Journal on Artificial Intelligence
DOI: 10.32604/jai.2022.032390
Article

Evaluating Neural Dialogue Systems Using Deep Learning and Conversation History

Inshirah Ali AlMutairi* **and Ali Mustafa Qamar**

Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia
*Corresponding Author: Inshirah Ali AlMutairi. Email: 411200124@qu.edu.sa
Received: 16 May 2022; Accepted: 25 July 2022

Abstract: Neural talk models play a leading role in the growing popular building of conversational managers. A commonplace criticism of those systems is that they seldom understand or use the conversation data efficiently. The development of profound concentration on innovations has increased the use of neural models for a discussion display. In recent years, deep learning (DL) models have achieved significant success in various tasks, and many dialogue systems are also employing DL techniques. The primary issues involved in the generation of the dialogue system are acquiring perspectives into instinctual linguistics, comprehension provision, and conversation assessment. In this paper, we mainly focus on DL-based dialogue systems. The issue to be overcome under this publication would be dialogue supervision, which will determine how the framework responds to recognizing the needs of the user. The dataset utilized in this research is extracted from movies. The models implemented in this research are the seq2seq model, transformers, and GPT while using word embedding and NLP. The results obtained after implementation depicted that all three models produced accurate results. In the modern revolutionized world, the demand for a dialogue system is more than ever. Therefore, it is essential to take the necessary steps to build effective dialogue systems.

Keywords: Seq2Seq; CNN; dialogue systems; NLP; RNN; transformer; GPT

1 Introduction

A dialogue system is a computer software that helps in conversational interplay with humans. Typically, there are sorts of talk structures assignment-oriented and non-assignment-oriented talk systems [1]. We propose a quit-to-give-up neural conversational model to generate responses based on the verbal exchange information, semantics, and area information. Ultimately, to avoid inconsistent dialogues, we undertake a deep reinforcement ultra-current method that debts for future rewards to optimize the conversational neural model [2]. The critical concept is that the agent conditions solutions are now based on communique statistics and the extracted information that applies to the contemporary-day context. We hire the hierarchical recurrent encoder-decoder (HRED) version for

modeling the communique approach due to the spine network [3]. We assume that the verbal exchange history serves as an anchor to the context in which the dialogue machine is attempting to provide awareness, especially in venture-impartial communication systems in which there is no clear purpose of obtaining, but a greater unique and meaningful communication is preferred [4].

Deep mastering can leverage many statistics for dialogue systems to research meaningful function representations and response technology techniques. Specifically, the dialogue systems are divided into undertaking-oriented and non-mission-oriented fashions depending on how deep mastering techniques assist them with representative algorithms. With the improvement of massive information and deep getting-to-know strategies, the intention of making an automated human-PC communique gadget as a chat accomplice is now not an illusion. In recent times, getting entry to "massive records" for conversations on the net is possible, and we might be capable of discovering ways to respond and what to reply to given (nearly any) inputs, which substantially lets us build statistics-driven, open-area communique structures between user and computers [5]. We are cognizant of training project-oriented speak systems through consumer interactions, wherein the agent improves through communicating with users and studying its mistakes. We suggest a hybrid learning approach using a quit-to-cease trainable neural community version for such structures. We develop a hybrid imitation and reinforcement mastering approach, where we, first of all, educate a talking agent in a supervised way by learning from dialogue corpora and continue to enhance it by consumer coaching and feedback with imitation and reinforcement learning [6].

In this study, we identify some of the most critical issues that must be addressed to develop a superior neural conversation system. We make a humble effort to list these challenges in an order that profiles the maturity level of each one of them. We aim to begin by overcoming the single-turn conversation challenge early in the development process and then address more complex challenges later. Humans can use cognitive higher-order reasoning through conditional logic and semantic inference, which may explain why we require so many arguments in our public discourse. However, natural language processing has not been able to catch up with the range of intelligence in human-to-human conversation. Although empirical knowledge is limited, current research has shown that existing neural models are far from perfect.

