Generating Personalized Food Recipes Using NLP Techniques

Varad Nevasekar
02006225
Dept. of Computer Science
University of Massachusetts Lowell
varad_nevasekar@student.uml.edu

Udith Lakshmi Narayan
02007636
Dept. of Computer Science

University of Massachusetts Lowell
fnu_udithlakshminarayan@student.uml.edu

Anuja Tidke
02081854
Dept. of Computer Science
University of Massachusetts Lowell
AnujaAnil_Tidke@student.uml.edu

***Abstract*— This project aims to develop a cutting-edge system that leverages Natural Language Processing (NLP) techniques to generate personalized recipes based on user-provided keywords, categories, and ingredients. The system will empower users to input their culinary preferences, dietary restrictions, and available ingredients, and it will swiftly deliver tailored recipe recommendations. By intelligently understanding user queries, the system will provide highly relevant and enticing recipes. This innovation not only streamlines the recipe discovery process but also encourages culinary exploration and learning. The project signifies a remarkable synergy between technology and gastronomy, making cooking more efficient, enjoyable, and personalized.**

Keywords*: *Natural Language Processing, Keywords, Keyword Based Recipe Generation, Text Generation

I. INTRODUCTION

The demand for personalized recipe generation, facilitated by NLP, arises from various factors. Diverse dietary preferences necessitate recipes tailored to individual needs. Ingredient constraints due to location, season, and personal choices require recipes that match available ingredients, reducing food waste. Health-conscious individuals aim for low-calorie, high-protein, or heart-healthy meals, emphasizing the need for personalized nutritional recipes. Additionally, busy lifestyles call for efficient cooking solutions with quick and convenient recipes. These diverse requirements underscore the significance of personalized recipe generation using NLP, offering solutions that cater to individual dietary, ingredient, health, and time constraints. This NLP project seeks to enhance semantic comprehension and natural language generation for recipe generation, aiming to deliver highly pertinent culinary instructions. By addressing challenges in data handling, scalability, and adaptability, we aim to advance NLP boundaries and open doors for personalized content generation and recommendations across domains. This innovation can benefit the food industry and researchers in NLP, AI, and culinary science by offering technology that fosters exploration of semantic comprehension, personalization, and user behaviour modelling[1]. The project leverages a range of Natural Language Processing (NLP) models to create personalized recipes. The initial baseline model, LSTM, undergoes pretraining with DistilGPT2 and is characterized by specific architecture parameters. The Falcon-7B model, known for its superior performance, is a causal decoder-only model trained on extensive data. Additionally, an Encoder-Decoder with Attention Model is utilized, with defined parameters for both the encoder and decoder. Task-specific models, namely T5 for Question-Answer and T5, are also incorporated.

II. RELATED WORK

[1] tackles a similar problem where the researchers are using user historical preferences to generate new recipes. They use an encoder decoder model and emphasize the importance of attention, while mentioning that larger scale model could improve performance. [2] is the original paper introducing sequence to sequence model and proposes the encoder decoder model, which helps further clarify this model. [3] covers the methodology behind the T5 model, exploring how Text to Text transformers can utilize transfer learning for several common tasks. The relevant tasks for this project are “question answering” and “summarization.” The limitation of using pre-trained models is that you must assume that your task is similar enough to the pre-trained task to meaningfully improve performance over training a new model. Furthermore, transfer learning in NLP enables the transfer of lexical knowledge, focusing on information related to words and their meanings. However, it may not necessarily extend to semantic knowledge, which entails a more profound comprehension of the meaning conveyed in phrases, sentences, or texts. Another relevant study, Chef Transformer [4], employs T5 for the summarization task, where the recipe corresponds to the summary, and the ingredients of the recipe corresponds to the original text.

III. METHOD

To survey several different methods for natural language generation, we divided our approaches into two different methods of managing data: causal modelling and marked modelling. A causal language model will predict next token in a sequence of tokens and can only read left to right, but a masked language model predict from a masked token sequence where the entire input is passed and then an inference is made before generating an output.

a Causal Models

For the preprocessing to train the causal models, we took the relevant recipes and joined the title to the ingredients and instructions to create a single string as an input. These strings were then tokenized using the appropriate tokenizer for each model and then target string was generated by offsetting the string by a single token.

i LSTM

For our baseline model, we created a model that consisted of a 512 size embedding layer, 2 LSTM layers with 256 hidden units, and an output layer the size of the vocabulary. Unlike the other models used, this LSTM model tokenizes on a character level. The tokenizer uses a dictionary to map a unique id to all possible characters. After some experimentation with smaller datasets, we settled on training the model for 5 epochs with a learning rate of 0.001.

