Enhancing a Common Sense Knowledge Graph with Intuitive Physics Content

DARPA Machine Common Sense

Tetherless World Constellation

Presentation by Jody Sunray

Mentors:
Deborah McGuinness
Minor Gordon

Research Statement

- Commonsense knowledge is universally accepted beliefs about the world.
  - Two important domains are needed to support commonsense reasoning:
    - Intuitive physics (e.g., understanding physical interactions)
    - Intuitive psychology (e.g., understanding human emotion, motives, and behavior)
- The Machine Common Sense (MCS) program broadly aims to improve the integration of commonsense knowledge in AI machines.
  - The goal of my project is to explore the extraction of knowledge from data sources that can be used to answer questions that require some notion of intuitive physics, more specifically spatial reasoning.
  - In order to achieve this goal, I need to collect data about relative size relationships between types of physical objects.
- Hypothesis:
  - It should be possible to extract relative size information about objects from semi-structured text, such as the Web Data Commons product corpus.

Work Plan

- Visually inspect the Web Data Commons offers corpus for products having dimensions that can easily be extracted.
  - The data is structured using the JSON Lines text format, making it relatively easy to process each product one line at a time.
  - Each line has various attribute-value pairs, like “title” and “description.”
  - Shown at the bottom of the slide is a sample product.
- Extract product dimensions from the Web Data Commons data set.
  - Use Python regular expressions to parse out the dimensions; for many products, the dimensions are found in the description (e.g., “15w x 10d x 17 5h in”).
- Use natural language processing (NLP) to map product titles and descriptions to generic object types in a taxonomy such as Wikidata.
  - spaCy is used for NLP and allows for easy part-of-speech tagging.
- Generate relationship statements in the form of triples such as (shredder, smallerThan, car).
  - The extracted relationships will get incorporated into the larger Common Sense Knowledge Graph (CSKG), which is a network of semantic relationships.

{**..."dimensions":"15w x 10d x 17 5h in"**,"shipping weight":"16 5 lbs","security level":"4","warranty":"2 year parts lifetime","product code":"shdeshsm1058"},"price":"usd 133 98","specTableContent":"hsm x6pro paper shredder model x6pro model mirco cut shred size 5 32 x 1 3 8 in sheet capacity 6 sheets feed opening 8 8 in shredder speed dimensions 15w x 10d x 17 5h in shipping weight 16 5 lbs security level 4 warranty 2 year parts lifetime product code shdeshsm1058",**"title":"hsm x6pro mirco cut paper shredder"**}

Method & Materials

- Materials:
  - offers_corpus_english_v2.jsonl contains corpus of product offers taken from the Web Data Commons knowledge source
  - Python package **json** to work with JSON data
  - Python module **re** for regular expression operations
  - Python module **random** to generate random sample of data
  - Open-source library **spaCy** for NLP, particularly part-of-speech tagging
- Methods:
  - Create a random sample of 100 products because the corpus is a very large data set.
  - Examine the product data and identify “good” examples having properly formatted dimensions that can be parsed.
    - Familiarizing myself with the data helped me to decide what the regular expressions should look like and make predictions about whether the algorithm will be able to parse out the correct product name and dimensions.
  - For each product in the JSON Lines file, extract the title and tokenize it (split the title into individual words and punctuation marks).
  - Tag each token with its part of speech and determine the simple product name by choosing the last noun that appears in the title.
  - Attempt to find dimensions in the product description or title using regular expressions.
    - The method `re.findall("\d+\s?\w+\sx\s\d+\s?\w", text)` will detect “200mm x 47mm”

Initial Results & Analysis

	Predicted		
Actual		Negative	Positive
	Negative	85 (TN)	7 (FP)
	Positive	0 (FN)	8 (TP)

Figure 1. Confusion matrix to measure dimension parsing performance for 100 items

- Human annotation enables comparison of the actual results of the algorithm to predicted outcomes.
- A confusion matrix is a table used to evaluate the performance of a machine learning program.
  - **True positive (TP):** Dimension parsed correctly as predicted
  - **True negative (TN):** No dimension/parsed incorrectly as predicted
  - **False positive (FP):** Dimension parsed incorrectly but predicted otherwise
  - **False negative (FN):** Dimension parsed correctly but predicted otherwise
- My algorithm successfully extracted only 8 dimensions out of a sample of 100 products, but 85 of the products do not have dimensions listed or have complex dimension structures that are difficult to parse.
  - The regular expressions are not applicable to all dimension formats.

Conclusion

- Web Data Commons is a useful source for gathering information about everyday objects.
- The data is semi-structured and requires additional parsing for specific purposes.
  - I focused on extracting object dimensions in order to support spatial reasoning.
- The same corpus could be used to support other types of reasoning, such as extracting prices in order to determine the relative value of products.
- Similar dimension parsing is being accomplished on unstructured product data by the data cleansing tool Data Ladder.
  - The ProductMatch feature enables product data standardization, attribute extraction, and pattern matching.
- Parsing semi-structured data is a common problem, but the MCS program's application is novel, as it supports commonsense reasoning.

This work is supported by DARPA MCS award number N660011924033. PI McGuinness.