The objective of open-domain dialogue systems is to chat with users in a friendly manner without restricting dialogue topics (or domains). For example, given the user intent, "I want to travel anywhere for spring break," we expect the system can respond as "Two of my favorite destinations are Mexico and Hawaii." With the growing interest in pursuing a human-level conversation, more and more research has been done on such open-domain dialogue systems in recent years [7]. The rest of the paper is organized as follows: Section 2 presents an overview of the state-of-the-art. We describe our recommended strategy or methodology in Section 3. We present the experimental outcomes and assessments against a range of standards in Section 4. Finally, in Section 5, we conclude the paper along with providing some future directions.

2 Literature Review

By supplying an herbal way for records looking, multimodal communication structures are gathering interest of numerous fields such as retail and travel. However, most of the current communication structures are restrained to textual modality, which can be quite challenging while trying to seize the rich semantics in the visual modality. Without this, the correspondence specialist may neglect to create legitimate reactions for clients. The paper by Liao et al. [8] shows an information mindful Multimodal Dialog (KMD) rendition to manage the difficulty of text-principally based correspondence structures.

It gives extraordinary thoughts to the semantics and area data found in a noticeable substance and is included with three essential added substances. First, they construct a taxonomy, primarily-based learning module to seize the high-quality grained semantics in pictures. Secondly, they advocate a quit-to-quit neural conversational version to generate responses based on the verbal exchange records, visible semantics, and area understanding. Ultimately, they undertake a deep reinforcement studying method, such as the neural conversational model [8].

We comprehensively review state-of-the-art research outcomes in dialogue systems and analyze them from two angles: model type and system type. The discussed models include: Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Vanilla Sequence-to-sequence Models, HRED, Memory Networks, Attention Networks, Transformer, Pointer Net and Copy Net, Deep Reinforcement Learning models, Generative Adversarial Networks (GANs), Knowledge Graph Augmented Neural Networks [7].

Another paper by Yang et al. [9] states that an intelligent personal assistant system that can have multi-turn conversations with human users has become more and more admired. Most primary research has focused on using two techniques together, namely retrieval-based and era-based, to enhance such structures. The recovery-based methods have to gain recovery and quick responses with brilliant variety. However, the performance of these techniques is evaluated by utilizing a method known as the reaction repository. On the other hand, strategies can produce immediate and effective responses to any query. However, the generated responses are frequently widespread and not informative due to the shortage of grounding expertise. The experiment was done on Twitter and other social sites' data, and the implemented model produced better results with both methods.

Knowledge-grounded talk structures are meant to deliver the data based on evidence provided in a given textual content supply. The authors in [10] speak about the challenges of training a generative neural dialogue model for such structures. Existing datasets include a combination of trustworthy conversational responses to choose proof and extra subjective or chit-chat fashion responses. The researchers advise specific evaluation measures to disentangle those exclusive forms of responses by quantifying in formativeness and objectivity. The study also investigates the use of extra controls when interpreting the use of resampling techniques. Furthermore, in this examination, they perform a human assessment and assess which raters choose the output of these managed technology models to be typically more objective and devoted to the proof than baseline dialogue systems.

Building open-area conversational designs that permit clients to draw in discussions on subjects in their domain is complex. Alexa, a voice assistant developed by Amazon, was released in 2016 to solve the problem of achieving regular, supported, and drawing open-end word exchanges. In a competition organized in 2018, school groups designed the best in class conversation model using a setting in discourse models, utilizing charts for language information, overseeing complex expressions, building concrete and various leveled discussions, and utilizing model alarms from individual reactions. The 2018 opposition additionally blanketed the provision of a suite of gear and fashions to the competition together with the CoBot (conversational bot) toolkit, subject matter, and conversation act detection models, so that the competing groups may want the consciousness of constructing understanding-rich, coherent, and attractive multi-turn dialogue structures [11].

The objective-oriented dialogue framework needs to be improved for calculating the talk speed and highlighting strong correspondence under different circumstances to meet the client's point-of-view. The customary way to deal with developing such talking machines is to take a pipelined measured structure, in which its modules are upgraded as one would see it. In any case, such an improvement plot does never again yield the total machine's general exhibition advancement. Then

again, exchange frameworks with solid structures are frequently prepared distinctly with enter-yield expressions, without thinking about the entire comments accessible inside the dialogue. This plan makes it extreme for objective-arranged exchanges wherein the interpreter wishes to incorporate outer structure or offer interpretable data about why the gadget created a chosen reaction [12].

Spoken discussion frameworks have nearly 30 years of history, which can be isolated into three categories: representative rule or format-based absolutely (before the late 90s), measurable research-based, and profound acquiring information fundamentally-based (since 2014). Nuruzzaman et al. [13] presented Normal Language Generator (NLG), also called reaction Generator. They provided insights-driven discussion frameworks notwithstanding intelligent discussion structures, in which a human is associated with acquiring information on exchange machine-added substances. The use of profound learning on genuine worldwide conversational discussion is communicated in language or discourse acknowledgment. This has a significant impact on the progress of the generally spoken discussion device. This front-quit thing incorporates various variables that make it intense for machines to comprehend discourse.

However, gathering facts to broaden intention-oriented communication systems that could help customers complete an undertaking in a particular domain remains hard. Consequently, in this paper, we comprehensively evaluate works in recent years with a focal point on deep ultra-modern primarily-based procedures and provide insight into studies from both version attitude and device perspective. Moreover, this research covers various hot topics in speaking systems. Most, if not all, dialogue systems are deep state-of-the-art-based systems. The fast boom trendy deep present-day improves the performance of modern-day systems. Deep state-of-the-art can be regarded as an illustration of ultra-modern with multilayer neural networks [7].

3 Proposed Methodology

3.1 Methodology Overview

The dataset for this implementation will primarily come from numerous movies. The features would be formed using natural language processing algorithms applied to a dataset consisting of movie scripts written in English, obtained from multiple sites. These movies were chosen due to the presence of long conversational speeches and the informative descriptions they provide. The modest number of language data collections brings about a low standard sequence, so preprocessing of language information is required. To ease this difficulty, numerous researchers have stated preprocessing techniques and used pre-training models in the word embedding system with the Seq2Seq model. Bidirectional Encoder Representations from Transformers (BERT) comprises a few degrees of self-consideration, and natural language understanding is quickly moving along. Even though it can perform classification undertakings, the BERT-based research has led to performing sequential problems in GPT. Transformer models have likewise accomplished better outcomes in the arrangement of sequential assignments. These pre-trained models will assume a more critical part in the dialogue exchange module. Be that as it may, they have moderately high necessities for equipment limits and limit how much info tokens are utilized [14]. The proposed approach is depicted in Fig. 1.

Recent research on the encoder-decoder model for creating reactions principally utilizes some implanted potential features to upgrade the variety of reactions. The quality of the reactions to the issues is missing, and it exhibits in the start to finish model. The search data of the information base and the data of the dialogue setting might have a decent impact on the dialogue made. According to Fig. 1, once the data is gathered from various sources, it moves to the preprocessing phase because, as previously stated, data is of low quality and this phase is required to make it valuable and useful.

After cleaning the data, it is divided into an 80:20 ratio for the training and testing phases. Because the neural network is entirely based on continuous training and testing, this phase is also critical. Once the data has been split, it is passed to the proposed models to determine which model is the best in terms of accuracy and reliability in the evaluation phase.

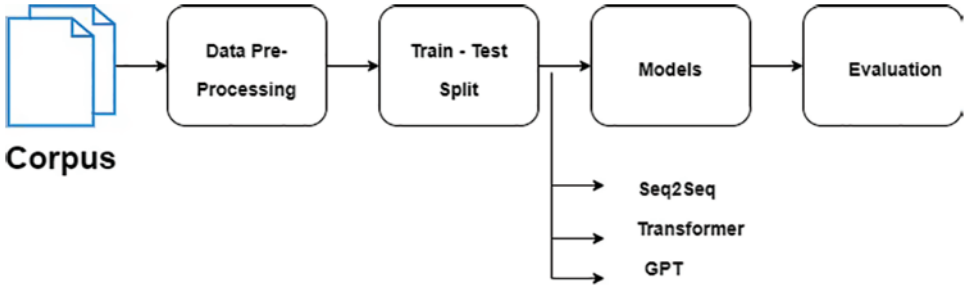

Figure 1: Proposed approach

3.2 Word Embedding

Word embedding is an integral part of DL and neural networks. It is essential for various NLP projects. A vital part of neural networks for NLP functions is the primary layer, or the word implanting layer, that guides input words to vectors. These embeddings, which represent a query table from words in jargon to digits, are regularly trained through preparing an artificial neural network to produce language displaying on a huge dataset and can then be utilized as prepared vectors to introduce the word inserting a layer of other NLP.

3.3 Seq2Seq Model

Seq2Seq model is a high-level neural network that intends to transform one sequence into another. Two enhanced neural networks are included in the Seq2Seq model: the encoder RNN (more frequently as LSTM or GRU to prevent gradient issues) encrypts the source sentences to supply the secret stage for the decoder RNN. However, the decoder RNN creates the objective sentences. Although the Seq2Seq model has tremendous advancement in NLP, yet it experiences the bottleneck of disappearing gradient issue. A popular strategy on top of Seq2Seq permits the decoder to analyze the specific pieces of the source, which eases the gradient descent issue by giving an easy route to distant states. NLP can allude from scratch for additional subtleties, which is a translation with a Seq2Seq Network [15]. Apart from this, Seq2Seq generation-based dialogue systems function poorly because the generation mechanism is uncontrollable. It responds to a response depending on the pattern learned from the training dataset. As a result of the high frequency of these characteristics in training data, the process is likely to provide an unexpected response with little interpretation, such as "I don't know" and "Okay" [16].

3.4 Transformers

There has been considerable growth in the subject of language modeling in recent years. Numerous language comprehension benchmarks have seen significant gains since the current generation of NLP models was announced at the end of 2018, with only a few obtaining fairly close to human-level precision. This rapid advancement can also be explained by more stringent benchmarks, as many models have exceeded existing standards. Transformer and Transfer learning are two significant contributors to this achievement. Transformers comprise a multi-head self-consideration system

joined with a mechanism that performs encryption and decryption. It can accomplish excellent results that are far better than those of models such as RNN or CNN, as far as assessment score and effectiveness is concerned. A crucial benefit of Transformer over other Neural Network structures is that a more drawn context and description around a word is considered more proficiently. It depends on self-consideration regarding process portrayals of its input and result without utilizing sequentially arranged Recurrent or Convolution neural networks.

Different transformer models should be implemented depending on the project to be performed using Natural Language Processing [17]. The transformer model design contains N stacked layers containing various self-attention heads trailed by a feed-forward layer. The system in the self-attention works by permitting instruments in each layer to concentrate on the situations from the layer before that are the closest, bringing about a weighted total sum of inputs from the past layer. It also constitutes a mechanism, same as the seq2seq model between encryption and decryption, compensating input with the required attention to produce the output. The self-attention architecture permits every layer to concentrate on the generally relevant layer, bringing about a weighted amount of the contributions from the last layer, where the loads are the consideration values. Likewise, the standard attention system is found in the arrangement of the seq2seq model between the encoder's result and the decoder, which realizes where to give the most consideration throughout the inputs while creating its result [18].

