Ontology Mining

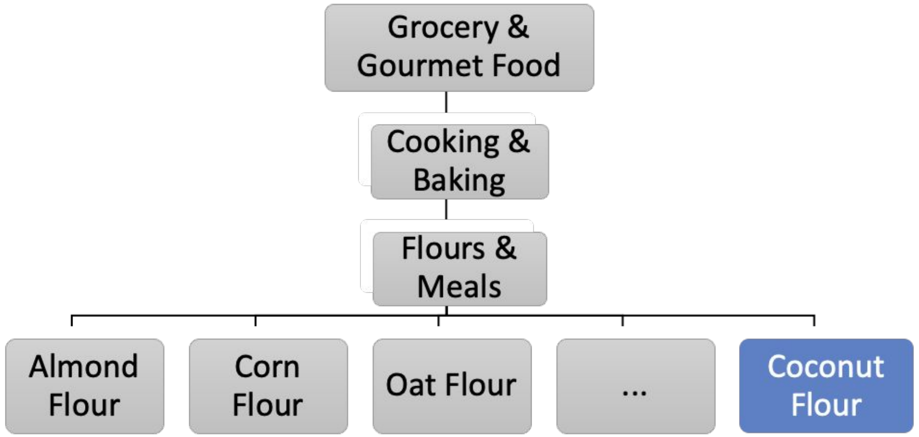
Categories (Sub-Categories), Attributes, and Values

Why Ontology Mining?

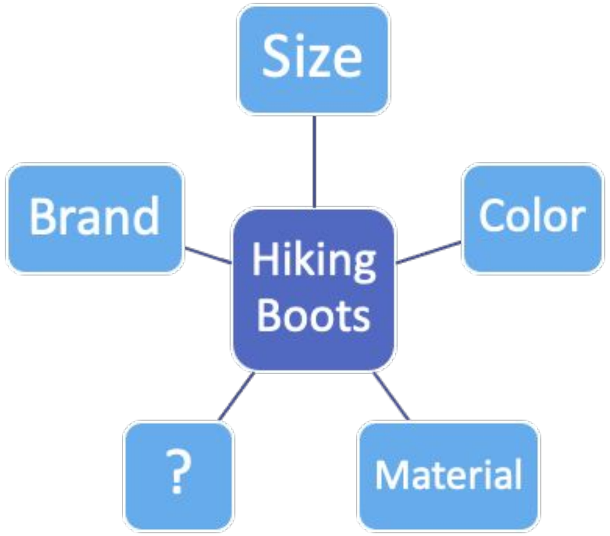
- Living in a world that can't be fixed...
  - Emerging Product Categories

Anthony's Organic Coconut Flour, 4lbs, Batch Tested Gluten Free, Non GMO, Vegan, Keto Friendly

  - Fine-grained Product Categorization



  - New product attributes to model



Such as Footwear Height (e.g. “ankle”)

Section Structure

- Problem Definition

  What is needed beyond techniques for building generic KGs?
- Short answer -- key intuition

  What are key intuitions for ontology mining?
- Long answer -- details

  What are practical tips?
- Reflection/short-answer

  Can we apply the techniques to other domains?

What is Product Ontology?

- Ontology defines...
  - The relationship between product categories -> Category Ontology
  - The relationship between product category and attribute -> Relation Ontology
- Questions to answer in this section

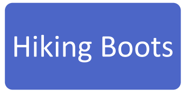
Category (Product Type) Ontology	Relation Ontology
“What is this product?” -> Product Type	“Can mug have a flavor?” -> Applicability
“What are the relationship between product types?” -> Type Hypernym/Synonyms	“Does size really matter, or is it only important for one product?” -> Importance
	“Are 2 ft and 24 inches the same length?” -> Value Variations/Synonyms

Techniques required

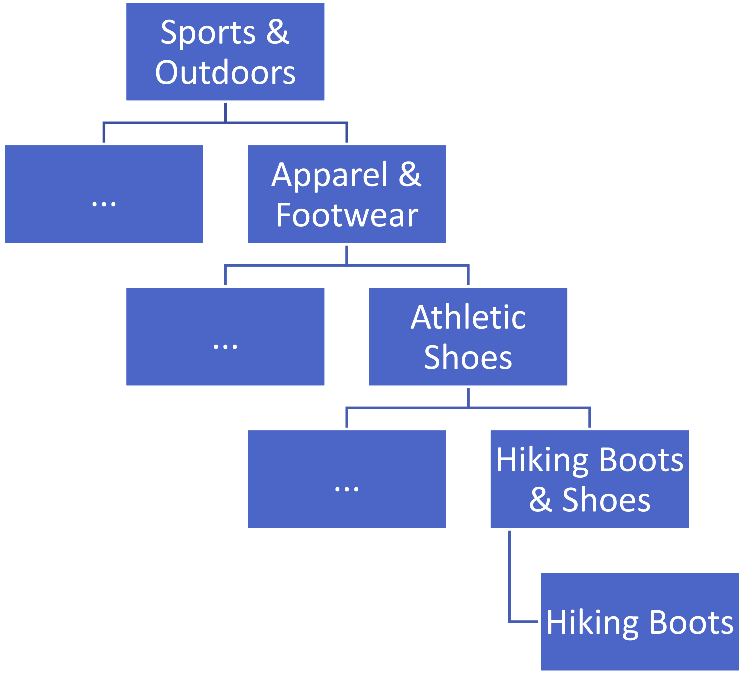
- Scarce direct supervisions
- Handling noisy labels
- Open-world learning

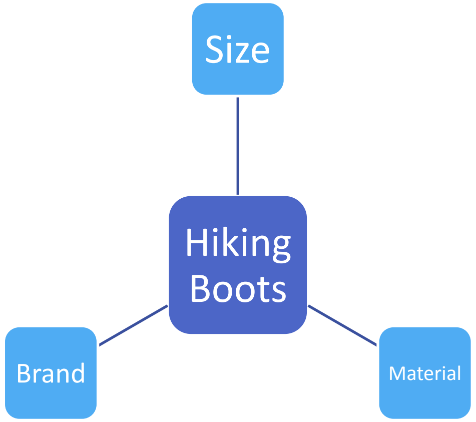
Ontology

- Categories & Sub-Categories
  - Category Name



  - Category Taxonomy





- Attributes
  - Facets/Asp

- Values
  - Possible va
    - E.g. for r
      fiberglas

Product Type Ontology

Product Type, Synonyms, Hypernyms, and Taxonomies

Product Type

- Product Type tells what a product is
  - E.g. “Coffee”:
    - Ground Coffee
    - Coffee Machine
    - Espresso Coffee Costume
    - Coffee-flavored Ice Cream



- A Product Type identifies a group of real-world products and defines their scope (i.e. "what the product is") based on visible and functional characteristics.
- Challenges
  - Emerging category, freshness is crucial
  - Lack of annotation, alleviate manual efforts

PT Extraction Methods – Closed-World

- Closed-word Learning Approaches
  - The product types are predefined
  - Formulated as a multi-class, multi-label classification problem
- Structured attribute values help define a product
  - Leveraging attribute information [Krishnan & Amarthaluri 2019]
- Product category label name possesses meaningful semantics
  - Leveraging label information [Meng+ 2020][Chen & Miyake 2021]

PT Extraction Methods – Open-World

- Open-world Learning Approaches
  - Product types not pre-defined
- Classification model with the capability to handle new classes
  - Classification-based with open-world learning [**Xu+ 2019**]
- Adopt advanced formulations (e.g. Seq-to-Seq model)
  - Generation-based [Li+ 2018] [**Umaashankar+ 2019**][Zahavy+ 2018] [Verma+ 2020]
- Utilize domain experts' knowledge
  - Active Learning [Zhu+ 2020]

PT Extraction Methods – Open-World

- Classification-based with open-world learning [Xu+ 2019]
  - Reject examples from unseen classes (not appeared in training)
  - Incrementally learn the new/unseen classes to expand the existing model.
- Maintains only a dynamic set of seen classes that allows new classes to be added or deleted with no need for model re-training.

