Food Recipe Alternation and Generation with Natural Language Processing Techniques

Yuran Pan
Dept. of computer and information sciences
Fordham University
New York, U.S.A.
ypan73@fordham.edu

Qiangwen Xu
Dept. of Computer and Information Sciences
Fordham University
New York, U.S.A.
qxu47@fordham.edu

Yanjun Li
Dept. of Computer and Informaiton Sciences
Fordham University
New York, U.S.A.
yli@fordham.edu

***Abstract*— We prefer to have more options when we cook. Choosing alternative ingredients or recipes, creating new recipes could be quite challenging. In this research project, we investigated how to apply the state-of-the-art natural language processing techniques such as word embedding to help people choose alternative ingredients/recipes and build language models – N-gram and neural network model to generate new recipes with authentic flavor of certain cuisine style.**

Keywords—Natural language processing (NLP), neural network, long short term memory (LSTM), N-gram, word embedding, food recipe, ingredient, recipe generation

I. Introduction

With the development of research in food and nutrition study, people have more options when they choose cooking recipes and ingredients based on their own needs and situations. One common problem in our daily life is when we cook, some required ingredients of a recipe are not available. To deal with this, we either replace the required ingredients with other ingredients, or replace the recipe with another recipe. We would like to propose a model to help people make an informed decision when choosing alternative ingredients or recipes. A novel neural network model proposed in [3] recommended ingredient substitution based on cuisine styles. Research in [6] investigated that cooking actions could be considered when choosing replaceable materials in cooking recipe. An ingredients network was built to measure the importance of ingredients and suggest modification of recipe ingredients based on users' preference [8]. Research in [1] proposed to utilize domain expert knowledge and word embedding to find ingredient substitutes.

In this research project, we adopted natural language processing (NLP) techniques to help people choose alternative ingredients/recipes. Skip-gram model with negative samples [4] was used to obtain the word embedding for ingredients and compare the similarity of different ingredients and recipes accordingly. Based on the similarity measurement, we could replace one ingredient with similar ingredients, or one recipe with similar recipes.

An interesting challenge in our cooking experience is how to create a new recipe which has the authentic flavor of certain cuisine style. Since recipes could be treated as text documents, we built two language models – traditional N-gram model and the state-of-the-art neural network language model to generate new recipes with authentic flavor of different cuisine styles.

This paper is organized as follows. Section II introduces how to apply NLP techniques to alternate ingredients/recipes, and generate new recipes. Section III describes the dataset collection and discusses the experimental results. Section IV covers the conclusion of this research and future work.

II. Study Food Recipes with NLP Techniques

A. Similarity Measurement Based on Embedding

Ingredient in recipes could be considered as word forms, and each recipe could be treated as text documents with ingredient word forms. The Skip-gram model with negative samples [4] of NLP is the model to predict the surrounding context words given a center word. This model is trained with a neural network and the embedding layer is extracted to represent the words in vector space, a.k.a. word embedding. In this research, we would like to get the embedding for all ingredients by performing the Skip-gram model on collected recipes. In this way, each ingredient is represented as a vector – ingredient embedding. Similar to the training of ingredient embedding, we also obtain vectors to represent recipes as recipe embedding [5].

Cosine similarity of two vectors is commonly used to measure the similarity of two words represented in vector space. In this research, we measured the similarity of two ingredients by calculating the cosine similarity of two ingredient embedding vectors. And the similarity of two recipes are the cosine similarity between two recipe embedding vectors.

B. Recipe Generation with Language Models

A language model is the model to compute the probability of a sequence of words/tokens or to predict the probability of an upcoming word for a certain language. In traditional NLP research, N-gram language model is implemented to generate new document after training on a given document dataset. In the state-of-the-art NLP research, neural network is broadly adopted to perform NLP tasks. Among many deep learning models, Recurrent Neural Network (RNN) with Long Short Term Memory (LSTM) layer is a popular model for language

2473-3490/20/$31.00 ©2020 IEEE
DOI 10.1109/ICDEW49219.2020.000-1

modeling [7]. This model could capture information stored in long sequences of text. Since recipe could be considered as a document with ingredients as words, we built a recipe generator with a language model. All recipes of a given cuisine were concatenated and a text generator was trained by feeding the entire text. Both N-gram language model and RNN with LSTM model were implemented to generate new recipes and their performance is reported in this research.

III. Experiment

Python packages and Google Colab were adopted to implement all the experiments for this research project.

A. Data Collection

Recipe data of different cuisine styles were collected from a website hosting thousands of recipes – Spoonacular (https://www.spoonacular.com). After removing duplicates, there are 3433 recipes across 15 cuisine styles in the dataset. The details of dataset could be found in Table I. The average number of recipes collected for each cuisine style is 229.

B. Data Preprocessing

To pre-process the recipe dataset, we performed traditional NLP techniques - tokenization, stemming (Porter Stemmer), and stop words removal. We also manually merged some ingredients of sub-types into a more general type, e.g. merging different kinds of tomato to one ingredient "tomato". To improve the performance of model training, we performed Part-Of-Speech tagging and only kept Verbs and Nouns in the cooking instruction steps. The exact amount of each ingredient in the recipes was also ignored.

After cleaning and pre-processing, we created a dataset of 3433 recipes with 9 features. Table II shows the sample of the dataset. The original cooking instructions were kept (it is not shown in Table II due to space limit) and we generated a new feature named "process" based on cooking instructions. The ingredient-process pairing was added to the dataset as an extra feature (not shown in Table II). This new feature is composed of all activities related to different ingredients. For example, the process - "lightly beaten" is paired with the ingredient "Eggs"; the process – "frozen" is paired with the ingredient "peas".

C. Ingredients Distribution Analysis

We studied the ingredient distribution of each cuisine style. American style has the highest number of ingredients – 683 and African style has the lowest number – 364. The average number of ingredients of 15 cuisine styles is 519. To observe ingredients frequency in each cuisine style, we created Word Clouds to visualize them. Common ingredients such as Salt, Olive Oil, and Garlic are used frequently across all cuisine styles, and we removed them from the lists. Please see Fig. 1 for the results. The size of the word indicates the frequency of the ingredient. The visualization shows that we could easily tell the cuisine style based on the frequently used ingredients with common cooking knowledge.

To study the shared common ingredients across 15 cuisine styles, we created a list of top 20 frequent ingredients of each cuisine style, and then compare with each other to get the shared ingredients. More shared ingredients indicate cuisine styles are more related. The maximum number of shared ingredients is 16

TABLE I. Dataset Information

Cuisine	count	Cuisine	count	Cuisine	count
African	123	German	250	Mediterranean	124
American	250	Indian	211	Mexican	250
Caribbean	250	Italian	242	Middle Eastern	250
Chinese	266	Japanese	249	Thai	250
French	227	Korean	241	Vietnamese	250

TABLE II. Sample of Processed Recipe Data

id	cuisine	title	ingredients	process	diets	nutrition
3147	Chinese	Kale Fried Rice	Cooked brown rice, garlic, ...	Cooked, finely minced, ...	Gluten free, dairy free,…	Percent Protein: 11.92,…
7985	Chinese	Chinese Beef & Broccoli	Broccoli, cornstarch, ...	Cut into florets, …	Gluten free, dairy free,…	Percent Protein: 41.62,…
0717	Italian	Easy Calzones	Bell pepper, butter,…	Sliced, whole, …		Percent Protein: 16.52,…

between the pair of Italian/Mediterranean, and the pair of Japanese/Korean share 15 ingredients. Chinese/Japanese and French/Mediterranean both share 14 ingredients. These findings are not surprising since these cuisine styles are well-known to be similar. An interesting finding is we found India and African share 14 among top 20 ingredients. The next tier is both Thai/Vietnamese and American/German share 12 ingredients. The least shared ingredient number is 1 between the pair of Vietnamese/German. This finding shows that Vietnamese is least related to German.

D. Ingredient/Recipe Similarity Measurement

Word2vec package of Gensim module was adopted to implement the Skip-gram algorithm with negative samples. In theory, word embedding vector with larger dimensions store more information, and usually it is set between 300-500. We have tested several settings and found there was no big difference for this project. Thus the dimensionality of word embedding was set as 100. The training was performed on the whole dataset of 3433 recipes. Each ingredient was represented with one embedding vector of 100-dimension. Fig. 2 shows the visualization of top 50 ingredients' embedding in a 2-D space.

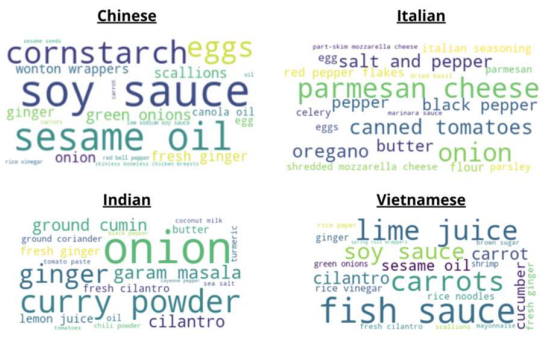

Fig. 1. Visualization of Frequent Ingredients of Different Cuisine Styles.

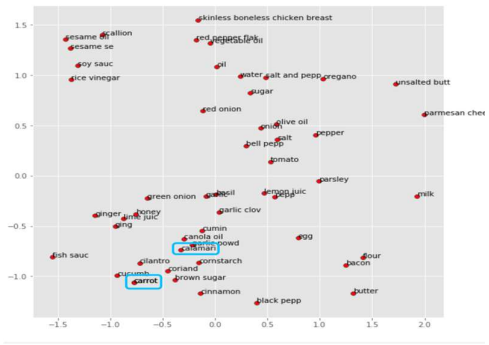

Fig. 2. Visualization of Ingredients' Embedding of 100-Dimension.

To find an alternative ingredient for a recipe when the required one is not available, our initial intuition is choosing the most similar ingredient as the substitute. The ingredient similarity was measured with cosine similarity of their ingredient embedding. However, the similarity measured based on the embedding obtained through Skip-gram algorithm reflects that two words are similar because they often show up together in the same context. In the other word, words in the same context neighborhood are considered similar. Two ingredients frequently used together in many recipes are not necessary good candidates as substitutes for each other. Thus, two similar ingredients obtained by measuring their embedding sometime could not be used to replace each other. The results in Table IV shows the findings of alternative ingredients across all cuisine styles with different window size and training epochs. An interesting finding is "Calamari" is found as the substitute of "Carrot" in all cases. It is observed in Fig 2. that ingredient embedding of "Carrot" and "Calamari" are close to each other.

A recipe is defined by its ingredients and the corresponding cooking steps. In the preprocessing step, we extracted a feature – process, which shows the cooking instructions for each ingredient. To assist the measurement of two recipes, we created a new feature – ingredient-process pairing. The intuition of measuring similarity of two recipes is to study the overlapping of pairings between two recipes. Doc2vec package of Gensim module [9] was adopted to obtain the recipe embedding. Thus, instead of ingredient embedding, each recipe was represented by its document embedding. To train the model with Skip-gram algorithm with negative samples, we set the parameter ns_exponent as -0.25 instead of the default value 0.75. The negative value means to give more weight to low frequency words.

During training, we found that doc2vec model required less training epochs, and results of 50 epochs and 200 epochs had no big difference. Even with limited amount of recipes collected, there are still some interesting findings. The performance of training with ingredient-process pairing is better than with original cooking instructions. For example, we found "Potato Beef Lasagna (Italian)" is quite similar to "Greek Patato Beef Lasagna (Mediterranean)" when the pairing of ingredients and process is included in the dataset (See Fig. 3). On the other hand, "Shrimp and Chive Spring Rolls (Vietnamese)" was found similar when the model was trained on cooking instructions only.

By comparing recipe embedding similarity, we could search for similar recipe from different cuisines or from different diet categories. The result in Table IV and Table V show the top 10 results of similar recipes for Vietnamese - Pork Meatballs among different cuisine styles and diet categories, respectively.

TABLE III. FINDING INGREDIENTS SUBSTITUTES BY MEASURING INGREDIENTS' SIMILARITY

Ingredient	Epoch=50		Epoch = 200	
	Window size=3	*Window size=5*	*Window size = 3*	*Window size = 5*
Onion	quick cooking oats	tomato juice	olive oil	garlic
Tomato	Focaccia	soy crumbles	roast pork	onion
Carrot	Calamari	calamari	calamari	calamari
Lemon Juice	marinated artichoke	whole wheat pita bread	baby green	baby green

Fig. 3. Similar Recipes for Potato Beef Lasagna (Italian Cuisine).

TABLE IV. SIMILAR RECIPE FROM DIFFERENT CUISINES FOR PORK MEATBALLS (VIETNAMES)

Cuisine Style	Most Similar Recipe	Cuisine Style	Most Similar Recipe
African	N/A	Italian	Pumpkin Risotto
American	South American Flank Steak	Korean	Korean Style Refried Beans
Caribbean	Jamaican Jerk Chicken	Mediterranean	N/A
Chinese	Schezwan Fried Rice	Middle Eastern	Jalapeno Honey Hummus with …
Indian	Oats muthia	French	Karen's Smooth French Milk …

TABLE V. SIMILAR RECIPE FROM DIFFERENT DIETS FOR PORK MEATBALLS (VIETNAMES)

Diet Category	Most Similar Recipe	Diet Category	Most Similar Recipe
Dairy Free	Caribbean Pineapple Tofu	Paleolith	French String Bean Salad
Foodmap Friendly	French String Bean Salad	Pescatarian	Paleo Salmon Tacos with …
Gluten Free	Caribbean Pineapple Tofu	Primal	French String Bean Salad
Ketogen	N/A	Vegan	Caribbean Pineapple Tofu
Lacto ovo veg.	Caribbean Pineapple Tofu	Whole 30	Vietnmese Pho Soup with Spicy…

E. Recipes Generation

1) N-gram Language Model

To build a recipe generator, we first trained the traditional N-gram language model (N=3) on the dataset. Cooking instructions of all recipes from the same cuisine style were concatenated to create one document as the training dataset. Since the amount of each ingredient was ignored during the embedding training, the new recipes do not have specific amount of each ingredient. Here are two recipe samples generated with max. 50 tokens by feeding "heat the water".

- **Chinese cuisine, vocab size: 2605**
  "heat the water and moisten all four edges of the wonton wrapper. fold the wrapper. fold the wrapper. fold the wrapper. ……"
- **American cuisine, vocab size: 2497**
  "heat the water and stir until the mixture is smooth. add the flour, baking powder, baking powder, baking powder, baking powder, ……"

Since N-gram model performs a greedy search of next word with highest probability, and it has no back-off mechanism, the generated text begins to repeat part of the phrase and is stuck in a loop.

2) LSTM Language Model

The RNN with LSTM model was implemented to perform a mini-batch training on sequences of 10 tokens. The training dataset kept the original format of recipes without doing traditional NLP pre-processing steps except tokenization. The original recipes were not trimmed and digits were not removed. Recipes were grouped according to its cuisine style, concatenated as a long sequence of text as the training dataset. The sample results of generated recipes with given input string "cooking until rice is heated through and peas are just cooked serve with" are shown as follows.

