also rigid, for the decoding another deep Long Short-Term Memory used to get output from that vector as a target sequence. We will be discussing in detailed about the sequence-to-sequence model using encoder-decoder technique in Methodology section. [12]

2.8 NLP Language Models

Natural language processing was the breakthrough that led the deep dive in machine learning research as NLP provides the platform that enables the machines and systems to mimic human characteristics like reading, understanding and responding based on the given scenario. Initial work started by training the machines using hard quoted rules or scripts and supervised training but with the passage of time NLP field has grown tremendously by enabling machines in predicting, sentiment analysis and speech recognition. These days we find large number of apps that uses the NLP building blocks but for the success we must have decent language models to support the sophisticated level of language translation. One must remember that creating new language model could be a risky and expensive task, thanks to the researchers work the pre-trained models were developed that allow them to use same models on different datasets for different purposes.

There are many pre-trained Language models exists in the market that serves the purposes required for different NLP problems, notable are BERT, roBERTa, Open AI GPT2, Open AI GPT3, DeBERTa and ALBERT, XLNet etc. Let’s discuss some of the widely used models in this section.

2.8.1 Bidirectional Encoder Representation (BERT)

Bert is a new technology and many of the new researcher are very excited about the use of this new machine learning technology. Bert, A Bidirectional Encoder Representation from the transformers, a technique that was developed by the Mr. Jacob li and his associates from the Google in 2018-2019. It is transformer-oriented Machine Learning technique that uses Natural Language Processing. Bert model was highly accepted after its presentation because of the new prospect unlike previously adopted machine learning methods it was aimed to organize, prepare and train the unlabelled text of the language. Moreover, it was jointly conditioned all the layers available, and it takes both sides of the context (left and right side) into considerations. [52]

To understand the BERT model, first we must have basic understanding about the concept behind its creating. Let’s begin with the Transformer model, In Transformer model the purpose was to provide solution for the problems occurring when it comes to a translation of languages. Each language has different way of formation, different context and when we translate the text from one language to the other it lacks the actual sense behind the meaning of the context. LSTM model was used to provide the solution of language translation, but it had its own limitations, it was bit slow and hard to train in many cases because it takes the words sequentially that means word by word translation. That makes it hard to train and time constraint is the basic concern here if the data is quite large. LSTM models learn sequentially, left to right or right to left, same effect can be seen during the output process word by word.

When it comes to transformers, they provide some luxury in this regard as they are fast to learn and core to directional approach.

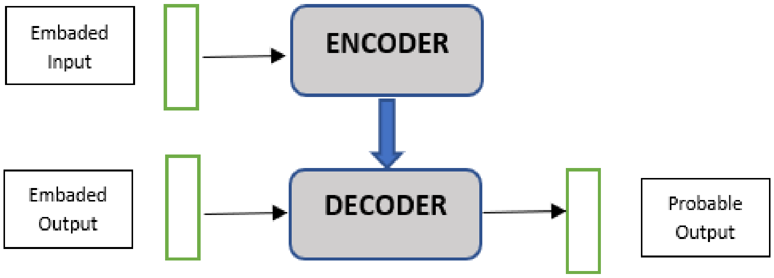


Figure 2.6: Transformers basic working model

In Figure 2.6 we can see that the embedded output is passed to the encoder, encoder takes the input, each word of the text simultaneously. These inputs tend to store in vectors and each vector or word on each vector has the different value but the words with the similar meanings have values close to each other. Therefore, the output becomes more generic and formal as it shows meaningful context even after generating the output simultaneously. This method also helps the encoder to learn with the passage of time and eventually it gets better and provide more meaningful outcomes.

In case of Bert, it goes one step further and it exclude the decoder and replace it with one encoder. As the main objective of the Bert model was to overcome the flaws in previous models by providing solution to these aspects.

- Natural Machine Translation
- Argumenta Debate
- Question and answer-based discussion
- Sentiment based analysis

Bert Model is based on two parts,

1. Pre train
2. Fine Tuning

In first part it is all about to provide the basic training about the understanding of the language and its context. To learn the basic about the language this model uses to un-supervised tasks (MLM and NSP) at the same time. MLM is “Masked language Model” and NSP is “Next sentence prediction”. Pre training is very important part because it enables the model to get to the point where it already understands the context and formalities of the natural language. In part 2 it is all about finding the best solution. [52]

2.8.2 RoBERTa Approach

We already discussed the BERT model. Bert is bidirectional Encoder representation from transformers, a technique that was developed by Google that uses transformers and neural networks heavy relies on self-attention mechanism for language translation [52]. RoBerta is in a way more refined and improved version of BERT, it builds on on masking strategy of BERT that allow the system to learn and get trained also on the hidden state of the text.

In simple words, it is better version of the Bert if we consider the results of this model comparing with Bert initial model as it changes the existing hyperparameters of the Bert.

In Bert model training was done on big mini batches but removing the objectives of next sentence probability and training on longer sequences makes Roberta model more sophisticated. Roberta was proposed as a “Robust optimized Bert approach” and after the defined changes in the actual model Roberta outperformed original Bert model by distance. [71]

2.8.3 ALBERT

ALBERT is another pre-trained NLP language model presented by Google as a lite version of BERT. In BERT section we briefly discussed about the data size of pre-trained language models that how the large data for training can give more prominent results in the long run but keeping all the aspects in mind it has some drawback we might consider important. When we use large size data in the pre-trained model in increases the actual size of model that cause slowness in the processing time, consume more resources and computational power. To counter those drawbacks in actual Bert model google introduced the ALBERT model.

The ALBERT model had two parameters, 1: Factorized embedding parameterization & cross layer sharing (to control the increase when the network grows). This is a very fine parameter reduction method, and it eventually helps to encounter the problem of large size language processing models and the computation resources require to deal with them. [72]

2.8.4 Embedding from Language Model (ELMo)

Elmo (embedding from language models) is state of the art language model developed by Allen institute of Artificial Intelligence. [70] where researcher worked on, they focused on the language modelling and proposed new representations for the pretrained language tasks called ELMo. The ELMo model starts with the pretrained well organized neural network language model based on the previously available language models. So the language model calculate the probability of the next word occurrence based on the previous execution in similar state. For example if there was sentence that states “my feet like soft shoes” and there is repetition “My feet like soft ‘ ’ ” , it will probably take shoes as a right feet instead bat or cat etc.

But with ELMo, this model will use two-layer bidirectional long, short term memory layer model and it contextualized the representation of the words, complexity in the context of word usage as well as model polysemy (linguistic context of word representation).

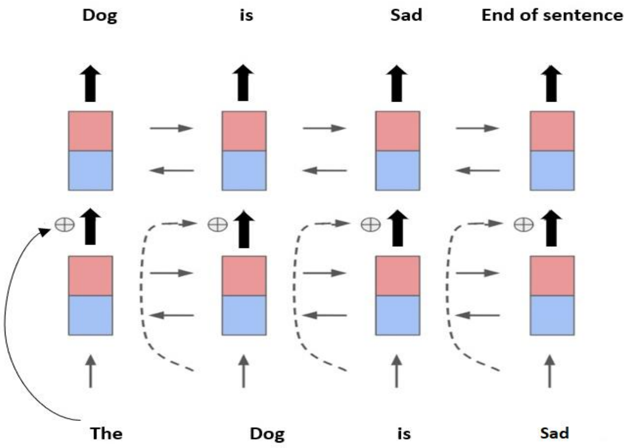


Figure 2.7: ELMo: Two-layer LSTM.

Source: https://www.mihaileric.com/posts/deep-contextualized-word-representations-elmo/

The above figure illustrates existing way of word embedding by neural language-based frameworks, the input is “the dog is sad” is divided into tokens and converted into a fixed length before being used as an input to a recurrent unit thanks to pre-trained embedding like Glove etc. But this is different when we deal the same setup using ELMo.

Elmo language model handle this situation more critically, instead of taking notice to the embedding word matrix like in above setup, it uses “Character embedding” by converting every vector into appropriate character embedding by the help of conventional layer. After that it runs through multi-layered (2-layer) highway networks and eventually presented to the LSTM layer as an input entity. [70]

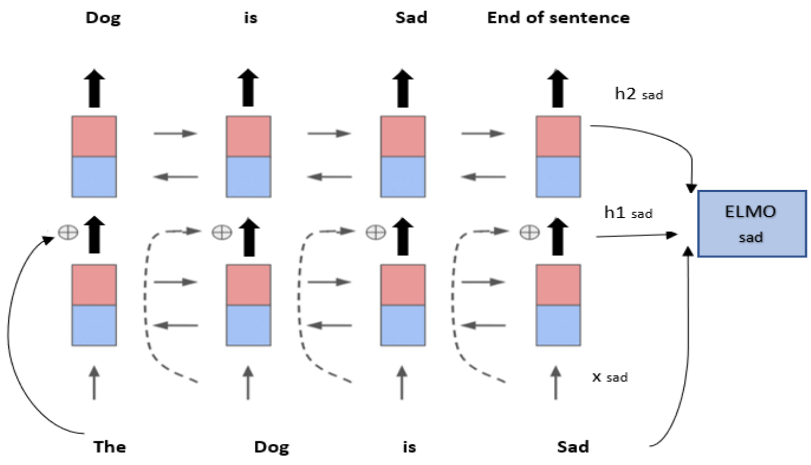


Figure 2.8: ELMo specific representation of "Sad" word

The Elmo language model became very successful as the experiments shows that it outperforms almost every NLP benchmarks.

2.8.5 XLNet

XLNet is Auto aggressive pretraining model that outperformed existing BERT language model in almost 20 tasks. In BERT (Auto-encoder AE) pre-training model it overlooks the existing dependencies within masked positions and that cause the discrepancy in finetuning. But when it comes to the generative natural language tasks the AR model (Auto-regressive) outperforms the AE (Auto-encoding) model such as BERT.

As we understand the XLNet is Auto-regressive (AR) model, it uses the context word to make prediction of the upcoming word, the context word can be in two directions, forward and backward. That formation makes the AR language models such as XLNet, GPT, GPT-2 ideal for generative language processing tasks. [74, 75]

2.9 Summary of Background

In this chapter we thoroughly discussed the work related to the development of chatbot technologies and frameworks. We started this section with documenting related work in the area of chatbot technologies and how these systems are adding values to encounter different existing scenario such as customer services help, product suggestion, ticket reservations, any many other routinely operations. And with the passage of time the shift in the approach of being simple retrieval-based models to advance generative NLP based frameworks was evident. We discussed the basic limitations we found by evaluating the research work provided by different researchers and organizations.

Most important part of this research was to provide a theoretical overview of this technology to the new tech audience, like how it came into existence and how the research community's perception and approach evolved with the passage of the time. We conducted research survey by evaluating hundreds of research papers and documented the finding in to two different phases. First in **Table 2.1** we discussed the early approaches taken by research community from 1950 to 2000. Where the rule based or scripted approach was used to train machines and predict based on the query asked and data provided in the script [23, 30] it was more than a revolution in the field of machine learning and based on these finding in 1971 the first chatbot "perry the doctor" [31] was introduced by using a model based on the behavior of a paranoid person using the base concept of disease. From the data we collected it was evident that the community worked on same kind of Chatbot frameworks with obvious improvements, but the actual magic happened in the beginning of 21st century. **Table 2.2** is the illustration of the advancement we had in last 2 decades, how the approach shifted with the introduction of advance machine learning algorithms and NLP models. The use of sequence to sequence and encoding decoding framework with the RNN networks was extremely successful for language translation tasks and next sentence generation.

We later discussed the Natural language processing models available these days and how they are providing the boost in terms of accuracy when it comes to sentence generation that would certainly help in terms of productive communication between machine and human. And same framework and models can help us to build advance AI chatbot that later can be used in interviews. This data is enough to support our first 2 research questions mentioned in the introduction chapter. In Natural Machine Translation section we provided all the NLP models that can provide better results when it comes to Interview related communication.

CHAPTER 3. METHODOLOGY

3 METHODOLOGY

Methodology part is the most important part of any technical writing, and it becomes more relevant in case of thesis as it set the theme for the whole development process. In this thesis one of the most important research question was “To find resources that can help us to build advance AI based smart chatbot machine”. To find the answer we divided my research work into two different segments. One method was purely theoretic where we conducted a technical survey from the very beginning when first time chatbot technology and the ideas of machine learning was introduced back in 1949-1950 [23]. For this purpose, I found and studied more than 100 relevant research papers in yearly order to analyse the development the mindset (regards with smart chatbot) that evolved throughout the last 7 decades.

