Semantics-Driven Ingredient Substitution in the FoodKG *

Sola S. Shirai[1 [0000-0001-6913-3598]], Oshani Seneviratne[1 [0000-0001-8518-917X]], Minor E. Gordon[1 [0000-0003-1928-4130]], Ching-Hua Chen[2 [0000-0002-1020-0861]], and Deborah L. McGuinness[1 [0000-0001-7037-4567]]

[1] Rensselaer Polytechnic Institute, Troy NY 12180, USA
{shiras2, senevo, gordom6}@rpi.edu
dlm@cs.rpi.edu
[2] IBM Research, Yorktown Heights NY 10598, USA
chinghua@us.ibm.com

Abstract. People who would like to improve their eating patterns can make small changes in their diet by substituting ingredients in recipes. "Good" substitutions may have multiple dimensions and definitions, and one definition is to maintain the recipe while adhering to constraints that may include personal preferences, allergies, and nutritional or other dietary considerations. Our proposed system automatically finds and ranks plausible substitution options using a knowledge graph of food information. We evaluate our substitute ranking heuristic using a novel data set of ground-truth substitutions, showing promising preliminary results.

1 Introduction

Eating habits can play an important role in improving personal health. For example, patients diagnosed with diabetes typically receive recommendations to monitor their intake of specific nutrients from foods (like carbohydrates and protein) to treat their condition. If patients find that some nutrients that they are monitoring need adjustments, they may choose to modify their diet by substituting ingredients in some of their recipes. Such substitutions can remove restricted types of ingredients (e.g., common allergens) or replace ingredients that most negatively affect patient health (e.g., replacing potatoes to reduce carbohydrate intake). Substituting individual ingredients rather than strictly following a new meal plan allows patients to eat familiar meals while maintaining their dietary goals.

Our work aims to empower people to improve their eating habits by simplifying the process of identifying ingredient substitutions that satisfy their dietary constraints. We facilitate this using our FoodKG [7], a knowledge graph of recipe and ingredient information. We use linked information about ingredient classification and nutrition to filter valid options for ingredient substitutions. We also

* Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-2721/paper561.pdf

automatically rank plausible ingredient substitutions using several heuristics for substitutability. Our approach differs from previous works in that others have tended to identify substitutions only using either similarity metrics [1, 2, 11] or explicit rules based on types of foods [6, 5, 12, 4]. In this paper, we present methods and preliminary results for our contributions of a novel heuristics-based approach to rank ingredient substitutions and an evaluation using a "ground-truth" data set of substitutable ingredients.

2 Ranking Plausible Ingredient Substitutions

Intuitively, our approach to ingredient substitution ranking should consider similarities in the properties of ingredients as well as the similarity of the recipes in which they are used. We also must consider how to determine whether ingredients that are "similar" are also good substitutes. Some combinations of ingredients (e.g., garlic and olive oil) may be similar in the sense that they co-occur frequently in recipes, but they may not be substitutable. On the other hand, some ingredients that often are good substitutes (e.g., potatoes and cauliflower) may be dissimilar in terms of some properties like their food classification.

In order to create a ranked set of substitution options, we seek to develop a heuristic combining several scores based on latent and explicit semantic information about ingredients. We use two sources of latent semantics in the form of word embeddings based on ingredient names. The first is a Word2Vec [10] model from Recipe1M [9], trained over ingredient names and recipe instructions. The second is a pre-trained word embedding model from Spacy [8]. Cosine similarity is used to compare ingredient names with the expectation that good ingredient substitutions would have similar embeddings.

For explicit semantic information, we compute two substitutability scores based on the intuitions that (1) good substitutions should pair well with similar ingredients, and (2) good substitutions should be used in similar recipes. Additionally, we use the linked ontology of food from FoodOn [3] to further generalize the ingredients and recipe. For example, rather than just analyzing whether ingredients occur together with "whole milk," we can further generalize the ingredient as "milk" or as a "dairy food product." We believe the rich semantic information from FoodOn will allow our approach to more effectively identify substitutions for a wider variety of ingredients than previous works.

Combined Heuristic and Filtering We use a heuristic combining the aforementioned scoring metrics to rank the substitutability of ingredients. We also experiment with employing a filtering strategy for substitution options based on the assumption that super- or sub-classes of the target ingredient are not the most useful "substitutes." For example, this filtering strategy would remove the option of "yellow onions" as a substitute for "onions" in a recipe. Linked information about ingredient nutrition can also be used to filter out substitution options matching certain nutritional criteria (e.g., potato substitutes with lower carbs). Figure 1 shows a high-level overview of the substitution ranking process.

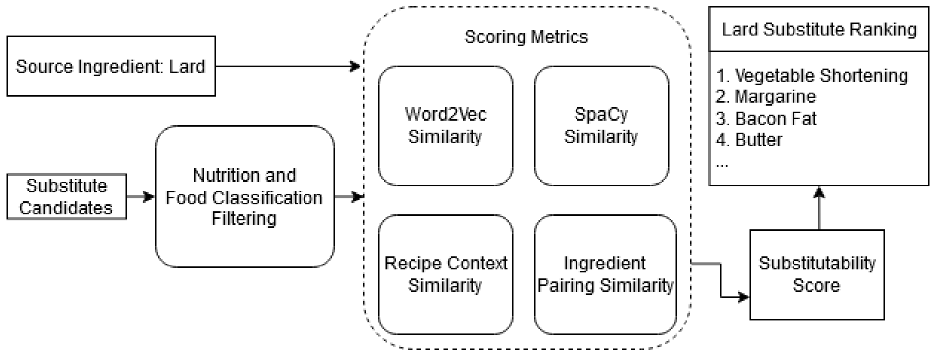

Fig. 1. Example of the process used to score and rank substitution ingredients for the target ingredient "Lard." Filtering is applied to ignore undesirable ingredients, and the output substitutability score is used to rank substitution options.

Ground-Truth Data Collection Because there is no widely accepted gold-standard data set for ingredient substitutions, we have taken the approach of selecting an online source to serve as a "ground-truth" for a wide variety of ingredient substitutions. We collected substitution data from The Cook's Thesaurus[3], yielding a set of 2,300 substitutions pairs (i.e., pairs of "target ingredient" to "substitute ingredient"). 1,161 unique ingredients were present in the data set, providing substitutions for 928 target ingredients.

3 Results

We evaluate our results by producing a list of ranked substitution options for each ingredient and comparing it against our ground-truth data. Since our method produces substitution rankings for each target ingredient, we frame our approach as an information retrieval problem and compute mean average precision (MAP) and mean reciprocal rank (MRR) as our evaluation metrics. MAP allows us to assess our algorithm's ability to rank all substitutions because the number of correct substitutes varied for each ingredient (e.g., in our evaluation data set, "potato" had 6 substitutes while "white asparagus" only had 1). MRR then gives us insight into the highest-ranked correct substitution for each ingredient. With our filtering strategy applied, our combined heuristic achieves a MAP of 0.418 and MRR of 0.499 over our ground-truth data. Comparing it against a simple baseline using word embedding similarity, our heuristic shows some promising improvement over the Word2Vec model (MAP 0.297 and MRR 0.385) and SpaCy (MAP 0.274 and MRR 0.371).

[3] foodsubs.com

4 Discussion

Although our approach was able to show better performance than simple baselines, there are many limitations in the methods and evaluation. Our filtering strategy was not always able to successfully filter out undesirable options because of incorrect links to FoodOn classes as well as limitations of the information captured in FoodOn. Sometimes the valid substitutions from the algorithm that were filtered out were not in line with our ground-truth data. For example, garlic powder is classified as a subclass of garlic in FoodOn, which led our approach to incorrectly filter it out of garlic's substitution options. This reflects our reuse of an ontology that perhaps had different intentions for how its subclass relationships may be used from our reuse expectations.