3.5 GPT

GPT is a huge transformer-based language model with 1.5 billion highlights developed using a dataset of eight million sections. GPT has a clear objective: predict the impending word, thinking about every one of the last words. The variety of the dataset makes this basic objective contain reoccurring demonstrations of many assignments across different domains. GPT is an immediate scale-up containing almost 10 times the features and a training set of more than 10 times of the original data [19]. GPT is a pre-formation language model built on an auto-regressive Transformer that has received great interest in the NLP area because of its revolutionary success in various applications. GPT's victory can largely be attributed to its extensive pre-formation of massive amounts of data and its high characteristics (ranging from 100 million to billions). Although GPT has improved performance (particularly in very few zero-shot setups), its over parameterized character makes it difficult to deploy on systems with low computing capabilities or storage [20].

4 Experiments and Results

4.1 Seq2Seq Model

Sequences are central to almost every data science domain, from speech and language processing (voice input, text translation) to time-series forecasting (weather forecasting, stock market prediction). A seq2seq model is an extremely effective architecture and has been used in various applications. This model was used in production by Google Translate as well in the year 2016. The Seq2seq model takes a sequence of the object which involves time series, alphabets, and letters. The output is represented by another sequence of objects. In the translation of the Neural Machine system, the series of alphabets, words, or letters are considered as an input and the translated word of the series obtained is considered as an output.

There are three graphs related to the seq2seq model in the context provided next. Fig. 2 represents the seq2seq model accuracy. In the beginning, the accuracy on the training set is low, but with the increase in the number of epochs, the accuracy of the training set increases and reaches up to 1.0, which is equivalent to 100% when the epoch value is at 100. After that, the accuracy becomes stable for the remaining number of epochs. In contrast, the accuracy of validation set (Val), in the beginning,

is low, and it increases up to 0.2, but as it reaches this maximum value, the accuracy starts dropping and becomes stable at less than 0.2. With the increase in epoch values, there is no increase in the accuracy of the Val.

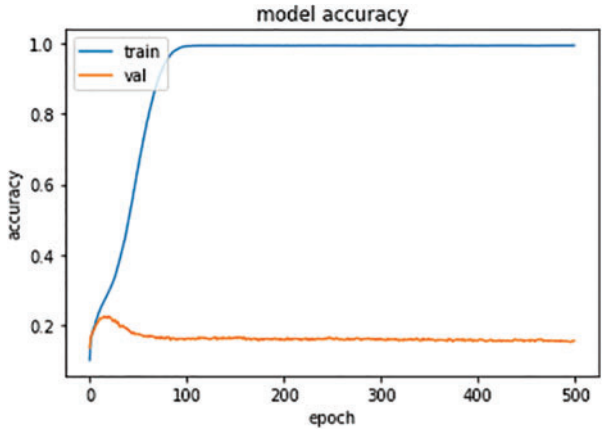

Figure 2: Seq2Seq model accuracy

Fig. 3 represents the loss for the seq2seq model. In the beginning, the loss of training set is around 0.6, but with the increase in the number of epochs, the losses of training set decrease and reach up to 0, when the epoch value is at 100. Once the losses reach 0, there is no increase or decrease in the loss, and it becomes stable for the remaining epochs. In contrast, the losses of Val, in the beginning, are also close to 0.6, but with the increase in the number of epochs, there is a significant increase in the losses of Val.