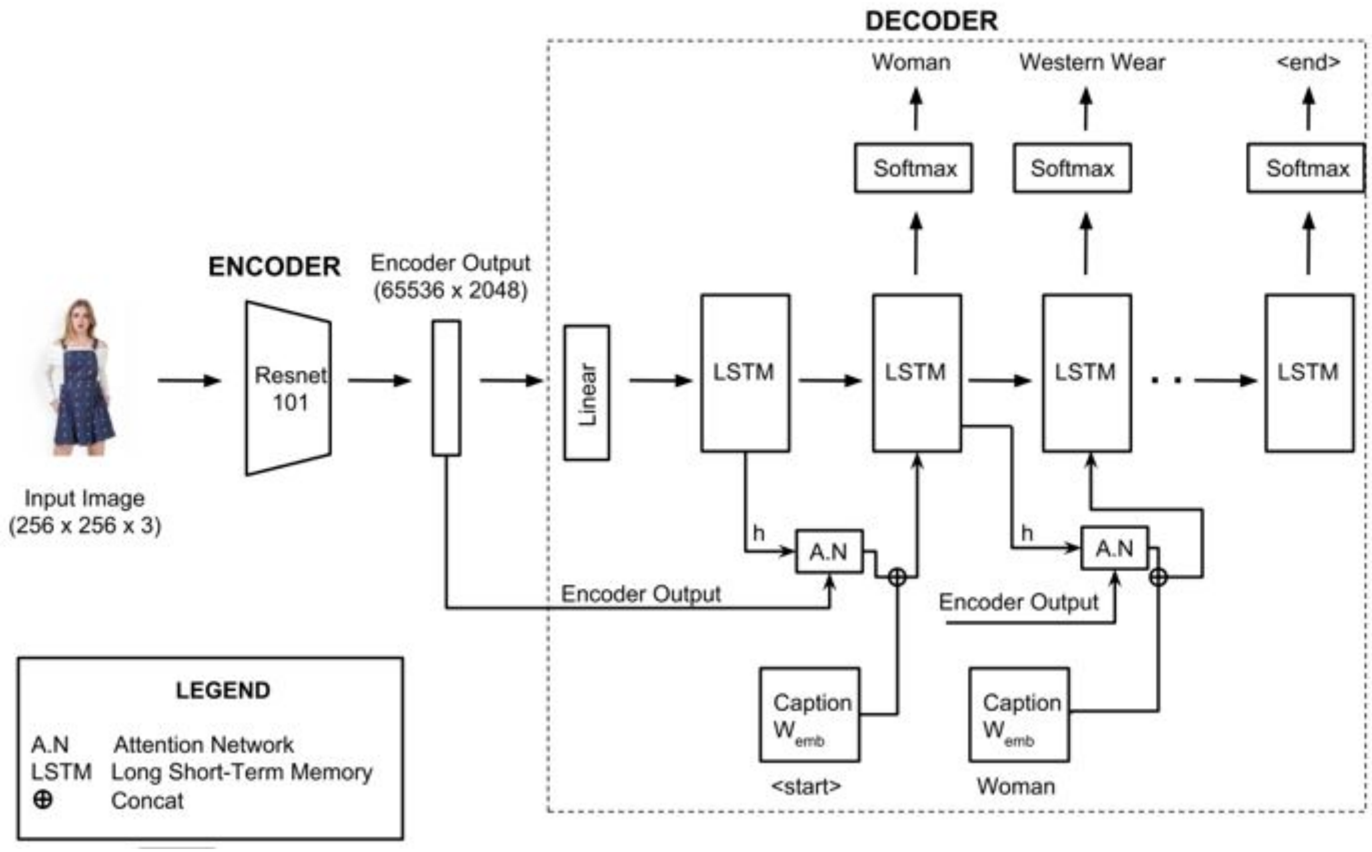
$$p(c|x_t, x_{a_{1:k}}) = \sigma\big(W \cdot \mathrm{BiLSTM}(r_{1:k}) + b\big).$$

$$\hat{y} = \begin{cases} \textit{reject}, \text{ if } \max_{c \in S} p(c|x_t, x_{a_{1:k}}) \leq 0.5; \\ \arg\max_{c \in S} p(c|x_t, x_{a_{1:k}}), \text{ otherwise.} \end{cases}$$

- Train a meta-classifier that uses the examples from seen classes (including the newly added classes) on-the-fly for classification and rejection

PT Extraction Methods – Open-World

- Generation-based multi-modal model [Umaashankar+ 2019]



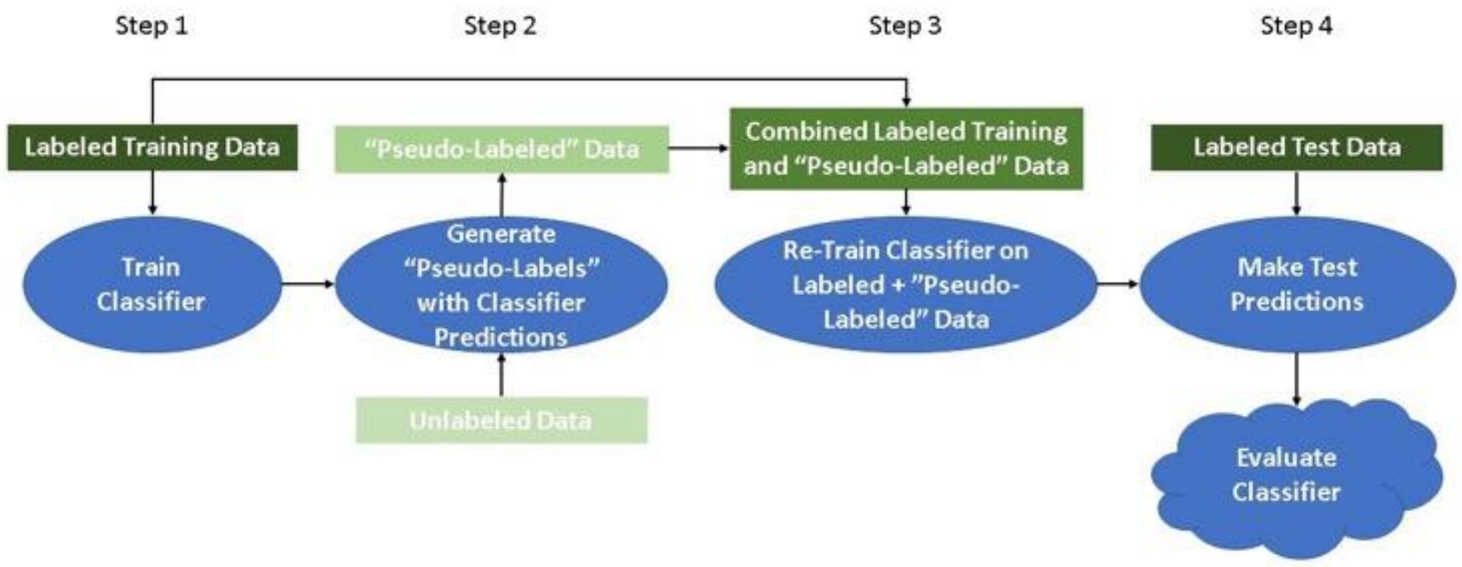
Valid Category Paths
Women >Western Wear >Jackets
Women >Western Wear > Blazers&Suits

