Advancement in Communication using Natural Language based VideoBot System

1st Flewin Dsouza
Dept. of AI
G. H. Raisoni College of Engineering
Nagpur, India
dsouza_flewin.ai@ghrce.raisoni.net

2nd Rushikesh Shaharao
Dept. of AI
G. H. Raisoni College of Engineering
Nagpur, India
shaharao_rushikesh.ai@ghrce.raisoni.net

3rd Yashsingh Thakur
Dept. of AI
G. H. Raisoni College of Engineering
Nagpur, India
thakur.yashsingh.ai@ghrce.raisoni.net

4th Pranav Agwan
Dept. of AI
G. H. Raisoni College of Engineering
Nagpur, India
agwan.pranav.ai@ghrce.raisoni.net

5th Gopal Sakarkar
D Y Patil Institute of Master of Computer
Applications Managament,Akurdi, Pune, India
g.sakarkar@gmail.com

6th Piyush Gupta
Assist 2 Path Tech Pvt. Ltd.
New Delhi, India
piyush@strideahead.in

***Abstract*—Natural Language Processing is a subset of Artificial Intelligence, which focuses more on the natural language communication and speech recognition. After evaluation of AI and its sub branches, the automated answering system or now it's called as chatbot is a very popular and widely used application. One of the limitations of this application is that it is text based. This application is not so effective when we want to develop a dynamic system. In this paper, authors have proposed a VideoBot application, which is a more effective way of communication while interacting with end users. While simple chatbots, which don't have any emotional attachment with end users, as compared to this Videobot have more effectively connected with end-users, as it has videos with emotional expressions.**

***Index Terms*—Natural Language Processing, Chatbot, Videobot, Artificial Intelligence.**

I. Introduction

The world is witnessing a booming growth in natural language processing and its applications in numerous areas. Natural Language Processing or NLP is a branch of artificial intelligence (AI) that concerns with computers to communicate with humans in their own language and perform other linguistic-oriented tasks like translation, information extraction, summarization, question answering, speech recognition, auto-correct and many more [1]. In this paper, the term Conversational AI is emphasized to introduce a new way of interacting with people with the help of AI. A conversational Agent or the dialogue system is a part of AI that deals with text-based or speech-based data. It uses various NLP techniques to understand the human language and respond accordingly. Conversational AI also helps to simulate human conversations. They are handled by conversational agents. An intelligent conversational agent would be any computer program which can mimic human-like conversations such that they are as human as possible [2]. Voice-enabled personal assistants such as Amazon Alexa, Google Assistant, and Apple Siri are popular with consumers because of their intuitive usage and similarities in human communication. However, most of these assistants are designed for short task-oriented dialogues, such as playing music or requesting information, rather than long free-form conversations. Achieving sustainable, consistent and engaging dialogue is the next frontier of conversational AI. This allows artificial intelligence agents to have natural and enjoyable conversations with human interlocutors [3].

In today's scenario, conversational AI agents seem a natural fit for counseling services. A wide range of applications in conversational AI with NLP has tremendous potential because of advancements in research topics such as semantic similarity, text classification, emotion recognition and transformer NLP models which is a deep learning model that uses BERT for language understanding [4]. BERT is a transformer-based model. A transformer is an effective architecture for text modeling and has been an essential component in many state-of-the-art NLP models like BERT. A transformer is an effective text modelling architecture that has been used in several state-of-the-art NLP models, such as BERT. The typical Transformer must calculate a dense self-attention matrix based on interactions between each pair of tokens in text, with computing complexity proportional to text length squared [5]. While dealing with the transformer model, we will encounter the term cosine similarity. In information retrieval and related research, cosine similarity is an extensively used statistic. A written document or a sentence is represented as a vector of terms in this metric. The cosine value between two documents' term vectors can be calculated using this model to determine the similarity between two documents or sentences. This metric's implementation can be done on any two texts (sentence, paragraph, or whole document). Similarity values between user queries and documents are ranked from highest to lowest in the search engine case. A higher similarity score between a document's term vector and a query's term vector indicates that the document and query are more relevant to each other

Thank you Mr. Piyush Gupta for their guidance and providing the resources to carry out the project.

978-1-6654-9291-1/22/$31.00 ©2022 IEEE

[6].

