Classifying Fruits and Vegetables with Word Embeddings

Kanitha Mann
W266, Section 2, Spring 2019
UC Berkeley School of Information
kanitha@berkeley.edu

Abstract

The classification of fruits and vegetables is not universal for culinary usage. This report explores an attempt to study word embeddings as a way to discern these classifications, both by how they differ across cultures such as the United States and Brazil, and how different corpora influence the discussion leading to the classifications. Using cosine similarity on eight specific food items compared to fruit and vegetable, it is found that word embeddings trained on both general corpora and recipe corpora generally align with the definitions that health organizations provide. General corpora word embeddings proved better at providing context for the classification, but the recipe corpora could have benefitted from a larger dataset for the word embeddings to become more effective. This work has shown that there is a literature around what constitutes a fruit or a vegetable that can be explored through NLP.

1 Introduction

The definition of what items are classified as fruits or vegetables are standardized for botany, but there is no such universal classification system agreed upon by health professionals and consumers. (Thompson et al., 2011) Instead, cultural knowledge and traditions heavily influence the classification system. Thus, what someone considers a fruit or a vegetable depends on their cultural knowledge, which may be different from someone else's views. Health professionals recognize the distinction, so national health guides are typically country specific and reflect local practices.

How different countries classify fruits and vegetables can also be inferred from natural language. How these items are communicated within communities can be measured through word embeddings, which have been shown to capture a range of semantic relationships within the body of text it is trained on. (Mikolov et al., 2013a)

This project explores how effective measuring semantic similarity of a food item to a fruit or a vegetable within word embeddings can be when the task is to understand how different countries define this classification. Two different countries, the United States and Brazil, will be explored for this analysis. Both a pre-trained English and Portuguese word embedding is evaluated to understand how the classification system is understood from a substantial amount of documents. Additionally, word embeddings trained on a set of English and Portuguese recipe corpora is evaluated to understand if looking at domain-specific corpora helps grant better understanding over general corpora.

2 Related Work

2.1 Classifying Fruits and Vegetables

Health organizations have been documenting how different items fit into various fruits and vegetable classification systems, and if there is enough similarities and positive benefits in creating a global standard. Thompson et al. conducted a survey study measuring how respondents classified a set of foods as a fruit, vegetable, or something else. These respondents' primary languages were collected in order to provide a comparison between responses, with English and Spanish speakers' views being the main difference discussed. Most respondents agreed that three items - corn, green pepper, and potatoes - were vegetables, but disagreed on most other items posed in the survey.

In an effort to promote fruit and vegetable consumption globally, the World Health Organization received survey results from around 80 countries about their thoughts on the subject, including how a fruit and a vegetable is defined. (Keller)

Fruits have similar definitions between the surveyed countries, but vegetables differ in regards to starchy tubers, dry pulses, and corn. The WHO used these preliminary findings to discuss whether a global definition for fruits and vegetables could be proposed. The meeting report subsequently stated that the definition varied from country to country, and more discussions were needed to reach a point where cross-country comparisons could eventually be made. (Organization)

2.2 Domain Specific Word Embeddings

A handful of studies have explored creating word embeddings on domain specific corpora. Nooralahzadeh et al. (2018) evaluated both CBOW and skip-gram architectures of *word2vec* on Oil & Gas domain corpora, then evaluated the efficacy of the resulting word embedding versus general word embeddings on intrinsic and extrinsic tasks. Mengnan et al. (2018) evaluated the skip-gram architecture of *word2vec* on drug-named entity corpora and showed the results performed better on tasks related to biomedical NLP.

2.3 Food Domain

Wiegand et al. (2012) proposed that preparing meals and health-related issues are the best problems to address with NLP techniques. Yet, there has been recent research in augmenting food image recognition tasks with NLP. Marín et al. (2018) created a joint embedding model of recipes and images for an image-recipe retrieval task, in which both ingredient and instruction text were used. Min et al. (2017) developed a similar image-recipe retrievel task, but used the recipe's cuisine type, course, and flavor information as attributed in their multi-modal embedding.

3 Methods

The *gensim* implementation of *word2vec* is used to process and measure the inference related to fruit and vegetable classifications. (Řehůřek and Sojka, 2010) Two countries' views on this classification system, the United States and Brazil, were chosen due to availability of data resources, and that there are already established differences in classifying items as a fruit or vegetable between the two countries in their primary languages. The analysis is benchmarked against these established differences as a way to measure how well the implementation performs.

Two pre-trained, widely available word embeddings were first evaluated to measure how the fruit and vegetable classification may be inferred from large, general corpora. To represent the U.S. and its primary language of English, the word embedding developed by Mikolov et al. (2013b) trained on Google News text was used. To represent Brazil and its primary language of Portuguese, the word embedding developed by Hartmann et al. (2017) is used. This particular word embedding was trained on over one billion tokens worth of both Brazilian and European Portuguese text, where the latter dominates the type of language encompassed by the word embedding.

Word embeddings trained on domain specific corpora were then evaluated to measure if the resulting inference proves better. English recipes were collected from AllRecipes (Vance, 2017), and Brazilian Portuguese recipes were collected from Tudo Gostoso (Ferreira, 2016). After isolating the datasets to each recipe's title, description where applicable, ingredient list, and instruction list, the AllRecipes data included 225,602 documents consisting of 27,024,084 tokens and 35,601 types, whereas the Tudo Gostoso data included 7,483 documents consisting of 246,198 tokens and 7,972 types. The skip-gram model of *word2vec* in *gensim* was used to train the recipe datasets separately in order to capture word embeddings.

4 Intrinsic Evaluation

Evaluation of how well the domain specific and general word embedding capture the classification between the United States and Brazil is done by measuring the semantic similarity between various food items to the categories *fruit* and *vegetable*. Cosine similarity between the food item and the potential categories will be the main metric used for this analysis and is provided by the *gensim* library.

Because there is no universal classification, there is no list of discernable fruits and vegetables that can be used to evaluate the efficacy of the word embeddings. Instead, the semantic similarity of eight food items will be used in this analysis. The decision to use these items is based on the literature set forth by Thompson et al., Keller, and the food-based dietary guidelines published by Brazil's Ministry of Health (2015) and the U.S. Department of Health and Human Services (2015):

- **açaí**: Listed as a fruit in the Brazil guide but not mentioned in the US guide
- **apple**: Listed as a fruit in both guides
- **avocado**: Listed as a fruit in the Brazil guide and a vegetable in the US guide
- **broccoli**: Listed as a vegetable in both guides
- **corn**: English and Spanish speakers mostly agree to it being a vegetable in Thompson et al., but Keller asserts that there is disagreement between countries with this classification
- **potato**: Listed as a tuber in the Brazil guide and a vegetable in the US guide
- **tomato**: Mainly defined as a vegetable by Spanish speakers with more divisive results for English speakers as asserted by Thomas et al.
- **watermelon**: Not mentioned in the Brazil guide and a fruit in the US guide

4.1 General Corpora Word Embedding

Table 1 displays the cosine similarity metrics of the food items as measured by the general corpora word embeddings. These results suggest that:

- açaí, apple, and watermelon is a fruit for both countries
- broccoli is a vegetable for both countries
- avocado, potato, corn, and tomato is a vegetable in the US and a fruit in Brazil

Of note is that avocado, abacate, brócolis, milho, tomato, tomate, and watermelon had less than 0.04 measurement difference when classifying between a fruit or a vegetable. While most words are in agreement with the literature specified earlier, the Portuguese word embedding classifying batata and milho as fruits prompts more exploring.

The most similar words to batata, in descending order, are mandioca (cassava), tomate (tomato), batata-doce (sweet potato), cenoura (carrot), aveia (oats), cebola (onion), alface (lettuce), feijão (bean), arroz (rice), and couve (cabbage). Most of these items are vegetables themselves, which would imply that batata would also be one. However, the Brazil guide explicitly listed batata as a tuber, and not a vegetable. Unfortunately, the cosine similarity between *batata* and *tubérculos* (the Portuguese word for tubers) is 0.28, significantly less than the measurements presented in Table 1.

The most similar words to milho, in descending order, are arroz (rice), soja (soy), mandioca (cassava), trigo (wheat), feijão (bean), sorgo (sorghum), centeio (rye), cevada (barley), tomate (tomato) and aveia (oats). Most of these items are grains. Thus, measuring the cosine similarity between *milho* and *grão* (the Portugese word for grain) yields 0.67, higher than the measurements presented in Table 1.

For the food items explored, the general word embeddings mostly aligned with the classifications given in the food-based dietary guidelines, indicating that general literature tends to assert similar classifications for food items as does health related governing bodies.

4.2 Domain Corpora Word Embedding

Table 2 displays the cosine similarity metrics of the food items as measured by the general corpora word embeddings. These results suggest that:

- apple, avocado, and watermelon is a fruit for both countries
- broccoli, potato, tomato, and corn is a vegetable for both countries
- açaí is a fruit in the US and a vegetable in Brazil

Of note is that the cosine similarity values are less than from the general word embedding values in Table 1, especially when considering the values generated from the English word embedding trained on AllRecipes recipes. The exception to this trend are some of the cosine similarities for the Portuguese word embedding trained on Tudo Gustoso recipes for melancia (watermelon), whose values are greater than from the general Portuguese word embedding. Tomato, potato, and batata were almost indistinguishable in classifications. These results contained two words that did not align with literature: açaí for Portuguese and avocado for English, which will be explored further.

The most similar words to açaí, in descending order, are brisas (breezes), maio (May), tropicália (a type of art movement in Brazil), charr (char), balanceado (balanced), iaiá (an honorific), basico

	United States		Brazil	
Food	**Fruit**	**Vegetable**	**Fruta**	**Verduras**
EN (PT)	*sim(food,fruit)*	*sim(food,vegetable)*	*sim(food,fruta)*	*sim(food,verduras)*
açaí (açaí)	0.42	0.35	0.63	0.58
apple (maçã)	0.64	0.45	0.73	0.51
avocado (abacate)	0.51	0.55	0.65	0.62
broccoli (brócolis)	0.50	0.61	0.60	0.62
corn (milho)	0.32	0.47	0.64	0.61
potato (batata)	0.51	0.63	0.69	0.60
tomato (tomate)	0.60	0.64	0.66	0.64
watermelon (melancia)	0.53	0.52	0.50	0.42

Table 1: Cosine similarity for general English and Portuguese word embeddings

	United States		Brazil	
Food	**Fruit**	**Vegetable**	**Fruta**	**Verduras**
EN (PT)	*sim(food,fruit)*	*sim(food,vegetable)*	*sim(food,fruta)*	*sim(food,verduras)*
açaí (açaí)	0.37*	0.02*	0.38	0.48
apple (maçã)	0.43	0.12	0.55	0.47
avocado (abacate)	0.17	0.07	0.50	0.27
broccoli (brócolis)	0.03	0.07	0.08	0.48
corn (milho)	0.04	0.19	0.08	0.17
potato (batata)	0.09	0.11	0.12	0.12
tomato (tomate)	0.09	0.09	0.01	0.13
watermelon (melancia)	0.35	0.06	0.84	0.58

Table 2: Cosine similarity for domain specific English and Portuguese word embeddings. *: *açaí* was not a recognized word in the AllRecipes dataset due to the accented characters, so *acai* was used instead.

(basic), esfihas (a type of pizza dough), tsurlhardo, and limonada (lemonade). Unlike in the general word embeddings, the most similar words are not fellow fruits or vegetables and do not provide much additional help. *Charr Tsurlhardo* is a name of a recipe in the Tudo Gostoso dataset, but unfortunately did not have any ingredients or instructions associated with it.

