Priyanka Upadhyay
2581714, s8prupad@stud.uni-saarland.de
Commonsense Knowledge extraction and Consolidation Seminar

Task 1

- Difference between Commonsense knowledge (CSK) and Encyclopedic knowledge

  - CSK goes through the general concept whereas Encyclopedic Knowledge refers to instances.
  - CSK: Generalizes across subjects
    Encyclopedic Knowledge: Typical Facts.

- Example: University

  - Commonsense Knowledge Statement:
    1. Students and Professors will be part of University.
    2. University will have lecture halls.
    3. University will organize events and Activities.

  - Encyclopedic Knowledge Statements:
    1. University positions (Dean of CS)
    2. AsTA is part of the University event organization.
    3. University has MPI-INF.

- General distinction between CSK and Encyclopedic Knowledge:

  "General knowledge" can be broken down into a few types. First, there is real world factual knowledge, the sort found in an encyclopedia. Second, there is common sense, the sort of knowledge that an encyclopedia would assume the reader knew without being told." [1]

  - "If any plural form of an entity exist then it is not an Encyclopedic knowledge"
    Example: Elephant is a Concept but Elephants exist which is commonsense knowledge entity
  - Commonsense Knowledge (CSK): More advanced human-like reasoning
    Commonsense Knowledge Base (CKB): putting people commonsense into KB, **combination of lexical and Common knowledge base**. Example: ConceptNet, WebChild, QUASIMODO etc.
  - Encyclopedic Knowledge: Factual knowledge
    Encyclopedic Knowledge Base (EKB): putting Factual Knowledge into KB, Example: Knowledge graphs like DBpedia, Freebase, or Yago etc.

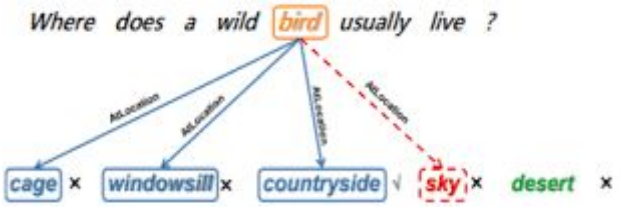

Fig 1. From ConceptNet to Commonsense

Table 1: Example of incorrect entity recognition results

Query	Stanford	Longest
邓卓棣美国国籍(Deng Zhuoli United States nationality)	邓卓棣(Deng Zhuoli)(PER), 美国(United States)(LOC)	邓卓棣(Deng Zhuoli), 美国国籍(United States nationality)
梦幻西游大唐符石(Tang runestone of the *Menghuanxiyou* computer game)	梦 幻 西(Meng Huanx-i)(PER)	梦幻西游大(*Menghuanxiyouda*-mobile game), 唐(the Tang dynasty), 符石(runestone)
爱回家2014过新年在第几集(Which episode of *Come Home Love* shows the 2014 New Year festival)		爱回家2 (*Come Home Love*2-book), 过新年(*New Year*-song)

Fig 2. Encyclopedic Knowledge

Task 2:

Semantic Network of Cooking Commonsense Knowledge Base:

There are such a big amount of different cuisines and Cooking preferences built by cultural, ethnical backgrounds, geographical locations and social classes. Meat eating habits, herbs, and crops – everything makes its contribution to the standard cuisine and culture. India is a land of spices, Africa is a continent of sauces, Europe discloses esthetical great things about Cooking and disclose new opportunities and inventions for those that value and luxuriate in eating.

My resource contribution is to create a Knowledge Base using Commonsense that may be accustomed to create a unified Cooking knowledge graph that links the various silos associated with Cooking while preserving the provenance information. The Cooking Commonsense KB is an Open Information Extraction problem and I can use publicly available data about Cooking for extracting and maintaining, and for constructing a knowledge graph that can be consumed by both humans and machines. I will make KG which is based on FAIR (Findable, Accessible, Interoperable, Reusable) principles. Cooking KG will have recipes, ingredients, and nutrients as resources and my work in this research is to consider the different datasets and link it with health concepts and the offering of a question-answering service as an application.

Some of the competency questions include questions such as: What are the ingredients and the total calorie count of a piece of a chocolate cake according to USDA(US Department of Agriculture) nutritional data?. The answer may include butter, eggs, sugar,flour, milk, and cocoa powder for the ingredients, and a calorie count of 424. For a diabetic who is trying to abide by the ADA (American Diabetes Association's) guidelines, a question like, How can I increase the fiber content of this cake? may be a natural follow-up question to ask. Similarly, a person suffering from lactose intolerance may ask What can I substitute for milk in chocolate cake?. Answering questions like this is not possible from sources such as DBpedia alone, because the information from those sources is not complete. Example, the dbo:ingredient for the resource dbr:Chocolate_cake6 contains only dbr:Cocoa_powder and dbr:Chocolate. CookingKG contains additional information from online recipe sites, along with the corresponding nutrient information from USDA, that has more relevant information than what is available on DBpedia. Therefore, to answer this question, I can use the semantic structure of my knowledge graph. I present a methodology that can be used to extract publicly available data on Cooking and construct a semantically meaningful knowledge graph that can power applications to help consumers understand their foods and discover substitutions.