- **Chinese cuisine, 50 epoch**
  "cooking until rice is heated through and peas are just cooked serve with pepper , hot add add sauce cooked). bowl are for with lime add white and and cook 6 cook , of , , parts to oil 30 seconds medium - and cook broccoli top chicken over are , translucent seconds medium beef and soy shallots 30 and and add to cook tender that high serve , brown 1 chicken with pour sauce sauté flakes 30 ribbons , and over oil teaspoon heat oil and , greens stir sesame up add of turn scallion resealable beef , add broccoli cook noodles heat chicken chiffonade steam dressing scallions and , cooked"
- **Chinese cuisine, 100 epoch**
  "cooking until rice is heated through and peas are just cooked serve with for to , 2 strips 1 pepper pan : stirring cooked in size pat , cooked cilantro cook for stirring in add , dry large cooked and of nonstick large steam over high heat how sauce rice 1 oil and ingredients salt of steam - over - serve another , together : until ingredients salt of fry bright oils heat add if in add salt to 1 over or until toss of beef turn are rice sauté temperature for wok steak marinade , seconds cook for seconds more toss let rest then just room coat , immediately container , whisk"
- **American cuisine, 50 epoch**
  "cooking until rice is heated through and peas are just cooked serve with heat boiling dressing 1 remaining processor low season , cut , about top with 5 your heat to processor tomatoes , 3 mixing to to to through water tablespoons ingredients are 1 heat mostly golden mixture until , , to to package dressing smooth with high ready to dressing with heated to , , minutes until line of , , salad on on on to remaining garlic over bowl are large and , , - heat high each and to over with - high through hours of with and blender , high garlic as crushed baking head in , romaine"

3) Discussion

On average, recipes generated by 3-gram model include five recognizable cooking steps and three ingredients, and recipes generated by LSTM model include seven recognizable cooking steps and eight ingredients. Based on the evaluation of six volunteers who have moderate cooking experience, the generated cooking instructions with LSTM did extract important phrases and ingredients of each cuisine style. In the future, we plan to collect more recipes and increase the training with more epochs since the RNN model has good performance with large dataset.

IV. Conclusion and Future Work

In this research project, we showed some interesting findings by performing NLP methods on food recipes data and providing a special view to this study field. Since the amount of recipes collected for each cuisine is limited, training ingredient embedding for each cuisine separately is not practical. In the future, we plan to collect more recipes and continue to study ingredient alteration option with different similarity measurement. And collaboration with native chefs of different cuisine types will help us to better interpret the model generated recipes.

References

[1] S. Dekkers, "Automatic ingredient replacement in digital recipes: combining machine learning with expert knowledge," Master Thesis, University of Amsterdam, 2017.

[2] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] M. Kazama, M. Sugimoto, C. Hosokawa, K. Matsushima, L. R. Varshney, and Y. Ishikawa, "A neural network system for transformation of regional cuisine style," *Frontiers in ITC*, vol. 5. 2018, pp. 14.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," International Conference on Learning Representations: Workshops Track, 2013.

[5] L. Quoc, T. Mikolov, "Distributed representations of sentences and documents," Proceedings of the 31th International Conference on Machine Learning, 2014, Beijing, China, pp. 1188–1196.

[6] Y. Shidochi, T. Tomokazu, I. Ichiro, and M. Hiroshi, "Finding replacabel materials in cooking recipe texts considering characteristic cooking actions," Proceedings of the ACM Multimedia 2009 Workshop on Multimedia for Cooking and Eating Activities, Beijing, China, pp.9-14.

[7] M. Sundermeyer, R. Schluter, and N. Hermann, "LSTM neural networks for language modeling," Thirteenth annual conference of the international speech communication association, 2012.

[8] C. Teng, Y. Lin, and L.A. Adamic, "Recipe recommendation using ingredient networks," Proceedings of the 4th Annual ACM Web Science Conference, 2012, pp.298-307.