A Word Embedding Model for Mapping Food Composition Databases Using Fuzzy Logic

Andrea Morales-Garzón(✉), Juan Gómez-Romero, and M. J. Martin-Bautista

Department of Computer Science and Artificial Intelligence, Universidad de Granada, Granada, Spain
andreamgmg@correo.ugr.es, {jgomez,mbautis}@decsai.ugr.es

Abstract. This paper addresses the problem of mapping equivalent items between two databases based on their textual descriptions. Specifically, we will apply this technique to link the elements of two food composition databases by calculating the most likely match of each item in another given database. A number of experiments have been carried by employing different distance metrics, some of them involving Fuzzy Logic. The experiments show that the mappings are highly accurate and Fuzzy Logic improves the precision of the model.

Keywords: Word embedding · Fuzzy distance · Database alignment

1 Introduction

Nutrition and health organizations offer specialized and curated resources describing food and food composition, often under open access licenses. The most widely used resource is the database of the United States Department of Agriculture (USDA), which collects and harmonizes food facts from academic and industrial sources [10]. In Europe, the primary reference is the European Food Information Resource Network (EuroFIR), which compiles data from different European countries' databases [15]. There are also private initiatives such as i-Diet [22], an information system addressed to nutritionists to create personalized diets and focused on Spanish cuisine. Along with their nutritional information, i-Diet includes food item labels in Spanish and English.

These resources differ in scope and focus and usually struggle to capture the peculiarities of regional cuisines and the specificity of local products. At the same time, diet recommendation systems must be localized to the patients' context and need to be effective. Based on this principle, the Stance4Health project[1] aims

[1] Stance4Health (Smart Technologies for Personalised Nutrition and Consumer Engagement) is a project funded by the European Union under the Horizon 2020 research and innovation programme. More information: https://www.stance4health.com.

Supported by the European Union under grant agreement No. 816303.

M.-J. Lesot et al. (Eds.): IPMU 2020, CCIS 1238, pp. 635–647, 2020.
https://doi.org/10.1007/978-3-030-50143-3_50

at developing a personalized and localized nutrition service that will optimize the gut microbiota activity and long-term consumer commitment. The absence of wide-scope large databases, including regional and local products at the European and Spanish levels [16], makes it necessary in Stance4Health to combine resources mentioned above, e.g., USDA and i-Diet. However, this task is not trivial since these databases have significant differences in structure, semantics, and coverage. The latter, along with the vagueness associated to the language (e.g., a mapping of two equivalent items with different level of specialization) calls for flexible approaches to calculate the mappings.

In this paper, we propose a methodology based on a word embedding model to map food items' databases from their respective short descriptions in English. Similarity between items is calculated by using a (fuzzy) distance metric. In particular, we use this methodology to map the i-Diet and USDA databases: given an i-Diet food item, we calculate the most similar USDA item by measuring the distance between their embedding representations, obtained after encoding the short text associated with each of them with the learnt model.

In contrast to similar works, we use a larger corpus to train the language model and consider the complete recipes instead of just the ingredient list. This approach allows us to find matches between items that are different but have a similar role in several preparations, e.g., hazelnut and almond butter. This contribution could be used to cross-link food items used in different regional cuisines and to propose ingredient substitutions (or even new fusion dishes). More importantly, we expect that the mapped databases will support personalized nutrition in Stance4Health, as well as other Food Computing applications such as recipe nutrients calculation before and after cooking.

The remainder of this paper is structured as follows. In the following section, we contextualize our work within the recent literature on food item mapping and Food Computing. In Sect. 3, we further describe the data sources used in the study: USDA, i-Diet, and the corpus of recipes. Afterward, we describe the methodological approach (Sect. 4) and the experiments carried out (Sect. 5). In the last section, we analyze the results and interpret them. The paper finishes unfolding the conclusions of the work and hinting some promising directions for future work.

2 Related Work

Food Computing researchers have long acknowledged the need for a standard and open food and food components resource considering regional cuisines and cultural differences [20]. Given the effort required for such development and the absence of a central organization, the usual procedure is to extend the USDA database according to application needs [11]. In this regard, database and ontology merging and alignment techniques can be applied to find similarities and links between item registries automatically [21].

