Part II: Taxonomy Construction and Enrichment

Automated Mining of Structured Knowledge from Text in the Era of Large Language Models

Yunyi Zhang, Ming Zhong, Siru Ouyang, Yizhu Jiao, Sizhe Zhou, Linyi Ding, Jiawei Han

Computer Science, University of Illinois Urbana-Champaign

KDD 2024 Tutorial, Aug 25, 2024

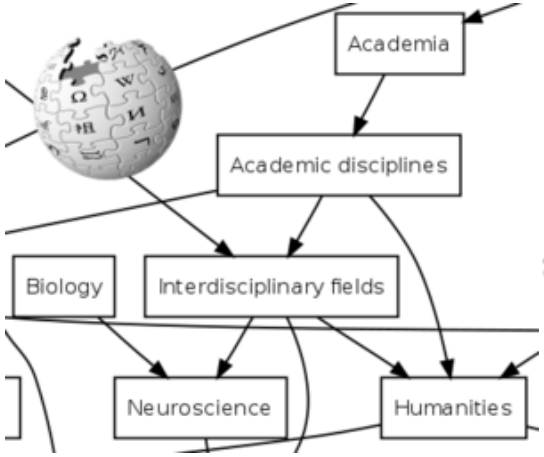
Tutorial Website:

Outline

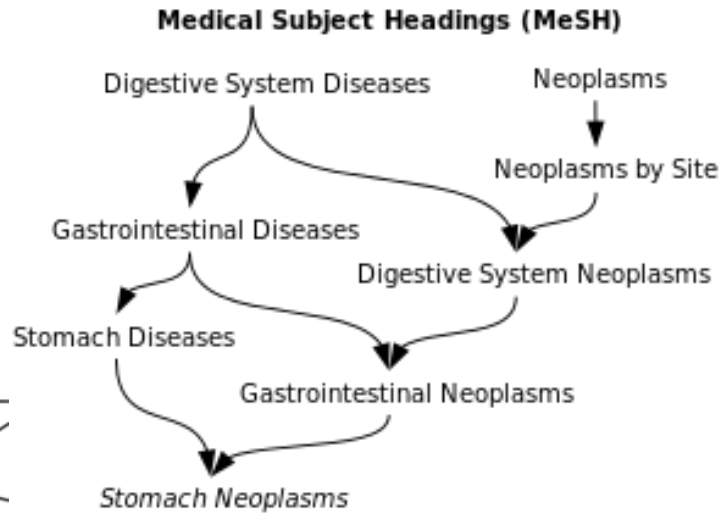
- Taxonomy basics and why do we need it?
- Taxonomy Construction
  - CGExpan [ACL'20], CoRel [KDD'20], TaxoCom [WWW'22], Chain-of-Layer [arXiv'24]
- Taxonomy Expansion
  - TaxoExpan [WWW'20], BoxTaxo [WWW'23], TaxoInstruct [arXiv'24]
- Taxonomy Enrichment
  - CatE [WWW'20], JoSH [KDD'20], SeedTopicMine [WSDM'23]

What Is Taxonomy?

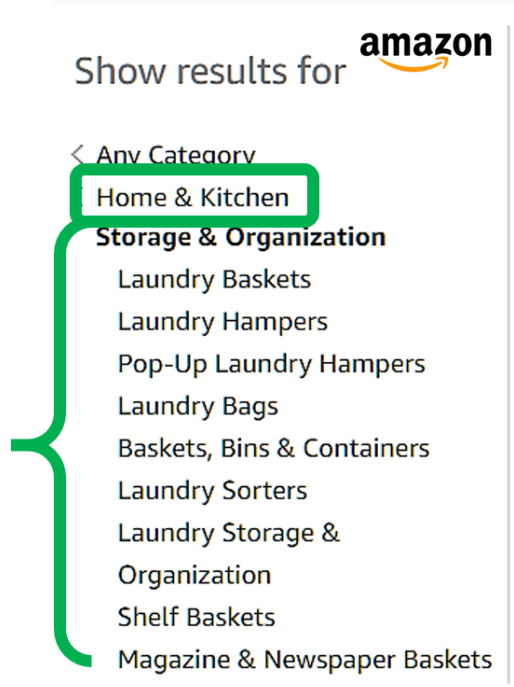
- Taxonomy is a hierarchical (or DAG) organization of concepts
  - Ex.: Wikipedia category, ACM CCS Classification System, Medical Subject Heading (MeSH), Amazon Product Category, Yelp Category List, WordNet, …



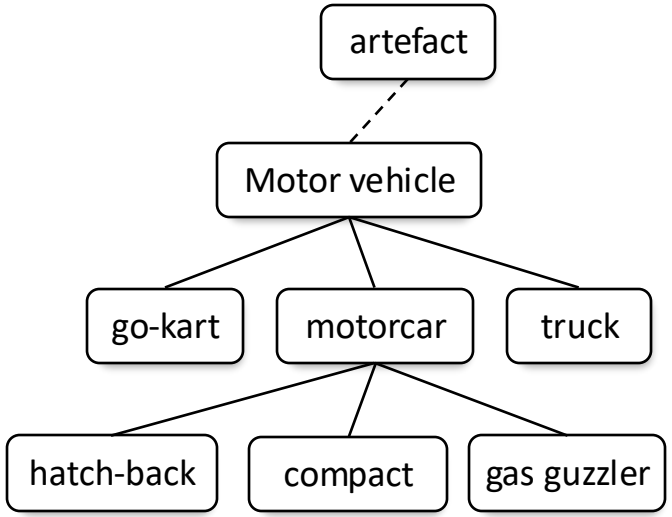
Wikipedia Category



MeSH: PubMed



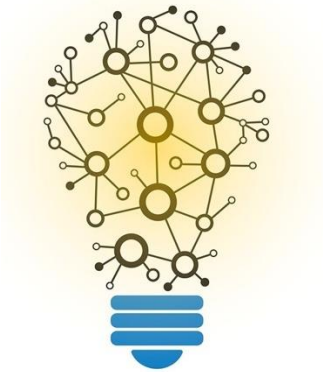
Amazon Product Category



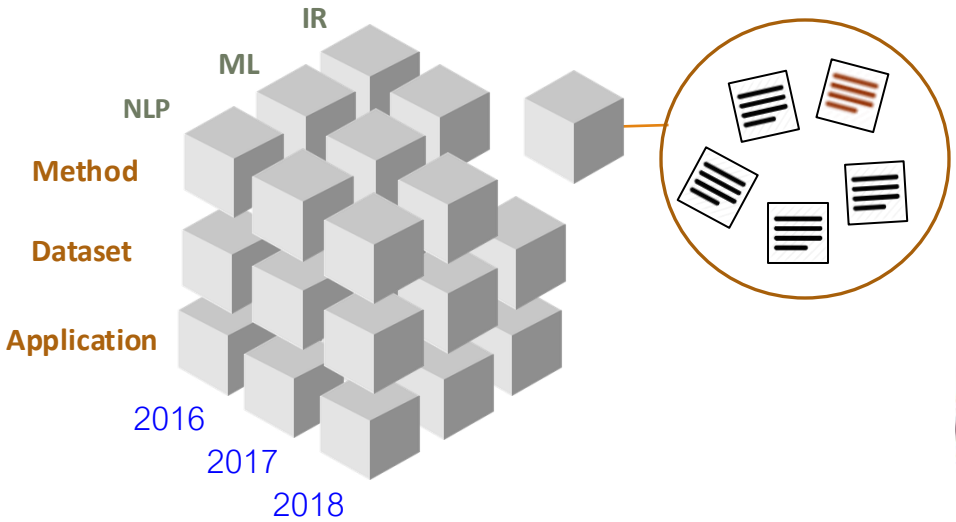
WordNet

Why Do We Need Taxonomy?

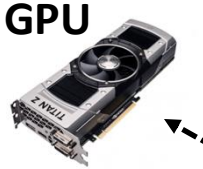
- Taxonomy can benefit many knowledge-rich applications
  - Text Understanding
  - Knowledge Organization
  - Document Categorization
  - Recommender System
  - ……



Corpus



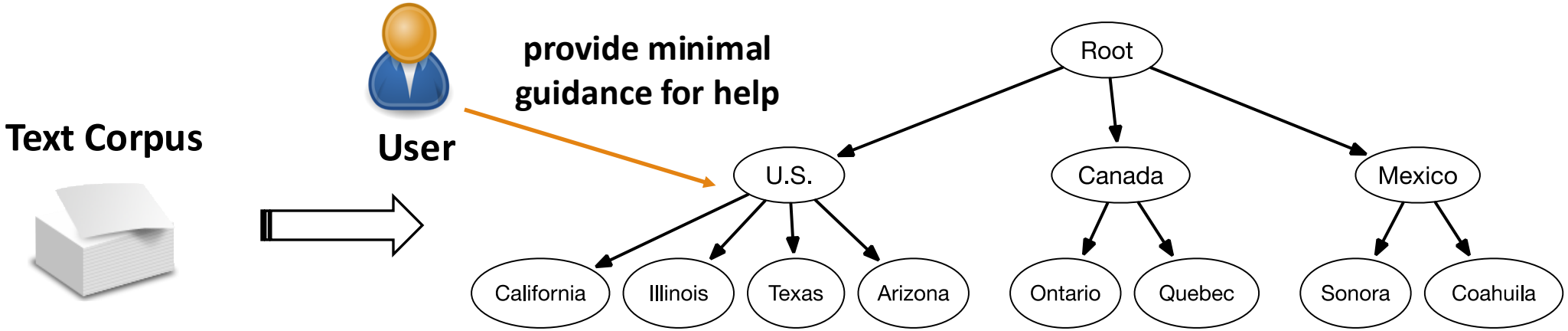
Multi-dimensional Corpus Index



How to Get Taxonomy: Manual vs. Automated?

- Manual Curation
  - Time-consuming
  - Tremendous **human (experts) efforts**
  - Examples
    - Medical Subject Heading (MeSH): 60+ years
    - ACM CCS Classification System: 40+ years
    - IEEE Taxonomy: 40+ years

- Automated taxonomy construction/enhancement from **text** is in great demand



Issues Related to Taxonomy Construction

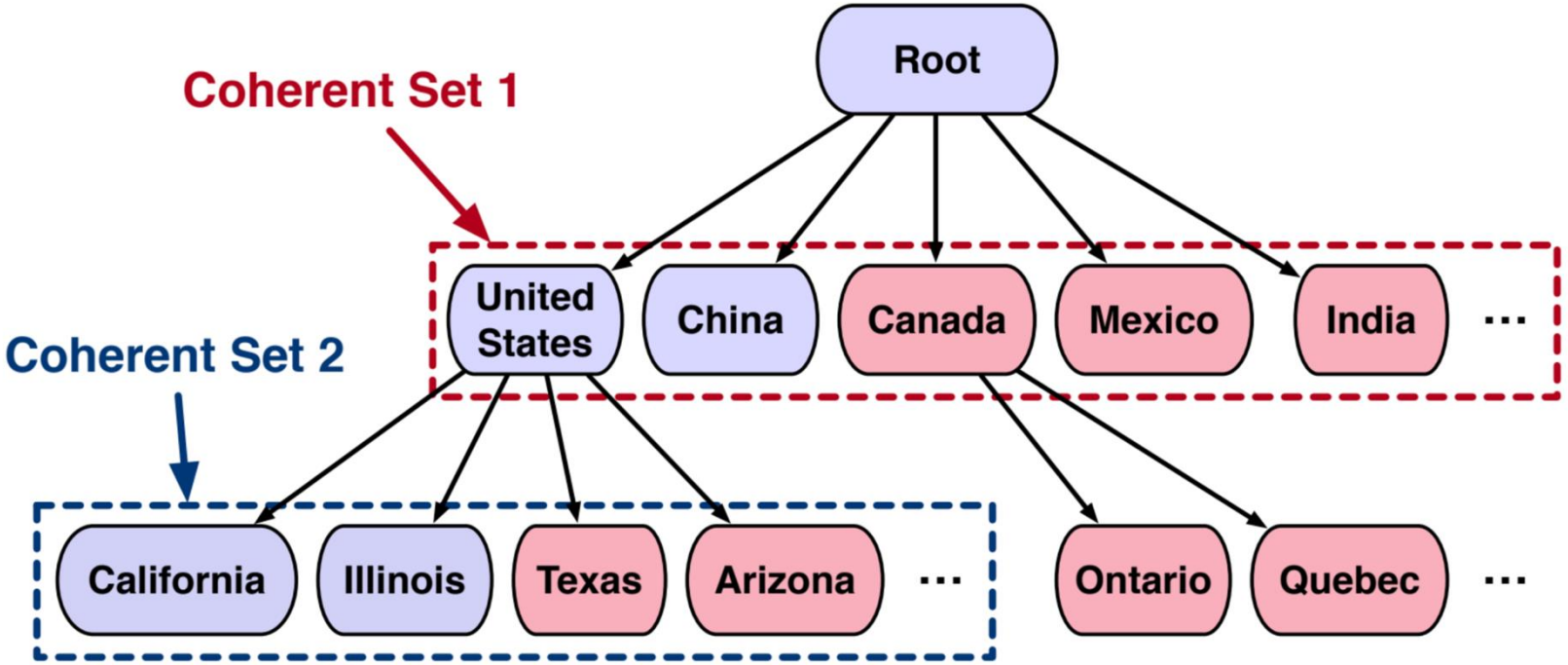
- Taxonomy Construction (with Minimal User Guidance)
  - User give a seed skeleton taxonomy (in a small scale) and text corpus to build a taxonomy organized by certain relations
- Taxonomy Expansion
  - Update an already constructed taxonomy by adding new items on the existing taxonomy
- Taxonomy Enrichment
  - Given a taxonomy, enrich each node with more indicative terms that distinctively represent this node

Outline

- Taxonomy basics and why do we need it?
- Taxonomy Construction
  - CGExpan [ACL'20], CoRel [KDD'20], TaxoCom [WWW'22], Chain-of-Layer [arXiv'24]
- Taxonomy Expansion
  - TaxoExpan [WWW'20], BoxTaxo [WWW'23], TaxoInstruct [arXiv'24]
- Taxonomy Enrichment
  - CatE [WWW'20], JoSH [KDD'20], SeedTopicMine [WSDM'23]

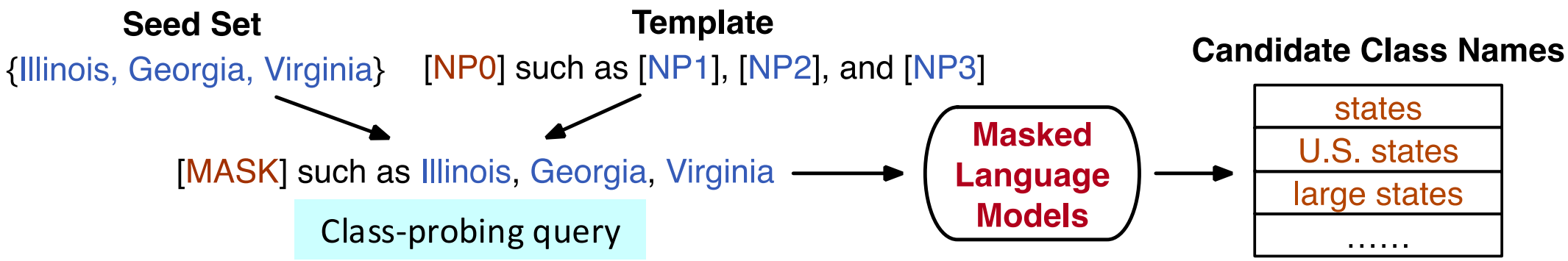
Entity Set Expansion

- Observation: in a taxonomy, each set of siblings form a coherent set
- An important subtask of taxonomy construction: Entity Set Expansion
  - Given a small set of seed entities (e.g., only 3-5 seeds), find more entities belonging to the same semantic class