STEP 1: Data Acquisition **[Time spent- 3 Months]:**

There are many challenges while gathering and integrating the data which has significant effect on consistency, accuracy, and completeness, and challenges are: **Invalid data, Incomplete data, Ambiguous Data**

Recipes: Few Online recipe sites allow users to browse and share recipes. Few others permit users to upload their own recipes.

Nutrients: I chose to use USDA's National Nutrient Database for Standard Reference (https://catalog.data.gov/dataset/food-and-nutrient-databasefor-dietary-studies-fndds), which

contains approximately 8,000 records for a variety of types of food and their nutrients. The majority of the foods are generic,rather than coming from a specific brand.

Food Ontologies: All the list of recipes and nutritional tables provides the huge information about millions entities but lack the ontology and I can solve this problem with Taxonomy. FoodOn provides an extensive taxonomy for foods, organizing them by source organism, region of origin, and so on.

Example: Indian Dish

↙ ↘

South Indian Dish (Dosa) | North Indian Dish (Chickpeas curry)

STEP 2: Knowledge Graph Construction [Time Spent- 5 months]:

A knowledge graph includes resources with attributes and entities, relationships between such resources, and annotations to express metadata about the resources. My complete Cooking knowledge graph contains several key components: i) Recipes and their ingredients, ii) Nutritional data for individual food items, iii) Additional knowledge about foods, and iv) Linkages between the above concepts.

Recipes: Each recipe describes the ingredients needed to produce a dish. Each recipe receives a unique identifier, which is accompanied by its name, any provided tags, and a set of ingredients. Each ingredient points to its name, unit, and quantity. **High-quality name recognition significantly improves the quality of later results.**

Nutrients: From recipes, I can produce a network of foods and their ingredients. However, without information about the nutritional content of each ingredient, I cannot make a meaningful health-related knowledge base. I use the USDA public nutrition dataset for this information.

Given with Recipes and Nutrients data, I can define the shape of the resulting knowledge graph via semantic relationships. After these annotations are completed, the semantic data dictionary conversion script is run to convert the tabular USDA data into quads, which are triples grouped into named graphs.

STEP 3: Knowledge Graph Augmentation [Time Spent- 4 months]:

With all of the data imported, I will work on the second phase which is construction of my knowledge graph is linkage. I leverage various entity resolution techniques to automatically connect various concepts together. To ensure that the dataset can be practically expanded and updated, I make use of well-studied linked data techniques to establish provenance of these derived relationships.

Entity Resolution: Names are the most obvious shared attributes between various domains of recipes, nutrients, and foods. For this reason, I have largely focused on entity resolution techniques that work on strings, such as cosine similarity, which performs quite well for matching by name, particularly after normalization. I also examined using word embeddings, such as word2vec and FastText, with a pretrained model to resolve names. However, results

were poor - likely a product of the embedding capturing only the general meaning of a statement.

Entity Selection: I found it beneficial to limit the domain of concepts to match against, both for the sake of performance and to maximize accuracy. Ex: For instance, many categories of food are rarely seen. ingredients - but, critically, have names that are similar to kinds of food that are relevant. Ex: baby food, with names such as `Babyfood, juice, apple' and `Babyfood, meat, lamb, strained'. I remove such entries, since they cause problems with linkage of ingredients. I also ignore text beyond the third comma, as I found that the distinction between entities becomes insignificant at that point; doing so also speeds up the linkage process.

Provenance and Publication Information: To provide clear provenance for every claim made in my knowledge graph - including both imported knowledge and inferred linkages - I have made extensive and consistent use of the RDF Nanopublication specification. Nanopublications represent atomic units of publishable information, attaching information about where it came from and who/what published it. They express this knowledge via linked data, using four named graphs: assertions graph, provenance graph, publication info, head graph.

Application of Cooking Commonsense Knowledge Base:

- Answering Competency Questions in SPARQL.
- Answering Competency Questions in Natural Language.

Notes:

- DBpedia - has structured content from the information created in the Wikipedia.
- The dbo prex refers to http://dbpedia.org/ontology and dbo:ingredient dereferences to http://dbpedia.org/ontology/ingredient.
- The dbr prex refers to http://dbpedia.org/resource and dbr:Chocolate_cake dereferences to http://dbpedia.org/resource/Chocolate_cake.

Reference:

[1] CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks