Part IV: Mining Entity Structures: Taxonomy and Knowledge Base Construction

EDBT 2023 Tutorial: Mining Structures from Massive Texts by Exploring the Power of Pre-trained Language Models

Yu Zhang, Yunyi Zhang, Jiawei Han

Department of Computer Science, University of Illinois at Urbana-Champaign

Mar 29, 2023

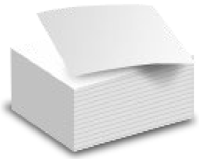
Outline

- Phrase Mining
  - UCPhrase: Unsupervised Context-aware Quality Phrase Tagging [KDD’21]
- Named Entity Recognition
- Taxonomy Construction
- Relation Extraction and Knowledge Graph Construction

Why Phrase Mining?

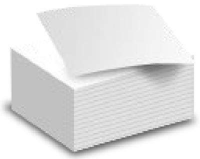
- Identifying and understanding quality phrases from context is a fundamental task in text mining.

Scientific Papers



Expected Results
data mining machine learning information retrieval … support vector machine …

News Articles



Expected Results
US President Anderson Cooper Barack Obama … Obama administration …

- Quality phrases refer to informative multi-word sequences that "*appear consecutively in the text, forming a complete semantic unit in certain contexts or the given document*" [1].

[1] Geoffrey Finch. 2016. Linguistic terms and concepts. Macmillan International Higher Education

Why Phrase Mining?

w/o phrase mining

- What's "United"?
- Who's "Dao"?

w/ phrase mining

- United Airline!
- David Dao!

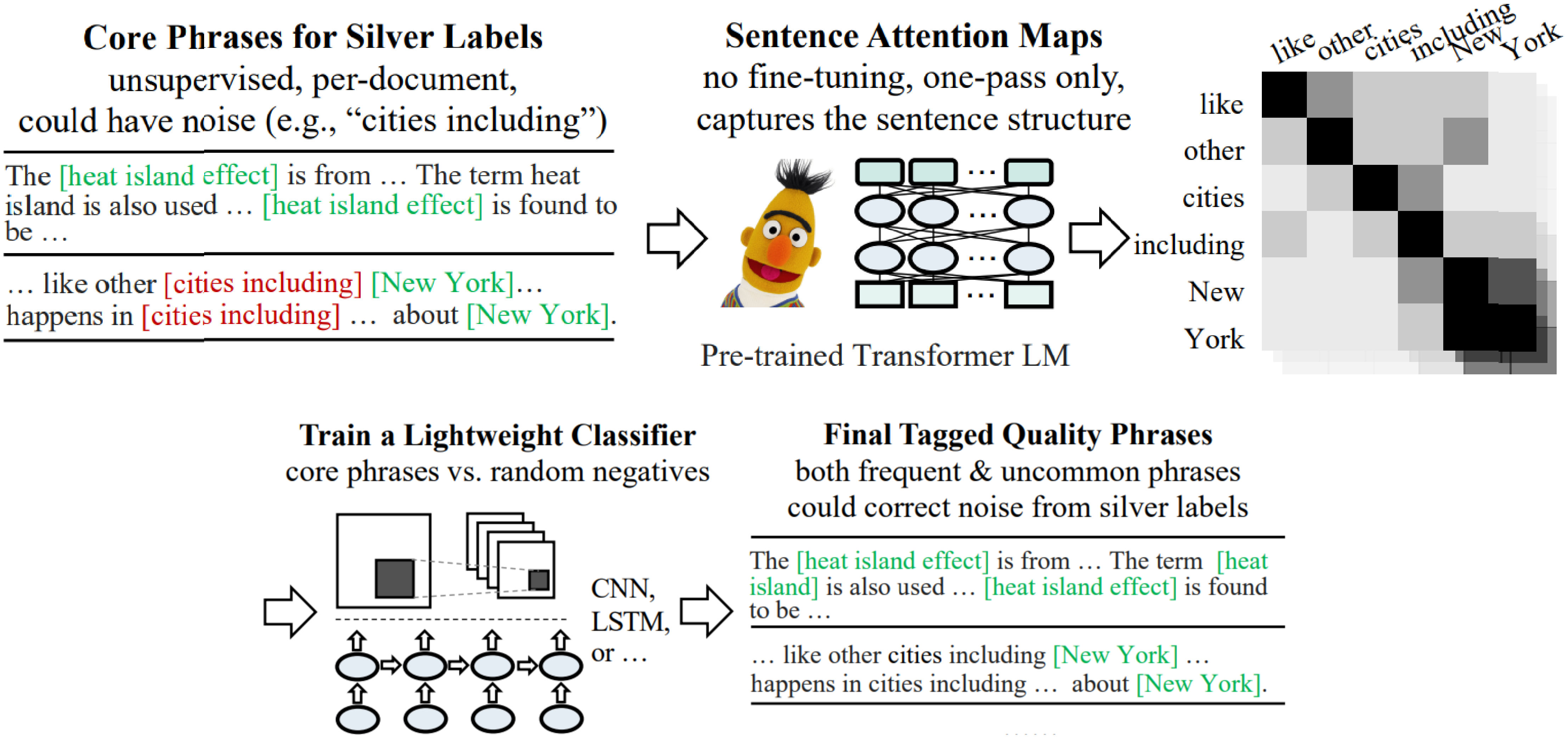
- Applications in NLP, IR, Text Mining
  - Text Classification
  - Indexing in search engine
  - Keyphrases for topic modeling
  - Text Summarization

Previous Phrase Mining/Chunking Models

- Statistics-based models (*TopMine, SegPhrase, AutoPhrase)*
  - only work for frequent phrases, ignore valuable **infrequent / emerging phrases**
- Tagging-based models (*Spacy, StanfordNLP*)
  - do not have requirements for frequency
  - require **expensive and unscalable** sentence-level annotations for model training

Framework of UCPhrase

- Silver Label Generation + Attention Map-based Span Prediction

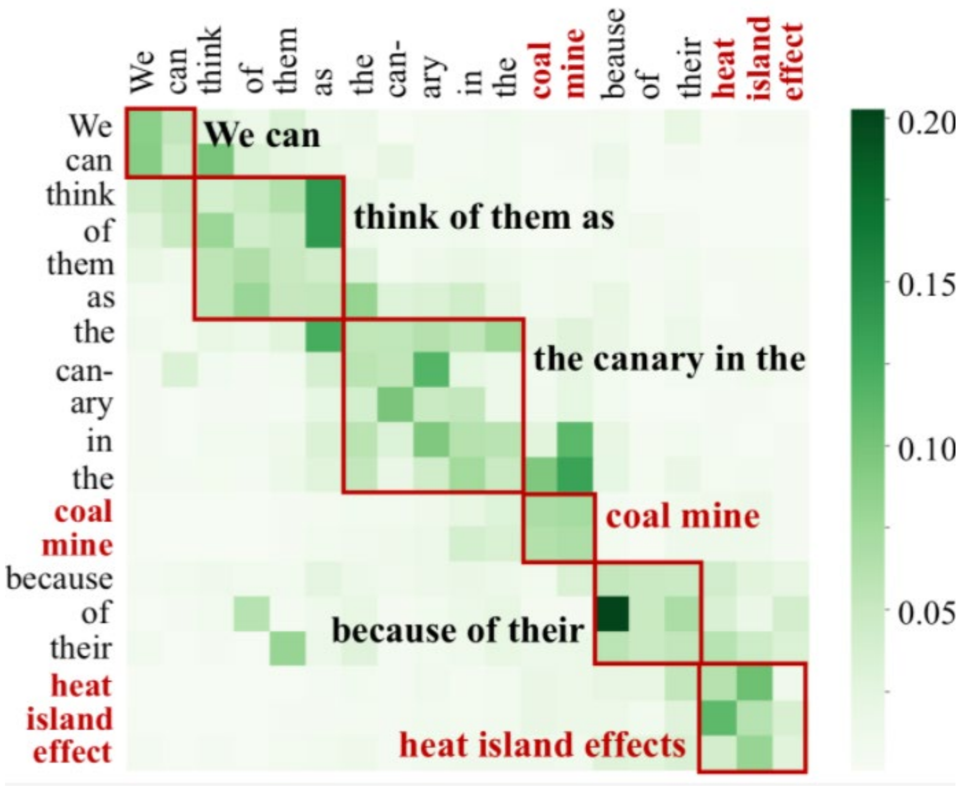


Silver Label Generation

- How do human readers accumulate new phrases?
  - even without any prior knowledge we can recognize these consistently used patterns from a document
  - e.g., *task name, method name, dataset name, concepts* in a publication
  - e.g., *human name, organization, locations* in a news article
- Mining core phrases as silver labels
  - independently mine **max word sequential patterns** within each document
  - with each document as context
    - preserve contextual completeness ("biomedical data mining" vs. "data mining")
    - avoid potential noises from propagating to the entire corpus

Attention Map as Surface-Agnostic Feature

- Good features for phrase recognition should be
  - agnostic to word **surface names** (so the model cannot rely on rigid memorization)
  - focusing on **sentence structure** rather than phrase names
- Extract knowledge directly from a pre-trained language model
  - the **attention map** of a sentence vividly visualizes its **inner structure**
  - high quality phrases should have **distinct attention patterns** from ordinary spans
- Phrase Tagging as Image Classification
  - train a lightweight 2-layer CNN model for binary classification: is a phrase or not



Quantitative Evaluation

Table 2: Evaluation results (%) of three tasks for all compared methods on datasets on two domains.