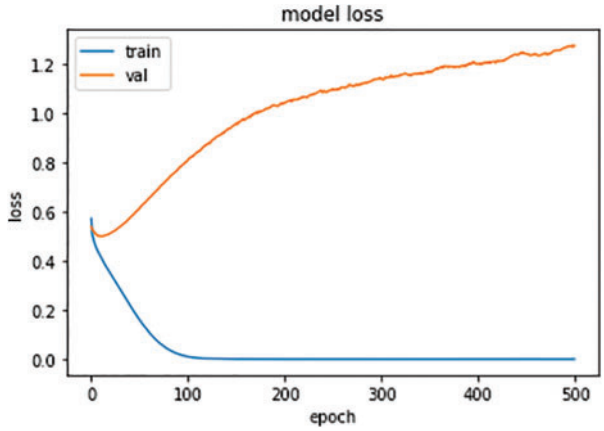

Figure 3: Seq2Seq model losses

Fig. 4 represents the seq2seq model's perplexity. The training perplexity does not increase with the increase in epochs. On the other hand, the validation perplexity increases with the increase in epochs and based on the data, the highest perplexity is observed around 450 epochs.

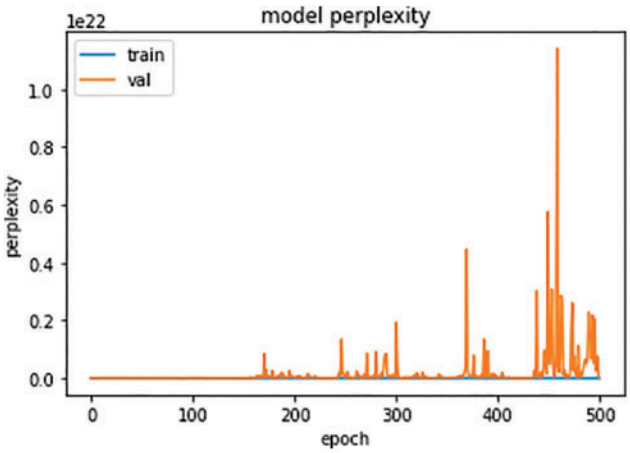

Figure 4: Seq2Seq model perplexity

4.2 Transformer Model

Without ever using sequentially ordered RNN or CNN, transformers rely nearly exclusively on self-attention to comprehend their input and output representations.

There are three graphs associated with the Transformer model in the context provided next. Here, the results show a correlation between the training values and the real values. Fig. 5 represents the Transformer model's accuracy. Based on the provided figure, in the initial stages, the training accuracy is low. Still, with the increase in epoch value, there is a significant increase observed in the training accuracy, and it becomes stable when the epoch value reaches 500. Similarly, in the beginning, the accuracy of Val is low, but it increases with the increase in epoch values and becomes greater than the training accuracy.

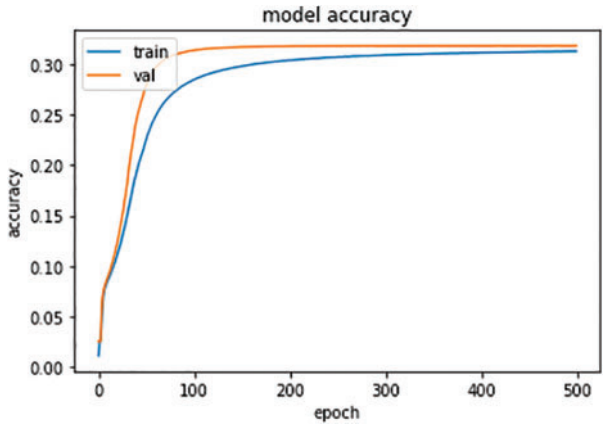

Figure 5: Transformer model accuracy

Fig. 6 represents the losses for the Transformer model. One can observe that in the beginning, the train losses are high. However, with the increase in epoch value, there is a significant decrease observed in the train losses, and as with the increment in epoch value, the decrement in the losses increase. Similarly, in the initial phase, the losses of Val are high, but it decreases with the increase in epoch values and becomes stable.

