A Literature review - Retrieve product knowledge graph from product description with natural language processing

Meiling Shi

Technical University of Darmstadt, Germany

Problem & Motivation

- Unstructured building product data
  - (PDF, textual descriptions & table on website...)
- Product search takes a long time and not intelligent
- Not all relevant information is displayed
- Manual converting product description to ontology instances is huge manual effort

- Searching results from https://ch.buildup.group/search

Goal

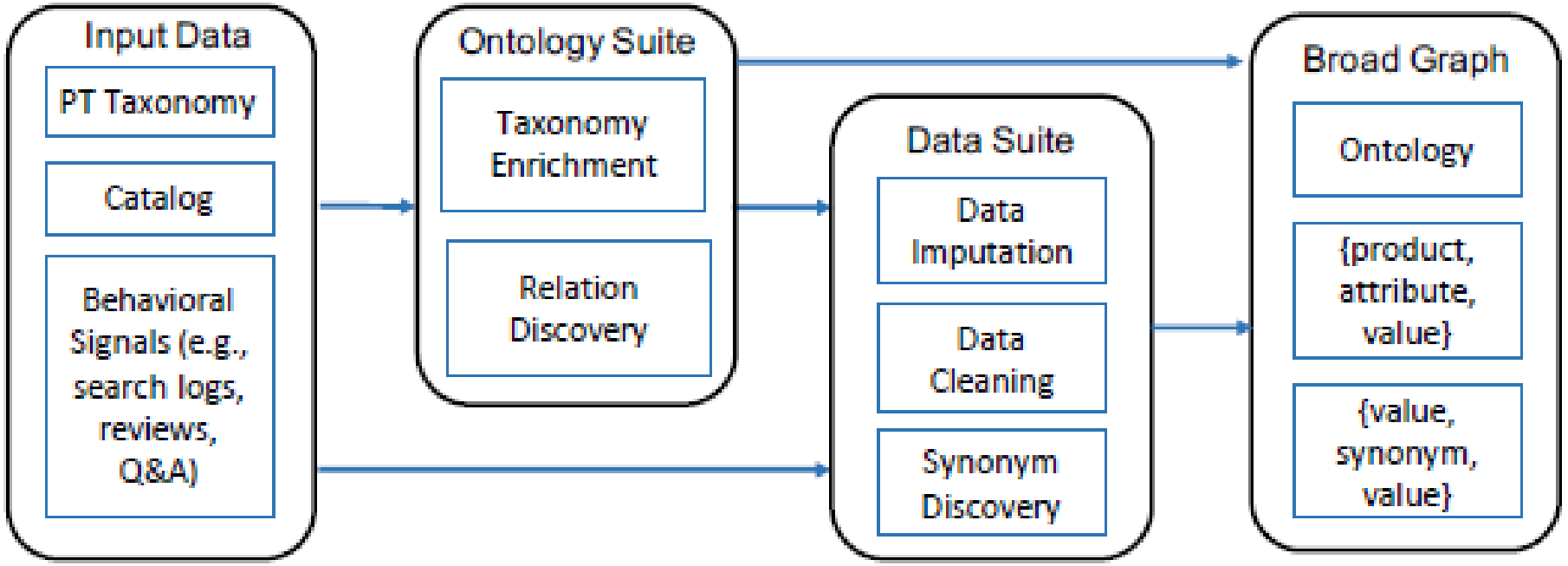
- Extract products and their attributes from unstructured data (data on website: PDF, text, table) and store it in a structured way with AI method.
  -> Significantly reduces manual work for creating a product Knowledge Graph
  -> Store Data in machine readable form
  -> improved search results, no rigid structure of data entry on web pages
  (-> Filter out products according to use case/project style/requirements