II. Literature Review

Conversational Agents (CA) have gained a lot of attention in recent years and when we see how they are evolving, it is certain that as a need of the hour many more CAs will be deployed in the market. Looking at the trends, CAs are supposedly the driving factor in the evolving fields like healthcare, education, business and various forms of customer services [4]. It is generally more valuable if problems of the people or the customers are resolved and delivered as quickly as possible with efficiency [7].

In this section, various outlooks of conversational agents and their use cases are highlighted. A paper on human-computer communication where there is contextual communication between the user and the artificial conversational entity. The dialogues become so realistic that it becomes difficult and incomprehensible to perceive the presence of an actual person or the CA [8]. The author of the paper Conversational ai: The science behind the Alexa prize briefs about the traditional and new approaches/methods to develop Conversational AI agents. The methods or approaches for developing the Dialogue Management Systems that are discussed in this paper include Switch Statements, Finite State Machines, Machine Learning based Approaches, Deep Reinforcement Learning based Approaches, Belief based DM Systems. Disadvantages of widely used methods have been reviewed too [9]. Thereafter, in “From Eliza to Siri and Beyond authors have introduced a virtual butler, Edgar Smith, which can be found at Monserrate Palace, in Sintra, Portugal, a knowledge-based conversational agent, with a domain of expertise in the palace. In this paper, we get to know how a virtual assistant considers several measures for addressing the problems the user asks and chooses an appropriate answer from the set of candidate answers. This paper illustrates the simple approach of the proof-of-concept chatbot which also deals with out-of-domain interactions [10]. In the paper, Survey on chatbot design techniques in speech conversation systems, the similarities and differences in particular with the Loebner prize-winning Chatbots are discussed. Also, how these chatbots are limited to particular applications and how over time how human-computer speech interaction strategies and techniques work [11]. A Counseling Service Using Emotional Response Generation, a chat assistant platform where natural language processing along with clinical psychiatric analysis is taken into consideration to recognize and monitor human emotions. This conversation model provides the user with psychiatric counseling services for people who need proper professional mental health advice [12]. In the proposed paper, it shows how the paraphrase detection works in the Malayalam language. Paraphrasing means suppose you have a sentence and then you create a new sentence that has the same meaning but has different words. The main use of paraphrasing is information retrieval systems, QA systems(question-answer), plagiarism checkers, etc. For computing, semantic similarity cosine similarity is used. However, other similarity scores are also compared. After calculating the similarity score system will decide if two sentences are paraphrased or not. A total of 900 pairs of Malayalam sentences are used for classification (P or NP), which gives an accuracy of 0.8. In conclusion, paraphrase detection is performed in the Malayalam language. For finding similarity score cosine similarity is used and not Jaccard similarity to get a higher F1 score [2]. The research in the paper Refining Word Embeddings Using Intensity Scores for Sentiment Analysis examines the word embedding techniques that are used in Natural Language Processing for different types of tasks such as sentiment analysis. Word embedding means representing words in a form of low-dimensional vectors. It shows how traditional word embedding techniques such as word2vec, GloVe failed to retrieve sentiment information. On the other hand, new sentiment embedding techniques will help classify the sentiment polarity. For example, good and bad [13].

III. Related Work

The visual representation of such robots can boost audience retention and speech understanding, just as we discussed with bots like Siri, Alexa, and Google. Consider the robot as a two-dimensional representation of ourselves; this could allow us to be present in places without physically being there. A person’s voice can be recreated with only a few seconds of their original vocal input using voice cloning. A module that accepts a voice clip and text input and synthesises a speech in the person’s voice that sounds exactly like them can be constructed by combining the Text-to-Speech converter with voice cloning [14]. A visual clone or digital human can be made that looks and expresses just like the person, with only one photograph of them. This digital human can be made to deliver the speech that was given as text input in their own synthesized voice from the Speech module. Here comes the concept of a video-bot. A video-bot is a video of a person that a chatbot displays to website visitors automatically. Video bots can help with personalisation and conversion automation by encouraging visitors to take action, whether on a website or in an app. Customers would communicate with a video bot using pre-recorded video with a human and the functionality of a chatbot [11]. VideoBot follows three principal parts of the architecture of conversational AI

- Natural Language Understanding (NLU). This step includes creating word embeddings.
- Finding a similar question in the database.
- Response Generation.

By using the above architecture VideoBot will understand the meaning of users’ queries and then it will show the response in video format.