Food databases' principal elements are meals and ingredients. Therefore, it is possible to leverage ingredient detection and cuisine prediction methods to match

food items based on their constituents. For instance, in [28] and [25], we can find algorithms to classify recipes by country from their ingredients. Similarly, in [19], the authors identified cuisine by using topics extracted from the recipe's text. Predictive models have also been used to translate typical dishes from one region to another by applying an encoder-decoder Deep Learning architecture [12].

From a broader perspective, other research works studied the relation among ingredients and cooking methods from food data descriptions, as in [1]. These relationships can be reused to match food items in different databases. Our work follows the same strategy, but we learn a language model based on embeddings instead of a network of ingredients that appear together in recipes. Our approach has some advantages over the latter, such as avoiding the need for precise ingredient identification in the texts. The latter problem has been extensively addressed in the literature, mostly by applying customized parsers statistical natural language processing, with limited results, e.g., [6,7,29].

Regarding the use of Deep Learning for recipe text processing, Food2Vec used a word embedding model trained only with the list of ingredients included in recipes [2]. In contrast, we also use the text describing the cooking instructions. Therefore, we obtain close encodings for ingredients that appear together in recipes (as in Food2Vec), but also for those that are involved in similar preparations (which is useful for cross-cultural item matching). The Recipe2Vec [5] tool does encode the whole text, although it focuses on recipe comparison and retrieval and not publicly available. Food images were used in [26] to enhance the embedding model. Since we do not have image information in the recipes of our corpus, analyzing the possible improvement after incorporating images remains as future work.

Furthermore, we must take into account that the food text includes the use of food brands, often replacing ingredients themselves. Moreover, brand information also appears in the USDA database. Consequently, our language model must be able to deal with such terms. We follow the guidelines of [8], which identified semantically-related terms with an embedding model, including brands.

Finally, we can use several metrics to measure the distance between two words encoded according to the model [9] and, more interestingly, between two short texts [13]. In this context, similarity techniques combining token-based similarity and Fuzzy Logic [30] can be applied to obtain the mappings. We leverage and validate these approaches to formulate a fuzzy distance metric to tackle both vagueness of the language and syntactic/semantic content within the tokens.

3 Data

We used the English recipe corpus published by archive.org[2] to build the word embedding model. This corpus collates recipes extracted from several websites, e.g., BBC Food Recipe, Epicurious, Cookstr, and AllRecipes. The final corpus includes 267,071 texts. The records corresponding to each recipe source can be seen in Table 1.

[2] https://archive.org/download/recipes-en-201706.

Table 1. Recipe corpus: sources and number of records

Data source	Records
BBC Food Recipe	10,679
Epicurious	20,111
Cookstr	225,602
AllRecipes	10,679
Total of records	**267,071**

As mentioned in the introduction, the databases used in this work are i-Diet and the USDA Food Composition Databases. i-Diet is a proprietary database that provides nutritional content of food items usually found in Spanish diets. The USDA database, in turn, contains more extensive and more detailed data, since its scope goes beyond the use in diet recommendations. Examples of their structure and fields are respectively shown in Tables 2 and 3. Due to the nature of the databases, item descriptions have a substantial variability.

Each register in the i-Diet Food Composition Database corresponds to a food item, which can be a complete meal or an ingredient. A food item register consists of an identification number, a description of the item in Spanish, the corresponding translation of the description into English, and the food group to which the item belongs in Spanish. Translations in i-Diet have been performed manually by nutritionists. Additionally, each register includes numerical fields corresponding to the nutritional values of the item. The mapping procedure only uses the English description field; others are discarded.

Table 2. Example of food items in the i-Diet Food Composition Database

ID	Description (ENG)	Group	...
96	Onion	HORTALIZAS BULBOSAS[a]	...
290	Apple	FRUTAS[b]	...

[a] Bulbous vegetables
[b] Fruits

The structure of the USDA Food Composition Database is similar. Each food item register in USDA encompasses an identification number, a short description of the item, a food group category, and the category description. The rest of the fields are related to the item nutritional facts (mostly major and minor nutrient values). The mapping only uses the description field; the others are discarded.

4 Methods

Our methodology is organized into four main steps: (1) data preprocessing, (2) word embedding model training and parameter tuning, (3) distance metrics,

Table 3. Example of food items in the USDA Food Composition Database

Food code	Main food description	WWEIA Category code	WWEIA Category description	...
75117020	Onions, mature, raw	6414	Onions	...
63101210	Apple, cooked or canned, with syrup	6002	Apples	...

(4) calculation of mappings by computing the Word Mover's Distance between pairs of short texts from the encodings obtained with the trained model, and (5) validation of the mappings. These steps are further described in the following sections.

4.1 Data Preprocessing

Although the recipe corpus was already collated and published on the web in a readable format, an extra preprocessing stage was required to prepare the data to train the model:

1. We extracted the data from the text files, i.e., the ingredient list and the cooking instructions. (Note that we did not consider ingredients and instruction separately.) These two pieces of data were filtered and saved in text files, one per recipe.
2. We performed a typical text cleaning process: conversion to lowercase; removal of punctuation marks, digits and special characters; removal of stop words; and lemmatization.
3. The clean data was used to train a bigram model to detect compound words. For this step, we used the Software Framework for Topic Modelling with Large Corpora [24]. English stop words were also imported from this module.

The steps above were applied to the cooking instructions presented in the recipes, e.g., the recipe text "*Combine nutritional yeast, salt, cumin, garlic powder, onion powder, paprika, chili powder, and cayenne pepper in a small bowl.*" is turned into "*combin nutrit yeast salt cumin garlic_powder onion powder paprika_chili powder cayenn pepper small_bow*" after the preprocessing phase.

4.2 Model Training and Parameter Tuning

We built the language model from a corpus of text recipes by using Word2Vec [17, 18], an unsupervised Deep Learning algorithm for the creation of word embeddings. An embedding is a set of numeric vectors, each one coding a feature, which represents a language unit preserving its semantics [3]. That is, two related language units (e.g., words) will have encodings located closely in the embeddings space. Therefore, they allow us to operate with the embeddings in a meaningful way; e.g., $\langle$ king $\rangle$ - $\langle$ man $\rangle$ + $\langle$ woman $\rangle$ = $\langle$ queen $\rangle$. There are other algorithms

for learning word embeddings that can be used with the same purpose, such as GloVe [23] and *fasttext* [4].

Since a generic word embedding model does not encompass such a specific domain as food from a nutritional context, a Word2Vec model was trained on the preprocessed corpus by using the Continuous Bag Of Words (CBOW) implementation, also provided by the Software Framework for Topic Modelling with Large Corpora [24]. We trained the model using the cooking instructions as a whole entry to the training model, instead of processing every sentence from each recipe separately. The nature of the text of the corpus, with short sentences and frequent anaphora, suggests that this is the most suitable approach. Experimental work and comparison to other works confirmed this assumption [27].

4.3 Distance Metrics

Let S_i be the textual representation of an item, and let $T_i = \{t_1, ..., t_n\}$ be the token set obtained as a result of the preprocessing task of such item; e.g., consider the item k whose textual representation is S_k= *"Canned fish, average"*, the corresponding T_k would be {"*can*", "$fish$", "*averag*"}.

We formulate the mapping problem between two items as finding the minimal distance of an item token set against every item token set from the other database. For that purpose, the different distance metrics listed below were compared.

Crisp Distance Metrics

- *Jaccard Distance: JACCARD* is a token-based distance metric which quantifies the distance based on the lexical difference between the token sets [30]:

$$JACCARD(S_1, S_2) = 1 - \frac{|T_1 \cap T_2|}{|T_1| + |T_2| - |T_1 \cap T_2|} \quad (1)$$

- *Word Mover's Distance: WMD* treats a text document as a cloud of words; each word represented as a point in the vector embeddings space [14]. The distance between two clouds is quantified by the minimum cumulative distance that words from one text document need to travel to match exactly the point cloud of the other text document. To calculate the distance between two single words, an Euclidean Distance between the corresponding vector representation is used. Therefore, WMD takes advantage from the semantic information provided by the word embedding model.
- *Hybrid Distance:* Preliminary studies within this work showed that using a unique distance measure, either lexical or semantic, strongly reduces the precision of the model. Therefore, we propose a hybrid distance measure formulated as a weighted combination of Jaccard and Word Mover's Distances.

$$HDISTANCE(t_1, t_2) = wJACCARD(t_1, t_2) + (1 - w)WMD(t_1, t_2) \quad (2)$$

$$\text{where } w \in \mathbb{R} \text{ and } 0 \leq w \leq 1$$

Fuzzy Distance Metrics

– *Fuzzy Jaccard Distance* [30]: This metric consists of a combination of token-based similarity and character-based similarity to determine the fuzzy overlap set. The Jaccard Distance described above is used to measure the distance between tokens, and a threshold determines which ones belong to the fuzzy overlap set. This latter parameter has been empirically tuned to 0.2.

$$FJACCARD_{\delta}(S_1, S_2) = \frac{\left|T_1 \widetilde{\cap}_{\delta} T_2\right|}{|T_1| + |T_2| - \left|T_1 \widetilde{\cap}_{\delta} T_2\right|} \tag{3}$$

$$\delta = 0.2$$

– *Fuzzy Document Distance:* We propose a fuzzy approach of the distance between short documents, considering each document as a token set. The distance between two sets is calculated as the Euclidean Distance between the vectors' tokens in both sets. These vectors correspond to the numerical representation obtained from the Word Embedding model previously trained. The fuzzy function is described as follows:

$$FDIST(S_1, S_2) = \frac{\sum_{x \epsilon T_1 \cup T_2} min(\mu S_1)x \times min(\mu T_2)x}{\sum_{x \epsilon T_1} (\mu T_1)(x) + \sum_{x \epsilon T_2} (\mu T_2)(x) - \sum_{x \epsilon T_1 \cup T_2} min(\mu S_1)x \times min(\mu T_2)x}$$

$$\mu_{T_i}(x) = \begin{cases} sigmoid(\frac{1}{distance(t_i,x)}) & 0 < distance(t_i, x) < \infty \\ 1 & distance(t_i, x) = 0 \\ 0 & distance(t_i, x) = \infty \end{cases} \tag{4}$$

where $distance(t_i, x)$ is the Euclidean distance between t_i and x

Noted that the membership of a token x to a set S_i is defined as the minimum distance of x to every token in S_i.

4.4 Mapping Food Items

Once the embedding model is available, it can be used to compare the similarity of two words. To this aim, as already introduced, we tested different metrics to get the most accurate results. Our mapping procedure calculated item mappings for each i-Diet register. That is, for each i-Diet item, we obtained the distance between its English description and the description of every USDA item. The algorithm finally returns the USDA item that minimizes the distance, i.e., the most likely match. Let us mention that we tackle the mapping as a multilabel classification problem, where there are many labels as USDA items apart from the "*No matches*" label (which represents the case where there is not a possible matching for an item between the databases).

4.5 Validation

A nutrition expert validated the quality of the mappings by verifying their exactness. Note that, in some cases, there may be more than one best candidate mapping (i.e., with the same quality). This situation typically happens when items in one database are more general (hypernym) than the corresponding items in the other one (hyponyms). In these cases, the validation labels the mapping as *correct* as long as one of the possible best mappings is retrieved.

Different flexibility levels have been considered to detect the robustness of the model. We obtained the number of i-Diet items where the best possible matching is achieved. We also calculate a less restrictive accuracy value, that allows us to determine the number of items whose best matching is reached between the first and the tenth candidate from the whole USDA database.

5 Experiments

The embedding model was trained during 30 epochs with vector dimensionality set to 300 and a window of size 5. Words that appear less than three times in the whole corpus are ignored. The final model yielded a vocabulary of 11,288 words. Mappings were calculated for every i-Diet food item (735 items). One human expert manually assigned the validation label of each mapping.

The results of the validation of the mappings with the different metrics are showed in Table 4. The first column "Top 1" shows, for each metric, the percentage of items whose best possible matching is achieved by the model. The rest of columns show, respectively, the percentage of items in which the best matching is found in the 2,3,5 or 10 best candidates. The weight parameter of (3) was empirically tuned to achieve the optimal performance ($w = 0.2$).