The technical Investigative survey segment was also divided into 2 separate phases. First phase registered the milestone in terms of rule-based chatbot models during the era of 1950-2000. **Table** 2.1: Summary of research work on smart chatbots (1950-2000)” has all the information’s that we find in that era. **Table** 2.2: Summary of research work in smart chatbot development (2010-2022)” contains the recent developments in chatbot, machine learning and natural language processing technologies from 2001 to the recent years.

In the second part we worked on finalizing the architecture of the model and how specific technologies can be glued together to get the desire product. Before moving to discuss how to build smart chatbot with the suitable architecture and frameworks lets first discuss the ACM classifications that can be used for technical part.

3.1 Overview

Before starting the development of the advanced software systems (chatbot), it was important to gain a deep understanding of the technology and its scope, as having enough information and data about the subject is always key to productive research practices. There are many theoretical and experiment-based research methodologies and classifications [24] aimed to assist the research like ours in more coherent and consistent manners, such as guidelines suggested by the "Association of Computing machinery" in 1989, [25] the ACM classifications. These guidelines helped to maximize the scope of "Computer science" as a scientific field and disapproved of the perception that Computer science is only programming. The ACM divided the computing subject into three somehow co-related "paradigms" mentioned in the report "Computing as discipline" [25], (i) Theory, (ii) Abstraction, (iii) Design. [25, 26]

3.1.1 Theory

The idea mainly comes from the field of "Applied mathematics" where the main concern lies towards the development of the accurate and viable "theory". In theory, we define the characteristics of the objects and their relation, then we hypothesize the relationships between the objects, later we evaluate the correctness of the relation between the objects, in the end, we collect the data in form of interpretation of results. [26]

3.1.2 Abstraction

According to the ACM report, the Abstraction paradigm is more like an "experimental scientific method". In this phase, we usually perform an investigative approach which involves multiple processes to investigate the theory, (i) at the beginning of the phase we establish the hypothesis, (ii) basing on the hypothesis statement we design the prediction model, then (iii) we work on designing the experiments and collect relevant data, In the end, (iv) the most important part is analysing and evaluation of collected data. [25, 26, 27]

3.1.3 Design

The “design” module is rooted in engineering within the computer science field. It is the most important of all as the success and the failure of the theory depends on it. It starts with (i) Finalising the Requirements and requirement elicitation, then (ii) stating the specifications of the product, (iii) designing and developing of the product/ system based on requirements and specifications, in the end (iv), most important part testing of the system and evaluation of the results. [25, 27, 28]

3.2 Framework and Architecture

Recent advancement in technology has made a great technological impact on the human way of interaction with robots and machines. This uprise in the ‘Human-computer Interaction’ segment leads to the development of dialogue-based and speech base engines like XiaoIce, Cortana, and Siri, etc. All these AI-based human assistant machines can interact through voice commands, hand gestures, text-based conversations, or with device touchscreen. [14]

In the near past, the development of intelligent conversational robotic systems and their successful integration with the enterprises initiated a new dimension in this research field. The scope of these intelligent chatbots is not limited to specific tasks or services such as “customer service agent”, “booking and reservations” or “shopping advice” it offers way more than that. As we understand that chatbot technology has reached the level where the development of human alike conversational machines is very possible, where these chatbots can assist us to handle a complex task such as decision making, teaching, or employee training. Employ training is one of the key factors in business, Organizations allocate a rich portion of resources towards employee training and education as the whole business architect is based on skilful employees. But as a matter of fact, technology is assisting us in almost every field of work including in the recruiting process to employ training. Studies have shown that with required functionalities to the chatbot system can be an alternative to this orthodox time consuming and expensive way of training.

Here we have documented the resources to build an architecture of smart Chatbot machines. In limited research work, in phase 1, 2 we tried to find the advance resources required for the development of human alike and smart chatbot systems. We know it requires more research work to finally reach the level of information to decide which model and technology will be ideal to counter these problems. But relying on our limited research work we find these below mentioned technologies as a best suit for our problem statement.

3.3 Proposed Methods and Technologies

3.3.1 Google's (NMT) Model

The Neural Machine Translation system was created to overcome the gaps and vulnerabilities of the existed phrase-based translation approaches. This NMT system was better in execution and results, but it was a very expensive system for deployment as it required more resources such as finance and very large datasets for training and then building interfaces for translation. To provide the solution to these problems Google came up with a new "Google's Neural Machine Translation system" based on deep "Long Short-Term Memory" LSTM along with 8 encoders and decoders. GNMT model is aimed to provide much-needed robustness and versatility where it requires less training time, and due to the use of the "Low-Precision "approach while computation the translation speed will be robust. [13] Google's Neural Machine Translation model is based on Sequence-to-sequence approach famous for text generation or dialogue generation. It also provides agility where its built-in techniques are essential to creating any advanced chatbot framework. [7]

3.3.2 Seq2seq Framework

Very popular and widely used end-to-end sequence model proposed by Google in 2014. [11] Unlike the Deep Neural Network (DNN) framework that requires rich training datasets for better results and shows limitation when the solution of the problem or task cannot be encoded into the fixed length. Sequence to Sequence model is more like an end-to-end sequence-based model using encoder-decoder architecture relying on RNN "Recurrent Neural Network", specifically used to handle tasks such as image captioning, human alike dialogue creation, machine translation, interrogative conversation, and machine translation. [7] Sequence to Sequence model used the "Long Short-Term Memory" (LSTM) system to make a fixed input sequence in a vector where the vector's dimensionality is also rigid, for the decoding another deep Long Short-Term Memory used to get output from that vector as a target sequence. This way it is not necessary to have an equal length of the output and the input. [7, 11, 12]

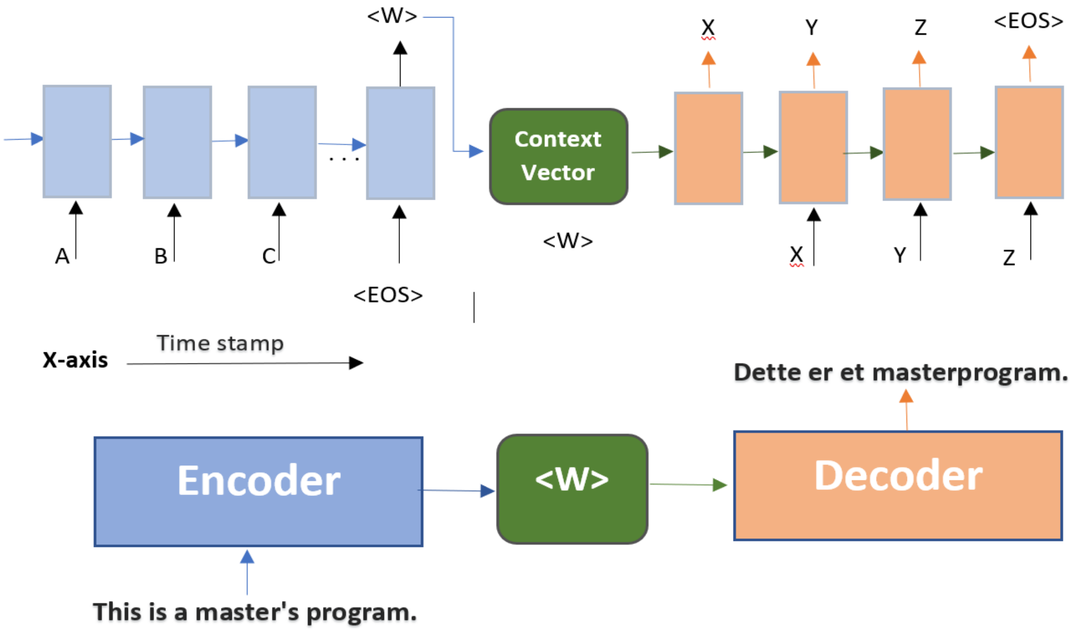


Figure 3.1: Seq-2-Seq model with encoder decoder technique

In Figure 3.1, we are illustrating the encoder-decoder framework [11], where we are giving input A, B, C (x1, x2, x3 up to xn) to a series of neural networks along the x-axis that represents 'timestamp', here <EOS> is the end of the stream where the input sentence ends and it provides output as a vector form representation of the sentences <W>, this is an encoder. Now the output of <EOS> which is <W> will be passed on to the first timestamp (neural network) 'X' of the decoder section and it gives us output y, that is used as an input to the next neural network until we get the output < EOS >. The best thing about this technique is that the output is flexible in length, it could contain more words in output than the input sentence length.

3.3.3 Recurrent Neural Network

Recurrent Neural Networks are deep learning model, recently became famous for most advance machine learning applications. The reason behind its wide applause is because it has very organized and simple structure in place for "Feedback loop" that provides the ability to perform as a forecasting engine. It was not the case until recently as previously recurrent neural networks had network system with built in "Feedforward loop" networks that are only allow the flow of the signals in a single direction from input to output. In RNN the output of previous layer is used as an input data for the next step and fed back to the same layer that was used at the first time and off course the only layer available in the network.

Recurrent neural networks are designed to provide solutions to the sequential problems. If we study the basics of this model, we will analyse that the core purpose behind their application is availability of recurrent connection nodes among the layers. [57] RNN takes an input from the sequence separately, one at a time that allows the hidden units in the network (Vectors) to remember the history of the previously used element in the sequence and that affects the outcome based on the data in the previous computation. This is the quality that the simple Recurrent network model lacks when it comes to word prediction in a sentence, but recurrent neural networks provide this agility in word prediction in a sentence. Here we would like to discuss the research work on the Uni-directional recurrent neural networks by Lechun. [57, 58]

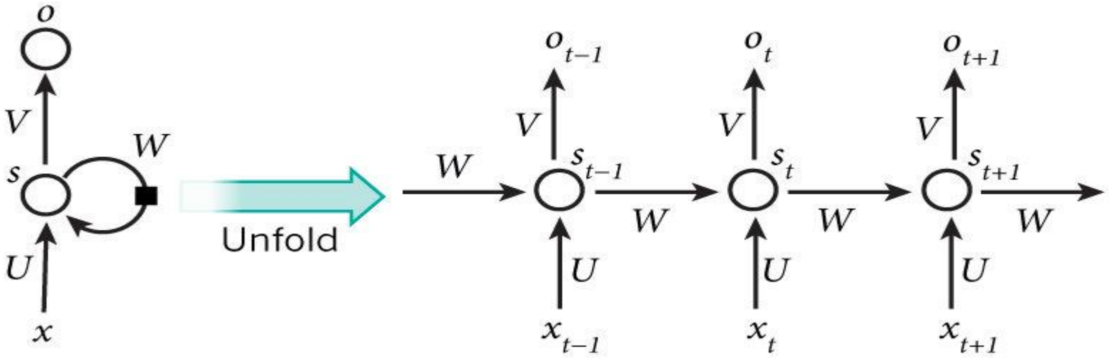


Figure 3.2: Unfolding RNN over 3-time step (source: Lechun 2015) [57, 58]

In the above figure (source: Lechun 2015) we can relate to the process, it shows unfolded RNN in a time to provide the accurate feed forwarding network. Where we can see that **xt** is time step in the network and **st** is the hidden step in the sequence of the **t** time step.

$$st = tanh\ (U\ xt + W\ st - 1)[59] \qquad \textit{Eq(3.1)}$$

As we describe earlier that in RNN takes an input from the sequence one at a time that allows the hidden units in the network (Vectors) to remember the history of the previously used element in the sequence and that affects the Output. And this is also known as a network memory. In the above mentioned equation and figure 7, the x is representing the input, o is representing the output, U,V,W, are representing the weights in the network and s is the output of the sequence in the network. [57,58]

3.3.4 Decoding Input Sequence

Sentence sequence is extremely important matter when we are dealing the natural language processing related or auto sentence related jobs. To be able to get most out of it there are many techniques and methods to perform decoding of the input sentence from probability of the words in the sentence. One thing to notice is the probability of getting the same output based on the same input is on higher side, hence it is labelled as greedy approach. Better alternative could be decoding the entire word sentence based on the highest probability instead of the singular word on each iteration. [60]

$$T' = arg\max p\ (TS|SS)\ \ [60] \qquad \textit{\textbf{Eq(3.2)}}$$

In the above formation TS is a target sentence and SS is representing Source sentence. As our priority is to perform decoding entire sentence based on the highest probability instead of the singular word, the beam search (left to right) will make the traceable computation and form a sequence of all possibilities. As mentioned in the research paper of Mr. Richard [60] during the first iteration the word with the highest probability of occurrence will be kept and allocation with the highest K score. During each iteration the K value will be assign to the most obvious words hence the pool of words with highest probability will be increasing. The iteration will be continued until it reaches to the end of the output sequence.