In this paper, the proposed VideoBot is a conversational AI system built for career counseling. We can say that offline mentoring is more impactful as students can connect with the mentor more easily, but in online mode i.e., a conventional chatbot it’s very hard to make an impact on students and also, it’s hard for students to connect with mentors. We need a way to close this gap. As a result, we created this VideoBot that responds to users' questions. Designed to closely resemble how a human would interact with a conversational partner. Our main objectives are:

- To provide a better, emotionally rich, trustworthy, platform for career guidance in the online mode
- Solving career-related queries with face-to-face interaction with experts in the field.
- To create a data-driven mentorship platform, where mentors share their experience of landing a dream career.

IV. Methodology

The main goal of this research is to develop an effective communication system for students and mentors. This system will help to make the process of mentorship, career guidance, counseling, etc. more effective. By using videobot system, creators or mentors will create videos of their expertise areas. Users can ask questions on the platform related to the topic of the video and videobot will display an answer on the screen. VideoBot coincides with chatbot, but videobot has a video interface that makes videobot more effective and engaging. Components of the videobot are shown in figure 1.

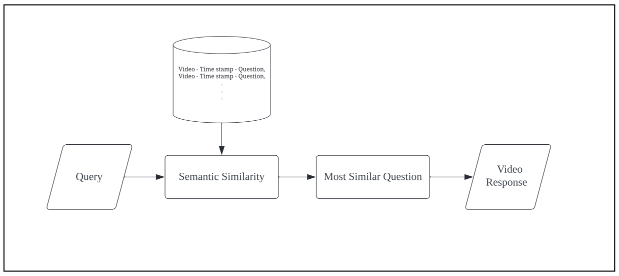

Video - Time stamp - Question,
Video - Time stamp - Question,
Query
Semantic Similarity
Most Similar Question
Video Response

Fig. 1. Architecture of VideoBot

A. Taking users query as an input

Users can ask their question either in text format or speech format. If the user asks a query in speech format it gets converted into text format. Speech to text helps to convert the speech into text format. Converting speech to text is important because semantic similarity function requires input in text format.

B. Database

Database comprises data like videos (links to videos or video files), timestamp and questions tagged on that particular timestamp of the video. Videobot has an interface where creators can add their video along with tags as questions to different timestamps. Supposedly, a question does not match with any relevant video playback in the database, a video with a message stating that the asked question could not be answered as its out of its domain knowledge.

C. Semantic Similarity

To understand the user's query and give the accurate answer, it is important to find a tagged question similar to the user's query in our database.

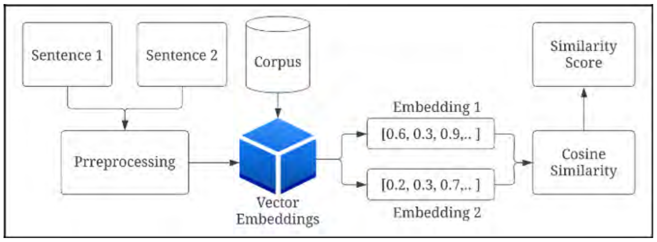


Fig. 2. Working of Semantic Similarity

This process includes components such as pre-processing, creating embeddings and similarity measures as shown in figure 2.

1) Text Preprocessing: Queries by users and questions tagged by creators may have some noise in the form of extra punctuations, inputs in different cases, extra spaces and many more unwanted things. Removing those noises from the text data is known as text preprocessing. To prepare the text data for creating embeddings it is necessary to preprocess the text data.

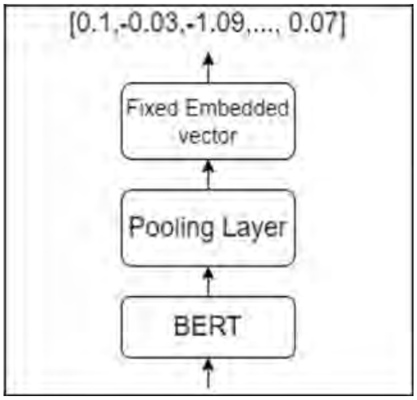


Fig. 3. Network Architecture - Sentence Transformer. How do embedding layers work?

a) Embedding: Embedding is the process of converting high-dimensional text data into low-dimensional data in the form of a vector in such a way that the two are semantically similar and the context information is not lost [15]. For creating embedding there are many BERT based pre-trained models available on Hugging face.

b) Sentence Transformers: Sentence-Transformers is a Python framework that allows you to create cutting-edge sentence, text, and image embedding. Video bot uses it as a NLU for understanding the meaning or context of the user input. Sentence transformers use Siamese BERT-Networks which contextualize words of sentence efficiently, BERT converts the given vector to embedding vector. Machines cannot understand words/texts they only understand the numbers therefore, the term word embedding has come, Word embedding are a way to represent words and the sentences in a numerical manner [16].

c) Network Architecture: We want to map a variable-length input text to a fixed-size dense vector for sentence /

text embedding. The basic network topology of the sentence transformers is shown in Figure 3.

