Part II: Multi-faceted Taxonomy Construction

KDD 2020 Tutorial

Embedding-Driven Multi-Dimensional Topic Mining and Text Analysis

Yu Meng, Jiaxin Huang, Jiawei Han

Computer Science, University of Illinois at Urbana-Champaign

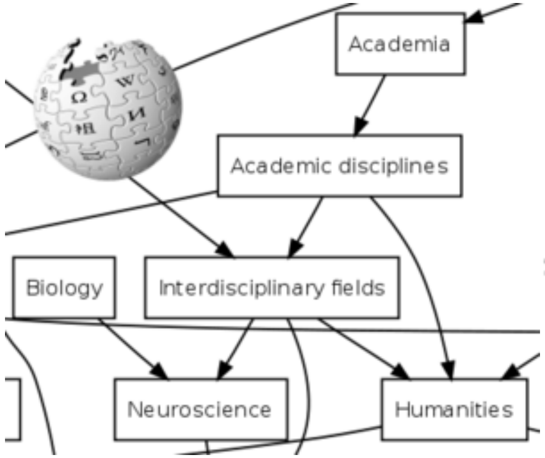
August 23, 2020

Outline

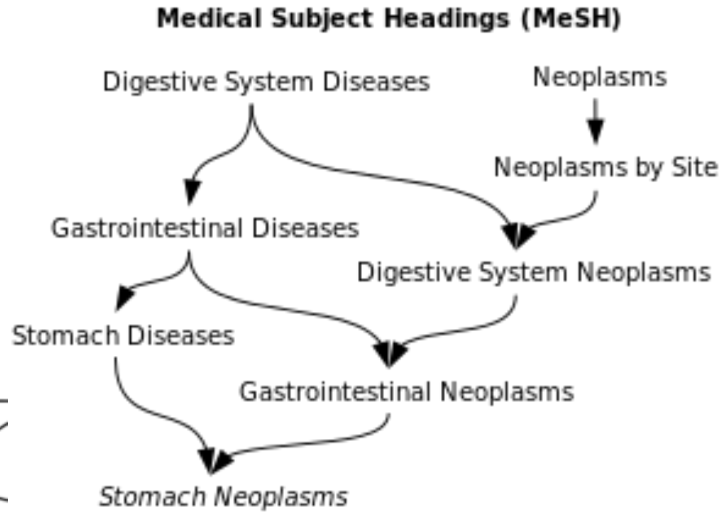
- Taxonomy Basics and Construction
  - What is taxonomy and why use taxonomy?
- Parallel Concept Discovery: Entity Set Expansion
- Taxonomy Construction from Scratch
- Taxonomy Expansion

What is a Taxonomy?

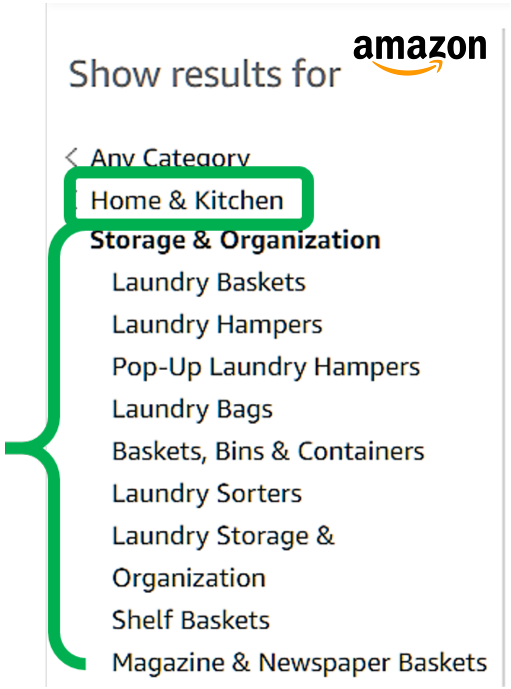
- Taxonomy is a hierarchical organization of concepts
  - For example: Wikipedia category, ACM CCS Classification System, Medical Subject Heading (MeSH), Amazon Product Category, Yelp Category List, WordNet, and etc.



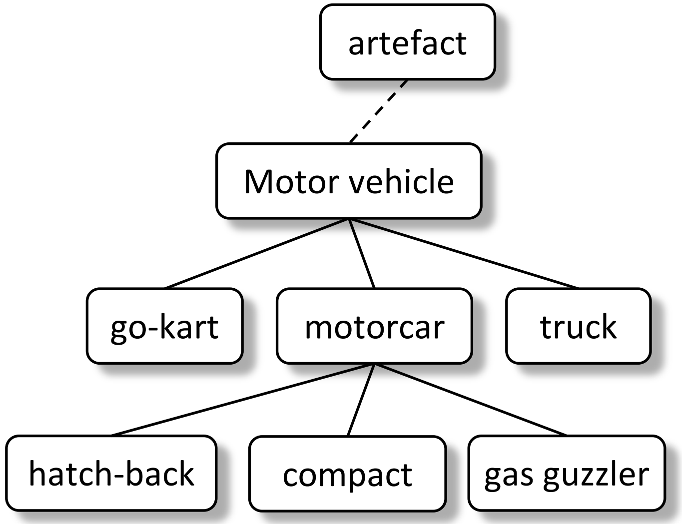
Wikipedia Category



MeSH



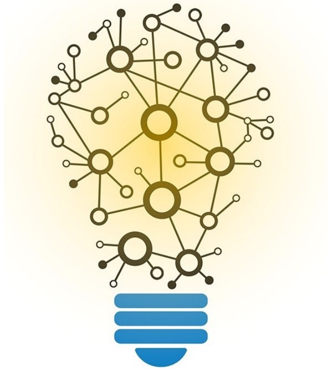
Amazon Product Category

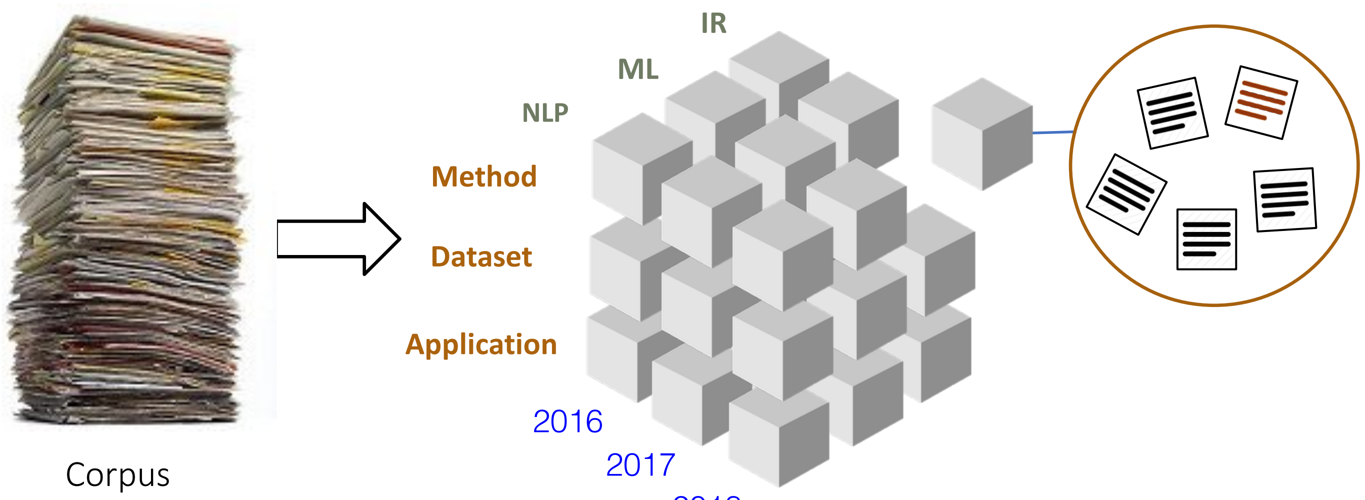


WordNet

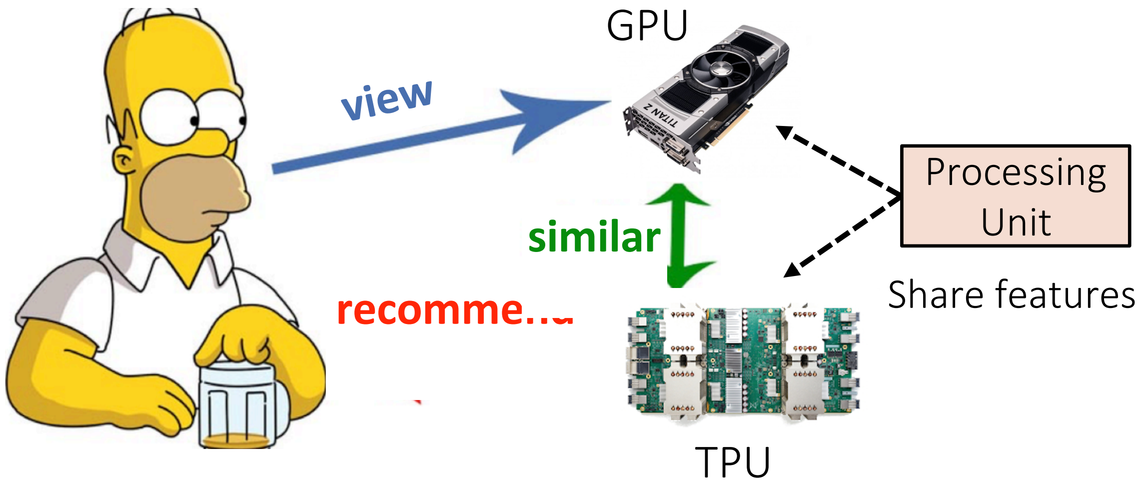
Why do we need a Taxonomy?

- Taxonomy can benefit many knowledge-rich applications
  - Question Answering
  - Knowledge Organization
  - Document Categorization
  - Recommender System



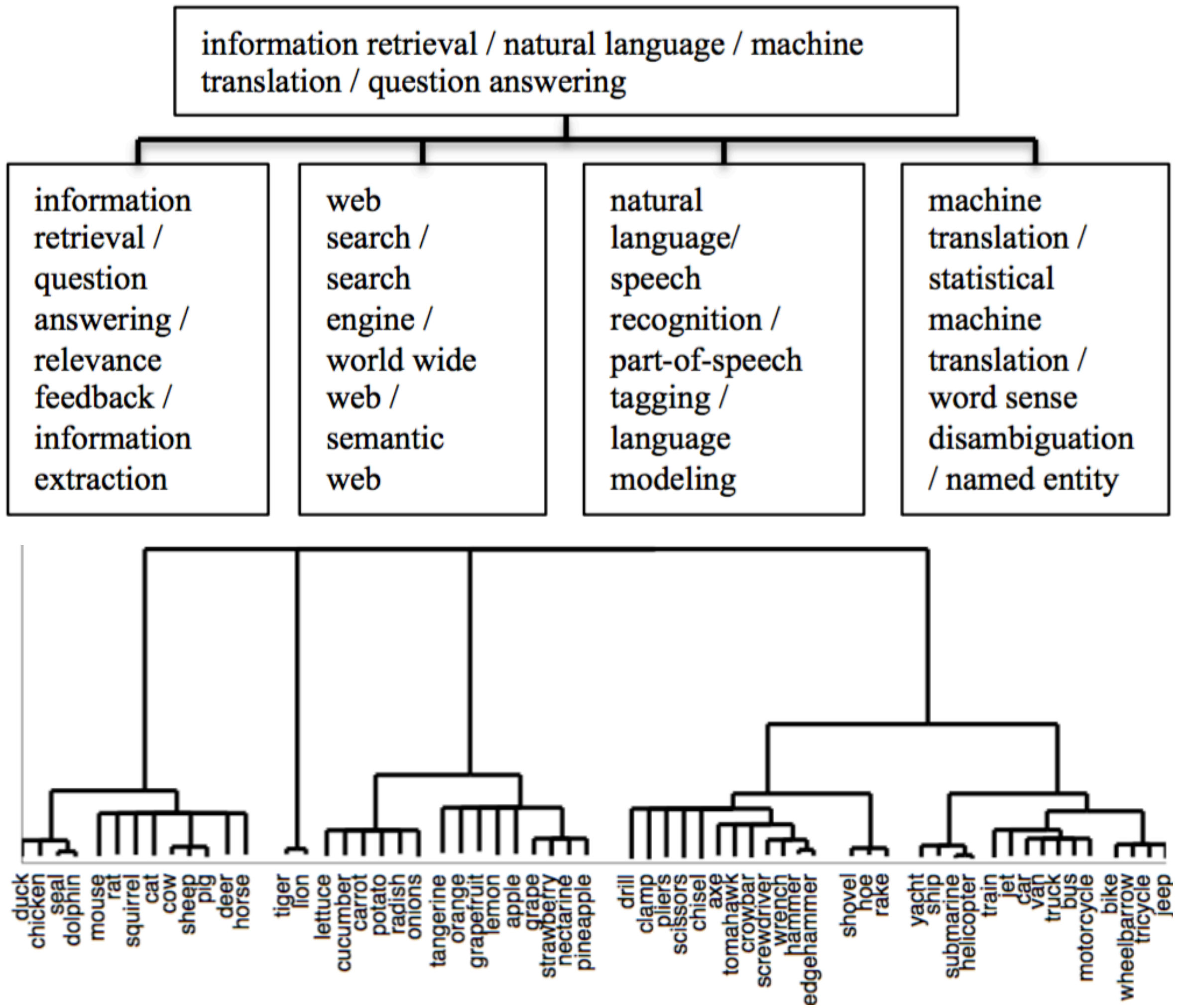


Multi-dimensional Corpus Index

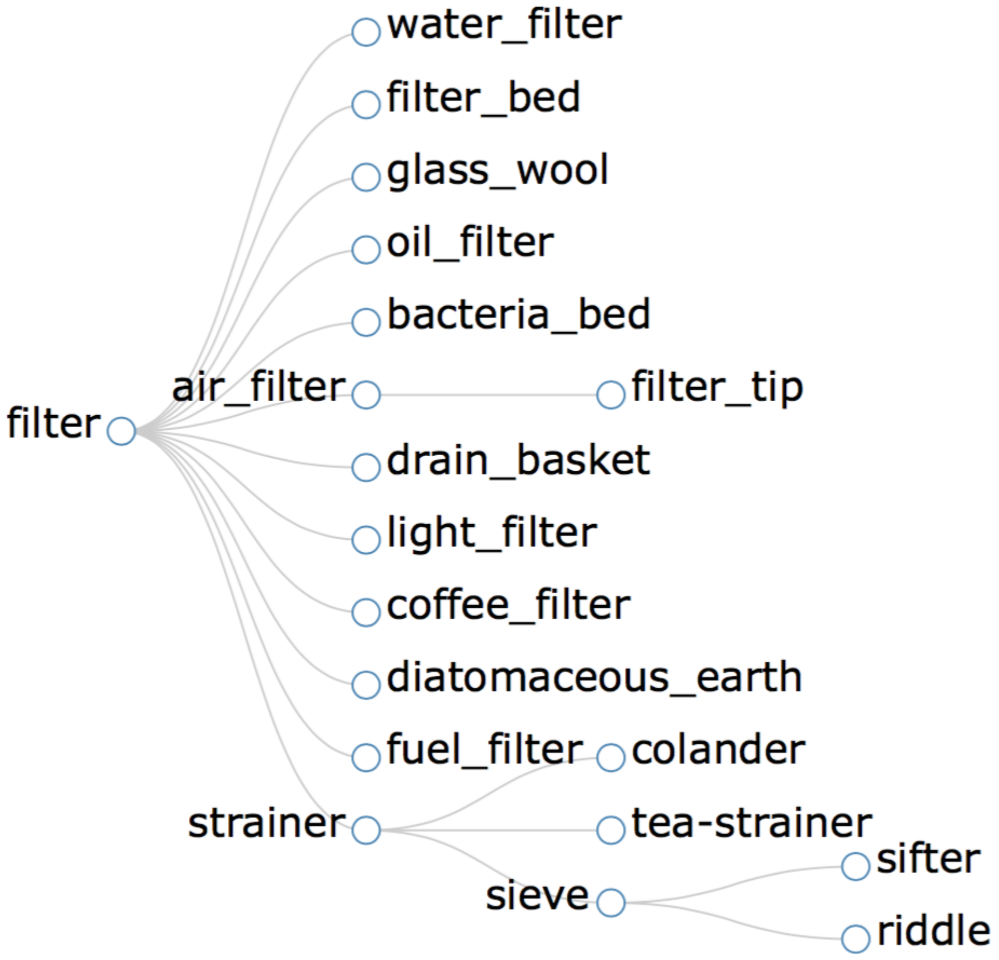


Two types of Taxonomy

- Clustering-based Taxonomy

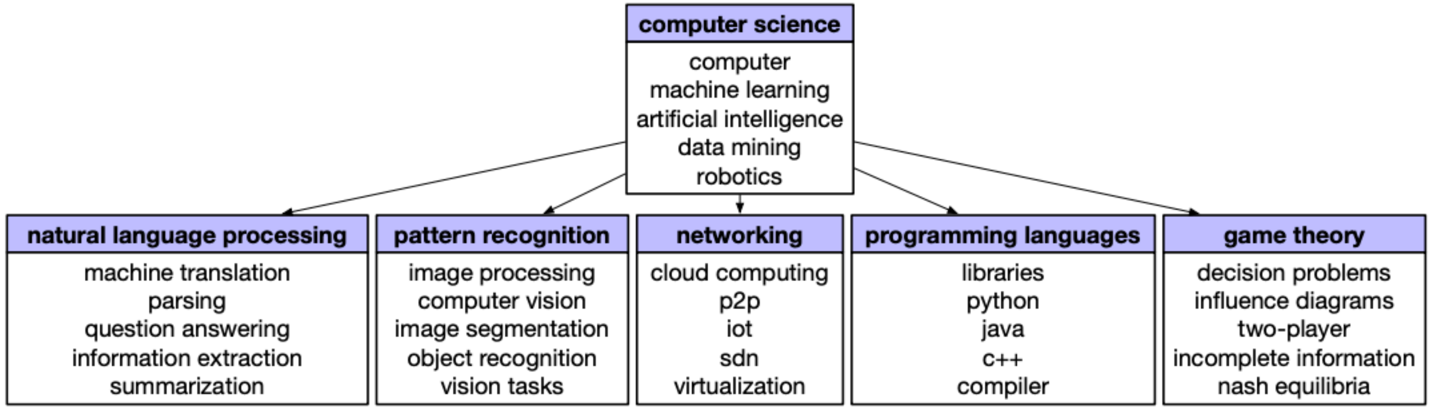


- Instance-based Taxonomy

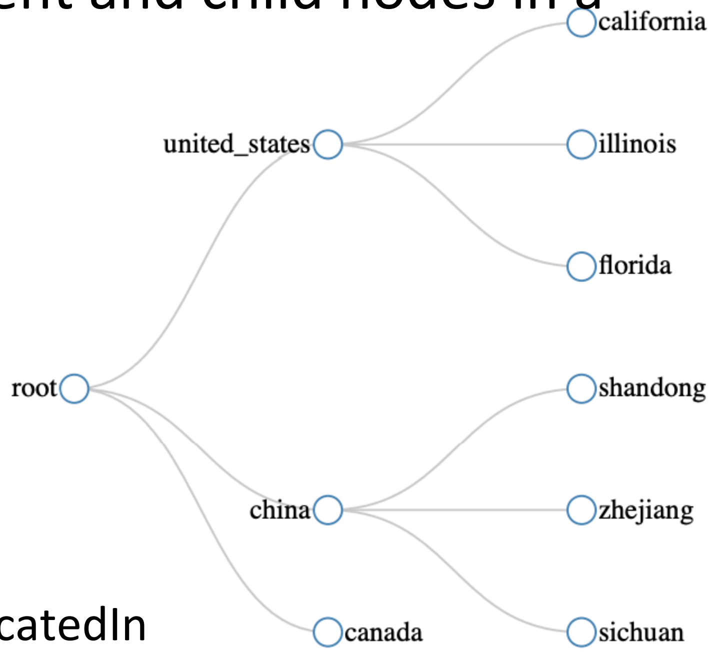


Multi-faceted Taxonomy Construction

- Limitations of existing taxonomy:
  - A generic taxonomy with fixed “is-a” relation between nodes
  - Fail to adapt to users’ specific interest in special areas by dominating the hierarchical structure of irrelevant terms
- Multi-faceted Taxonomy
  - One facet only reflects a certain kind of relation between parent and child nodes in a user-interested field.



Relation: IsSubfieldOf



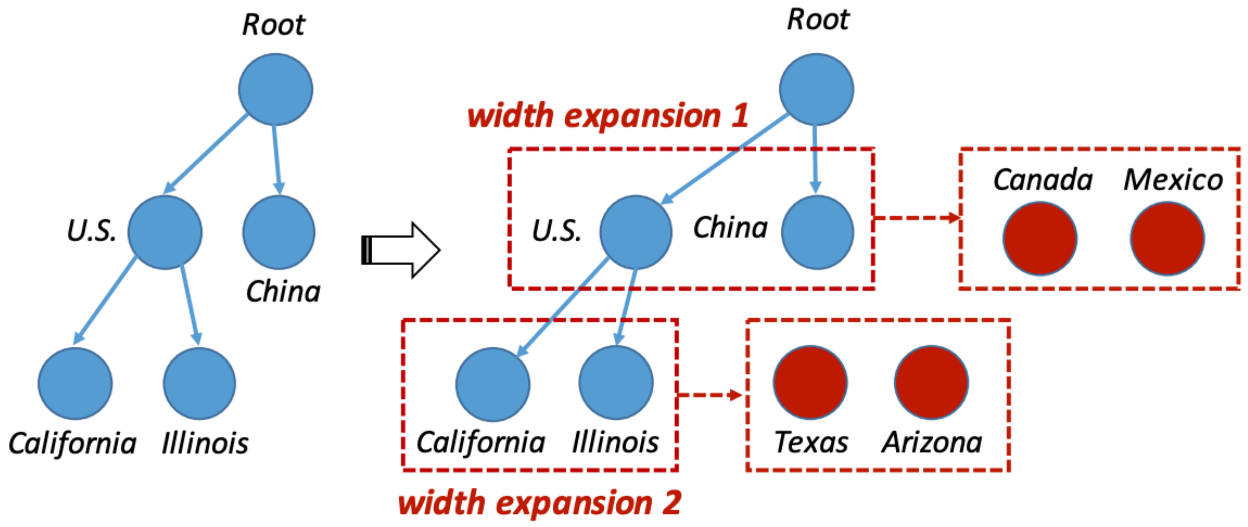
Relation: IsLocatedIn

Two stages in constructing a complete taxonomy

- Taxonomy Construction from Scratch
  - Use a set of entities (possibly a seed taxonomy in a small scale) and unstructured text data to build a taxonomy organized by certain relations
- Taxonomy Expansion
  - Update an already constructed taxonomy by attaching new items to a suitable node on the existing taxonomy. This step is useful since reconstructing a new taxonomy from scratch can be resource-consuming.

Concept Expansion as a Flat Version

- If a seed taxonomy is provided by user, then we can gradually expand a hierarchical structure by the following two sub-tasks:
  - (1) concept expansion as a flat version to expand a wide range of entities on the same level;
  - (2) taxonomy construction as a hierarchical version to capture user-interested relations.



Outline

- Taxonomy Basics and Construction
- Parallel Concept Discovery: Entity Set Expansion
  - EgoSet [WSDM' 16]
  - SetExpan [ECML PKDD’17]
  - SetCoExpan [WWW’20]
  - CGExpan [ACL’20]
- Taxonomy Construction from Scratch
- Taxonomy Expansion

Automated Corpus-Based Set Expansion

- **Corpus-based set expansion***:* Find the "complete" set of entities belonging to the same semantic class, based on a given corpus and a tiny set of seeds
  - Ex. 1. Given {Illinois, Maryland}, derive all U.S. states
  - Ex. 2. Given {machine learning, data mining,…}, derive CS disciplines
- Challenges: Deal with noisy context feature derived from free-text corpus, which may lead to entity intrusion and semantic drifting

Previous Work on Set Expansion

- Search engine-based, online processing: E.g., *Google Set, SEAL, Lyretail*
  - A query consisting of seed entities is submitted to a search engine
  - Good quality but time-consuming and costly
- *Corpus-based* set expansion: offline processing based on a specific corpus
  - Two approaches: *One-time entity ranking* and *iterative pattern-based bootstrapping*
- *One-time entity ranking:* Similar entities appear in similar contexts
  - One-time ranking of candidates based on their distributional similarity with seeds
  - One-time is hard to obtain the full set; *Entity intrusion* error: wrong one intruded
- *Iterative pattern-based bootstrapping:*
  - From seeds to extract quality patterns, based on a predefined scoring mechanism
  - Then apply extracted patterns to obtain even higher quality entities using another scoring method.
  - Semantic drifting: Non-perfect extraction leads to drifting

EgoSet: Methodology

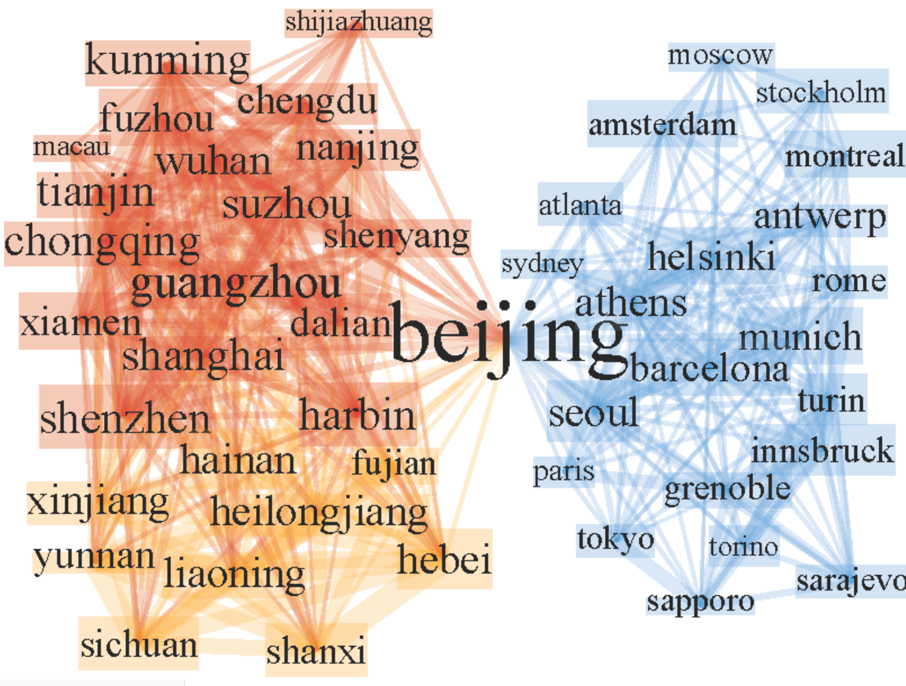
- Ontologies and skip grams: Combine existing **user-generated ontologies** (Wikipedia) with a novel word-similarity metric based on **skip-grams**
- **Ego-network generation:** Treat words that are distributionally similar to the seed (the ego) as nodes and use the pairwise similarity between those words to create weighted edges, thereby forming an "**ego-network**"
- **EgoSet discovery**: Use the ego-network to find the initial clusters for a seed, and align those clusters with user-created ontologies

Wikipedia lists

word ego-network

Olympic Host Cities

City	Country
Albertville	France
Amsterdam	Netherlands
Antwerp[g]	Belgium
Athens	Greece
Atlanta	United States
Barcelona	Spain
Beijing	China
Berlin	Germany

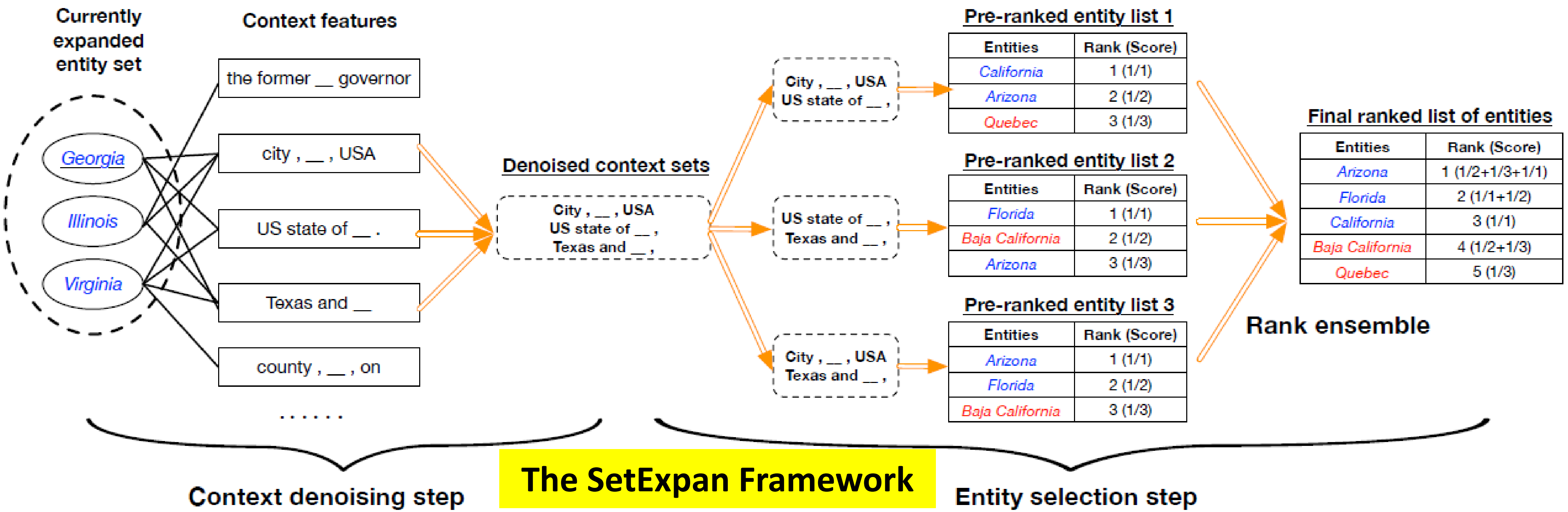


- Ontologies are a natural source for set expansion
- **List-of pages of Wikipedia** were found to have the right combination of being prevalent and relatively "clean"

Outline

- Taxonomy Basics and Construction
- Parallel Concept Discovery: Entity Set Expansion
  - EgoSet [WSDM' 16]
  - SetExpan [ECML PKDD’17]
  - SetCoExpan [WWW’20]
  - CGExpan [ACL’20]
- Taxonomy Construction from Scratch
- Taxonomy Expansion

SetExpan: Context Feature Selection and Rank Ensemble



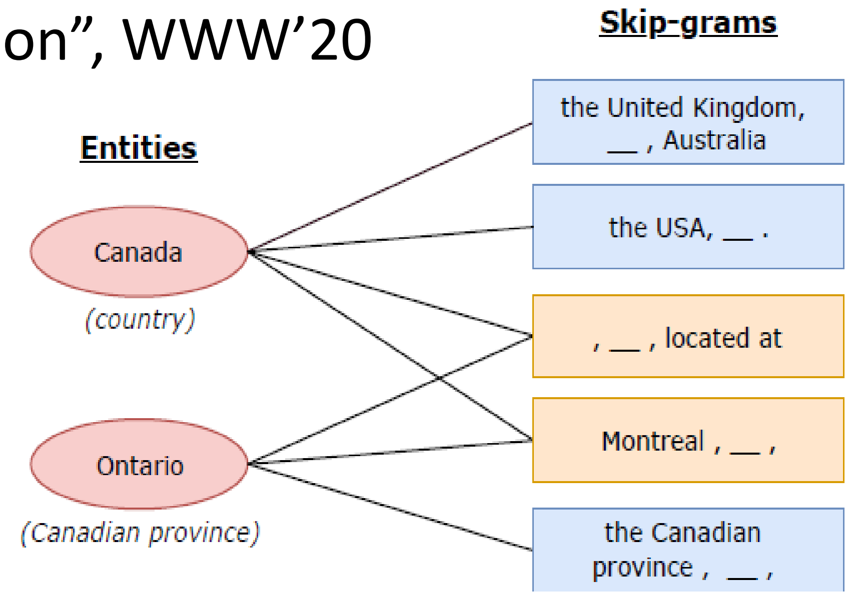
- Instead of using all context features, context features are carefully selected by calculating distributional similarity
- High-quality feature pool will be reset at the beginning of each iteration
- Unsupervised ranking-based ensemble method at each iteration: robust to noisy or wrongly extracted pattern features

Outline

- ❑ Taxonomy Basics and Construction
- ❑ Parallel Concept Discovery: Entity Set Expansion
  - ❑ EgoSet [WSDM' 16]
  - ❑ SetExpan [ECML PKDD’17]
  - ❑ SetCoExpan [WWW’20]
  - ❑ CGExpan [ACL’20]
- ❑ Taxonomy Construction from Scratch
- ❑ Taxonomy Expansion

SetCoExpan: Auxiliary Sets Generation and Set Co-Expansion

- J. Huang, Y. Xie, Y. Meng, J. Shen, Y. Zhang and J. Han, "Guiding Corpus-based Set Expansion by Auxiliary Sets Generation and Co-Expansion", WWW'20



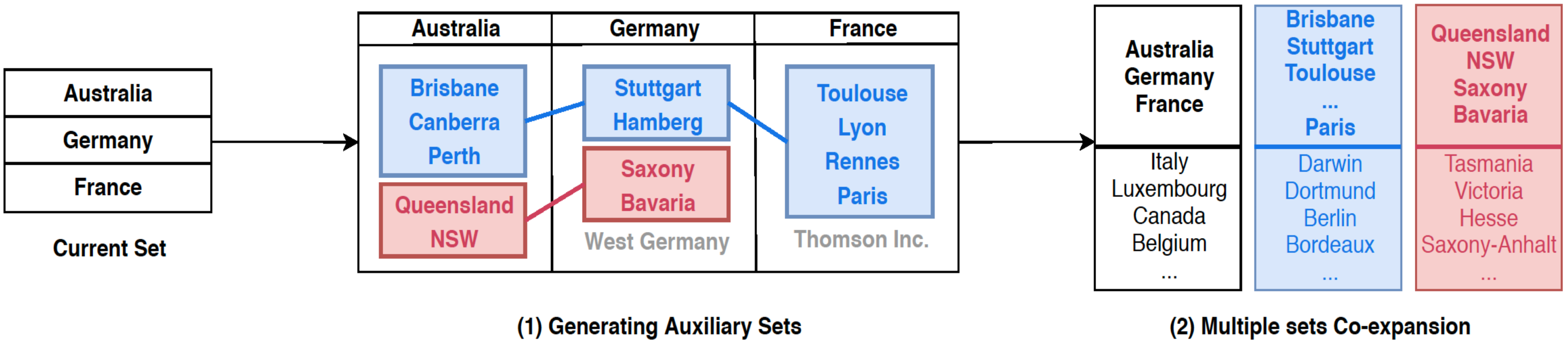
- Semantic drifting problem:
  - Existing set expansion algorithms typically bootstrap the given seeds by incorporating lexical patterns and distributional similarity
- Typical errors
  - Similar concept but wrong granularity
    - Countries: (United States, Canada, China) → Ontario, Illinois, California
  - Related but not similar concepts
    - Sports_leagues: (NHL, NFL, NBA) → Chicago Rush, San Francisco Giants, Yankee
    - Companies: (Google, Apple Inc. , Microsoft) → google maps, ios, android
    - Diseases: (lung cancer, AIDS, depression) → chest pain, headache, fever

Resolving the Semantic Drift Issue by Set Co-expansion

- Automatically generate *auxiliary sets* as negative sets that are *closely related to the target set of user's interest*
  - Auxiliary set: holding certain subtle relations in the embedding space with the user-interested semantic class
- Co-expand multiple coherent sets that are distinctive from one another
  - Expand all the sets in parallel to mutually enhance each other by finding the most contrastive features to tell their difference
- Example:
  - User input: Australia, France, Germany
  - Generated Auxiliary sets:
    - (Provinces): Queensland, New South Wales, Saxony, Bavaria
    - (Cities): Brisbane, Canberra, Rennes, Hamburg

General Framework of Set Co-expansion

- At each iteration of expanding the seed set, perform two steps
  - **Auxiliary Sets Generation**: Run CatE then Clustering to automatically generate auxiliary sets (related to but different from the semantic class of user input)
  - **Multiple Sets Co-Expansion**: Expand multiple sets simultaneously by extracting the most contrastive features, and expand each set in the direction away from other sets



