Automated Identification of Food Substitutions Using Knowledge Graph Embeddings

Julie Loesch[1], Louis Meeckers[1], Ilse van Lier[2,3][0000−0001−8381−1252], Alie de Boer[2][0000−0002−6500−4649], Michel Dumontier[4][0000−0003−4727−9435], and Remzi Celebi[4][0000−0001−7769−4272]

[1] Data Science and Knowledge Engineering, Maastricht University, Netherlands
j.loesch,l.meeckers@student.maastrichtuniversity.nl
[2] Food Claims Centre Venlo, Campus Venlo, Maastricht University, Venlo, Netherlands
[3] Chair Youth, Food, and Health, Maastricht University Campus Venlo
i.vanlier,a.deboer@maastrichtuniversity.nl
[4] Institute of Data Science, Maastricht University, Netherlands
michel.dumontier,remzi.celebi@maastrichtuniversity.nl

Abstract. Healthy eating is a daily challenge for many, which is influenced by various factors such as taste, accessibility, price, and the food environment. Consumers often are insufficiently informed about healthier options for the foods they consume. Being able to identify healthy alternatives for foods according to similarities in nutritional value will help consumers choose products that they prefer. This work aims to identify healthy alternatives to foods that also have similar nutritional characteristics through the use of knowledge graph embeddings (KGEs). The quality of the KGEs is assessed against a newly created ground truth, which is verified by two domain experts. Hence, this work presents a newly created ground truth food substitution data set and describes the development of a food recommender system that identifies healthier alternatives to foods.

Keywords: Healthy food choice · nutritional profile · ingredient substitution · Knowledge graph embedding · Food similarity.

1 Introduction

An unhealthy diet is associated with an increased risk on a range of health issues and diseases. Multiple studies have shown that chronic diseases such as cardiovascular disease, high blood pressure, type 2 diabetes, some cancers, and poor bone health are linked to poor dietary habits [1]. At the same time, health crises such as the COVID-19 pandemic highlight the importance of a healthy diet, as dietary and health status have been shown to influence people's ability to prevent, combat, and recover from infections [2]. Even though no specific foods or dietary supplements can prevent or cure infections such as COVID-19, healthy diets are important to support an individual's immune system [3].

While healthy diets are known to be important, it is known that individuals do not always make healthy dietary choices. Even though information about nutritional values, ingredients, and even health effects of foods is made available on food labels, this information is not always used to make healthy dietary decisions [4,5]. There are various factors that influence the food choices individuals make, which are not limited to social, political, cultural, and individual factors (e.g., habits). General knowledge of nutritional aspects of food plays an important role as well [6]. Studies show a relation between nutrition knowledge of individuals and their overall diet quality [7,8]. Providing individuals with tools to select unfamiliar foods that are similar to, or even have a better nutritional value, than the ones they are familiar with, could increase the quality of their diet and subsequently, their overall health. To this extent, it is important to create a system that provides individuals with personalized dietary information [6].

Previous efforts to automate the selection of food substitutions have been limited by the absence of an accepted data set of valid substitutions. For this reason, Shirai and colleagues [9] proposed to scrape online resources for a ground truth food substitution data set and developed a heuristic that ranks plausible food substitutions. The researchers created semantically interlinked food information by linking USDA[5], FoodOn Ontology [10] and FoodKG [11]. Moreover, the authors incorporated "healthy" ingredient substitution options into their work as previous works did not consider personal dietary constraints on nutritional information. Shirai and colleagues [9] considered two categories of dietary constraints, namely restrictions on the types of ingredients that may be consumed (e.g., replacing meat-based ingredients for vegetarian alternatives or replacing allergens such as peanuts), and limitations on the consumption of certain nutrients (e.g., replacing high-carb ingredients with low-carb alternatives). However, their "healthy" ingredient substitution options are limited, which is why our work explores the use of knowledge graph embeddings to identify a broad range of food substitution options.

More precisely, this study presents an approach to find alternative food products with comparable or more favourable nutritional profiles that fall within a similar product category using knowledge graph embeddings. With this, a recommender system is built that suggests healthier substitutes for the ingredients and food products to its user. The knowledge graph of food is based on two open data sets, namely OpenFoodFacts[6], which is a food products database, and USDA, which provides nutritional information of food products. Furthermore, due to the low quality and unavailability of the existing ground truths (food review and cook thesaurus, used in the work of Shirai, et al., 2021 [9]), we curated an expert-verified data set for the evaluation of food substitution recommendations. The data and code to generate the analysis are made available at our Github repository[7].

[5] `https://fdc.nal.usda.gov/index.html`
[6] `https://world.openfoodfacts.org/`
[7] `https://github.com/MaastrichtU-IDS/healthy-food-subs`

2 Background

2.1 Knowledge Graphs

A knowledge graph is a graph, composed of a set of assertions (edges labeled with relations) that are expressed between entities (vertices). A knowledge graph is made up of three main components: nodes, edges, and labels. Any object, place, or person can be a node, while an edge defines the relationship between the nodes. The directed edges are often called triplets and are represented as a (h, r, t) tuple, where h is the head entity, t is the tail entity, and r is the relation associating the head with the tail entities. For instance, the triplet (banana, contains, protein) would describe the fact that protein is contained in a banana.

2.2 KG Embeddings and Similarity

Knowledge graph embeddings are low-dimensional representations of the entities and relations in a KG. Compared to high-dimensional representations of KGs such as the adjacent matrix, these representations are more efficient at identifying the semantic similarities. There are many popular KGE models, such as TransE [12] and Complex [13]. Essentially, what most methods do is to create a vector for each entity and each relation. These embeddings are then generated in such a way that they capture latent properties of the semantics in the knowledge graph, that is, similar entities and similar relationships will be represented with similar vectors. Thus, these KGE models differentiate by their scoring function, which measures the distance of two entities relative to its relation type in the low-dimensional embedding space. These score functions are used to train the KGE models so that the entities connected by relations are close to each other, while the entities that are not connected are far away.

3 Related Work

Eftimov and colleagues [14] showed the utility of representing food data as embeddings, which are in the form of vectors of continuous numbers. The authors used the FoodEx2 data, which is a comprehensive system for classifying and describing food items developed by the European Food Safety Authority (EFSA) [15] to learn vector representations by using the Pointcaré graph-embedding learning method [16]. The authors showed the utility of such vector representations on four different problems: i) automated determination of different food groups, ii) automated detection of the food class for each food concept (raw, derivative or composite), iii) identification of most similar food concepts for a given food concept, and iv) qualitative evaluation by a food expert. Hence, the authors introduced the concept of vector representations for food, or food embeddings, that can be used for downstream food data analysis and is available as an open-source resource. Moreover, their experiments have shown that the FoodEx2vec embeddings outperformed traditional feature representations for food data analysis.

One common problem when people prepare food is that some required ingredients of a recipe are not available. In order to deal with this issue, Pan and colleagues [17] collected recipe data of different cuisine styles from a website hosting thousands of recipes (Spoonacular[8]) to generate ingredient and recipe embeddings. Calculating the cosine similarity (i.e. the measure of similarity that computes the cosine of the angle between two non-zero vectors) of two ingredients or two recipes enables people to choose alternative ingredients, or even recipes. For instance, the authors found out that "Calamari" is the substitute of "Carrot". However, no formal evaluation of the results is provided by the authors.

A promising way to find food substitutes is to use the vast amounts of (mostly textual) cooking-related data to draw conclusions about which food items can replace one another. For that reason, Pellegrini and colleagues [18] exploited NLP techniques and trained two models, namely word2vec [19] (named Food2Vec) and BERT [20] (named FoodBERT) on recipe instructions from the Recipe1M+ dataset[9]. The Food2Vec approach is divided in two parts. The first part calculates text-based embeddings for all ingredients and optimally concatenates them with image-based embeddings. In the second part, these embeddings are used in addition with KNN to predict food substitutes. The only difference to the FoodBERT approach is that the latter calculates text-based embeddings for up to 100 occurrences of every ingredient and adds a further scoring and filtering step before predicting food substitutes. The authors evaluated their results by human evaluation and created a list of ground truth substitutes for a subset of ingredients, showing good performance.

Transey and colleagues [21] presented diet2vec, which is a scalable and robust approach for modeling nutritional diaries from smart phone apps. The authors analyzed massive amounts of nutritional data generated by 55k active users of a diet tracking app, called LoseIt[10]. To model the foods, the authors first ran word2vec [19] on the names of the food and subsequently ran weighted k-means to cluster the foods into $5,000$ "food words", placing 20% of the weight on the name and 80% of the weight on the nutrients. The authors then generated meal vectors via the DBOW model of paragraph2vec [22]. Similar to the foods, the authors clustered the meal vectors to get "meal words". The authors then represented each user's diet as a bag of meal words and again generated diet vectors, which were clustered into 100 diet words. The clusters generated by the authors are interpretable: however, no formal evaluation of the results is provided.

4 Methodology

The first step was to construct knowledge graph data in RDF format and create semantically interlinked food knowledge by linking OpenFoodFacts and USDA.

[8] `https://spoonacular.com/`
[9] `http://pic2recipe.csail.mit.edu/`
[10] `lhttps://www.loseit.com/`

In the second step, food substitution recommendations were extracted using the knowledge graph by applying different graph embedding approaches, namely, TransE [12], Complex [13] and RDF2Vec [23].

4.1 Datasets

USDA USDA consists of 8,618 different foods and provides the information on both macronutrients and micronutrients. To incorporate the USDA data set into a knowledge graph, we used the previous work (also known as FoodKG) of Haussmann et al., 2019 [11].

OpenFoodFacts OpenFoodFacts is an open and collaborative database which gathers more than $1,600,000$ products from over 150 countries. For each food product, information such as categories, nutritional data, Nutri-Score, ingredients, origin, and allergens were retrieved.

Ground Truth To create ground truth substitution data, we first looked at accessible substitution data from Food.com reviews[11]. We used the script provided by [9][12] to scrape the substitutions from Food.com reviews. We linked the ingredients to the USDA food items via Limes framework (see Section Linking for the details). The linking was reviewed manually and incorrect matches for the ground truth ingredients were removed from the ground truth. After cleaning and linking, 1,841 candidate substitute pairs remained from 3,846 samples in this dataset. We built an additional candidate food substitution list to increase the amount of available substitutions. We used the RDF2Vec-based similarity algorithm (see Section Embeddings) for the most commonly consumed foods to generate candidate substitutions and took the top 20 foods with the highest similarity scores for each food. Two domain experts (nutrition scholars and co-authors AdB and IvL) were asked to annotate these candidate food pairs as being a correct substitution or not, based on a pre-determined set of criteria.

Before labeling, the experts compiled a list of criteria for nutritional content similarity [13] based on data about macronutrients and various micronutrients and then applied this list to the candidate substitution dataset. Two researchers reviewed the list of 3,344 candidate substitutions between 966 unique food items independently and labeled all items based on the criteria defined. The annotation results were compared with each other and the inter-agreement between the two experts was computed using Cohen Kappa score. The Kappa score for inter-agreement between these two experts was 0.88, which indicates a strong agreement. In total, 1,847 substitutions spanning 786 unique food items approved by both experts were added to the ground truth.

[11] `https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions`

[12] `https://github.com/solashirai/FoodSubstitutionDataScripts`

[13] Supplementary material: `https://doi.org/10.6084/m9.figshare.16658284.v1`

4.2 Linking

We used Limes[14], a discovery framework for linking the Web of Data, to create relations between the food ingredients of USDA and OpenFoodFacts using a cosine similarity measure. More precisely, the metric employed evaluates the similarity between two input strings, taking an inner product space that measures the cosine of the angle between their vector representations. We set a threshold of 0.8 to accept results from linked ingredients based on manual inspection.

4.3 Enrichment of Knowledge Graph

The KG was enriched by tagging the ingredients based on the nutritional content we calculated according to the U.S. FDA's Recommended Dietary Allowances (RDAs)[15]. The tags that indicate the presence of rich mineral or vitamin content were added to the knowledge graph. Each food was tagged as high in a nutrient if the level of that nutrient contained in the food per serving is more than 30% of its respective RDA. This is the cut-off point that is used for nutritional content claims in the EU. In the EU, a nutritional content claim that a food is high in a certain vitamin or mineral, and any claim likely to have the same meaning for the consumer, may only be made where the product contains at least twice the value of 'source of (NAME OF VITAMIN/S) and/or (NAME OF MINERAL/S)'. In other words, the food should contain at least 30% of the RDA of a specific mineral/vitamin to be tagged as 'high in'. The distribution of the generated tags from the USDA dataset is depicted in Figure 1.