PT Extraction

- Dealing with scarce annotations
- Abundant unlabeled data is available
  - Semi-supervised Self-training approaches [**Liu+ 2018**]
- Different modeling has unique inductive biases, thus good to combine together
  - Co-training with text and graph information [**Zhang+ 2021**]
- Augment existing data when unlabeled data is not available
  - Data augmentation methods [Du+ 2021]
- Improve model generalization so it handles unseen classes
  - Zero-shot learning [Ye+ 2020]

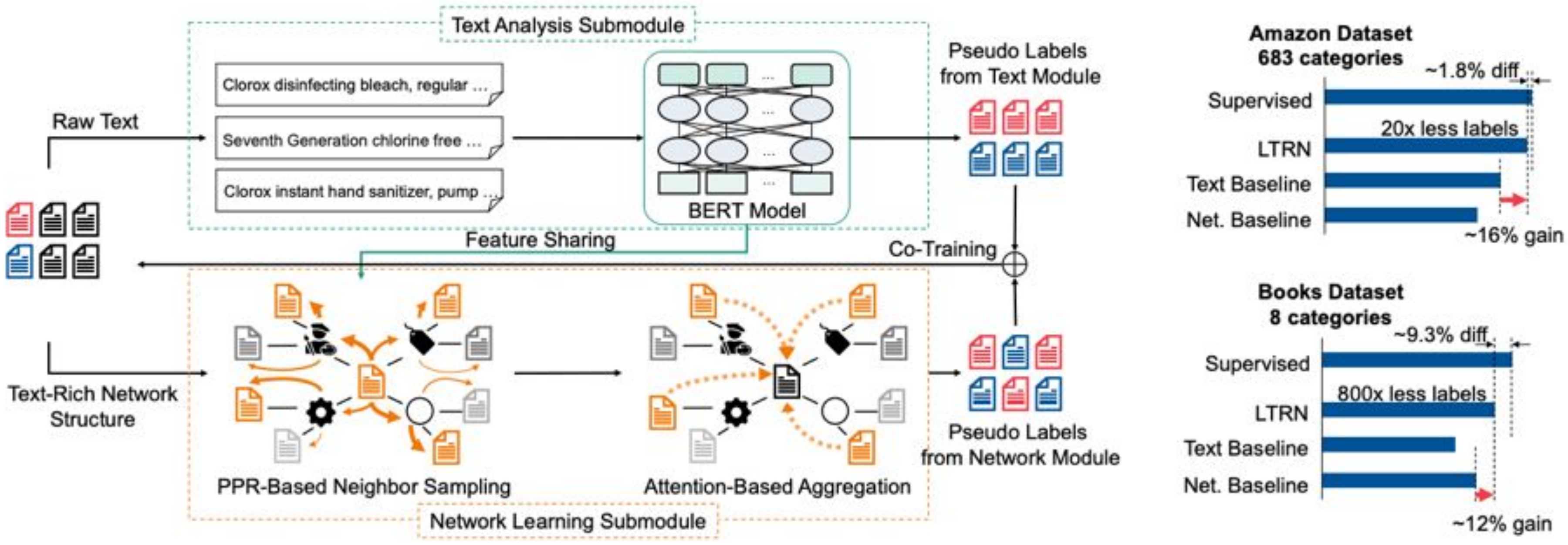
Self-training based methods

- Used the trained model to generate *pseudo labels* on unlabeled data



Co-training based methods

- Co-training with text and networked-text (graph) information [Zhang+ 2021]



Resources

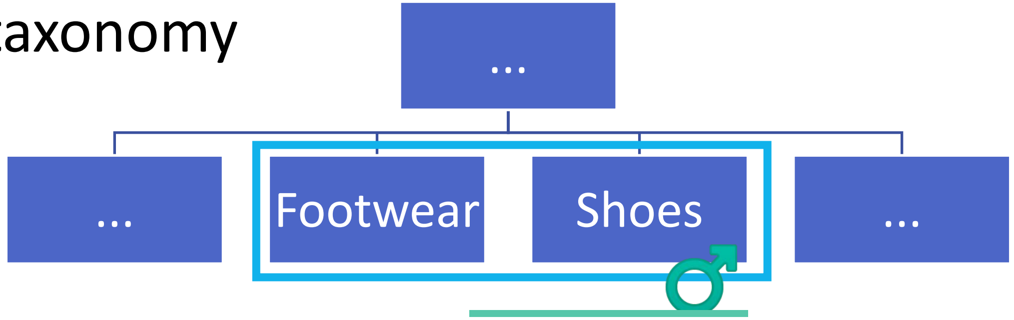
- Rakuten SIGIR2018 Dataset
- icecat
- WDC Product Data Corpus and Gold Standard for Large-Scale Product Matching - Version 2.0
- WDC-25 Gold Standard for Product Categorization
- WDC-222 Gold Standard for Hierarchical Product Categorization
- Amazon
- Amazon Berkeley Objects (ABO) Dataset

Product Taxonomy

- We now have Product Type terms extracted/classified from new products. Why it is important to have a product taxonomy?
  - Organize emerging Product Types
  - Capture relations between Product Types
- Challenges
  - Taxonomy not incomplete
  - Manual curation for thousands of PTs
- Taxonomy Enrichment
  - Synonym
  - Hypernym

Type Synonym

- Why?
  - A duplication-free taxonomy



  - Improve downstream applications, e.g. recall in product search

- [Fei+ 2019] [Boteanu+ 2019]

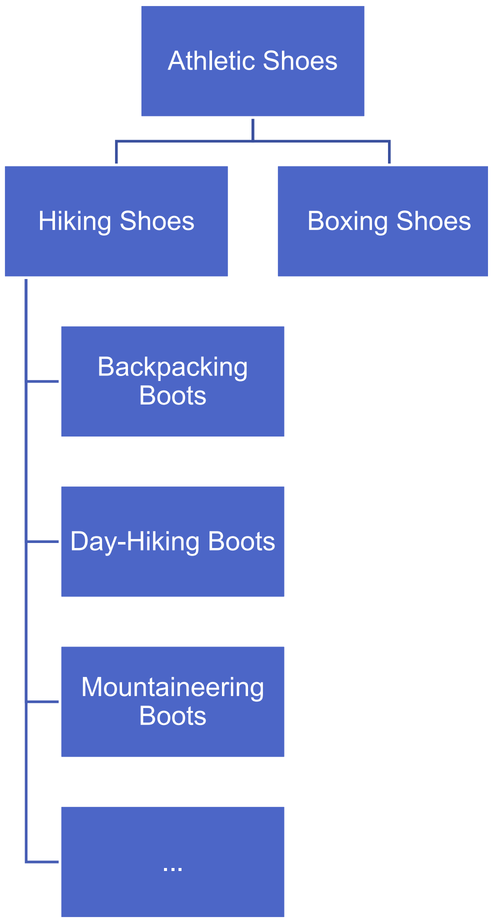
Type Synonym

- Leveraging local and structure features to train a synonym classifier [Boteanu+ 2019]
- Local features:
  - Word frequency in search queries
  - Character and word edit distance
  - Cosine similarity
- Structure features:
  - The node's parent name
  - The name of the taxonomy root
  - The average distance to the node's direct children

	Local Features			Structural Features			Metrics		
Row	WF	$edit(SC, N)$	$d(SC, N)$	$d(SC, P)$	$\bar{d}(SC, ci)$	$d(SC, R)$	P	R	F-1
1	x	x	x	x	x	x	0.92	0.79	0.85
2	x	x	x				0.87	0.73	0.79
3				x	x	x	0.90	0.41	0.56
4	x	x	x	x	x		0.94	0.90	0.92
5	x	x	x	x		x	0.94	0.90	0.92
6	x	x	x		x	x	0.94	0.88	0.91
7	x	x		x	x	x	0.95	0.67	0.79
8	x		x	x	x	x	0.94	0.89	0.92
9		x	x	x	x	x	0.94	0.89	0.91
10	x	x	x		x		0.95	0.87	0.91
11			x				0.87	0.70	0.77
12	x						1.0	0.07	0.13

Type Hypernym

- Why?
  - Capture the *is-a* relation between products
  - Help navigate customers thru the shopping funnel
- Different Settings



Hypernym Detection Methods

- Leveraging lexical patterns such as "Hiking shoes ***is a type of*** athletic shoes".
  - Pattern-based [Hearst 1992][Jurgens & Pilehvar 2015][Roller+ 2018]
- Leveraging distributed semantic embeddings
  - Distributional models [Wang+ 2014] [Yamane+ 2016][Espinosa-Anke+ 2016] [Nguyen+ 2017][Chang+ 2018][Le+ 2019][Mao+ 2020][Manzoor+ 2020]
- Leveraging user search behaviors
  - Search query based [Liu+ 2012] [Shalom+ 2019]

Hypernym Detection Methods

- Pattern-based [Roller+ 2018]
  - /[NP] *such as* [NP] (and [NP])?/
  - animals *such as* cats and dogs
  - Cats, dogs*, and other* animals
- Provide high-quality and robust predictions on large corpora by capturing important contextual constraints, which are not yet modeled in distributional methods.

Pattern
X which is a (example|class|kind|...) of Y
X (and|or) (any|some) other Y
X which is called Y
X is JJS (most)? Y
X a special case of Y
X is an Y that
X is a !(member|part|given) Y
!(features|properties) Y such as X_1, X_2, ...
(Unlike|like) (most|all|any|other) Y, X
Y including X_1, X_2, ...

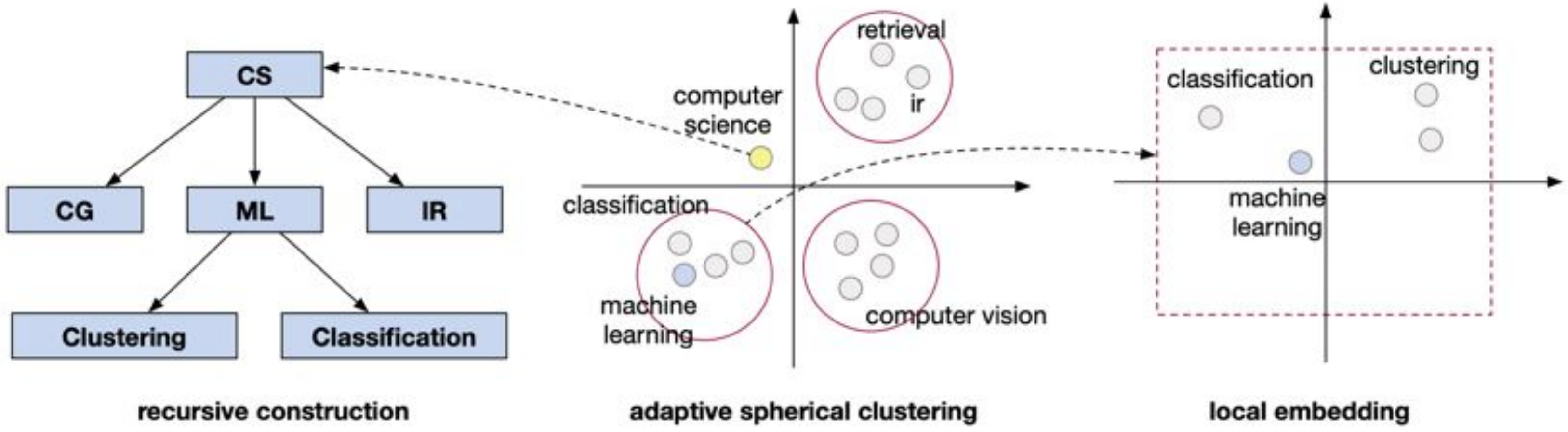
- Advantage: leveraging co-occurrence information, robust, high precision
- Disadvantage: may not able to find match in e-commerce corpus, low recall
  - Hard to collect corpus containing "athletic shoe *including* hiking boots"

Hypernym Detection Methods

- Embedding-based distributional models
- Unsupervised methods mostly based on *Distributional Inclusion Hypothesis* [Weeds+ 2004] [Geffet & Dagan 2005] [Kotlerman+ 2010] [Santus+ 2014] [Lenci and Benotto 2012] [Shwartz+ 2017]
  - We assume that more general words like “animal” appear in a variety of different contexts, while more specific words like “cat” appear in a few specific contexts.
  - When the contexts of “animal” include all the contexts of “cat”, we can assume that “animal” is a hypernym of “cat”.

