Improvements in Semi-Structured Text Generation for Cooking Recipes

Joel Sageau
Université de Montréal

Kevin Lessard
Université de Montréal

Mathieu Péloquin
Université de Montréal

{joel.sageau; kevin.lessard; mathieu.peloquin.1}@umontreal.ca

Abstract

Producing cohesive semi-structured texts is a challenging task. In recent years, a lot of researchers have suggested new approaches to accomplish this task. However, these techniques still have difficulty producing high quality context-aware texts for cooking recipes. The objective of our project is to make use of the most recent advancements in the field of natural language processing to generate recipes based on a user-supplied list of ingredients. We want to generate high-quality instructions and surpass the suggested baseline using the most recent high quality dataset of culinary recipes. Past papers have produced poor results using inferior datasets. By using the latest advances in text generation, we will be able to produce better formed culinary recipes. We will compare our results with the dataset's baseline using the same metrics.

1 Introduction

There are several recipe websites that provide instructions to prepare delectable dishes. However, these websites often provide static recipes, thus it is the user's responsibility to discover the right recipe for the ingredients they wish to use. Wouldn't it be helpful to input a list of items and use natural language processing to return a recipe based solely on your ingredients?

This brings us to the task of semi-structured text generation, where we enter components and obtain a final list of ingredients with valid quantities and a corresponding complete recipe with instructions as output. To help us in this task, we were introduced to a brand-new, sizable, state-of-the-art dataset from the paper called *RecipeNLG: A Cooking Recipes Dataset for Semi-Structured Text Generation* (Bień et al., 2020), which contains complete cooking recipes that are suitable for semi-structured text generation. More specifically, this paper provides a useful dataset and establishes the baseline for generating culinary recipes from the names of input ingredients.

Our first order of business would be to make a pipeline that trains models for the task of cooking recipe generation, and then to reproduce the experiments of the previously mentioned paper. Afterwards, our objective would be to investigate more modern models and techniques that would be compatible with the structure of our dataset in order to develop a new model that outperforms the original baseline from (Bień et al., 2020). In their work, (Bień et al., 2020) generate recipes with a restricted list of ingredients, but also recipes with not only the input list of ingredients but also additional ingredients. Since we are trying to reproduce the study, our main objective would be to do both generative tasks. Furthermore, if we have enough time, we would like to implement an intuitive user interface to test our models.

2 Related work

It is necessary to first talk about (Marín et al., 2018) which proposes the *Recipe1M+* dataset. This was the largest collection of cooking recipes that could be used to generate texts. However, it also has multiple shortcomings. Several data points have duplicated or misaligned recipe structures. Generally, recipes steps were created from segmented sentences and the fraction symbol is not present in the quantities. This results in a dataset that cannot be used to generate feasible recipes.

The enhanced version of the dataset, *RecipeNLG* (Bień et al., 2020), which was produced using *Recipe1M+*, addresses the issues mentioned above and is approximately double the size of its predecessor. Combined with this new revised dataset, (Bień et al., 2020) experimented on the task of recipe generation, using a version of GPT-2 that was fine-tuned to create a baseline for this dataset.

Their model can be tested online[1].

Another related article, (Majumder et al., 2019), took a similar approach to the challenge of generating personalized recipes from user historical preferences. This research also presents a new dataset of 180K+ recipes and 700K+ user reviews coming from Food.com [2] which is useful for this sort of task. The model they suggest is an older combination of bidirectional Gated Recurrent Units (BI-GRU) as the encoder, using Badhanau Attention over the encoded ingredient, and a GRU for the decoder.

The work (Li et al., 2022), which offers the fundamental ideas, key methods, and most recent advancements in text generation based on pre-trained language models (PLM), is a crucial study to explore in order to find potential solutions to our problem.

3 Task and dataset

Our model's job is to take as input a list of food components functioning as entities and produce a complete culinary recipe with a title, an ingredient list with quantities and step-by-step instructions. This task goes hand in hand with the dataset.

The dataset we intend on using is *RecipeNLG* (Bień et al., 2020), a cooking recipe dataset which was specifically built for semi-structured text generation. There are over 2 million cooking recipes in this collection. More than one million of them are updated, deduplicated versions of recipes from *Recipe1M+* (Marín et al., 2018). About 1.6 million of these 2 million recipes are of higher quality. Cooking directions follow a precise structure that includes a title, an ingredient list with given quantities and step by step instructions. Entities like the quantity, its unit name and the component name are included in the ingredient list. Quantities should be aligned with the serving size offered by the result of the recipe.

Overall, the dataset has respected the logic of the recipe in terms of the components, the proportions and the states of those ingredients. The model's innovative output must reflect this. (Bień et al., 2020) had to sanitize the recipes taken from *Recipe1M+* (Marín et al., 2018) by removing the excessive amount whitespace characters and replacing Unicode symbols with their ASCII equivalents. They also removed duplicate recipe entries and fixed common amount fraction errors such as writing one half as "12" rather than "1/2." This dataset will be essential to the accomplishment of our goal.

4 Methods

We plan to first reproduce the results of the paper and then apply latest techniques around text generation in order to provide the next state-of-the-art model for recipe generation. Then, in an effort to further enhance recipe development, we would try to apply fresh concepts directly to the dataset.

(Bień et al., 2020) employed a named entity recognition (NER) to teach a model to identify what an ingredient is by making them entities. Additionally, they carried out a number of post-processing procedures to make sure the data was prepared to be used as an input by their model. Afterwards, they applied the Hugging-Face implementation of the pre-trained GPT-2[3]. Then, the model was given a set of ingredients and instructed to come up with complete, logical recipes. Additionally, a collection of control tokens was prepared and included in the dataset. To expedite the training process, they also combined numerous tokenized recipes into a single context.

Hardware wise, (Bień et al., 2020) used Cloud TPUs from Google[4] in order to train their NER and Hugging Face GPT-2 model.

To reach our goal, we want to investigate the newest text creation strategies and advancements. We have a decent chance of acquiring better coherent output recipes and evaluation outcomes since there are newer, better existing models than GPT-2.

5 Baselines and evaluation

We will compare to (Bień et al., 2020)'s results. The cosine similarity determined using the TF-IDF format is what we'll be employing. The similarity between a created recipe and its gold standard equivalent might then be evaluated. This is the evaluation method used by (Bień et al., 2020). We will also use it to contrast our findings with theirs. Additionally, we want to determine how many grammatical errors there are in our recipes using the LanguageCheck spell and grammar checker. Additionally, if required, we may develop additional assessment metrics that are unique to our dataset and use them to evaluate our models.

[1] https://recipenlg.cs.put.poznan.pl/generator

[2] https://www.kaggle.com/datasets/shuyangli94/food-com-recipes-and-user-interactions

[3] https://huggingface.co/gpt2

[4] https://sites.research.google/trc/about/

References

Michał Bień, Michał Gilski, Martyna Maciejewska, Wojciech Taisner, Dawid Wisniewski, and Agnieszka Lawrynowicz. 2020. RecipeNLG: A cooking recipes dataset for semi-structured text generation. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 22–28, Dublin, Ireland. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2022. A survey of pretrained language models based text generation. *CoRR*, abs/2201.05273.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian J. McAuley. 2019. Generating personalized recipes from historical user preferences. *CoRR*, abs/1909.00105.

Javier Marín, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2018. Recipe1m: A dataset for learning cross-modal embeddings for cooking recipes and food images. *CoRR*, abs/1810.06553.