4.4 Embeddings

TransE Translation based embedding model (TransE) [12] is a representative translational distance model that represents entities and relations as vectors in the same semantic space. A relational fact is represented as a triplet (h, r, t) where h stands for the head, r represents the relation, and t denotes the tail. A vector representation of every entity and relation in the knowledge graph can be computed by training a neural network model, which minimizes the energy function $f(h, r, t) = ||h + r - t||$. The key idea is to make the sum of the head vector and the relation vector as close as possible to the tail vector.

Complex Complex [13] scoring function is based on the Hermitian dot product, meaning that it involves the conjugate-transpose of one of the two vectors. Consequently, the dot product is not symmetric anymore, which is why complex vectors can effectively capture anti-symmetric relations.

RDF2Vec RDF2vec [23] is a tool for creating vector representations of RDF graphs by creating a numeric vector for each node in an RDF graph. Thus, RDF2Vec [23] generates (random) walks on the knowledge graph data to be used as input for word2vec [19] neural networks. Word2vec [19] represents each word

[14] `https://github.com/dice-group/LIMES/releases`

[15] Food Component: `https://www.fda.gov/media/99059/download` and Nutrient: `https://www.fda.gov/media/99069/download`

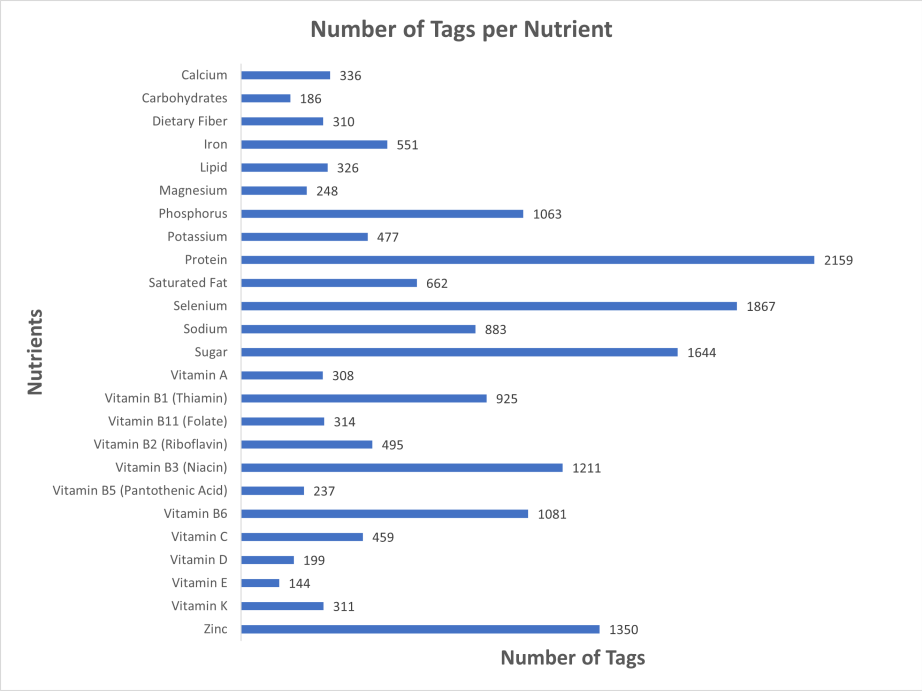

Fig. 1: Number of Tags per Nutrient.

with a low-dimensional vector, called word embeddings, where semantically and syntactically closer words appear closer in the vector space. Thus, word2vec [19] trains a neural network model to learn vector representation of words to predict a target word from its surrounding words.