The most similar words to avocado, in descending order, are avocados, hass, cucumber, mango, papaya, haas, guacamole, mangos, chilean, and mangoes. Many of the similar words are fruits themselves, which may have lent weight to the cosine similarity between avocado and fruit from the AllRecipe dataset.

Overall, the domain specific word embeddings held promise in aligning with the classifications given in the food-based dietary guidelines, but did not hold substance when exploring most similar words. While the Tudo Gostoso dataset could have done with more documents, the much larger AllRecipe dataset had lower cosine similarity values.

5 Conclusion

Presented in this report is an attempt to take a classification that is not universally agreed upon - what item is a fruit and what item is a vegetable - and attempt to discern those classifications through word embeddings that have been trained on either general corpora or recipe corpora. Additionally, because cultural norms heavily influence perception on this classification, English and Portuguese word embeddings were explored to understand classifications in the United States and Brazil, as referenced by each country's food-based dietary guidelines, as well as additional studies about fruit and vegetable classification. Cosine similarity was the main metric used to identify if the word embeddings could produce similar classifications of eight items as documentation from the country's health organizations. Both general corpora and domain specific (recipe) word embeddings produced classifications similar to what the health organizations recommended. However, the general corpora word embeddings was able to provide better context of why an item was classified the way it was as opposed to the domain specific corpora. The recipe datasets used to train the domain specific word embeddings would have benefitted from more documents in order for the method to perform better. This work has shown that reaching a universal classification of what is a fruit and what is a vegetable has a ways to go, but that there is benefit in exploring how these items are discussed, and thus exploited with NLP.

References

Ministry of Health of Brazil and Secretariat of Health Care. 2015. *Dietary Guidelines for the Brazilian population*. Ministry of Health of Brazil.

Adrianos Ferreira. 2016. Afrodite - o maior livro de receitas culinárias em língua portuguesa. `https://github.com/adrianosferreira/afrodite.json`.

Nathan Hartmann, Erick R. Fonseca, Christopher Shulby, Marcos Vinícius Treviso, Jessica Rodrigues, and Sandra M. Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *CoRR*, abs/1708.06025.

U.S. Department of Health, Human Services, and U.S. Department of Agriculture. 2015. *Dietary Guidelines for Americans*. Health and Human Services Dept. and Agriculture Dept.

Ingrid Keller. Preliminary results to the who fruit & vegetable survey. Presentation of Ms I. Keller, Technical Officer, Noncommunicable Disease Prevention and Health Promotion Department, World Health Organization, Geneva, Switzerland, given during the WHO Fruit and Vegetable Initiative Expert Meeting.

Javier Marín, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2018. Recipe1m: A dataset for learning cross-modal embeddings for cooking recipes and food images. *CoRR*, abs/1810.06553.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Neural and Information Processing System (NIPS)*.

Weiqing Min, Shuqiang Jiang, Shuhui Wang, Jitao Sang, and Shuhuan Mei. 2017. A delicious recipe analysis framework for exploring multimodal recipes with various attributes. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, pages 402–410, New York, NY, USA. ACM.

Farhad Nooralahzadeh, Lilja Øvrelid, and Jan Tore Lønning. 2018. Evaluation of domain-specific word embeddings using knowledge resources. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

World Health Organization. Fruit and vegetable promotion initiative.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Frances E Thompson, Gordon B Willis, Olivia M Thompson, and Amy L Yaroch. 2011. The meaning of 'fruits' and 'vegetables'. *Public Health Nutrition*, 14(7):1222–1228.

Zachary Vance. 2017. 140,000 english recipes in computer-readable form with photos and crawl. `https://archive.org/details/recipes-en-201706`.

Michael Wiegand, Benjamin Roth, and Dietrich Klakow. 2012. Knowledge acquisition with natural language processing in the food domain : Potential and challenges.

Mengnan Zhao, Aaron J. Masino, and Christopher C. Yang. 2018. A framework for developing and evaluating word embeddings of drug-named entity. In *Proceedings of the BioNLP 2018 workshop*, pages 156–160, Melbourne, Australia. Association for Computational Linguistics.