W3C LBD Community Group
Minutes - Call 05/10/2021

Attendees:

- Mathias Bonduel (KU Leuven & Neanex Technologies, Belgium)
- Alex Donkers (Eindhoven University of Technology)
- Ranjith Soman (Imperial College London)
- Serge Chávez-Feria (Universidad Politécnica de Madrid)
- Kevin Luwemba Mugumya (University of Nottingham Malaysia)
- Hazar Karadag
- Katja Breitenfelder (Fraunhofer IBP & TU Munich, Germany)
- Edlira Vakaj (Birmingham City University, UK)
- Joel Bender (Cornell University, Ithaca, NY, USA)
- Pierre Bourreau (Nobatek)
- Calin Boje (LIST, Luxembourg)
- Louise Dam (Semmtech)
- Meiling Shi (TU Darmstadt)

Presentation slides

- https://drive.google.com/file/d/1285mujIBPvsG5CJ1ISHdZGMaQTUH4FMb/view?usp=sharing

Date and time

- 05/10/2021, Tuesday, 15:00-16:30@UTC/ 17:00-18:30@CEST/ 08:00-09:30@PST

Moderator

1. Mathias Bonduel

Agenda

1. Introduction of new/returning members
2. Presentation by Meiling Shi (TU Darmstadt): “A literature review - Retrieve product knowledge graph from product description with natural language processing”
3. Q&A
4. Presentation by Ranjith Soman:”AEC Production control room- Role of linked data”
5. Q&A
6. Call for elevator pitches and announcement of next meeting

Minutes

1. Introduction of new/returning members
    a.
2. Presentation by Meiling Shi (TU Darmstadt): “A literature review - Retrieve product knowledge graph from product description with natural language processing”
    a. New in ontology engineering, but background with Anna Wagner - Product data modeling.
    b. Research of machine learning techniques for analyzing text
    c. Problem and motivation:
        i. Catalogs are text files > analyzing and finding the right product requires a large amount of time
        ii. Product platforms such as Buildup > not so intelligent, German only
            1. Text search also returns not wanted sub elements (e.g. wooden door returns also a doorbell)
            2. Not all relevant info shown
        iii. Manually converting product descriptions in ontologies takes also a large amount of effort > can this be improved using AI, with Natural Language Processing (NLP)
    d. Source data: unstructured data on a website (PDF, text, tables) and structure it in a knowledge graph to improve search results
    e. Related research fields:
        i. Ontology definition > reuse existing instead of developing new ones
        ii. AutoKnow:
            1. At Amazon
            2. Inputs for Machine Learning models
                a. implement taxonomy and product catalogs
                b. customer behaviour signals for supervision (e.g. did they click on the product on the Amazon website)
            3. Ontology suite: enrich taxonomies
            4. Data suite: instantiate product instances in triple form + data cleaning + synonym discovery
            5. Broad graph: combining the outputs
            6. Overview of the model in a diagram
            7. Taxonomy enrichment:
                a. Type extraction: Open-world tagging model BIOE (begin inside outside end) > useful is product name consists of multiple parts
                    i. Distant supervision to generate training labels
                b. Type attachment: adding names by hypernym classification
                    i. Graph neural network
            8. Relation discovery
                a. Classification model
                b. Regression model
                c. Approach: random forest (thus no deep learning)

9. Data imputation: BIOE combined with active learning + human in the loop: ask for feedback if unsure
    a. Got better results compared to SOTA, by using taxonomy-aware sequence tagging
10. Data cleaning: control if attribute-value are correct or incorrect for the product
    a. Transformer-based neural net model > suitable for scaling and good results (precise)
11. Synonym finding:
    a. Collaborative filtering on co-view behaviour
    b. Logistic regression
12. Datasets
    a. 1 bill product knowledge facts - 11k distinct product types

iii. Unsupervised construction of a product knowledge graph
1. Microsoft
2. Challenges: dynamic domain, not always clean data, no large sources for brands and products
3. Brands generation, domains, categorization and product identification
    a. Product identification is most difficult > focused on this in literature review
4. Inputs
    a. Retailer catalogs:
        i. Convolutional deep semantic similarity model + product clustering using K-Means with heuristics
        ii. Take semantic similarity into account during training
5. Methods:
    a. aggregation functions > uniform results
    b. Tag brands with domains and categories > similar to approach at Amazon (case 1)

iv. Building, Maintaining and using knowledge bases: A report from the trenches
1. Walmart
2. Use Wikipedia and convert into a knowledge base + enrich with other data sources
3. Building the knowledge base
    a. Ontology-like KB > manual input from a domain expert
        i. Constructing taxonomy tree from Wikipedia
            1. Delete cycles
            2. Extraction from info boxes
            3. Adding metadata: synonyms, etc
    b. Source-specific KB > a lot of data but no domain expert
4. Maintaining and curating KB