3.3.5 Vocabulary Formation

After discussing the basic decoding model of the input sequence for the sequence-to-sequence method, let’s consider another very important aspect of the sequence-to-sequence model. It is obvious that whenever there is language aspect, we must consider its most important component “Vocabulary”. Vocabulary contains different word and various symbols and if we are dealing with high end dataset, we might encounter enormous amount of data that will increase the parameters used for the models proportional to the size of vocabulary. This could create a bit of problem as all the attributes in the dataset does not contain the vocabulary, one way can be embedding the most frequently used words in the dataset and replacing all other unimportant symbols and signs as an unknown word represented by common tokens as mentioned in the research paper by Mr. Luong. [60, 61]

3.4 Architecture of Smart Chatbot

3.4.1 LSTM (Long Short-Term Memory)

LSTM “Long Short-Term Memory” are a type of Recurrent neural networks specialized in long term sequence of dialogues generations. we can describe the RNN as special functions that has ability to match one kind of variables with another kind of variables, during the process of these sequence mix we end up getting different useable architectures such as vector-sequence model, or sequence-to-sequence model. Sequence to sequence model is the architect of our interest as we are using this model for the development of Intelligent Chatbot. We explained seq2seq model with details earlier in “Proposed Methods and technologies” section. LSTM are a gated recurrent network, and for example if we have vanilla recurrent network where we use hidden units, if we replace the units with LSTM cells and it will become LSTM RNN. The basic purpose of LSTM was to handle the vanishing & gradient problem as each, and every hidden LSTM cell maintain the cell state of the vector v and every next cell can decide to read from the right of the cell or completely reset the cell by gating mechanism.

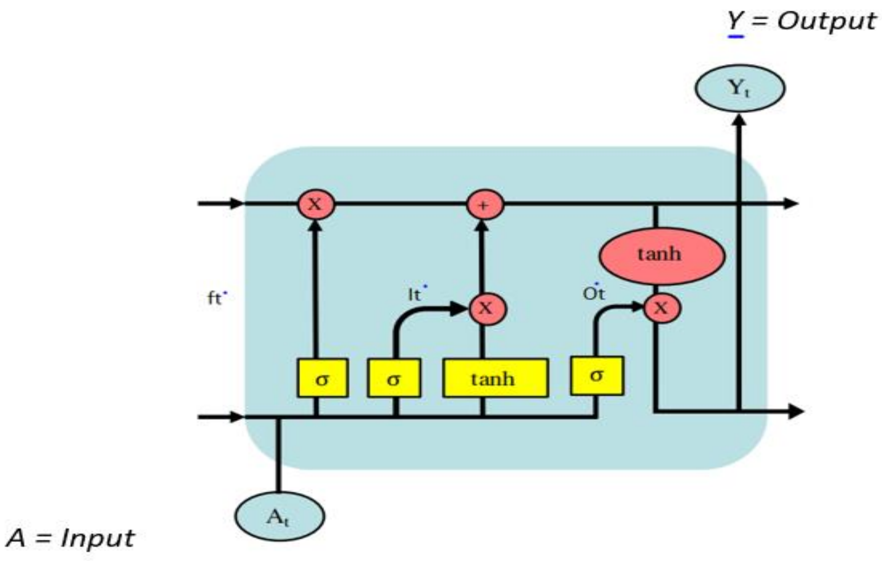


Figure 3.3: Cell structure of LSTM by "S.R Bukka", 2020. Source: [66]

In the diagram we can see that all the RNN networks has form a repeating circle of the networks. LSTM cells maintain the cell state of the vector and every next cell can decide to read from the right of the cell or completely reset the cell by gating mechanism. All the unit has 3 gates or binary gates, the input gate A is responsible about the current update in the cell, F forget gate checks either the gate is set to zero or not, Y gate is A output gate responsible for the visibility of the output sequence.

3.4.2 Gated Recurrent Unit

We know that basic recurrent neural networks highly suffer from the memory problems especially the short-term memory. Usually, the RNN only remember so little that sometimes won't provide better results. To handle the deficiencies long-short term memory or LSTM RNN were introduces where these neural networks had better chances to remember the previous states. GRU or gated recurrent unit are light weight form of LSTM, and it actually combines the short term and the long term memory into a hidden state. If we investigate the LSTM section above, we can see that in Figure 3.4 that LSTM had 3 gates (input, output and forget) gate. But GRU is different the formation and it has only 2 gates (update gates and reset gate).

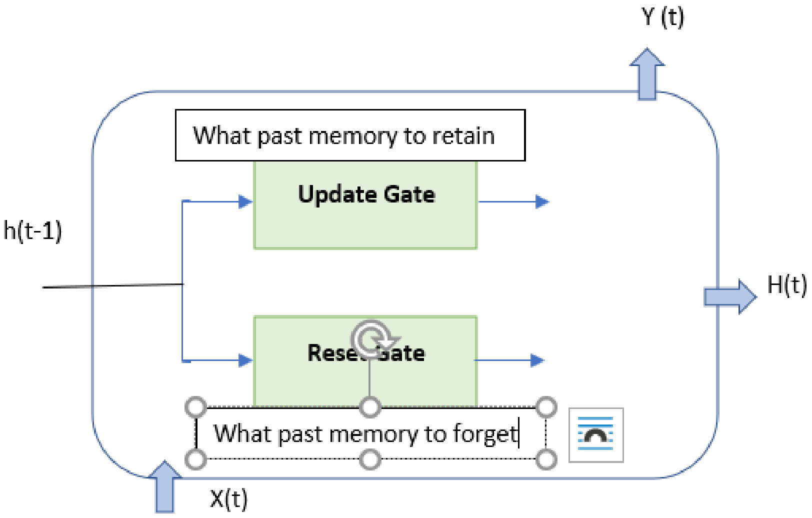


Figure 3.4: GRU gate representation [77]

Figure 3.4 shows that comparing with LSTM RNN, GNU only has 2 gates. first gate is “update gate”, it decides that what should be retain and the second gate is” Forget gate” responsible to decide how much past data should be forget. [77]

3.4.3 Attention Mechanism

The neural attention mechanism was first introduced in 2015, [48] It can build excellent systems for neural, machine translation. The idea behind the attention model is to somehow develop an attention configuration model that will help to understand that what input data from a source sentence is important to focus on to generate corresponding output against that data. During translation, it forms short-cut connections between output and input based on the attention factor. [49] An image from Bahdanau [50] is a fine example of alignment matrix visualization between the input source and target output.

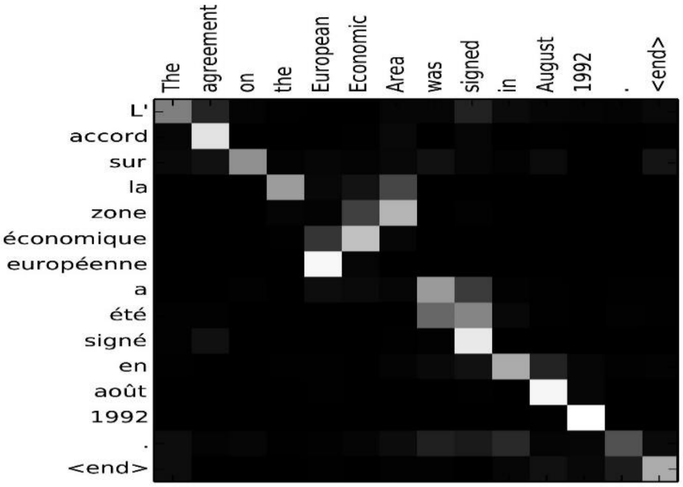


Figure 3.5: Attention visualization of Alignment matrix: Bahdanau [50]

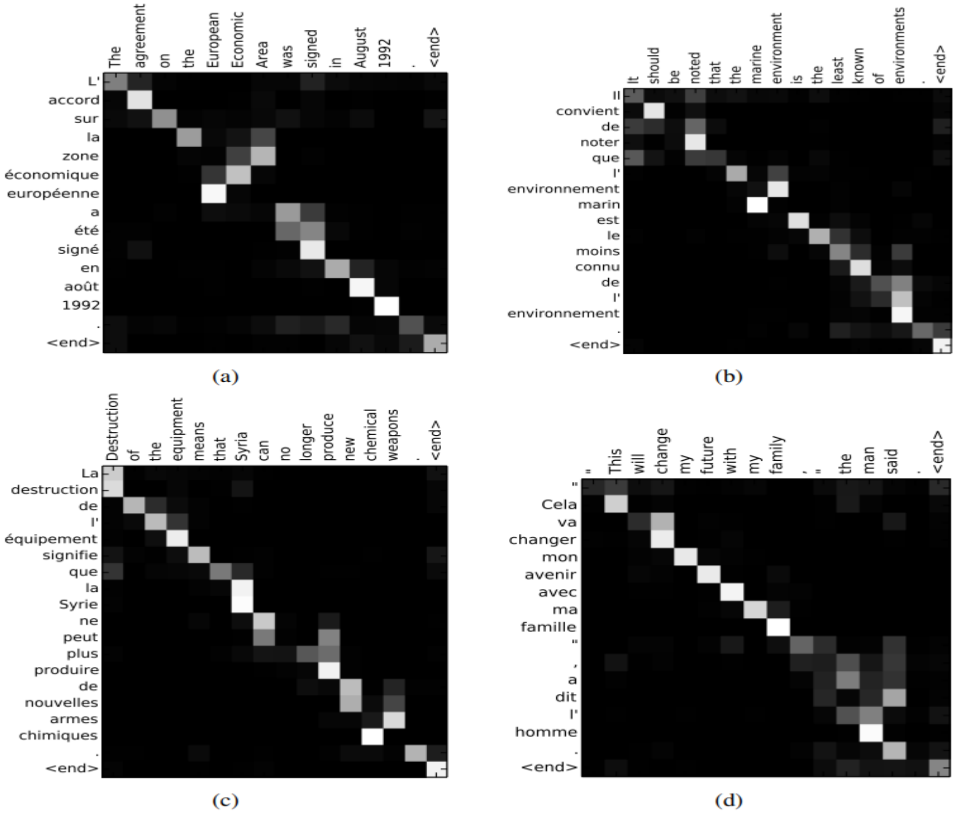


Figure 3.6: Attention visualization of Alignment matrix

Source: (Bahdanau et al., 2015) https://arxiv.org/pdf/1409.0473.pdf [50]

Above matrix produce by (Bahdanau et al., 2015) is a sample model for 4 alignments produced by RNNsearch where x-axis and y axis hold the English words of the source sentence and respectively the French translation. One thing to notice in this formation the use of fixed length vector for the basic context relying on sequence-to-sequence based encoder-decoder approach. This is good formation, but one possible disadvantage could be the limitation when it comes to the long sentences of the source sequence. It can affect the overall accuracy of the encoder-decoder model.

