Homework 1: Reddit Cooking NER and Data Cleanup

Name entity recognition (NER) is an important step in text mining and information extraction. The goal of the task is to identify and classify entities found in text data to specific labels and categories that are predetermined by the analyst. The goal of this assignment is to create and train an NER model to, as accurately as possible, identify and predict which entities belong to which labels. We are looking at the text data found on Reddit cooking forums to teach our model how to label entities as either a ‘DISH’, ‘INGREDIENT’, or ‘EQUIPMENT’.

The first model I create would just be trained by manual annotations using prodigy to establish a benchmark that I can then use to improve off of using different annotation techniques. After annotating 60 different text samples, I trained the model to achieve an average accuracy (based on the f1-score) of 6%, with some iterations going as low as 0% and as high as 24%. After establishing this benchmark, my next goal was to find other possible techniques to improve the accuracy. I decided that the next best method to implement was active learning where we are given a binary choice of whether the model is annotating correctly or not. By emphasizing the annotations that are more ambiguous, wrong, or partially correct, I hoped to create a much more accurate model. I also felt like the model needed many more manual annotations than I gave it from the original baseline of 60 examples. Thus, I annotated more than 200 more entities and trained the model for an average f1-score of approximately 54%, where some were as low as 40% and as high as 61%. This large increase in accuracy was thanks in part to the large number of annotations I completed manually and the active learning that helped improve the model.

I believed that the model could be improved just a bit more so I decided to implement another technique by correcting annotations my model has done using “ner.correct”. After training the model on 114 corrections, that average accuracy increased from 54% to 62% with one being as low as 44% and one as high as 70%.

Besides the annotated dataset that I created, there was another dataset that was annotated with ChatGPT version 3.5. I did not want to use that dataset because I did not trust the annotation of a large language model; but just to cover all bases, I decided to merge the dataset that I improved with the ChatGPT dataset to see if it might help improve the accuracy. After merging the two datasets, I trained the model again and achieved an average accuracy of approximately 33%. This a significant downgrade so I decided to exclude the ChatGPT dataset from my model. This is most likely because my model as already been refined and correct more than the ChatGPT dataset, making their combination worse than my own dataset. With more time, I could have implemented one final technique by using patterns in a jsonl file to make rules that the model can easily follow. These rules are generated based on simple, hard rules that apply in almost every single situation. For example, one pattern that I would implement would be {"patterns": {"lower": "salt"}, "label": "INGREDIENT"}, which means that every time “salt” appears, it is labeled as an ingredient.

My best model, with an average f1-score of 62%, involved a large number of manual annotations, active learning, and then correcting those annotations. Unfortunately, merging with the ChatGPT dataset made it worse. Next time, I can implement patterns and rules for the model to follow.

ANNOTATION GUIDLINE:

TASK - My goal is to identify and annotate dish, ingredient, and equipment entities found in Reddit cooking forums.

A **dish** is food prepared and served after combining ingredients and following a recipe.

**Ingredients** are food items that are used to prepare a dish. They can range from whole food items, to seasonings, garnishes, spices, and herbs.

**Equipment** refers to tools used on ingredients to help prepare dishes and helps serve the meal.

To determine which label that entity belongs to, we need to carefully scrutinize the entire context; if that entity is used as a food item to help prepare another food item, it is an ingredient. Conversely, if that food item is the end product of a process using other ingredients, that entity is a dish. If the entity is a nonedible tool that is used in preparation of a food item or used to serve that food item, that entity is equipment.

The largest problem when annotating with the aforementioned labels is how to completely differentiate some dishes and ingredients. Depending on the context, some entities can be correctly labeled as either a dish or an ingredient.

- “It’s a good sandwich topping, it’s good with baked brie, it’s great in smoothies”

This example is difficult because each of the 3 entities are used in different ways as a food item with another food item that is not mentioned. How they are used determines what label to annotate it with. In the first example, the unnamed food item is used as a topping on the sandwich, since it used as an ingredient to make the sandwich, the sandwich is a dish. For the second example, the baked brie is served *with* the other food item so it is considered an ingredient. The last example, the unnamed food item is used *in* the smoothies so the smoothies are clearly dishes.

As I mentioned before the best way to differentiate between the two would be to see if they are the end product or part of the process. Another way to determine very quickly is to see if there is a measurement associated with the entity; with measurements, the entity is most likely an ingredient because that is indication that it is part of a recipe to create a dish.

- “Strained and blotted. I used a jar of pizza sauce. 14oz.”

With the use of the measurement (14oz), it is clear to see that pizza sauce is the ingredient used in this recipe.

- “2. Sauce. 1 can san marzano tomatoes, salt, pepper, olive oil, basil (optional).”

Sometimes the measurements do not have to metric or imperial units; they can be ways to measure that specific ingredient (i.e., 1 can, 3 bags, or 2 spoons). Usually if food items are in a long list, they are most likely ingredients.

Sometimes, the entity includes words that help specify more of what type of dish, ingredient, or equipment it is. It is important to annotate the whole name:

- “I have the Kuhn Rikon peeler. It is excellent!”

In this example, annotating the correct annotation includes the brand of the peeler as well.

- “Although it is actually more a ham and potato stew with flat dumplings”

This example is more ambiguous because of the mixture of possible dishes and ingredients. Normally, both ham and potato are ingredients, but since it is attached as a description for the stew, the dish that should be annotated is “ham and potato stew”. Furthermore, flat dumplings are generally dishes, but since it is a part of making the stew dish, this classifies the flat dumplings, in this case, as an ingredient.

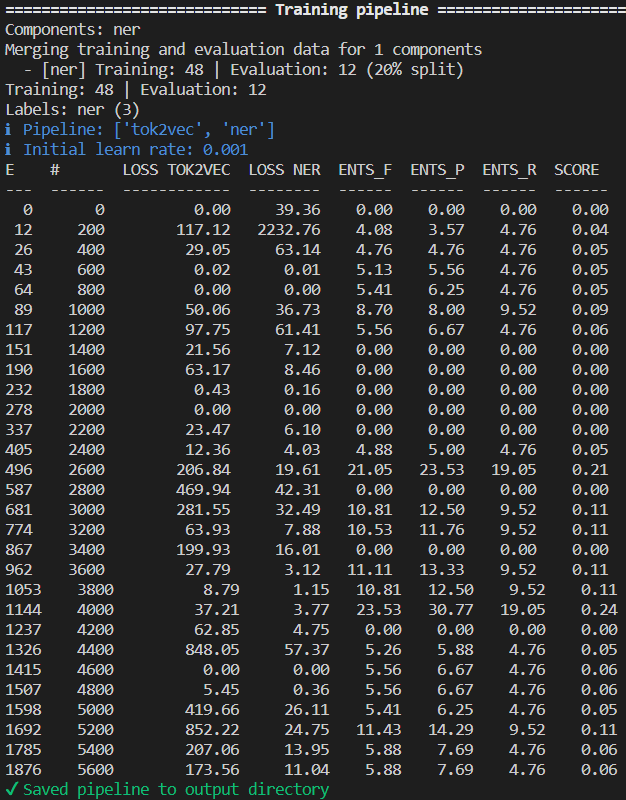
Compared to the other two labels, equipment is much easier to identify because it is a nonedible tool that prepares ingredients or serves dishes

- “Mine’s pretty much the same except the icemaker is in the freezer, not the door.”

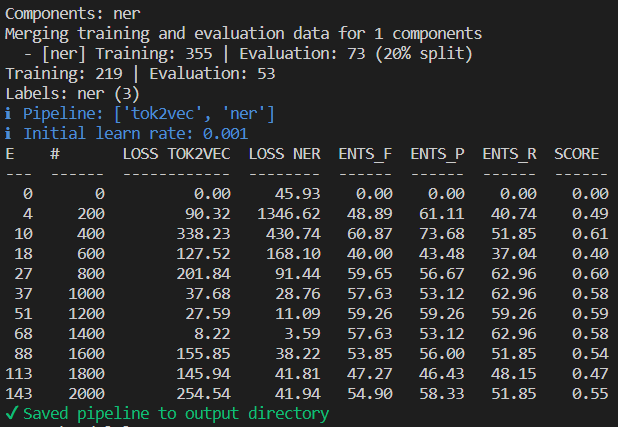
The equipment above is easily identifiable and relatively unambiguous.

APPENDIX

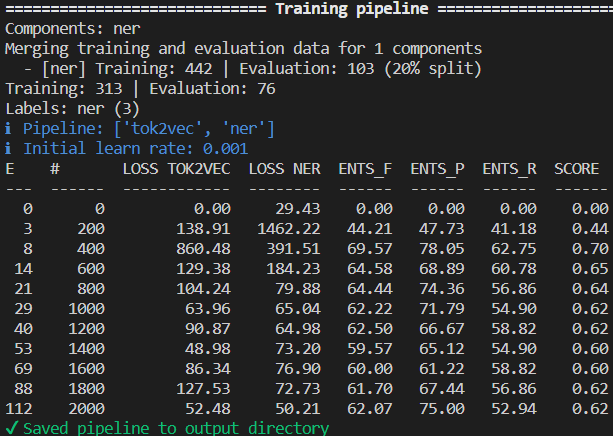
Baseline:



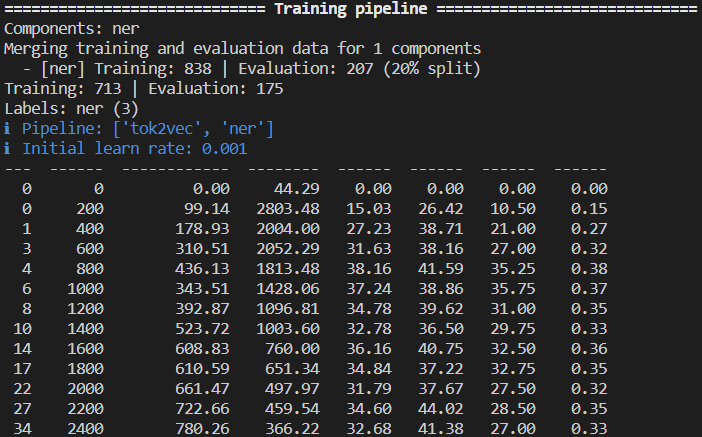
Post “ner.teach” output:



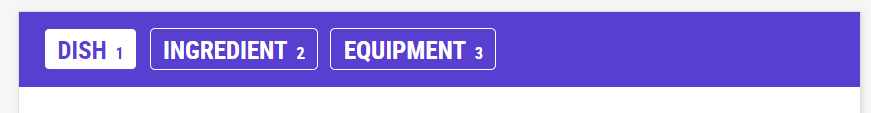
Post “ner.correct” output:



Merging with the ChatGPT dataset:



Labels used to categorize entities



Potential Implemental Patterns