CGExpan: Probing Language Model for Guidance

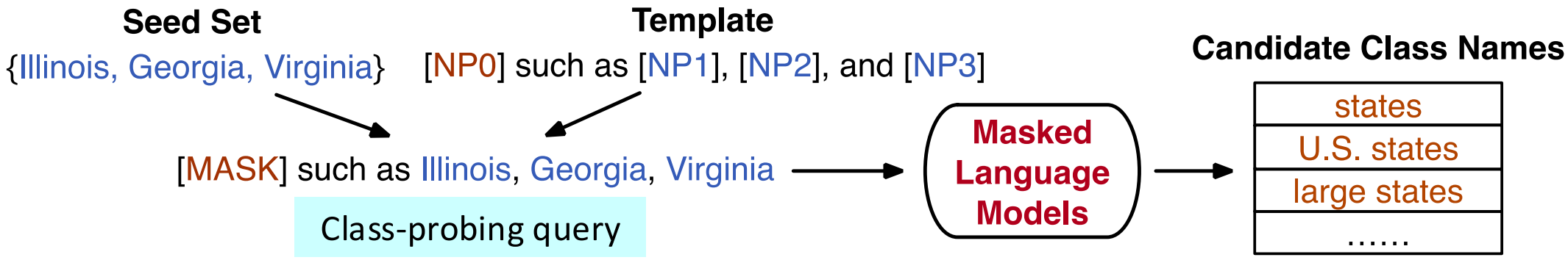
- Generating the **target class names** by probing a language model



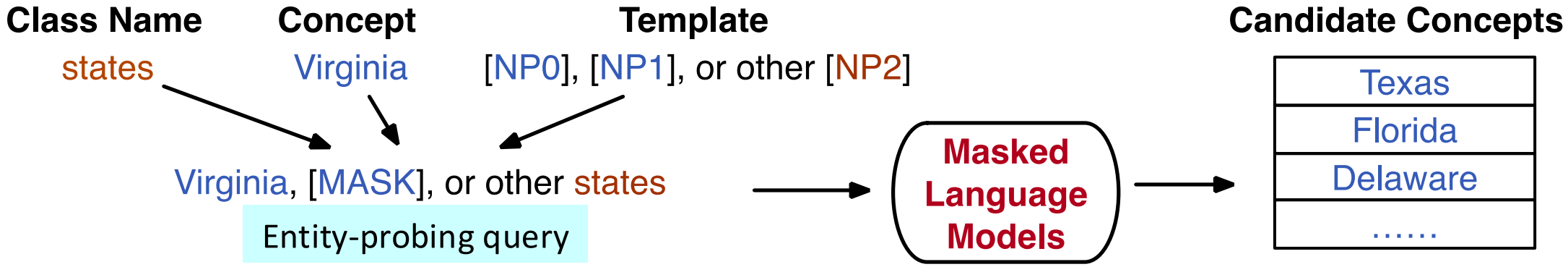
Yunyi Zhang, Jiaming Shen, Jingbo Shang, Jiawei Han, "Empower Entity Set Expansion via Language Model Probing", ACL'20

CGExpan: Probing Language Model for Guidance

- Generating the **target class names** by probing a language model



- Preventing concept drifting with **Class Guided Expansion (CGExpan)**



Yunyi Zhang, Jiaming Shen, Jingbo Shang, Jiawei Han, "Empower Entity Set Expansion via Language Model Probing", ACL'20

CGExpan: Quantitative Results

	Methods	Wikipedia		APR	
		MAP@20	MAP@50	MAP@20	MAP@50
Bootstrapping	**Egoset** (Rong et al., WSDM'16)	0.877	0.745	0.710	0.570
Bootstrapping	**MCTS** (Yan et al., ACL'19)	0.930	0.790	0.900	0.810
One time text ranking	**SetExpander** (Mamou et al., EMNLP'18)	0.439	0.321	0.208	0.120
One time text ranking	**CaSE** (Yu et al., SIGIR'19)	0.806	0.588	0.494	0.330
Our solutions	**SetExpan** (ECMLPKDD'17)	0.921	0.720	0.763	0.639
Our solutions	**SetCoExpan** (WWW'20)	0.964	**0.905**	0.915	0.830
Our solutions	**CGExpan** (ACL'20)	**0.978**	0.902	**0.990**	**0.955**

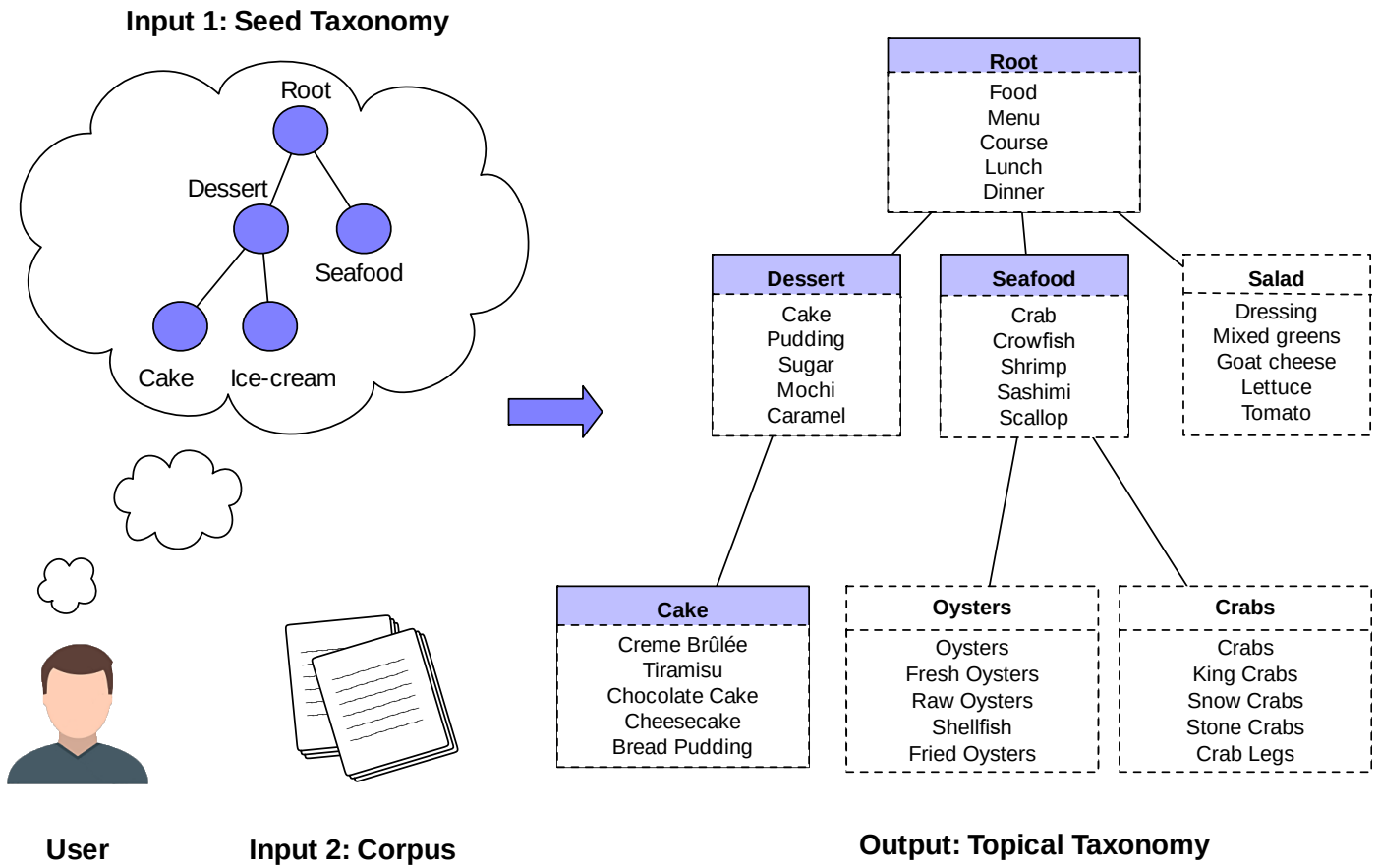
MAP@K: Mean Average Precision truncated at position K

- **vs. Bootstrapping**: better address the concept drifting issue
- **vs. One time text ranking**: better leverage seed supervision iteratively

Wikipedia: 1.5M Wikipedia article sentences (20 semantic classes manually labeled for evaluation);
APR: 1.1M news article sentences (40 semantic classes manually labeled for evaluation)

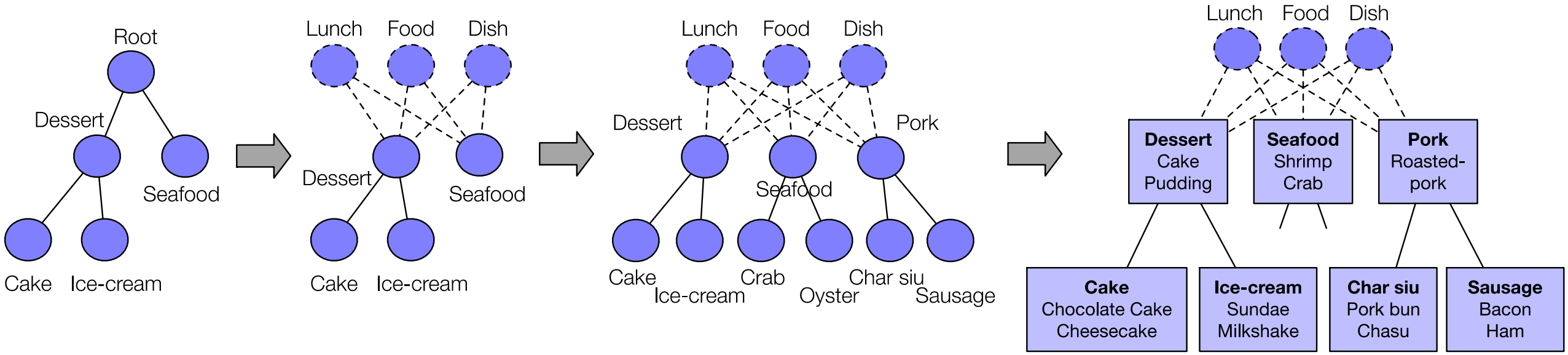
Seed-Guided Topical Taxonomy Construction

- User gives a seed taxonomy as guidance
- 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: Relation transferring upwards

Step 2: Relation transferring downwards

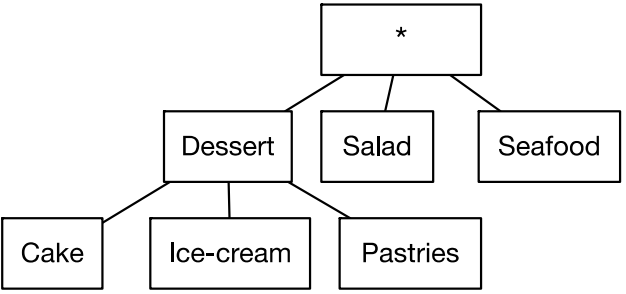
Step 3: Concept learning for generating topical clusters

Three Steps:

1. Learn a relation classifier and transfer the relation upwards to **discover common root concepts** of existing topics
2. Transfer the relation downwards to **find new topics/subtopics** as child nodes of root/topics
3. Learn a discriminative embedding space to **find distinctive terms for each concept** node in the taxonomy

Jiaxin Huang, Yiqing Xie, Yu Meng, Yunyi Zhang and Jiawei Han, "CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring", KDD'20

Qualitative and Quantitative Results



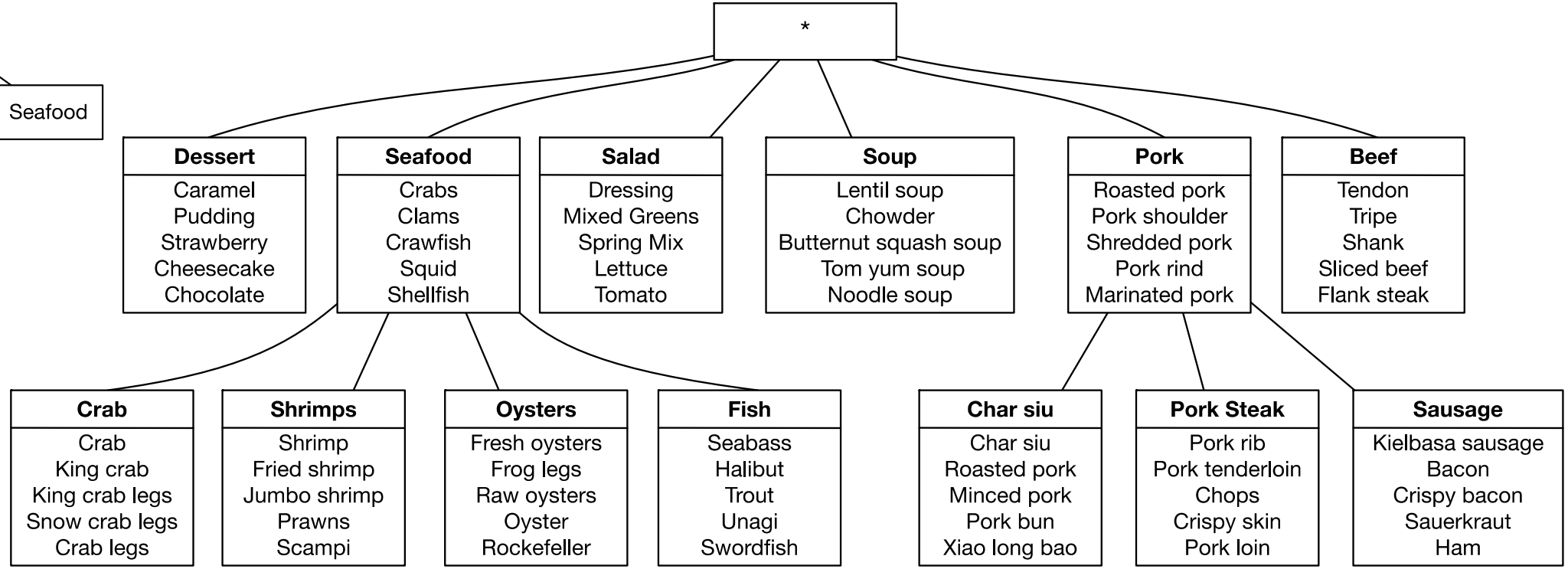
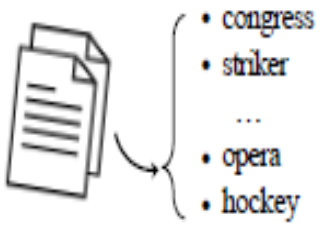


Table 5: Quantitative evaluation on topical taxonomies.