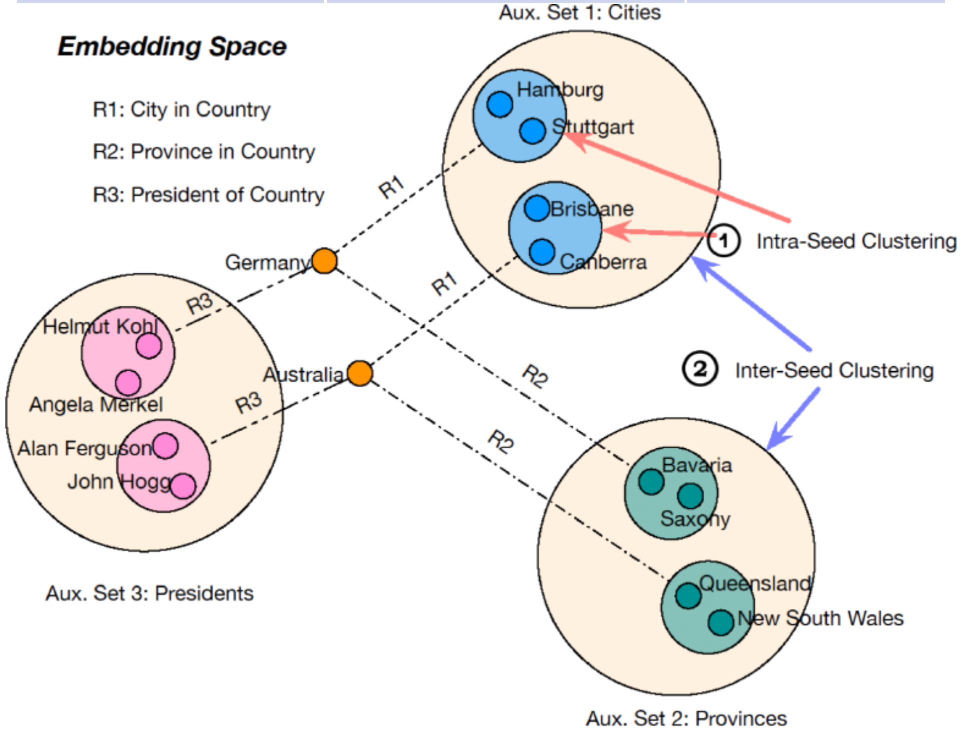
Auxiliary Sets Generation

- Semantic Learning and Related Terms Retrieval for Seed Entities
  - Generate representative terms for each entities in the seed set
- Cross-Seed Parallel Relations Clustering
  - Intra-seed clustering: Cluster terms related to each seed into initial semantic groups
  - Inter-seed clustering: Merge initial groups across different seeds using the equation:

$$Relation(e_1 \in C_T, g_1) \approx Relation(e_2 \in C_T, g_2)$$

  - Remove groups that cannot match in different sets—retain only the cross-seed groups

australia	germany	france
queensland	west_germany	provence
nsw	bavaria	montpellier
brisbane	saxony	rennes
canberras	hamburg	lyon
perth	stuttggart	toulouse



Multiple Sets Co-Expansion

- Iteratively refine *feature pool* and *candidate pool* in set expansion
  - **Feature pool** stores common context features of seed entities, that best distinguish the target semantic class from auxiliary ones
    - Skip-grams that make each set coherent while distinguishing different sets are encouraged
  - **Candidate pool** stores the possible candidate entities to be expanded, and they are narrowed down by co-occurrence with features in the feature pool

$$F^* = \arg\max_{|F|=Q} \frac{2}{|S| * (|S|-1)} \sum_{e_i, e_j \in S} Sim(e_i, e_j | F)$$
$$- \sum_{S_k, S_{k'} \in C_{aux}} \frac{1}{|S_k| * |S_{k'}|} \sum_{e_i \in S_k, e_j \in S_{k'}} Sim(e_i, e_j | F)$$
$$+ \sum_{S_k \in C_{aux}} \frac{2}{|S_k| * (|S_k|-1)} \sum_{e_i, e_j \in S_k} Sim(e_i, e_j | F)$$

Skip-grams shared by entities in different sets are scored lower.

Skip-grams shared by entities in the same set are scored higher.

Experiments and Performance Study

- Experiment data sets:
  - Each class includes 5 queries
    - from the same semantic class (e.g., Countries, Companies, Sports Leagues)

Dataset	# classes	# queries	entity vocabulary size	# documents
Wiki	8	40	41242	780556
APR	3	15	71707	1014140

- Evaluation Metric: Mean Average Precision (MAP)
- Methods compared
  - SetExpan (ECMLPKDD'17)
  - SetExpander (EMNLP'18)
  - CaSE (SIGIR'19)
  - BERT (NAACL'19)

Table 3: Mean Average Precision across all queries on *Wiki* and *APR*.

Methods	*Wiki*			*APR*		
	MAP@10	MAP@20	MAP@50	MAP@10	MAP@20	MAP@50
CaSE	0.897	0.806	0.588	0.619	0.494	0.330
SetExpander	0.499	0.439	0.321	0.287	0.208	0.120
SetExpan	0.944	0.921	0.720	0.789	0.763	0.639
BERT	0.970	0.945	0.853	0.890	0.896	0.777
Set-CoExpan (no aux.)	0.964	0.950	0.861	0.900	0.893	0.793
Set-CoExpan (no flex.)	0.973	0.961	0.886	0.927	0.908	0.823
Set-CoExpan	**0.976**	**0.964**	**0.905**	**0.933**	**0.915**	**0.830**

When the ranking list is longer (i.e., when the seed set gradually grows out of control and more noises appear), SetCoExpan is able to steer the direction of expansion and set barriers to prevent out-of category words from coming in

Case Studies

Negative seeds (auxiliary sets) generated for various queries (in Wiki Dataset)

Table 4: Auxiliary sets generated for various queries.

Class	Query	Auxiliary sets
Companies	Myspace, Youtube, Twitter	**(Products):** flickr, wordpress, google earth, gmail, google maps
Countries	Australia, France, Germany	**(Provinces):** Queensland, New South Wales, Saxony, Bavaria, Thuringia **(Cities):** Brisbane, Canberra, Rennes, Hamburg, Stuttgart
TV Channels	ESPN News, ESPN Classic, ABC	**(TV Programmes):** the young and the restless, all my children, guiding light, general hospitale
Sports Leagues	national football league, national hockey league, major league baseball	**(Sports Teams):** new york jets, ottawa senators, chicago white sox, dallas cowboy, st.louis hawks
Political Parties	new democratic party, liberal party of canada, northern ireland labour party	**(Elections):** 1980 federal election, 1997 federal election, 1980 election, 1962 election, 2008 provincial election
Chinese Provinces	jiangsu, liaoning, sichuan	**(China Cities):** xi'an, hangzhou, shanghai, chengdu, beijing
Diseases	tuberculosis, parkinson's disease, esophageal cancer	**(Symptoms):** tumor, dehydration, dementia, muscle stiffness
US States	Texas, Florida, New Mexico	**(US Cities):** fort worth, san antonio, jacksonville, tampa, orlando

Co-Expansion vs. Separate Expansion of Target Set and Auxiliary Sets

Table 5: Results of Co-Expansion and Separate Expansion of Target Set and Auxiliary Sets.

seeds	seeds from Target Set: Australia, France, Germany		seeds from Aux. Set 1: Queensland, Saxony, New South Wales		seeds from Aux. Set 2: Brisbane, Canberra, Stuttgart	
Multiple Sets Co-Expansion	Italy	Luxembourg	Baden-Wurttemberg	Hesse	Berlin	Hanover
	Canada	Belgium	Baden	Saxony-Anhalt	Dortmund	Frankfurt
	Norway	Spain	Schleswig-Holstein	Silesia (✗)	Heidelberg	Strasbourg
	The Netherlands	Denmark	Rhineland-Palatinate	WestPhalia	Munich	Bonn
	England	Switzerland	Mecklenburg-Vorpommern	Saarland	Cologne	Mannheim
Separate Expansion of Each Set	Italy	Luxembourg	Baden-Wurttemberg	WestPhalia	Strasbourg	Berlin
	Canada	Belgium	Hesse	Saxony-Anhalt	Marseille	Hanover
	Spain	Brussels (✗)	Baden	Berlin	Auxerre	Lyon
	England	Paris (✗)	Wurttemberg	Munich (✗)	AS Saint-Etienne (✗)	Nancy
	Switzerland	Ireland	Franconia (✗)	Stuttgart (✗)	Paris Saint-Germain (✗)	Lens

Outline

- Taxonomy Basics and Construction
- Parallel Concept Discovery: Entity Set Expansion
  - EgoSet [WSDM' 16]
  - SetExpan [ECML PKDD’17]
  - SetCoExpan [WWW’20]
  - CGExpan [ACL’20]
- Taxonomy Construction from Scratch
- Taxonomy Expansion

Class-guided Set Expansion: Overall Idea

- We propose to empower entity set expansion with **class names** automatically generated from pre-trained language models, which can help us identify **unambiguous patterns** and eliminate erroneous entities
- CGExpan: Class-Guided Set Expansion

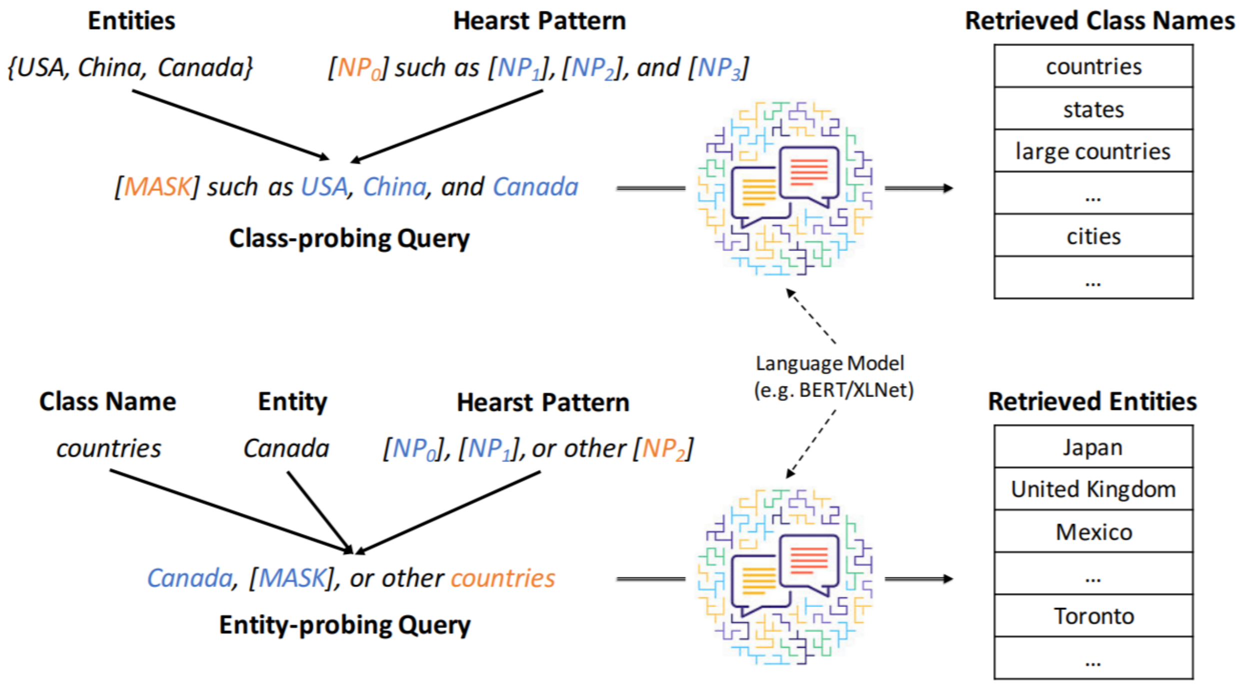


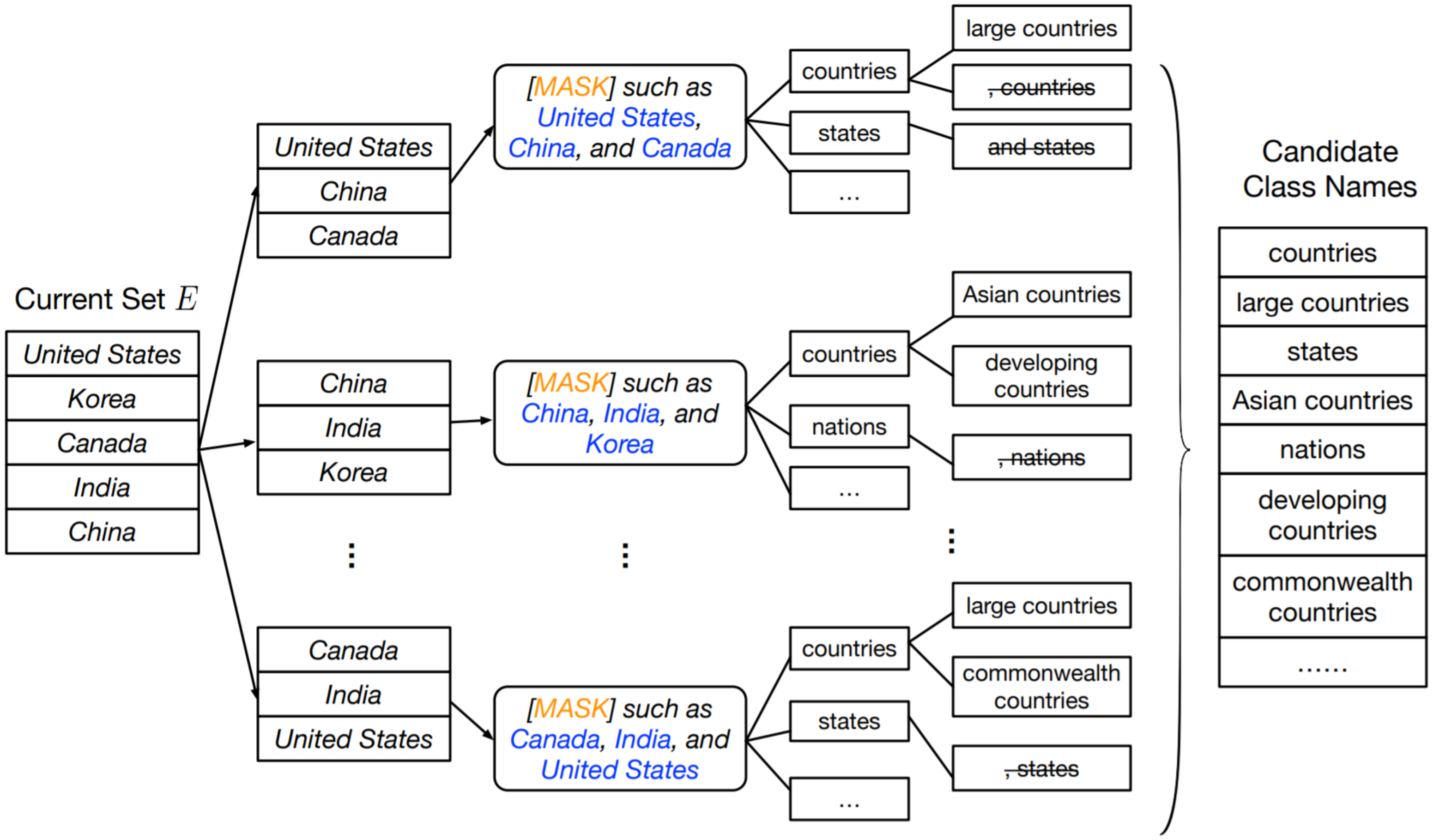
Figure 1: Examples of class-probing and entity-probing queries generated based on Hearst patterns.

Probing Queries

- Hearst Patterns: a set of lexico-syntactic patterns inducing hypernym relations
  - E.g. "<u>countries</u> such as <u>US</u> and <u>China</u>" -> China, US ∈ Countries
- Probing Query: a word sequence containing one [MASK] token
  - **Class-probing query**: predict class name of some given entities
    - E.g. "[MASK] such as USA, China, and Canada"
  - **Entity-probing query**: retrieve entities given class name and some seed entities
    - E.g. "Canada, [MASK], or other countries"
- By inputting a probing query, we can get the contextualized embedding of the [MASK] token and let MLM predict the missing word

CGExpan: Class Name Generation

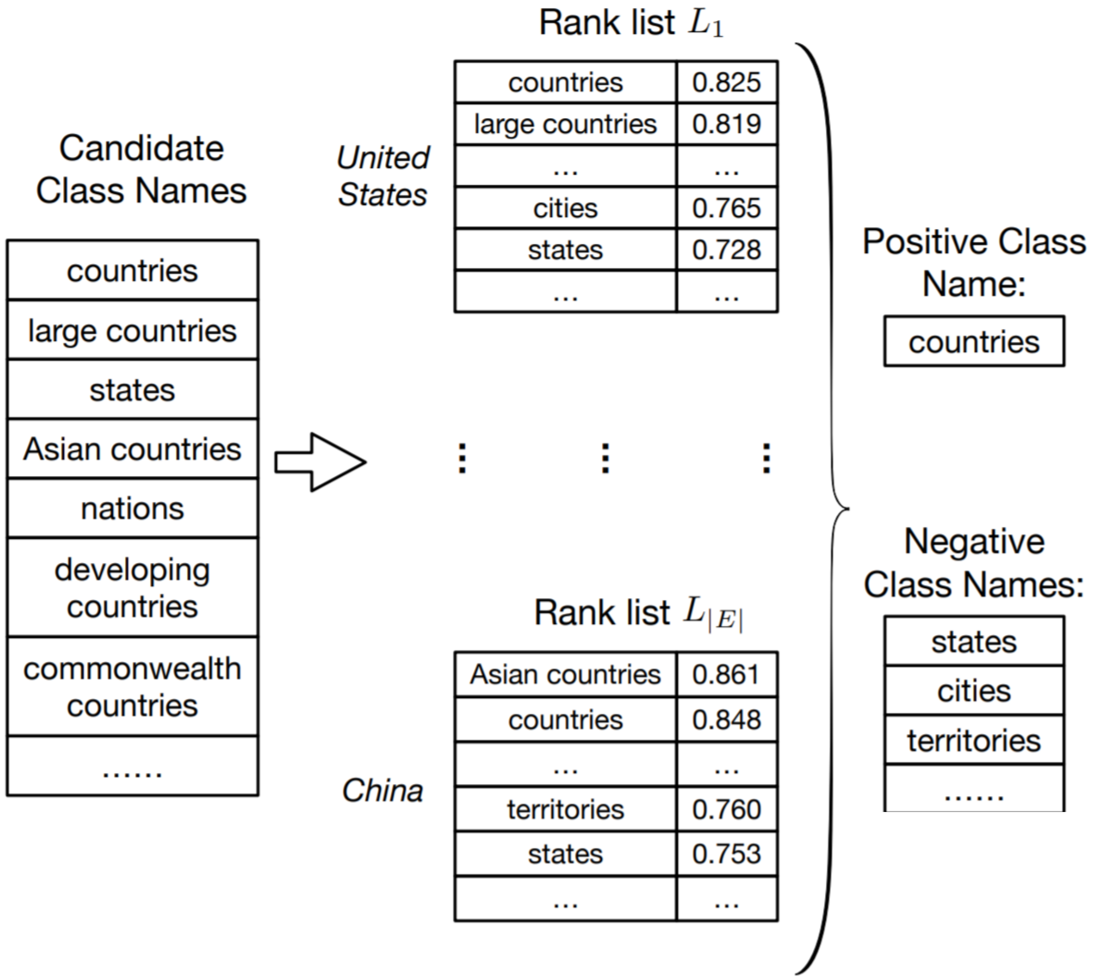
- Iteratively submit class-probing queries to a language model to get multi-gram class names
- Repeat the process by randomly sampling entities
- Keep all generated class names that are noun phrases



26

CGExpan: Class Name Ranking

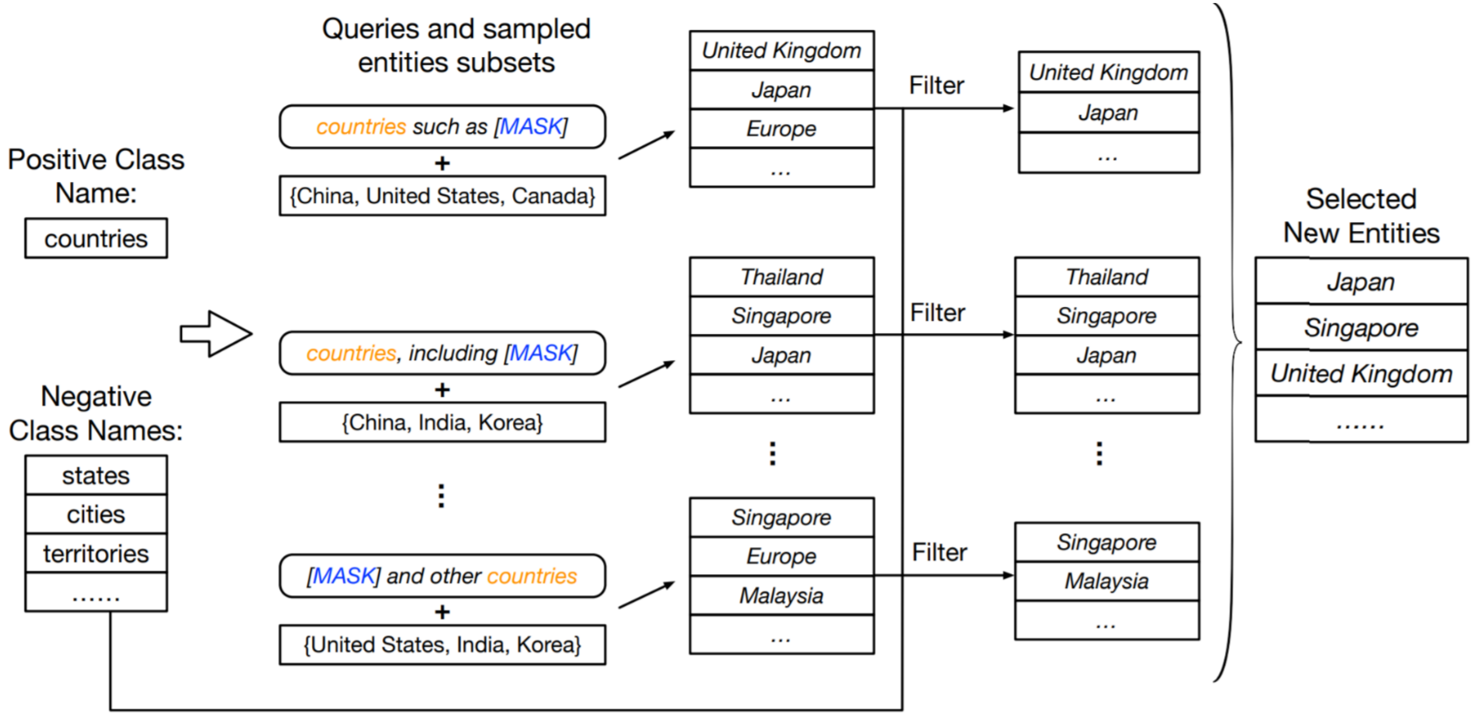
- Identify the top-k most similar occurrences of an **entity** with the embedding vector of an **entity-probing query** and take their average as the similarity between the entity and a class name
- Aggregate all ranked lists (one for each entity) and select the top one as the positive class name, $\boldsymbol{c_p}$
- Select class names ranking lower than c_p in **all** lists corresponding to the **initial seed set** as negative class names, $\boldsymbol{C_N}$



CGExpan: Class-Guided Entity Selection

- Prefer entities that appear at top position in multiple entity rank lists
- Filter out entities that are more similar to any $c' \in C_N$ than c_p
- Assign higher score to entities currently in the set

$$mmrr(e_i) = \sum_{t=1}^{T} \left(\mathbb{1}(e_i \in E) + \frac{1}{r_i^t} \right) \times \mathbb{1}(r_{c_p}^i < \min_{c' \in C_N} r_{c'}^i),$$



28

Case Study: Class Name Selection

Seed Entity Set	Ground True Class Name	Positive Class Name	Negative Class Names
{*“Intel”*, *“Microsoft”*, *“Dell”*}	`company`	`company`	`product`, `system`, `bank`, ...
{*“United States”*, *“China”*, *“Canada”*}	`country`	`country`	`state`, `territory`, `island`, ...
{*“ESPNews”*, *“ESPN Classic”*, *“ABC”*}	`tv channel`	`television network`	`program`, `sport`, `show`, ...
{*“NHL”*, *“NFL”*, *“American league”*}	`sports league`	`professional league`	`sport`, `competition`, ...
{*“democratic”*, *“labor”*, *“tories”*}	`party`	`political party`	`organization`, `candidate`, ...
{*“Hebei”*, *“Shandong”*, *“Shanxi”*}	`Chinese province`	`chinese province`	`city`, `country`, `state`, ...
{*“tuberculossi”*, *“Parkinson’s disease”*, *“esophageal cancer”*}	`disease`	`chronic disease`	`symptom`, `condition`, ...
{*“Illinois”*, *“Arizona”*, *“California”*}	`US state`	`state`	`county`, `country`, ...

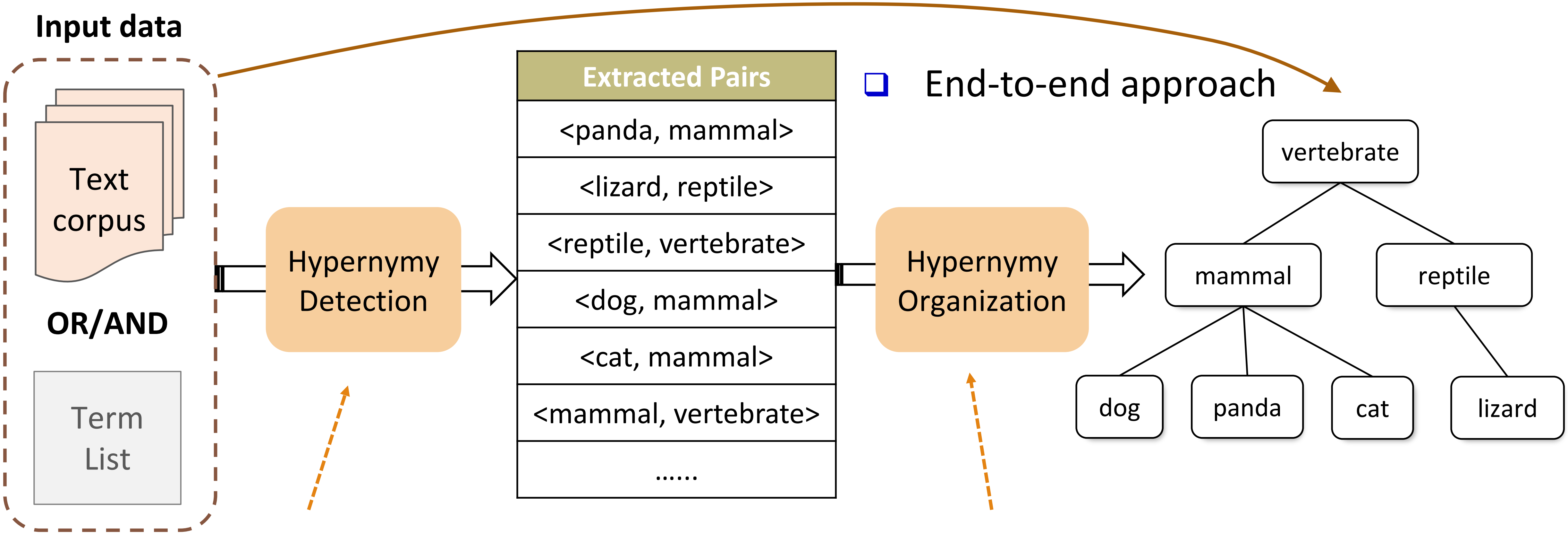
Table 5: Class names generated for seed entity sets. The 2nd column is the ground true class name in the original dataset. The 3rd and 4th columns are positive and negative class names predicted by CGExpan, respectively.

Outline

- Taxonomy Basics and Construction
- Parallel Concept Discovery: Entity Set Expansion
- Taxonomy Construction from Scratch
  - Instance-based Taxonomy Construction
    - Hypernym-hyponym detection
    - HiExpan: Task-guided Taxonomy Construction by Hierarchical Tree Expansion [KDD’18]
  - Clustering-based Taxonomy Construction
- Taxonomy Expansion

Instance-based Taxonomy Construction: Overview

- ❑ Decompose taxonomy construction into multiple subtasks



- ❑ Pattern-based approach
- ❑ Supervised approach

- ❑ Simple pruning heuristics
- ❑ Graph-based approach

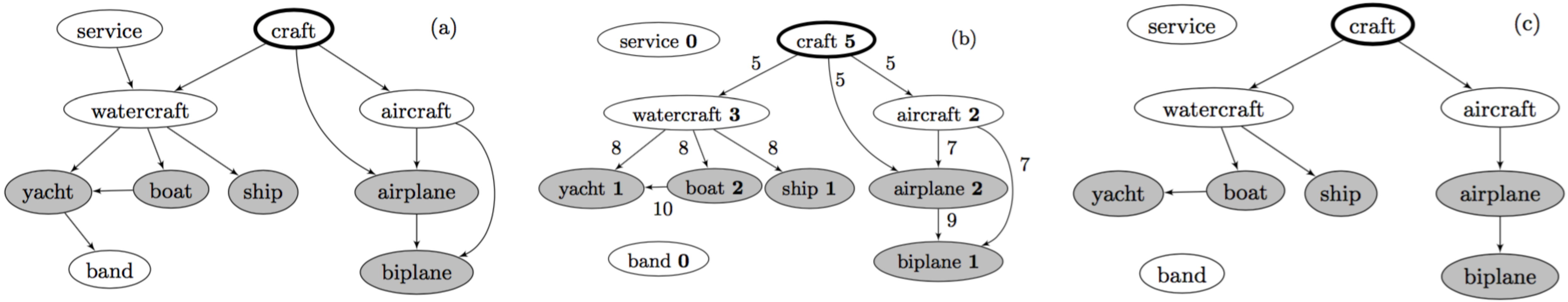
Hypernymy Detection

- Pattern-based approach: use patterns to extract hypernym-hyponym relations from raw text
  - Lexical-syntactic pattern [Hearst'92] [Kozareva and Hovy'10], [Luu et al.'14]
- Supervised approach: train a classifier to predict whether two terms in vocabulary hold hypernymy relation
  - Leverage multiple features:
    - Term embedding: [Fu et al.' 14] [Yu et al.15] [Luu et al.'16] [Weeds et al.'16]
    - Dependency path: [Snow et al.'04] [Snow et al.'06] [Shwartz et al.'16] [Mao et al.'18]

Hypernymy Organization

- Simple pruning heuristics:
  - Remove cycle [Kozareva and Hovy'10] [Faralli et al.'15]
  - Retain longest-path [Kozareva and Hovy'10]
- Graph-based approach:
  - Maximum Spanning Tree [Paola et al.'13] [Bansal et al.'14] [Zhang et al.'16]

Figure credits to [Paola et al.'13]



Noisy Graph **Trimmed Graph with Edge Weights** **Induced DAG**

Limitations of Existing Methods

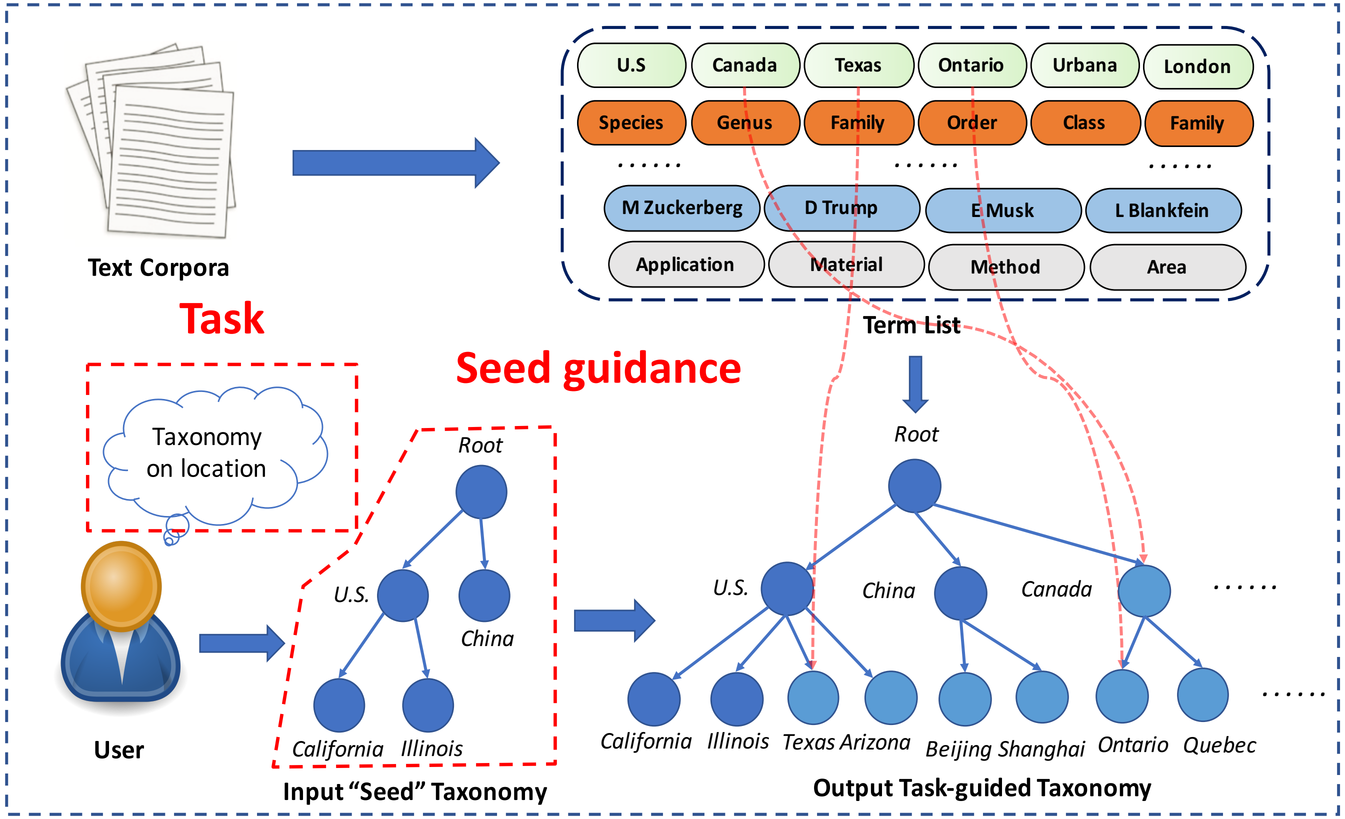
- Limitations: Build a corpus-agnostic, task-agnostic taxonomy with mainly is-A relation
  - ***Inflexible semantics***: cannot model flexible edge semantics (e.g., “country-state-city”)
  - ***Limited applicability***: cannot fit user-specific application tasks

Outline

- Taxonomy Basics and Construction
- Parallel Concept Discovery: Entity Set Expansion
- Taxonomy Construction from Scratch
  - Instance-based Taxonomy Construction
    - Hypernym-hyponym detection
    - HiExpan: Task-guided Taxonomy Construction by Hierarchical Tree Expansion [KDD'18]
  - Clustering-based Taxonomy Construction
- Taxonomy Expansion

HiExpan: User/Task-Guided Taxonomy Construction

- Input: A user provides:
  - a domain-specific corpus, and
  - a seed taxonomy as task guidance
- Model outputs:
  - A corpus-dependent taxonomy tailored for user's task
- Distinction: <u>Task-guided</u> taxonomy construction
  - Corpus-dependent
  - Leverage user's seed guidance

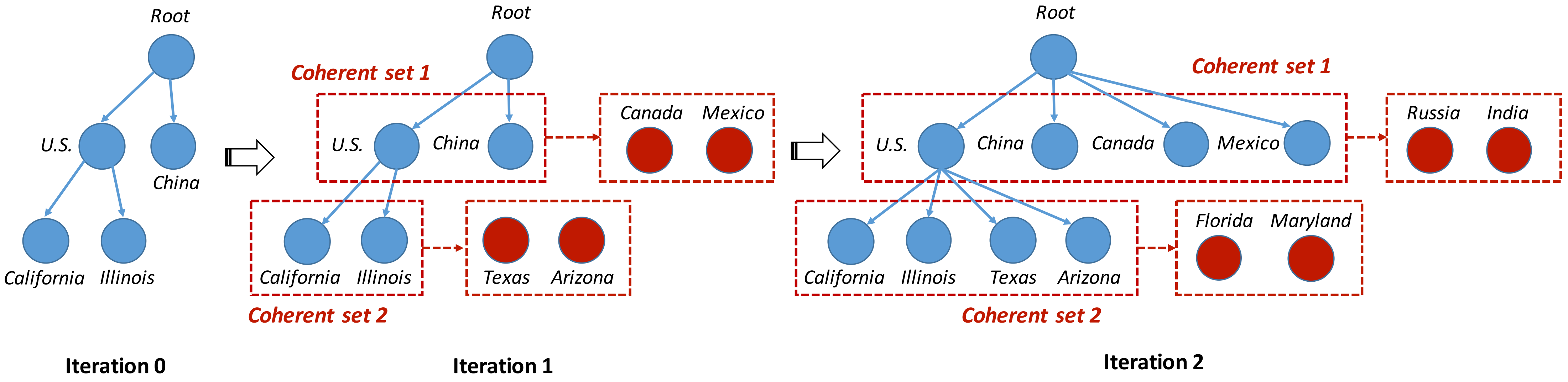


Shen, Jiaming, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle Vanni, Brian M. Sadler and Jiawei Han. "HiExpan : Task-Guided Taxonomy Construction by Hierarchical Tree Expansion." KDD (2018)