ii Pretrained - distilgpt2

DistilGPT2[5] is the smallest version of GPT2 trained on supervised tasks, so we used the packaged tokenizer and trained this model for 2 epochs using a learning rate of $2*10^{-5}$ with 0.01 weight decay. DistilGPT2 is the most widely used causal language model on Huggingface for text generation, so it felt important to use it in a trial.

iii Pretrained - falcon-7b

Falcon-7b[6] is a large language model and a causal encoder-decoder model that was trained on billions of web pages. It was the highest-scoring LLM on Huggingface's Open LLM Leaderboard at the time of this project and could be feasibly loaded on the available hardware. Due to the immense size of the model in comparison to the processing power we had access to there was not much experimentation with hyperparameters done, so we chose to use the same parameters as DistilGPT2.

b Marked Models

For the preprocessing to train the marked models, we took the relevant recipes and joined the ingredients and instructions to create the target text and used the title as an input. These strings were then tokenized using the appropriate tokenizer for each model.

i Encoder-Decoder with Attention

We used the same preprocessing that we implemented with the LSTM. The model itself followed the encoder-decoder architecture described in [2]. We had an encoder with an LSTM unit with 128 hidden units and a decoder with attention where the decoder also used an LSTM with 128 hidden units. While considering training time, we trained this model for 2 epochs using a learning rate of 0.001.

ii Pretrained – t5 (question answering)

Our initial intuition suggested that t5's question-answering task would be the best suited for the project's task, as one can see the title as a question that the recipe is the answer to. T5 similar to the other pretrained model was used with data tokenized with the packaged tokenizer. It was also trained for 2 epochs using a learning rate of $2*10^{-5}$ with 0.01 weight decay.

iii Pretrained – t5 (summarization)

To replicate the results of [4], we applied the summarization task of t5 to our data as well with the same parameters for training of 2 epochs using a learning rate of $2*10^{-5}$ with 0.01 weight decay.

To generate outputs from each of the six models we trained we sampled 25 titles from our dataset and then generated an output which we compared using the metrics described below.

Our methodology surveys a wide range of approaches to doing NLG with recipe data and can indicate which approach might be worth developing further.

IV. DATASET

Our dataset is called Recipe Box and it consists of approximately 125,000 recipes from various food websites. Recipes typically consist of several components: a recipe title, a list of ingredients and measurements, instructions for preparation, the website link, and a picture of the resulting dish. The website link and the image of the dish were dropped. Additionally, a limit of 2000 was set on the total recipe length, which is the sum of ingredients and the instructions. The histogram shows the recipe length of the entire dataset, which shows that most of the recipes fall under the 2000 length category. This limit dropped the data to approximately 80,000 recipes. Table 1

shows dataset summary statistics,

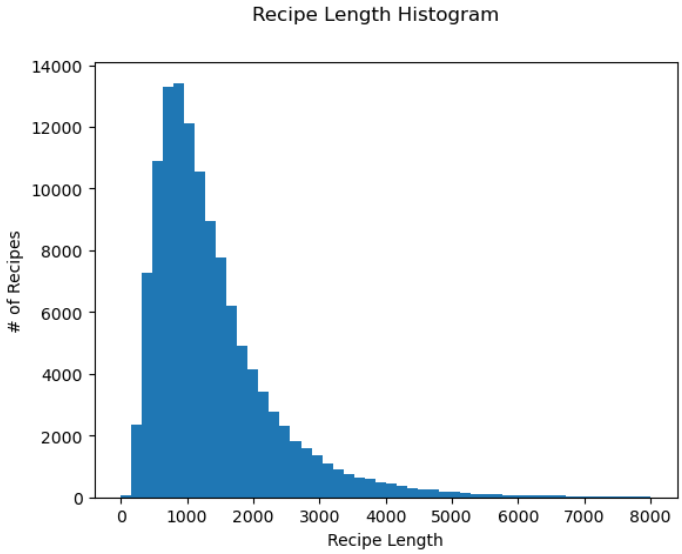

Table 1	
	#Recipes
Raw	122,938
Processed	80,784

V. RESULT

We detail our evaluation approach and the metrics used to assess the performance of our recipe generation models. We employed the following key evaluation components:

BLEU Score: We used the BLEU (Bilingual Evaluation Understudy) metric to evaluate the quality of generated recipes quantitatively. BLEU measures the similarity between generated and human-authored recipes, providing an indication of the model's language generation capabilities. We calculated the BLEU score by first calculating precision scores for N-grams (word sequences) in the generated text and reference texts. Then, we apply a brevity penalty to account for overly short generated texts. Finally, we compute the geometric mean of N-gram precision scores, weighted by the brevity penalty[7].

Human Analysis: We manually verify recipe coherence for 25 generated recipes from each model. We verify the recipe coherence based on 5 metrics and their scores, the generated ingredients(0 or 1), the generated instructions(0 or 1), the ingredients mentioned in the instructions(0 to 5), relevance(0 or 1) and the overall quality of the generated recipe(0 to 5).

Model	BLEU Score	Human Analysis Score
LSTM	0.03035	3.41667
DistilGPT2	0.01409	10.12
Falcon-7B	0.00747	2.64
Encoder-Decoder with Attention	0.00615	1.0882
T5(Q&A)	0.02153	3.58333
T5(Summarize)	0.01360	5.08333

Model Comparison: We compared the performance of our six models based on the BLEU scores. This allowed us to identify which models excelled in generating recipes that closely matched human-authored ones.

Human Verification: Human evaluators provided valuable feedback on the quality of the generated recipes, which helped validate the model-generated content.

Insights: As compared to other models, the DistilGPT2 model demonstrated a strong correlation between recipes and their ingredients in the generated output. Additionally, it created a fictional website layout that included designated spaces for photographs. For LSTM, it was noted that there was no discernible correlation between the instructions and the ingredients. For falcon-7b, Ingredients were absent in the generation of almost all recipes, while the instructions fared well for the test set. The T5 (question answer) model generates outputs in a variable manner, providing only ingredients in some instances, only instructions in others, and both ingredients and instructions at times. For T5 summarize, the set of instructions was the longest when compared to other models.

VI. CONCLUSION

From our results, we were able to conclude that DistilGPT2 was able to produce the most relevant recipes since it performed the highest on the Human Scorer metric, while the LSTM was able to replicate the database the best lexically. If taken at face value, the results suggest that the simpler models are better adapted to the problem. However, our research capabilities were limited by the resources available to us. For instance, when training the LLM (Falcon-7b) on the GPU servers available to us, the training was interrupted due to the checkpointing filling the available hard disk space, so the tested model is a recovered checkpoint from an unknown point in training. Furthermore, BLEU is a flawed metric for this task as it is heavily dependent on the reference, because it can rate a semantically correct result low due to large differences in the generated sentence structure. Interestingly, of the pretrained transformers, DistilGPT2 was the only model that is trained on just English. The other pretrained transformers would have had to unlearn the other languages they were trained on to generate more relevant results to our

English dataset. Also, all the causal models except Falcon-7b performed better than the marked models, which means feeding in the input with causality seems to provide a significant improvement in recipe relevancy. If we account for these limitations of our experiment, we can conclude that pre-trained causal transformers trained on an English dataset are most likely to generate relevant recipes.

Future work could involve finetuning an LLM trained on english for longer epochs on this recipe dataset to avoid having to unlearn other features.

VII. CONTRIBUTION CHART

Task/Sub-Task	Student ID	Contribution
-Pre-trained Transformers -Encoder Decoder with Attention	02006225	
-LSTM -Preprocess -Metric Analysis	02081854	
-LSTM -Preprocess -CUDA implementation	02007636	

REFERENCES

[1] B. P. Majumder, S. Li, J. Ni, and J. McAuley, "Generating Personalized Recipes from Historical User Preferences," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 5976-5982, 2019.

[2] Sutskever, Ilya, et al. Sequence to Sequence Learning with Neural Networks. arXiv:1409.3215, arXiv, 14 Dec. 2014. arXiv.org, https://doi.org/10.48550/arXiv.1409.3215

[3] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," J. Mach. Learn. Res., vol. 21, no. 140, pp. 1-67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html J. Mach. Learn. Res., vol. 21, no. 140, pp. 1-67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[4] "Chef-transformer (chef transformer)," chef-transformer (Chef Transformer), https://huggingface.co/chef-transformer (accessed Dec. 9, 2023).

[5] Hugging Face, "DistilGPT2 Model," Hugging Face. [Online]. Available: https://huggingface.co/distilgpt2. Accessed: December 9, 2023.

[6] Tiiuae, "Falcon-7B Model," Hugging Face, https://huggingface.co/tiiuae/falcon-7b. Accessed: Dec. 9, 2023.

[7] BLEU score calculation: https://www.nltk.org/_modules/nltk/translate/bleu_score.html

[8] R. Lee, "Recipe box," 2022. [Online]. Available: https://eightportions.com/datasets/Recipes/#fnref:2. [Accessed: December 9, 2023].