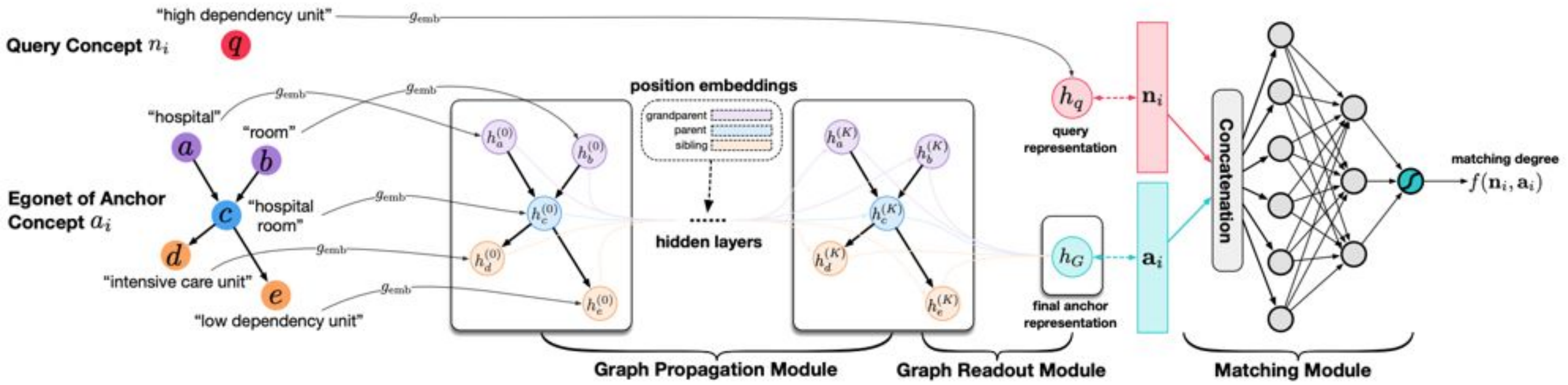
Hypernym Detection Methods

- Embedding-based distributional models
- Unsupervised methods mostly based on *Distributional Inclusion Hypothesis* [Weeds+ 2004] [Geffet & Dagan 2005] [Kotlerman+ 2010] [Santus+ 2014] [Lenci and Benotto 2012] [Shwartz+ 2017]
- Clustering based methods [Zhang+ 2018]



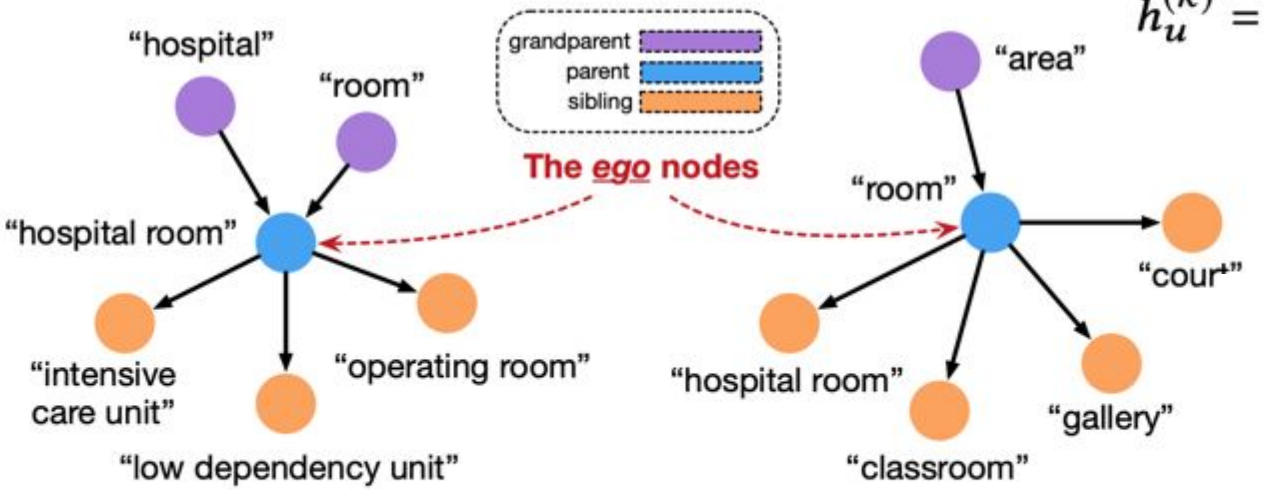
Hypernym Detection Methods

- Distributional models
- Supervised methods leverage various signals to improve embeddings
- Graph Neural Network models
- [Shen+ 2020][Zeng+ 2021]



Hypernym Detection Methods

- Distributional models
- Supervised methods leverage various signals to improve embeddings
- Graph Neural Network models [Shen+ 2020][Zeng+ 2021]



A graph propagation module

$$h_u^{(k)} = \mathrm{AGG}^{(k)}\left(\{h_v^{(k-1)} | v \in \widetilde{N(u)}\}\right), \quad k \in \{1, \ldots, K\},$$

A graph readout module

$$\mathrm{READOUT}(\{h_u^{(K)} | u \in G\}) = \sum_{u \in G} \frac{\log(1+\exp(\alpha_{p_u}))}{\sum_{u' \in G} \log(1+\exp(\alpha_{p'_u}))} h_u^{(K)},$$

Resources

- SemEval-2015 Task 17: Taxonomy Extraction Evaluation
- SemEval-2016 Task 13: a Taxonomy Induction Method based on Lexico-Syntactic Patterns, Substrings and Focused Crawling
- SemEval-2016 Task 14: Semantic Taxonomy Enrichment
- SemEval-2018 Task 9: Hypernym Discovery

Overall workflow

- Data preprocessing
- Taxonomy Construction
  - Extraction: PT Phrase Extraction/Classification
  - Deduplication: Product type similarity calculation
  - Induction: clustering, attachment
- Mapping product to categorization labels

Product Type ontology workflow

Data preprocessing

Taxonomy Enrichment

Segmentation & Labeling

Building Types

Building Type Relations

Induction

Term Extraction/ Classification

Deduplication

Classification for term relations

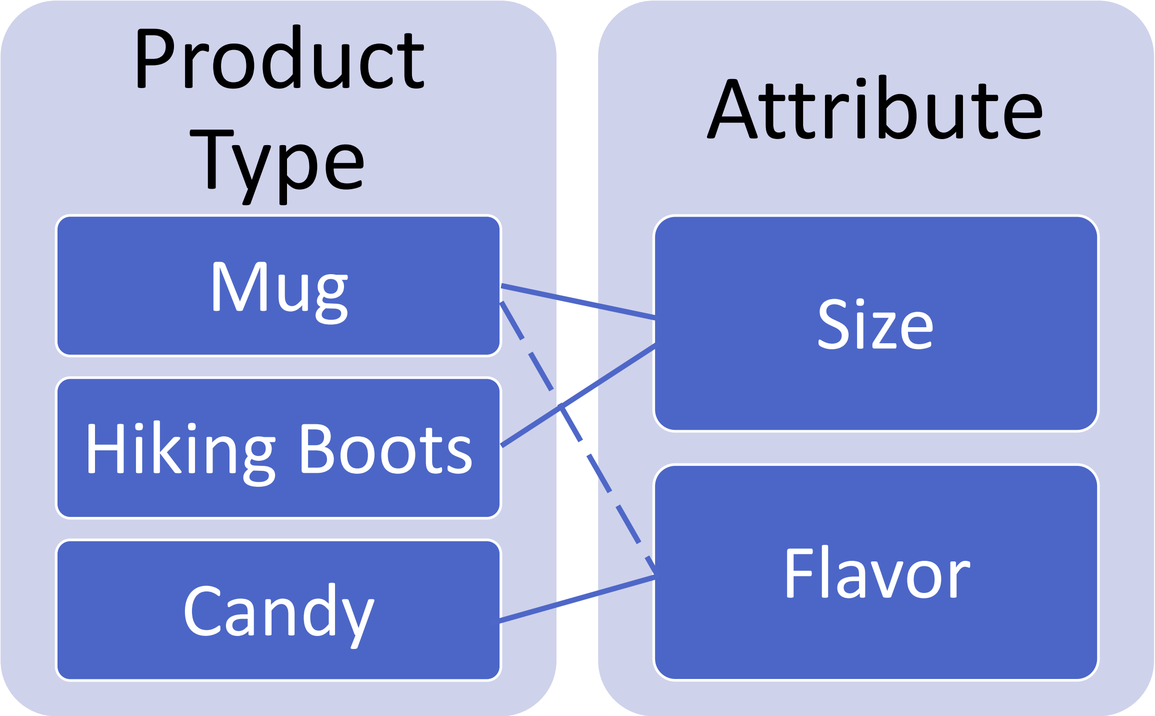
Clustering/Attachment

Relation Ontology

Applicability, Importance, Variations/Synonyms

Attribute Applicability

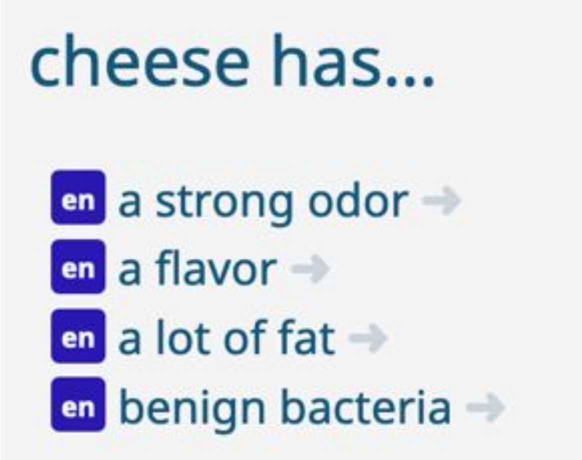
- "Can mug have a flavor?"



- Why?
  - Discover a new aspect of a product
  - Understand attribute applicability for downstream applications
  - Regularize attribute value extraction results

Attribute Applicability prediction methods

- Applicability Prediction [Rukat+ 2017]
- Aspect Extraction [Ramezani+ 2020] [Tian+ 2020]
- Commonsense Knowledge [Luo+ 2020]
  - ConceptNet [Speer+ 2017]

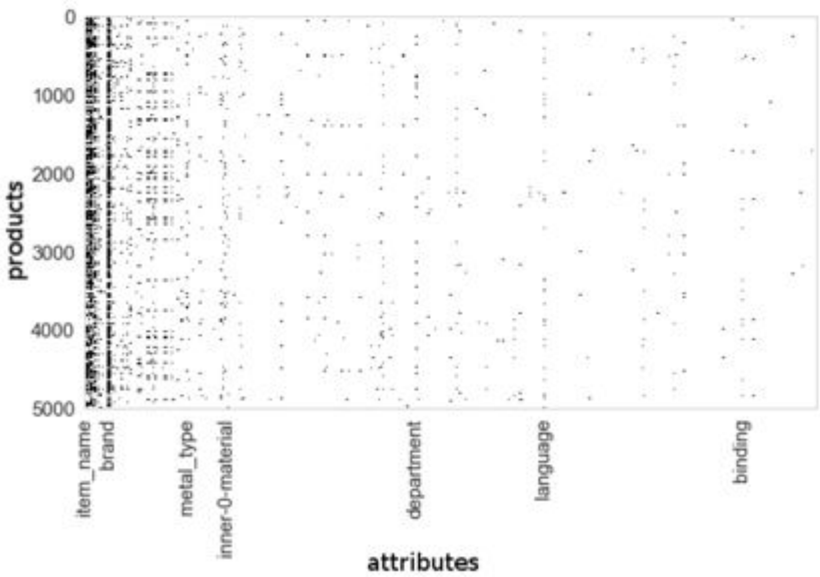


Attribute Applicability

- Applicability Prediction methods [Rukat+ 2017]
- As a multi-label classification problem
- Works on a binary matrix between products and attributes
- Binary matrix factorization

$$\begin{array}{c} A\left\{\begin{array}{c} \\ \\ \\ \\ \end{array}\right. \\ B\left\{\begin{array}{c} \\ \\ \\ \\ \end{array}\right. \end{array} \begin{pmatrix} 1&1&1&1&1&1&0&0&0\\ 0&0&1&0&1&1&0&0&0\\ 1&1&0&1&0&1&0&0&0\\ 1&1&1&1&1&1&0&0&0\\ 0&0&0&0&0&0&1&1&0\\ 1&1&0&0&0&1&1&1&1\\ 1&1&0&0&0&1&0&0&1\\ 0&0&0&0&0&0&1&1&0 \end{pmatrix} \approx \begin{pmatrix} 1&1&0\\ 0&1&0\\ 1&0&0\\ 1&1&0\\ 0&0&1\\ 1&0&1\\ 1&0&0\\ 0&0&1 \end{pmatrix} \cdot \begin{pmatrix} 1&1&0&0&0&1&0&0&0\\ 0&0&1&0&1&1&0&0&0\\ 0&0&0&0&0&0&1&1&0 \end{pmatrix}$$

Attribute	cup-size	closure-type	leather-type
Product types with largest p(apply)	Bra 22(10)% Swimwear 3(2)% Underwear 3(2)% Shoes 2(2)% Suit 2(2)%	Shoes 48(18)% Pants 24(10)% Shorts 6(3)% Outerwear 4(2)% Bra 2(2)%	Shoes 48(15)% Outerwear 3(3)% Shorts 2(2)% (< 1%)

