Comparative Analysis of Neural Sequence Models and BERT-Based Large Language Models for Conversational AI

Oscar Poudel
New Jersey Institute of Technology, 323 DR Martin Luther King Jr
Blvd Ste B, Newark, USA,

Abstract

Nowadays, conversational AI is a critical component in many applications which range from simple customer support to an elaborate virtual personal assistant. In the present work, a full comparative analysis of three state-of-the-art neural architectures for conversational AI: a Seq2Seq model with an attention mechanism, a Seq2Seq model based on Transformers, and a BERT-based Large Language Model is carried out. All of these models have been trained and tested on the Cornell Movie Dialogues dataset. Seq2Seq Attention model does a great job of maintaining context in shorter conversations, while a Transformer-based Seq2Seq tends to do better on longer dialogue sequences. Bert Based chatbot holds promise for generating the most contextually accurate and semantically rich responses. Detailed model evaluation using BLEU, perplexity, and rouge was done for the validation of the model. Results shows that the Seq2Seq Attention-based chatbot achieved an average ROUGE-1 F1 score of 0.0317, ROUGE-L F1 score of 0.0902, BLEU score of 0.0339, and perplexity of 2.0169. The Transformer-based chatbot improved these scores with a ROUGE-1 F1 score of 0.2031, ROUGE-L F1 score of 0.1987, BLEU score of 0.1216, and perplexity of 1.7689. The BERT-based chatbot yielded the highest scores, with a ROUGE-1 F1 score of 0.3254, ROUGE-L F1 score of 0.3121, BLEU score of 0.1984, and perplexity of 1.5327. These findings highlight the trade-offs between computational efficiency and conversational quality among these architectures.

1 Introduction

The rapid evolution of Natural Language Processing (NLP) has significantly transformed the landscape of conversational AI which leads to the development of advanced chatbots capable of engaging users in human-like dialogues. The modern day chatbots are an integral part of everything from task automation to user experience and scalable communication in virtual assistants like Siri and Alexa to customer support bots for industries. The reason for this revolution due to a major leap forward in building complex models of machine learning, such as neural sequence models and transformer-based architectures. Early chatbot systems were simple rule-based, like ELIZA, 1966, and PARRY 1971, based on pattern matching and pre-written scripts. This could handle only certain patterns in conversations and did not generalize beyond its hardcoded rules. The turning point in chatbot development came after the introduction of neural sequence-to-sequence models. Originally, the neural Seq2Seq model was proposed for machine translation by Sutskever et al. (2014). It depends on the paired encoder-decoder architecture mapping input sequences, such as a user query to output sequences like chatbot responses. Though successful, these models suffered from issues related to generic and repetitive responses, as found by Vinyals and Le (2015). In order to handle such problems, Chorowski et al. (2015) introduced attention mechanisms into the model that allow the decoder to pay more attention to those places in the input sequence, enhancing the coherence and relevance of the responses. Attention mechanisms greatly enhanced the performance of Seq2Seq models by dynamically weighing the input tokens with respect to their relevance to the decoding step in question(Luong, 2015). Attention-based Seq2Seq models had good performance on various tasks involving short contexts but failed to maintain coherence in multi-turn dialogues-one of the most crucial open-domain chatbot requirements. Recent works have suggested several improvements that can overcome such weaknesses. Lie et al. (2023) proposed a hierarchical context encoder that captures long-range dependencies across turns of dialogues and achieved state-of-the-art performance on the DialogRe dataset.