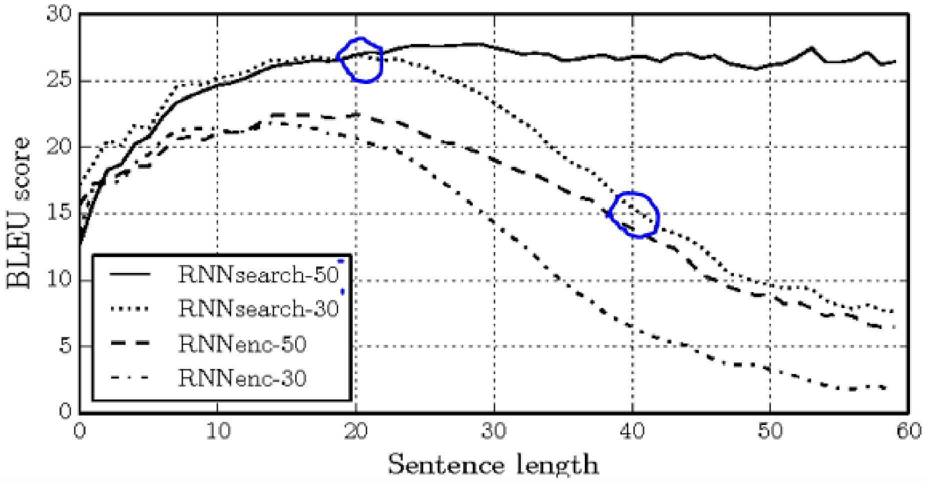


Figure 3.7: Attention model with blue score

Source: (Bahdanau et al., 2014), https: semanticscholar.org/paper/Neural-Machine-Translation-by-Jointly-Learning-to-Bahdanau-Cho [50, 62]

In the graph representation by Bahdanau [50] we can clearly analyze that how our previous argument has weight where we stated that the increase in the length of the sentence directly affect the attention mechanism as it will require more data to handle. In the graph it is evident that lengthy sentence in the given data resource outperform the encoder-decoder model. But as our main agenda for the development of the chatbot is to build a system that may be used for interview related scenarios, and we assume that most of the direct questioning does not have lengthy outline. Most probably this model in sequence-to-sequence framework will perform as per our desires.

3.4.4 ECM Model

With the advancement in NLP and machine learning technologies, now researchers are working more and more on emotion and sentiment aspect of the machine understanding. Emotions play a vital role in any dialog-based conversation or discussion. "ECM" ‘Emotional chatting machine’. [2] were proposed by **H. Zhou et al** where three new mechanisms for the development of sufficient Emotion based chatting machine were discussed.

- Analysing the expression of emotional conversation and create categories.
- Record the fluctuation & change in an internal emotional state.
- The vocabulary which used to show explicit external emotions.

As the idea was new and with collaboration of Microsoft China, they were able to build good model as per the experiment analysis, the results were encouraging, the ‘ECM’ system was able to respond to the questions more consistently and formed way, not only content-wise also the structure of emotional conversation and out was well formed with proper vocabulary based on the situation.

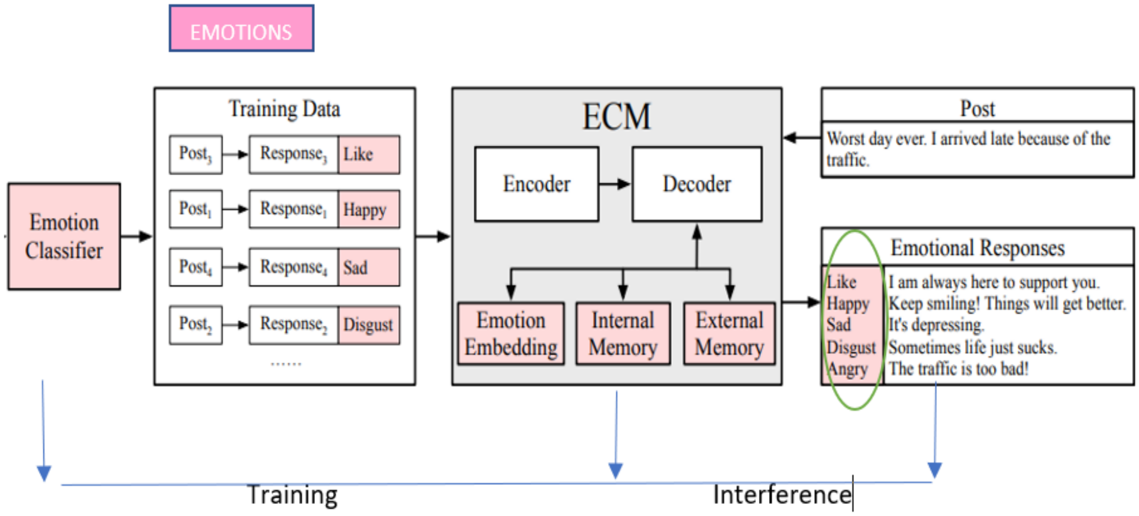


Figure 3.8: ECM model: by H. Zhou [2]

source: h arXiv:1704.01074v4 [cs.CL] [2]

In the above figure, the Pink colored areas are representing the emotion classifier in the ECM model. This model is unique in its built as it was not only taking care of the proper usage of the emotion-oriented/ library but also the aspect and scenario. As the same words can be interpreted differently based on the situation and sentence.

3.4.5 Beam Search

Unlike the "Best First search" algorithm, beam search is a more efficient technique when it comes to handling space complexity, and it is very important to take care of memory state when we deal with large data sets. Beam search provides better translation quality even better than greedy search. In beam search, we set the "Size of the beam" or "Beamwidth", where we can decide the number of top candidates against each word during translation. For example, we want to translate a sentence "This is master thesis!" into other language and we have set beam width = 3, with a vocabulary dataset of 1000 words, during translation it will initiate 3 copies of these networks against each word into the memory, and if we count all the options for the first word it would be 3000. This is how we can handle the memory complexity when it comes to big datasets, where other search algorithms can make the system inefficient and slow. [50]

3.4.6 Greedy Search

Greedy search is method that generally involved to handle the optimization problems. Optimization problems usually requires extreme case or minimal case results. Majority of the NLP based application or models or chatbot frameworks they end up producing a raw output initially, but most important part in those NLP based applications is to produce final output. Beam search and greedy search are couple of those algorithms that are being used for the purpose.

In beam search we understand that the computational requirements are high, and we need extra time and resources, but in greedy search we often get the correct results without getting the hustle of additional resources and time. Greedy search algorithm just takes the variable (word) with the highest probability based on the scenario and provide the prediction with taking too much time and additional computing and rendering resources. [67]

3.5 Importance of Datasets

Efficient Training material for machine learning frameworks plays a vital role in the success of these ML models as their performance is dependent on the quality of training we provide. Therefore, preparation of these training datasets is also a very tricky task to achieve, however, in the field of machine learning especially related to Chatbot training one can find a collection of many available public datasets that can be used for training & testing purposes. One thing to notice is that the availability of these data collections is mostly in “English” & “Mandarin” languages. According to the research conducted by Sutoyo et al. the summary of these widely available datasets is shown the table 3. [4]

Dataset	Topic	Dialog Quantity	Language
Facebook Dataset [44]	Persona-based	700 million	English
Twitter Dataset [45]	General	1.3 million	English
Wibo Corpus [46]	General	4.5 million	Mandarin
Cornell Dialogue Corpus (Movies) [47]	Film Dialogue	220 k	English

Table 3.1: Summary of widely available data sets. [4]

3.5.1 Data Formation and Processing

For our project we have used Popular movie subtitle data set Cornell Dialogue Corpus (Movies) [47]. There are many other available Datasets that could be used but due to the enriched constancy and coherent plot available in the movies scripts makes this Cornell Dialogue corpus extremely useful when it comes to chatbot training.

The Cornell Movie Dataset contains,

- More than 220,000 conversational contexts
- Around 11,000 paired characters
- Contains around 9000 Characters (600 movies)
- Around 300000 proper sentences

That is the reason mostly chatbots are trained initially with the Cornell Data set. As we explained above that why we are choosing the "Cornell Dialogue Corpus" as it is formed in very sophisticated way. In the latest available corpus, it contains the id numbers of all components, it contains ID of the Movie, ID of the character in the movie script, ID of the line with the dialogue. All these components are separated by "+++++". For the purpose of data formation for pre-processing, it was very import to clean the data. To achieve that the conversational data was cleaned as well as the meta data (IMBD rating, year, votes, score etc) it also included all the components of the data such as (ID's of movie, character) etc. in the dataset there were also elements with UTF-8 standards and it was very difficult to process for the model hence those components were also eliminated.

As our model was based on sequence-to-sequence model so the data needed to be formatted in a way that serves the purpose of training. For that we have formed the data into lists with proper values where the first list was formed as a dialogue and the second list represents as response to the first list. We can also assume the first list as a question or the seconds list as answer to the question. These formations were the most important one to be able to train our model on clean data. To make the data more reliable we also remove the punctuations (? , ! , – , “ “) and some special symbols as they were not as import in small questioning as compare to long engaging conversation. For better performance the punctuations were parted with the single space before and after except (‘ ’). All the additional spaces in the dataset were removed to avoid the overload on the initial model. The availability of multi utterance from the same user in a same situation were also taken care off by elimination the irrelevant and adventitious dialogues.

As we documented earlier that Cornell Movie Dataset contains more than 220,000 conversational contexts and around 300000 number of sentences, that is a lot for this kind of model. So, we formed the training data in a way where we used small length sentences (typically between 2 to 6 words) as a training context. Moreover, this famous cornel movie dataset contains around 6000 most frequently used words (at least two times), and we are taking them as our vocabulary stock. Further, to simplify we used different token for different purposes for example if there is requirement for padding the sequence of incoming data to the same length and size <Tx> was used, for singling the end of the sentence token <ty> was used. As the exception of having not proper words for the situation we have used token <Tz> for the purpose of registering the unknown words in the vocabulary backlog. [64]

Tokens:

- <Tx> = padding in the sequence
- <Ty> = End of the sentence
- <Tz> = Unknow words
- <Tg> = Go, Kick starts the sentence

By the use you these tokens it became easy to work with fixed sequence. All these steps will make it simple for the process of model training. Let's, discuss an example of the benefits of this approach, assume we want fixed length of the question and answer in a sequence to set equal to 8, and the question contains less words, (what is your name?). It contains only five words but to fulfil the requirement it needed to have 8 words to fulfill the sequence length. By using above mentioned tokens will help us in this situation.

Input: what is your name?

It will transform to fixed sequence length using the tokens, Tx, Ty, Tz or Tg.

Input: [Tx, Tx, Tx, "?", "name", "your", "is", "what"]

Same formation will happen during the output sequence for answering the question.

Output: I am Sheraz.

Output: [Tg, "I", "am", "Sheraz", ".", Ty, Tx, Tx, Tx, Tx,]

It is fascinating that how this padding or use of tokens can make things simpler for us, but what if the length of the sentences is not within our comfortable spots, what if instead of 5 words we end up with sentences of the length as 50 or more. How will we handle the scenario as if the largest sentence or dialogue in our dataset is of 50 words our system must be able to encode everything based on that fixed length only than it won't have any issues dealing with all inputs and response generation. But if we take our above discussed test case where we delt with the input of 5 words and fix sequence of 8 in the vector, what is the longest sentence is of 50 words and we encode 5-word input in respect with fixed length vector of 50 words. It will be a lot to deal, and it will create the extra memory by adding 45 padding tokens <Tx> in the end. It will not look appropriate as the essence of the actual text will be lost within the formalities. [64]

To save the model from overloading and overwhelming the formation, there is concept of "Bucketing". What bucking do is create a list of buckets with appropriate classification of the length such as [(5,10), (11, 20), (21, 30), … (41, 50)]. As per our test case performed above it will be dealt with the bucket length 5 to 10 and the formation will be the same.

Input: [Tx, Tx, Tx, "?", "name", "your", "is", "what"]

Output: [Tg, "I", "am", "Sheraz", ".", Ty, Tx, Tx, Tx, Tx]

3.6 A Prototype of Smart Chatbot with Emotion classifier

With the help of earlier conducted investigative research on smart chatbot frameworks, now we have enough data that suggests the encoder-decoder based Seq-2-Seq architecture is ideal to fulfil our requirements of a smart chatbot.

The encoding-decoding-based Seq-2-Seq framework enriched with the ability to learn smartly and able to provide quality responses constantly. Encoder decoder model consist of neural networks (RNN, LSTM, GRU) etc. In this case bidirectional LSTM neural networks, LSTM works on the probability of getting the output based on the input training data. During this process of prediction to avoid the loss Adam optimizer can reduce the loss using back propagation technique. Furthermore, it has attention mechanism to provide productive results during the training stages.