Although relying on a single website as our ground-truth allowed our evaluation to avoid subjectivity issues that may arise from user studies, it is also limited in its scope. While the source of our evaluation data provides a wide variety of substitution options, it does not cover all possible substitutions that people may use (e.g., many websites cite zucchini as a valid substitute for potatoes, but it was not in our ground-truth data). Finally, our current evaluation method does not assess other factors that may be involved in judging the goodness of substitutions (e.g., the fitness of a substitution for a particular type of diet, or the fitness of a substitute in vastly different types of recipes).

5 Conclusion and Future Work

Using FoodKG and its linked information about ingredient classification, our approach for identifying ingredient substitutions shows promise when compared against simple baselines. Our current best heuristic achieves a MAP of 0.418 and MRR of 0.499 in our collected ground-truth substitution data set. Work remains to improve both our strategies for ingredient substitute ranking as well as its evaluation. Moving forward, we are also expanding our approach to include more explicit semantics about the "healthiness" of substitutions in combination with more sources to develop a more comprehensive evaluation.

Acknowledgements

This work is partially supported by IBM Research AI through the AI Horizons Network.

References

1. Achananuparp, P., Weber, I.: Extracting food substitutes from food diary via distributional similarity. CoRR (2016)
2. Akkoyunlu, S., Manfredotti, C., Cornuéjols, A., Darcel, N., Delaere, F.: Investigating substitutability of food items in consumption data p. 5 (2017)

3. Dooley, D.M., Griffiths, E.J., Gosal, G.S., Buttigieg, P.L., Hoehndorf, R., Lange, M.C., Schriml, L.M., Brinkman, F.S.L., Hsiao, W.W.L.: FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. npj Science of Food **2**(1), 1–10 (Dec 2018)
4. Gaillard, E., Infante-Blanco, L., Lieber, J., Nauer, E.: Tuuurbine: A Generic CBR Engine over RDFS. In: Case-Based Reasoning Research and Development, vol. 8765, pp. 140–154. Springer International Publishing, Cham (2014)
5. Gaillard, E., Lieber, J., Nauer, E.: Improving ingredient substitution using formal concept analysis and adaptation of ingredient quantities with mixed linear optimization. In: Computer Cooking Contest Workshop, Frankfurt, Germany, September 28-30, 2015. CEUR Workshop Proceedings, vol. 1520, pp. 209–220. CEUR-WS.org (2015)
6. Gaillard, E., Lieber, J., Nauer, E.: Adaptation of TAAABLE to the CCC'2017 Mixology and Salad Challenges, Adaptation of the Cocktail Names. In: International Conference on Case-Based Reasoning (ICCBR) Computer Cooking Contest Workshop. pp. 253–268. Trondheim, Norway (Jun 2017)
7. Haussmann, S., Seneviratne, O., Chen, Y., Ne'eman, Y., Codella, J., Chen, C.H., Mcguinness, D., Zaki, M.: FoodKG: A Semantics-Driven Knowledge Graph for Food Recommendation, pp. 146–162 (10 2019)
8. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017)
9. Marin, J., Biswas, A., Ofli, F., Hynes, N., Salvador, A., Aytar, Y., Weber, I., Torralba, A.: Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Bengio, Y., LeCun, Y. (eds.) 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings (2013), http://arxiv.org/abs/1301.3781
11. Pan, Y., Xu, Q., Li, Y.: Food recipe alternation and generation with natural language processing techniques. In: 2020 IEEE 36th International Conference on Data Engineering Workshops (ICDEW). pp. 94–97 (2020)
12. Skjold, K., Øynes, M., Bach, K., Aamodt, A.: Intellimeal - enhancing creativity by reusing domain knowledge in the adaptation process. In: Proceedings of ICCBR 2017 Workshops (CAW, CBRDL, PO-CBR), Trondheim, Norway, June 26-28, 2017. CEUR Workshop Proceedings, vol. 2028, pp. 277–284. CEUR-WS.org (2017)