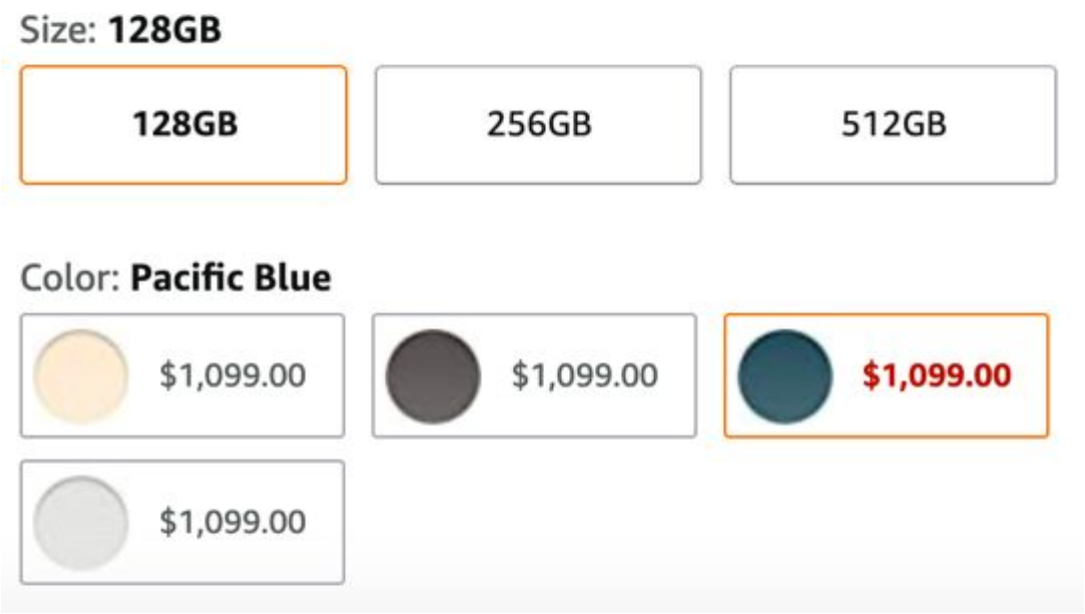


Attribute Value Variations/Synonyms

- “2 ft” vs. “24 inches”?
  - DALS Lighting 6000-ACCE24 **24"** Extension Cord for PowerLED Linear, Black
  - FIRMERST 1875W Flat Plug Extension Cord White **2ft** UL Listed (15A 14AWG)
  - Value Synonyms [Shinzato & Sekine 2013]
- “64Gb” vs. “256Gb”?
  - Apple iPhone 12, **64G**B, White - Fully Unlocked (Renewed)
  - Apple iPhone 12 Pro, **128GB**, Graphite - Fully Unlocked (Renewed)
  - Value Variations [Embar+ 2020]

Value Variations

- Variations are the options that a customer can choose from when purchasing a product.
  - Size: 128GB, 256GB, 512GB
  - Color: Gold, Graphite, Pacific Blue, Silver



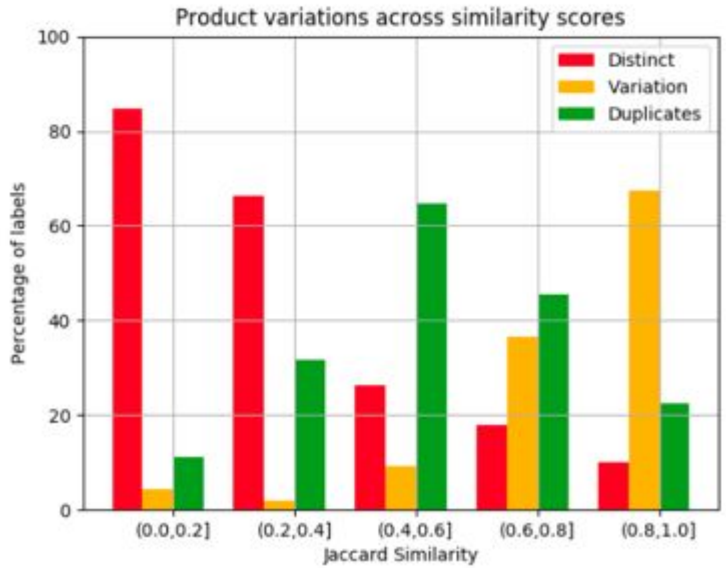
Value Variations

- Product with similar titles are likely variations
- Utilizing contrast features as value variation candidates [Embar+ 2020]

Apple iPhone 11 Pro **64 GB**

Apple iPhone 11 Pro **256 GB**

- Improve product duplication performance



Magellan

Software		Without contrast features	CEL
Duplicates	F1	0.785	0.81
	APS	0.877	0.897
Variations	F1	0.677	0.695
	APS	0.761	0.777

Value Synonyms

- Feature-based Matching [Elmagarmid+ 2007]
  - Edit distance, Affine Gap Distance, Smith-Waterman Distance, Jaro Distance Metric, and Q-Gram Distance
- DL-based Matching [Mudgal+ 2018]

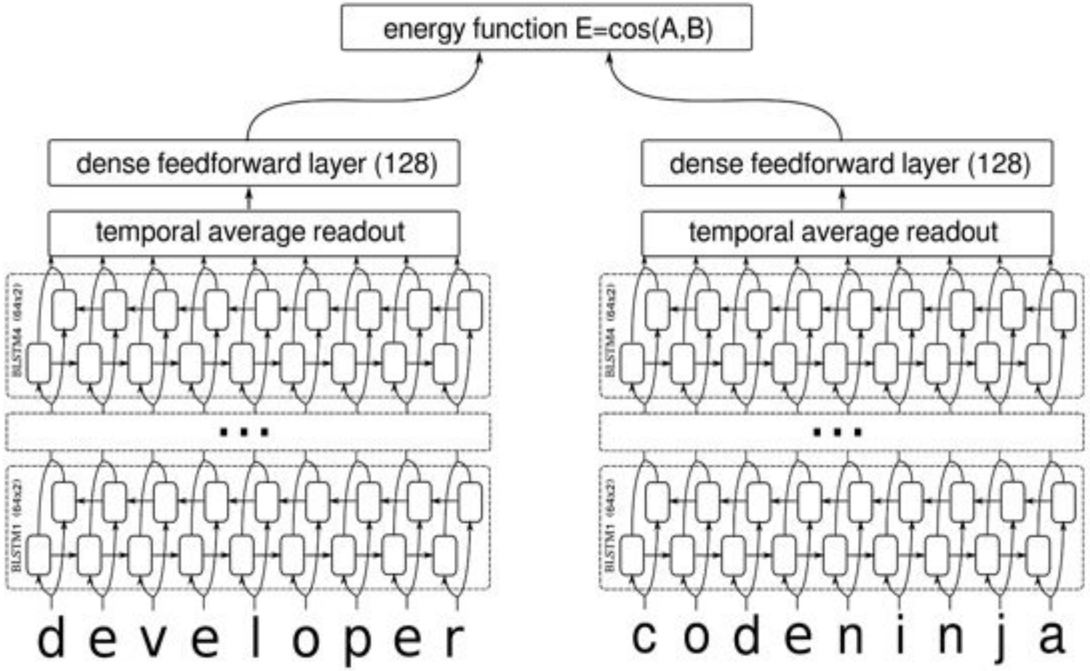
Architecture module		**Options**	
Attribute embedding		*Granularity:* (1) Word-based (2) Character-based	*Training:* (3) Pre-trained (4) Learned
Attribute similarity representation	(1) Attribute summarization	(1) Heuristic-based (2) RNN-based (3) Attention-based (4) Hybrid	
	(2) Attribute comparison	(1) Fixed distance (cosine, Euclidean) (2) Learnable distance (concatenation, element-wise absolute difference, element-wise multiplication)	
Classifier		NN (multi-layer perceptron)	