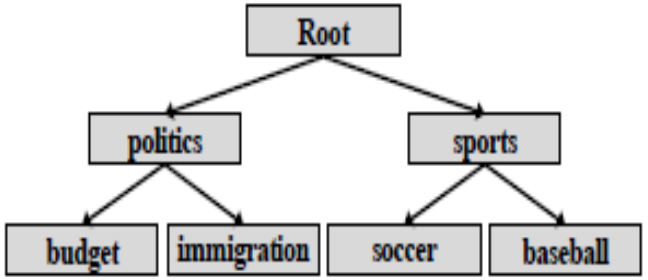
Methods	DBLP					Yelp				
	TC	SD	Precision$_r$	Recall$_r$	F1-score$_r$	TC	SD	Precision$_r$	Recall$_r$	F1-score$_r$
HLDA	0.582	0.981	0.188	0.577	0.283	0.517	0.991	0.135	0.387	0.200
HPAM	0.557	0.905	0.362	0.538	0.433	0.687	0.898	0.173	0.615	0.271
TaxoGen	0.720	0.979	0.450	0.429	0.439	0.563	0.965	0.267	0.381	0.314
Hi-Expan + CoL.	0.819	0.996	0.676	0.532	0.595	0.815	**1.000**	0.429	0.677	0.525
CoRel	**0.855**	**1.000**	**0.730**	**0.607**	**0.663**	**0.825**	**1.000**	**0.564**	**0.710**	**0.629**

TaxoCom: Topic Taxonomy Completion with Hierarchical Discovery of Novel Topic Clusters

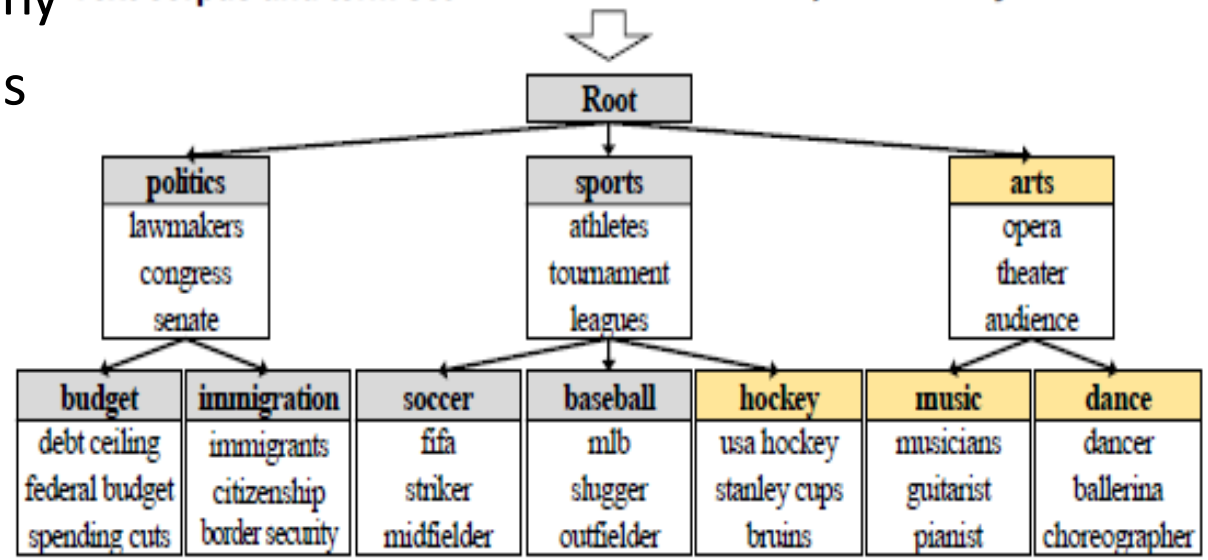
- Topic taxonomy completion: Task ≈ CoRel
- Results: Better quality than Corel
- Method:
  - Recursive expansion of a given topic hierarchy
  - Discovering novel sub-topic clusters of terms and documents



Text corpus and term set



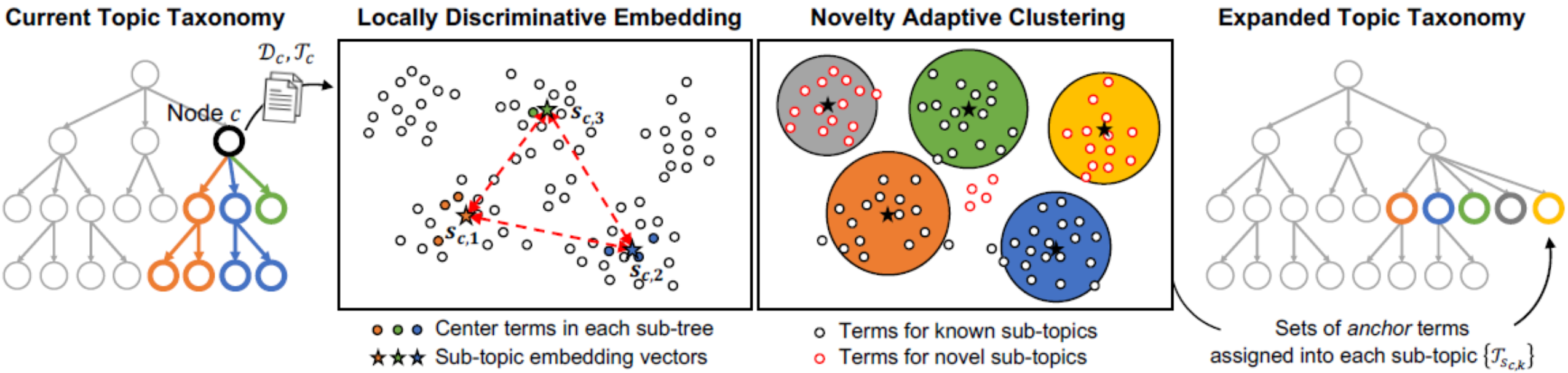
Partial topic hierarchy



	dance	surveillance	number theory	accelerator physics
CoRel	dance dancers new york city ballet american ballet theater choreography choreographer	surveillance national security agency intelligence snowdennational security counterterrorism	number theory birch mathematicians pure mathematics number fields class numbers	accelerator physics particle accelerators linear accelerator conceptual design mechanical design power converters
TaxoCom	dance choreography ballet dancers pas de deux balanchine ballets	surveillance eavesdropping spying national security agency phone records patriot act	number theory modular form number fields iwasawa theory elliptic curves prime number theorem	accelerator physics synchrotron particle accelerators linear accelerator storage ring tevatron

Dongha Lee, Jiaming Shen, SeongKu Kang, Susik Yoon, Jiawei Han, Hwanjo Yu, "TaxoCom: Topic Taxonomy Completion with Hierarchical Discovery of Novel Topic Clusters", WWW'22

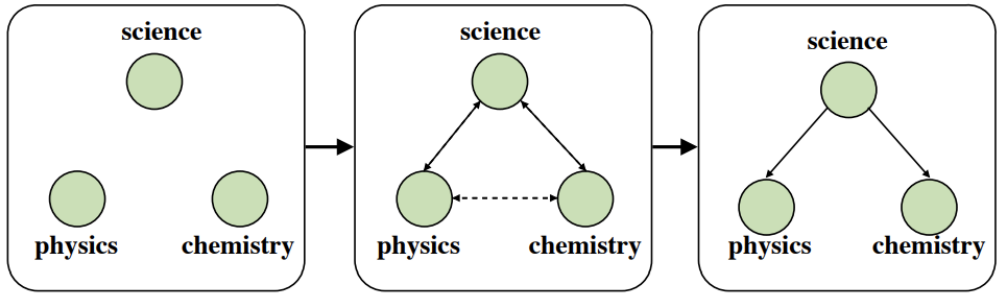
TaxoCom: Hierarchical Discovery of Novel Topic Clusters



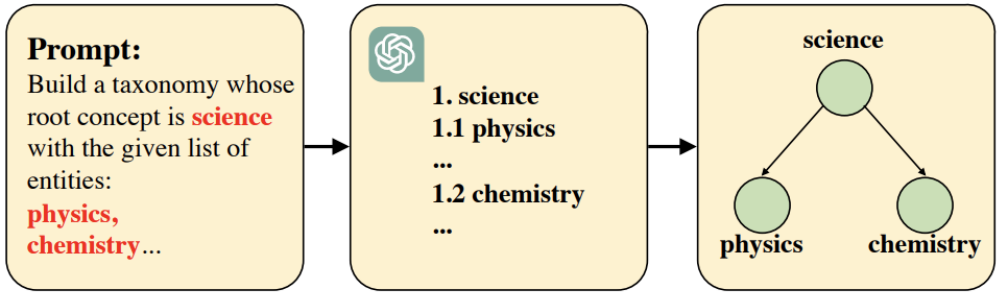
- Starting from the root node, it performs (i) locally discriminative embedding, and (ii) novelty adaptive clustering, to selectively assign the terms (of each node) into one of the child nodes
  - Locally discriminative embedding optimizes the text embedding space to be discriminative among known (i.e., given) sub-topics
  - Novelty adaptive clustering assigns terms into either one of the known sub-topics or novel sub-topics

LLM for Taxonomy Construction

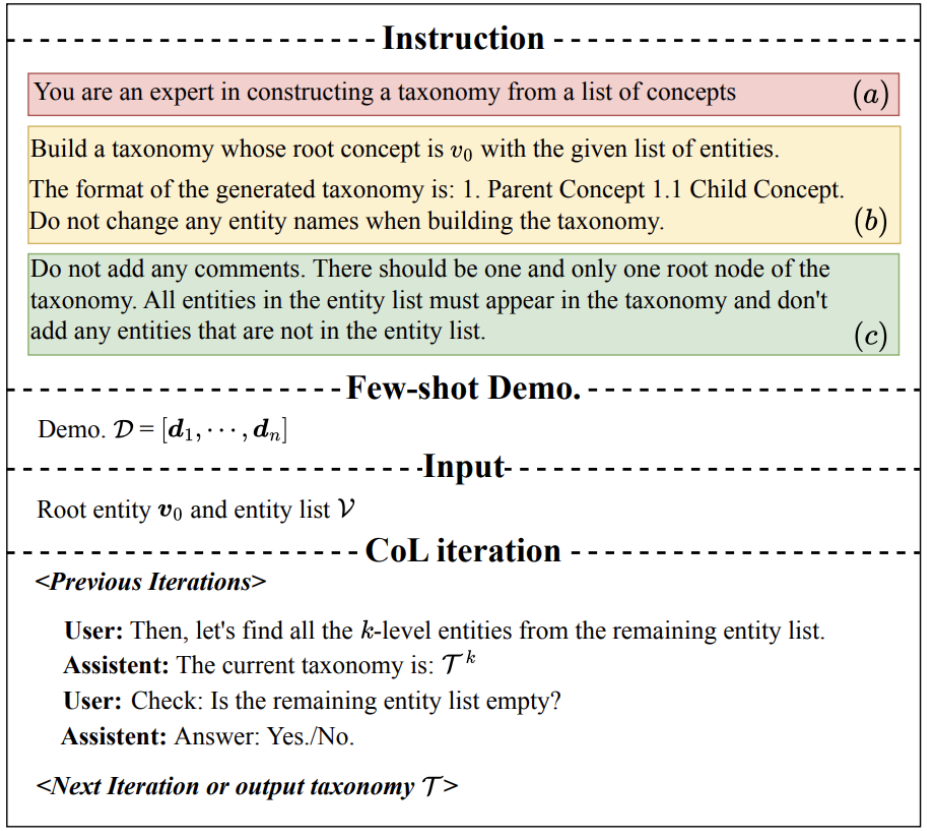
- The general knowledge of LLMs makes it possible to prompt an LLM for taxonomy construction



(a) Discriminative Methods: Scoring each entity pair and pruning to taxonomic structure [6, 26]



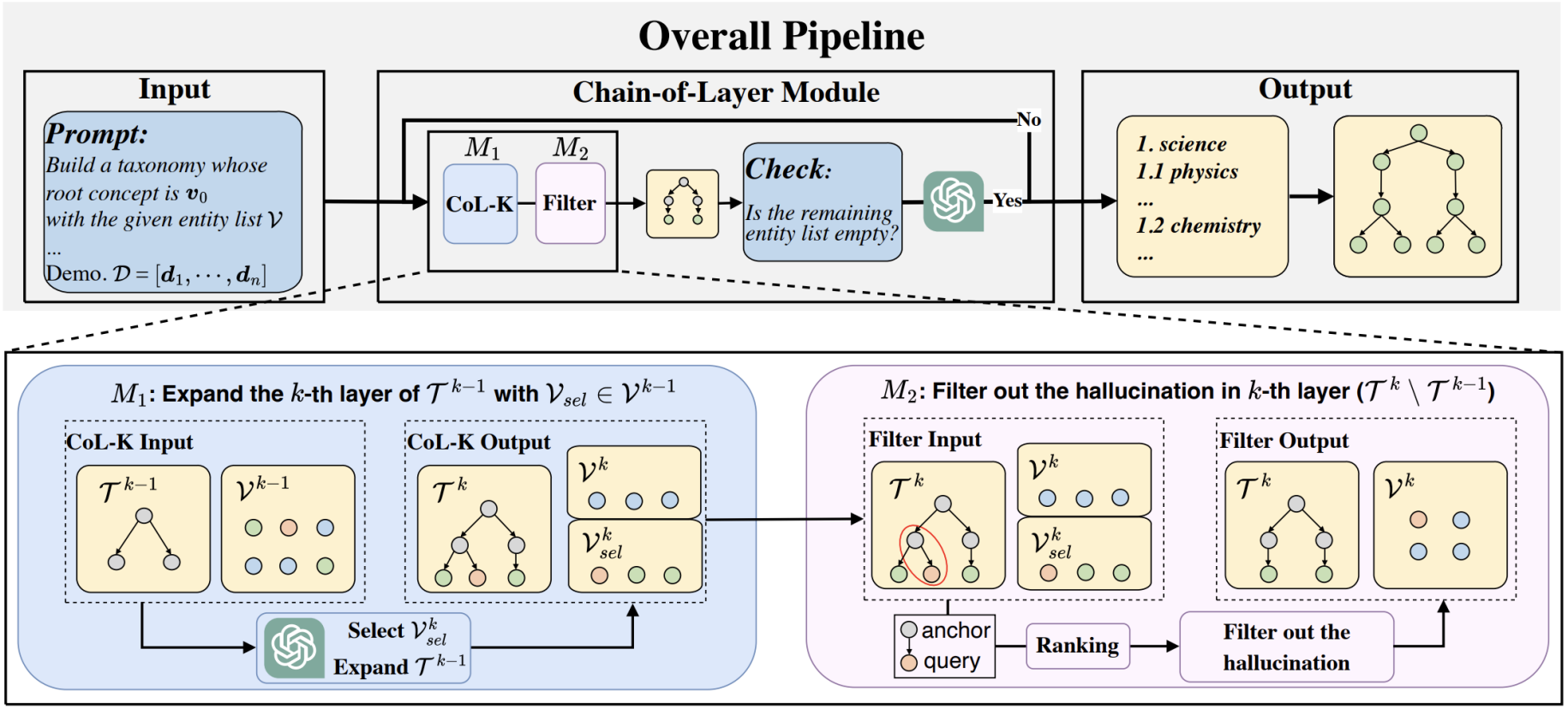
(b) Generative Methods: Prompting LLMs to generate taxonomy



Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Shangbin Feng, Zhenwen Liang, Zhihan Zhang, Meng Jiang, "Chain-of-Layer: Iteratively Prompting Large Language Models for Taxonomy Induction from Limited Examples", arXiv'24

Chain-of-Layer

- Chain-of-layer prompting to build new layers from provided candidates
- Filtering out hallucination by LLMs



Outline

- Taxonomy basics and why do we need it?
- Taxonomy Construction
  - CGExpan [ACL'20], CoRel [KDD'20], TaxoCom [WWW'22], Chain-of-Layer [arXiv'24]
- Taxonomy Expansion
  - TaxoExpan [WWW'20], BoxTaxo [WWW'23], TaxoInstruct [arXiv'24]
- Taxonomy Enrichment
  - CatE [WWW'20], JoSH [KDD'20], SeedTopicMine [WSDM'23]

Taxonomy Expansion: Motivation

- Why taxonomy expansion 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

TaxoExpan: Self-supervised Taxonomy Expansion with Position-Enhanced Graph Neural Network

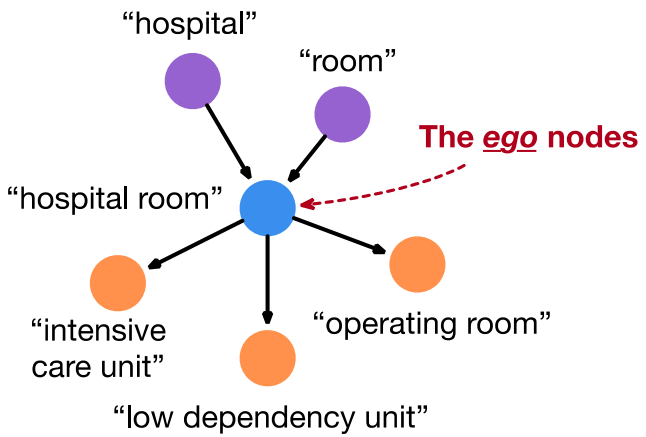
- **Two steps** in solving the problem:
  - Self-supervised term extraction
    - Automatically **extracts emerging terms** from a target domain
  - Self-supervised term attachment
    - A multi-class classification to match a new node to its potential parent
    - Heterogenous sources of information (structural, semantic, and lexical) can be used

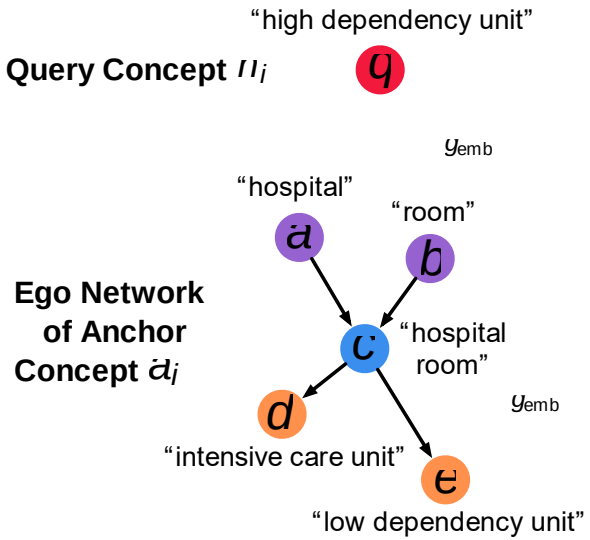
Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang and Jiawei Han. "TaxoExpan: Self-supervised Taxonomy Expansion with Position-Enhanced Graph Neural Network", WWW'20

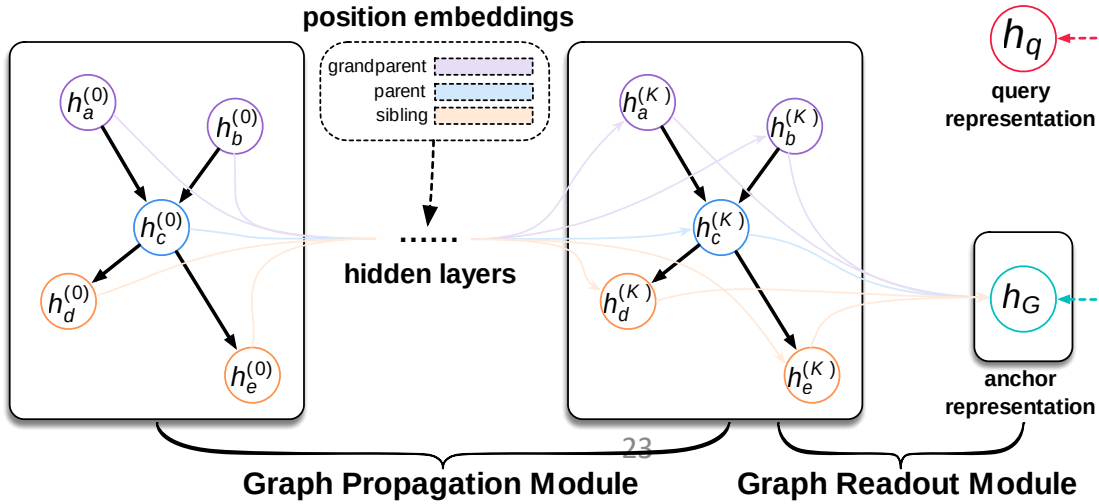
Self-supervised Term Attachment

- **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"



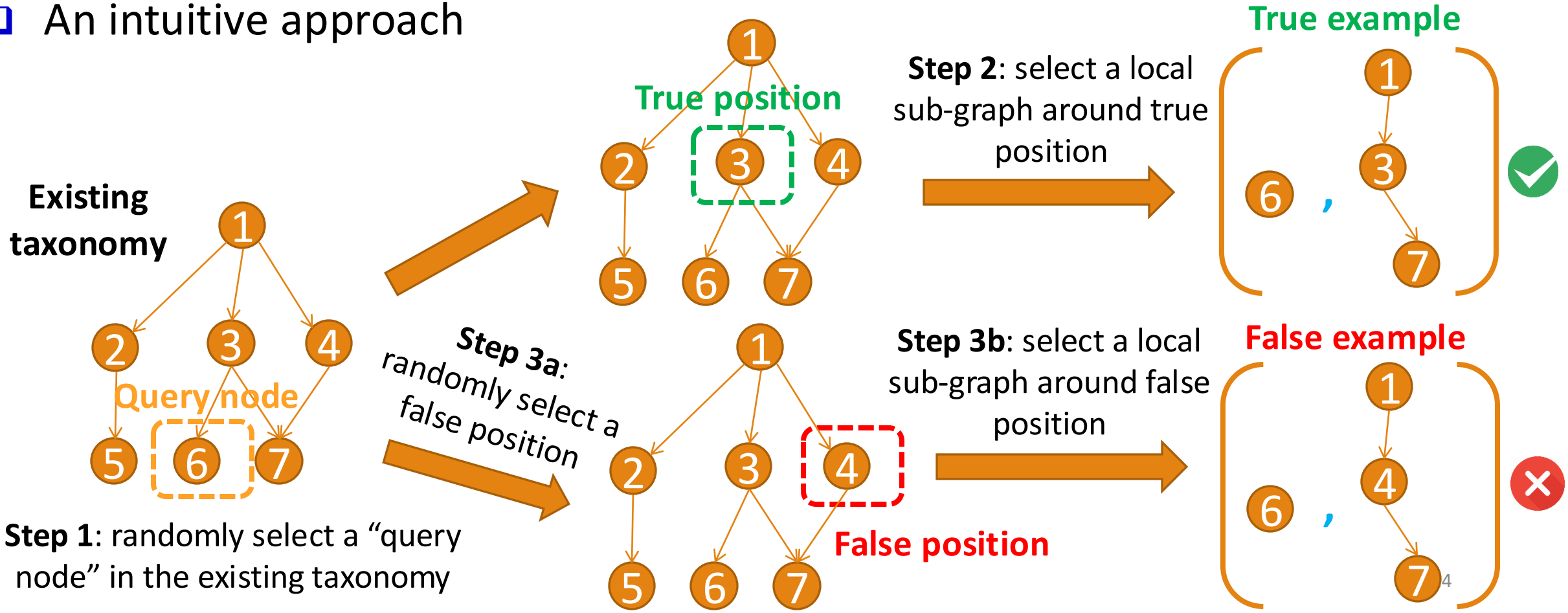




23

Leveraging Existing Taxonomy for Self-supervised Learning

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



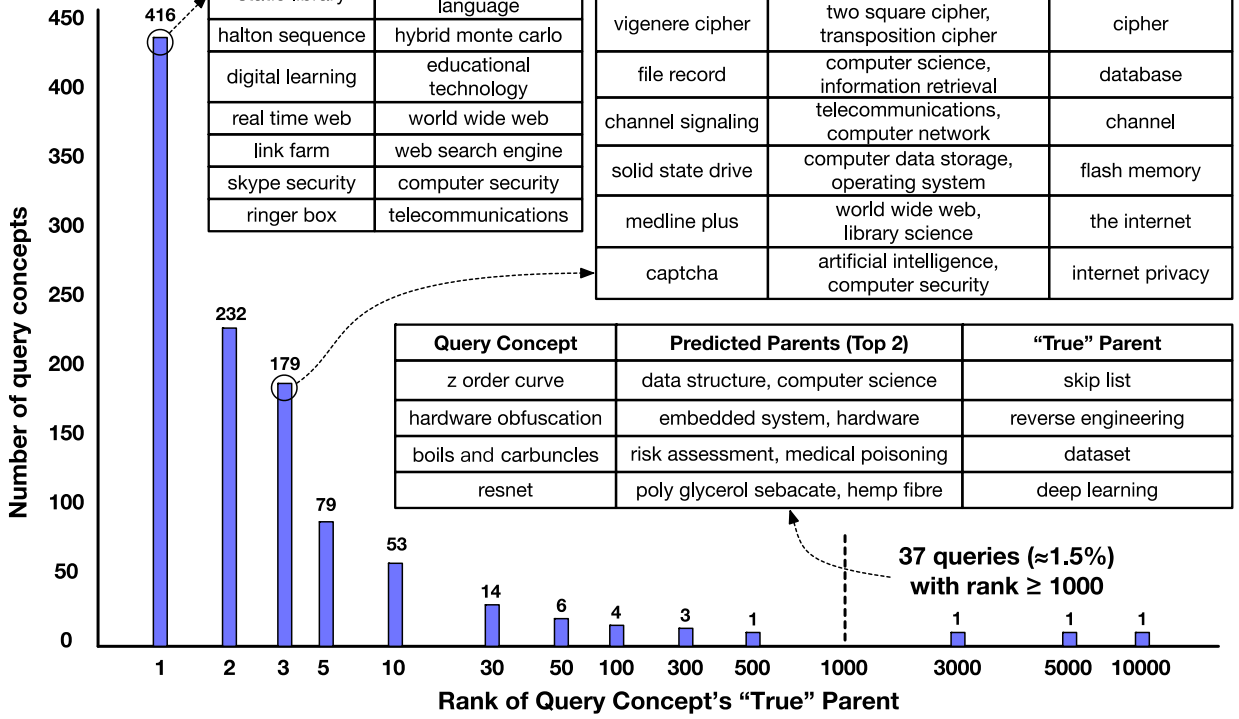
TaxoExpan Framework Analysis

- 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

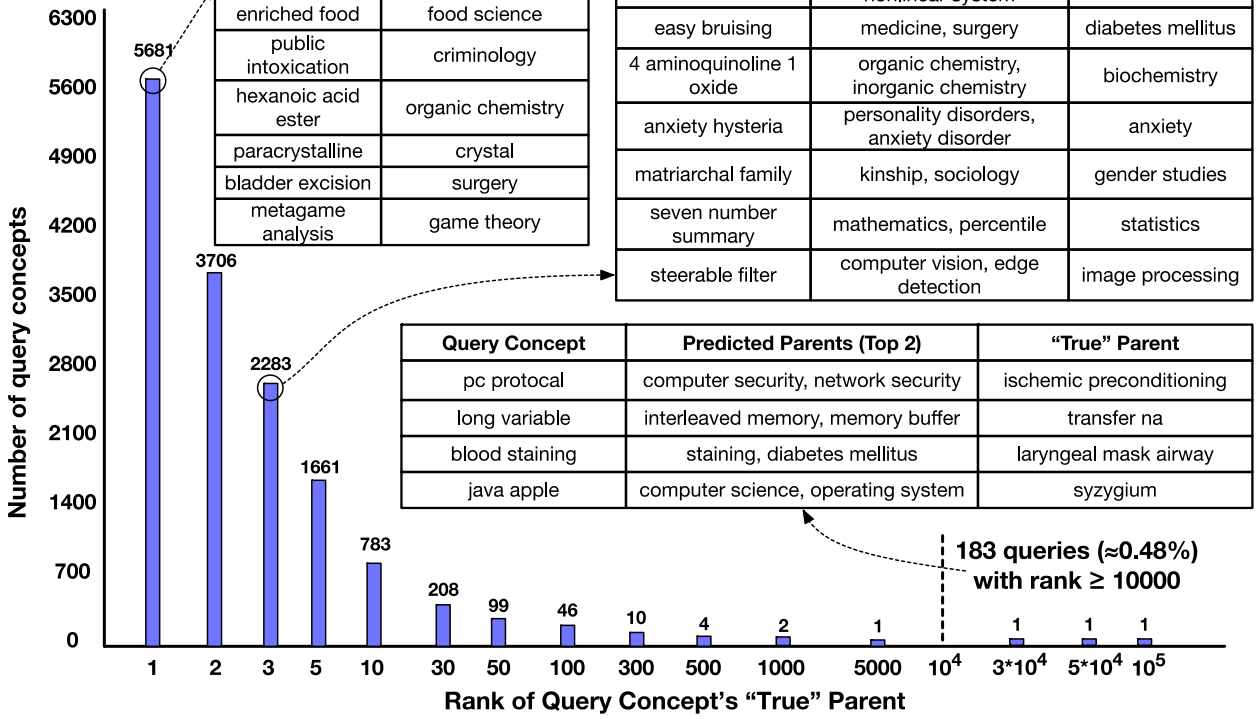


(a) MAG-CS Dataset (totally 2450 query concepts)

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

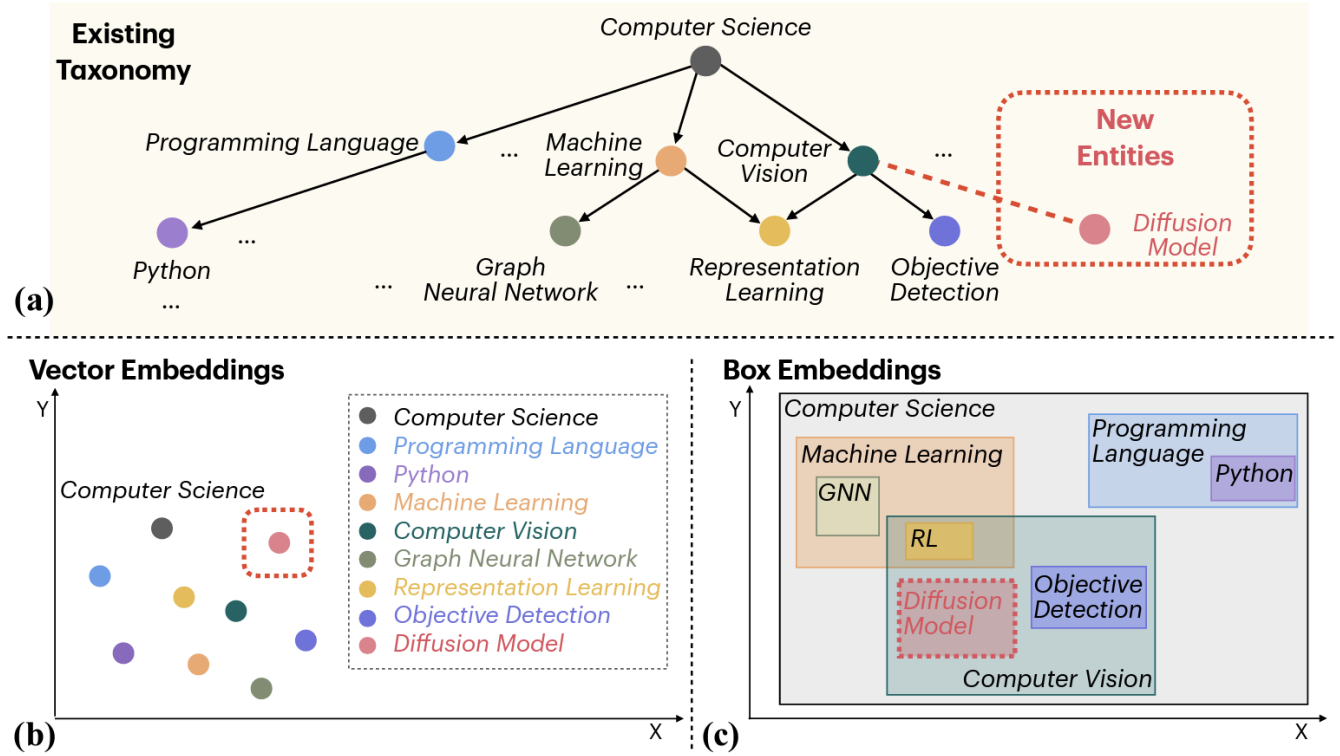
Query Concept	Predicted Parents (Top 2)	"True" Parent
pc protocal	computer security, network security	ischemic preconditioning
long variable	interleaved memory, memory buffer	transfer na
blood staining	staining, diabetes mellitus	laryngeal mask airway
java apple	computer science, operating system	syzygium