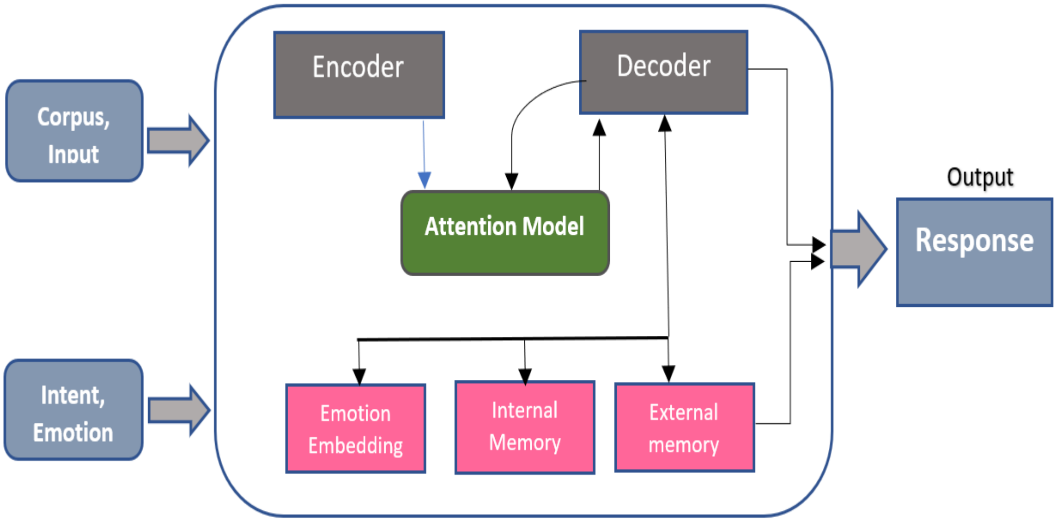


Figure 3.9: Encoder-decoder architecture with an emotion classifier. [51]

In Figure 3.9, we are feeding input to our machine and using an emotion classifier that will help us to detect or identify the intent/ emotion in the given dataset, we are using seq2seq framework that has encoder-decoder architecture, so this data will be passed to encoder, then from encoder to "Attention mechanism" for efficiency and then to a decoder. here pink units are illustrating the mechanism to generate emotional expressions proposed by Zhou et. al [2, 51] that includes "Emotion embedding" that shows the high-level abstraction of emotion state than "internal memory" and "external memory" units for making a balance between internal emotion state and explicit emotion state by external memory module.

Furthermore, it has attention mechanism [48] that can build excellent systems for neural, machine translation. The attention model is ideal to develop an attention configuration model that will help to understand that what input data from a source sentence is important to focus on to generate corresponding output against that data. During translation, it forms short-cut connections between output and input based on the attention factor. One advantage of this attention mechanism is the formation, the use of fixed length vector for the basic context relying on sequence-to-sequence based encoder-decoder approach.

3.7 Summary of Methodology

In this section we started with the explaining the ACM classification that used during the development of the technical projects, The ACM divided the computing subject into three co-related “paradigms” mentioned in the report “Computing as discipline” [25], (i) Theory, (ii) Abstraction, (iii) Design that makes easy to form and plan the development model. [25, 26]

Our aim of this paper was to find the possible resources vital to build smart chatbot system that can be trained later for interview related tasks. As stated in our research question “What are the resources available to build an advanced chatbot system?”. We provided detailed overview based on the technical survey we conducted in chapter 2 about the models and frameworks can be used to develop the smart chatbot system, such as encoder decoder framework using Bi-directional LSTM with attention model.

In Architecture section we provided the components that can run on sequence-to-sequence model to build a smart chatbot system. In seq2seq model we use encoder decoder framework that consist of neural networks (RNN, bi-LSTM, GRU) these networks work as a training unit, and then attention model with the greedy search adds the required efficiency by deciding what input data from a source sentence is important to focus on to generate corresponding output against that data.

CHAPTER 4. EXPERIMENTS

4 FINDINGS & EXPERIMENTS

4.1 Initial Findings

As we define in the INTRODUCTION chapter section “Research Questions” that what are the resources available to build smart chatbot system? And how smart chatbot can be built? we conducted a research-based survey in BACKGROUND chapter section “Investigative Research Survey”. This research aimed specifically the Chatbot technology, how it started, when it started and how it evolves to this very day to provide the insight to the audience that want to pursue their research in this exciting field. Initially we divided the research-based survey into two different stages, first stage was to study the chatbot technology in 20th century, specifically (1950 -2000) and second stage was to address recent developments in this field.

Year	Topic	Model Type
1950	A.M Turing Computing Machine	Rule-based
1966	Aliza Chatbot, Pattern matching conversation based chatbot	Rule-based,
1973	Perry the doctor (Psychiatrist)	Script-based
1995	ALICE (The Chatbot)	Script-based

Table.4.1: Chatbot model early approaches (1950 - 2000)

In the above table that based on the survey in Background chapter section “2.3.1 Initial Research Approaches”, we can easily determine that how the Chatbot technology came into existence and what development models were initially used. The findings registered above are the major breakthroughs that happened in previous century in the regard of Chatbot technology. One important take away in these findings is the “Model type”. “Rule-Based and Scripted model” were used to develop chatbot throughout the last century.

Year	Topic	Model Type
2014	Smart sentiment based Chatbot	Seq2Seq with GRU RNN
2015	Chatbot based on neural conversational Model	Seq2Seq with LSTM RNN
2016	Personal Based chatbot model	Seq2Seq with bi-LSTM RNN
2019	XiaoIce chatbot (IQ + EQ)	Seq2Seq with GRU RNN
2020	Chatbot for education (Interview)	Seq2Seq with RNN

Table 4.2: Chatbot models in recent years

In the above table, there is the representation of the some of the important chatbot related work done by researchers in last decade. If see the "Model type" column we can easily get the idea that all these chatbots were built on Sequence-to-sequence based encoder decoder model. Only difference is the used of neural networks. Some uses the long-short term memory neural networks some gated recurrent unit neural network. Based on these findings we have enough data to support our research question "What are the resources (framework) available to build an advanced chatbot system?"

4.2 Basic Smart Chatbot Model

Based on the data collected in Chapter 2 and Chapter 3 and the above section "Initial findings" we were very much convinced that Sequence to sequence model is the best suit for deep learning based advance chatbot system.

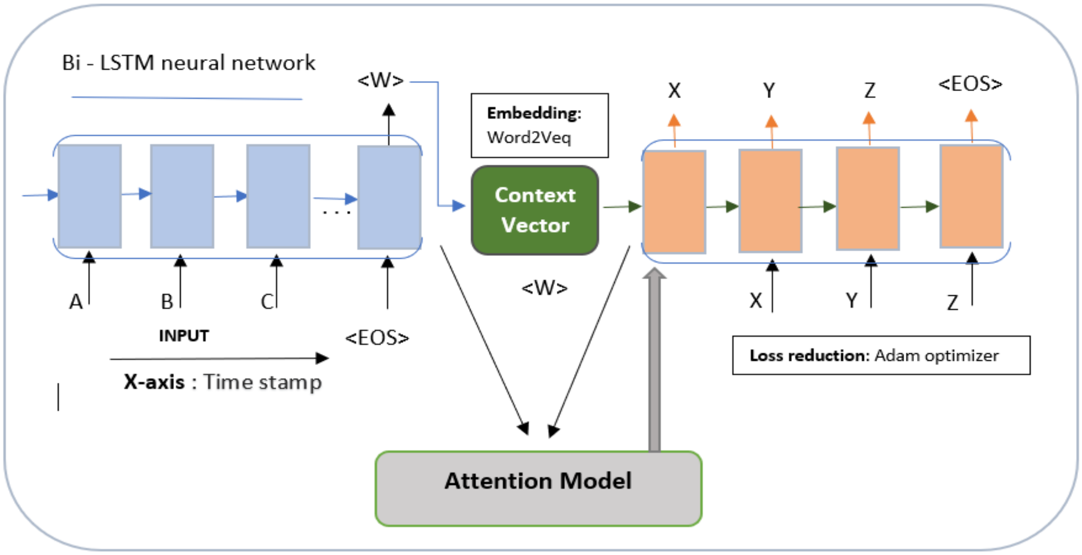


Figure 4.1: Seq2Seq model with attention mechanism

In the figure above there is illustration of how encoding-decoding-based Seq-2-Seq framework using RNN can be developed. Encoder decoder model consist of neural networks (RNN, LSTM, GRU) etc. In this case bidirectional LSTM neural networks, LSTM works on the probability of getting the output based on the input training data. During this process of prediction to avoid the loss Adam optimizer can reduce the loss using back propagation technique. Furthermore, it has attention mechanism to provide productive results during the training stages.

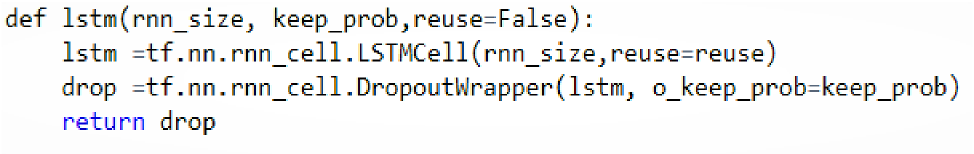
4.2.1 Overview:

Basic Seq2Seq model contains:

- Bidirectional LSTM or GRU
- Dataset formation (Chapter 3 section 3.5)
- Attention mechanism
- Loss reduction (Adam optimizer)
- Word Embedding (Word2Vec)
- Greedy Search or binary search

LSTM “Long Short-Term Memory” are a type of Recurrent neural networks specialized in long term sequence of dialogues generations. we can describe the RNN as special functions that has ability to match one kind of variables with another kind of variables, during the process of these sequence mix we end up getting different useable architectures such as vector-sequence model, or sequence-to-sequence model. [66]

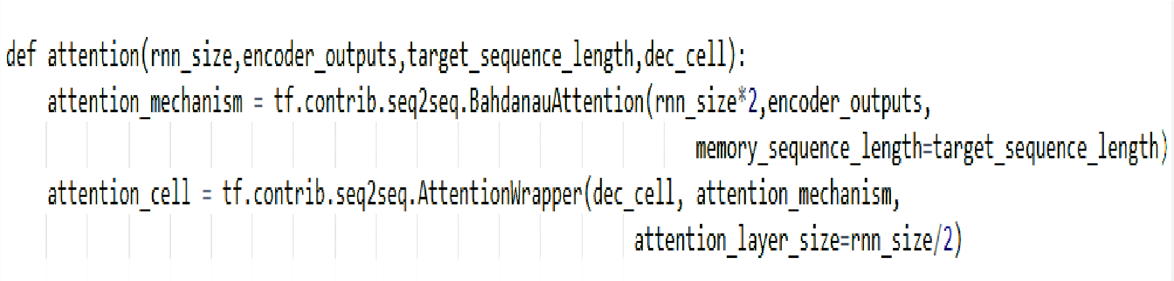
The LSTM class,



```
def lstm(rnn_size, keep_prob,reuse=False):
    lstm =tf.nn.rnn_cell.LSTMCell(rnn_size,reuse=reuse)
    drop =tf.nn.rnn_cell.DropoutWrapper(lstm, o_keep_prob=keep_prob)
    return drop
```

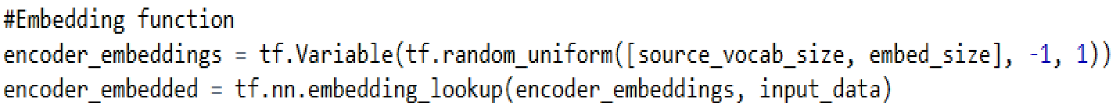
Here you can see the drop out function that add dropouts the cell (both input and output).

We thoroughly discussed Attention Mechanism proposed by Bahdanau [50] in Chapter 3. The idea behind the attention model is to somehow develop an attention configuration model that will help to understand that what input data from a source sentence is important to focus on to generate corresponding output against that data. It produces hidden state in the input sequence of the encoder.