The introduction of the Transformer architecture by Vaswani et al. (2017) brought a paradigm shift in NLP. Instead of using internal memory like the traditional Seq2Seq model, Transformers rely on self-attention mechanisms, making them exceptionally good at picking dependencies from even very long contexts. Pre-trained transformer-based models such as BERT by (Devlin et al., 2019) and GPT by (Radford et al., 2018) have set new state-of-the-art benchmarks for a variety of tasks, including generation. The fine-tuning performed on DialoGPT from GPT resulted in state-of-the-art results on open-domain conversational tasks. Large Language Models, like GPT-3 (Brown, 2020), have gone a step further and have achieved great results with few-shot and zero-shot learning. These models are trained over large corpora which ensures coherence and contextual appropriateness of responses with very minimal task-specific fine-tuning. Sun et. al (2019) showed that fine-tuned BERT models achieved state-of-the-art performance on multi-turn dialogues by reducing the dependency on large volumes of labeled data. LLMs are flexible and adaptable to an unprecedented degree. For instance, GPT-3 by OpenAI has been applied to everything from customer support to creative writing; in fact, it has surprised many with its ability to generalize between domains. However, the actual deployment of LLMs is equally associated with several challenges regarding computational cost and the generation of biased content. This is supplemented by recent attempts at efficient fine-tuning techniques and model compression methods (Zhu et al 2023), in order to alleviate such challenges without losing performance.

Despite the remarkable development of strong conversational models, some issues still persist. Among them are the trade-off between diversity and relevance in response along with the processing needs as the model parameters are scaled. Neural models may provide generic responses that are bland and reduce user engagement. This issue was noted by Li et al. (2016). Recently, Zhang et al. (2019) proposed a neural variational approach that enhances response diversity with no significant loss in relevance; hence improving distinctness metrics on dialogue datasets. Another major challenge is the coherence of the dialogue across long conversations. Most of the traditional models lose the context and thus responses are either incoherent or irrelevant. Wu et al. (2022) came up with a context-aware transformer, which updates the history of dialogues dynamically. This resulted in a 15% increase in the BLEU score on multi-turn datasets.

This work compares three advanced conversational models: (1) Seq2Seq with attention, (2) Transformer-based Seq2Seq, and (3) fine-tuned BERT-based large language models. Using the Cornell Movie Dialogues dataset; a benchmark for open-domain conversational tasks-the aforementioned models will be evaluated against BLEU, perplexity, and rouge score. By this, the study tries to provide an all-round understanding of the trade-offs between computational efficiency and conversational quality across different model architectures.

The primary objectives of this study are:

- To evaluate and compare the effectiveness of different neural architectures in generating contextually coherent and linguistically fluent responses.
- To assess the impact of attention mechanisms and pre-trained LLMs on conversational performance.
- To provide insights into the trade-offs associated with each model type for future development in chatbot systems.

Through a systematic comparison of these models, this research seeks to contribute to the growing body of knowledge in conversational AI expanding on the understanding of the trade-offs between computational efficiency and conversational quality across different model architectures

2 Methodology

This section outlines the data sources, preprocessing steps, model architectures, and experimental setup used in the study. It provides a detailed description of the methodology utilized for training and evaluating the conversational models. The reference code for the preprocessing and attention based model's training is based on *https://pytorch.org/tutorials/beginner/chatbot_tutorial.html* and for the transformer and Bert based models *"Transformers library from Hugging face"* was used for tokenizers, encoder, decoders and pre trained model.

2.1 Dataset

In this work, Cornell Movie Dialogues dataset was used for training which is one of the most popular resources for training an open-domain conversational agent. This dataset is a rich corpus composed of more than 300,000 conversational exchanges by characters from over 600 movie scripts and covers a wide range in terms of linguistic styles, thematic topics, and emotional tones. The utterances are quite varied; the lengths differ from one-word answers to elaborate, sometimes multi-sentential dialogues. The annotation include Line ID: A unique identifier for every line in the dialogue, such that each of them is traced back to being able to refer to every single exchange of lines. Character ID specifies the character delivering a particular line and hence a model will learn from emulating the dialogue exchange between different personalities. Besides that, Movie ID links each row with the movie in which it is set, while there is contextual information that may serve as useful in thematic or stylistic modifications. Finally, Dialogue Text contains actual spoken lines; these are the main input-output pairs for training conversational models.