Related research fileds

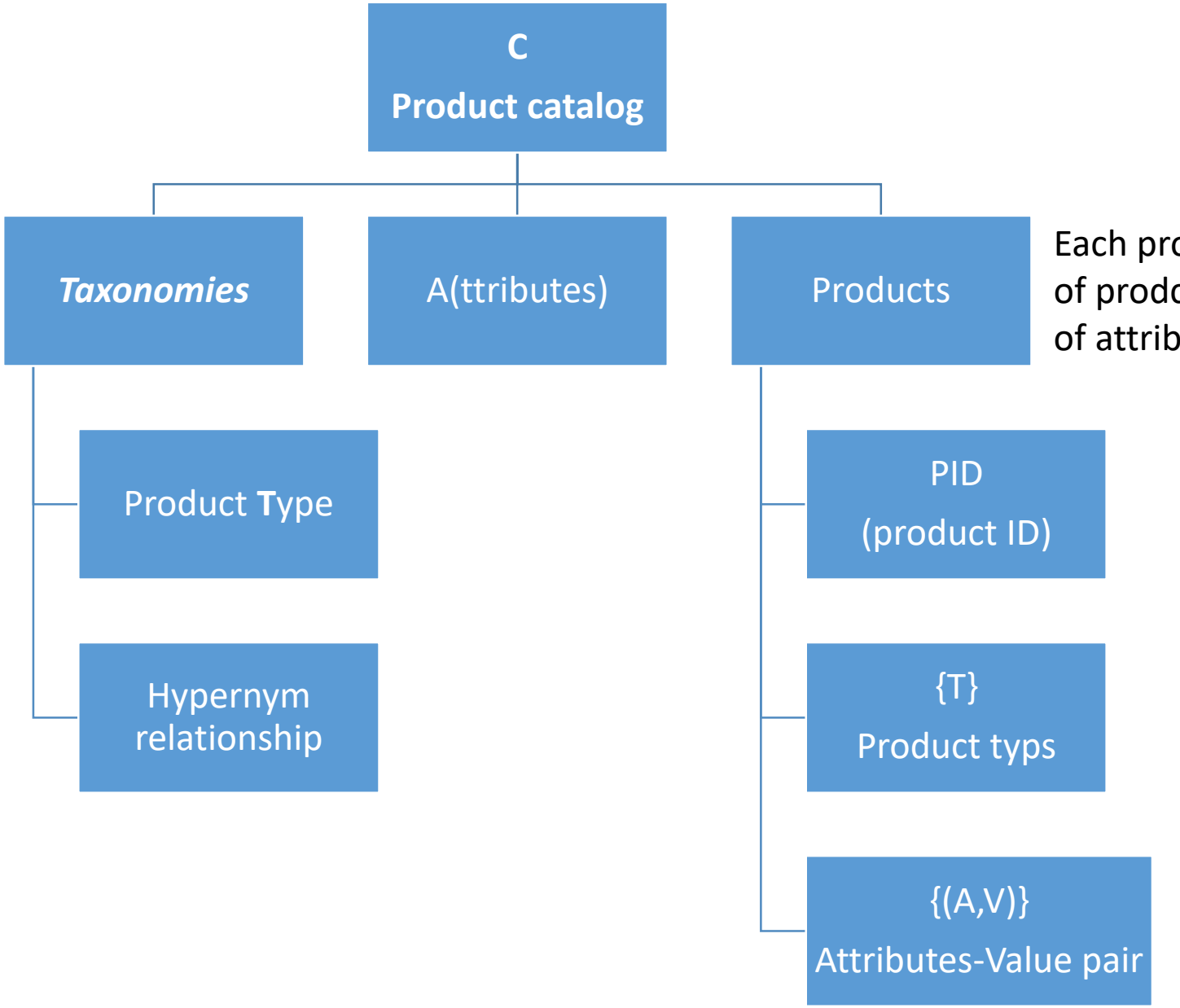
- Ontology definition
- **Entity identification**
- **Relation extraction**
- Linkage
- Knowledge fusion

1. AutoKnow: Self-Driving Knowledge Collection for Products of Thousands of Types -Amazon

- Incorporating taxonomy knowledge into machine learning models
- Ultilizing customer behaviour signals



Source:AutoKnow



Each product (PID) has a set of prodct typs and a set of attribut-value pairs

Input: C (product catalog) & L (costomer behaviour), enriches the product knowledge by adding new types and hypernym relations to T

Taxonomy enrichment

1. Type extraction
   - Discover products‘ types from titles and search queries
   - Open-world tagging model BIOE (begin inside outside end) with distant supervision to genrate training labels
2. Type attachment
   - Attach the discovered types to known taxonomies by solving hypernym classification problem
   - Graph Neural Network
   - Distant supervision

Relation Discovery

- Whether attribute *A* applies to product typ T (classification model)
- How important *A* is for purchase decisions (regression model)
- Technique: Random forest

Data imputation

- Extract new Atrribute-Value pairs for each product from product profiles
- State of the Art: BIOE sequential labeling (BiLSTM, CRE) combined with active learning
- Method of Amazon: Taxonomy-aware sequence tagging that makes predictions conditioned on the product type.
  - Conditional self attention layer, allows hierachical relations btw. Taxonomy nodes to influnce attention weights
  - Multi-Task learning, distant supervision to generate training sequence labels