The HiExpan Framework & Width Expansion

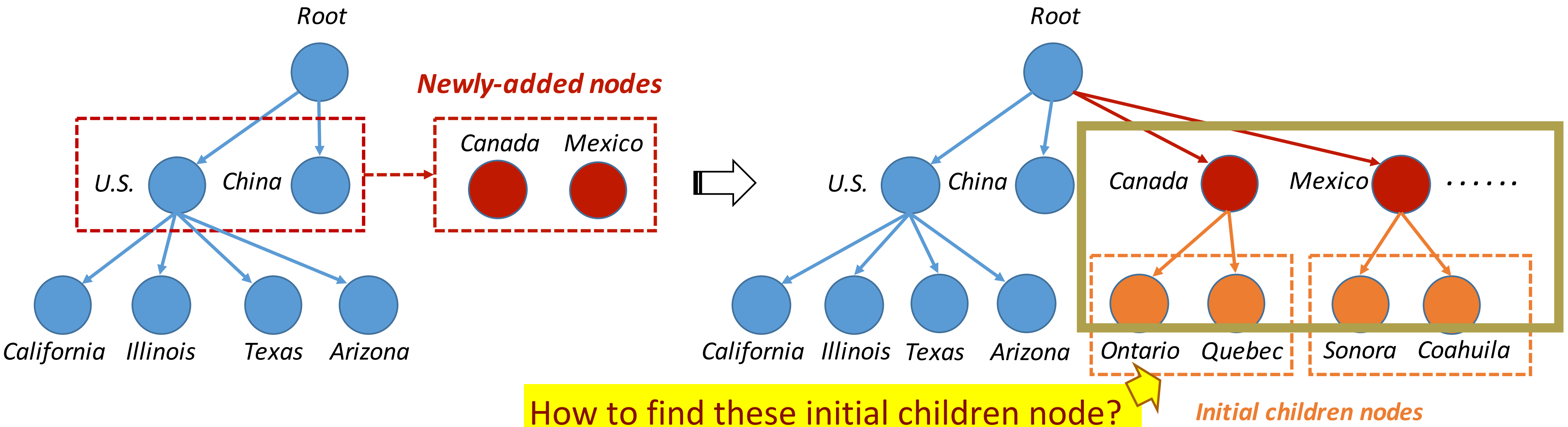
- The HiExpan core idea: View all children under each taxonomy node forming *a coherent set* and build the taxonomy by expanding all these sets
  - Use set expansion algorithm to expand all sets
  - Recursively expand the sets in a top-down fashion

Width expansion: The width of taxonomy tree increases (i.e., expanded)



How to Dig Deeper? Cold-Start with Empty Initial Seed Set

- Newly-added nodes in taxonomy tree do not have any child node
  - How to acquire a target node's initial children?
- Depth Expansion
  - Based on US (California, Illinois, …), find Canada (Ontario, Quebec, …), Mexico (…)
  - Based on term embedding and embedding vector similarity



How to find these initial children node?

Outline

- Taxonomy Basics and Construction
- Parallel Concept Discovery: Entity Set Expansion
- Taxonomy Construction from Scratch
  - Instance-based Taxonomy Construction
  - Clustering-based Taxonomy Construction
    - Hierarchical Topic Models
    - TaxoGen: Constructing Topical Concept Taxonomy by Adaptive Term Embedding and Clustering [KDD’18]
    - CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring [KDD’20]
    - NetTaxo: Automated Topic Taxonomy Construction from Text-Rich Network [WWW’20]
- Taxonomy Expansion

Hierarchical Topic Model

- Use a cluster of terms (i.e., a topic) to represent a concept and organize topics in a hierarchical way
- Pose different statistical assumptions on the data generation process
  - Nested Chinese Restaurant Process:
    - hLDA [Blei et al.'03], hLDA-nCRP [Blei et al.' 10]
  - Pachinko Allocation Model:
    - PAM [Li and McCallum'06], hPAM [Mimno et al.'07]
  - Dirichlet Forest Model：
    - DF [Andrzejewski et al.'09], Guided HTM [Shin and Moon'17]

Example: hLDA

- Assume documents are generated by a nested Chinese Restaurant Process

1. Let c_1 be the root restaurant.
2. For each level $\ell \in \{2, \ldots, L\}$:
   (a) Draw a table from restaurant $c_{\ell-1}$ using Eq. (1). Set c_ℓ to be the restaurant referred to by that table.
3. Draw an L-dimensional topic proportion vector θ from $\mathrm{Dir}(\alpha)$.
4. For each word $n \in \{1, \ldots, N\}$:
   (a) Draw $z \in \{1, \ldots, L\}$ from $\mathrm{Mult}(\theta)$.
   (b) Draw w_n from the topic associated with restaurant c_z.

Generates

We develop an approach to risk minimization and stochastic optimization that provides a convex surrogate for variance, allowing near-optimal and computationally efficient trading between approximation and estimation error.

“Observed” documents

Inference

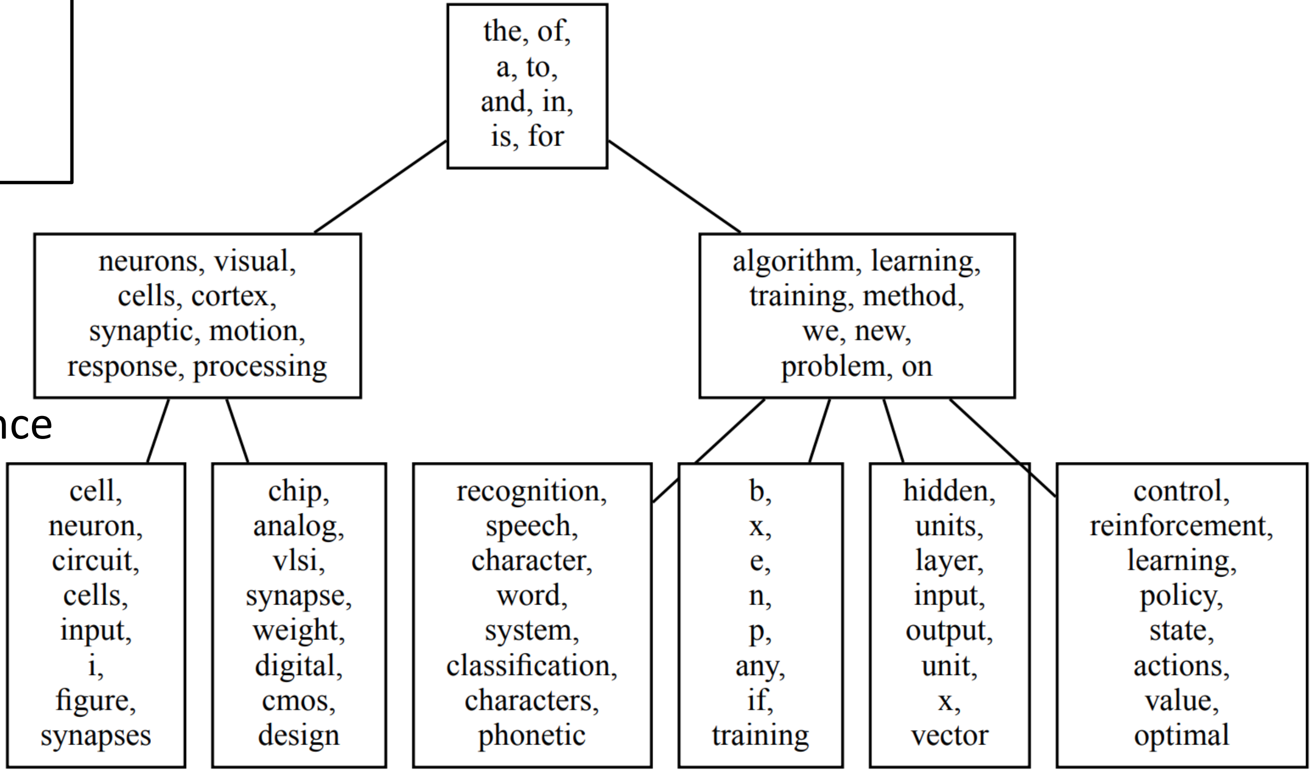


Figure credits to [Blei et al.’03]

Example: hPAM

- Assume documents are generated by a mixture of

1. For each document d, sample a distribution θ_0 over super-topics and a distribution θ_T over sub-topics for each super-topic.
2. For each word w,
   (a) Sample a super-topic z_T from θ_0.
   (b) Sample a sub-topic z_t from θ_{z_T}.
   (c) Sample a level ℓ from $\zeta_{z_T z_t}$.
   (d) Sample a word from ϕ_0 if $\ell = 1$, ϕ_{z_T} if $\ell = 2$, or ϕ_{z_t} if $\ell = 3$.

Generates

We develop an approach to risk minimization and stochastic optimization that provides a convex surrogate for variance, allowing near-optimal and computationally efficient trading between approximation and estimation error.

"Observed" documents

Inference

writes article don time apr **super-topic**
god jesus christ people christian
faith wrong read spiritual passage
agree reason matter statement means
history support community house involved
key government encryption president clipper **sub-topic**
agree reason matter statement means
power arms president home vote
history support community house involved
israel jews israeli jewish arab
history support community house involved
side left happened committee region
agree reason matter statement means
turkish armenian armenians people turkey
side left happened committee region
history support community house involved
hundred clothes tyre bosnians origin
file ftp windows window image
bit fax manager lib uk
site dec sources key public
release size function appreciated box

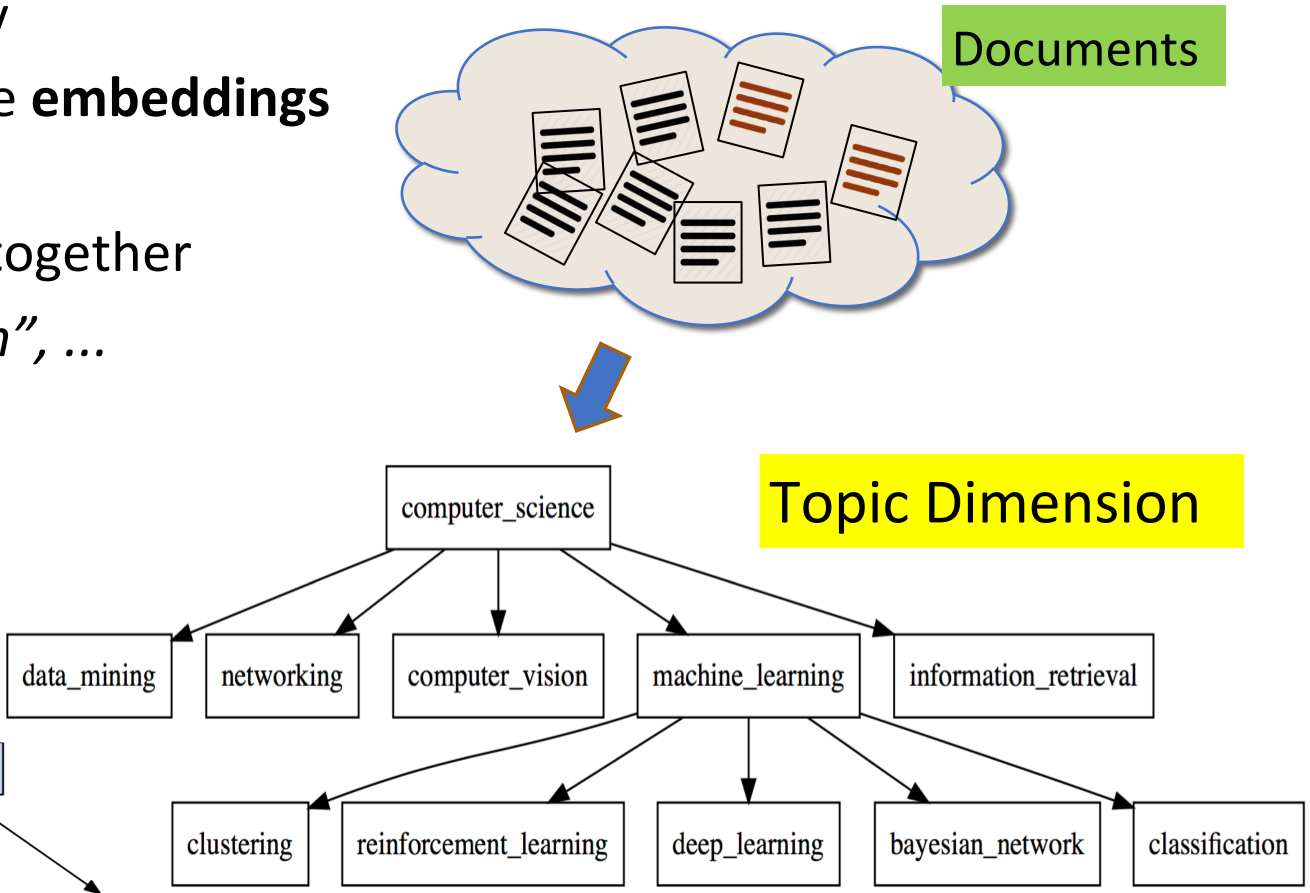
Figure credits to [Mimno et al.'07]

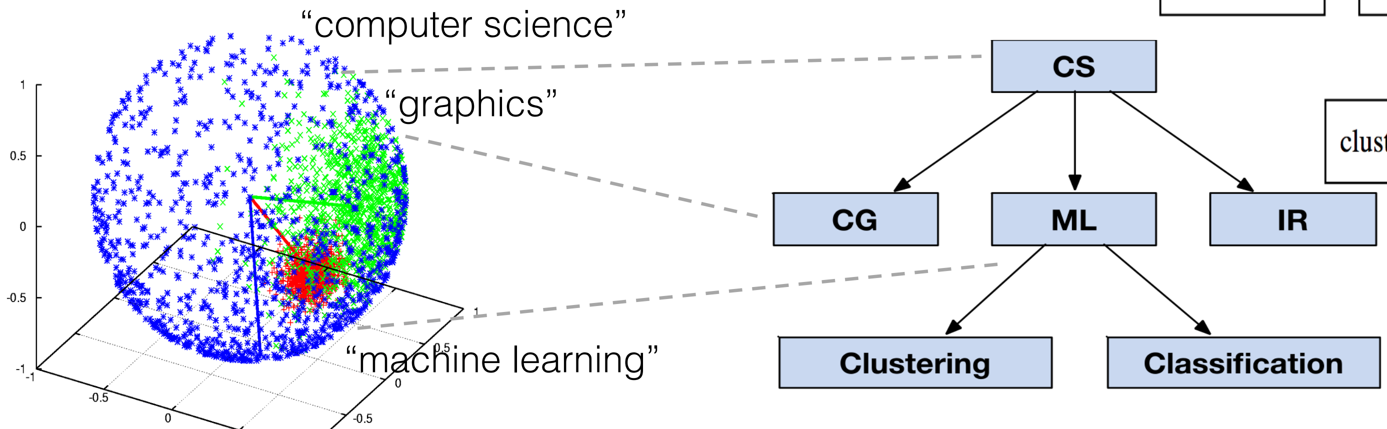
Outline

- Taxonomy Basics and Construction
- Parallel Concept Discovery: Entity Set Expansion
- Taxonomy Construction from Scratch
  - Instance-based Taxonomy Construction
  - Clustering-based Taxonomy Construction
    - Hierarchical Topic Models
    - TaxoGen: Constructing Topical Concept Taxonomy by Adaptive Term Embedding and Clustering [KDD’18]
    - CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring [KDD’20]
    - NetTaxo: Automated Topic Taxonomy Construction from Text-Rich Network [WWW’20]
- Taxonomy Expansion

TaxoGen: Unsupervised Construction with Term Embedding

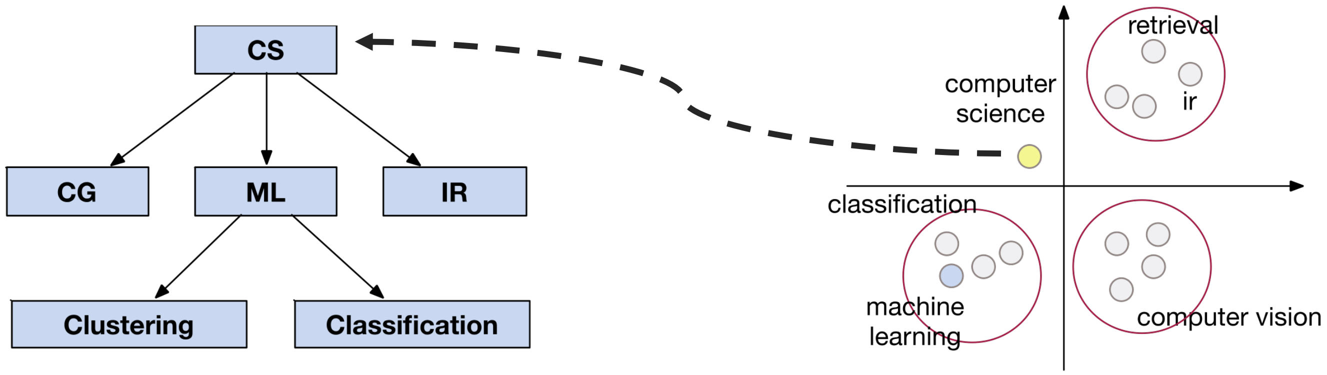
- Automated construction of topic taxonomy
- Selected method: **spherical clustering**—use **embeddings** to find semantically consistent clusters
  - Domain-specific terms can be clustered together
    - *"machine learning", "learning algorithm", ...*
  - Where do the general terms go?
    - *"computer science", "method", "paper*





recursive construction

Spherical Clustering + Local Embedding



recursive construction

adaptive spherical clustering

After pushing up general terms, the remaining terms become more separable

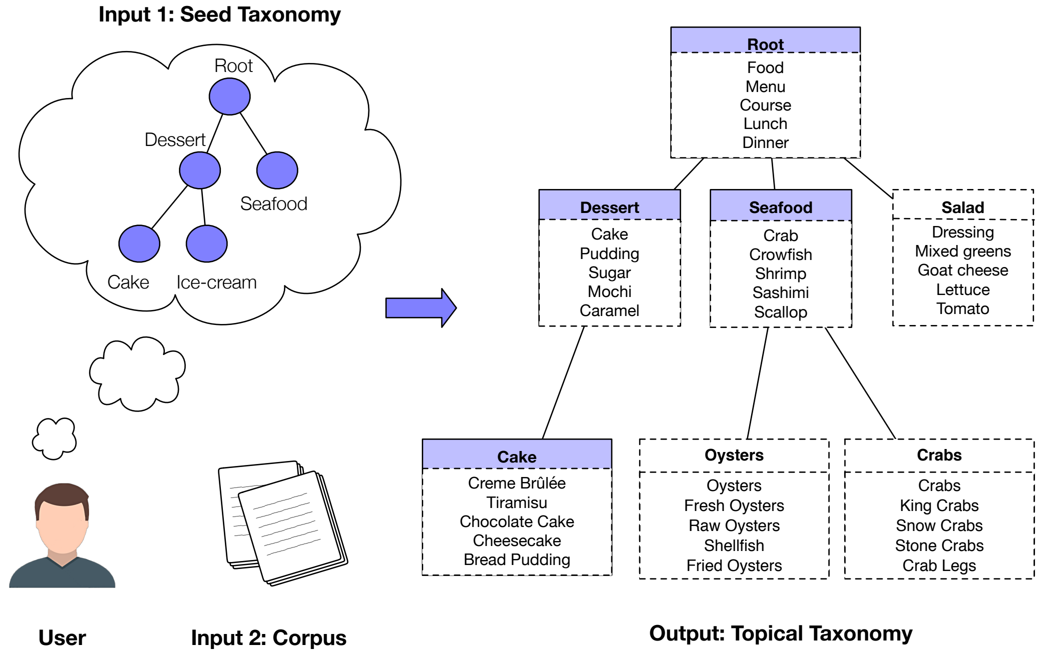
- Design a ranking module to select ***representative phrases*** for each cluster
  - Conduct comparative analysis (combining **popularity** and **concentration)**
    - Does this phrase better fit my cluster or my siblings'?
- Push the ***background phrases*** back to the general node
  - "computer science", "paper" → the higher-level node (root node)
  - "machine learning", "ml", "classification" → the "ML" node
- Local embedding:
  - For each "sub-topic" node, learn ***local embedding*** only on relevant documents
  - Only perserve information relevant to the "sub-topic"

Outline

- Taxonomy Basics and Construction
- Parallel Concept Discovery: Entity Set Expansion
- Taxonomy Construction from Scratch
  - Instance-based Taxonomy Construction
  - Clustering-based Taxonomy Construction
    - Hierarchical Topic Models
    - TaxoGen: Constructing Topical Concept Taxonomy by Adaptive Term Embedding and Clustering [KDD’18]
    - CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring [KDD’20]
    - NetTaxo: Automated Topic Taxonomy Construction from Text-Rich Network [WWW’20]
- Taxonomy Expansion

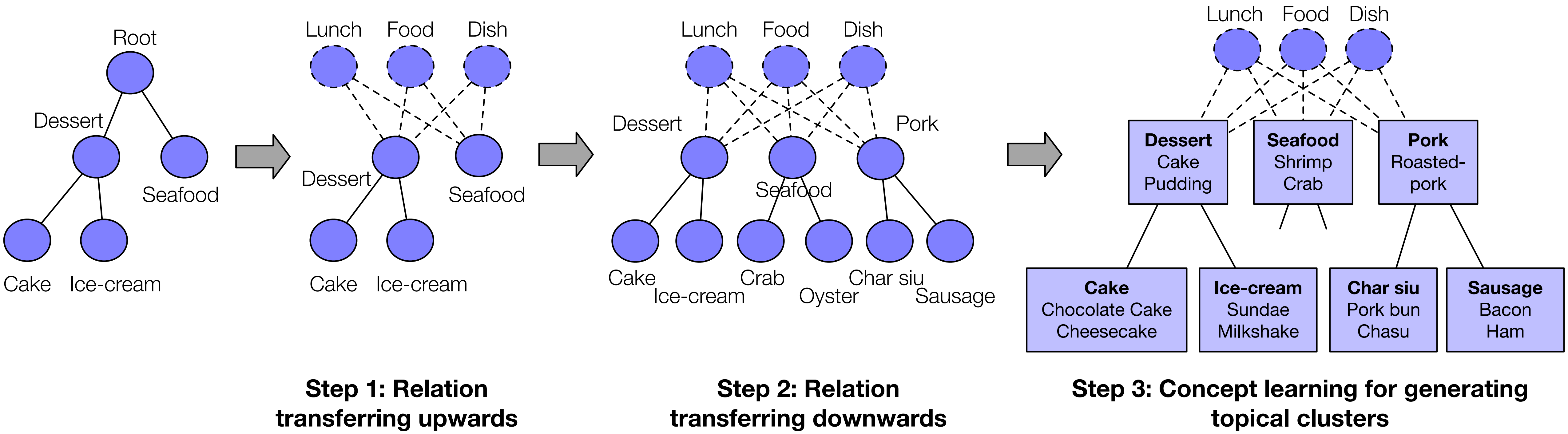
Seed-Guided Topical Taxonomy Construction

- Previous clustering-based methods generate generic topical taxonomies which cannot satisfy user's specific interest in certain areas and relations. Countless irrelevant terms and fixed "is-a" relations dominate the instance taxonomy.
- We study the problem of seed-guided topical taxonomy construction, where user gives a seed taxonomy as guidance, and a more complete topical taxonomy is generated from text corpus, with each node represented by a cluster of terms (topics).



A user might want to learn about concepts in a certain aspect (e.g., *food* or *research areas*) from a corpus. He wants to know more about other kinds of food.

CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring



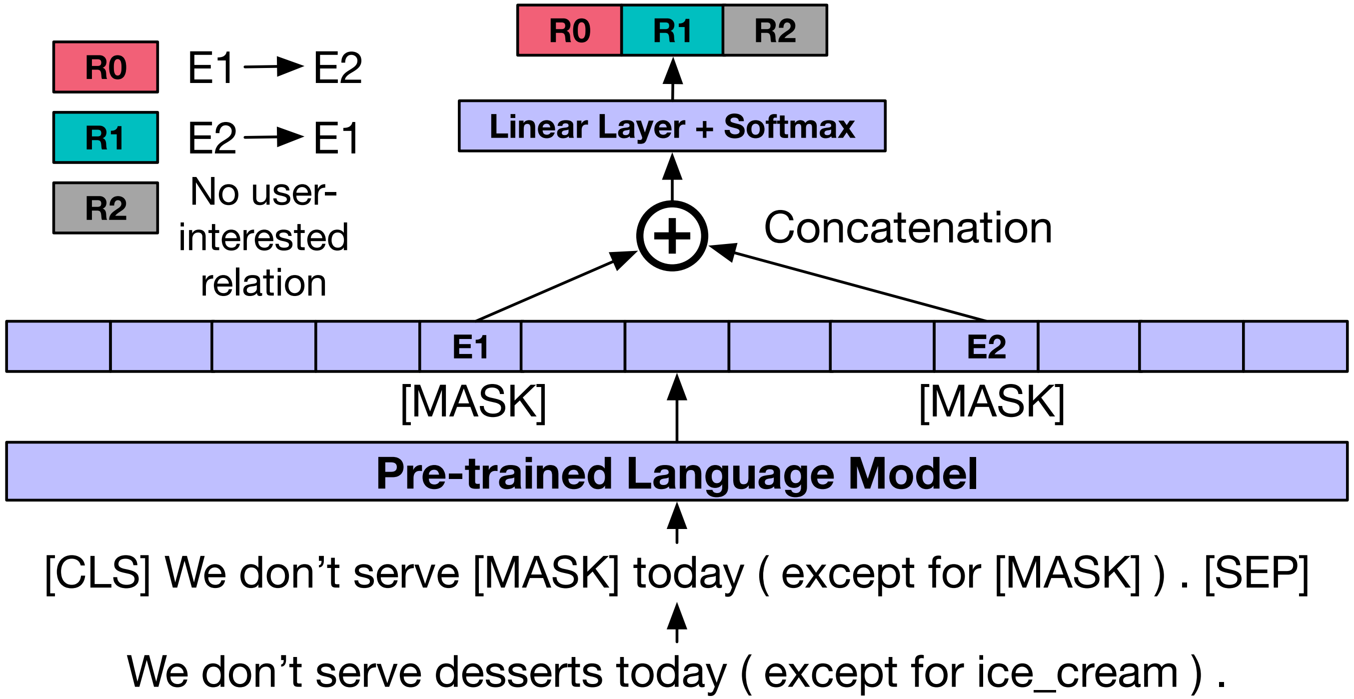
Step 1: Learn a relation classifier and transfer the relation upwards to **discover common root concepts** of existing topics.

Step 2: Transfer the relation downwards to **find new topics/subtopics** as child nodes of root/topics.

Step 3: Learn a discriminative embedding space to **find distinctive terms for each concept** node in the taxonomy.

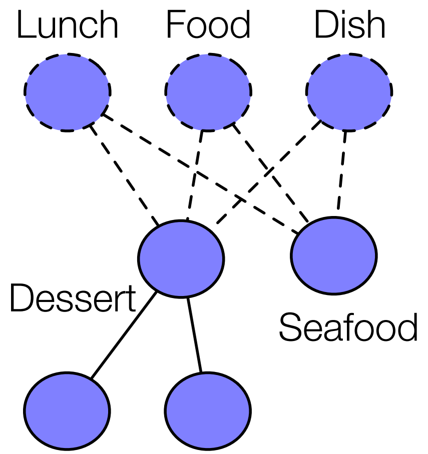
Relation Learning

- We adopt a pre-trained deep language model to learn a relation classifier with only the user-given parent-child (<p,c>) pairs.
- **Training samples**: We generate relation statements from the corpus as training samples for this classifier. We assume that if a pair of <p,c> co-occurs in a sentence in the corpus, then that sentence implies their relation.



Relation Transferring

- We first transfer the relation upwards to discover possible root nodes (e.g., “Lunch” and “Food”). This is because the root node would have more general contexts for us to find connections with potential new topics.

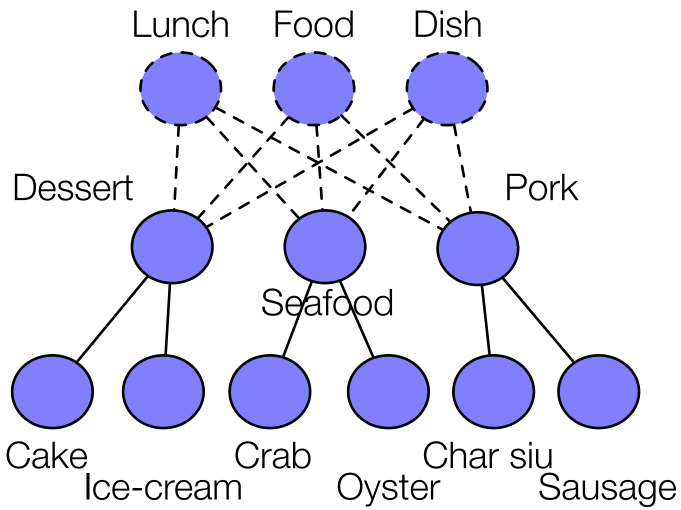


- We extract a list of parent nodes for each seed topic using the relation classifier. The common parent nodes shared by all user-given topics are treated as root nodes.
- To discover new topics (e.g, Pork), we transfer the relation downwards from these root nodes.

Relation Transferring

- We then transfer the relation downwards from each internal topic node to discover their subtopics.
- Since each candidate term has multiple mentions in the corpus, leading to multiple relation statements. We only count those confident predictions, and if the majority of these predictions judge the candidate term w as the child node of e, we retain the candidate term to be clustered later.

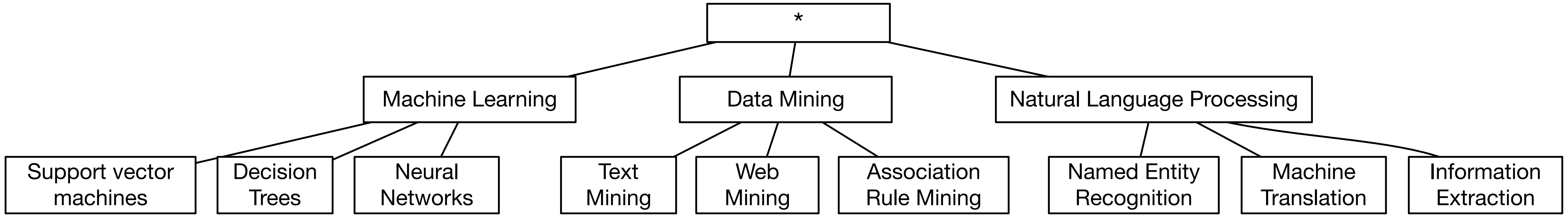
$$\text{Score}(\boldsymbol{e} \rightarrow \boldsymbol{w}) = \frac{\sum_{s_{e \rightarrow w}} \mathbb{1} \boxed{\left(KL\left(\boldsymbol{l} \| \boldsymbol{p}_w\right) > \delta\right)}}{\sum_{q \in Q} \sum_{s_q} \mathbb{1}\left(KL\left(\boldsymbol{l} \| \boldsymbol{p}_w\right) > \delta\right)}$$

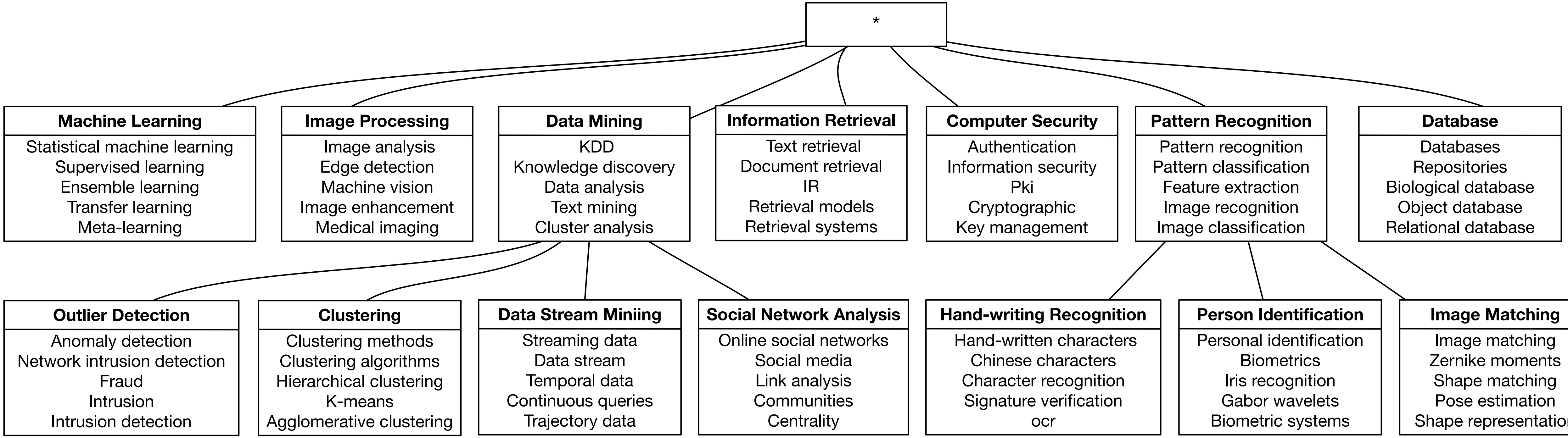


Concept Learning

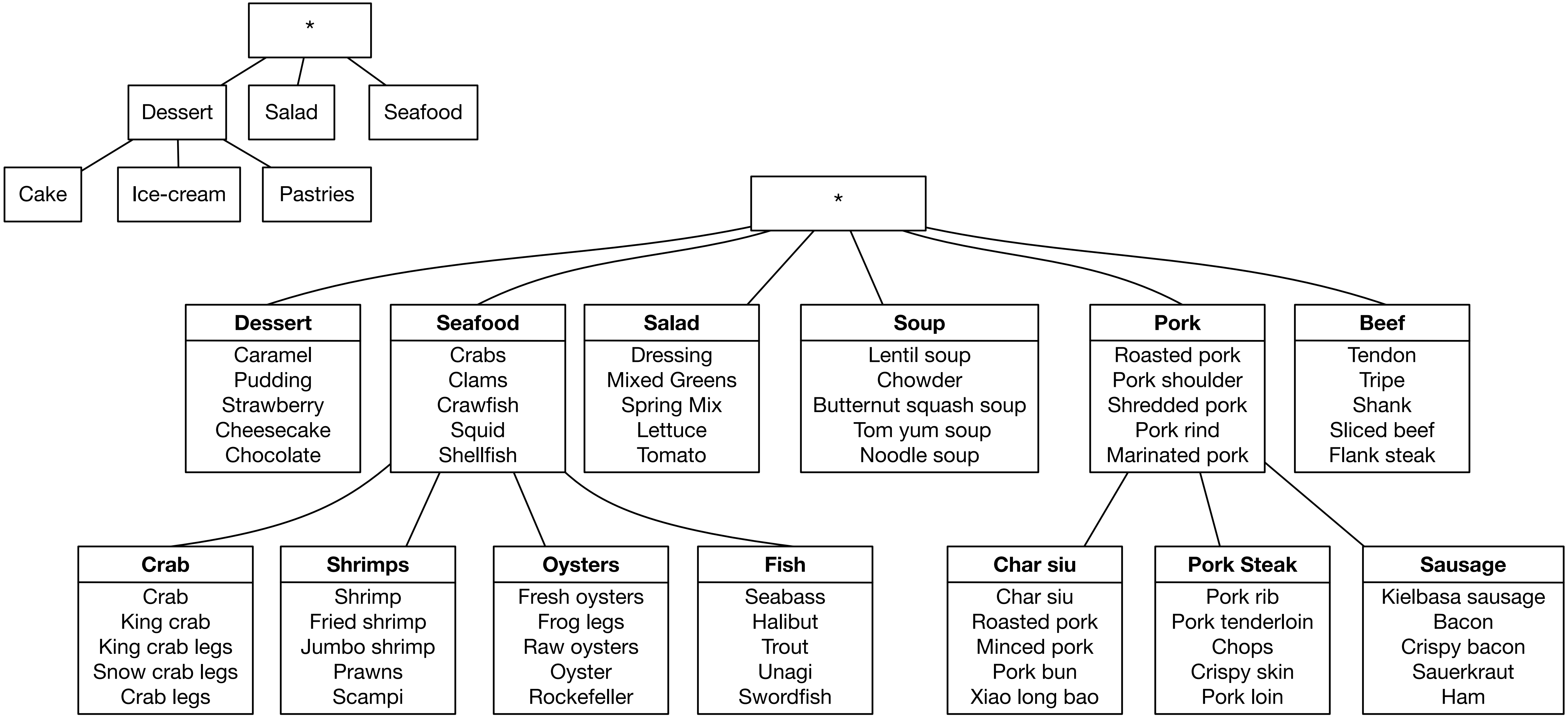
- Our concept learning module is used to learn a discriminative embedding space, so that each concept is surrounded by its representative terms. Within this embedding space, subtopic candidates are also clustered to form coherent subtopic nodes.
- Fine-grained concept names can be close in the embedding space, and directly using unsupervised word embedding might result in relevant but not distinctive terms (e.g., ``food" is relevant to both ``seafood" and ``dessert").
- Therefore, we leverage a **weakly-supervised text embedding framework** to discriminate these concepts in the embedding space, and this algorithm will be introduced in the next section.
- Subtopics should satisfy the following two constraints:
  - 1. must belong to representative words of that parent topic.
  - 2. must share parallel relations with given seed taxonomy.