Table 4. Accuracy of the model (%) obtained with the different metrics

Distance metric	Top 1	Top 2	Top 3	Top 5	Top 10
(1) Jaccard Distance	16.75	20.16	22.20	25.20	27.52
(2) Word Mover's Distance	30.65	35.55	36.92	40.87	44.82
(3) Hybrid Distance	32.15	37.12	40.19	43.05	47.41
(4) Fuzzy Jaccard Distance	23.84	29.70	33.37	39.23	45.64
(5) Fuzzy Document Distance	35.55	40.46	43.46	47.00	53.26

A sample of the final results is provided in Tables 5 and 6. In both cases, matching are carried out using the distance metric with the best performance (see Table 4). Both tables have the same structure. In the first column, we show the original i-Diet item name. Columns from 2 to 4 show the results of the mapping: from left to right, the English description of the source i-Diet item, the description of the mapped USDA item, and the distance between both of them. The last column corresponds to the most accurate mapping identified manually.

6 Discussion

Table 5 shows a selection of successful mappings between i-Diet and USDA, i.e., mappings labeled as correct. Rows (1) to (3) show that when equivalent items had a similar text description in both databases, the model was able to match them properly. Note that a lower distance value of a mapping with respect to another one does not necessarily entail that it is better. The relative values of the distance metric are useful to select the best match for a given item, but not to compare different mappings.

Table 5. Selected examples of correct mappings

	Mapping			
	◇	△	▽	◯
1	Pickle, cucumber, sour	Cucumber pickles, sour	0.0	Cucumber pickles, sour
2	All Bran Kellogg's	(Kellogg's All-Bran)	0.25	(Kellogg's All-Bran)
3	Sunflower seeds	Sunflower seeds, NFS	0.333	Sunflower seeds, NFS
4	Meat extract 'Bovril'	Meat, NFS	0.578	Meat, NFS
5	Blue Cheese	Cheese, Blue or Roquefort	0.333	Cheese, Blue or Roquefort
6	Pears canned	Pear, cooked or canned, in light syrup	0.571	Pear, cooked or canned, in light syrup
7	Canned fish, average	Fish, NS as to type, canned	0.6	Fish, NS as to type, canned
8	Chicken giblets	Chicken liver, fried	0.5	Chicken liver, fried
9	Pate liver not specified	Liver paste or pate, chicken	0.6	Liver paste or pate, chicken

◇ i-Diet text (ENG) △ USDA mapped item ▽ WMD ◯ Best USDA mapping

Rows (4) to (6) illustrate more difficult mappings that were correctly solved by the procedure. In these cases, the model was capable of matching item descriptions even though one of them was slightly less specific than the other. In particular, row (4) includes a commercial brand. Rows (5) and (6) correspond to cases in which the model can map a broad (i-Diet) description with a more precise one (in USDA). Last but not least, the rows (7) to (9) show correct mappings that were not as obvious as the previous ones.

We also found some limitations to our approach, as depicted in Table 6, largely due to the coverage of the corpus and errors in translations in i-Diet

from the original Spanish item description into the English one. First, in rows (1) and (2), we can see that items with no real translation and that are never used in the English recipe corpus were not mapped. We expected this behavior since there is no proper embedding for the terms used in the description. Accordingly, a more diverse corpus should be used, including recipes for local cuisines.

Table 6. Selected examples of not found, acceptable, approximate, and wrong mappings

	Mapping			
	◇	△	▽	◯
1	Salchichon	No matches	1.0	Sausage, NFS
2	Morcilla asturiana (38,5%H)	No matches	0.793	Blood Sausage
3	Fondu cheese	Cheese fondue	0.0	Cheese fondue
4	Cottage cheese	Cheese, cottage, NFS	0.0	No matches
5	cocoa and hazelnut butter, Nocilla, Nutela	Almond butter	0.8	No matches
6	Wine Special	Wine, nonalcoholic	0.666	No Wine, table, white
7	Strawberry mermelade	Strawberries, raw	0.666	Jam, preserve, all flavors
8	Low fat sausage	Buttermilk, low fat (1%)	0.5	Sausage, NFS / Pork sausage
9	Scallop	Potato, scalloped, NFS	0.66	Scallops, cooked, NS as to cooking method

◇ i-Diet text (ENG) △ USDA mapped item ▽ WMD ◯ Best USDA mapping

Besides, rows (3) and (4) illustrate mappings where the Spanish text is poorly translated, and therefore the mapped item has a slightly different meaning. In these cases, mappings are marked as acceptable because, despite their similar semantics, there is a better match in USDA. These problems could be addressed by manually editing the translations or by using a (more accurate) machine translation system.

