Natural language processing approaches applied to finding ingredient substitutions and complementing ingredients

Zach Casler, Isabella Eriksen, Ken Masumoto, Maxwell, Sanyam Savla

Abstract

Natural language processing (NLP) techniques and machine understanding of language is increasingly becoming a larger part of the general public's everyday lives. From virtual voice assistants on smartphones to customer service chatbots to parsing Chat-GPT prompts, NLP is used in a wide variety of contexts to understand and transform human language input into something a machine can process. One area of research that might prove to be especially useful moving forward is using NLP techniques to parse things other than speech or text input. Using these NLP techniques in contexts other than speech or text input could provide insights into those contexts as well as insights into understanding why these NLP techniques work. In this paper, we outline our research of a dataset of over 170,000 recipes and their ingredients and the NLP techniques we used to try to identify ingredient substitutions. We show that using sparse matrix representations of recipes falls short of being able to allow machines to parse recipes for adequate ingredient substitutions. However, this research was not without promise, as the limitations of our research could have contributed to this negative result.

Introduction

With the advancement of artificial intelligence, NLP and other related fields have also seen rises in popularity. Research into understanding these NLP techniques and how they might be able to be applied in situations outside of human speech or text is an interesting field of study that could provide insights into better understanding these contexts in addition to possibly providing a better understanding of the NLP techniques themselves. One such context in which NLP techniques might be applied are in ingredients in culinary recipes. There is importance in researching ingredient embeddings. They are beneficial for various goals, such as nutrient optimization or avoiding allergens and they provide insight on how ingredients are related to each other. Research in this field could allow for individuals to make quick exchanges for ingredients they may not have, as well as provide overall insights into using NLP techniques in contexts outside of human speech or text.

In this paper, we researched the following question: **Can we use sparse matrix representations of recipes to provide adequate ingredient substitutions for ingredients?** To answer this question, we used a dataset consisting of over 170,000 recipes scraped from posts spanning over 18 years on Food.com. We used sparse matrix representations of this data and calculated cosine similarities between ingredients to identify similar ingredients that might be used as substitutes for one another in a recipe. We found that for a curated list of common ingredients and their replacements found on popular food site allrecipes.com, the known replacement did not appear in the top five cosine similarities in 88% of cases. In addition, using that list of known ingredient substitutions, we found that the cosine similarity between known substitutions were statistically significantly lower than other ingredients with high cosine similarities that weren't necessarily adequate ingredient substitutes (co-occurrence matrix $p=0.02$; PPMI matrix $p=0.0003$). As a result, we concluded that using cosine similarities of sparse matrix representations such as co-occurrence and PPMI matrices fell short of being able to accurately identify substitute ingredients in recipes in the same ways it is able to identify synonyms in language.

Related Work

Our work is an extension of prior work on using NLP techniques in non-text, non-speech contexts, specifically in the field of recipes and ingredients. While previous work focuses on state-of-the-art or dense matrix approaches to finding ingredient replacements, our work focuses on using specifically sparse matrix representations to do so, as sparse matrix representations are often more computationally efficient, and thus allow for more accessible models to be created.

In *Exploiting Food Embeddings for Ingredient Substitutions*, Pellegrini, et al. used dense matrix approaches to try to find adequate replacement ingredients for a recipe. It was found that the lack of standardized evaluation metrics made it difficult to compare their results between methods to each other, which was also a limitation of our research. Despite this, they were able to create FoodBERT, a vector representation model for ingredient embeddings that they concluded was "well suited for substitute recommendations in dietary use cases." This suggests that using advanced, computationally-heavy, state-of-the-art NLP techniques are able to give adequate ingredient substitution recommendations.

In *Food Recipe Alternation and Generation with Natural Language Processing Techniques,* Pan, et al. used neural nets trained on word embeddings to try to provide ingredient replacements and generate entirely new recipes. Similar cosine similarity techniques were used to try to find similar ingredients, but they used recipe embedding vectors instead of co-occurrence or PPMI matrices. Results from this study were inconclusive, as the training data to create recipe embeddings was too limited.

These past works show that the use of computationally-heavy, state-of-the-art techniques to provide ingredient recommendations is a promising field. However, as shown by the conclusions of Pan, et al.'s work, ingredient or recipe vector embeddings require extremely large sets of data that might not be feasibly available for practical applications. Our work investigates if adequate ingredient substitutions can be found without the need to create computationally expensive ingredient vector embeddings, instead using much more available and easily-processed data for sparse matrix representations.

Data

The dataset that we used was found on Kaggle.com, but was originally acquired by scraping over 170,000 recipes from Food.com posts spanning over 18 years. Each recipe had information about the steps required as well as the ingredient list, though we did not take recipe steps into consideration for our research. Each ingredient in a recipe was given an ID, which could be mapped to its human-readable name through a mapping dataset. Ingredient names for this dataset were pre-categorized to fit into larger, less specific ingredient groupings. For example, the ingredients “iceberg lettuce” and “butterhead lettuce” were both categorized as just “lettuce” and were treated as the same ingredient in our dataset. While this allowed us to focus on larger ingredient groupings rather than minute differences between ingredients that are ultimately very similar, this also introduced room for error, as these groupings were created at the determination of the original dataset creator. Despite this room for error, this dataset was chosen over other similar datasets containing recipe information because of the breadth of the data provided (170,000 recipes was the largest amongst public and easily-accessible datasets) and the ingredient groupings allowed us to try to find substitutions between similar-but-not-identical ingredients (as ultimately, getting results that tell us that “purple grapes” and “green grapes” are very similar do not provide practically useful insights).

Approach and Experiments

Approach

The main idea of our approach was to treat ingredients in recipes in the same way words are treated in a corpus. We would use cosine similarity, a metric used in NLP to determine how similar two words are, to determine how similar two ingredients in our data were. In text data, two words with very high cosine similarities could be classified as synonyms and thus could be replaced with one another in many contexts. If our hypothesis was true, ingredients in recipes would act in a similar way, with ingredients with high cosine similarity scores with one another acting as replacements for each other in recipes.

As our approach focuses primarily on using sparse matrix representations to find ingredient substitutions, our first step was to create a co-occurrence matrix representation of our recipe data. This was done by parsing through each ingredient of each recipe and creating a count of each ingredient that appeared alongside that ingredient in a recipe. From there, we calculated the PPMI scores for each ingredient pairing in our dataset and stored those values in a PPMI matrix.

With the two sparse matrix representations created, we were able to move on to calculating cosine similarity between ingredients in line with our experiments and test the evaluation criteria that we developed.

Experiments and Evaluation Criteria

As noted by Pellegrini, et al., evaluation criteria in this field is especially difficult because of a lack of standardized evolution metrics. For the purposes of our research, we used two primary evaluation criteria. For the first evaluation criteria, we conducted a t-test between high cosine similar ingredients from our sparse matrix representations and the cosine similarities of known substitutions. We began with a list of 20 known ingredient substitutions provided by allrecipes.com. For each of those ingredients, we found the average of the top 5 cosine similarity scores across all other ingredients in the co-occurrence matrix that we created. Then we calculated the cosine similarity score between that ingredient and the known substitution provided by allrecipes.com. We then ran a one-tailed t-test between the top 5 averages and the known substitutions. We then repeated this process for the PPMI matrix instead of the co-occurrence matrix. The second evaluation criteria involved taking 20 common ingredients in recipes and finding the top 20 ingredients with the highest cosine similarities with that ingredient. Then we manually marked the number of those top 20 ingredients that could act as adequate substitutions. This manual analysis was subjective, as is unavoidable for non-binary and subjective outcomes that pertain to things like ingredients and human taste.

Results

Our results were decidedly mixed. Our experiment comparing top 5 cosines to known substitutions produced no meaningfully different results between raw co-occurrence and PPMI (both failed about 88% of tests). Further, we conducted a t-test that suggested that the cosine similarity of the average of the top 5 was statistically significantly higher than the known substitution. When ranking all ingredients by closeness to the target ingredient, the known substitutes had an average of 500th rank, out of the ~8000 total ingredients. Contextualizing this is the relatively large spread of these ranks, with a standard deviation of more than 900. Upon inspection, several known substitutes ranked much higher than the average, but the distribution is heavily skewed by a few distinct values. The median rank of known substitutes for the co-occurrence matrix was 35, but interestingly, the median for PPMI was 140. This despite the presumed increase in sophistication of the PPMI metric over raw co-occurrence. At first glance this suggests that neither sparse matrix approach is equipped to suggest substitutes with reliable accuracy.

Insights

To arrive at any insights, we drew inspiration from qualitative research and analyzed the top 20 nearest (by cosine similarity) ingredients to our targets. We labeled the ingredients in the top 20 that would be acceptable substitutes and observed their relative distances. The co-occurrence matrix surfaced an average of 4.5 out of 20 justifiable substitutes while the PPMI surfaced just

3.3 out of 20. While analyzing the results this way, it became clear that the raw matrix was repeatedly able to surface ‘similar’ ingredients better than the PPMI matrix, but only for common ingredients or ingredients with lots of substitutes(eg. broth and rice). In contrast, the PPMI matrix appeared better able to perform on niche or context dependent substitutions, outperming the raw matrix on surfacing turmeric in the top 20 for saffron, where the former is *only* a substitute for the latter in terms of the color it adds to a dish. Similarly, the PPMI matrix surfaced applesauce as a substitute for vegetable oil with a much higher rank than the raw matrix, where the former is only a substitute for the latter in the context of baking. This implies that while PPMI may not surface substitutes as consistently as the raw matrix, it navigates the complexity of the real world model more successfully.

A further note about PPMI is that it seemed unable to distinguish between ‘substitutes’ and ‘pairs.’ For example: garlic, ginger and onions are not substitutes for vegetable oil, but they are commonly fried in it and are thus ‘pairs’. Said another way, the PPMI matrix outperformed the co-occurrence matrix in ingredient pairings. A final note is that the average of cosine similarity between *any* (of the entire ~8000) ingredient pairs in the PPMI matrix was about 10x smaller than in the raw matrix, suggesting the raw matrix space is approximately 10x narrower].

Limitations and Future Work

Limitations

The scope of this project and the overall availability of data created notable limitations on this project that could have affected the results. The first major limitation was the size of the dataset. Language corpuses that are used effectively in practical applications are usually orders of magnitude larger than the recipe dataset that we had access to. As a result, these sparse matrix representations of the recipes are incomplete and it is possible that larger datasets of recipes could lead to more effective ingredient substitution recommendations. Another major limitation of our research was the evaluation criteria. As noted above, there is no standardized evaluation criteria to measure the effectiveness of the substitution recommendations given by our cosine similarity results. As a result, we had to develop our own evaluation criteria, which makes it difficult to compare our results to a standardized baseline or existing research. In addition, our evaluation criteria required us to manually verify some of our results, which could lead to human error in the evaluation process.

Future Work

One future area of work could be to include the recipe steps as features in our matrix representations. In text and/or speech input commonly used with NLP techniques, the semantic meaning of a sentence or idea can often be captured within a relatively small context window. However, for recipes and ingredients, the steps of a recipe add entirely new context to the ingredients that can’t necessarily be taken from the ingredients themselves.For example, the ingredient “cod” can taste very differently in a recipe depending on if it is fried or baked. As a result, incorporating the steps of the recipe in future work might lead to more accurate and meaningful results.

Another future area of work would be to explore other NLP modeling techniques. NLP is a rich feld, and just because sparse matrix representations did not produce accurate results does not rule out other NLP techniques from being practically applied to recipes and ingredients and producing meaningful results. In particular, semi-supervised methods might be useful in creating accurate ingredient substitutions.

Ethical Considerations

This work is meant to highlight ways to provide ingredient substitutions in recipes. This can be useful in providing substitutions for ingredients that someone might be allergic to, or a part of other dietary restrictions. This work can also help make recipes and cooking more equitable, as some ingredients might not be readily available for people in certain geo-locations or in particular income brackets. Providing ingredient substitutions for these groups that are more accessible or cheaper alternatives can help bridge the gap and allow for healthier home-cooked meals to be accessible in more homes.

Another ethical consideration of this work is surrounding the overarching idea of using NLP in non-human-text, non-speech contexts. The context of our research is relatively low-stakes and intended for good. However, there are applications of NLP in outside contexts that should be met with considerable ethical discussion before exploration. In particular, any application where the results could impact real people and their livelihoods should be met with extreme caution and consideration before proceeding.

Conclusions

The first major conclusion of our work is that our sparse matrix representations of recipe data did not adequately identify ingredient substitutions. There are a variety of possible explanations for this, as outlined in the Limitations section, with one of the most likely reasons being that the training data was not comprehensive enough for the methods chosen. This means that we were unable to confirm our hypothesis that sparse matrix approaches can be applied analogously between the NLP and recipe spaces. However, this does not mean that this research space is without promise. As outlined in the Future Work section, there are many factors that can be changed within our experiments that might provide better and more insightful results. Larger datasets might provide different results for this research question and entirely different analogous research questions between the NLP space and recipe ingredient space could also be explored.

The second major conclusion of our work is that the inclusion of recipe steps into future work should be one of the primary inclusions for any additional work in this field. There is no analogy between recipes and recipe steps for text and sentence data. In addition, the context of any given ingredient might change drastically depending on the cooking technique used. As a result, we've concluded there are likely to be massive blind spots in any future work that does not take into account the steps or cooking techniques used in a recipe.

References

Hu, G., Ahmed, M., & L'Abbé, M. R. (2023). Natural language processing and machine learning approaches for food categorization and nutrition quality prediction compared with traditional methods. *The American Journal of Clinical Nutrition*, *117*(3), 553–563. https://doi.org/10.1016/j.ajcnut.2022.11.022

Pan, Y., Xu, Q., & Li, Y. (2020). Food recipe alternation and generation with natural language processing techniques. *2020 IEEE 36th International Conference on Data Engineering Workshops (ICDEW)*. https://doi.org/10.1109/icdew49219.2020.000-1

Pellegrini, C., Özsoy, E., Wintergerst, M., & Groh, G. (2021). Exploiting food embeddings for ingredient substitution. *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies*. https://doi.org/10.5220/0010202000670077