5. Using KB

v. Automatic MEP knowledge acquisition based on Documents and NLP
   1. Dataset:
      a. Industry spec, literature, encyclopedia > 70 mb thus small dataset
   2. Entity extraction
      a. named entity extraction (NER) > supervised learning, well-established technique
      b. Dictionary mapping to auto label text
      c. BI (also used in Amazon paper)
   3. Relations between entities
      a. similarTo, hasAttribute, instanceOf, include > residual
      b. Construct neo4'j KG

vi. conclusions
   1. Challenges in product knowledge graph
      a. Sparse
      b. Noise in structured data from product owner
      c. Increasing number of products, dynamic
   2. Creating a KG is a long process
   3. Big data and labelling

vii.
   1. Main techniques seen
      a. Graph neural network
      b. Weak supervision (distant supervision)
      c. Deep learning
   2. Research field of Meiling:
      a. Use more available data from Internet, weak supervised learning
      b. Store the results in graph

3. Q&A
   a. Q (Kevin Luwemba): seen several methods from the SOTA (e.g. GNN). Did you look into the older graph learning methods (knowledge graph embeddings, etc)?
      i. A: focused on the newer works. GNN also takes relations between nodes into account and good results were reported
      ii. Kevin: when injecting new nodes, do we need to retrain? How generalizable are the GNN approaches?
         1. A: not that deep in GNN; think you need to retrain them entirely or you might need to rewrite parts (only new nodes)
   b. Q (Mathias): In case you would use inductive reasoning on your training data, would this help with improving the results of the NLP process?
      i. A: have not looked into inductive reasoning in detail, but it was used in the research done by Amazon (AutoKnow)

c. Q (Mathias): In the Walmart case, the researchers extracted knowledge from Wikipedia: Why didn't they reuse Wikidata in combination with the natural text on wikipedia pages?
    i. A: the researchers wanted to have the entire wikipedia dataset
    ii. Mathias: might also have been something to do with the researchers not being used to knowledge graphs in Linked Data form?
d. Q (Alex Donkers): Are you aware of the BauDataWeb? It contains thousands of building data entries that could be valuable for training your model.
    i. A: not yet aware of this, but will consider it.
e. Q (Alex Donkers): Reusing existing ontologies (BPO, BOT) might be a good idea (at a later stage when generalizing the output of your work to be used for different building types for example)?
    i. Not planning to reuse BOT, but considering ontologies from HVAC domain in the future..
f. (Mathias): A bot:Element already indicates a building element. It wouldn't make sense to create for example subclasses of bot:Element with product classes, as that would mean that each product is also a building element (e.g. a door product can be part of a building, but it can also be part of a bridge which is not a building). You would always need a double classification in order to link it to product classification. You'll need to carefully consider how the outputs of the machine learning algorithms should be structured (which ontologies to use), as this will influence the uptake by e.g. a designer, engineer, contractor when they are using BOT, REC, Bricks, etc for describing their construction. This could be solved by providing different outputs based on the needs of the user. Recommendation to focus on compatibility with existing ontologies for describing buildings. The BPO from Anna Wagner could be considered.

4. Presentation by Ranjith Soman:"AEC Production control room- Role of linked data"
    a. The presentation **will be postponed** in order to provide sufficient time for presenting and Q&A.
5. Call for elevator pitches and announcement of next meeting
    a. Call for elevator pitches: 10-15 minutes at the beginning of a call, e.g. for Phd research or industry work.

Next Call

- 19/10/2021, Tuesday, 15:00-16:30@UTC/ 17:00-18:30@CEST/ 08:00-09:30@PST

  Agenda: TBD

  We are interested in getting suggestions from the community about potential agenda items and **Elevator Pitches** for the following calls. Please send your suggestions to the chairs or to internal-lbd@w3.org, whether you have a short presentation to bootstrap the discussion, and an approximate duration you think the discussion will last.

Previous minutes

https://www.w3.org/community/lbd/meeting-minutes/
https://github.com/w3c-lbd-cg/lbd/tree/gh-pages/minutes