FoodKG: A Semantics-Driven Knowledge Graph for Food Recommendation

Steven Haussmann [1] Oshani Seneviratne [1] Yu Chen [1]
Yarden Ne'eman [1] James Codella [2] Ching-Hua Chen [2]
Deborah L. McGuinness [1] Mohammed J. Zaki [1]

[1]Rensselaer Polytechnic Institute

[2]IBM

Oct. 30, 2019

Motivation

What Ails Us

Many major modern health problems are tied to food:

- Obesity
- Diabetes
- Hypertension

An Overabundance of Choices

Ironically, these are often the product of modern abundance

- Excessive caloric intake causes obesity
- Sugary foods can cause type II diabetes
- Too much salt raises blood pressure

An Overabundance of Information

Just as abundant as the food is the information

- The latest fad diet
- Nutritional labels
- That one food you should never eat
- Guidelines from health organizations
- ...

Wrangling the Facts

We want to bring some order to these oceans of data

- . . . without tremendous time investment
- . . . without losing sight of where it came from

Applying the Facts

We'd like to support the following use-cases:

- Describing the nutritional content of a recipe
- Providing suggestions for ingredient alternatives
- Suggesting foods that fit the user's personal diet

Sources

Recipes

Recipes describe . . .

- what foods are made of
- how foods are made

Sometimes, they also tell us . . .

- where the food comes from
- what other people think

Recipes

We decided to use the Recipe1M dataset [2]

- Contains approx. one million recipes from various sources
- Includes ingredients, steps, titles, and sources

Recipes

We're interested in additional information, though:

- Tags - region of origin, preparation time, etc.
- Serving sizes - important for calculating nutritional values

We acquired these via web scraping

Nutrients

Nutrition is, in theory, straightforward:

- Most packaged food has a nutritional label
- Plenty of well-regarded data for "standard" ingredients

However, recipes are a bit challenging

Nutrients

In theory, a recipe's nutrients are the sum of its ingredients' nutrients - but challenges exist . . .

- Some recipes use uncommon ingredients
  - Specific brands, specialized ingredients, etc.
- Some recipes use difficult names
  - Adjectives, comments, and more
- Some recipes have strange or ambiguous units
  - One package, a pinch, "some", . . .

Nutrients

We obtained nutritional data from the USDA's Legacy Standard Reference dataset

- Covers several thousand food items of various types
- Includes several dozen nutrients
- Includes at least one measure - e.g. number of grams in one tablespoon of butter

Nutrients

The last step is to find the links between recipe ingredients and the USDA's ingredients

- Butter = "Butter, Salted"
- Pasta sauce = "Sauce, pasta, spaghetti/marinara, ready-to-serve"

This is done via fuzzy string matching

- Rather error-prone, due to disagreeing vocabularies, choices of names, etc.
- Some human oversight would help here - manually identify the most common ingredients?

Derived Data

We can thus compute the nutritional profile of an entire recipe.

Meaning

We might have recipes and nutrition, but we're missing something:

- Apples and pears are both fruits
- Chicken is not vegetarian
- Walnuts are a kind of nut

. . . we're missing the *meaning* of foods

Meaning

Therefore, we have linked our ingredients to FoodOn - a "field-to-fork" ontology [1]

- FoodOn provides a hierarchy for foods, covering . . .
  - properties
  - type
  - origin
- FoodOn is human-curated, providing very high accuracy

Creating these linkages enhances the *meaning* of our ingredients

Implementation

Philosophy

We want to build the FoodKG in a reproducible, extensible manner

- Should be *buildable from sources*
- Should be built deterministically
- Should be easy to run at any scale, small or large

Architecture

Our Python-based framework takes the form of a pipeline.

- Data is brought in via **sources**
- Data is transformed via **processors**
- Data is written out via **sinks**
- All of these produce and consume streams of data

Architecture

The system is decidedly object-oriented:

- Each stream of data is a list of objects
- Objects can refer to other objects
  - Recipes have ingredients
  - Ingredients have names/quantities/units
- Objects define how they are identified . . .
  - URI
- . . . and how they are serialized
  - RDF triples
  - Nanopublications

Architecture

Nanopublications [3] attach provenance information to RDF triples

- Who claimed the facts
- When they claimed the facts
- Where they got the information from

Each class (Recipe, Ingredient, etc.) stores provenance data, and includes it when serialized

Architecture

Therefore, objects record . . .