Qualitative Results





Qualitative Results

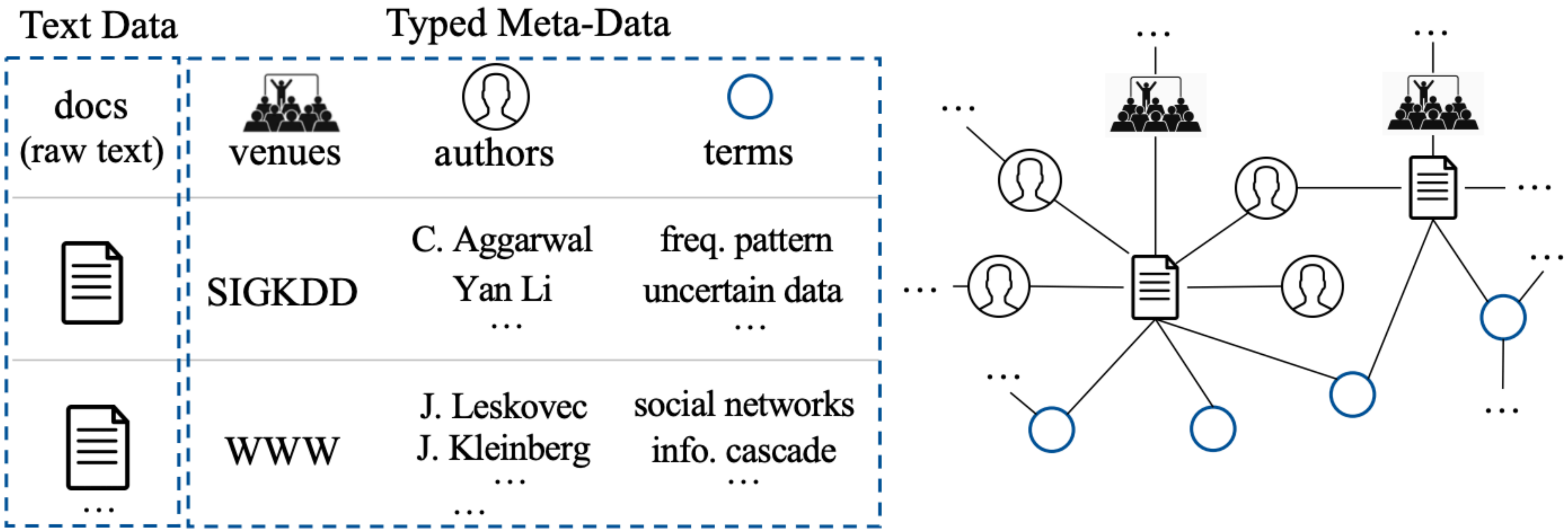


Outline

- ❑ Taxonomy Basics and Construction
- ❑ Parallel Concept Discovery: Entity Set Expansion
- ❑ Taxonomy Construction from Scratch
  - ❑ Instance-based Taxonomy Construction
  - ❑ Clustering-based Taxonomy Construction
    - ❑ Hierarchical Topic Models
    - ❑ TaxoGen: Constructing Topical Concept Taxonomy by Adaptive Term Embedding and Clustering [KDD'18]
    - ❑ CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring [KDD'20]
    - ❑ NetTaxo: Automated Topic Taxonomy Construction from Text-Rich Network [WWW'20]
- ❑ Taxonomy Expansion

NetTaxo: Automated Topic Taxonomy Construction from Text-Rich Network

- Besides leveraging unstructured text data, we can take the meta-data of documents into consideration and view the corpus as a text-rich network.
- Terms in scientific papers linked by the same venue or author can belong to the same research field, such as “social network” and “information cascade”.

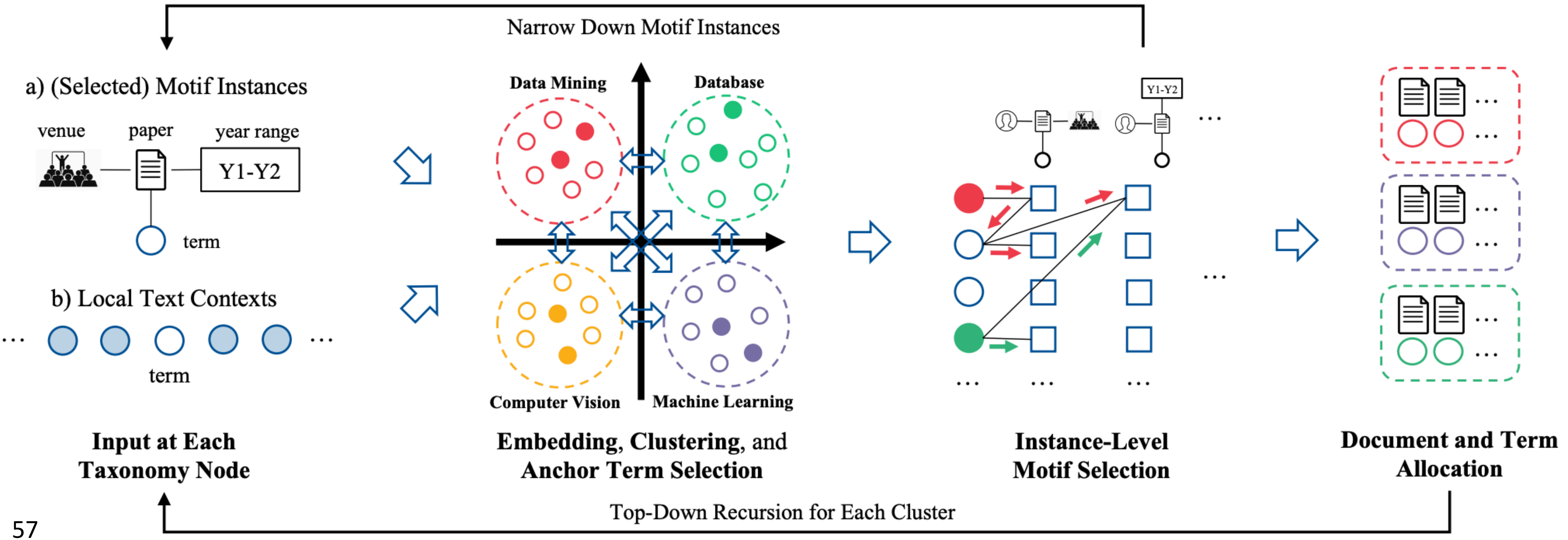


(a) An example digital collection of massive scientific papers.

(b) An text-rich network view of the example digital collection.

NetTaxo: Automated Topic Taxonomy Construction from Text-Rich Network

- A motif pattern Ω refers to a subgraph pattern at the meta level (i.e., every node is abstracted by its type).
- NetTaxo conducts a motif instance-level selection to pick the most informative network structures for better topic taxonomy construction.



Outline

- Taxonomy Basics and Construction
- Parallel Concept Discovery: Entity Set Expansion
- Taxonomy Construction from Scratch
- Taxonomy Expansion

Taxonomy Enrichment: Motivation

- Why taxonomy enrichment instead of construction from scratch?
  - Already have a decent taxonomy built by experts and used in production
  - Most common terms are covered
  - New items (thus new terms) incoming everyday, cannot afford to rebuild the whole taxonomy frequently
  - Downstream applications require stable taxonomies to organize knowledge

Taxonomy Enrichment: Motivation

- Why taxonomy enrichment instead of construction from scratch?
  - Already have a decent taxonomy built by experts and used in production
  - Most common terms are covered
  - New items (thus new terms) incoming everyday, cannot afford to rebuild the whole taxonomy frequently
  - Downstream applications require stable taxonomies to organize knowledge
- What is missing then?
  - Emerging terms take time for humans to discover
  - Long-tail / fine-grained terms (leaf nodes) are likely to be neglected

Three Assumptions in Taxonomy Expansion

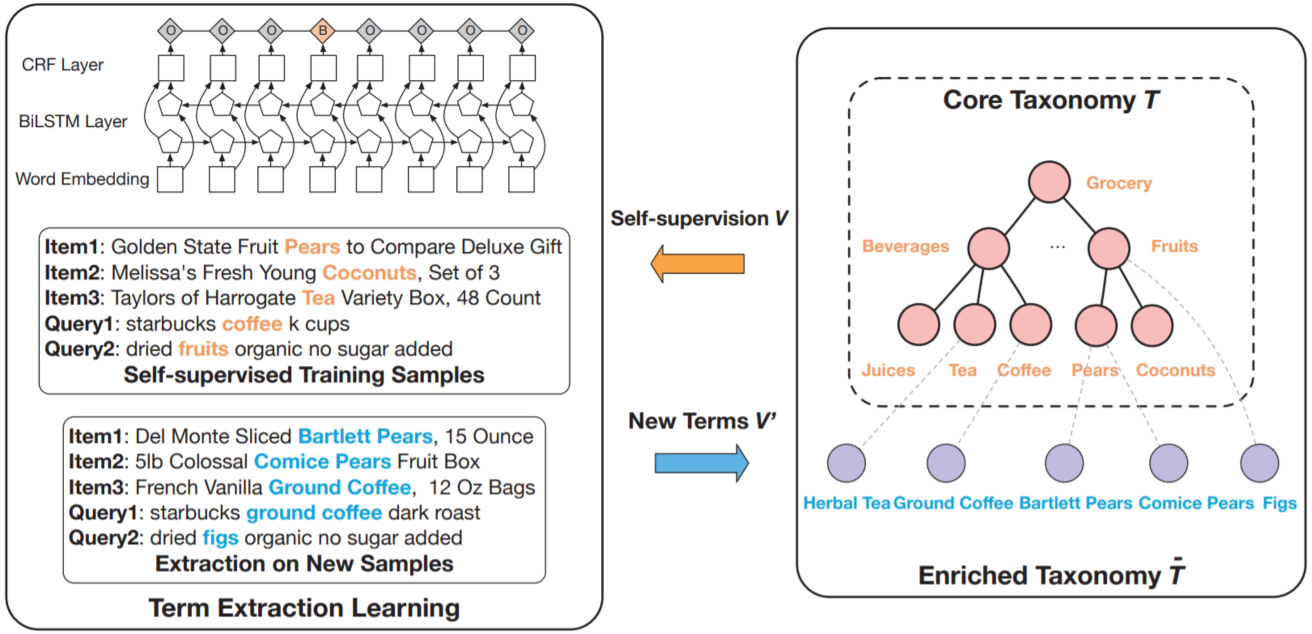
- First, we assume each concept will have a textual name
  - Therefore, we can get the *initial feature vector* of each concept in the existing taxonomy and of each new concept
- Second, we do not modify the existing taxonomy
  - Modification of existing relations happens less frequently and usually requires high cautiousness from human curators
- Third, we focus on finding parent node(s) of each new concept
  - New concept's parent node(s) typically appear in the existing taxonomy but its children node(s) may not exist the taxonomy

Taxonomy Expansion: Octet and TaxoExpan

- TaxoExpan: Self-supervised Taxonomy Expansion with Position-Enhanced Graph Neural Network [WWW' 20]
- Octet: Online Catalog Taxonomy Enrichment with Self-Supervision [KDD' 20]
- **Two steps** in solving the problem:
  - Self-supervised term extraction
    - Automatically **extracts emerging terms** from a target domain
  - Self-supervised term attachment
    - A multi-class classification to match a new node to its potential parent
    - Heterogenous sources of information (structural, semantic, and lexical) can be used

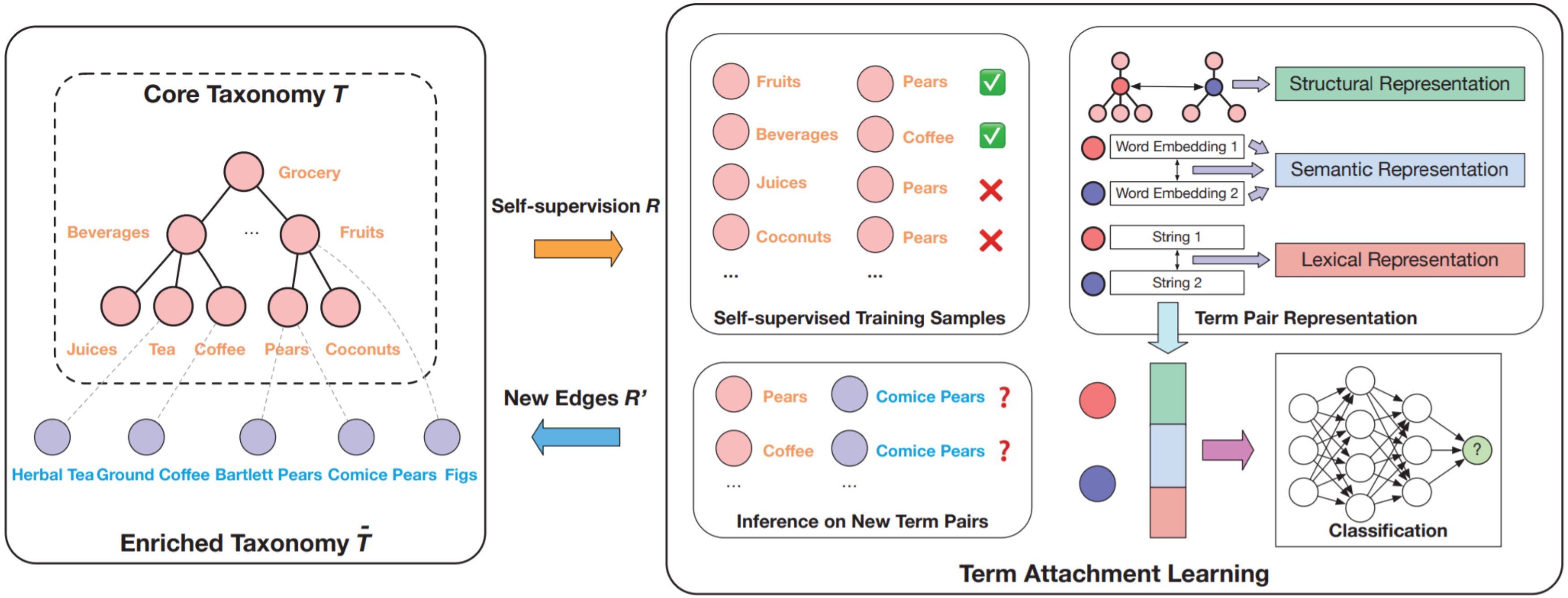
Self-supervised Term Extraction

- Octet adapts state-of-the-art sequence labeling method w. BiLSTM-CRF + Attention (Zheng et al, KDD'18)
- **Self-supervision**
  - Use existing nodes as desired terms to be extracted
  - No human efforts needed



Self-supervised Term Attachment

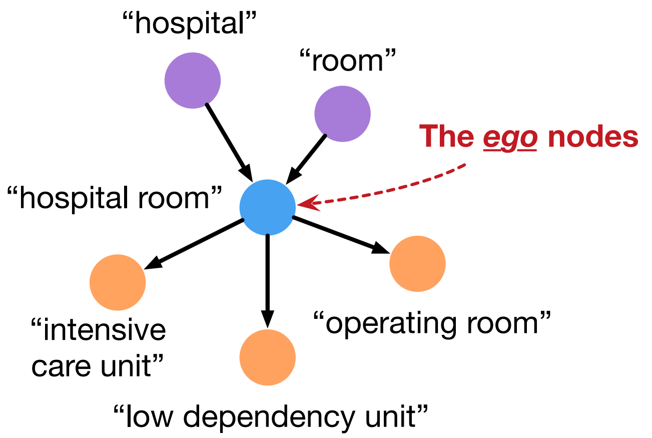
- **Octet** combines structural, semantic and lexical representation to learn a term-pair representation and feeds it into a two-layer network.
- Structural Representation: Interactions among taxonomy nodes, items, and queries
- Semantic Representation: Word embedding-based features
- Lexical Representation :Surface string-level features (Ends with, Contains, Suffix match, …)

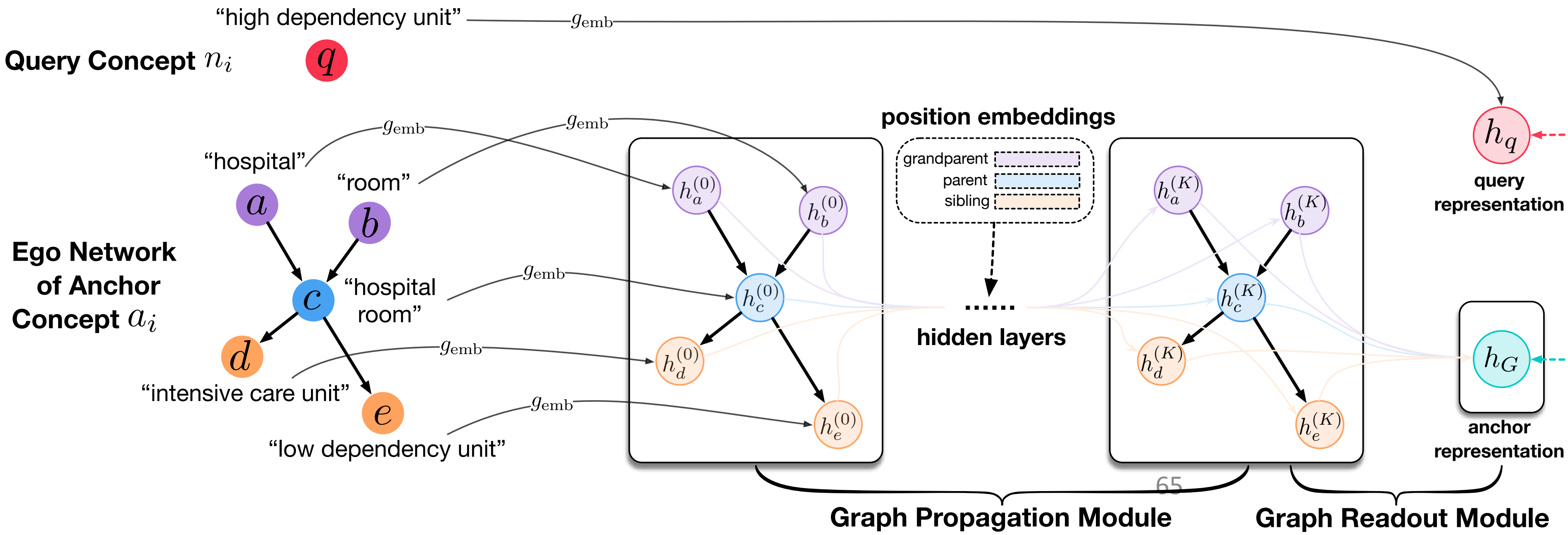


Self-supervised Term Attachment

- **TaxoExpan** uses a matching score for each <*query*, *anchor*> pair to indicate how likely the *anchor concept* is the parent of *query concept*
- Key ideas:
  - Representing the *anchor concept* using its ego network (egonet)
  - Adding position information (relative to the *query concept*) into this egonet