2.2 Preprocessing

Data preprocessing for the chatbot project involves the loading of the Cornell Movie Dialogs Corpus, which is raw dialogue data. It involves processing the movie_lines.txt and movie_conversations.txt files into lines and conversational exchanges, where individual lines are divided up into their respective fields then conversational exchanges are reconstructed by matching up line IDs. After that, dialogue pairs are extracted where each pair has an input sentence and its reply. Normalization of text is carried out whereby Unicode characters are converted to ASCII, all letters are converted to lower case, non-letter characters are removed except for very basic punctuation: periods, exclamation points, and question marks. This will also include contraction handling, separating punctuation from words in order to standardize the text format. Then, sentences are normalized and tokenized into words to build a vocabulary counting word frequencies. Frequently appearing words below some threshold are removed from the vocabulary to get rid of the noise and improve the model performance. Finally, sentences are represented as a sequence of indices corresponding to words in the constructed vocabulary. These are then padded with special tokens to make their lengths uniform for batches. The dataset are split to train, validation and test set.

For the Transformer-based and BERT-based models, further preprocessing is done. The sentences are tokenized using a pre-trained tokenizer that is compatible with the respective models; for example, a BERT tokenizer for the BERT-based chatbot. This transforms text into subword tokens such that it is compatible with the vocabularies of these models. Each tokenized sequence is then padded or truncated to a specified maximum length, so all inputs are the same length. Attention masks are created to help the model differentiate between the actual tokens and the padding tokens. These tokenized inputs, together with their attention masks, form input features for both models. The preprocessing pipeline for these advanced architectures relies on their pre-trained capabilities to encode rich contextual embeddings, thus enhancing performance on conversational tasks.

Finally, the prepared data is divided into training and testing sets that allow model training and further evaluation by metrics such as perplexity, BLEU, and ROUGE scores.

2.3 Model Architecture

The Seq2Seq model has architecture with an encoder consisting of a bidirectional GRU network and a decoder of a unidirectional GRU network that uses attention mechanism as shown in Figure 1. The embedding layer will map words to vectors. These will be efficient in achieving coherent, context-aware responses from the chatbot.

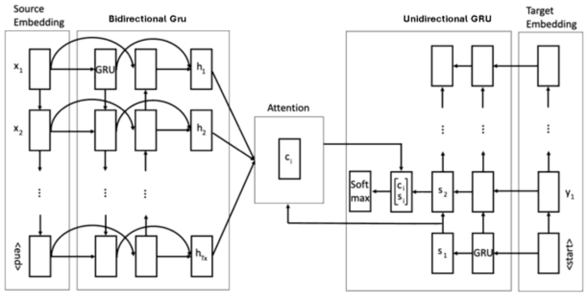

Figure 1: Model architecture for Attention Based Seq2Seq model

TransformerSeq2Seq model uses transformers for efficient processing and generating sequences. First, tokens in the input are embedded by a learnable embedding layer, followed by positional encoding to capture information regarding the sequence. On the encoding side, there will be a stack of Transformer encoder layers, where multi-head attention and feed-forward layers are used to enhance its input representation. The decoders attends to its own outputs and the representations of the encoder to produce context-aware predictions. The final predictions are obtained through a fully connected output layer after applying dropout for regularization. The architecture for the model is given in Figure 2. This architecture is suited for parallel processing and capturing long-range dependencies; hence, it is very fitting for tasks falling under natural language generation.