Rows (5) to (7) show approximate mappings in which the link USDA is semantically related, but the association is not correct or can be improved. It is interesting to highlight that row (5) include a food brand that is correctly identified. Also, row (5) shows a case of mapping a local item and a replacement with similar usage.

Finally, rows (8) and (9) depict incorrect mappings due to the limitations of the corpus and the (unfrequent) case that the i-Diet item is more specific than

the possible USDA candidates. The last column of (9) shows that dealing with hypernym and hyponyms is difficult, and can lead to several possible candidate mappings in USDA for one item in i-Diet.

As shown in Table 4, the fuzzy metrics improved the outcomes obtained with crisp approaches. From the obtained results we can draw the conclusion that vagueness of the language can make Fuzzy Logic a suitable option to tackle the matching task. Given the dimensionality and complexity of the problem, the results are reasonably accurate.

7 Conclusions and Future Work

This research work was motivated by the need for mapping two food composition databases with different scopes. This problem poses additional obstacles when the food items correspond to different regions and local cuisines. We created a word embedding model to address these issues and showed that this technique has the potential to facilitate working with non-overlapping data resources in the Food Computing domain. Our model worked well with regional brands and was able to some extent to identify substitute items used in similar preparations. Fuzzy distance metrics showed better performance than crisp alternatives.

For the future, we plan to improve the mappings by training the embedding model with a larger-scale recipe corpus and by improving the translations of Spanish item descriptions into English in i-Diet. A relevant aspect of our approach that can be further explored is the capability for finding ingredient replacements in recipes, which also entails using more imprecise knowledge. These replacements can either refer to the same item expressed differently, or to similar ingredients more often used in a particular region or cuisine. This kind of situation cannot be addressed by more traditional techniques –e.g., regex and concordances– without resorting to a specialized and comprehensive knowledge base. The absence of such resources is indeed the original motivation for our work. This same idea can be applied to recipe retrieval and automatic generation of recipes.

This work only considered English text recipes from the web. Consequently, some bias is introduced, since the popular dishes from other countries could not have sufficient representation in the collected corpus. Nevertheless, since international dishes have been introduced in cuisines from all over the world we consider that this corpus is suitable to generate useful word embeddings. We acknowledge that including typical recipes from other cuisines would help to improve the model performance. As well, more sophisticated measures can be added as well as combined with the implemented ones. Additionally, Machine Translation techniques can be applied to the Spanish text descriptions in order to reduce the errors generated by the manual translations. Also, we plan to research a multi-modal extension of this work, combining short text embeddings with the numerical fields from Food Composition Databases and other media resources, e.g., images.

Acknowledgements. This work was supported by the European Union under grant agreement No. 816303 (Stance4Health).