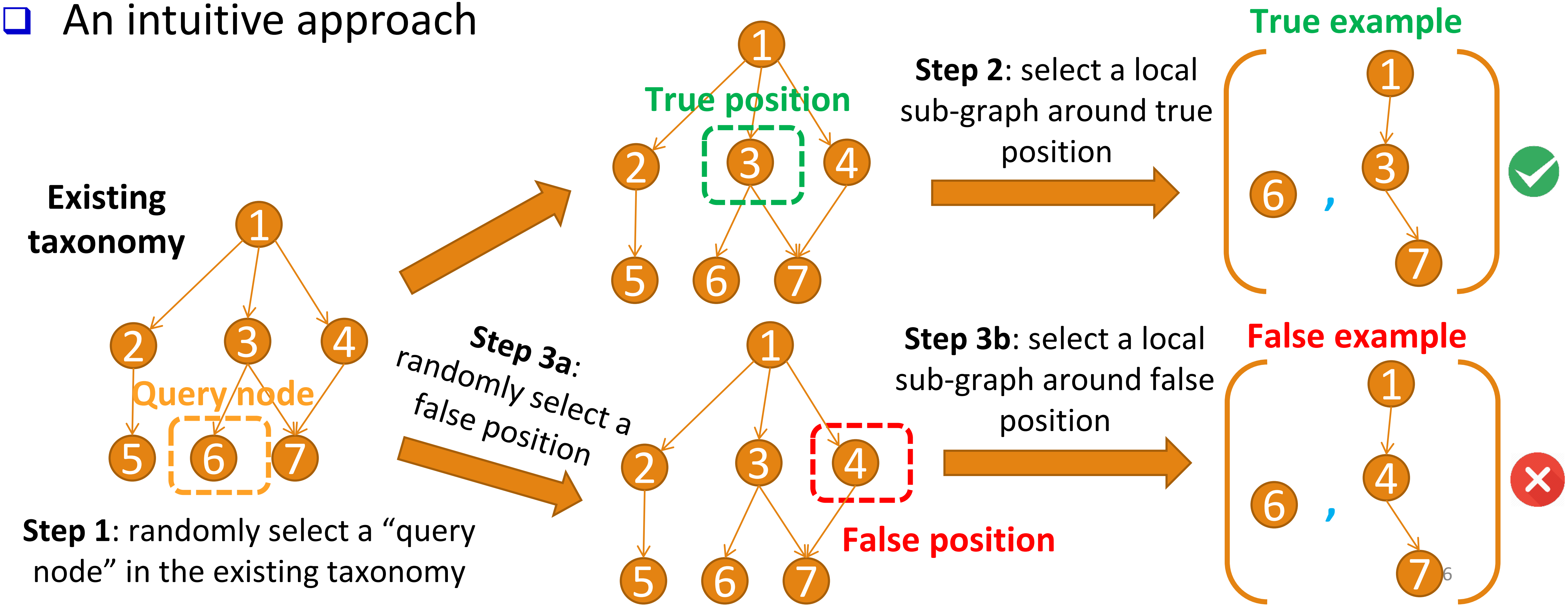
Query: "high dependency unit"





Leveraging Existing Taxonomy for Self-supervised Learning

- How to learn model parameters without relying on massive human-labeled data?
- An intuitive approach



Octet Framework Analysis

Performance Trade-off

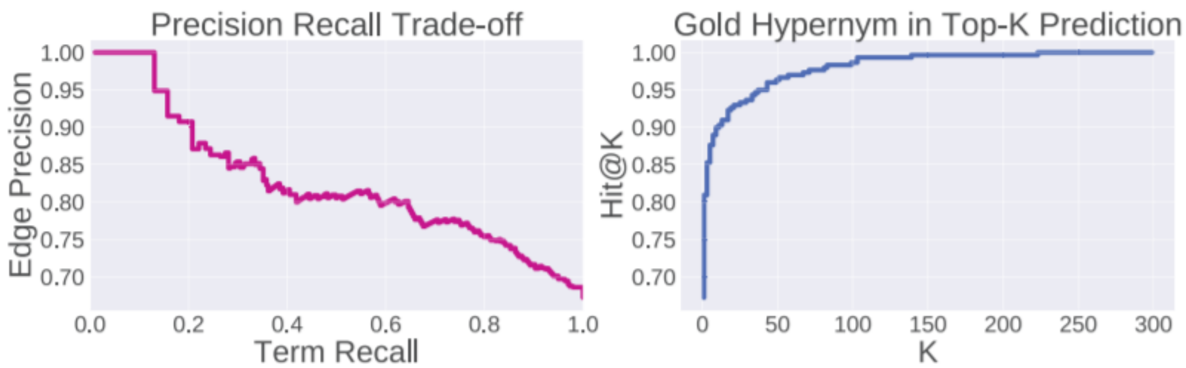


Figure 4: The precision recall trade-off *(Left)* and performance of term attachment in Hit@K *(Right)*.

how many terms can be attached if a specific precision of term attachment is required?

What if we relax the task to top-K prediction (instead of top-1 in Edge-F1)?

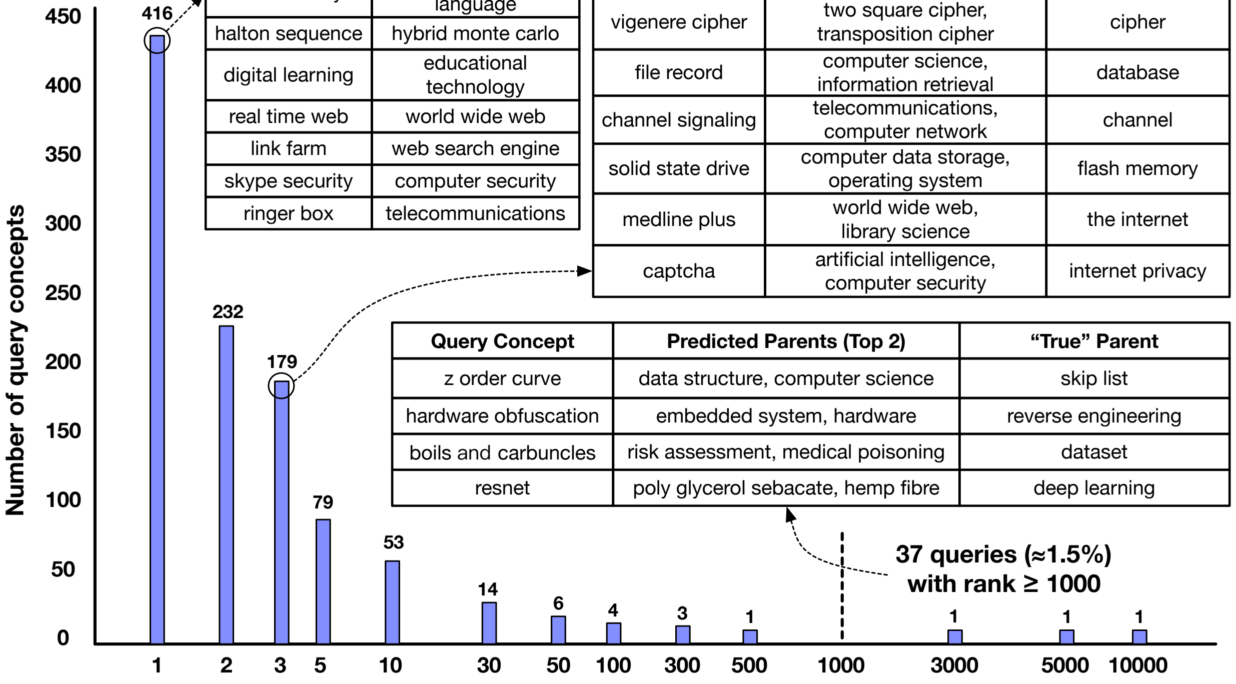
Case studies

Table 10: Case studies of term attachment. Correct and incorrect cases are marked in green and red, respectively.

Query Term	Gold Hypernym	Top-3 Predictions
fresh cut carnations	fresh cut flowers	fresh cut flowers, fresh cut root vegetables, fresh cut & packaged fruits
tilapia	fresh fish	fresh fish, liquor & spirits, fresh shellfish
bock beers	lager & pilsner beers	**W/O structural representation:** ales, beer, tea beverages **Full Model:** lager & pilsner beers, porter & stout beers, tea beverages
fresh russet potatoes	fresh potatoes & yams	fresh fingerlings & baby potatoes, fresh root vegetables, fresh herbs
pinto beans	dried beans	canned beans, fresh peas & beans, single herbs & spices

TaxoExpan Framework Analysis

- Case studies on MAG-CS and MAG-Full datasets

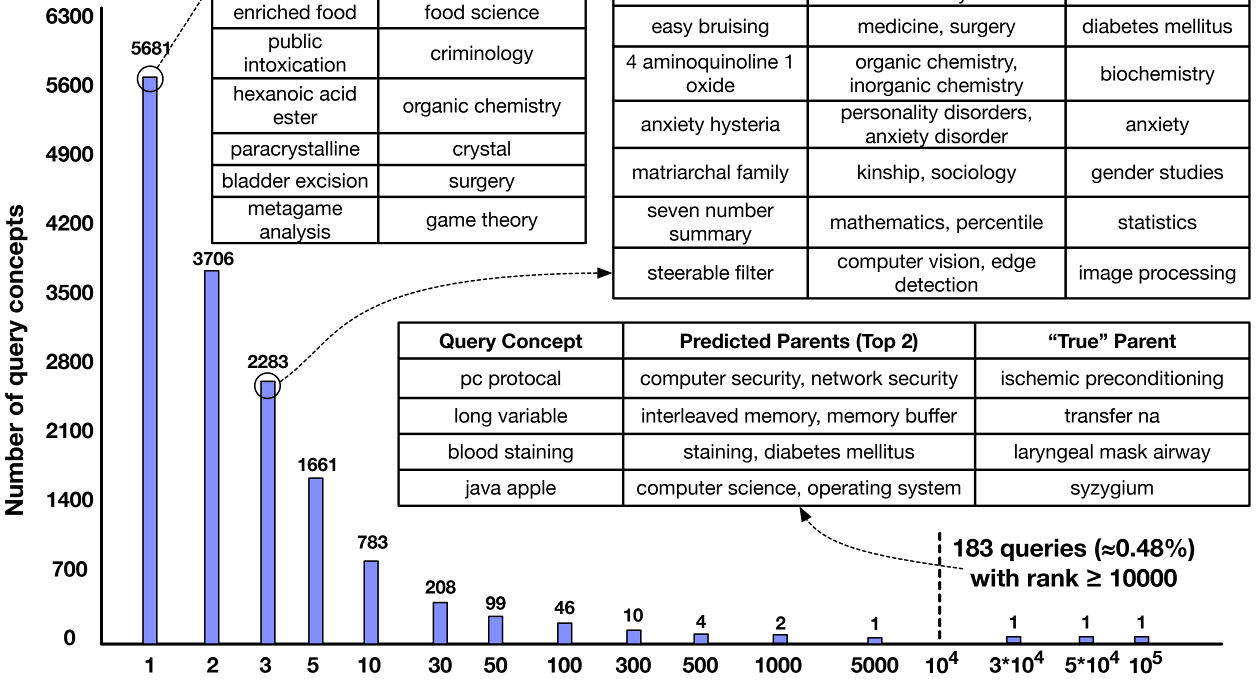


Query Concept	Predicted Parent = "True" Parent
archival science	library science
static library	programming language
halton sequence	hybrid monte carlo
digital learning	educational technology
real time web	world wide web
link farm	web search engine
skype security	computer security
ringer box	telecommunications

Query Concept	Predicted Parents (Top 2)	"True" Parent
email hacking	internet privacy, hacker	computer security
social graph	world wide web, the internet	social network
vigenere cipher	two square cipher, transposition cipher	cipher
file record	computer science, information retrieval	database
channel signaling	telecommunications, computer network	channel
solid state drive	computer data storage, operating system	flash memory
medline plus	world wide web, library science	the internet
captcha	artificial intelligence, computer security	internet privacy

Query Concept	Predicted Parents (Top 2)	"True" Parent
z order curve	data structure, computer science	skip list
hardware obfuscation	embedded system, hardware	reverse engineering
boils and carbuncles	risk assessment, medical poisoning	dataset
resnet	poly glycerol sebacate, hemp fibre	deep learning

(a) MAG-CS Dataset (totally 2450 query concepts)



Query Concept	Predicted Parent = "True" Parent
hindi language	linguistics
dyssodia	botany
enriched food	food science
public intoxication	criminology
hexanoic acid ester	organic chemistry
paracrystalline	crystal
bladder excision	surgery
metagame analysis	game theory

Query Concept	Predicted Parents (Top 2)	"True" Parent
syndactyla	ecology, biology	zoology
m matrix	symmetric matrix, nonlinear system	matrix
easy bruising	medicine, surgery	diabetes mellitus
4 aminoquinoline 1 oxide	organic chemistry, inorganic chemistry	biochemistry
anxiety hysteria	personality disorders, anxiety disorder	anxiety
matriarchal family	kinship, sociology	gender studies
seven number summary	mathematics, percentile	statistics
steerable filter	computer vision, edge detection	image processing

Query Concept	Predicted Parents (Top 2)	"True" Parent
pc protocal	computer security, network security	ischemic preconditioning
long variable	interleaved memory, memory buffer	transfer na
blood staining	staining, diabetes mellitus	laryngeal mask airway
java apple	computer science, operating system	syzygium

(b) MAG-Full Dataset (totally 37804 query concepts)

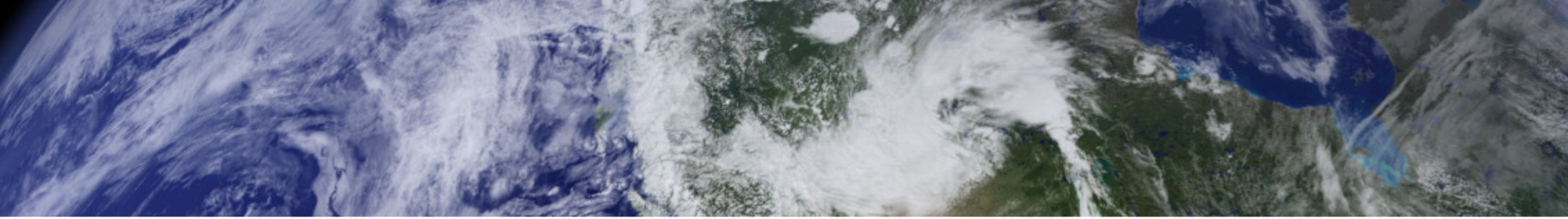
68

References: Concept Expansion + Hierarchical Topic Modeling

- Xin Rong, Zhe Chen, Qiaozhu Mei, and Eytan Adar. “EgoSet: Exploiting Word Ego-networks and User-generated Ontology for Multifaceted Set Expansion.” (WSDM’16)
- Jiaming Shen, Zeqiu Wu, Dongming Lei, Jingbo Shang, Xiang Ren, Jiawei Han, "SetExpan: Corpus-based Set Expansion via Context Feature Selection and Rank Ensemble", (ECMLPKDD'17)
- Jiaxin Huang, Yiqing Xie, Yu Meng, Jiaming Shen, Yunyi Zhang and Jiawei Han, ”Guiding Corpus-based Set Expansion by Auxiliary Sets Generation and Co-Expansion”, (WWW’20)
- Yunyi Zhang, Jiaming Shen, Jingbo Shang and Jiawei Han, “Empower Entity Set Expansion via Language Model Probing”, (ACL’20)
- David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. 2003. Hierarchical Topic Models and the Nested Chinese Restaurant Process. In NIPS.
- David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2010. The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies. Journal of ACM (2010).
- Wei Li and Andrew D McCallum. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations. In ICML.
- D. M. Mimno, W. Li, and A. McCallum. Mixtures of hierarchical topics with pachinko allocation. In ICML, pages 633–640, 2007

References: Taxonomy Construction and Expansion

- Li, Wei, David M. Blei and Andrew McCallum. “Nonparametric Bayes Pachinko Allocation.” *UAI* (2007)
- David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. ICML (2009)
- Su-Jin Shin and Il-Chul Moon. 2017. Guided HTM: Hierarchical Topic Model with Dirichlet Forest Priors. TKDE (2017)
- Zhang, Chao, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian M. Sadler, Michelle Vanni and Jiawei Han. “TaxoGen : Unsupervised Topic Taxonomy Construction by Adaptive Term Embedding and Clustering.” KDD (2018)
- Jiaming Shen, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle T. Vanni, Brian M. Sadler and Jiawei Han, "HiExpan: Task-Guided Taxonomy Construction by Hierarchical Tree Expansion", KDD (2018)
- Jiaxin Huang, Yiqing Xie, Yu Meng, Yunyi Zhang and Jiawei Han, “CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring”, KDD (2020)
- Jingbo Shang, Xinyang Zhang, Liyuan Liu, Sha Li and Jiawei Han, “NetTaxo: Automated Topic Taxonomy Construction from Large-Scale Text-Rich Network”, (WWW’20)
- Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang and Jiawei Han ”TaxoExpan: Self-supervised Taxonomy Expansion with Position-Enhanced Graph Neural Network”, (WWW’20)
- Yuning Mao, Tong Zhao, Andrey Kan, Chenwei Zhang, Xin Luna Dong, Christos Faloutsos and Jiawei Han, “Octet: Online Catalog Taxonomy Enrichment with Self-Supervision”, (KDD’20)



Q&A