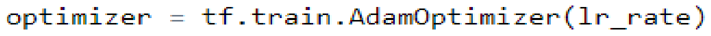
```
def attention(rnn_size,encoder_outputs,target_sequence_length,dec_cell):
    attention_mechanism = tf.contrib.seq2seq.BahdanauAttention(rnn_size*2,encoder_outputs,
                                                                memory_sequence_length=target_sequence_length)
    attention_cell = tf.contrib.seq2seq.AttentionWrapper(dec_cell, attention_mechanism,
                                                          attention_layer_size=rnn_size/2)
```

Word embedding is very important part of the vectorized data, and in encoder decoder framework when all the input comes to the context vector than for word embedding one can use the most popular word embedding technique “WordtoVec”.



```
#Embedding function
encoder_embeddings = tf.Variable(tf.random_uniform([source_vocab_size, embed_size], -1, 1))
encoder_embedded = tf.nn.embedding_lookup(encoder_embeddings, input_data)
```

After test and train setup, during the prediction of the next data in LSTM based encoder decoder model to check the probability of the loss is usually calculated by back propagation by using some optimizer. In this case “Adam Optimizer”.



```
optimizer = tf.train.AdamOptimizer(lr_rate)
```

This is how the basic advance sequence to sequence based model works for a smart chatbot. [48,50, 79]

4.3 Experimental setup

In experiments especially related to the NLP or machine learning based models, datasets hold the outmost importance as the complete scenario and procedures are based on that. As we thoroughly explained our dataset and data preparation model in previous chapter (3.6) where we explained the dataset formation and processing. we have used movie subtitle data set Cornell Dialogue Corpus (Movies) [47]. There are other Datasets like this one but due to the enriched constancy and coherent plot available in the movies scripts makes this Cornell Dialogue corpus extremely useful

The Cornell Movie Dataset contains, more than 220,000 conversational contexts, 11,000 paired characters and around 300000 proper sentences.

In experiment section we will document some of the experiments conducted by other researchers based on Seq2Seq model with different frameworks with different formation (NLP models) and we will compare the results for future work.

4.3.1 Training Model and Test Questions

In methodology chapter we have discussed in detailed manners that how we prepared our dataset for the training purposes. we have formed the data into lists with proper values where the first list was formed as a dialogue and the second list represents as response to the first list. We can also assume the first list as a question or the seconds list as answer to the question. In short, we have prepared a filtered form of question-and-answer list that has length of more than 22000.

As in this paper our focus was to conduct a technical research survey and gather the research data for the audience in regard with smart chatbot model. But we also have two basic models with same formation presented above. One model has encoder decoder LSTM neural network and other model has encoder decoder Gated recurrent unit. [] We will try to run some basic test on both models.

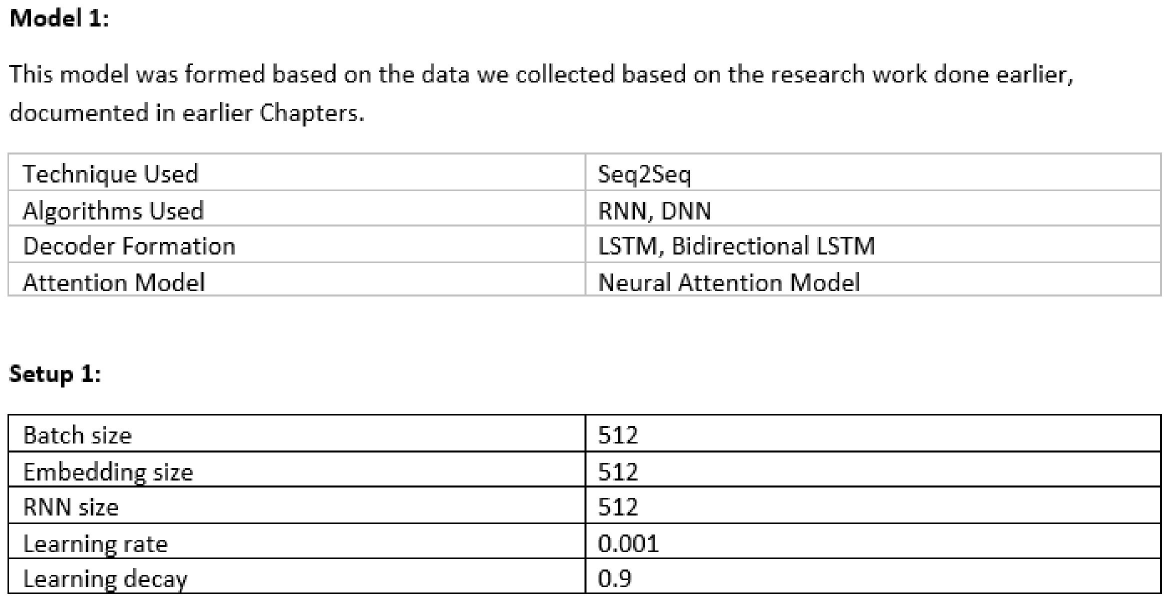
Here are the couple of question samples that we will use during our experiment phase.

- Hi
- Who are you
- How are you
- Who I am
- What you do
- How is the weather
- Is it raining today

4.4 Hardware Setup:

System Model	HP Zbook
Processor	10th generation
Installed Ram	16 GB
Graphic Card	Nvidia T500

4.4.1 Experiment and Results



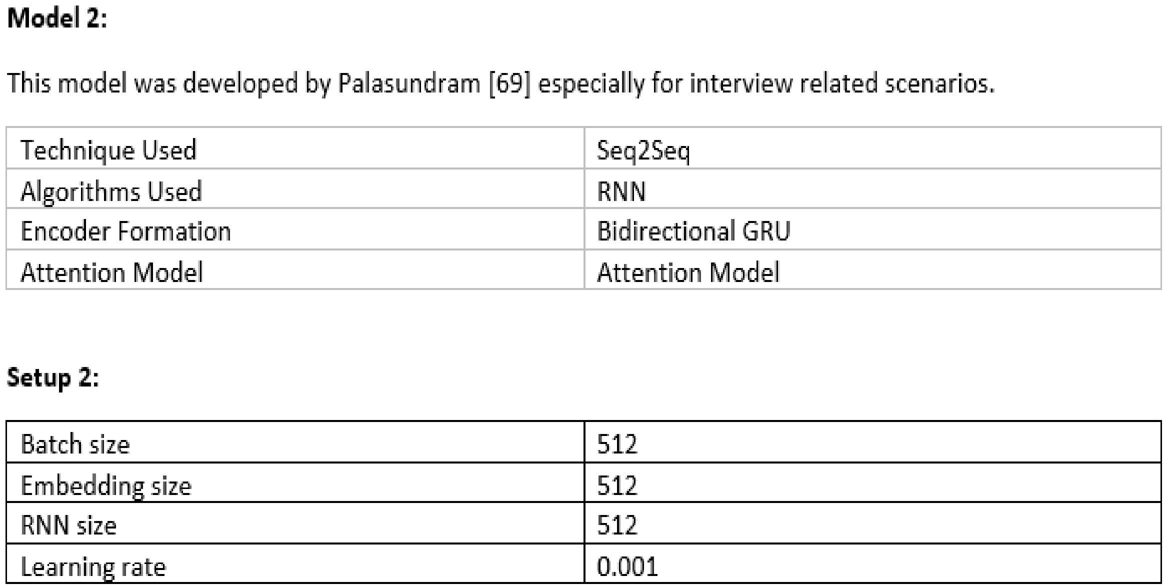
Model 1:

This model was formed based on the data we collected based on the research work done earlier, documented in earlier Chapters.

Technique Used	Seq2Seq
Algorithms Used	RNN, DNN
Decoder Formation	LSTM, Bidirectional LSTM
Attention Model	Neural Attention Model

Setup 1:

Batch size	512
Embedding size	512
RNN size	512
Learning rate	0.001
Learning decay	0.9



Model 2:

This model was developed by Palasundram [69] especially for interview related scenarios.

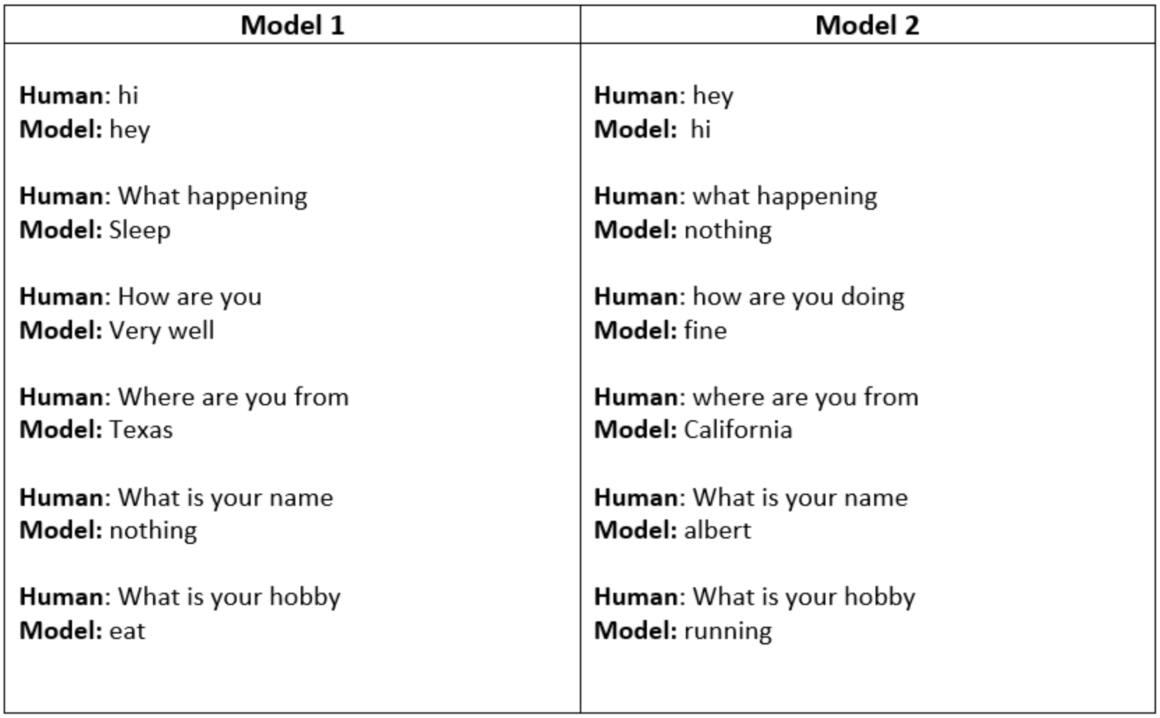
Technique Used	Seq2Seq
Algorithms Used	RNN
Encoder Formation	Bidirectional GRU
Attention Model	Attention Model

Setup 2:

Batch size	512
Embedding size	512
RNN size	512
Learning rate	0.001

Results:

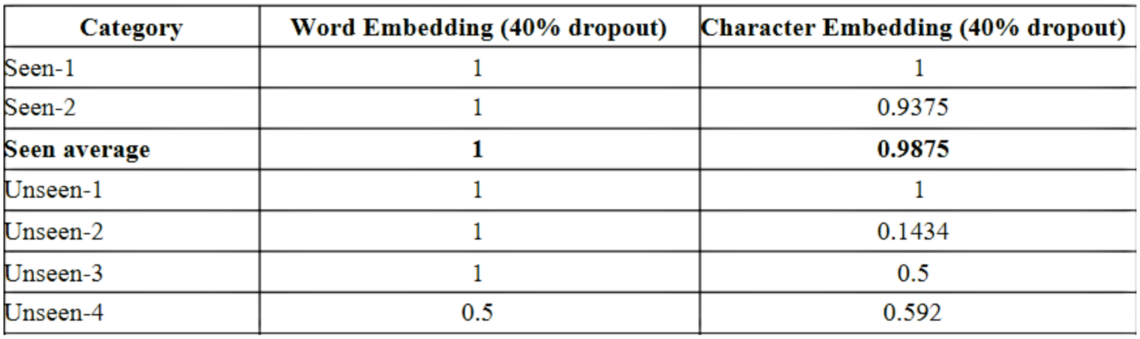
As both models are identical in every aspect of the structure except the model 2 that has Bidirectional GRU encoder. [69]



Model 1	Model 2
Human: hi **Model:** hey	**Human**: hey **Model:** hi
Human: What happening **Model:** Sleep	**Human**: what happening **Model:** nothing
Human: How are you **Model:** Very well	**Human**: how are you doing **Model:** fine
Human: Where are you from **Model:** Texas	**Human**: where are you from **Model:** California
Human: What is your name **Model:** nothing	**Human**: What is your name **Model:** albert
Human: What is your hobby **Model:** eat	**Human**: What is your hobby **Model:** running

Table 4.3: Chat results with both models

The table above shows almost identical results with minimal training. But as per the paper published by Palasundram [69] shows that the GRU bidirectional encoder-based model performed well when it comes to interview-based questions. As they created a special dataset model with the characteristics of "mapping cardinality", "synonym words", "similarity in questions". Then they divide the training model based on seen and unseen questions. Initially the training was done on seen question which were mapped one to one and one to many.