Data cleaning

- Given product information, classifies incorrect Attribute-Value pair for the product
- Transformer-based neural network
  - Process product profiles and product taxonomy
  - via multi-head attention mechanism to
  - decide if a triple (Product, attribute, value) is correct
  - Ideal for scaling
- Distant supervision for generating tarining labels from input Catalog

Synonym finding

- Identify synonyms with the same sematic meaning (lite vs low sugar)
- Collaborative filtering on co-view behaviour->product pairs with high similarity –> take their A-V pairs as candidate
- Logistic regression to check if the candidate A-V pairs has the same meaning

Dataset & Techniques

- 1 billion product knowledge facts
- 11K distinct product types
- Techniques:
  - Graph neural network
  - Multi-task training
  - Weak supervision (distant supervision) & Semi-supervidsed learning
  - Transformer
  - Combine both facts and heterogeneous expressions

2. Unsupervised Construction of a Product Knowlwdge Graph - Microsoft

- Challenges:
  - No major sources that contain clean information about brands and products
  - Dynamic domain
  - Providers do not always provide clean data
- Process of genrating product graph
  - Brands generation
  - Brand domains generation
  - Brands Categorization
  - Products identification (most difficult part)

Product identification

- Products from retailer catalogs
  - Convolutional Deep Semantic Similarity Model (CDSSM) on product corpus -> product name in lower dimensions
  - Product clustering using K-Means with heuristics -> product hierachie
  - Gnerate a set of M representative labels for each cluster using generative algorithms, also using CDSSM encoding
- Products from bidded keywords

Methods

- Generating lists of brands from multiple sources using Aggregation functions
- Tag brands with domains and categories using query log mining and modeling
- Drive products using multiple modeling approaches on advertiser provided data

3. Building, Maintaining, and Using Knowledge Bases: A Report from the Trenches -Walmart

- Building KB
  - Convert Wikipedia into a KB
  - Then integrate the KB with additional data sources
- Maintaining KB
  - Rerunning the whole process from scratch
  - Only update changes
- Using KB

-> No concrete ML/NLP methods, but process & data processing

Special aspects of the paper

- Ontology-like KB
  - Data source acquisition problem
- Source-specific KB
  - Data integration problem from given data sources
  - Much easier if we haven ontology-like KB

Building the knowledge base

- Constructing taxonomy tree from Wikipedia
  - Edmonds‘ algorithm: find the max & min number of optimum branching in directed trees
- Construct DAG (directed acyclic graph) on top of T-box
  - Delete cycles in graph such as „category-subcategory“
- Extracting relationships from Wiki
  - Extract free-form relationship instances without pre-defined relations
  - Extraction from infobox
- Adding metadata
  - Adding synonyms, homonyms, web URLs, page traffic…
- Adding other data sources

Maintaining & Curating

- Maintaining:
  - Rerunning the KB construction pipeline from scratch
  - Performing incremental updates
- Curating the knowledge base
  - Manually evaluating graph quality
  - Correct false nodes and relations at scale

4. Automatic MEP Knowledge Acquisition based on Documents and NLP (2019)

- Dataset:
  - Industry specifications, research literature, encyclopedia, Q&A corpus, MEP component library (PDF, Word, HTML)
  - 70 mb
  - More than 14 million words

Entity Extraction

- Named Entity Recognition (NER) (supervised learning)-> 3 Categories of entities (Electic, HVAC, water supply)
- Dictionary mapping method which automatically labels text
- Manual screening process for deleting mistake entities
- -> 11332 MEP entities were extracted
- Bidirectional Long Short Term Memory network with Conditional Random Field (Bi-LSTM-CRF)

Entity Relationship Discovery

- Entity Relationship Discovery (similarTo, hasAttribute, instanceOf, include)
  - Sentences contains >=2 entities were selected
  - 900 sentences were manuelly labeled for training
  - Residual Convolutional Neural Network (ResCNN)
  - Syntactic analysis
  - Manual inspection process
  - -> 9439 relations were discovered
- Construct knowledge graph
  - Neo4j

My takeaways

- Problems of generating product knowledge graph compered to general KG
  - Data sparsity
  - Noise of structured data for products
  - Complexity of the domain with numerous products and attributes
  - Constantly growing number of products
- Building product knowledge graph is a long process including maintaining & curating
- Big dataset & labeling is necessary

Techniques

- Graph neural network
- Weak supervision (distant supervision)
- Deep learning: BiLSTM, CRE, Transformer