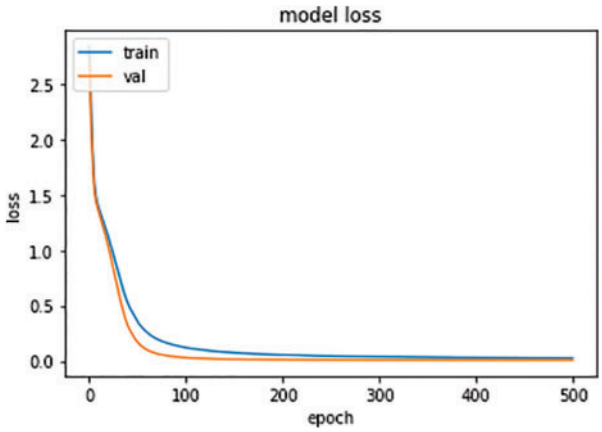

Figure 6: Transformer model loss

Fig. 7 depicts the Transformer model perplexity. Based on the figure, the perplexity of Val is higher as compared to the train.

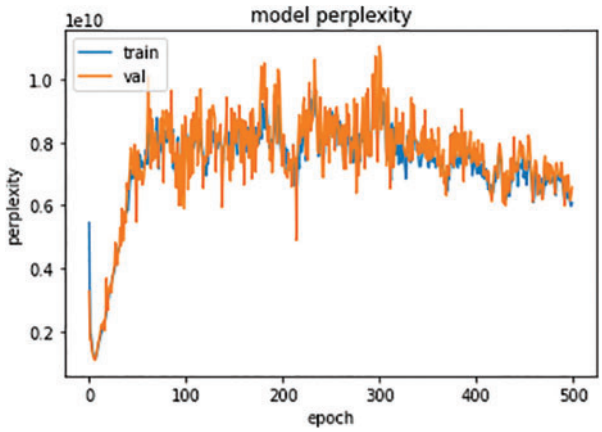

Figure 7: Transformer model perplexity

4.3 GPT

In the GPT pre-trained model, it is impossible to generate model accuracy, losses, and perplexity. Therefore, for the comparisons with the other models, we can only utilize the scores of BLEU and ROUGE.

Based on Table 1, the Bleu score of the Dialo GPT model is close to 0 which shows that this model is not perfect compared to the other models. Moreover, the Rouge score of the Dialo GPT model is also close to 0, indicating that the model is not as good as the other two models.

Table 1: Comparison chart

Model	Loss	Accuracy	Perplexity	Bleu	Rouge
Seq2seq	0.001	1.0000	1.0000	0.734	0.068
Transformer chatbot	0.0286	0.3123	6071625728.000	0.803	0.789
DialoGPT	—	—	—	0.015	0.0039

5 Conclusion

This study concludes that a communication device is a software for computers that allows humans to converse with each other in a verbal, word-based, or multimedia manner. Assignment-oriented and non-assignment-oriented talk systems are two common discussion structures. We assume that the history of verbal exchanges exists as a pillar to the frame of reference on which the dialogue machine intends to get awareness. This is how it has been so widely used, particularly in venture-impartial communication networks with no obvious objective to meet. Still, a significantly larger distinctive and impactful communication is recommended. We utilize the hierarchical recurrent encoder-decoder (HRED) model to design the verbal exchange approach.

For this purpose, we have selected three models: sequence-to-sequence (seq2seq), transformer, and GPT. Based on the analysis, every model is distinctive from the other. Combining these models can help us build the desired system because the accuracy of the seq2seq model is relatively high compared to the other models. However, we can reduce the system's losses by implementing the transformer model in terms of losses. In terms of Bleu and Rouge scores, the results obtained with the transformer model are closer to the desired results than other models.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. McTear, "Conversational AI: Dialogue systems, conversational agents, and chatbots," *Synthesis Lectures on Human Language Technologies*, vol. 13, no. 3, pp. 1–251, 2020.

[2] C. Sankar, S. Subramanian, C. Pal, S. Chandar and Y. Bengio, "Do neural dialog systems use the conversation history effectively? An empirical study," in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy, 2019.