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

A Single Vector Is Not Enough: Taxonomy Expansion via Box Embeddings

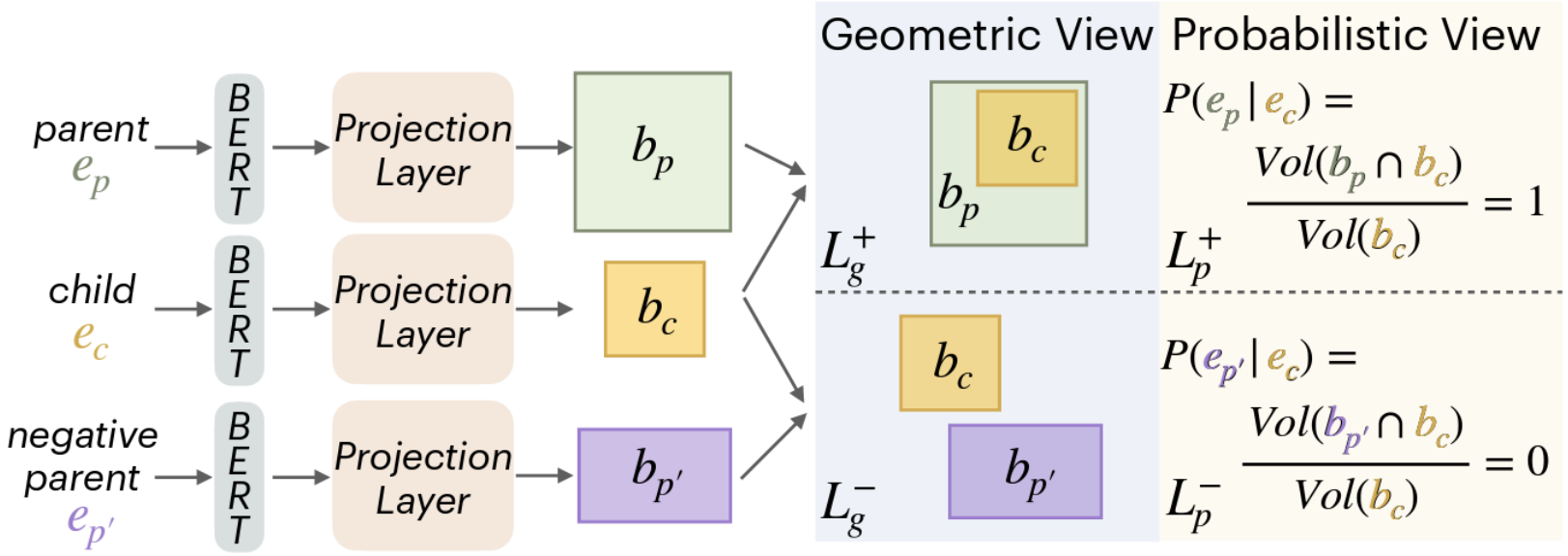
- Vector embeddings can only represent similarity/dissimilarity
- Box embeddings can represent entailment relations



Song Jiang, Qiyue Yao, Qifan Wang, Yizhou Sun. "A Single Vector Is Not Enough: Taxonomy Expansion via Box Embeddings", WWW'23

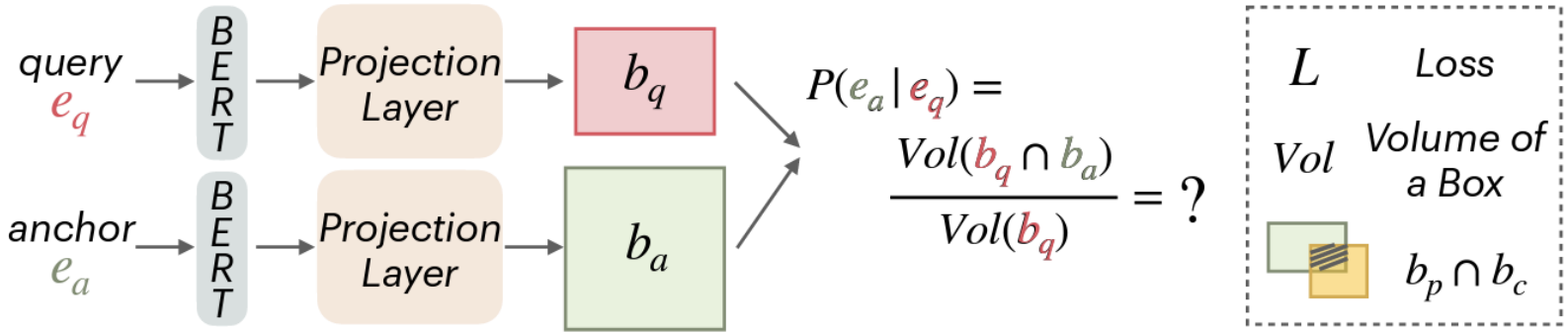
Box Training

- Training: the box embeddings are optimized to accurately represent the taxonomic hierarchies.



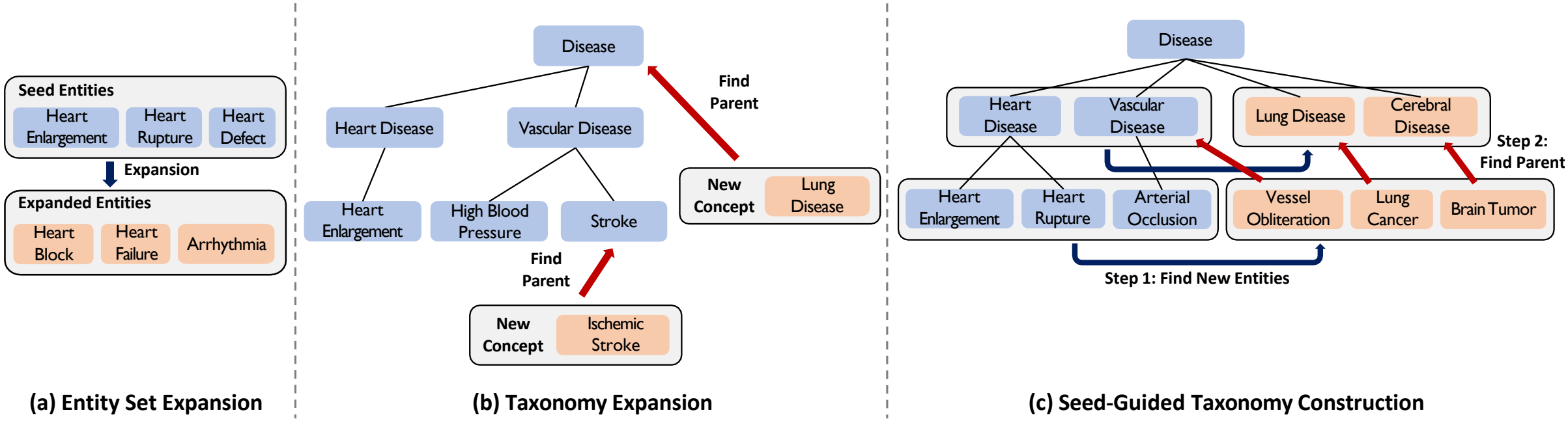
Inference with Box

- Inference: check whether a query's box is enclosed by the candidate anchor's box in a probabilistic way.



TaxoInstruct: Unifying Entity Set Expansion and Taxonomy Expansion

- Entity Set Expansion, Taxonomy Expansion, and Seed-Guided Taxonomy Construction all focus on populating an existing taxonomy with new concepts
- Previous approaches only work on one specific task
- Can we build a unified framework that offers solution to all three tasks?



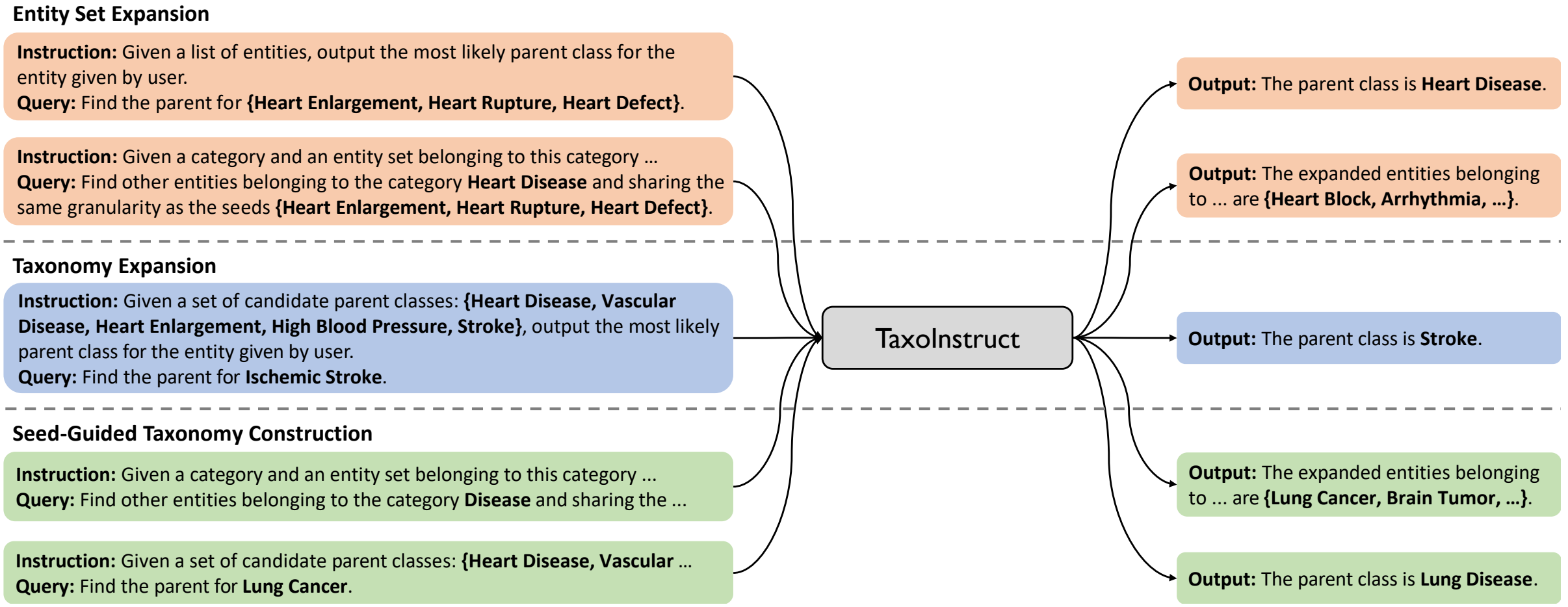
(a) Entity Set Expansion (b) Taxonomy Expansion (c) Seed-Guided Taxonomy Construction

Yanzhen Shen, Yu Zhang, Yunyi Zhang, Jiawei Han. "A Unified Taxonomy-Guided Instruction Tuning Framework for Entity Set Expansion and Taxonomy Expansion", arXiv'24

TaxoInstruct: A Unified Framework

- All three tasks are entities that share a relationship with the given entities
  - Entity Set Expansion: **Finding siblings**
  - Taxonomy Expansion: **Finding parents**
  - Seed-Guided Taxonomy Construction: **Finding siblings** + **Finding parents**
- Utilize task-specific instructions to teach LLM to generate siblings and find parents
- A Joint training process facilitates the mutual enhancement of both skills

TaxoInstruct: Taxonomy-Guided Instruction Tuning



TaxoInstruct: Experiment Result

Entity Set Expansion

Method	APR		Wiki	
	MAP@10	MAP@20	MAP@10	MAP@20
EgoSet [33]†	0.758*	0.710*	0.904*	0.877*
SetExpan [37]†	0.789*	0.763*	0.944*	0.921*
SetExpander [24]†	0.287*	0.208*	0.499*	0.439*
CaSE [56]†	0.619*	0.494*	0.897*	0.806*
SetCoExpan [12]‡	0.933*	0.915*	0.976*	0.964*
CGExpan [62]†	0.992	0.990*	0.995	0.978*
SynSetExpan [35]▷	0.985*	0.990*	0.991*	0.978*
ProbExpan [19]▷	0.993	0.990*	0.995	0.982
TaxoInstruct	**0.9956**	**0.9928**	**0.9957**	**0.9875**
NoParentPretrain	0.9867*	0.9689*	0.9746*	0.9720*

Seed-Guided Taxonomy Construction

Method	DBLP		PubMed-CVD	
	Sibling nDCG@50	Parent nDCG@50	Sibling nDCG@50	Parent nDCG@50
HSetExpan [37]	0.8814*	0.8268*	0.6515*	0.5085*
NoREPEL [38]	0.8830*	0.8152*	0.6705*	0.6216*
NoGTO [38]	0.9527*	0.8855*	0.7395*	0.6428*
HiExpan [38]	0.9524*	0.9045	0.7365*	0.7132*
TaxoInstruct	**0.9817**	**0.9210**	**0.9220**	**0.8034**
NoParentPretrain	0.9668*	0.7836*	0.8920*	0.7864
NoSiblingPretrain	0.9425*	0.9114	0.7930*	0.6838*

Taxonomy Expansion

Method	Environment		Science	
	Acc	Wu&P	Acc	Wu&P
TAXI [30]†	0.167*	0.447*	0.130*	0.329*
HypeNET [42]†	0.167*	0.558*	0.154*	0.507*
BERT+MLP [8]†	0.111*	0.479*	0.115*	0.436*
TaxoExpan [36]†	0.111*	0.548*	0.278*	0.576*
Arborist [26]‡	0.4615*	–	0.4193*	–
Graph2Taxo [34]‡	0.2105*	–	0.2619*	–
STEAM [57]†	0.361*	0.696*	0.365*	0.682*
TMN [60]‡	0.3793*	–	0.3415*	–
TEMP [22]▷	0.492*	0.777*	0.578*	0.853
GenTaxo [58]‡	0.4828*	–	0.3878*	–
BoxTaxo [15]†	0.381*	0.754*	0.318*	0.647*
TaxoInstruct	**0.5115**	**0.8300**	**0.6165**	0.8480
NoSiblingPretrain	0.4616*	0.7911*	0.5953*	**0.8559**

Outline

- Taxonomy basics and why do we need it?
- Taxonomy Construction
  - CGExpan [ACL'20], CoRel [KDD'20], TaxoCom [WWW'22], Chain-of-Layer [arXiv'24]
- Taxonomy Expansion
  - TaxoExpan [WWW'20], BoxTaxo [WWW'23], TaxoInstruct [arXiv'24]
- Taxonomy Enrichment
  - CatE [WWW'20], JoSH [KDD'20], SeedTopicMine [WSDM'23]

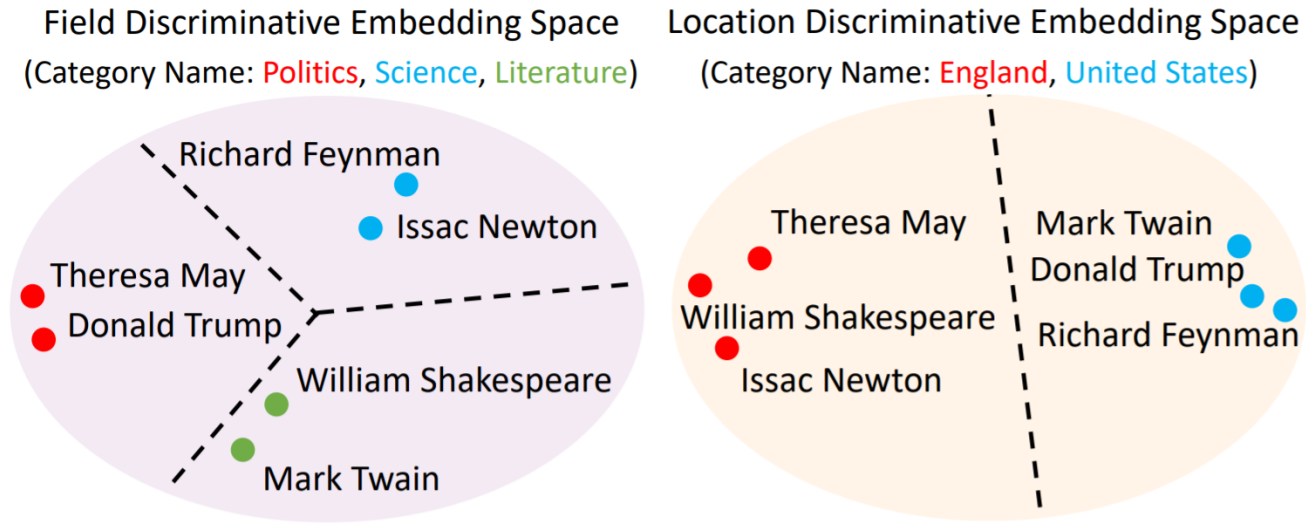
Taxonomy Enrichment: Discriminative Topic Mining

- **Discriminative Topic Mining:** Given a text corpus and a set of (hierarchically organized) **category names**, retrieve a set of terms that **exclusively belong to** each category
  - E.g., given c_1: "The United States", c_2: "France", c_3: "Canada"
    - Yes to "Ontario" under c_3: (a province in Canada and exclusively belongs to Canada)
    - No to "North America" under c_3: (a continent and does not belong to any countries (reversed belonging relationship))
    - No to "English" under c_3: (English is also the national language of the United States (not discriminative))
- Difference from (hierarchical) topic modeling [1, 2]
  - requires **a set of category names** and only retrieves terms belonging to the given categories
  - imposes strong discriminative requirements that each retrieved term under the corresponding category must **belong to and only belong to** that category semantically

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research.
[2] Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2003). Hierarchical topic models and the nested Chinese restaurant process. NIPS.

Discriminative Topic Mining via CatE

- Word embeddings capture word semantic correlations via the distributional hypothesis
  - captures local context similarity
  - not exploit document-level statistics (global context)
  - not model topics
- **CatE: Category Name-guided Embedding:** leverages *category names* to learn word embeddings with discriminative power over the specific set of categories
- CatE: Inputs
  - Category names + Corpus
- CatE: Outputs (see figure)
  - The same set of celebrities are embedded differently given different sets of category names



Meng, Y., Huang, J., Wang, G., Wang, Z., Zhang, C., Zhang, Y., & Han, J. (2020). Discriminative topic mining via category-name guided text embedding. WWW.

CatE Embedding: Objective

- Objective: negative log-likelihood

$$P(\mathcal{D} \mid C) = \prod_{d \in \mathcal{D}} p(d \mid c_d) \prod_{w_i \in d} p(w_i \mid d) \prod_{\substack{w_{i+j} \in d \\ -h \le j \le h, j \neq 0}} p(w_{i+j} \mid w_i)$$

1. Topic assignment
2. Global context
3. Local context

$$p(d \mid c_d) \propto p(c_d \mid d)p(d) \propto p(c_d \mid d) \propto \prod_{w \in d} p(c_d \mid w),$$

Decompose into word-topic distribution

- Introducing specificity

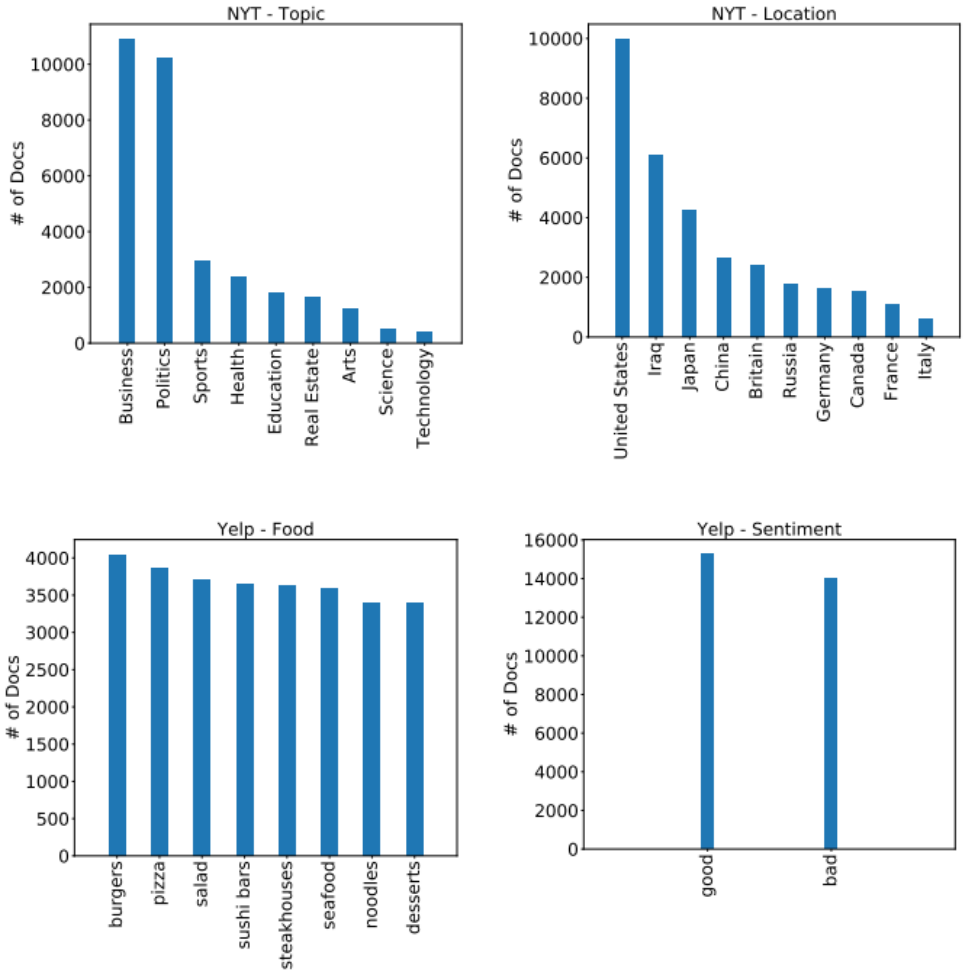
Definition 2 (Word Distributional Specificity). We assume there is a scalar $\kappa_w \geq 0$ correlated with each word w indicating how specific the word meaning is. The bigger κ_w is, the more specific meaning word w has, and the less varying contexts w appears in.

  - E.g., "seafood" has a higher word distributional specificity than "food", because seafood is a specific type of food

Quantitative Results

- Two datasets:
  - New York Times annotated corpus (NYT)
    - Two categories: topic and location
  - Yelp Dataset Challenge (Yelp)
    - Two categories: food type and sentiment

Methods	**NYT**-Location		**NYT**-Topic		**Yelp**-Food		**Yelp**-Sentiment	
	TC	MACC	TC	MACC	TC	MACC	TC	MACC
LDA	0.007	0.489	0.027	0.744	-0.033	0.213	-0.197	0.350
Seeded LDA	0.024	0.168	0.031	0.456	0.016	0.188	0.049	0.223
TWE	0.002	0.171	-0.011	0.289	0.004	0.688	-0.077	0.748
Anchored CorEx	0.029	0.190	0.035	0.533	0.025	0.313	0.067	0.250
Labeled ETM	0.032	0.493	0.025	0.889	0.012	0.775	0.026	0.852
CatE	**0.049**	**0.972**	**0.048**	**0.967**	**0.034**	**0.913**	**0.086**	**1.000**



Dataset stat: # of docs by category name

Qualitative Results

Methods	**NYT**-Location		**NYT**-Topic		**Yelp**-Food		**Yelp**-Sentiment	
	britain	canada	education	politics	burger	desserts	good	bad
LDA	company (×)	percent (×)	school	campaign	fatburger	ice cream	great	valet (×)
	companies (×)	economy (×)	students	clinton	dos (×)	chocolate	place (×)	peter (×)
	british	canadian	city (×)	mayor	liar (×)	gelato	love	aid (×)
	shares (×)	united states (×)	state (×)	election	cheeseburgers	tea (×)	friendly	relief (×)
	great britain	trade (×)	schools	political	bearing (×)	sweet	breakfast	rowdy
Seeded LDA	british	city (×)	state (×)	republican	like (×)	great (×)	place (×)	service (×)
	industry (×)	building (×)	school	political	fries	like (×)	great	did (×)
	deal (×)	street (×)	students	senator	just (×)	ice cream	service (×)	order (×)
	billion (×)	buildings (×)	city (×)	president	great (×)	delicious (×)	just (×)	time (×)
	business (×)	york (×)	board (×)	democrats	time (×)	just (×)	ordered (×)	ordered (×)
TWE	germany (×)	toronto	arts (×)	religion	burgers	chocolate	tasty	subpar
	spain (×)	osaka (×)	fourth graders	race	fries	complimentary (×)	decent	positive (×)
	manufacturing (×)	booming (×)	musicians (×)	attraction (×)	hamburger	green tea (×)	darned (×)	awful
	south korea (×)	asia (×)	advisors	era (×)	cheeseburger	sundae	great	crappy
	markets (×)	alberta	regents	tale (×)	patty	whipped cream	suffered (×)	honest (×)
Anchored CorEx	moscow (×)	sports (×)	republican (×)	military (×)	order (×)	make (×)	selection (×)	did (×)
	british	games (×)	senator (×)	war (×)	know (×)	chocolate	prices (×)	just (×)
	london	players (×)	democratic (×)	troops (×)	called (×)	people (×)	great	came (×)
	german (×)	canadian	school	baghdad (×)	fries	right (×)	reasonable	asked (×)
	russian (×)	coach	schools	iraq (×)	going (×)	want (×)	mac (×)	table (×)
Labeled ETM	france (×)	canadian	higher education	political	hamburger	pana	decent	horrible
	germany (×)	british columbia	educational	expediency (×)	cheeseburger	gelato	great	terrible
	canada (×)	britain (×)	school	perceptions (×)	burgers	tiramisu	tasty	good (×)
	british	quebec	schools	foreign affairs	patty	cheesecake	bad (×)	awful
	europe (×)	north america (×)	regents	ideology	steak (×)	ice cream	delicious	appallingly
CatE	england	ontario	educational	political	burgers	dessert	delicious	sickening
	london	toronto	schools	international politics	cheeseburger	pastries	mindful	nasty
	britons	quebec	higher education	liberalism	hamburger	cheesecakes	excellent	dreadful
	scottish	montreal	secondary education	political philosophy	burger king	scones	wonderful	freaks
	great britain	ottawa	teachers	geopolitics	smash burger	ice cream	faithful	cheapskates

Hierarchical Topic Mining

- Mining a set of meaningful topics organized into a **hierarchy** is intuitively appealing and has broad applications
  - Coarse-to-fine topic understanding
  - Hierarchical corpus summarization
  - Hierarchical text classification
  - …
- Hierarchical topic models discover topic structures from text corpora via modeling the text generative process with a latent hierarchy

Meng, Y., Zhang, Y., Huang, J., Zhang, Y., Zhang, C., & Han, J. (2020). Hierarchical topic mining via joint spherical tree and text embedding. KDD.

JoSH Text Embedding

- Modeling Text Generation Conditioned on the Category Tree (Similar to CatE)
- A three-step process:

1. A document d_i is generated conditioned on one of the n categories

1. Topic assignment

$$p(d_i \mid c_i) = \text{vMF}(\boldsymbol{d}_i; \boldsymbol{c}_i, \kappa_{c_i}) = n_p(\kappa_{c_i}) \exp\left(\kappa_{c_i} \cdot \cos(\boldsymbol{d}_i, \boldsymbol{c}_i)\right)$$

2. Each word w_j is generated conditioned on the semantics of the document d_i

2. Global context

$$p(w_j \mid d_i) \propto \exp(\cos(\boldsymbol{u}_{w_j}, \boldsymbol{d}_i))$$

3. Surrounding words w_{j+k} in the local context window of w_i are generated conditioned on the semantics of the center word w_i

3. Local context

$$p(w_{j+k} \mid w_j) \propto \exp(\cos(\boldsymbol{v}_{w_{j+k}}, \boldsymbol{u}_{w_j}))$$

JoSH Tree Embedding

- **Intra-Category Coherence**: Representative terms of each category should be highly semantically relevant to each other, reflected by high directional similarity in the spherical space

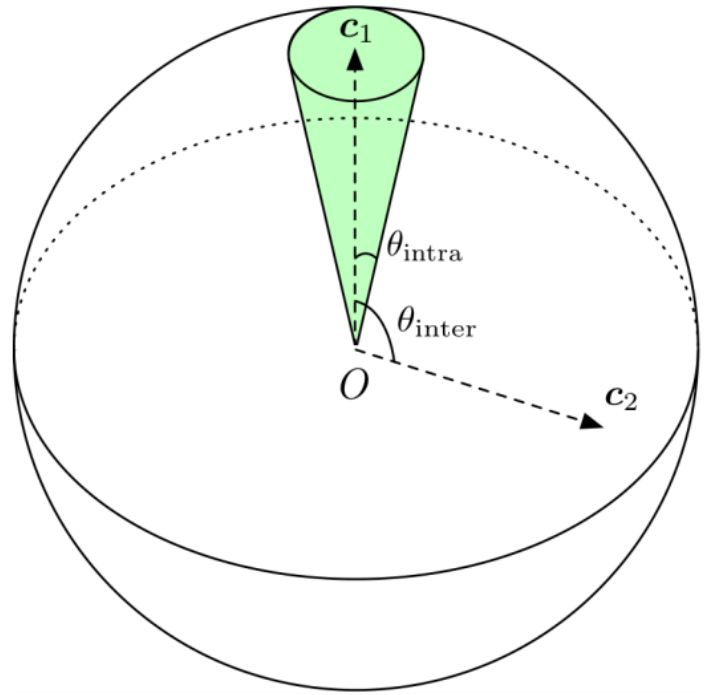
$$\mathcal{L}_{\text{intra}} = \sum_{c_i \in \mathcal{T}} \sum_{w_j \in C_i} \min(0, \boldsymbol{u}_{w_j}^{\top} \boldsymbol{c}_i - m_{\text{intra}}),$$

- **Inter-Category Distinctiveness**: Encourage distinctiveness across different categories to avoid semantic overlaps so that the retrieved terms provide a clear and distinctive description

$$\mathcal{L}_{\text{inter}} = \sum_{c_i \in \mathcal{T}} \sum_{c_j \in \mathcal{T} \setminus \{c_i\}} \min(0, 1 - \boldsymbol{c}_i^{\top} \boldsymbol{c}_j - m_{\text{inter}}).$$

$\theta_{\text{intra}} \leq \arccos(m_{\text{intra}})$

$\theta_{\text{inter}} \geq \arccos(1 - m_{\text{inter}})$

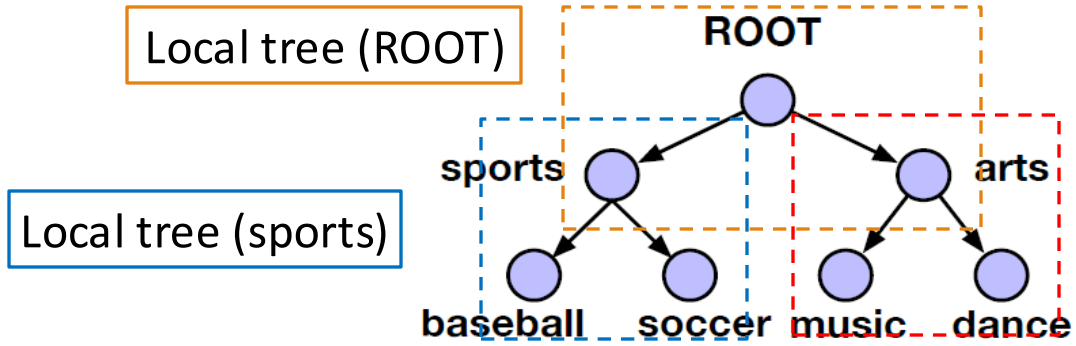


(a) Intra- & Inter-Category Configuration.

JoSH Tree Embedding

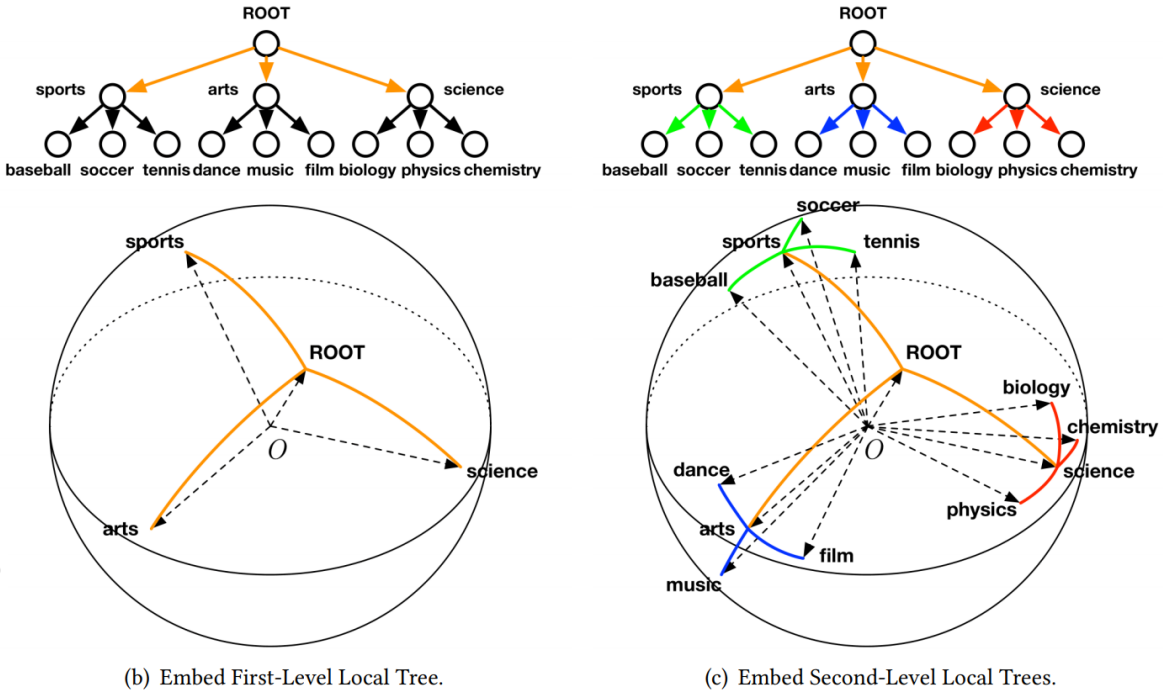
- **Recursive Local Tree Embedding:** Recursively embed local structures of the category tree onto the sphere

Local tree: A local tree T_r rooted at node $c_r \in T$ consists of node c_r and all of its direct children



- **Preserving Relative Tree Distance within Local Trees**: A category should be closer to its parent category than to its sibling categories in the embedding space

$$\mathcal{L}_{\text{inter}} = \sum_{c_i \in \mathcal{T}_r} \sum_{c_j \in \mathcal{T}_r \setminus \{c_r, c_i\}} \min(0, \boldsymbol{c}_i^\top \boldsymbol{c}_r - \boldsymbol{c}_i^\top \boldsymbol{c}_j - m_{\text{inter}}),$$



(b) Embed First-Level Local Tree.

(c) Embed Second-Level Local Trees.

Experiments: Qualitative Results on NYT

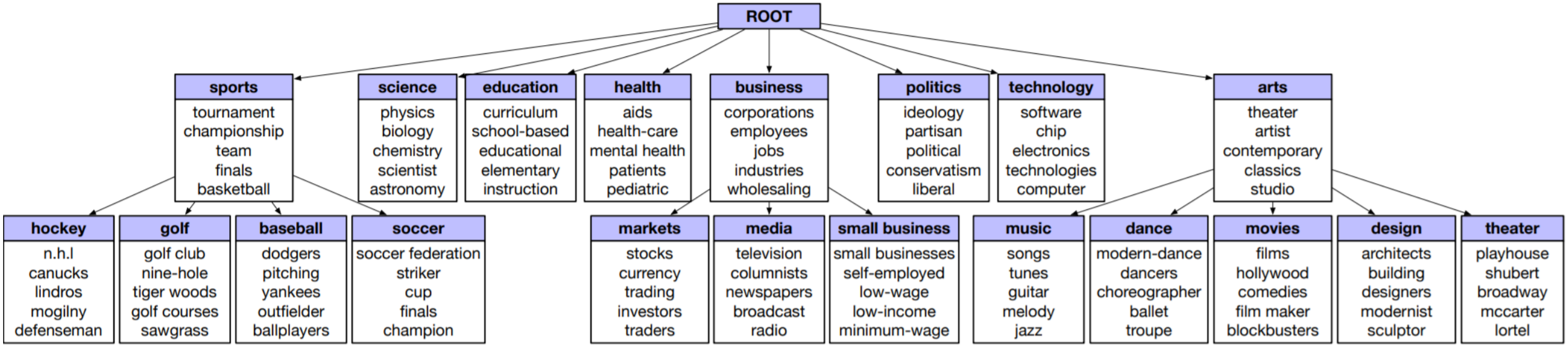
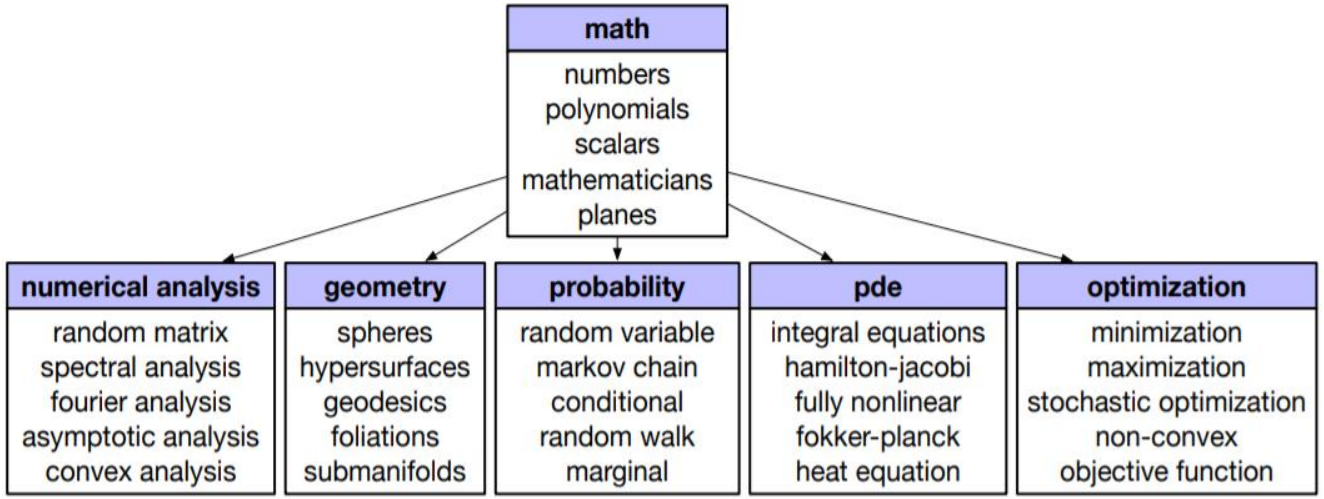
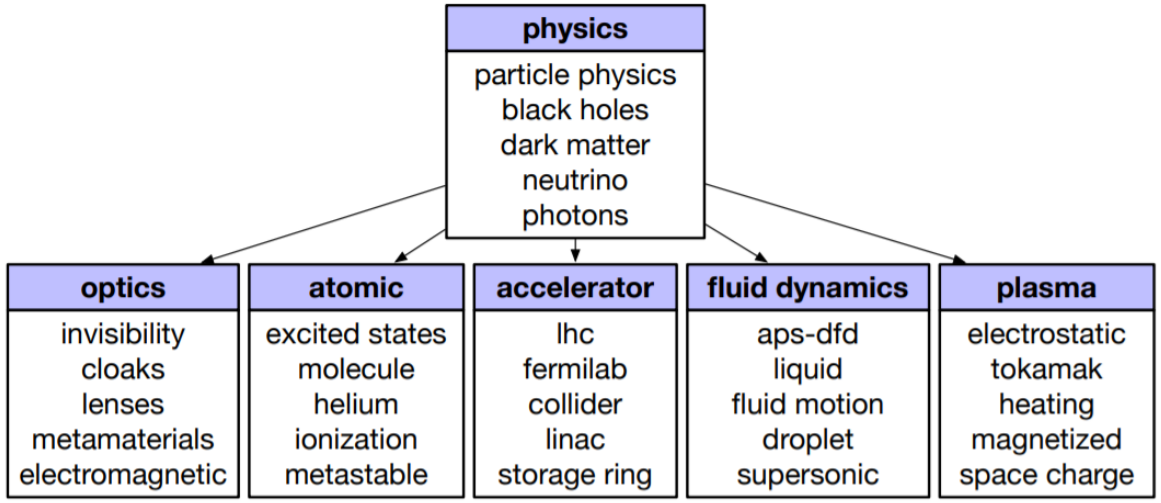


Figure 3: Hierarchical Topic Mining results on NYT.

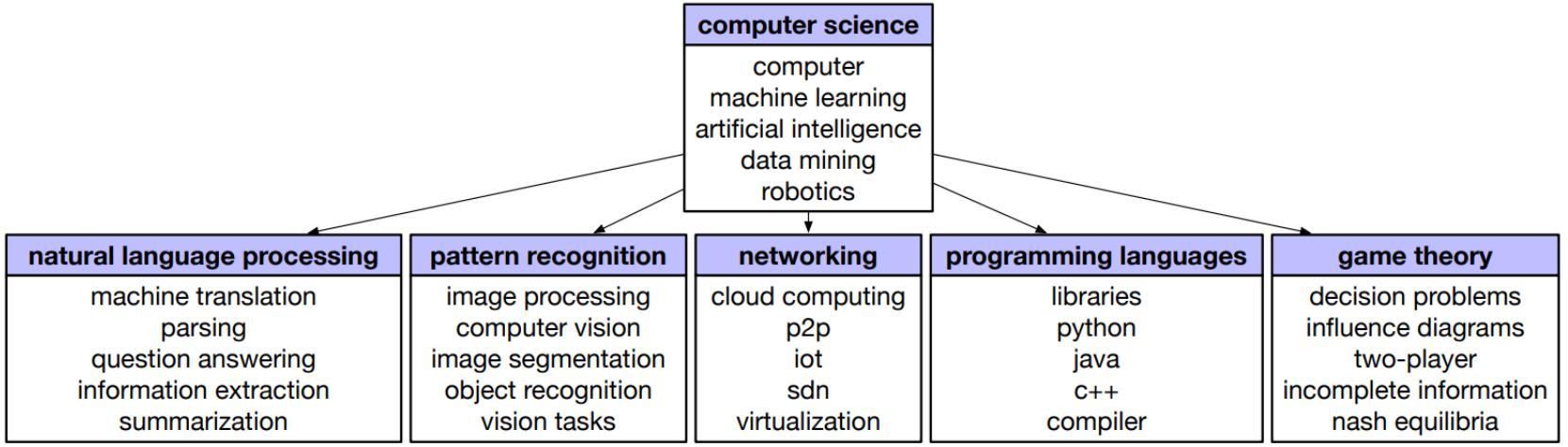
Experiments: Qualitative Results on ArXiv and Quantitative Results



(a) "Math" subtree.



(b) "Physics" subtree.

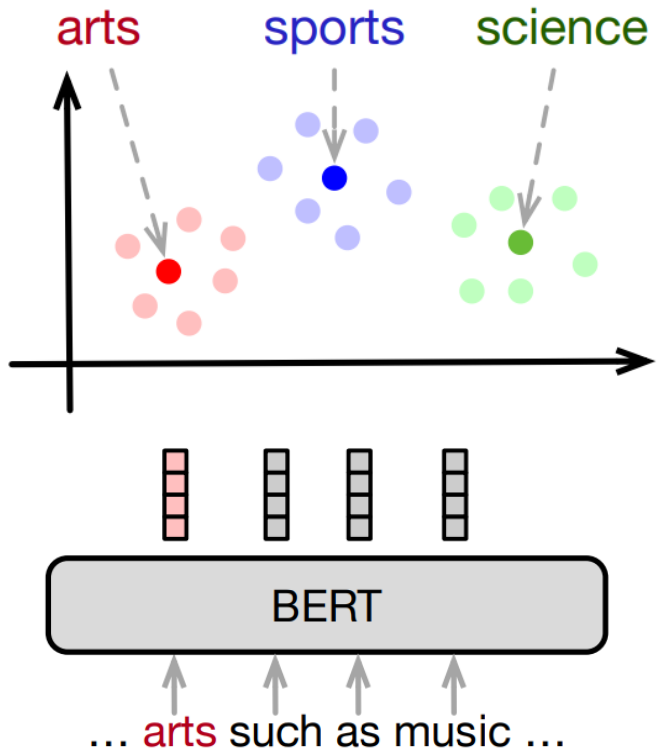


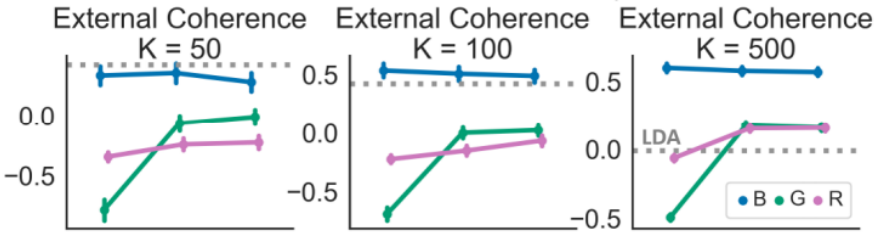
(c) "Computer Science" subtree.