Category	Word Embedding (40% dropout)	Character Embedding (40% dropout)
Seen-1	1	1
Seen-2	1	0.9375
Seen average	**1**	**0.9875**
Unseen-1	1	1
Unseen-2	1	0.1434
Unseen-3	1	0.5
Unseen-4	0.5	0.592

Table 4.4: GRU based encoder for unseen questions BLEU SCORE.
Source: Seq2Seq model for education chatbot [69]

As per the above table data [69] it is evident that the seq2seq model with GRU neural networks performed well when it comes to the seen questions. Their model also performed reasonably well with the unseen questions as well except when the new vocabulary was used. This model is evident that if created a special dataset based on the case study and the test cases, their chances that it will perform well when it comes to the interview related questions.

This model can answer our research question no 3 “Can this sequence-to-sequence based chatbots models be used for question answering?”. This model has shown good BLEU score when countering the unseen questions. Only thing that will be needed is the use of proper dataset creation and better test and train planning.

4.5 Summary of Findings and experiments

In this chapter we started with addressing one of the main aspects of the study by providing all the findings we had based on the technical research survey we conducted in chapter 2. Where we address in first and second research question. We provided an overview to the audience that what approaches (rule-based, script based) were adopted in the early era of the chatbot technology (1950 to 2020) **Table 4.1.** Later with the second phase of the technical survey our aim was to provide best deep learning-based model that can be used to form advance chatbot systems. Based on the research we submitted our finding in the Table 4.2.

Based on the findings in Table 4.2 we were able to be formed basic sequence to sequence model with bidirectional LSTM and attention mechanism with couple of other ingredients required to form a basic model that later can be trimmed and used for more advance purposes.

Then we conducted some basic experiments with two identical models but different in encoder decoder formation. Model 1 was formed based on the finding we have documented above (Using Bidirectional LSTM RNN) and model 2 was formed with same sequence to sequence model but with Gated recurrent Units. The performance of both models is similar as they have the same settings but the model two that was prepared with different settings and special datasets based on the different test case [8, 69] had better performance in question answer-based scenarios.

4.6 Limitations

Based on the research we conducted, and the findings documented above, it is obvious that these smart chatbot systems are far from perfect. The truth is these are machines not the actual human beings. They can learn from the data and the vocabulary provided in the data to some extent but not be the alternative. However, we can equip them with the better training techniques and based on well transformed data, so they can generate more accurate responses. The use of these chatbots in terms of interview-based question answers will certainly rely on the quality of the dataset provided and its formation according to the domain. The recent research work on sentiment analysis and emotion detection is very impressive such as ECM machine with emotional quotient [2] and XiaoIce an emphatic social Chatbot by Microsoft China [14] are the step forward into right direction but it will take some time. Existing models and the model used in this thesis also lacked the ability to maintain the long conversations, but they were reasonable with short questions.

The training process of these models is hard task especially when we don't have specific domain related rich datasets. Furthermore, the use of the encoder-decoder model also shows some consequences such as the way it handles sequences of variable length and vocabulary size, it becomes slow during handling huge vocabulary size. Also, the resources and computational power requires for the training these models are huge concern for many users as it require certain level of computational power that every individual don't have access too.

CHAPTER 5. CONCLUSION

5 CONCLUSION

5.1 Summary

The focal point of this thesis was to conduct a technical research survey to find the resources that can help to build a smart AI based chatbot system with an ability to perform coherent meaningful conversation during human interaction. Later that model can be used to conduct interviews with children and possibly for the training purposes. The goal of this study was to find answers to the pre-defined research questions and our findings are aligned.

Based on the finding we had after the extensive technical survey conducted in phase-2 where we studied all the developments related to the chatbot technology from 1950 to 2022, we were able to demonstrate that how the chatbot idea, for the matter of fact how the machine learning concept introduced and how it led to the development of chatbot in 20th century using skilled based and script-based frameworks. The finding presented in Table 2.2 and Table 4.2 we were convinced that sequence to sequence based encoder-decoder method is the industry accepted method to build smart chatbot system. As majority of the advance chatbots were built on Sequence-to-sequence based encoder decoder model. Only difference is the use of neural networks in encoder. Some uses the long-short term memory neural networks some gated recurrent unit neural network.

Later, we did experiment on two identical encoder-decoder sequences to sequence models but with different in neural network formation. Model 1 was formed based on the settings documented in this study using bi-direction neural network and Model 2 was developed by Palasundram [69] where it has GRU neural network. The results with both models were identical but the model 2 [69] had better results documented when it was formed with better dataset formations. The results conducted by the author were interesting while dealing with unseen question categories during interview questions. With further development it is possible to use these basic setups for more advance communications with better test case formation and using emotion classifiers.

5.2 Future Work

Chatbot technology is changing rapidly as now organizations are taking chatbots as an alternative to many existing static informative jobs as well as multifunction assisting entities. That increase in the business aspect of chatbot technology has led the additional resources into the research and development. With each passing day new frameworks and models are being introduced based on advance machine learning techniques. We have used encoding decoding based attention mechanism using long, short-term cells in the development of the chatbot, we can also replace this attention model Bahdanau [50] with the better attention model proposed by M. Luong [69]. Here we will mention some exited work that is being done by researchers to make the chatbot system more engaging by working on, attention, sympathy and emotion factor.

5.2.1 Emotion Classifier

In this study we were optimistic about the use of emotion classifier presented by (Zhou et al.) [2] where they proposed an "ECM" ‘Emotional chatting machine’ by introducing 3 new mechanisms for the development of ECM using sequence to sequence framework. Due to lack of the data available and complex code provided by the authors for me it was extremely hard to replicate the experiments in the lab as the initial project was developed with highly renowned researcher from Microsoft China. Initially they started by creating emotion-oriented categories in the dataset, then compare that data with the actual communication with human basing on the change in emotion state of the person. And lastly Vocabulary for emotion expression. As experiment suggested, that the results were encouraging, the ‘ECM’ system was able to respond to the questions more consistently and structurally, not only content-wise but also the response was showing the factor of emotion by using the wording from the vocabulary.

5.2.2 Persona based Model

Having in mind the parent project I believe that model based on human persona could be more reliable during the child interviews as it gives the conversation more realistic aspect. A human persona-based encoding decoding evaluation method proposed by Jiwei Li et al. [3] aimed to encode the participant's persona and try to figure the characteristics of the person such as his speaking signature and person’s psyche which reflect in his conversation style. Human judges were presented to oversee the results of the experiments. This persona-based model was supposed to overcome the problems which the data-driven model had such as consistency in the results. After the research and experiments, this approach proved more consistent than seq2seq based models but still the results are not extraordinary. As the scope of the paper was not handling the other characteristics of human nature just like emotions, psyche, and mood. Still, results were consistent, but the author suggested future work to make this model ideal.

5.2.3 Interview based Training

Sequence to sequence framework can be trained in a way that could be used for interview related scenarios. In a research work presented by Palasundram [69] a chatbot for education was developed based on Seq2Seq framework using (Bidirectional Encoder integrated GRU with beam search). There were two separate types of questions seen and unseen were used to determine the results. The end results were better when it comes to direct and seen questions. In future for making the chatbot for interview-based training we can prepare a dataset and train the model as per this suggested method. This chatbot framework could be an answer to one of our research questions “Can advance chatbots be a vital component for interview training?”. [69]

References

[1] Vinyals, Oriol, and Quoc Le. “A Neural Conversational Model.” ICML Deep Learning Workshop 2015, 19 June 2015, https://doi.org/https://doi.org/10.48550/arXiv.1506.05869.

[2] Zhou, H., M. Huang, T. Zhang, X. Zhu, and B. Liu. “Emotional Chatting Machine: Emotional Conversation Generation With Internal and External Memory”. Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, Apr. 2018, https://ojs.aaai.org/index.php/AAAI/article/view/11325.

[3] Li, Jiwei, et al. "A persona-based neural conversation model." arXiv preprint arXiv:1603.06155 (2016).

[4] Sutoyo, R., Chowanda, A., Kurniati, A., Wongso, R. Designing an emotionally realistic chatbot framework to enhance its believability with aiml and information states. Procedia Computer Science ;157:621–628. 2019.

[5] Pamungkas E. Wahyu. “Emotionally-Aware Chatbots: A Survey. In Proceedings of ACM Conference (Conference’17)”. ACM, New York, NY, USA, 8 pages, 2018. DOI: https://doi.org/10.1145/1122445.1122456

[6] Przegalinska, Aleksandra, Ciechanowski Leon , Anna Stroz, Peter Gloor “In Bot We Trust: A New Methodology of Chatbot Performance Measures.” Business Horizons, vol. 62, no. 6, 2019, pp. 785–797., 2019. DOI:10.1016/j.bushor.2019.08.005.

[7] Tiha, Anjana. “Intelligent Chatbot using Deep Learning”. University of Memphis, Memphis, Tennessee, USA. 10.13140/RG.2.2.14006.75841. 2018. Source: https://www.researchgate.net/publication/32858298

[8] Yakkundi, Sshriniket, Vanjare Amey, wavhal Vinay, Patankar Shreya. “Interactive Interview Chatbot”. International Research Journal of Engineering and Technology,” 06(04). 2019.

[9] Romero, Miriam Romero Casadevante Cristina Montoro, Montoro Helena. “How To Create A Psychologist-Chatbot”. Universidad Autónoma de Madrid. C/Iván Pavlov, nº 6. 28049 Madrid. España. 2020. DOI: 10.23923/pap.psicol2020.2920.

[10] Gulenko, Iwan. “Chatbot for IT Security Training: Using Motivational Interviewing to Improve Security Behaviour.” The Technical University of Munich, Munich, Germany. 2019. Source: https://docplayer.net/1261298-Chatbot-for-it-security-training-using-motivational-interviewing-to-improve-security-behaviour.html

[11] Sutskever, I., Vinyals, O, V. Le, Q. “Sequence to Sequence Learning with Neural Networks.” Google Research. 2014. doi: arXiv:1409.3215 [cs.CL].

[12] Kostadinov, Simeon. “Understanding the Encoder-Decoder Sequence to Sequence Model.” 2019. Source: https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346.

[13] Wu, Y., Schuster, M., V. Le Quoc, M. Norouzi & Chen, Z. et al. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. Google research. 2016 ArXiv:1609.08144v2 [Cs.CL].

[14] Li Zhou, J. Gao, Di Li, and Heung-Yeung Shum. “The Design and Implementation of XiaoIce, an Empathetic Social Chatbot”. 2018. arXiv preprint arXiv:1812.08989.

[15] Hasan, Mohammad & Yu, Hongnian. “Innovative developments in HCI and future trends”. International Journal of Automation and Computing. 14. 2016. arXiv: 10.1007/s11633-016-1039-6.

[16] Juniper research. “How Chatbots Will Facilitate the Omnichannel Experience.” Whitepapers, 15 May 2020, https://www.juniperresearch.com/document-library/white-papers/how-chatbots-will-facilitate-the-omnichannel.

[17] Cui, Lei & Huang, Shaohan & Wei, Furu & Tan, Chuanqi & Duan, Chaoqun & Zhou, Ming. “SuperAgent: A Customer Service Chatbot for E-commerce”. 2017. Websites. 97-102. 10.18653/v1/P17-4017.

[18] Go, Eun & Sundar, S. Shyam. “Humanizing Chatbots: The effects of visual, identity and conversational cues on humanness perceptions”. Computers in Human Behavior. 97. 2019. arXiv: 10.1016/j.chb.2019.01.020.

[19] M. Powell, B. Guadagno, M. Benson. “Improving child investigative interviewer performance through computer-based learning activities,” Policing and Society, vol. 26, pp. 365–374. 2016.

[20] C. S. Widom. “Longterm consequences of child maltreatment,” in Handbook of child maltreatment. Springer, pp. 225–247. 2017.

[21] L Dixon, D. F. Perkins, C. Hamilton-Giachritsis, L. A. Craig. “The Wiley Handbook of what Works in Child Maltreatment: An Evidence based Approach to Assessment and Intervention in Child Protection. john Wiley & Sons.” 2017.

[22] World health organization. “European report in preventing child maltreatment”. 2013

[23] Turing, A. M. Computing machinery and intelligence. Mind, pp. 433–460, 1950.

[24] Hossein Hassani. “Research Methods in Computer Science: The Challenges and Issues”. arXiv:1703.04080 [cs] (Mar. 12, 2017). 2017. URL: http: //arxiv.org/abs/1703.04080 (visited on 02/21/2019).

[25] D. E. Comer, D. Gries, M. C. Mulder, A. Tucker, A. J. Turner, and P. R. Young, “Computing as a discipline. Communications of the ACM”, 32(1):9–23, 1989.

[26] Sharanan Kulam: "Time-Series Classification with Uni-Dimensional Convolutional Neural Networks - An Experimental Comparison with Long Short-Term Memory Networks", November 2019 [Thesis]

[27] Edvarda Regine Winlund Eriksen: "A Machine Learning Approach To Improve Consistency In User-Driven Medical Image Analysis", May 2019 [Thesis].

[28] Håkon Kvale Stensland: "Processing Multimedia Workloads on Heterogeneous Multicore Architectures", February 2015 [Thesis]

[29] Saygin, A. P.; Cicekli, I.; Akman, V. "Turing Test: 50 Years Later" (PDF), Minds and Machines, 10 (4): 463–518. 2000.

[30] Weizenbaum, Joseph . "ELIZA--A Computer Program for the Study of Natural Language Communication Between Man and Machine" (PDF). Communications of the ACM. 9: 36–35 – via universelle-automation. 1966.

[31] Cerf, V. "PARRY encounters the DOCTOR". 1973. DOI:10.17487/rfc0439

[32] Carpenter, R. "Jabberwacky, the smart chatbot". Source: http://www.jabberwacky.com/j2about. 1988. Retrieved September 20, 2020.

[33] Wallace R.S. "The Anatomy of A.L.I.C.E. In: Epstein R., Roberts G., Beber G. Parsing the Turing Test." Springer, Dordrecht. 2009. DOI: https://doi.org/10.1007/978-1-4020-6710-5_13

[34] Naraine, R. "ActiveBuddy's Patent Win Riles IM Bot Developers". 2009. Retrieved September 20, 2020, from http://www.internetnews.com/bus-news/article.php/1446781/ActiveBuddys+Patent+Win+Riles+IM+Bot+Developers.htm

[35] Mavridis, Nikolaos & Petychakis, Michael. "Human-like memory systems for interactive robots: Desiderata and two case studies utilizing grounded situation models and online social networking. 46-51". 2010.

[36] N. Mavridis, D. Roy, "Grounded situation models for robots: Where words and percepts meet", in Proc. of IEEE IROS. 2006

[37] N. Mavridis, C. Datta, Emami S, Tanoto A, BenAbdelkader C, Tabie T. "Facebots: Social robots utilizing and publishing social information in Facebook", IEEE HRI. 2009

[38] Paiva Ana. "Empathy in Social Agents" In International Journal of Virtual Reality, 10, 1-4. 2011. DOI, https://doi.org/10.20870/IJVR.2011.10.1.2794

[39] S. Rodrigues, S. Mascarenhas, J. Dias and A. Paiva, "I can feel it too! Emergent empathic reactions between synthetic characters" in Affective Computing and Intelligent Interaction Conference, ACII, IEEE Press, 2009.

[40] Rui Zhang, Zhenyu Wang, and Dongcheng Mai. "Building emotional conversation systems using multi-task Seq2Seq learning". In National CCF Conference on Natural Language Processing and Chinese Computing. Springer, 612–621. 2017

[41] Xiao Sun, Xinmiao Chen, Zhengmeng Pei, & Fuji Ren. E"motional Human Machine Conversation Generation Based on SeqGAN". In 2018 First Asian Conference on Affective Computing and Intelligent Interaction "ACII Asia" bejing. 2018.

[42] Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, Satoshi Nakamura. "Positive Emotion Elicitation in Chat-Based Dialogue Systems". In IEEE/ACM Transactions on Audio, Speech, and Language Processing 27, 4 (2019). 2019.

[43] Minlie Huang, Zuoxian Ye, and Hao Zhou, "Shared Task: Emotion Generation Challenge. Overview of the NLPCC 2017". In National CCF Conference on Natural Language Processing and Chinese Computing. Springer, 926–936. 2017.

[44] P. Mazare, S. Humeau, M. Raison, A. Bordes. "Training Millions of Personalized Dialogue Agents, Facebook". 2018. arXiv:1809.01984v1 [cs.CL].

[45] Ritter, A., Cherry, C., Dolan, B. “Unsupervised modeling of Twitter conversations. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics; p. 172–180.” 2010.

[46] Shang, L., Lu, Z., Li, H.. ”Neural responding machine for short-text conversation”. arXiv preprint arXiv:150302364. 2015.

[47] Danescu-Niculescu-Mizil, C., Lee, L. “Chameleons in imagined conversations: A new approach to understanding the coordination of linguistic style in dialogs”. In: Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics. Association for Computational Linguistics; 2011, p. 76–87. 2011.

[48] Luong minh thang, Pham hieu, Manning Christopher. “Effective Approaches to Attention-based Neural Machine Translation”. Standford Computational dept. 2015. DOI: https://doi.org/10.48550/arXiv.1508.04025

[49] Luong minh thang, Pham hieu, Manning Christopher. “Neural Machine Translation (seq2seq) Tutorial”. Google research, https://github.com/tensorflow/nmt#background-on-the-attention-mechanism , 2015.

[50] Bahdanau Dzmitry, Cho Kyunghyun , Bengio Yoshua. “Neural Machine Translation by Jointly Learning to Align and Translate”. 2015. DOI: https://doi.org/10.48550/arXiv.1409.0473.

[51] Zhou Hao , Huang Minlie , Zhang Tianyang , Xiaoyan Zhu, Liu Bing. “Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory” In National Laboratory for Information Science and Technology. 2018. DOI: https://doi.org/10.48550/arXiv.1704.01074.

[52] Devlin Jacob , Chang Ming-Wei , Lee Kenton , Toutanova Kristina. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. Google Research. 2019. DOI: https://doi.org/10.48550/arXiv.1810.04805

[53] M Zahangir Alom et al. “The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches”. 2018. arXiv:1803.01164 [cs]. URL: http://arxiv.org/abs/1803.01164 (visited on 01/17/2019).

[54] G James et al. “An introduction to statistical learning: with applications in R. Springer texts in statistics 103”. New York: Springer, 2013. 426 pp. ISBN: 978-1-4614- 7137-0. 2013.

[55] J Hopfield. Hopfield J. J. “ Neural networks and physical systems with emergent collective computational abilities”. In Proceedings of the National Academy of Sciences of the United States of America, 79(8), 2554–2558. 1982. DOI: https://doi.org/10.1073/pnas.79.8.2554.

[56] Cho Knghyun, Bahdanau Dzmitry, Schwenk holger, Bengio yoshua. “Learning Phrase Representations using RNN Encoder and Decoder for Statistical Machine Translation”. 2014. URL: DOI: https://arxiv.org/pdf/1406.1078.pdf

[57] R Subburam et al., Neural Architectures for Relation Extraction Within and Across Sentence Boundaries in Natural Language Text. 10.13140/RG.2.2.36649.70245. 2018.

[58] LeCun, Bengio, and G. Hinton. “A recurrent neural network and the unfolding in time of

the computation involved in its forward computation”. 2015. Doi: 10.1038/nature14539.

[59] R. Subburam et al., “Neural Architectures for Relation Extraction Within and Across Sentence Boundaries in Natural Language” Text. 10.13140/RG.2.2.36649.70245. 2018.

[60] Richard Krisztian. “Deep Learning Based Chatbot Models Scientific". In Department Of Automation And Applied Informatic. Students’associations Report”. 2017. URL: https://tdk.bme.hu/VIK/DownloadPaper/asdad.

[61] Luong minh thang, Pham hieu, Manning Christopher. “). Effective Approaches to Attention-based Neural Machine Translation”. 2015. arXiv: https://arxiv.org/pdf/1508.04025.pdf.

[62] Bahdanau Dzmitry, Cho Kyunghyun , Bengio Yoshua. “Neural Machine Translation by Jointly Learning to Align and Translate”. 2014. Source: https://www.semanticscholar.org/paper/Neural-Machine-Translation-by-Jointly-Learning-toBahdanauCho/fa72afa9b2cbc8f0d7b05d52548906610ffbb9c5/figure/1.

[63] Sutskever, I., Vinyals, et al., “Sequence to sequence learning with neural networks”. In Advances in Neural Information Processing Systems pp. 3104–3112. 2014.

[64] Suresh Raj, Vivek. I. “Performance of Seq2Seq learning Chatbot with Attention layer in Encoder decoder model”. 2021. Source: 10.13140/RG.2.2.33355.92961.

[65] p Hochreiter, Jurgen Schmidhuber. Long Short-Term Memory, (LSTM). In Neural Computation 9(8):1735 – 1780. 1997. Sorce: www.bioinf.jku.at/publications/older/2604.pdf.

[66] Bukka, Sandeep & Gupta, Rachit & Magee, Allan & Jaiman, Rajeev. “Assessment of unsteady flow predictions using hybrid deep learning based reduced order models”. 2020.

[67] Shao Chenze , Feng Yang , Chen Xilin. “Greedy Search with Probabilistic N-gram Matching for Neural Machine Translation”. 2018. DOI: https://doi.org/10.48550/arXiv.1809.03132

[68] Luong minh thang, Pham hieu, Manning Christopher. “). Effective Approaches to Attention-based Neural Machine Translation”. 2016. DOI: https://doi.org/10.48550/arXiv.1508.04025.

[69] Palasundram Kulothunkan , M Nurfadhlina, kasmiran Khairul, azeem azreen. (2020), “Sequence to Sequence Model Performance for Education Chatbot”. In iJET. DOI: https://doi.org/10.3991/ijet.v14i24.12187.

[70] Peters Matthew , Neumann Mark, Iyyer Mohit ,Gardner Matt , Clark Christopher, Lee Kenton, Zett Luke. “Deep contextualized word representations” .2018. DOI: doi.org/10.48550/arXiv.1802.05365

[71] Liu Yinhan ,Ott Myle ,Goyal Naman ,Du Jingfei ,Joshi Mandar ,Chen Danqi ,Levy Omer ,Lewis Mike ,Zettlemoyer Luke , Stoyanov Veselin. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. 2019. DOI: https://doi.org/10.48550/arXiv.1907.11692

[72] Lan Zhenzhong ,Chen Mingda ,Goodman Sebastian ,Gimpel Kevin ,Sharma Piyush , Soricut Radu. “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations”._2019. DOI: https://doi.org/10.48550/arXiv.1909.11942

[73] Yang Zhilin ,Dai Zihang ,Yang Yiming, Carbonell Jaime ,Salakhutdinov Ruslan , Le Quoc V. "XLNet: Generalized Autoregressive Pretraining for Language Understanding". 2019. DOI: https://doi.org/10.48550/arXiv.1906.08237

[74] Jovanović Mlađan, Baez Marcos, Casati Fabio. "Chatbots as Conversational Healthcare Services," in *IEEE Internet Computing*, vol. 25, no. 3, pp. 44-51. June 2021, DOI: 10.1109/MIC.2020.3037151.

[75] Følstad, A., Araujo, T., Law, E.LC. et al. Future directions for chatbot research: an interdisciplinary research agenda. Computing 103, 2915–2942 (2021). 2021. DOI: https://doi.org/10.1007/s00607-021-01016-7

[76] Qi Peng ,Huang Jing ,Wu Youzheng , He Xiaodong ,Zhou Bowen. "Conversational AI Systems for Social Good: Opportunities and Challenges". 2021. DOI: https://doi.org/10.48550/arXiv.2105.06457

[77] Chung Junyoung ,Gulcehre Caglar , Cho KyungHyun, Bengio Yoshua "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". 2014. DOI: https://doi.org/10.48550/arXiv.1412.3555

[78] Malasundram, Kulothunkan & Sharef, Nurfadhlina & Nasharuddin, Nurul Amelina & Kasmiran, Khairul & Azman, Azreen. "Sequence to Sequence Model Performance for Education Chatbot". In International Journal of Emerging Technologies in Learning (iJET). 14. 56. 10.3991. 2019. arXiv.1412.3555

[79] Abonia Sojasingarayar. "Seq2Seq AI Chatbot with Attention Mechanism". In IA Universit-GEMA. 2020. DOI: doi.org/10.48550/arXiv.2006.02767