5 Evaluation and Results

We first applied TransE [12], Complex [13], and RDF2Vec [23] models on different subsets of the knowledge graph. The results of the experiments are shown in Table 1a. We evaluated the performance of the models by using Mean Reciprocal Rank (MRR), Mean Average Precision (MAP), and Recall Rate at k (RR@k). The MRR is the average of the reciprocal rank, which measures the reciprocal of the rank (multiplicative inverse of the rank) at which the first relevant ingredient was retrieved. The MAP is the average of the average precision, which is the mean of the precision after each relevant food is retrieved. The RR@k is the proportion of relevant ingredients found in the top-k recommended food substitutions.

Table 1a shows that the best performance results were obtained with the RDF2Vec method. RDF2Vec achieved a Recall Rate of 0.33 and 0.4 for the top 5 and top 10 results respectively, indicating a significant performance for a recommender system. While the MAP and MRR values seem relatively low, 0.133 and 0.234, it should be noted that these metrics were calculated by looking

Table 1: Results of experiments

(a) Without filtering

	Method	MAP	MRR	RR@5	RR@10
USDA KGE	TransE	0.057	0.136	0.195	0.259
	Complex	0.057	0.141	0.195	0.265
	RDF2Vec	0.083	0.180	0.259	0.332
USDA + Tags KGE	TransE	0.071	0.158	0.238	0.297
	Complex	0.081	0.185	0.246	0.300
	RDF2Vec	0.101	0.199	0.276	0.365
USDA + Tags + OpenFoodFacts KGE	TransE	0.093	0.202	0.286	0.362
	Complex	0.079	0.179	0.262	0.311
	RDF2Vec	0.133	0.234	0.330	0.400

(b) With filtered ranking using food category

	Method	MAP	MRR	RR@5	RR@10
USDA KGE	TransE	0.121	0.216	0.297	0.386
	Complex	0.113	0.211	0.305	0.400
	RDF2Vec	0.115	0.212	0.305	0.414
USDA + Tags KGE	TransE	0.125	0.211	0.303	0.438
	Complex	0.135	0.242	0.330	0.438
	RDF2Vec	0.136	0.235	0.330	0.430
USDA + Tags + OpenFoodFacts KGE	TransE	0.144	0.253	0.351	0.454
	Complex	0.140	0.247	0.327	0.414
	RDF2Vec	0.154	0.259	0.359	0.438

at the rank order of the substitute foods among all food items in the USDA database (8,618 ingredients), not only ground truth foods.

In order to see how food category information affects the results, we restricted the recommended substitutes to be in the same food category as the query food. More precisely, we made sure to filter out substitutes that were not in the same food category as the query food. The results in Table 1b show that all metrics have improved significantly with this filtering strategy.

6 Discussion

Overall, Table 1a and Table 1b show encouraging results from our objective to build a recommender system for substituting food products. Table 1b shows an improvement over the results shown in Table 1a by including food category information in the ranking calculation. It is logical to consider category information in ranking substitutes as most of the foods in the same category have similar nutritional profiles. However, the ranking might not be practical for some specialized diets. For example, the ranking may fail to recommend meat substitutions for specialized diets such as vegan or vegetarian diets, because their diet will not permit the recommendations from the meat category. On the other hand, it should be noted that the similarities between foods are mainly based on nutritional values.

This study describes the development of a food recommender system that identifies healthier alternatives to target foods. These healthier alternatives are

food products that have a more favourable nutritional profile within their product category, based on key macro- and micronutrients. However, when searching for food substitutes, people often focus on other factors such as taste, functionality, accessibility, or dietary restrictions [6]. For example, some people may wish to replace potatoes to reduce carbohydrate intake, or replace peanuts because of allergens. This is not yet included in the ground truth. These mentioned factors, that are known to affect food product selection and dietary choices, are a good direction for future work.

7 Conclusion

In this work, an unsupervised method using the knowledge graph embedding based similarity for food substitution is presented. The quality of knowledge graph embeddings for this task was assessed against a newly created ground truth which was verified by two domain experts. Even though the ground truth can be further optimised and the recommender system can be further developed by also including other variables to compare food products with each other, this ground truth is one of the first steps in making it easier to let people identify alternative food products. We believe that KGE based recommender can be improved further with existing supervised methods such as Graph Neural Network since a training dataset (ground truth) is now made available. As a future work, we would like to extend the recommender system by using an actual nutrient profiling system that is currently being used in specific countries to identify foods as being healthy or not. We also plan to use and compare the state-of-the-art supervised methods to train on ground truth data created.

References

1. Jill Jin. Dietary Guidelines for Americans. *JAMA*, 315(5):528–528, 02 2016.
2. *Maintaining a healthy diet during the COVID-19 pandemic*. FAO, 2020.
3. Michael J. Butler and Ruth M. Barrientos. The impact of nutrition on covid-19 susceptibility and long-term consequences. *Brain, Behavior, and Immunity*, 87:53–54, 2020.
4. Alie de Boer. Fifteen years of regulating nutrition and health claims in europe: The past, the present and the future. *Nutrients*, 13(5), 2021.
5. A.C. Hoek, D. Pearson, S.W. James, M.A. Lawrence, and S. Friel. Healthy and environmentally sustainable food choices: Consumer responses to point-of-purchase actions. *Food Quality and Preference*, 58:94–106, 2017.
6. Christoph Trattner and David Elsweiler. Food recommender systems: Important contributions, challenges and future research directions. 11 2017.
7. Dahyun Park, Yoo Kyoung Park, Clara Yongjoo Park, Mi-Kyung Choi, and Min-Jeong Shin. Development of a comprehensive food literacy measurement tool integrating the food system and sustainability. *Nutrients*, 12(11), 2020.
8. Maartje Poelman, S. Dijkstra, Hanne Sponselee, Carlijn Kamphuis, Marieke Battjes-Fries, Marleen Gillebaart, and Jaap Seidell. Towards the measurement of food literacy with respect to healthy eating: The development and validation of the self perceived food literacy scale among an adult sample in the netherlands. *International Journal of Behavioral Nutrition and Physical Activity*, 15, 06 2018.

9. Sola S. Shirai, Oshani Seneviratne, Minor E. Gordon, Ching-Hua Chen, and Deborah L. McGuinness. Identifying ingredient substitutions using a knowledge graph of food. *Frontiers in Artificial Intelligence*, 3:111, 2021.
10. E. Griffiths, Damion M. Dooley, P. L. Buttigieg, R. Hoehndorf, F. Brinkman, and W. Hsiao. Foodon: A global farm-to-fork food ontology. In *ICBO/BioCreative*, 2016.
11. Steven Haussmann, O. Seneviratne, Yu Chen, Yarden Ne'eman, James Codella, Ching-Hua Chen, D. McGuinness, and Mohammed J. Zaki. Foodkg: A semantics-driven knowledge graph for food recommendation. In *SEMWEB*, 2019.
12. Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, page 2787–2795, Red Hook, NY, USA, 2013. Curran Associates Inc.
13. Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction, 2016.
14. Tome Eftimov, Gorjan Popovski, Eva Valenčič, and Barbara Koroušić Seljak. Foodex2vec: New foods' representation for advanced food data analysis. *Food and chemical toxicology : an international journal published for the British Industrial Biological Research Association*, 138:111169, April 2020.
15. European Food Safety Authority (EFSA). The food classification and description system foodex 2 (revision 2). *EFSA Supporting Publications*, 12(5):804E, 2015.
16. Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
17. Yuran Pan, Qiangwen Xu, and Yanjun Li. Food recipe alternation and generation with natural language processing techniques. In *2020 IEEE 36th International Conference on Data Engineering Workshops (ICDEW)*, pages 94–97, 2020.
18. Chantal Pellegrini., Ege Özsoy., Monika Wintergerst., and Georg Groh. Exploiting food embeddings for ingredient substitution. In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF,*, pages 67–77. INSTICC, SciTePress, 2021.
19. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.
20. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
21. Wesley Tansey, Edward W. Lowe Jr. au2, and James G. Scott. Diet2vec: Multi-scale analysis of massive dietary data, 2016.
22. Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents, 2014.
23. Petar Ristoski and Heiko Paulheim. Rdf2vec: Rdf graph embeddings for data mining. In Paul Groth, editor, *The Semantic Web - ISWC 2016 : 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part I*, volume 9981, pages 498–514, Cham, 2016. Springer International Publishing.