JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Predictive Text generator

Khushi Solanki, Sonali Patil, Asst. Prof. Ms. Pranali Patil

[1,2]Keraleeya Samajam's Model College, Khambalpada Road, Thakurli, Dombivli (East), Kanchangaon, Maharashtra

[3]Guide, Keraleeya Samajam's Model College, Khambalpada Road, Thakurli, Dombivli (East),Kanchangaon, Maharashtra

This paper explores the evolution of predictive text generation, focusing on state-of-the-art techniques in natural language processing (NLP), including machine learning algorithms and deep learning models. We provide a detailed analysis of the current landscape, describe the development of our own predictive text model, and assess its real-world applications. The results indicate that our model improves text completion efficiency, showcasing its potential in enhancing user interaction across various domains.

I. INTRODUCTION

Predictive text generation has become an essential feature in modern communication tools, from smartphones to virtual assistants. This technology leverages natural language processing (NLP) techniques to predict and suggest the next word or phrase, streamlining the typing process and enhancing user efficiency. Early models, like N-gram-based systems, relied on simple statistical patterns, predicting the next word based on frequency of usage in a given dataset. However, advancements in machine learning and the introduction of deep learning architectures, such as Long Short-Term Memory (LSTM) and Transformer-based models like GPT, have revolutionized predictive text by improving contextual understanding and accuracy.

Despite these improvements, challenges remain. Predictive text models must handle the complexities of human language, including context shifts, slang, and idiomatic expressions, while maintaining efficiency in real-time applications. Additionally, ensuring that predictions are accurate and contextually relevant across diverse languages and user inputs is a significant hurdle. This paper explores the evolution of predictive text generation, discusses current state-of-the-art techniques, and presents a custom-built model that seeks to address some of these challenges. The objective is to contribute to the growing body of research aimed at enhancing the accuracy, efficiency, and applicability of predictive text systems across various platforms.

II. METHODOLOGIES

1. Material and Procedure

To build the predictive text generator, we first collected a diverse dataset from publicly available sources, including Wikipedia, news articles, and conversational datasets such as OpenSubtitles and Reddit discussions. This dataset was preprocessed through tokenization, lowercasing, and the removal of special characters. Subword tokenization (e.g., Byte-Pair Encoding) was used to handle rare words and ensure the model could effectively predict both common and uncommon words. This preprocessing step was crucial for ensuring that the model could generalize across various text inputs and contexts.

We employed a Transformer-based architecture, specifically using a Generative Pre-trained Transformer (GPT) model due to its proven ability to handle long-range dependencies in text prediction tasks. The model was pre-trained on a large corpus for general language understanding and then fine-tuned on smaller, domain-specific datasets for enhanced contextual prediction. The model's performance was evaluated using perplexity, top-K accuracy (K=1, 3, and 5), and response time to ensure it met the requirements of real-time text prediction. We also conducted qualitative analysis to assess the fluency and relevance of generated text, ensuring the model performed well across a variety of input sequences.

Finally, the model's results were compared against baseline models, including traditional N-gram models and an LSTM-based model, which were trained on the same dataset. This comparison helped to demonstrate the improvements in contextual accuracy and efficiency brought about by the Transformer architecture. Ethical considerations were also addressed by ensuring that the data used did not contain personally identifiable information (PII), and efforts were made to mitigate biases in the dataset to ensure fair and accurate predictions.

III. MODELING AND ANALYSIS

The predictive text model developed in this study was based on the Transformer architecture, specifically the Generative Pre-trained Transformer (GPT). We selected GPT for its ability to capture long-term dependencies in text, allowing it to predict words based on an entire sequence rather than just the immediate preceding words. This makes it superior to traditional models like N-gram or LSTM, which are limited by shorter context windows. The model's architecture included 12 attention layers, 12 attention heads, and a hidden size of 768, allowing it to efficiently process text and generate predictions with high contextual relevance. Additionally, we used subword tokenization to handle rare and out-of-vocabulary words effectively, ensuring that the model could predict across a wide range of language structures.

Training was conducted in two stages. First, the model underwent pre-training on a large corpus of general text data, such as Wikipedia and news articles. This phase enabled the model to learn grammar, syntax, and common word patterns. The pre-trained model was then fine-tuned on more specialized datasets like conversational data from Reddit and OpenSubtitles, allowing it to adapt to different language registers and styles. During training, we optimized the model using the Adam optimizer, with a learning rate of 1e-4 and a batch size of 64. Early stopping and dropout regularization were employed to prevent overfitting, ensuring the model could generalize well to unseen text data.

To evaluate the model's performance, we used perplexity as a key metric, measuring how well the model predicted the next word in a sequence. A lower perplexity score indicated better prediction accuracy. The model achieved a perplexity score significantly lower than the N-gram and LSTM baselines, demonstrating its superior ability to generate contextually relevant text. We also calculated top-K accuracy, where K=1, 3, and 5, to measure the percentage of cases where the correct next word was among the top predicted words. Our model outperformed traditional methods, achieving over 80% accuracy for top-3 predictions. Additionally, response time was evaluated to ensure the model's real-time applicability, with average prediction times remaining within acceptable limits for use in mobile or web-based applications.

We conducted qualitative analysis by feeding the model various test inputs and analyzing its outputs. The Transformer-based model exhibited a robust understanding of context, generating highly relevant and grammatically accurate predictions across different scenarios, such as formal and informal language use. However, in some edge cases—such as highly idiomatic expressions or complex sentence structures—the model occasionally made irrelevant or off-topic predictions. Despite these limitations, the overall performance demonstrated a significant improvement over baseline models, confirming the effectiveness of the Transformer architecture in predictive text generation.

IV. CONCLUSION

In this study, we explored the evolution and development of predictive text generators, focusing on the capabilities of modern Transformer-based models like GPT. Our custom-built model demonstrated superior performance in predicting contextually relevant text, outperforming traditional methods such as N-gram and LSTM models in terms of accuracy and efficiency. While the model excelled in most cases, challenges remain in handling complex sentence structures and idiomatic expressions, presenting opportunities for future improvements. Overall, the advancements achieved in this research contribute to enhancing real-world applications like typing assistants and chatbots, paving the way for more efficient and intelligent user interaction systems.

V. REFERENCES

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1706.03762
2. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
3. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... & Amodei, D. (2020). *Language models are few-shot learners*. In Advances in Neural Information Processing Systems, 33. https://arxiv.org/abs/2005.14165
4. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781. https://arxiv.org/abs/1301.3781
5. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence to sequence learning with neural networks*. In Advances in Neural Information Processing Systems (pp. 3104-3112). https://arxiv.org/abs/1409.3215
6. Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). *OpenNMT: Open-source toolkit for neural machine translation*. arXiv preprint arXiv:1701.02810. https://arxiv.org/abs/1701.02810
7. Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory*. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735
8. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186). https://doi.org/10.18653/v1/N19-1423
9. Goldberg, Y. (2016). *A primer on neural network models for natural language processing*. Journal of Artificial Intelligence Research, 57, 345-420. https://arxiv.org/abs/1807.10854
10. Jurafsky, D., & Martin, J. H. (2019). *Speech and Language Processing* (3rd ed.). Pearson.