Method Type	Method Name	Task I: Phrase Ranking				Task II: KP Extract.				Task III: Phrase Tagging					
		KP20k		KPTimes		KP20K		KPTimes		KP20k			KPTimes		
		P@5K	P@50K	P@5K	P@50K	Rec.	F_1@10	Rec.	F_1@10	Prec.	Rec.	F_1	Prec.	Rec.	F_1
Pre-trained	PKE [3]	–	–	–	–	57.1	12.6	61.9	4.4	54.1	63.9	58.6	56.1	62.2	59.0
	Spacy [16]	–	–	–	–	59.5	15.3	60.8	8.6	56.3	68.7	61.9	61.9	62.9	62.4
	StanfordNLP [26]	–	–	–	–	51.7	13.9	60.8	8.7	48.3	60.7	53.8	56.9	60.3	58.6
Distantly Supervised	AutoPhrase [33]	97.5	96.0	96.5	95.5	62.9	18.2	77.8	10.3	55.2	45.2	49.7	44.2	47.7	45.9
	Wiki+RoBERTa	**100.0**	**98.5**	**99.0**	**96.5**	**73.0**	19.2	64.5	9.4	58.1	64.2	61.0	60.9	65.6	63.2
Unsupervised	TopMine [8]	81.5	78.0	85.5	71.0	53.3	15.0	63.4	8.5	39.8	41.4	40.6	32.0	36.3	34.0
	UCPhrase (ours)	96.5	96.5	96.5	95.5	72.9	**19.7**	**83.4**	**10.9**	**69.9**	**78.3**	**73.9**	**69.1**	**78.9**	**73.5**

Outline

- Phrase Mining
- Named Entity Recognition (NER)
  - Few-shot NER and Entity Typing
    - Few-Shot Fine-Grained Entity Typing with Automatic Label Interpretation and Instance Generation [KDD' 2022]
  - Distantly-supervised NER
- Taxonomy Construction
- Relation Extraction and Knowledge Graph Construction

Named Entity Recognition (NER)

- A **named entity** typically refers to a sequence of words that correspond to a specific entity in the real world (i.e., an entity with a *name*) (e.g., "*Bill Clinton*")
- **Named-entity recognition (NER)** is a subtask of **information extraction (IE)** that seeks to **locate** and **classify named entities** in text into **pre-defined categories**
  - Given a sentence, NER is to first *segment which words are part of entities*, and then *classify each entity by type* (person, organization, location, and so on)
  - Example
    - Input: Jim bought 300 shares of Acme Corp. in 2006
    - Output: [Jim]$_{\text{Person}}$ bought 300 shares of [Acme Corp.]$_{\text{Organization}}$ in [2006]$_{\text{Time}}$
- Most NER methods focus on three types of entities: *person*, *location*, and *organization. Some also include dates*, *times*, *monetary values*, and *percentages*
  - Also, *biological entities* (in bio-domain), or *product names* (for online advertising)

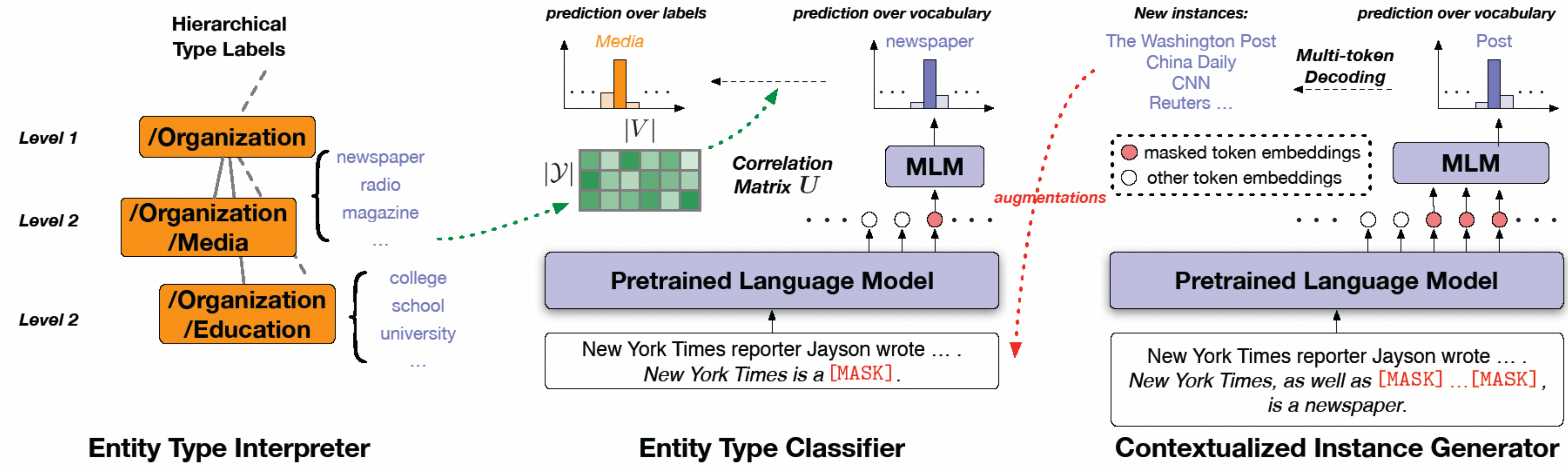
Motivation

- Deep neural models have achieved enormous success for NER
- However, a common bottleneck of training deep learning models is the acquisition of abundant high-quality human annotations
- **Few-shot NER** learns to transfer to new domains/categories with only a few training examples.

Limitations of current pipeline

- ❑ Current approaches have not fully utilized the power of PLMs
- ☑ **representation** models that predict entity types based on entity instance representations
- ☐ the **generation** power of PLMs acquired through extensive general-domain pretraining can be exploited to generate new entity instances
- ☐ model can be trained with more instances for better generalization

Overall Framework of ALIGNIE (Automatic Label Interpretation and Generating New Instance for Entity typing)



(Left): With a given type label hierarchy, an entity type interpretation module relates all the words in the vocabulary with the label hierarchy by a correlation matrix.

(Middle): An entity typing classifier maps the word probability at the [MASK] position to type probability using the correlation matrix.

(Right): A type-based contextualized instance generator uses an entity mention and its predicted type to construct a template for new instance generation to augment the training set.

PLM-based Instance Generator

- E.g., a *newspaper* entity "New York Times" → more newspaper names

Generation Template :

$\underbrace{\text{[Context]. } \textbf{New York Times}}_{\text{Entity Mention}}$, as well as $\underbrace{\text{[MASK] [MASK] [MASK]}}_{\text{\# ranges from 1 to the length of original entity mention}}$, is a $\underbrace{\textit{newspaper}}_{\text{Predicted by Entity Type Classifier}}$.

Multi-Token Instance Generation

- We randomly choose one [MASK] token at each step, and sample from its output token probability to fill in a word.

E.g.

New York Times, as well as the$_1$ [MASK] [MASK] is a newspaper.
New York Times, as well as the$_1$ Washington$_2$ [MASK] is a newspaper.
New York Times, as well as the$_1$ Washington$_2$ Post$_3$ is a newspaper.

$$\mathrm{Score}(\widetilde{\boldsymbol{m}}) = \sum_{i=1}^{|\widetilde{\boldsymbol{m}}|} \log(s_i)$$

The conditional probability at each step

The next blank to be filled in is randomly selected, therefore the order is not always from left to right.

Generated New instances based on predicted types of example entities

- Multi-token instances

Generation from **multi-token** entities		
Context & **entity mention**	MLM predicted type	Generated new instances
The album also included the song "Vivir Lo Nuestro," a duet with **Marc Anthony**.	singer	Beyonce, Jennifer Lopez, Rihanna, Taylor Swift, Lady Gaga, Michael Jackson, ...
The film was released on August 9, 1925, by **Universal Pictures**.	company	Warner Brothers, Paramount Pictures , Columbia Pictures, Lucasfilm, Hollywood Pictures, ...
Everland hosted 7.5 million guests in 2006, ranking it fourth in Asia behind the two **Tokyo Disney Resort** parks and Universal Studios Japan, while Lotte World attracted 5.5 million guests to land in fifth place.	park	Lotte World, Universal Studios Japan, Shanghai Disney World , Orlando Universal Studios, ...
The site of Drake's landing as officially recognised by the **U.S. Department of the Interior** and other agencies is Drake's Cove.	government agency	the Department of Homeland Security, the Bureau of Land Management, the Federal Bureau of Investigation, the United States Forest Service, the National Institutes of Health, ...
Pikmin also make a cameo during the process of transferring downloadable content from a **Nintendo DSi** to a 3DS, with various types of Pikmin carrying the data over.	handheld	3DS, 2DS, Wii U, Nintendo Switch, the PSP, PlayStation Vita, ...

Main Results

Method	OntoNotes (Acc.)	OntoNotes (Micro-F1)	OntoNotes (Macro-F1)	BBN (Acc.)	BBN (Micro-F1)	BBN (Macro-F1)	Few-NERD (Acc.)	Few-NERD (Micro-F1)	Few-NERD (Macro-F1)
5-Shot Setting									
Fine-tuning	28.60	50.70	51.60	51.03	60.03	58.22	36.09	48.56	48.56
Prompt-based MLM	32.62	60.97	61.82	67.00	75.23	73.55	44.69	59.24	59.24
PLET	48.57	70.63	75.43	71.23	79.22	78.93	56.94	68.81	68.81
ALIGNIE (- hierarchical reg.)	52.74	**77.55**	79.72	72.15	80.35	80.40	59.01	70.91	70.91
ALIGNIE (- new instances)	51.10	72.91	76.88	73.50	81.62	81.31	57.41	69.47	69.47
ALIGNIE	**53.37**	77.21	**80.68**	**75.44**	**82.20**	**82.30**	**59.72**	**71.90**	**71.90**
Fully Supervised Setting									
Fine-tuning	56.70	75.21	78.86	78.06	82.39	82.60	79.75	85.74	85.74
Prompt-based MLM	55.18	74.57	77.47	77.10	81.77	82.05	77.38	85.22	85.22

- Prompt-based results have higher performance than vanilla fine-tuning in few-shot settings. In fully supervised settings, however, fine-tuning performs a little better than prompt-based MLM.
- ALIGNIE can even outperform fully supervised setting on OntoNotes and BBN, but cannot on Few-NERD. This is because the training set of OntoNotes and BBN are automatically inferred from external knowledge bases, and can contain much noise.

Outline

- Phrase Mining
- Named Entity Recognition (NER)
  - Few-shot NER
  - Distantly-supervised NER
    - Distantly-Supervised Named Entity Recognition with Noise-Robust Learning and Language Model Augmented Self-Training [EMNLP’2021]
- Taxonomy Construction
- Relation Extraction and Knowledge Graph Construction

Challenge

- The biggest challenge of distantly-supervised NER is that the distant supervision may induce **incomplete and noisy labels,** because
  - the distant supervision source has **limited coverage** of the entity mentions in the target corpus
  - some entities can be matched to multiple types in the knowledge bases---such **ambiguity** cannot be resolved by the context-free matching process
- Straightforward application of supervised learning will lead to deteriorated model performance, as neural models have the strong capacity to fit to the given (noisy) data