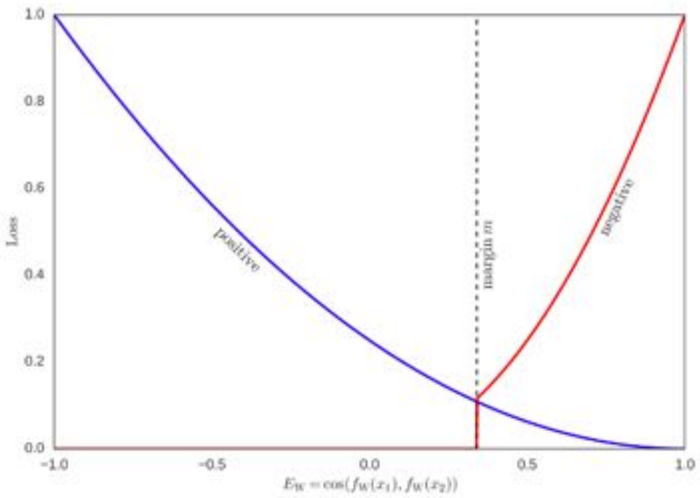
DL-based Matching

- Siamese network for learning good entity embeddings in vector space using contrastive loss. [Neculoiu+ 2016]

$$L_+(x_1, x_2) = \frac{1}{4}(1 - E_W)^2$$

$$L_-(x_1, x_2) = \begin{cases} E_W^2 & \text{if } E_W < m \\ 0 & \text{otherwise} \end{cases}$$



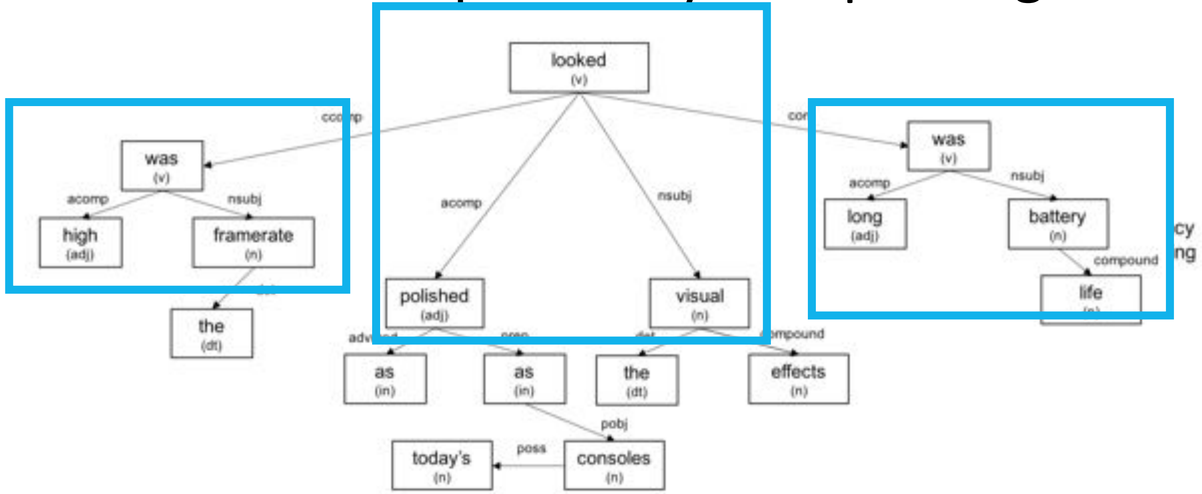


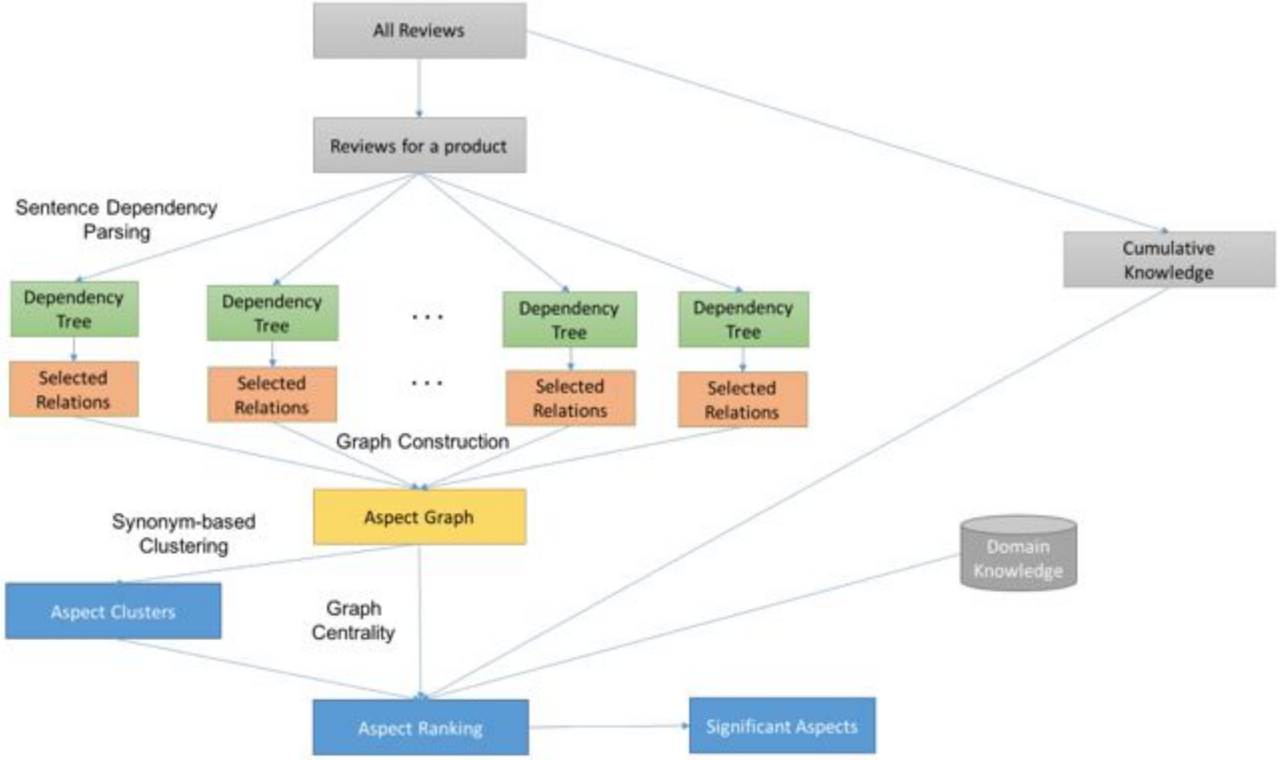
Attribute Importance

- “Does size really matter, or is it only important for one product?”
- From seller metadata
  - Manufacturer-provided attributes associated with product catalog
- From customer reviews
  - E.g. for an iPhone “old body style with a larger screen! ”
  - Rule-based [Popescu & Etzioni 2005]
  - Graph-based ranking methods: using PageRank [Yan+ 2015] [Indrakanti & Singh 2018]
  - Fine-grained LDA with keywords [Wang+ 2014]
  - Opinion mining with Fuzzy-c-means clustering [Zimmermann+ 2016]
- From customer search query log [Pound+ 2011], and behavior data [Zhou+ 2020]

Attribute Importance

- Aspect Extraction from reviews
  - Aspect extraction
    via Dependency tree pruning





  - Aspect graph construction
  - Aspect ranking

Resources

- Web Data Commons - Gold Standard for Product Matching and Product Feature Extraction