2) Cosine Similarity: Following the creation of embedding, the semantic similarity between the two embedding vectors is measured. Cosine similarity is a tool for determining how similar two phrases (vectors) are, regardless of their size. The cosine of the angle between two sentences (vectors) is used to determine their similarity. Cosine similarity is calculated mathematically by multiplying the dot products of the two vectors by their magnitude as shown in the equation 1.

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

The smaller the angle between the two vectors, the lower the degree of resemblance between two sentences. Conversely, the smaller the angle between the two vectors, the greater the degree of resemblance between the two sentences [17]. Cosine similarity gives a similarity score between those two sentences. Similar process is used to find the similar question in the database with the user's query. The representation is shown as in fig. 4.

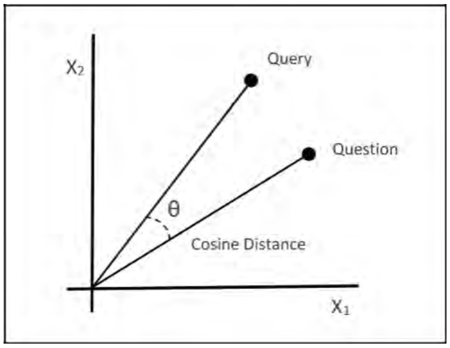


Fig. 4. Cosine Similarity - Representation

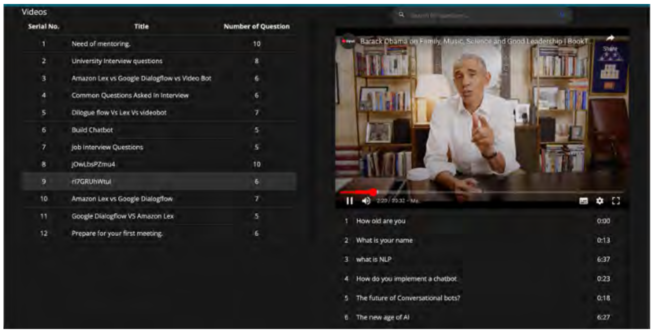
D. Similar Question and Video Response Generation

After getting similarity scores of the user's query with all the questions tagged to the videos in the database, the question which will have the highest similarity score will be treated as the most similar question.

Each question in the database is mapped with corresponding timestamps of the video. Response will be generated by playing video on the interface of videobot from the timestamp of the most similar question.

V. Results and Discussion

We need to understand that the proposed model of an interactive video aided bot is not a fully-fledged alternative to a human agent. At times, the idea of delivering the model of a question answering system is appreciated, but when working on the model and while approaching the problem of solving the users' queries, there are other implications to be confronted first. The application uses pre-recorded videos chiefly, to work and respond to the user. So, when setting up the application on a large scale, there is a fundamental requirement for resources. Here the resources being said are in aspects of storage, computation as well as time. The application runs on a deep learning model which makes it quite difficult to operate. A need for recurrent optimization and maintenance is demanded to make the implementation of such a system sensible in terms of feasibility.


Videos
Barack Obama on Family, Music, Science and Good Leadership | BookT...
Share

Serial No.	Title	Number of Question
1	Need of mentoring.	10
2	University Interview questions	8
3	Amazon Lex vs Google Dialogflow vs Video Bot	6
4	Common Questions Asked In Interview	6
5	Dilogue flow Vs Lex Vs videobot	7
6	Build Chatbot	5
7	Job Interview Questions	5
8	jOwLbsPZmu4	10
9	rI7GRUhWtuI	6
10	Amazon Lex vs Google Dialogflow	7
11	Google Dialogflow VS Amazon Lex	5
12	Prepare for your first meeting.	6

1	How old are you	0:00
2	What is your name	0:13
3	what is NLP	6:37
4	How do you implement a chatbot	0:23
5	The future of Conversational bots?	0:18
6	The new age of AI	6:27

Fig. 5. A Snapshot of the application, outlining an interactive video interface.