Models	NYT		arXiv	
	TC	MACC	TC	MACC
hLDA	-0.0070	0.1636	-0.0124	0.1471
hPAM	0.0074	0.3091	0.0037	0.1824
JoSE	0.0140	0.6818	0.0051	0.7412
Poincaré GloVe	0.0092	0.6182	-0.0050	0.5588
Anchored CorEx	0.0117	0.3909	0.0060	0.4941
CatE	0.0149	0.9000	0.0066	0.8176
JoSH	**0.0166**	**0.9091**	**0.0074**	**0.8324**

Commonly Used Context Information

- **Context Type I - Skip-Gram Embeddings**
  - Previous slides have shown that seed-guided topic discovery, can leverage seeds in the embedding learning process by viewing the category that a term belongs to as its context
- **Context Type II - Pre-trained Language Model Representations**
  - The generic knowledge learned by PLMs from web-scale corpora (e.g., Wikipedia) can complement the information one can get from the input corpus
  - BERT representations suffer from the curse of dimensionality and may not form clearly separated clusters [1]

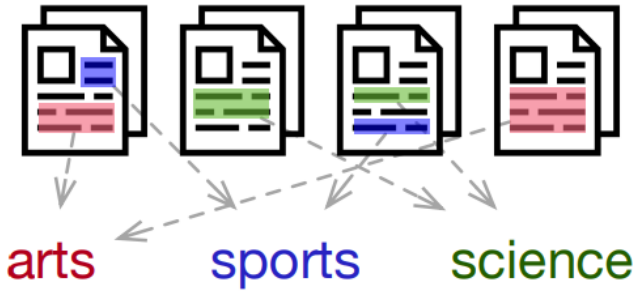




[1] Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, Jiawei Han (2022). Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations. WWW'22.

Commonly Used Context Information

- **Context Type III - Topic-Indicative Documents**
  - Supervised topic models [1] propose to leverage document-level training data. However, such information relies on **massive human annotation**, which is not available under the seed-guided setting.



  - A document may be **too broad** to be viewed as a context unit because each document can be relevant to multiple topics simultaneously.

- **Each type of context signals has its specific advantages and disadvantages.**
  - A topic discovery method purely relying on one type of context information may not be robust across different datasets or seed dimensions.
  - Meanwhile, the three types of contexts strongly **complement each other**.

[1] Blei, D., and McAuliffe, J. (2007). Supervised topic models. NIPS.

SeedTopicMine: Overview

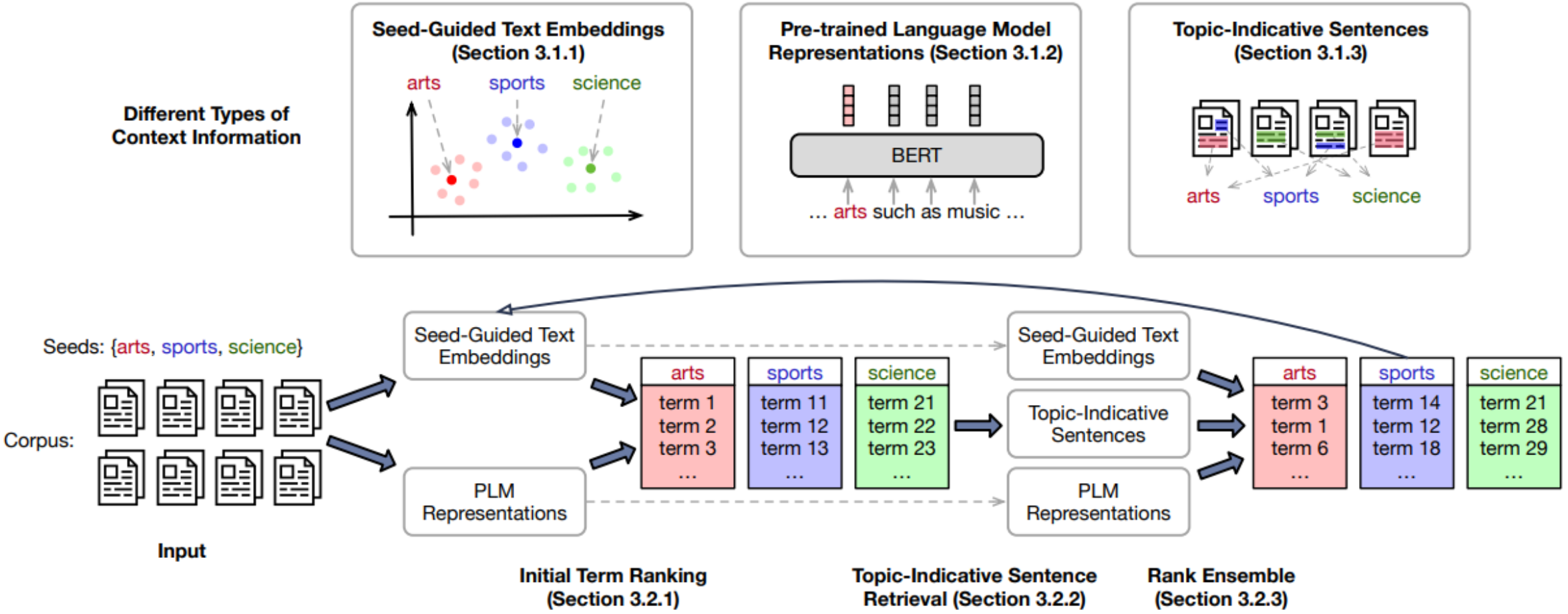
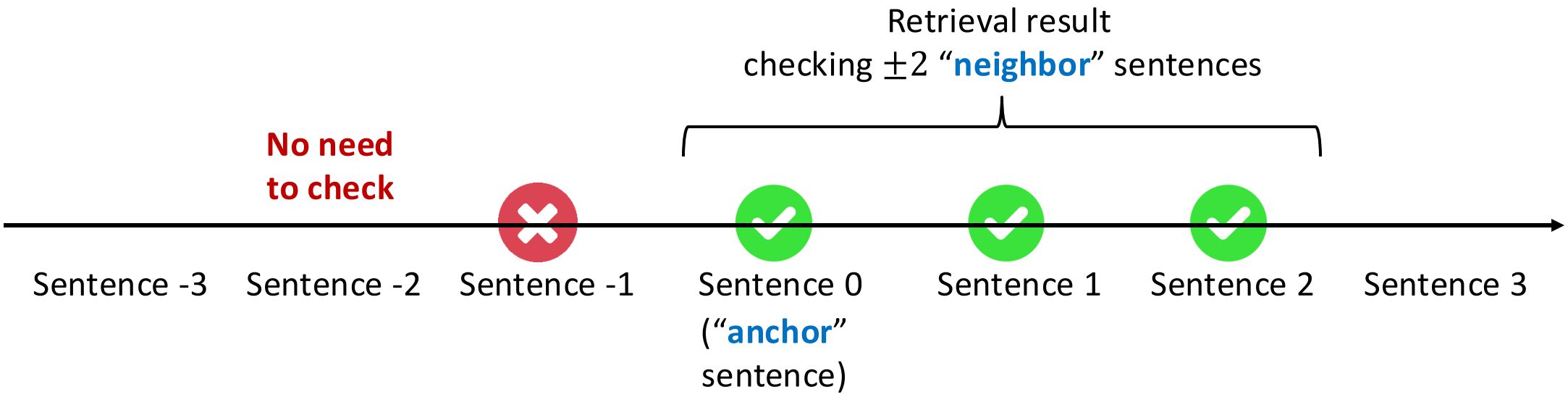


Figure 1: Overview of the SeedTopicMine framework.

Zhang, Y., Zhang, Y., Michalski, M., Jiang, Y., Meng, Y., & Han, J. (2023). Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts. WSDM.

SeedTopicMine: Topic-Indicative Sentence Retrieval

- The sentences containing many topic-indicative terms from one category and do not contain any topic-indicative term from other categories should be topic-indicative sentences. We call such sentences "**anchor**" sentences.
- The "**neighbor**" sentences of topic-indicative "anchor" sentences should be included in topic-indicative sentences as well if they do not contain topic-indicative terms from other categories.



Quantitative Results

Table 2: NPMI, P@20, and NDCG@20 scores of compared algorithms. NPMI measures topic coherence; P@20 and NDCG@20 measure term accuracy.

Method	**NYT**-Topic			**NYT**-Location			**Yelp**-Food			**Yelp**-Sentiment		
	NPMI	P@20	NDCG@20	NPMI	P@20	NDCG@20	NPMI	P@20	NDCG@20	NPMI	P@20	NDCG@20
SeededLDA [15]	0.0841	0.2389	0.2979	0.0814	0.1050	0.1873	0.0504	0.1200	0.2132	0.0499	0.1700	0.2410
Anchored CorEx [10]	0.1325	0.2922	0.3627	0.1283	0.2040	0.3003	0.1204	0.3725	0.4531	0.0627	0.1200	0.1997
KeyETM [13]	0.1254	0.1589	0.2342	0.1146	0.0700	0.1676	0.0578	0.1788	0.2940	0.0327	0.4250	0.4994
CatE [27]	0.1941	0.8067	0.8306	0.2165	0.7480	0.7840	**0.2058**	0.6812	0.7312	**0.1509**	0.7150	0.7713
SeedTopicMine	**0.1947**	**0.9456**	**0.9573**	**0.2176**	**0.8360**	**0.8709**	0.2018	**0.7912**	**0.8379**	0.0922	**0.9750**	**0.9811**

Method	**Yelp**-Food		**Yelp**-Sentiment	
	P@20	NDCG@20	P@20	NDCG@20
SeedTopicMine	**0.7912**	**0.8379**	**0.9750**	**0.9811**
SeedTopicMine-NoEmb	0.4488	0.5335	0.9550	0.9646
SeedTopicMine-NoPLM	0.6962	0.7602	0.7550	0.8029
SeedTopicMine-NoSntn	0.7488	0.8029	0.9500	0.9631

- Three types of contexts all have positive contribution.
- Even for the same dataset (i.e., Yelp), the contribution of a certain type of context information varies significantly with the input seeds. Therefore, it becomes necessary to **integrate them together** to make the framework more robust.

Qualitative Results

Table 3: Top-5 terms retrieved by different algorithms. ×: At least 3 of the 5 annotators judge the term as irrelevant to the seed.

Method	**NYT**-Topic		**NYT**-Location		**Yelp**-Food		**Yelp**-Sentiment	
	health	business	france	canada	sushi	desserts	good	bad
SeededLDA	said (×)	said (×)	said (×)	new (×)	roll	food (×)	place (×)	food (×)
	dr (×)	percent (×)	new (×)	city (×)	good (×)	us (×)	food (×)	service (×)
	new (×)	company	state (×)	said (×)	place (×)	order (×)	great	us (×)
	would (×)	year (×)	would (×)	building (×)	food (×)	service (×)	like (×)	order (×)
	hospital	billion (×)	dr (×)	mr (×)	rolls	time (×)	service (×)	time (×)
Anchored CorEx	case (×)	employees	school (×)	market (×)	rolls	also (×)	definitely (×)	one (×)
	court (×)	advertising	students (×)	percent (×)	roll	really (×)	prices (×)	would (×)
	patients	media (×)	children (×)	companies (×)	sashimi	well (×)	strip (×)	like (×)
	cases (×)	businessmen	education (×)	billion (×)	fish (×)	good (×)	selection (×)	could (×)
	lawyer (×)	commerce	schools (×)	investors (×)	tempura	try (×)	value (×)	us (×)
KeyETM	team (×)	percent (×)	city (×)	people (×)	sashimi	food (×)	great	food (×)
	game (×)	japan (×)	state (×)	year (×)	rolls	great (×)	delicious	place (×)
	players (×)	year (×)	york (×)	china (×)	roll	place (×)	amazing	service (×)
	games (×)	japanese (×)	school (×)	years (×)	fish (×)	good (×)	excellent	time (×)
	play (×)	economy	program (×)	time (×)	japanese	service (×)	tasty	restaurant (×)
CatE	public health	diversifying (×)	french	alberta	freshest fish (×)	delicacies (×)	tasty	unforgivable
	health care	clients (×)	corsica	british columbia	sashimi	sundaes	delicious	frustrating
	medical	corporate	spain (×)	ontario	nigiri	savoury (×)	yummy	horrible
	hospitals	investment banking	belgium (×)	manitoba	ayce sushi	pastries	chilaquiles (×)	irritating
	doctors	executives	de (×)	canadian	rolls	custards	also (×)	rude
SeedTopicMine	medical	companies	french	canadian	maki rolls	cheesecakes	great	terrible
	hospitals	businesses	paris	quebec	sashimi	croissants	excellent	horrible
	hospital	corporations	philippe (×)	montreal	ayce sushi	pastries	fantastic	awful
	public health	firms	french state	toronto	revolving sushi	breads (×)	delicious	lousy
	patients	corporate	frenchman	ottawa	nigiri	cheesecake	amazing	shitty

Taxonomy Enrichment with LLMs

- LLMs can also enrich a taxonomy structure with its general knowledge
- Hierarchy-aware prompting for taxonomy enrichment:
  - For each subtree structure, instruct LLM to enrich a node under the context of its parent and contrast with its siblings

Instruction: [Target Class] is a product class in Amazon and is the subclass of [Parent Class]. Please generate 10 additional key terms about the [Target Class] that is relevant to [Target Class] but irrelevant to [Sibling Classes]. Please split the additional key terms using commas.

Class	Enrichment
hair care	hair mask, hair oil, hair styling, …
shampoos	hair hygiene, scalp care, dandruff, …
conditioners	hydrating, nourishing, smoothing, …
health care	pharmaceuticals, health insurance, …

Zhang, Y., Yang, R., Xu, X., Li, R., Xiao, J., Shen, J., Han, J., "TELEClass: Taxonomy Enrichment and LLM-Enhanced Hierarchical Text Classification with Minimal Supervision", arXiv'24

References

- Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2003). Hierarchical topic models and the nested Chinese restaurant process. NIPS.
- Blei, D. M., & McAuliffe, J. D. (2007). Supervised topic models. NIPS.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research.
- Huang, J., Xie, Y., Meng, Y., Zhang, Y., & Han, J. (2020). CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring. KDD.
- Jiang, S., Yao, Q., Wang, Q., & Sun, Y. (2023). A Single Vector Is Not Enough: Taxonomy Expansion via Box Embeddings. WWW.
- Lee, D., Shen, J., Kang, S., Yoon, S., Han, J., & Yu, H. (2022). TaxoCom: Topic Taxonomy Completion with Hierarchical Discovery of Novel Topic Clusters. WWW.
- Meng, Y., Huang, J., Wang, G., Wang, Z., Zhang, C., Zhang, Y., & Han, J. (2020). Discriminative topic mining via category-name guided text embedding. WWW.
- Meng, Y., Zhang, Y., Huang, J., Zhang, Y., Zhang, C., & Han, J. (2020). Hierarchical topic mining via joint spherical tree and text embedding. KDD.
- Meng, Y., Zhang, Y., Huang, J., Zhang, Y., & Han, J. (2022). Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations. WWW.

References

- Shen, J., Shen, Z., Xiong, C., Wang, C., Wang, K., & Han, J. (2020). TaxoExpan: Self-supervised Taxonomy Expansion with Position-Enhanced Graph Neural Network. WWW.
- Shen, Y., Zhang, Y., Zhang, Y., & Han, J. (2024). A Unified Taxonomy-Guided Instruction Tuning Framework for Entity Set Expansion and Taxonomy Expansion. arXiv
- Thompson, L., and Mimno, D. (2020). Topic modeling with contextualized word representation clusters. arXiv.
- Zeng, Q., Bai, Y., Tan, Z., Feng, S., Liang, Z., Zhang, Z., Jiang, M. (2024) Chain-of-Layer: Iteratively Prompting Large Language Models for Taxonomy Induction from Limited Examples. arXiv
- Zhang, Y., Zhang, Y., Michalski, M., Jiang, Y., Meng, Y., & Han, J. (2023). Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts. WSDM.
- Zhang, Y., Shen, J., Shang, J., & Han, J. (2020). Empower Entity Set Expansion via Language Model Probing. ACL.
- Zhang, Y., Yang, R., Xu, X., Li, R., Xiao, J., Shen, J., Han, J., "TELEClass: Taxonomy Enrichment and LLM-Enhanced Hierarchical Text Classification with Minimal Supervision", arXiv'24

Q&A

Tutorial Website: