The Semantics and Collocations Relation in Food Reviews

Ledong Shi , Syed Ahmad Chan Bukhari and Fazel Keshtkar
Division of Computer Science, Math & Science
St. John's University
{ledong.shi17, bukharis, keshtkaf}@stjohns.edu

Abstract

Finding our favorite dishes have became a hard task since restaurants are providing more choices and varieties. On the other hand, comments and reviews of restaurants are a good place to look for the answer. The purpose of this study is to use computational linguistics and natural language processing to categorise and find semantic relation in various dishes based on reviewers' comments and menus description. Our goal is to implement a state-of-the-art computational linguistics methods such as, word embedding model, word2vec, topic modeling, PCA, classification algorithm. For visualizations, t-Distributed Stochastic Neighbor Embedding (t-SNE) was used to explore the relation within dishes and their reviews. We also aim to extract the common patterns between different dishes among restaurants and reviews comment, and in reverse, explore the dishes with a semantics relations. A dataset of articles related to restaurant and located dishes within articles used to find comment patterns. Then we applied t-SNE visualizations to identify the root of each feature of the dishes. As a result, to find a dish our model is able to assist users by several words of description and their interest. Our dataset contains 1,000 articles from food reviews agency on a variety of dishes from different cultures: American, i.e. 'steak', hamburger; Chinese, i.e. 'stir fry', 'dumplings'; Japanese, i.e., 'sushi'.

Introduction

With the improvement of people's life quality and consumption capacity, the restaurant industry is growing rapidly. Therefore, there is a growing diversity of restaurants with numerous dishes to choose from. And on the other hand, customers have even less time to choose and try different restaurant and food. In other words, it is harder and harder for customers to choose right restaurant dine in and favorite foods to enjoy.

The good point now days is that, as the growing of the restaurant industry, the development of restaurant advertisements, the same time, "food reviews" are being more regular and popular. Our dataset contains "Food reviews" articles, as it writes, they are basically reviews of food, mainly focusing on one restaurant per article. Each article describes about the dishes they offered, the culture background, names of the dish, what the dish made of, tastes of dish and so on. We applied and analyzed our model on this dataset using NLP algorithm and we found some interesting facts that we explore in this paper. The rest of paper organized as follow: In Section Related Work we explore some relevant research, Section Methodology describe our model and methods. Section Methodology explain our model. In Section Experiments and Results we demonstrate the results of the model.

Related Work

In this section we review the related work. Based on our review of previous research, studies in this area belong to food name recognition and food description understanding. Wiegand, Roth, and Klakow (2014) studied the the features of food terms and assign semantic information to food items using weakly-supervised induction. In another study done by Popovski, Seljak, and Eftimov (2019) and Gorjan Popovski1 and Eftimov (2019), they focused on annotate food terms, extract food information with a rule based named-entity recognition. In recent study by Popovski, Seljak, and Eftimov (2020), the researchers reviews the several study on food entity recognition and compared with another.

Wiegand, Roth, and Klakow (2012) proposed another research related to food knowledge acquisition with natural language processing. They discussed three possible method of extracting knowledge: Statistical Co-occurrence, Pattern-based Approaches, and the method with further linguistic analysis. Another study proposed by Dong, Zhong, and Huang (2018) presented a practice of food knowledge extraction: study on Chinese people's perception of spicy and numbing food with a corpus-based method.

Based on above studies, the difference between our research and related works is that in our model we mainly focus on the semantics relations, for example: similarity, between different food entities, and collocation semantics relations.

Methodology

In this section we discuss the details of our methodology, dataset, pre-processing, feature extraction, and findings that came out of this research.

Dataset

We collect our dataset using BeautifulSoup and requests package in python from eater.com, a website focus on food, dining and severs as a restaurant guide for the customers that launched in 2009. They cover nearly in 20 cities by 2012. The whole dataset comprised of 400 foods review articles. Each of the article introduce 1-2 restaurants and 6-10 dishes severed by the restaurant. The content is mainly about the food. In other word, the dishes, including their look, taste, price, also include some culture background. Table 1 illustrates the an example paragraph of the food review of some articles.

Table 1 shows an paragraph example of the food review in each article.

Table 1: Some sample paragraph of food review in each article from dataset.