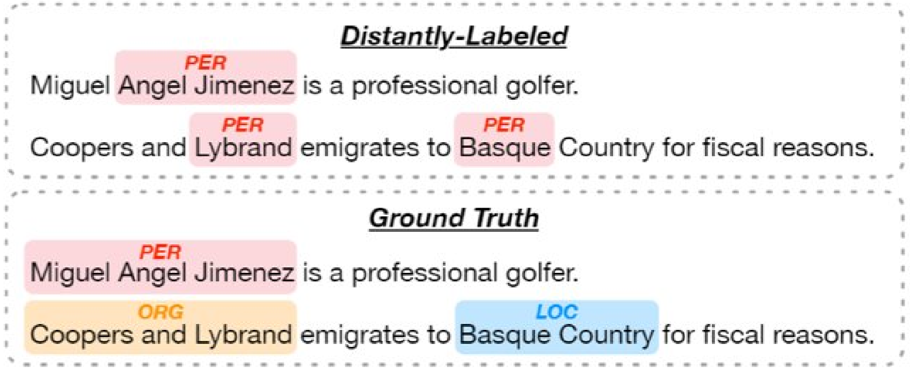
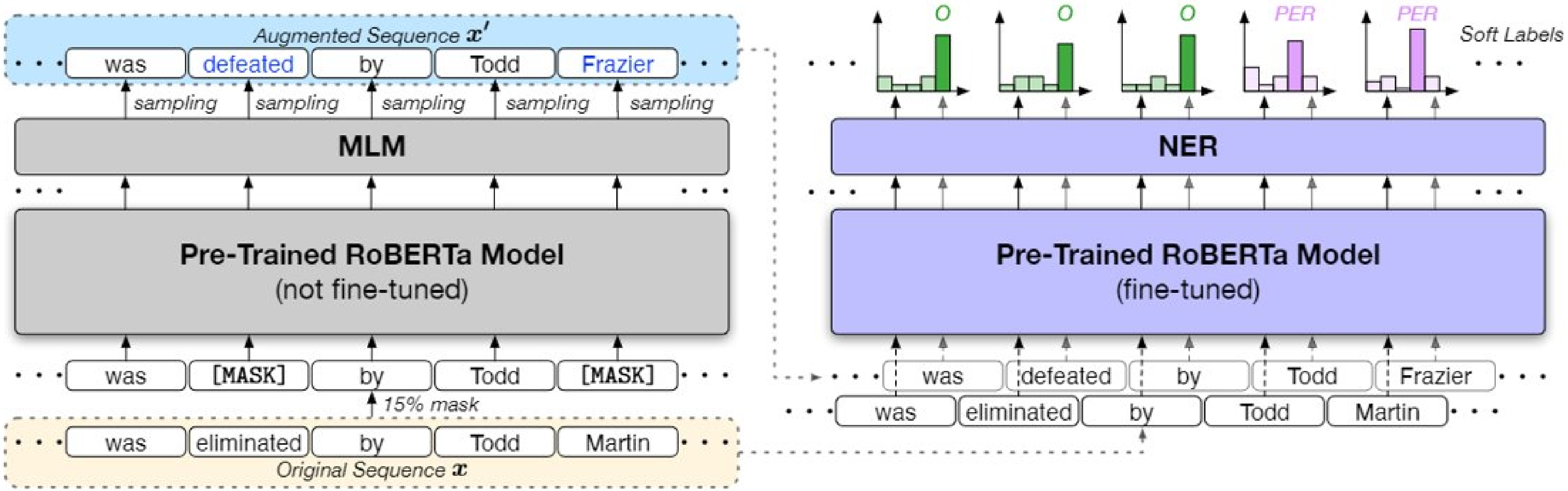


Figure 1: Distant labels obtained with knowledge bases may be incomplete and noisy, resulting in wrongly-labeled tokens.

RoSTER

- RoSTER: Distantly-Supervised Named Entity Recognition with Noise-Robust Learning and Language Model Augmented Self-Training [EMNLP'21]



Method

- Noise-Robust Learning: Why straightforward application of supervised NER learning on noisy data is bad?
- When the labels are noisy, training with the Cross Entropy (CE) loss can cause **overfitting** to the **wrongly-labeled** tokens
- Generalized Cross Entropy Loss (GCE)

$$\mathcal{L}_{\mathrm{GCE}} = \sum_{i=1}^{n} w_i \frac{1 - f_{i,y_i}(\boldsymbol{x};\boldsymbol{\theta})^{1-q}}{1-q} \qquad w_i = \mathbb{1}\left(f_{i,y_i}(\boldsymbol{x};\boldsymbol{\theta}) > \tau\right)$$

Only use reliable labels
(model prediction agrees)

- Rationale: Since our loss function is noise-robust, the learned model will be dominated by the **correct majority** in the distant labels instead of quickly overfitting to label noise; if the model prediction disagrees with some given labels, they are potentially wrong

Method

- Contextualized Augmentations with PLMs
- Randomly mask out 15% of tokens in the original sequence
- Feed the partially masked sequence into the pre-trained RoBERTa model
- Augmented sequence is created by sampling from the MLM output probability for each token
- Further enforce the label-preserving constraint:
  - sample only from the top-5 terms of MLM outputs
  - if the original token is capitalized or is a subword, so should the augmented one

Experiment Results

- Main Results

	Methods	CoNLL03			OntoNotes5.0			Wikigold		
		Pre.	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**
Distant-Sup.	**Distant Match**	0.811	0.638	0.714	0.745	0.693	0.718	0.479	0.476	0.478
	Distant RoBERTa	0.837	0.633	0.721	0.760	0.715	0.737	0.603	0.532	0.565
	AutoNER	0.752	0.604	0.670	0.731	0.712	0.721	0.435	0.524	0.475
	BOND	0.821	0.809	0.815	0.774	0.701	0.736	0.534	0.686	0.600
	RoSTER (Ours)	**0.859**	**0.849**	**0.854**	**0.803**	**0.775**	**0.789**	**0.649**	**0.710**	**0.678**
Sup.	**BiLSTM-CNN-CRF**	0.914	0.911	0.912	0.888	0.887	0.887	0.554	0.543	0.549
	RoBERTa	0.906	0.917	0.912	0.886	0.890	0.888	0.853	0.876	0.864

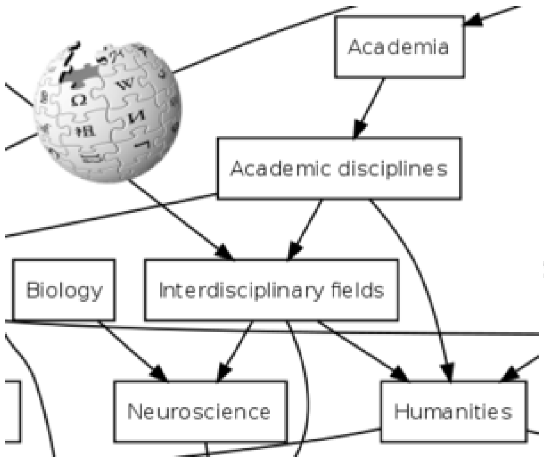
Table 2: Performance all methods on three datasets measured by precision (Pre.), recall (Rec.) and F1 scores.

Outline

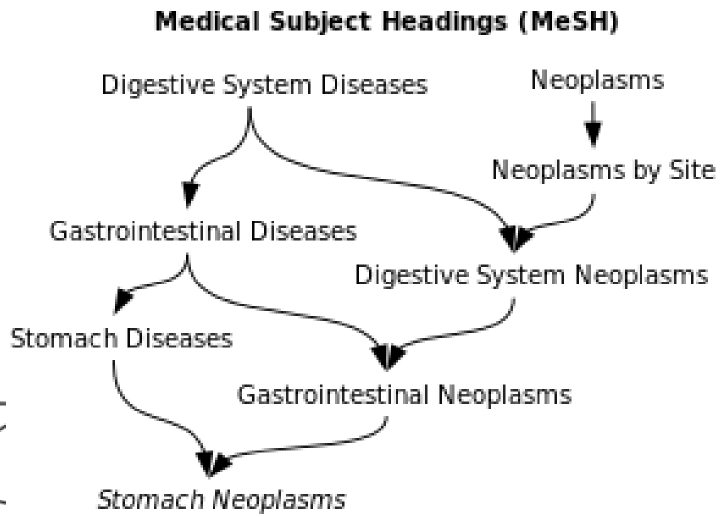
- Phrase Mining
- Named Entity Recognition
- Taxonomy Construction
  - Taxonomy Basics and Construction
  - Set Expansion
  - Taxonomy Construction (with Minimal User Guidance)
  - Taxonomy Expansion & Enrichment
- Relation Extraction and Knowledge Graph Construction

What Is Taxonomy?

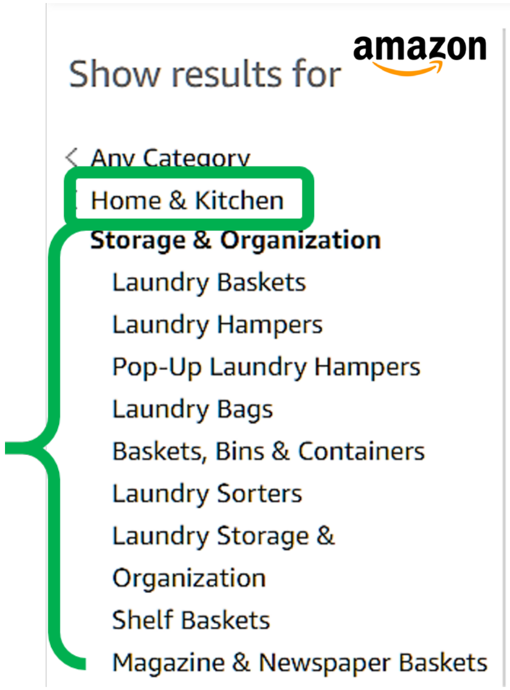
- Taxonomy is a hierarchical (or DAG) organization of concepts
  - Ex.: Wikipedia category, ACM CCS Classification System, Medical Subject Heading (MeSH), Amazon Product Category, Yelp Category List, WordNet, …