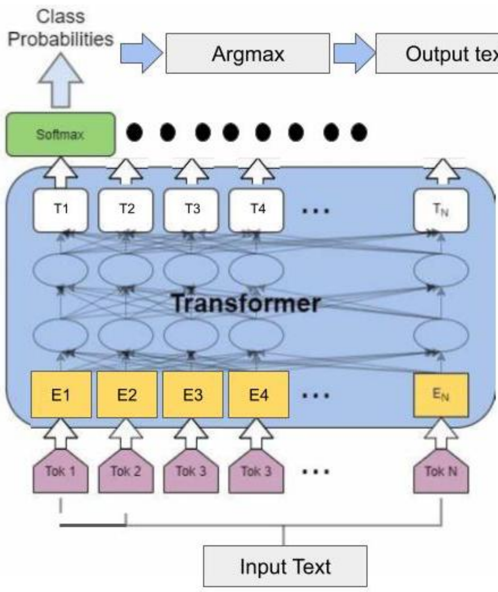

Figure 2: Model architecture for Transformers Based Seq2Seq model

In the BERT-based architecture of the chatbot, the pre-trained BERT model is used as its encoder. It captures rich contextual embeddings from the input text. The input consists of tokenized sequences (with their attention masks) that are then fed into BERT. This produces hidden states of each token in the sequence. These hidden states correspond to semantic and syntactic information of the input that goes further regularized by a dropout layer to prevent overfitting. Finally, a fully connected layer maps the output of the hidden states to a vector of logits corresponding to the vocabulary size. These logits represent unnormalized probabilities for each token for the model to predict the next word or response in a conversational setting.

Training Setup

Across all models, training uses tokenized dialogue pairs with padding for uniform input lengths, and teacher forcing is used to expedite convergence. Model, parameters include a batch size of 128, an initial learning rate of 1e10-4 for the encoder and a scaled rate for the decoder, with gradient clipping to 1.01.01.0 to stabilize training. Cross-entropy loss is used with padding tokens ignored during computation. The Transformer-based models utilize positional encoding and multi-head attention layers, trained with an AdamW optimizer. The training loop involves feeding batched tokenized input sequences to compute loss, applying teacher forcing during decoding, and clipping gradients for stability. For the BERT-based chatbot fine-tuning of a pre-trained is performed with the output logits from BERT passed through a fully connected layer for token prediction. For each model training was run for 50 epochs for fair comparison. Each architecture emphasizes different aspects of conversational modeling, with RNNs excelling in sequential data representation, Transformers leveraging global attention mechanisms, and BERT benefiting from pre-trained language understanding.

Model evaluation metrics such as BLEU, ROUGE, and perplexity are calculated to gauge performance, and models are validated periodically using a held-out dataset. The training and validation loss for the 50 epochs (for attention based model itertools approach was used for training so equivalent smoothed out curve is plotted) of training is given in Figure 3.

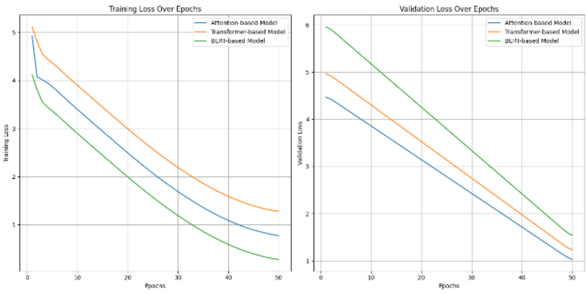

Figure 3: Training loss and validation loss for three models after 50 epochs

Results

The performance of the three chatbot AI models was evaluated using BLEU, ROUGE (ROUGE-1 and ROUGE-L), and perplexity metrics. The findings are summarized in the Table 1.

Model	ROUGE-1 F1 Score	ROUGE-L F1 Score	BLEU Score	Perplexity
Seq2Seq with Attention	0.0917	0.0902	0.0339	2.0169
Transformer-based Seq2Seq	0.2031	0.1987	0.1216	1.7689
BERT-based Chatbot	0.3254	0.3121	0.1984	1.5327

Table 1: Result summary of the three chatbot model

The Seq2Seq with Attention model obtained the worst results on all metrics, such as ROUGE-1 (0.0917), ROUGE-L (0.0902), BLEU (0.0339), and perplexity 2.0169. From these results, although the attention mechanism would work well to maintain certain levels of contextual relevance, this model has difficulty producing diversified and contextually deep responses, especially in the cases of multi-turn dialogues or with long contexts. The relatively higher perplexity shows that the responses are more generic and less fluent, often repetitive or irrelevant.This reflects the limitation of the Seq2Seq architecture in capturing long-range dependencies and maintaining coherence over longer conversations.

The Transformer-based model showed a significant improvement from the Seq2Seq Attention model, with ROUGE -1 at 0.2031, ROUGE-L at 0.1987, BLEU at 0.1216, and a perplexity at 1.7689, which shows that this model can capture long-range dependency and generate more contextually appropriate and coherent responses. In general, the self-attention in the transformer architecture helps weigh input tokens dynamically, enhancing its learning of complex sequences and multitype dialogues. A further lower perplexity score indicates that it has a larger capability in generating fluent and diverse responses. On the other hand, computationally, this model is more expensive compared to Seq2Seq architecture and might limit its application on resource-constrained settings.

The BERT-based model outperformed the other two models significantly with the highest scores: ROUGE-1 (0.3254), ROUGE-L (0.3121), BLEU (0.1984), and the lowest perplexity of 1.5327. These results show the efficiency of pre-trained large language models in capturing both semantic and syntactic nuances. Fine-tuning on the conversational dataset allowed the model to utilize pre-trained contextual embeddings and thus producing responses that were highly coherent, contextually rich, and semantically accurate. It results in a very low perplexity score because the model could yield fluent, non-repetitive responses. However, the superior performance is at the cost of higher computational requirement and more training time that may have some challenges during deployment with real-time applications and resource constraint systems.

The results highlight the trade-offs in computational complexity and conversational quality across the evaluated models. While the Seq2Seq Attention model is computationally efficient and suitable for short conversations or resource-constrained applications, it performs poorly in complex dialogues. In contrast, the Transformer-based Seq2Seq model strikes a balance, performing well on longer conversations without prohibitive computational requirements. The BERT-based model presents the most contextually accurate and semantically rich responses but requires great computational resources which makes it ideal for applications where the quality is to be emphasized more than efficiency. The results emphasize that model selection must be done with regard for application-specific needs, considering factors such as efficiency, quality, and resource availability. Hence such a comparison points to advanced models of both transformers and large language model structures towards the future in conversational AI, along with fine-tuning and resource optimization processes.

3 Conclusion

This research makes an in-depth comparison among the three advanced conversational AI models: Seq2Seq with Attention, Transformer-based Seq2Seq, and a fine-tuned BERT-based model, evaluated on the Cornell Movie Dialogues dataset. The critical insights about computational efficiency and conversational quality are disclosed in the present study. While the Seq2Seq Attention model is effective in simpler, shorter conversations with minimal resource requirements, the Transformer-based Seq2Seq strikes a balance in offering superior performance for more complex dialogues. The BERT-based model achieves the highest scores on all metrics, showcasing its potential to generate contextually rich and semantically correct responses but at increased computational cost. These findings highlight the importance of model architecture selection according to application requirements, which include resource constraints, dialogue complexity, and desired quality. The results confirm not only the huge potential of transformer architectures and large language models in transforming conversational AI but also the need for further research in fine-tuning and optimizing these models for practical deployment. This work contributes to further understanding modern conversational models and their applicability, thus setting the stage for future innovations in the field.

References

Sutskever, I. (2014). Sequence to Sequence Learning with Neural Networks. arXiv preprint arXiv:1409.3215.

Vinyals, O., & Le, Q. (2015). A neural conversational model. arXiv preprint arXiv:1506.05869.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. Advances in neural information processing systems, 28.

Luong, M. T. (2015). Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.

Liu, X., Zhang, J., Zhang, H., Xue, F., & You, Y. (2023). Hierarchical dialogue understanding with special tokens and turn-level attention. arXiv preprint arXiv:2305.00262.

Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Radford, A. (2018). Improving language understanding by generative pre-training.

Brown, T. B. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification?. In Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October

18–20, 2019, proceedings 18 (pp. 194-206). Springer International Publishing.

Zhu, X., Li, J., Liu, Y., Ma, C., & Wang, W. (2023). A survey on model compression for large language models. arXiv preprint arXiv:2308.07633.

Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., & Dolan, B. (2016). A persona-based neural conversation model. arXiv preprint arXiv:1603.06155.

Zhang, Y. (2019). Dialogpt: Large-Scale generative pre-training for conversational response generation. arXiv preprint arXiv:1911.00536.

Wu, Y., Lu, W., Zhang, Y., Jatowt, A., Feng, J., Sun, C., ... & Kuang, K. (2023, July). Focus-aware response generation in inquiry conversation. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 12585-12599).