References

1. Ahn, Y.Y., Ahnert, S.E., Bagrow, J.P., Barabási, A.L.: Flavor network and the principles of food pairing. Sci. Rep. **1**, 96 (2011)
2. Altosaar, J.: Augmented cooking with machine intelligence. https://jaan.io/food2vec-augmented-cooking-machine-intelligence/
3. Bengio, Y., Ducharme, R., Pascal Vincent, C.J.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003)
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
5. BuzzFeed: Recipe2vec: How word2vec helped us discover related tasty recipes. https://pydata.org/nyc2017/schedule/presentation/65/
6. Chang, M., Guillain, L.V., Jung, H., Hare, V.M., Kim, J., Agrawala, M.: RecipeScape: an interactive tool for analyzing cooking instructions at scale. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, p. 451. ACM (2018)
7. Chen, Y.: A Statistical Machine Learning Approach to Generating Graph Structures from Food Recipes, August 2017
8. Fan, Y., Pakhomov, S., McEwan, R., Zhao, W., Lindemann, E., Zhang, R.: Using word embeddings to expand terminology of dietary supplements on clinical notes. JAMIA Open (2), 246–253 (2019). https://doi.org/10.1093/jamiaopen/ooz007
9. Farouk, M.: Measuring sentences similarity: a survey. ArXiv (2019)
10. Gebhardt, S., et al.: USDA national nutrient database for standard reference, release 21. United States Department of Agriculture Agricultural Research Service (2008)
11. Ispirova, G., Eftimov, T., Korošec, P., Koroušić Seljak, B.: MIGHT: statistical methodology for missing-data imputation in food composition databases. Appl. Sci. (19), 4111 (2019). https://doi.org/10.3390/app9194111
12. Kazama, M., Sugimoto, M., Hosokawa, C., Matsushima, K., Varshney, L.R., Ishikawa, Y.: A neural network system for transformation of regional cuisine style. Front. ICT **5**, 14 (2018). https://doi.org/10.3389/fict.2018.00014, https://www.frontiersin.org/article/10.3389/fict.2018.00014
13. Kenter, T., Rijke, M.D.: Short text similarity with word embeddings. In: International Conference on Information and Knowledge Management, Proceedings, pp. 1411–1420 (2015). https://doi.org/10.1145/2806416.2806475
14. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances, pp. 957–966 (2015)
15. Laboratory, U.N.D., Consumer, U., Institute, F.E.: USDA nutrient database for standard reference (1999)
16. Lupiañez-Barbero, A., Blanco, C., De Leiva, A.: Spanish food composition tables and databases: need for a gold standard for healthcare professionals (review). Endocrinología, Diabetes y Nutrición (English ed.), July 2018. https://doi.org/10.1016/j.endien.2018.05.011
17. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at ICLR, January 2013

18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, October 2013
19. Min, W., Bao, B.K., Mei, S., Zhu, Y., Rui, Y., Jiang, S.: You are what you eat: exploring rich recipe information for cross-region food analysis. IEEE Trans. Multimed. **20**(4), 950–964 (2018). https://doi.org/10.1109/TMM.2017.2759499
20. Min, W., Jiang, S., Liu, L., Rui, Y., Jain, R.: A survey on food computing, August 2018. http://arxiv.org/abs/1808.07202
21. Noy, N.F., Musen, M.A.: Prompt: algorithm and tool for automated ontology merging and alignment. In: Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, pp. 450–455. AAAI Press (2000). http://dl.acm.org/citation.cfm?id=647288.721118
22. Gestión de Salud y Nutrición, S.: I-diet food composition database, updated from original version of g. martín peña fcd (2019)
23. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014). http://www.aclweb.org/anthology/D14-1162
24. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, pp. 45–50, May 2010
25. Sajadmanesh, S., et al.: Kissing cuisines: exploring worldwide culinary habits on the web. In: 26th International World Wide Web Conference 2017, WWW 2017 Companion, pp. 1013–1021 (2019). https://doi.org/10.1145/3041021.3055137
26. Salvador, A., et al.: Learning cross-modal embeddings for cooking recipes and food images. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3068–3076, July 2017. https://doi.org/10.1109/CVPR.2017.327
27. Sauer, C.R., Haigh, A.: Cooking up food embeddings understanding flavors in the recipe-ingredient graph (2017)
28. Singh, R., Arora, H.: CSE 255 assignment 2 cuisine prediction/classification based on ingredients (2015)
29. Takahashi, J., Ueda, T., Nishikawa, C., Ito, T., Nagai, A.: Implementation of automatic nutrient calculation system for cooking recipes based on text analysis. In: Anthony, P., Ishizuka, M., Lukose, D. (eds.) PRICAI 2012. LNCS (LNAI), vol. 7458, pp. 789–794. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-32695-0_74
30. Wang, J., Li, G., Fe, J.: Fast-join: an efficient method for fuzzy token matching based string similarity join. In: 2011 IEEE 27th International Conference on Data Engineering, pp. 458–469. IEEE (2011)