Compared with the existing models of videobot we are first understanding the intend of the user using NLP techniques which includes text processing and cosine similairty so that we can get an accurate output and the response is generated based on this intend of the user. Since the area is specific, users will get a clear-cut video response based on their query. In our model, user can ask their question in both text or speech format. Since the area of application is specific, users will get an output in less time but if the application is set on a large scale with a vast amount of data, the response time will increase. Since we are using NLP techniques like cosine similarity it will take time to compare the user's query with each sentence present in the database as there are too many queries stored in it. Studies have shown that how an inclusion of audio and video along with text have a drastic impact on the solving the problems on advising, mentoring and counselling. The known gap of potential problems can be bridged by means of using digital technology [18].

VI. Conclusion

In our work, we present the idea of an interactive videobot, a conversational interface which can be used in businesses, education, healthcare sector and many more. If the user has any query or doubt, they can make use and access this system anywhere anytime. The goal here is to provide the user the sense of being able to converse with an artificial entity with the same intent of emotional and responsive interaction that of any real-time conversation.

Acknowledgement

Authors are very much thankful to Assist 2 Path Tech Pvt. Ltd., New Delhi and CoE AIML, GHRCE, Nagpur for their continuous technical support for implementation of this innovative Video Chatbot project.

REFERENCES

[1] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia Tools and Applications*, pp. 1–32, 2022.

[2] L. Bradeško and D. Mladenić, "A survey of chatbot systems through a loebner prize competition," in *Proceedings of Slovenian language technologies society eighth conference of language technologies*, pp. 34–37, Institut Jožef Stefan Ljubljana, Slovenia, 2012.

[3] F. Guo, A. Metallinou, C. Khatri, A. Raju, A. Venkatesh, and A. Ram, "Topic-based evaluation for conversational bots," *arXiv preprint arXiv:1801.03622*, 2018.

[4] M. Wahde and M. Virgolin, "Conversational agents: Theory and applications," *arXiv preprint arXiv:2202.03164*, 2022.

[5] C. Wu, F. Wu, T. Qi, and Y. Huang, "Hi-transformer: hierarchical interactive transformer for efficient and effective long document modeling," *arXiv preprint arXiv:2106.01040*, 2021.

[6] F. Rahutomo, T. Kitasuka, and M. Aritsugi, "Semantic cosine similarity," in *The 7th international student conference on advanced science and technology ICAST*, vol. 4, p. 1, 2012.

[7] J. Feine, S. Morana, and U. Gnewuch, "Measuring service encounter satisfaction with customer service chatbots using sentiment analysis," 2019.

[8] M. J. Pereira, L. Coheur, P. Fialho, and R. Ribeiro, "Chatbots' greetings to human-computer communication," *arXiv preprint arXiv:1609.06479*, 2016.

[9] A. Ram, R. Prasad, C. Khatri, A. Venkatesh, R. Gabriel, Q. Liu, J. Nunn, B. Hedayatnia, M. Cheng, A. Nagar, *et al.*, "Conversational ai: The science behind the alexa prize," *arXiv preprint arXiv:1801.03604*, 2018.

[10] L. Coheur, "From eliza to siri and beyond," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 29–41, Springer, 2020.

[11] S. A. Abdul-Kader and J. C. Woods, "Survey on chatbot design techniques in speech conversation systems," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 7, 2015.

[12] D. Lee, K.-J. Oh, and H. Choi, "The chatbot feels you - a counseling service using emotional response generation," *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 437–440, 2017.

[13] L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Refining word embeddings using intensity scores for sentiment analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 671–681, 2017.

[14] W. Dhokley, B. Petiwala, M. Sitabkhan, and M. Suratwala, "Video creation using facial animation and speech synthesis," *Available at SSRN 3867650*, 2021.

[15] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[17] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," in *2016 4th International Conference on Cyber and IT Service Management*, pp. 1–6, IEEE, 2016.

[18] J. P. Sampson Jr, R. W. Kolodinsky, and B. P. Greeno, "Counseling on the information highway: Future possibilities and potential problems," *Journal of Counseling & Development*, vol. 75, no. 3, pp. 203–212, 1997.