Sample 1: 'The chef grills cumin lamb skewers in the style of a studied Xinjiang hangout, offsetting the juicy meat with marshmallow-like fat. He sends out chilled bang bang chicken that balances the poultry's clean punch with the searing heat of chiles.'
Sample 2: 'Chilaquiles: Traditionally, the dish consists of tortilla chips simmered in salsa. That much is true here, except Au Cheval stacks everything high and adds sour cream, guacamole, egg, and jalapenos. The chips and toppings are arguably more in conversation with American nachos than the Mexican staple, but it doesn't suffer from the fusion-y sensibilities. The salsa is applied generously enough to offset all the richness.'
Sample 3: 'Hash browns with duck hearts: This isn't a typical shaved potato hash; Au Cheval instead uses diced and roasted potatoes. The crispiness factor is more subdued. In any case, cooks slather the nuggets in mornay and duck gravy. This would be a reasonably nourishing dish at a roadside diner in, let's say, Montana, after a long day of skiing and hunting elk. But one might argue such organ-y and carb-y indulgences don't quite jibe with a hot and humid New York summer.'

The articles were written all in English and almost identical in format since we collect all of them from eater.com. All the articles consists of $606,244$ unique words with the lexical diversity of 20.21% before pre-processing. After pre-processing and removing stop words and other unnecessary tokens, it contains 286,654 unique words with the lexical diversity of 15%. Each article consists of almost 4800 words. For example, top 10 frequently used word in this dataset is 'restaurant', 'like', 'dish', 'good', 'menu', 'food', 'chef', 'come', 'flavor', 'chicken'. We used topic modeling, and categorized the restaurant based on the food the serve. Table 2 show the 17 categories of food or restaurant and the identical words in them find from our article.

Table 2: 17 categories of food/restaurant

food categories	**identical words**
Steak house	restaurant, beef, bread, meat, steak, steakhouse, meal, new york
Fast food	chicken, sample dish, drink, salad, sandwich, cheese, beer, pizza, fry, noodle, shrimp
Pizza restaurant	pizza, pie, wine, salad, cheese
Mexican	taco, price, bar, nachos
Vietnamese	chicken, rice, Vietnamese, crab, tasting, beef, noodle
IPA Beer bar	ipa, beer, brewery, dumpling, ale, soup
Sri Lankan	curry, Sri Lankan, rasa (Indonesian), chicken, fish, roti(Flatbread), lamprais(Sri Lankan dish), starch, juice, tea, rice, kottu, lamb
Irani	irani, chai, indian, curry, prune, bitter, butter, pav, roll
Tibetan	pork, beef, chicken, rice, soup, Tibetan, tea
Japanese sushi	sushi, price, fish
Beer bar	beer, lager, yeast, brett, ale, flavor, brewery, schlafly, ferment, allagash, brew, unfiltered, ipa

Pre-Processing

Pre-processing is the technique of cleaning and normalization of data which may consist in removing less important tokens (called stop words), words, or characters in a text such as 'a', 'and', '@','the', and other unnecessary stop words and lowering capitalized words like 'APPLE'. We also remove all the "'s" (and also "'s") in the article end with it for example: "it's" and "lifestyle's".

The texts contained several unimportant tokens, for instance, URLs, numbers, HTML tags, and special characters which caused noise in the text for analysis. We cleaned the data first using NLTK (Natural language and Text Processing Toolkit) Bird and Loper (2004) Porter stemmer and stopwords package. Here is an example of transformation of text before and after pre-processing:
before cleaning: 'You slurp the meat almost as easily as you would a noodle. Want to combine hot pot with a few hours of sake-fueled karaoke? ';
after cleaning: 'slurp meat easily noodle want combine hot pot hour sake fuel karaoke'.

Features Selection

Feature extraction is an accurate and concise reduction process of raw data to some grouped data (Features). In this section we describe the features that extracted from the dataset for further processing. Features included are W2V(Word2vec), bigram, PoS (Part-of-Speech), Similarity function. We applied Word-Embedding representation along with PCA scores in order to explore the characteristics of foods and dishes.

Bigram Features Bigram is a pair of words adjacent to each other which form a phrase. By applying bigrams, bigram helps to find the common phrase in the dataset articles. For example; name of dishes. There are many dish names with combinations of two or even more words. After using bigram, the dish name with two words will be put into the dictionary. For example, "hot pot" will be saved as "hotpot" in the dictionary.

Part-of-Speech Features Part-of-Speech (PoS) are classes or lexical representations which have similar grammatical properties. For the purposes of this research, we used Spacy[1] part of speech tagging package to sort out the most used descriptions for each dish in each article.

t-SNE Visualization

t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised, non-linear technique primarily used for data exploration and visualizing of high-dimensional data. In simpler terms, t-SNE provides an intuition of how the data is arranged in a high-dimensional space. It was developed by Laurens van der Maatens and Geoffrey Hinton in 2008 van der Maaten and Hinton (2008).

Model

In this section, we introduce some model and algorithms that have been used in this research, such as PCA, t-SNE, and Word2Vec.

t-SNE and PCA Model Principal Components Analysis (PCA) and t-SNE have some differences. The first thing to note is that PCA was developed in 1933 while t-SNE was developed in 2008. A lot has changed in the world of data science since 1933 mainly in the realm of compute and size of data. Second, PCA is a linear dimension reduction technique that seeks to maximize variance and preserves large pairwise distances. In other words, things that are different end up far apart. This can lead to poor visualization especially when dealing with non-linear manifold structures. Think of a manifold structure as any geometric shape like: cylinder, ball, curve, etc.

Word2vec Model

We propose a Word2vec technique to learn how a food term associates with its descriptions from a large corpus of text. Word2Vec utilizes two architectures: CBOW (Continuous Bag of Words) and Skip Gram, Mikolov et al. (2013). We applied the similarity function for several food terms to find out the which food is most similar to the other and then build t-SNE plot maps to visualize the similarities and differences for each food.

Collocations of Food Relation and Categories

In our dataset, each article has a theme, and so do each restaurant. The articles always has a main focus on one "topic" and they only introduce foods about or around that topic and the words they use are pointed to that topic. Due to this, we apply an LDA model (Blei, Ng, and Jordan (2003)) with 20 topics, and 10 passes, to our dataset which successfully generated 17 categories of different foods. Table 3 shows an example of one categories which is about pizza.

Table 3: An example similarity for Pizza based on semantic collocations.

Original word	**Similar words for pizza**
pizza	'0.007*"pizza" + 0.006*"like" + 0.005*"pie" + 0.005*"restaurant" + ' '0.005*"wine" + 0.005*"menu" + 0.004*"open" + 0.004*"good" + 0.004*"chef" + ' '0.004*"expect" + 0.003*"come" + 0.003*"salad" + 0.003*"cheese" + ' '0.003*"red" + 0.003*"bar" + 0.003*"player" + 0.003*"sauce" + 0.003*"dish" + ' '0.003*"white" + 0.003*"emmy"'

As it can be seen in Table 3, there are 'pizza', 'pie', 'wine', 'salad' which means in this category certain kinds of food always shows up.

Experiments and Results

In this section, we prsent some results that produced by our models in previous section.

Similarity of Terms

We first created three word2vec model with the same size of 300 and different window of 2, 5 and 10 for our dataset. The size in word2vec means the amount of dimensions model use to describe the words and the window means the mount of word the model will use for each calculation. So the window of 2 consider the only one word around the main word, the window of 5 considers 4 words around the main word and the window of 10 considers 9 words around the main word. We then applied the most similar function to find the top 20 most similar words of the main word, for example we use "beef", to test and describes better.

Staring with the window of 2, we build the word2vec model and applied the most similar function as shown in Table 4.

We then used PCA and t-SNE to visualize the main word entries (here we use example of "beef") with its similar words in Figure 1.

In Figure 1, "beef" in red is the main word, the word we use to search for its similar words. X-axis and Y-axis are the dimension of similarity which was lower by PCA and t-SNE from 300 dimension calculated by word2vect model. In this figure, the closer the two words get to each other, the more similar they are.

In this particular example, we can already see that there are several similar words with our target word "beef": "pork", "lamb", "fish", "chicken" and "duck". "pork" and "lamb" stands much closer to "beef" than "duck" and "chicken" in Figure 1. Therefore, this shows that in the article from dataset, the surrounding words of "pork" and

[1] www.spacy.io

Table 4: Top 20 most similar words for beef

original word	similar words
beef	lamb, pork, chicken, roast, rib, grill, duck, kebab, short rib, heart, stew, fry, sauteed, fish, bao, chicken wing, meatball, tasty, meat, pork chop
sushi	omakase, ya, nigiri, sashimi, kaiseki, toro, bluefin, eel, sake, wasabi, nori, shiso, ramirez, urasawa, uni, maison, sea, tuna, masa
curry	roti, masala, coconut milk, dosa, turmeric, goat, tamarind, biryani, mellow, pao, paneer, casserole, mango, lentil, naan, tandoori, stew, larb, papaya, rice

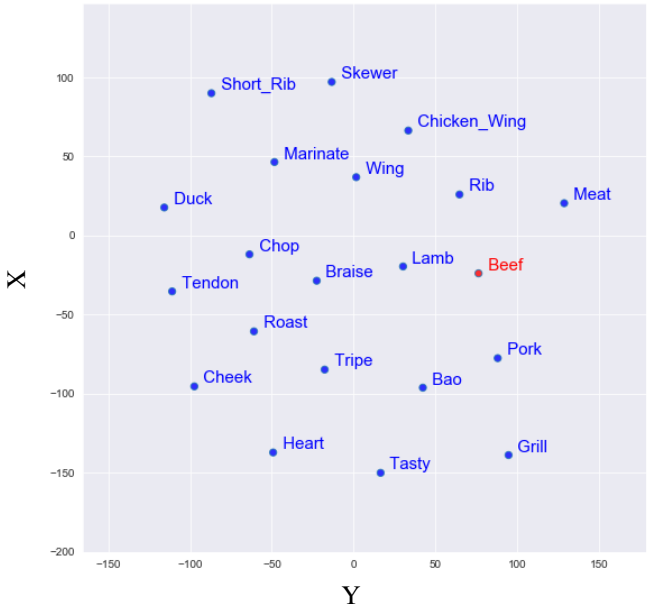

Figure 1: visualization for 20 most similar words for "Beef" by window of 5

"lamb" is more similar to the surrounding of "beef" than the surrounding word of "chicken". This suit the fact that, compare to "chicken" and "duck", "pork" and "lamb" is more similar to "beef".

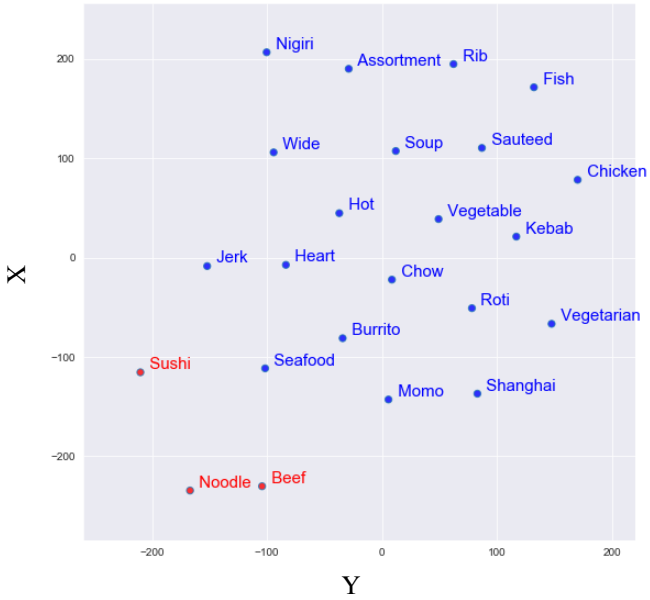

Figure 2: Similarity Visualization For "Beef", "Noodle" & "Sushi"

We have been able to find similar food words by a single food word. In Figure 2 shows multiple food terms at the same time while matching the results trying to find the common characteristics of multiple food words that users like.

After this, Figure 3 shows the foods that users don't like. According to below function, we try to see if it can be use negative parameters to filter out the features that users don't like.

Similarity of Topics and Types of Restaurant

We used LDA topic modeling to separate each different kinds of restaurant. Figure 4 is an overall visualization of 20 different topics, or types of restaurant. The figure represent the inter topic distance, when two topic is close to each other, that means the word they use are similar. Below is the relevance formula that we use to determine the relevance of each word in each topic:

$$r(w|t) = \lambda * p(w|t) + (1 - \lambda) * p(w|t)/p(w)$$

r(w|t) means the relevance of term w in topic t. λ determines the weight given to the probability of term w under topic k relative to its lift (Carson Sievert (2014)) measuring both on the log scale). Setting λ = 1 results in the familiar ranking of terms in decreasing order of their topic-specific probability, and setting λ = 0 ranks terms solely by their lift. Here we set λ = 1 to get the results in the familiar ranking of terms in decreasing order. Figure 4 represents the top 30 most relevant terms for this topic.

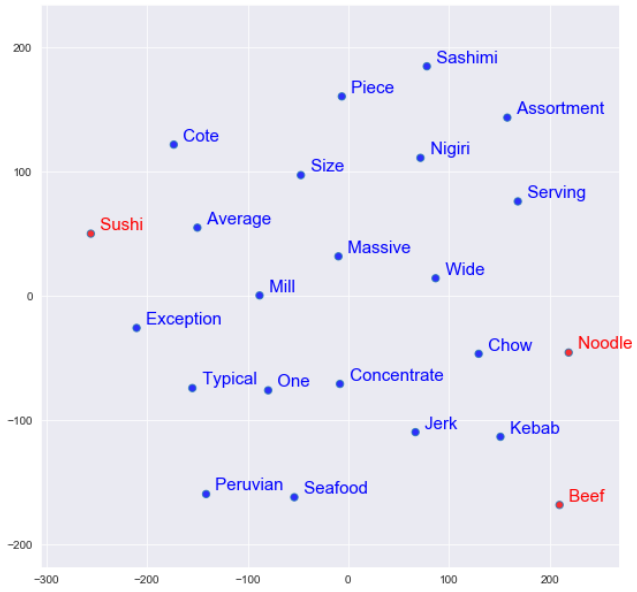

Figure 3: visualization for 20 most similar words for "Beef", "Noodle" & "Sushi" with an Negative "Chicken"

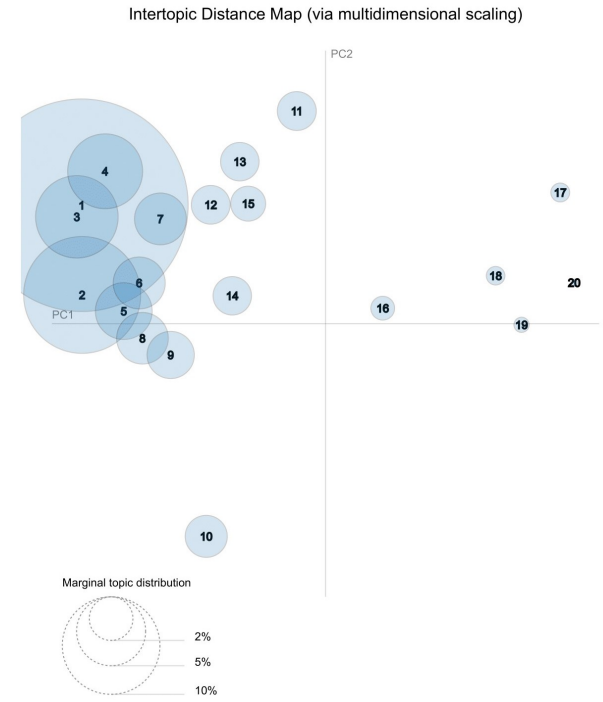

Figure 4: visualization 20 different kinds of restaurant article

By clicking on certain topics on the left, we can select certain topics, and the graph on the right will show the top-30 most relevant terms for this topic and also compare it with the overall frequency of this term in the whole data set. The red bar shows the frequency of this term in the selected topic and the blue bar shows the overall frequency of this term.

Then we select three relate topic, topic 5, 8 and 9, to show the details of this visualization. The reason of choosing these three topics is that they are close with each other on the overall visualization of 20 topics. It means they share some food terms but also have some different between each other.

Figure 5 visualize the top-30 most relevant terms for topic 8. It can be seen that in topic 8, the term "pizza" is also the most relevant food term just as topic 5. But the second most relevant food terms here is "pie". By looking at other relevant words, "wine", "bar", "salad", "sandwich", we can visualize this type of restaurant as an bar with pizza and pie.

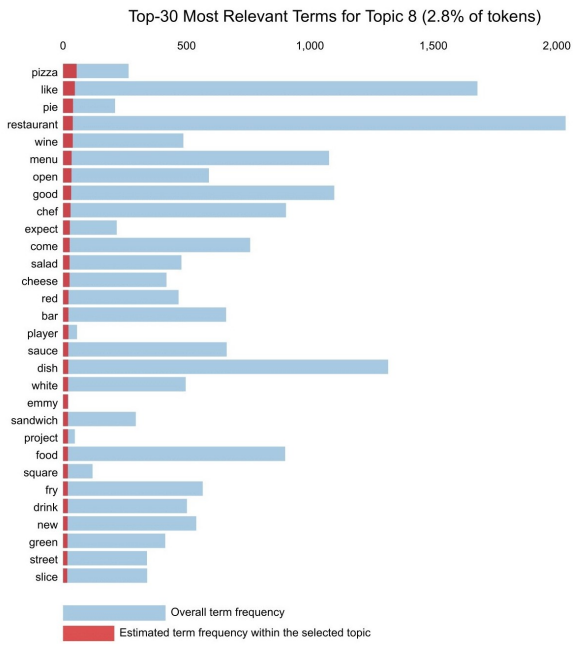

Figure 5: visualization of relevant terms in topic 8

Figure 6 visualize the top-30 most relevant terms for topic 9. We can see that in topic 9, the term "pizza" is also one of the most relevant food term just as past two topics but in a lower rank. The most relevant food terms here is "noodles", "pork", "beef" and "chicken". By looking at other relevant words, "rice", "Tibetan", "Indian", and "Chinese", we can see this type of restaurant as an Asian restaurant mainly focus on noodles and meat, but also served with pizza. With these visualization we can wisely decide which food is in which kind of restaurant.

Conclusion and Future Work

In this research we applied various semantic analysis model to investigate relation between foods and resultant with different categories. As shown, in our former graphs, there

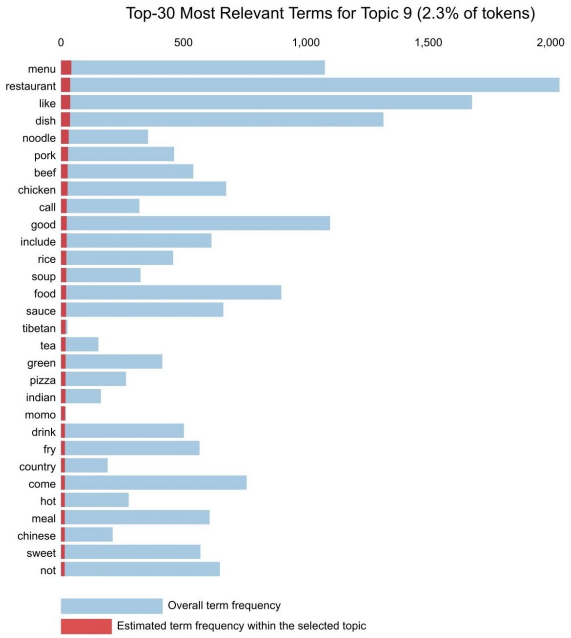

Figure 6: visualization of relevant terms in topic 9

are many common words, like "restaurant", "dish", "good", "like". It is hard to decide whether we should remove them or not. On the one hand, almost all food reviews has these words, but on the other hand, some of them, like "good", 'like", etc, represent people's favour of particular restaurant.

One of the option to investigate in future work is to improve these types conflicts. Also, some dish name won't contain any food in them, for example "hot-pot", the program is hard to identify these kind of dish names. This is another research to study. We also aim to investigate more dataset and reviews for larger foods and restaurants.

References

Bird, S., and Loper, E. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, 31. Association for Computational Linguistics.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Carson Sievert, K. S. 2014. Ldavis: A method for visualizing and interpreting topics.

Dong, S.; Zhong, Y.; and Huang, C.-R. 2018. How do non-tastes taste? a corpus-based study on chinese people's perception of spicy and numbing food.

Gorjan Popovski1, Stefan Kochev1, B. K. S., and Eftimov, T. 2019. Foodie: A rule-based named-entity recognition method for food information extraction.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In Burges, C. J. C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Popovski, G.; Seljak, B. K.; and Eftimov, T. 2019. Foodbase corpus: a new resource of annotated food entities. Brussels, Belgium: Oxford University Press.

Popovski, G.; Seljak, B. K.; and Eftimov, T. 2020. A survey of named-entity recognition methods for food information extraction.

van der Maaten, L., and Hinton, G. 2008. Visualizing high-dimensional data using t-sne. In *Journal of Machine Learning Research 9(Nov):2579–2605*.

Wiegand, M.; Roth, B.; and Klakow, D. 2012. Knowledge acquisition with natural language processing in the food domain: Potential and challenges.

Wiegand, M.; Roth, B.; and Klakow, D. 2014. Automatic food categorization from large unlabeled corpora and its impact on relation extraction.