[3] L. Liao, Y. Zhou, Y. Ma, R. Hong and T. -S. Chua, "Knowledge-aware multimodal fashion chatbot," in *Proc. of the 26th ACM International Conf. on Multimedia*, Seoul, Republic of Korea, pp. 1265–1266, 2018.

[4] Y. Ma, K. L. Nguyen, F. Z. Xing and E. Cambria, "A survey on empathetic dialogue systems," *Information Fusion*, vol. 64, pp. 50–70, 2020.

[5] H. Chen, X. Liu, D. Yin and J. Tang, "A survey on dialogue systems: Recent advances and new frontiers," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 2, pp. 25–35, 2017.

[6] B. Liu, G. Tür, D. Hakkani-Tür, P. Shah and L. Heck, "Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems," in *Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, LA, USA, pp. 2060–2069, 2018.

[7] J. Ni, T. Young, V. Pandelea, F. Xue and E. Cambria, "Recent advances in deep learning based dialogue systems: A systematic survey," *Artificial Intelligence Review*, 2022.
[8] L. Liao, Y. Ma, X. He, R. Hong and T. -S. Chua, "Knowledge-aware multimodal dialogue systems," in *Proc. of the 26th ACM International Conf. on Multimedia*, Seoul, Republic of Korea, pp. 801–809, 2018.
[9] L. Yang, J. Hu, M. Qiu, C. Qu, J. Gao *et al.,* "A hybrid retrieval-generation neural conversation model," in *Proc. of the 28th ACM Int. Conf. on Information and Knowledge Management (CIKM '19)*, Beijing, China, pp. 1341–1340, 2019.
[10] H. Rashkin, D. Reitter, G. S. Tomar and D. Das, "Increasing faithfulness in knowledge-grounded dialogue with controllable features," in *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. on Natural Language Processing*, Online, pp. 704–718, 2021.
[11] C. Khatri, B. Hedayatnia, A. Venkatesh, J. Nunn, Y. Pan *et al.,* "Advancing the state of the art in open domain dialog systems through the ALEXA prize," arXiv preprint, arXiv:1812.10757, pp. 1–28, 2018.
[12] D. Ham, J. -G. Lee, Y. Jang, and K. -E. Kim, "End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2," in *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 583–592, 2020.
[13] M. Nuruzzaman and O. K. Hussain, "IntelliBot: A dialogue-based chatbot for the insurance industry," *Knowledge-Based Systems*, vol. 196, pp. 105810, 2020.
[14] Y. Fan, X. Luo and P. Lin, "On dialogue systems based on deep learning," *International Journal of Computer and Information Engineering*, vol. 14, no. 12, pp. 525–533, 2020.
[15] J. Hui, "[Deep learning] How to build an emotional chatbot," (Part 2 the dialogue system). 2020. [Online] Available: https://towardsdatascience.com/deep-learning-how-to-build-an-emotional-chatbot-part-2-the-dialogue-system-4932afe6545c (Accessed 26-March-2022).
[16] L. Yao, Y. Zhang, Y. Feng, D. Zhao and R. Yan, "Towards implicit content-introducing for generative short-text conversation systems," in *Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 2190–2199, 2017.
[17] J. Nikulski, "How to use transformer-based NLP models," 2021. [Online] Available: https://towardsdatascience.com/how-to-use-transformer-based-nlp-models-a42adbc292e5 (Accessed 26-March-2022).
[18] M. B. Korpusik, "Deep learning for spoken dialogue systems: Application to nutrition," Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 2019.
[19] A. Radford, "Better language models and their implications," 2019. [Online] OpenAI. Available: https://openai.com/blog/better-language-models (Accessed 26-March-2022).
[20] A. Edalati, M. Tahaei, A. Rashid, V. P. Nia, J. J. Clark *et al.,* "Kronecker decomposition for GPT compression," in *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics Volume 2: Short Papers*, Dublin, Republic of Ireland, pp. 219–226, 2022.