- . . . their own data
- . . . their own provenance
- . . . the data and provenance of their children

Architecture

We can easily make incremental improvements to the knowledge graph

- Build the core FoodKG with recipes, ingredients, and linkages
- Query for each recipe's ingredients/nutrients and compute recipe-level nutrition

Application

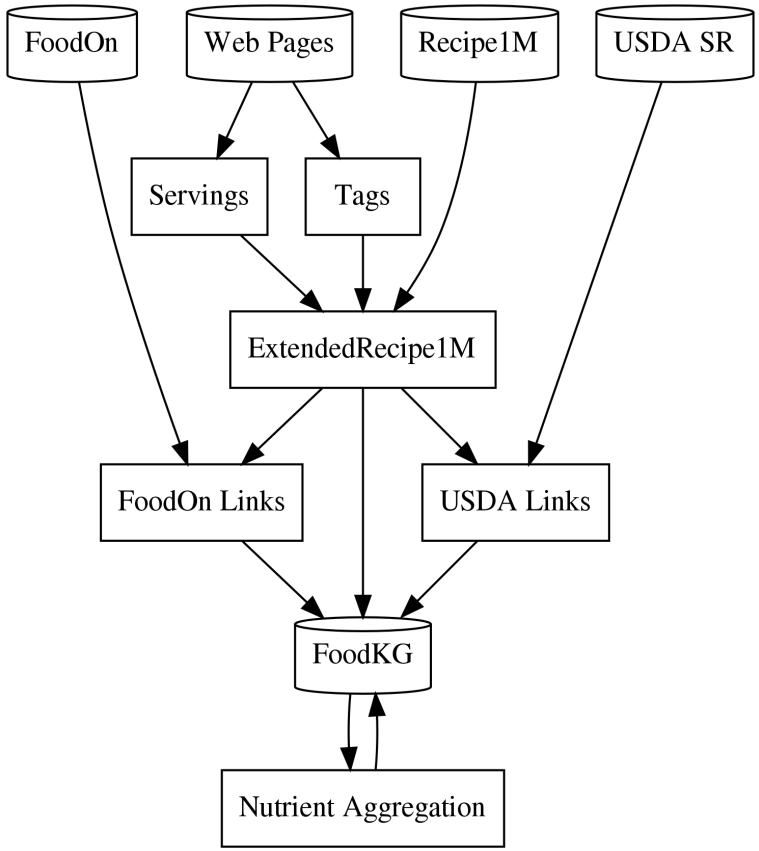

Figure: The construction of the FoodKG

Usage

Applications

- The FoodKG itself is a *resource*, not an application in its own right
- We'll briefly examine one of its current uses.

KBQA

A **Knowledge-Base Question Answering** system answers natural-language questions

- Learns to provide answers from a knowledge base

KBQA

We built a simple web application to . . .

1. Identify a question's intent with the IBM Watson Assistant (e.g. "recipe with constraints")
2. Pass the parsed question into the KBQA system, returning one or more URIs from the FoodKG
3. Execute a SPARQL query to transform the answer into HTML

In Action

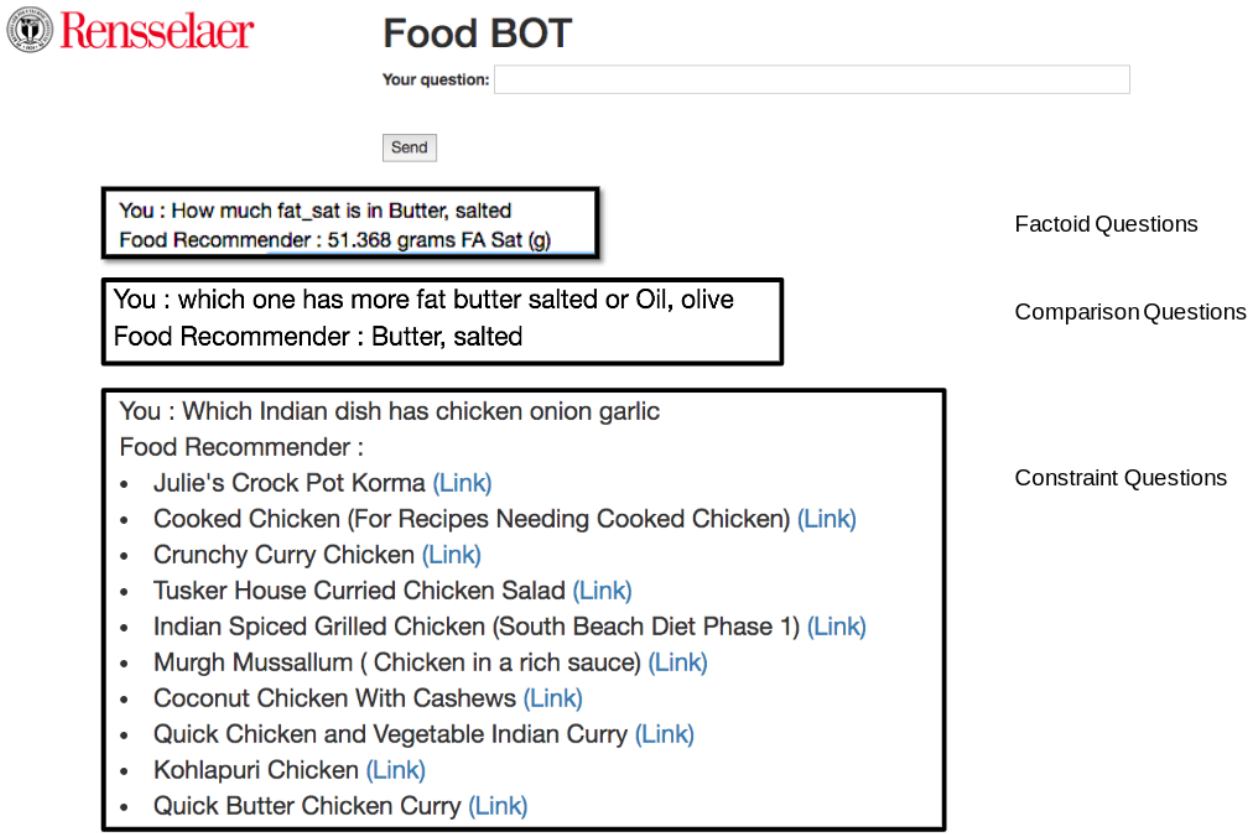

Figure: An example of the KBQA system in action (taken from slides by O. Seneviratne)

Statistics

Currently, the full FoodKG covers . . .

- 1,029,720 recipes
- 18,253 unique ingredient names
- 7,793 USDA nutrient records
- 7,315 FoodOn classes

Resources

The FoodKG can be constructed via the tools presented below:

- Website: `http://foodkg.github.io`
- GitHub Repository: `https://github.com/foodkg/foodkg.github.io/`

References I

Damion M. Dooley et al. "FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration". en. In: *npj Science of Food* 2.1 (Dec. 2018), p. 23. ISSN: 2396-8370. DOI: 10.1038/s41538-018-0032-6. URL: http://www.nature.com/articles/s41538-018-0032-6 (visited on 10/27/2019).

Javier Marin et al. "Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images". In: *IEEE Trans. Pattern Anal. Mach. Intell.* (2019).

What is a Nanopublication. URL: http://nanopub.org/wordpress/?page_id=65.

Acknowledgments

Thanks to . . .

- The Rensselaer Polytechnic Institute and IBM for organizing the *Health Empowerment by Analytics, Learning, and Semantics* project
- Mohammed J. Zaki, my research advisor
- Oshani Seneviratne, Yu Chen, Yarden Ne'eman, and Deborah L. McGuinness at RPI
- James Codella and Ching-Hua Chen at IBM

Funding by:

- IBM, supporter of *HEALS*