CIS 530 Milestone 2 Report

Leon Zhou, Anshul Wadhawan, Anni Pan

December 3, 2021

1 GitHub Link

https://github.com/annypan/CIS-530-final-project

2 Evaluation Measure

Our first and most important metric is the **overall precision** and the **overall recall**. The overall precision measures the total percentage of predicted substitutes that are in the ground truth data, measured by the number of correct substitutes over the number of total substitutes predicted. The overall recall measures the total number of substitutes in the ground truth that appear in the prediction over the number of substitutes in the ground truth.

Besides having an overall precision and recall, we also have precision and recall for each ingredient that measures the precision and recall of the predicted substitutes for each individual ingredient among the selected 37 ingredients. This would be helpful later when we investigate into on what ingredients did our model perform well and didn't perform well.

3 Simple baseline

Our simple baseline is a KNN model that finds K nearest neighbors for a particular ingredient in the FoodBERT embedding space, and shortlists them on the basis of a thresholding criteria. The final approach is separated into two parts: The first part calculates text-based embeddings for up to 100 occurrences of every ingredient and optionally concatenates them with image-based embeddings. The second part employs these embeddings together with KNN and a further scoring and filtering step to predict substitutes.

Instructions to run the provided code in simple-baseline.py, are given in simple-baseline.md. Sample substitutes for four ingredients (salt, sugar, honey, pepperoni) are produced. If the test file path is given to the python file as arguments, the code finds out the substitutes for the provided ingredients and saves it in the provided prediciton file. On our test set, the model has a Precision of 0.784 and Recall of 0.139.

4 Strong baseline from paper

The strong baseline we used is GPT3 prompting. Specifically, we treated the ingredient replacement task as a text completion problem using the "davinci" engine. Under this few-shot learning setting, where the model are presented as few as 7 examples (carefully selected to not have overlap with the test set), the large language model was able to perform with a relatively high precision: Precision: 0.894, Recall:0.052. Since the question is framed as following:

Q:What is a good replacement for ingredient1?

A: ingredient2.

the gpt3 model is prompted to return only a single ingredient replacement. Consequently, the recall rate can be much higher once we compile a better set of few shot examples with multiple ground truth replacements. As a next step, we will fix the above mentioned problem to increase recall rate as well as try to use other language models that have few-shot capabilities.