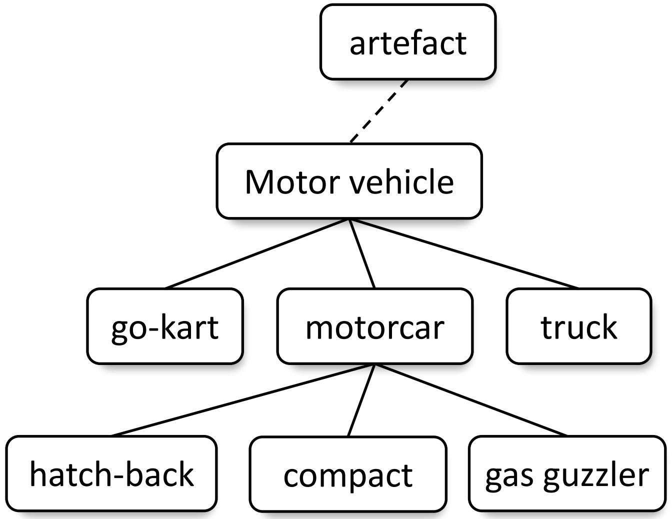
Wikipedia Category



MeSH: PubMed



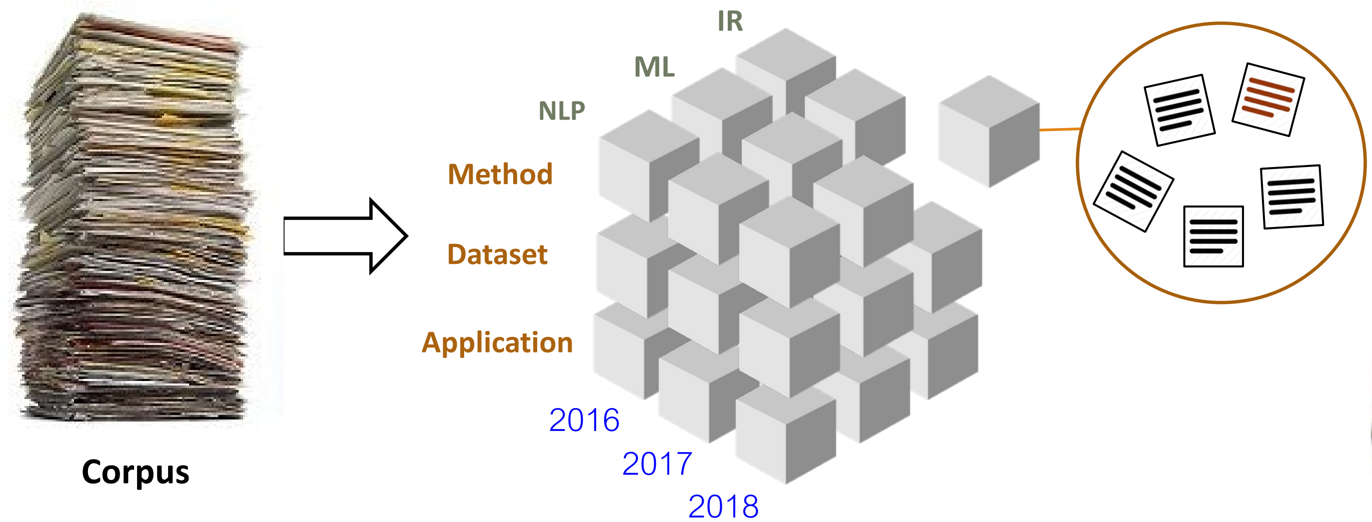
Amazon Product Category



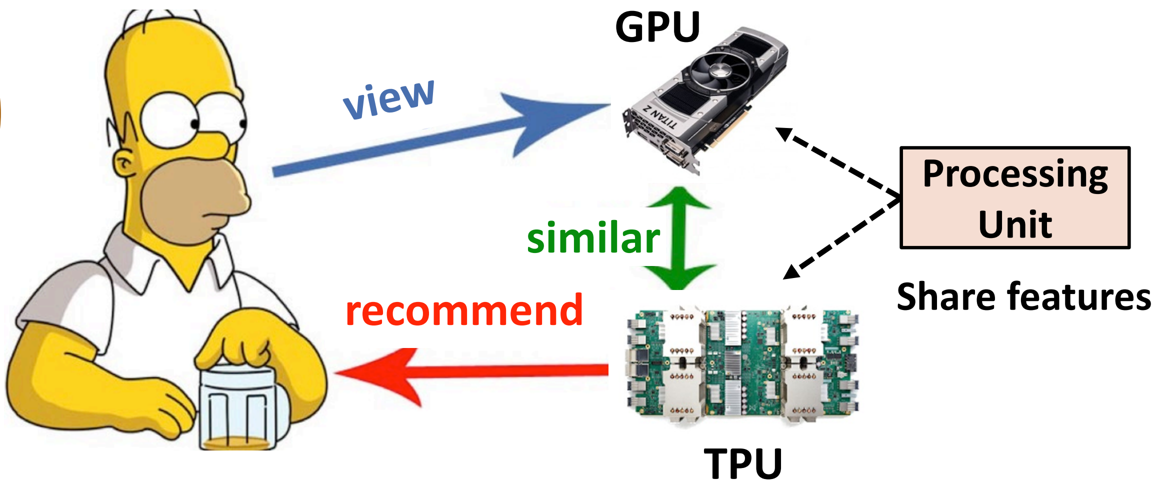
WordNet

Why Do We Need Taxonomy?

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

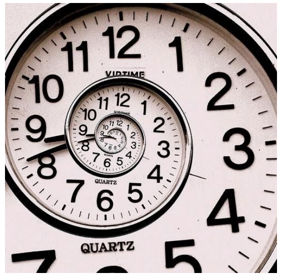


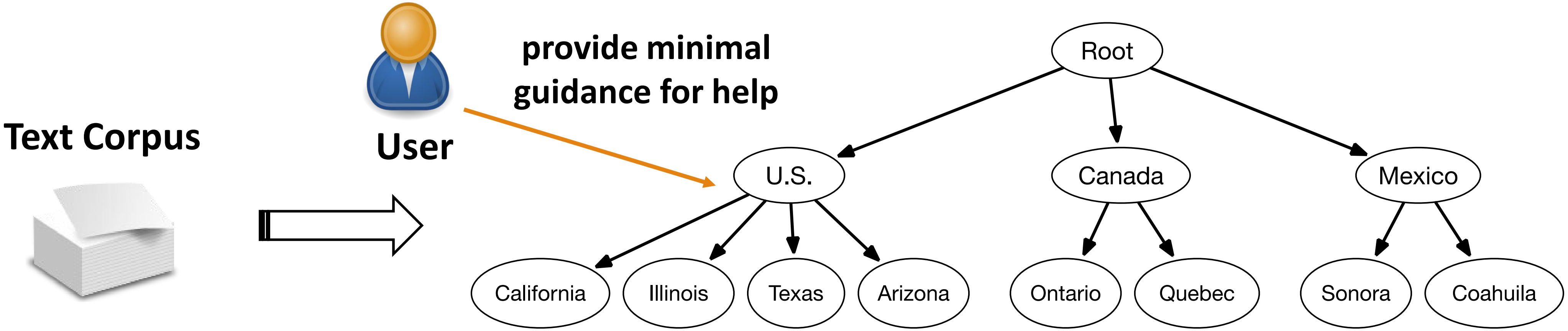
Multi-dimensional Corpus Index



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

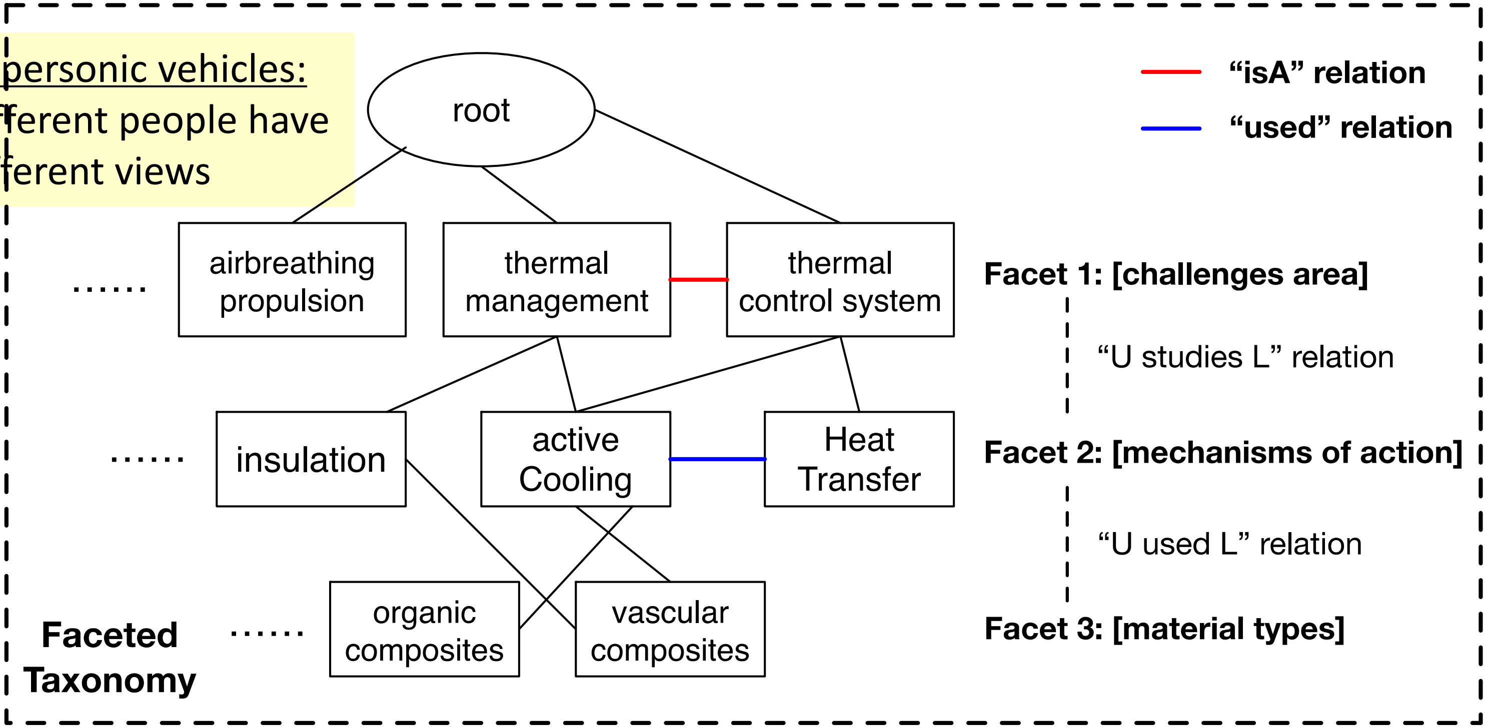


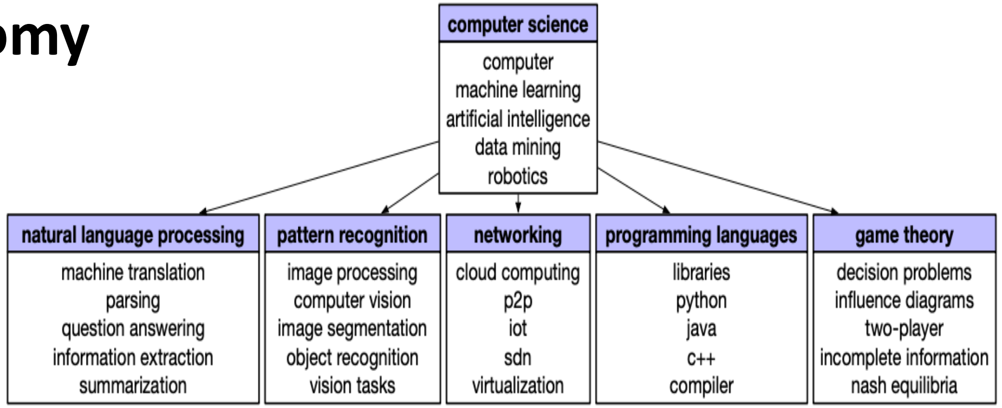


Multi-Faceted Taxonomy

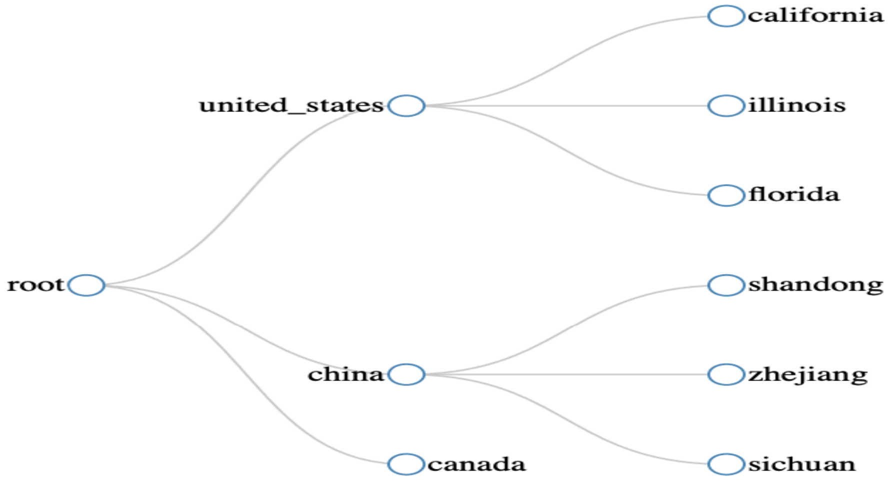
- One facet only reflects a certain kind of relation between parent and child nodes
- Real-world applications need **multi-faceted taxonomy**

Hypersonic vehicles:
Different people have different views





Relation: IsSubfieldOf



Relation: IsLocatedIn

- Help organize, index, and retrieve documents
- Facilitate multi-faceted search
- Conduct analysis at meaningful levels of abstraction

Issues Related to Taxonomy Construction

- Set Expansion
  - Given a few seeds as a set, find other items and expand the set
  - For example, given {*Illinois*, *Maryland*}, derive all U.S. states
- 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 & Enrichment
  - Update an already constructed taxonomy by adding new items on the existing taxonomy

Outline

- Phrase Mining
- Named Entity Recognition
- Taxonomy Construction
  - Taxonomy Basics and Construction
  - Set Expansion
  - Taxonomy Construction (with Minimal User Guidance)
  - Taxonomy Expansion & Enrichment
- Relation Extraction and Knowledge Graph Construction

CGExpan: Probing Language Model for Guidance

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

Seed Set
{Illinois, Georgia, Virginia}

Template
[NP0] such as [NP1], [NP2], and [NP3]

[MASK] such as Illinois, Georgia, Virginia → **Masked Language Models** →

Class-probing query

Candidate Class Names

states
U.S. states
large states
……

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

Class Name
states

Concept
Virginia

Template
[NP0], [NP1], or other [NP2]

Virginia, [MASK], or other states → **Masked Language Models** →

Entity-probing query

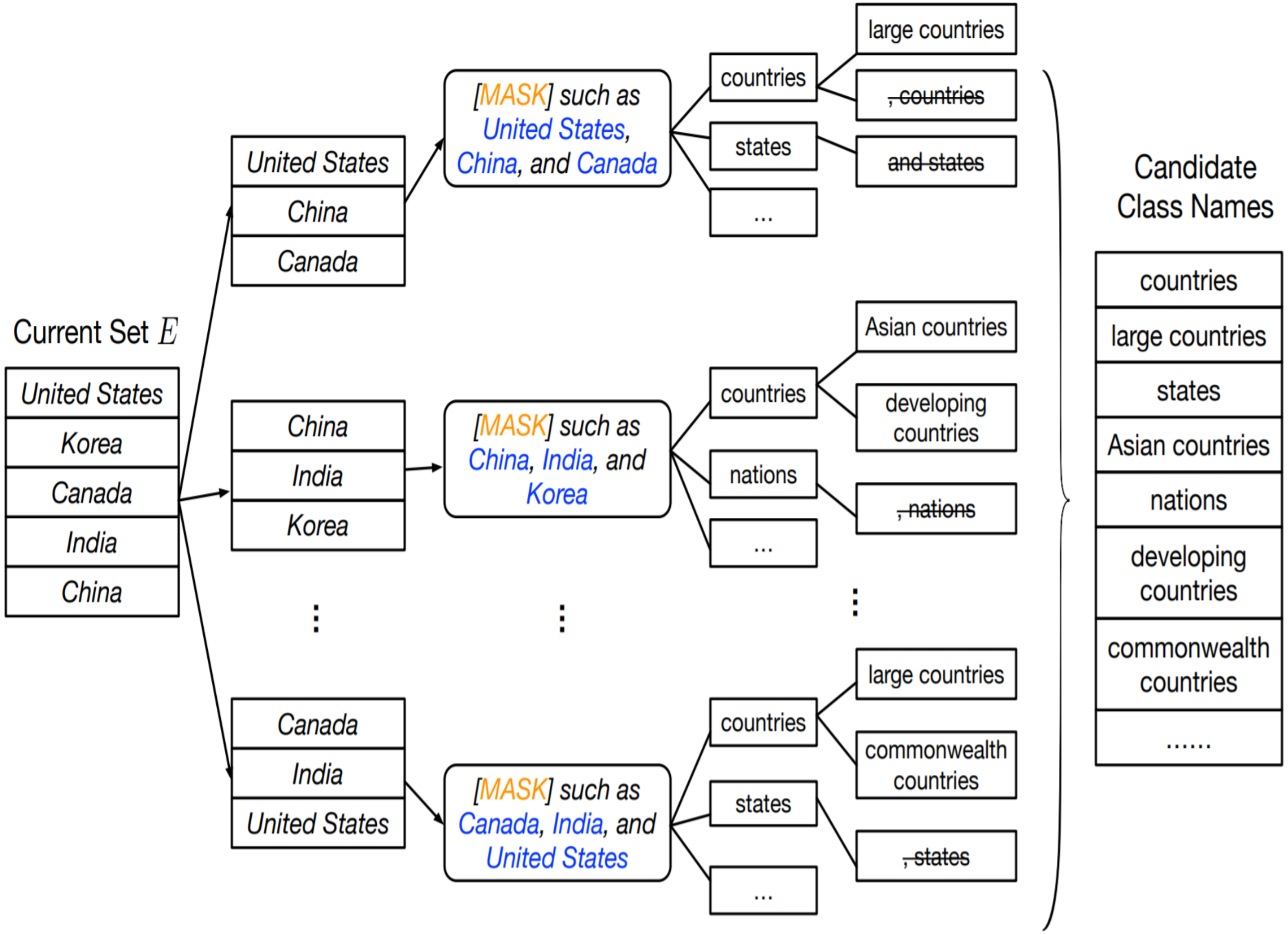
Candidate Concepts

Texas
Florida
Delaware
……

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

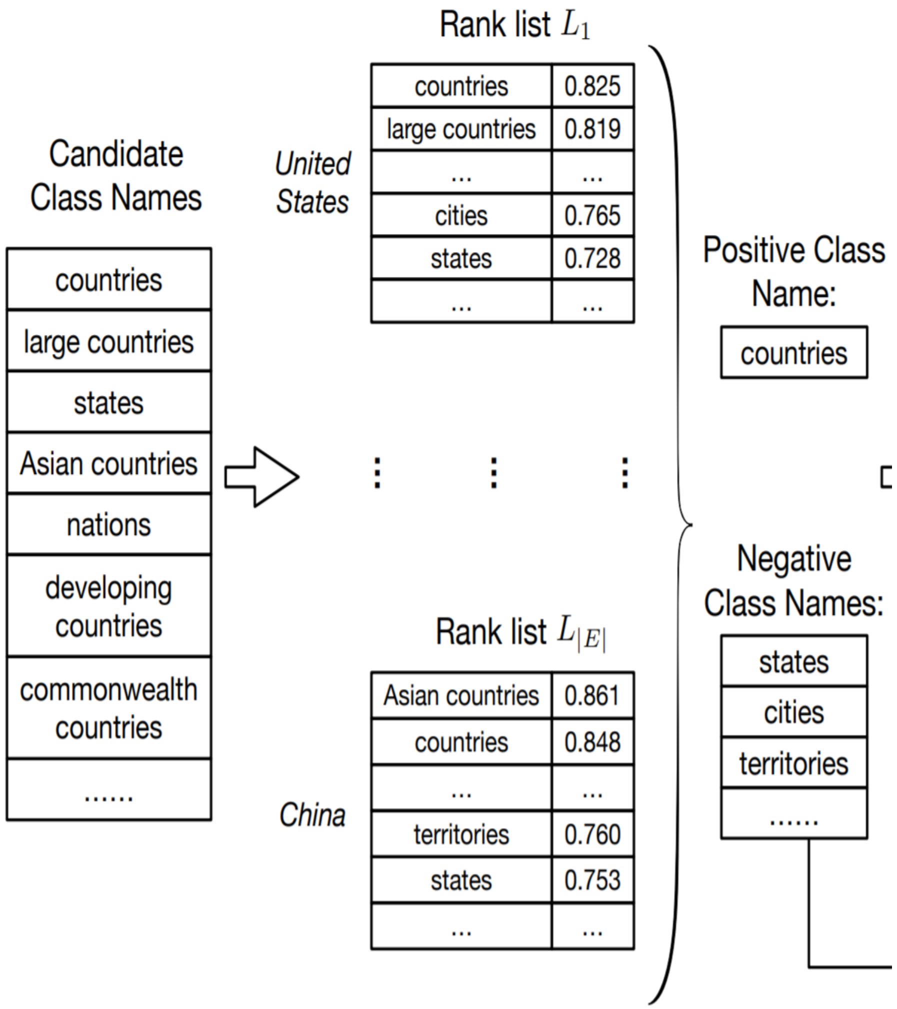
CGExpan 1: Class-Name Generation

- **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



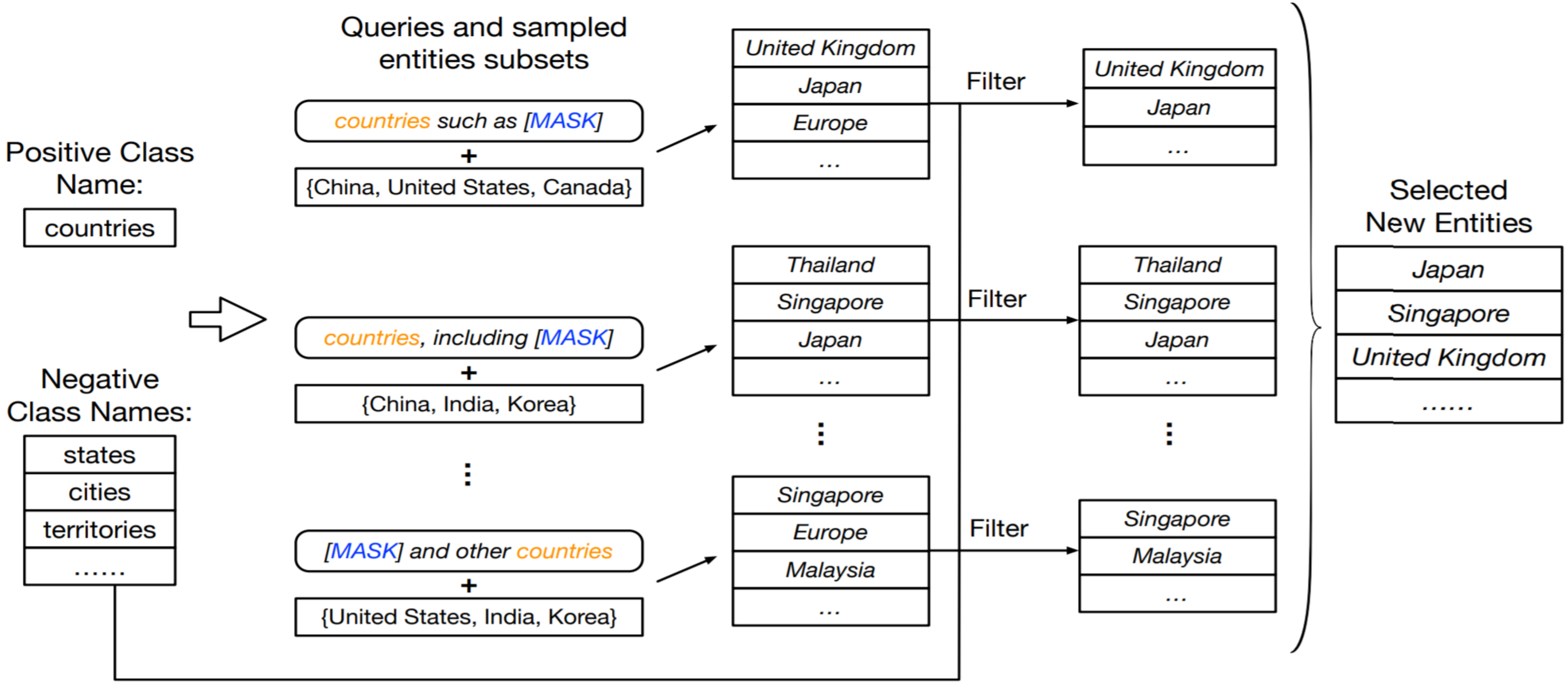
CGExpan 2: Class-Name Ranking

- **Class name ranking:**
  - Build <u>entity-probing queries</u> for each candidate class
  - Compare the retrieved results with seed set to score each class name
  - Rank the class names: select one best class name and several negative ones



CGExpan 3: Class-Guided Entity Selection

- **Class-guided entity selection** (by Rank ensemble)
  - Retrieve and score entities (including those currently in the expanded set) based on <u>entity probing queries</u> and selected class names
  - Select top-rank entities to expand the set



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
	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
	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
	SetCoExpan (WWW'20)	0.964	**0.905**	0.915	0.830
	CGExpan (ACL'20)	**0.978**	0.902	**0.990**	**0.955**

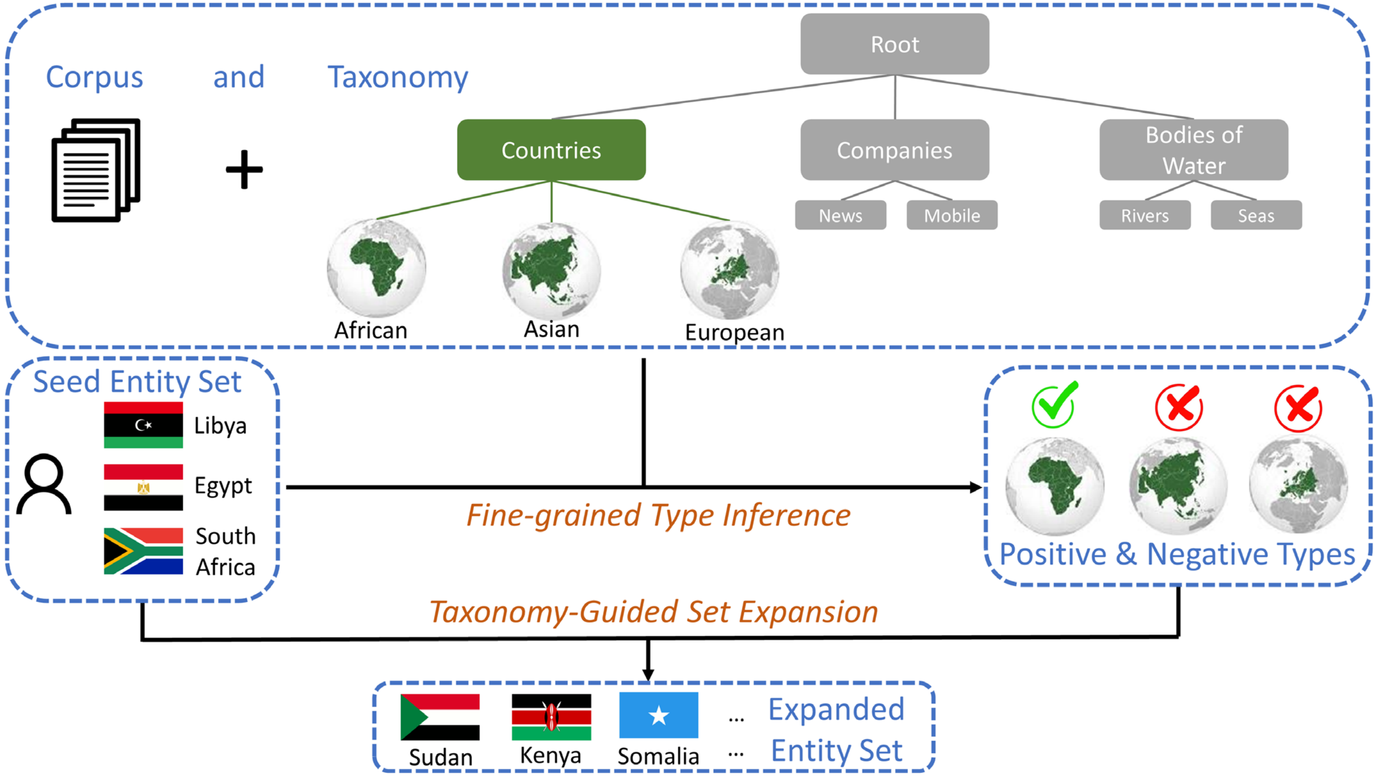
MAP@K: Mean Average Precision truncated at position K

- **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)

FGExpan: Fine-Grained Set Expansion

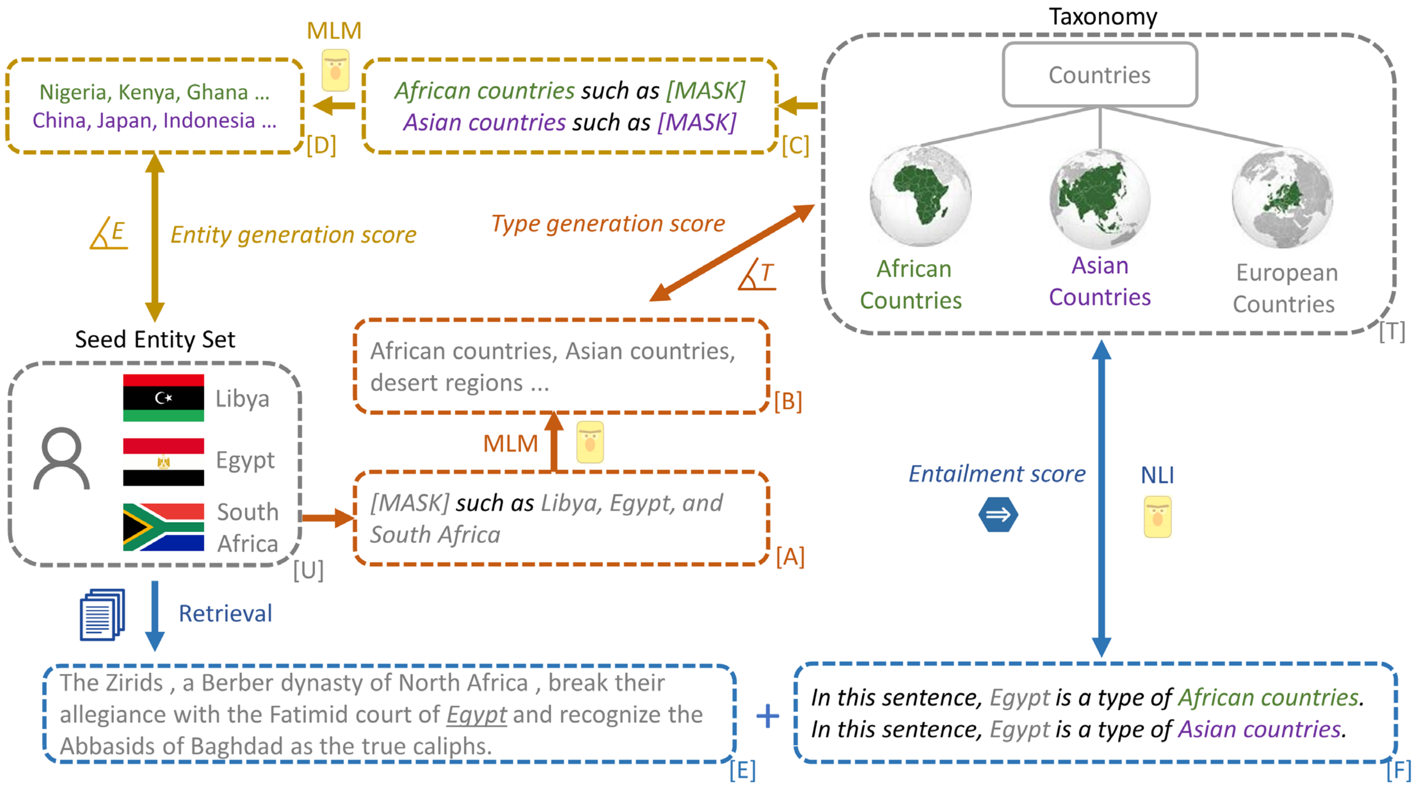
- Expanding entity sets at the finest possible granularity on a type taxonomy
  - E.g., If the seeds are all African countries, then we should not add countries on other continents into the expanded set



Jinfeng Xiao, Mohab Elkaref, Nathan Herr, Geeth De Mel, and Jiawei Han. "Taxonomy-Guided Fine-Grained Entity Set Expansion" SDM'23

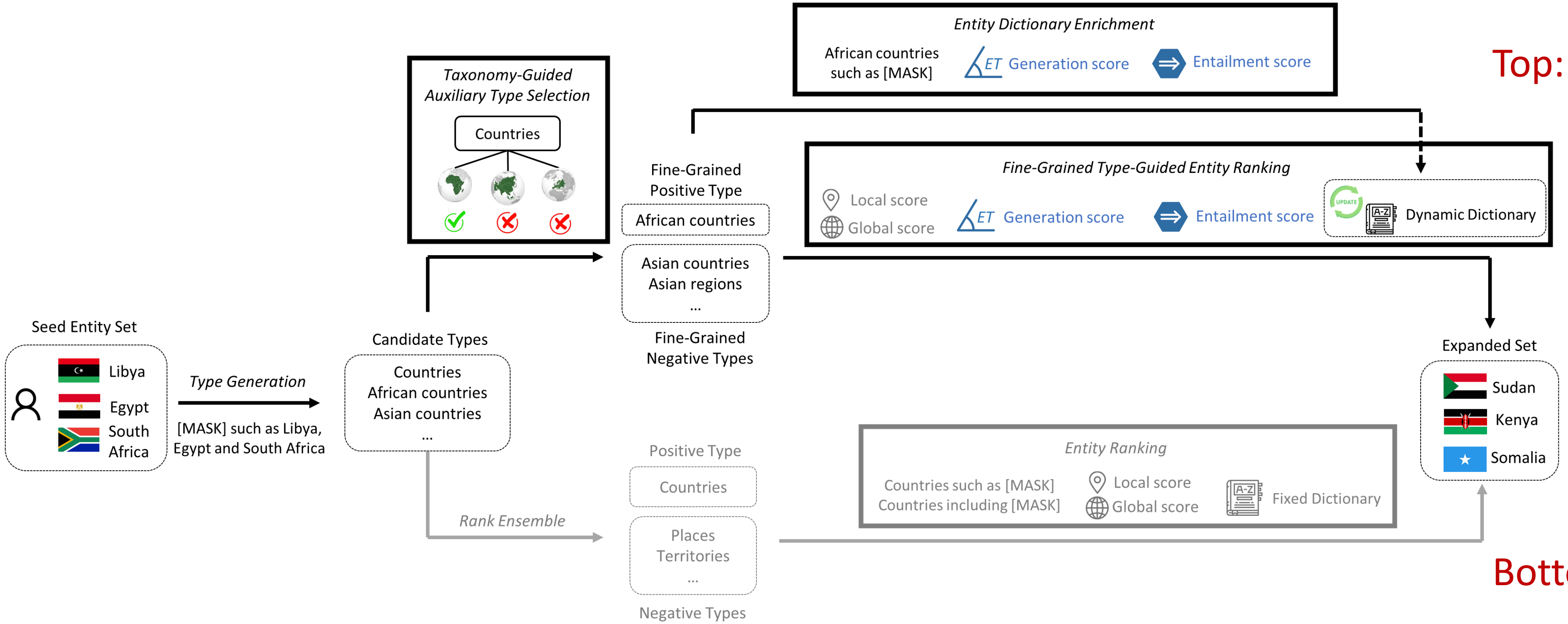
FGExpan: Fine-Grained Type Inference

- Combine three scores to infer the fine-grained type of a seed set
  - Entity generation score: Generate entities for each type and compare to the seed set
  - Type generation score: Generate types for seeds and compare to the taxonomy
  - Entailment score: Test if the types are supported by the corpus context



FGExpan: Taxonomy-Guided Expansion

- *Taxonomy-guided auxiliary type selection*: Use the type taxonomy to sharpen the distinctiveness between positive and negative types
- *Entity dictionary enrichment*: Dynamically add new entities to the vocabulary
- *Fine-grained type-guided entity ranking*: Use generation and entailment scores to tighten the semantic boundary of fine-grained types



Top: FGExpan components

Bottom: CGExpan components

FGExpan: Quantitative Results

Prevents critical failures due to semantic drifts in the inferred type of the entity set

Table 3: Fine-Grained Set Expansion Results

Taxomony Path	Positive Type		AP@10	
	FGExpan	CGExpan	FGExpan	CGExpan
loc → celestial	celestial objects	planets	0.678	0.3
loc → city	cities	cities	1.0	1.0
loc → geo → body of water → river	rivers	places	0.7	0.033
loc → geo → body of water → sea	seas	oceans	1.0	0.767
loc → geo → body of water → lake	lakes	lakes	0.89	0.879
org → Co. → broadcast	broadcasting companies	channels	0.89	0.707
org → Co. → entertainment	entertainment companies	companies	0.737	0.2
org → Co. → mobile phone maker	mobile phone makers	companies	1.0	0.753
loc → country → European	European countries	countries	0.707	0.643
loc → country → Asian	Asian countries	Asian countries	1.0	1.0
loc → country → African	African countries	countries	0.776	0.308
loc → country → Americas	countries in Americas	countries	0.653	0.45
loc → country → Oceanian	Oceanian countries	countries	0.581	0.193
org → education	educational institutes	universities	0.7	0.7
org → government	government agencies	agencies	1.0	1.0
org → military	military units	military forces	0.737	0.538
org → political party	political parties	opposition parties	0.879	0.852
org → sports team	sports teams	baseball teams	0.483	0.3
other → body part	body parts	facial features	0.879	0.879
other → currency	currencies	currencies	0.89	0.89
other → event → holiday	holidays	festivals	0.3	0.25
other → food	foods	foods	1.0	0.879
other → health → malady	diseases	physical symptoms	0.9	0.448
other → language	languages	languages	0.753	0.657
other → living thing → animal	animals	animals	1.0	0.866
other → product → car	cars	small cars	0.4	0.355
other → product → weapon	weapons	weapons	0.762	0.523
person → title	titles	positions	1.0	1.0
overall (MAP@10)			0.796	0.620

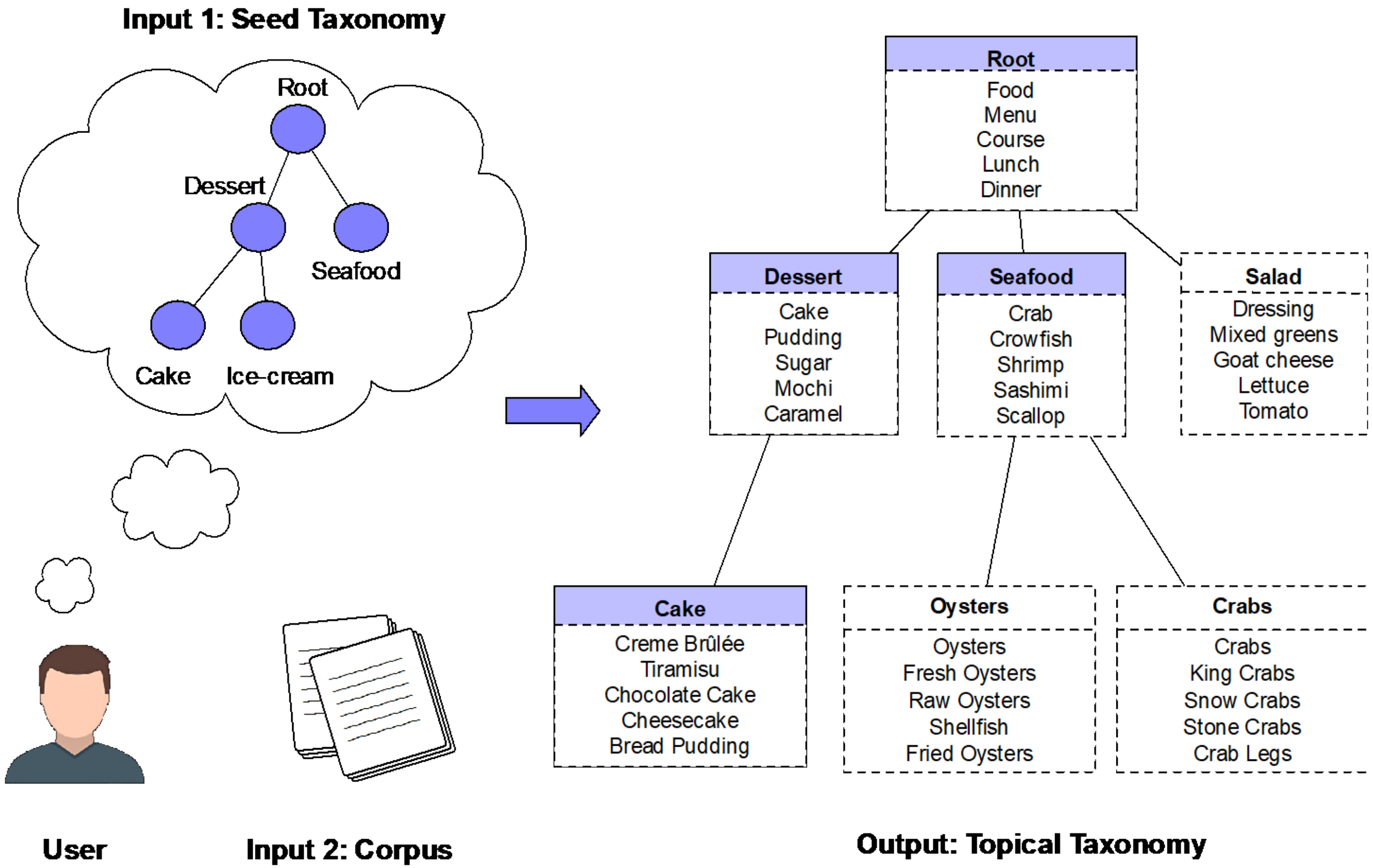
MAP up by 0.176

Outline

- Phrase Mining
- Named Entity Recognition
- Taxonomy Construction
  - Taxonomy Basics and Construction
  - Set Expansion
  - Taxonomy Construction (with Minimal User Guidance)
  - Taxonomy Expansion & Enrichment
- Relation Extraction and Knowledge Graph Construction

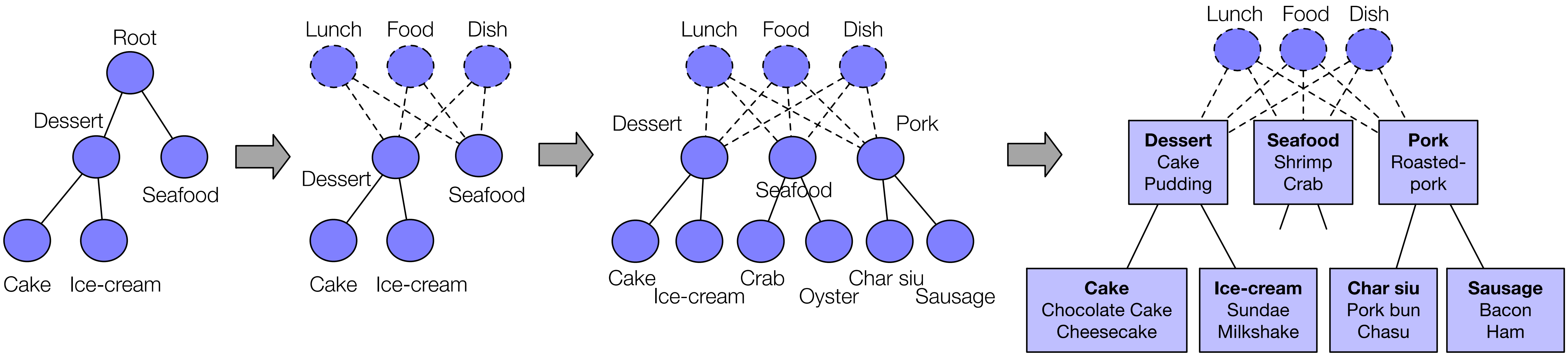
Seed-Guided Topical Taxonomy Construction

- 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

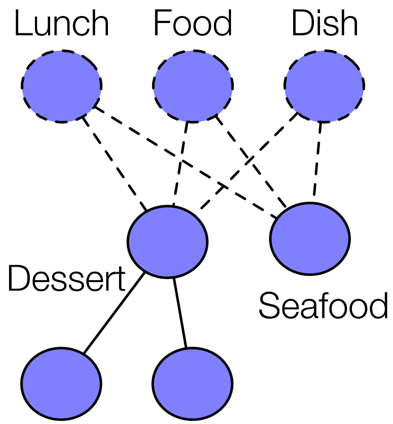
Three Steps:

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 (2020)

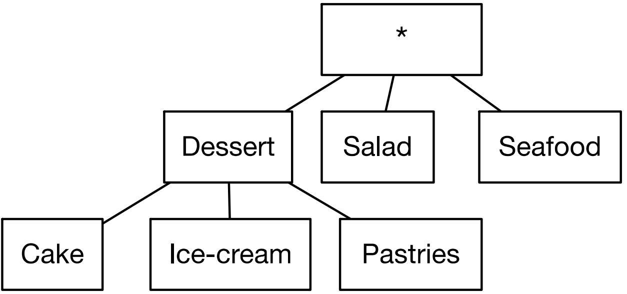
Relation Learning and Transferring

- Learn a relation classifier using pretrained language model (e.g., BERT)
  - Using a weakly-supervised text embedding framework
- Transfer the relation upwards to discover possible root nodes (e.g., “Lunch” and “Food”)
  - The root node would have more general contexts for us to find connections with potential new topics



- 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 the root nodes

Qualitative and Quantitative Results



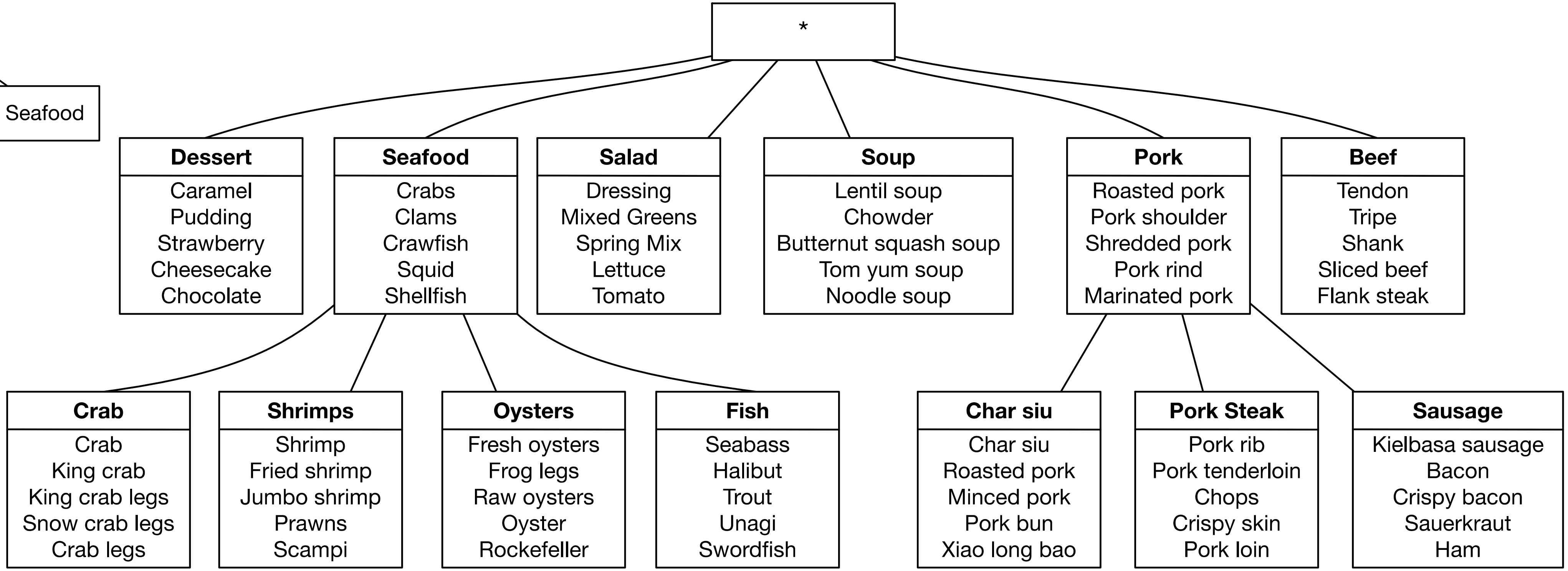


Table 5: Quantitative evaluation on topical taxonomies.

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**

Outline

- Phrase Mining
- Named Entity Recognition
- Taxonomy Construction
  - Taxonomy Basics and Construction
  - Set Expansion
  - Taxonomy Construction (with Minimal User Guidance)
  - Taxonomy Expansion & Enrichment
- Relation Extraction and Knowledge Graph Construction

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

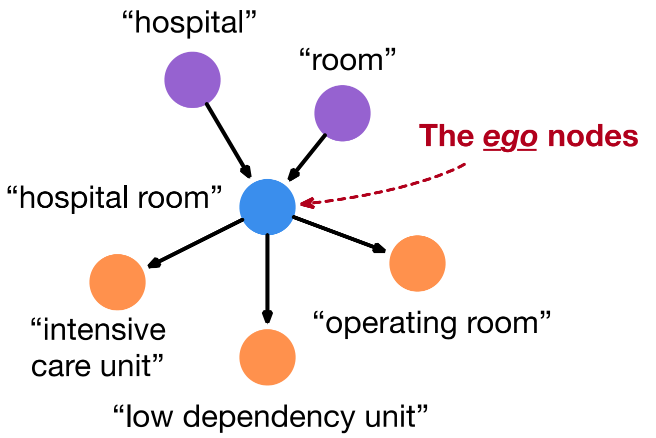
TaxoExpan: Self-supervised Taxonomy Expansion with Position-Enhanced Graph Neural Network [WWW' 20]

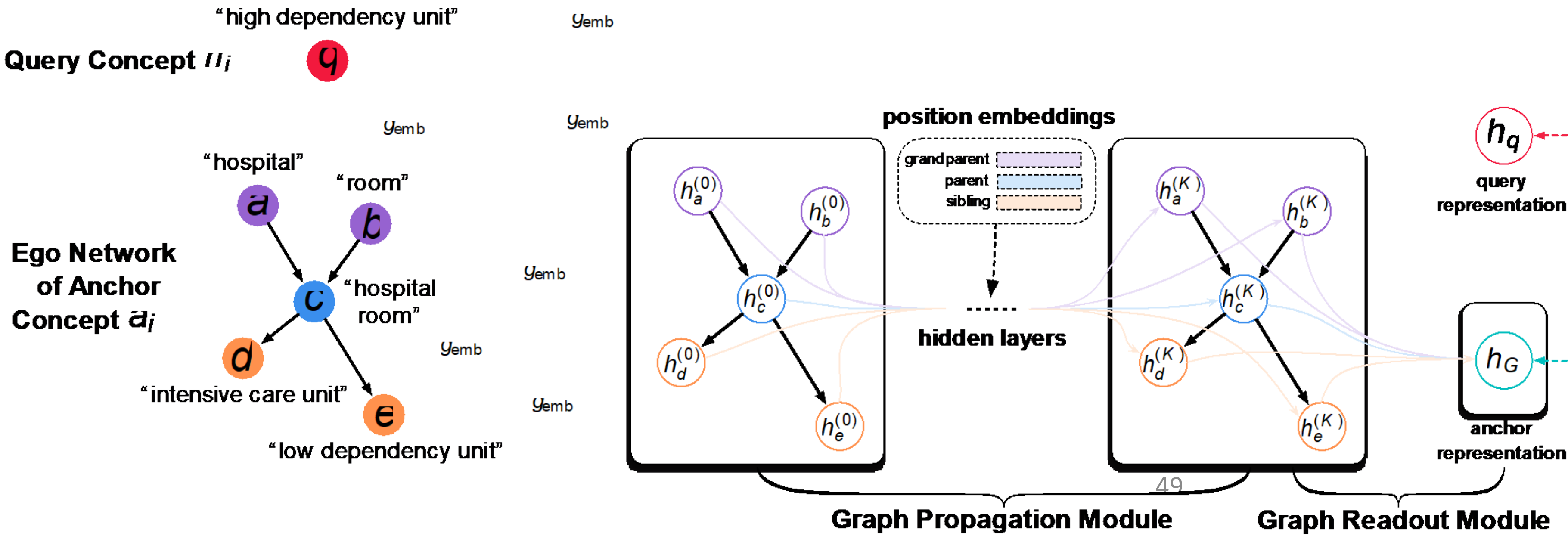
- **Two steps** in solving the problem:
  - Self-supervised term extraction
    - Automatically **extracts emerging terms** from a target domain
  - Self-supervised term attachment
    - A multi-class classification to match a new node to its potential parent
    - Heterogenous sources of information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

Self-supervised Term 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

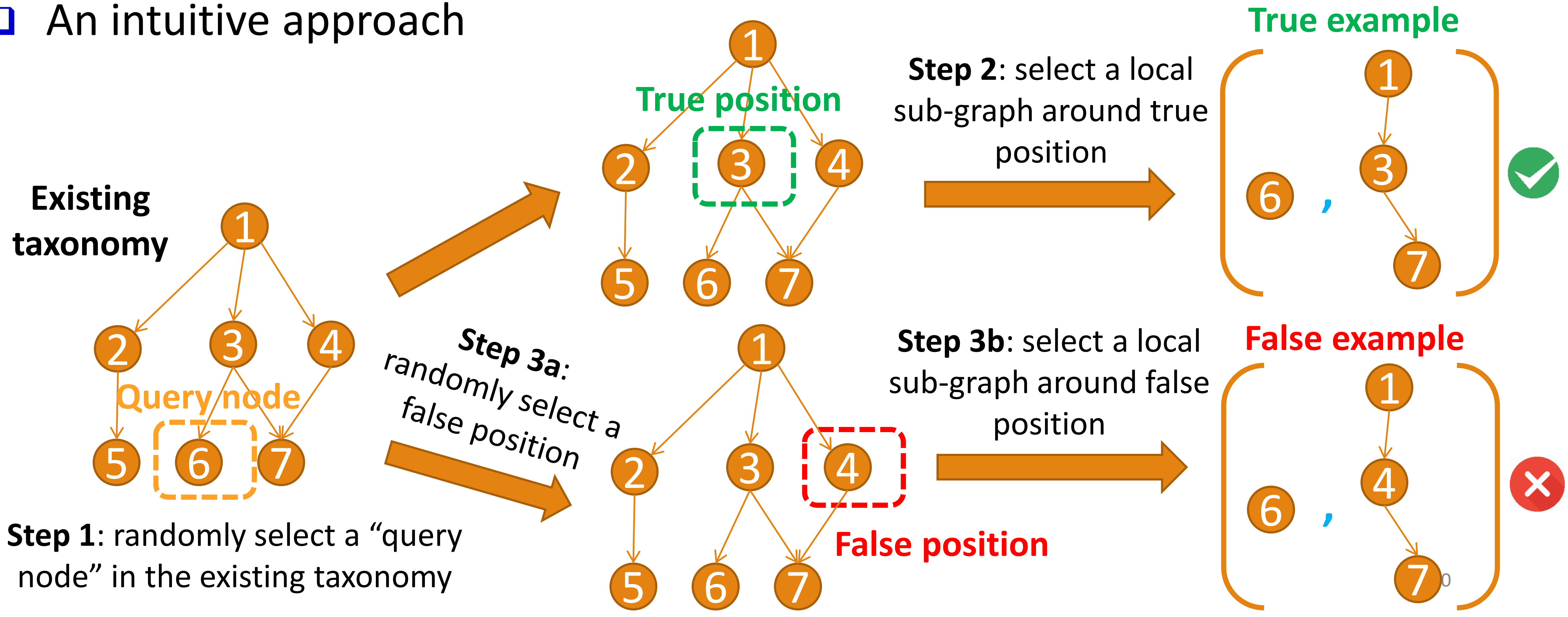
Query: "high dependency unit"





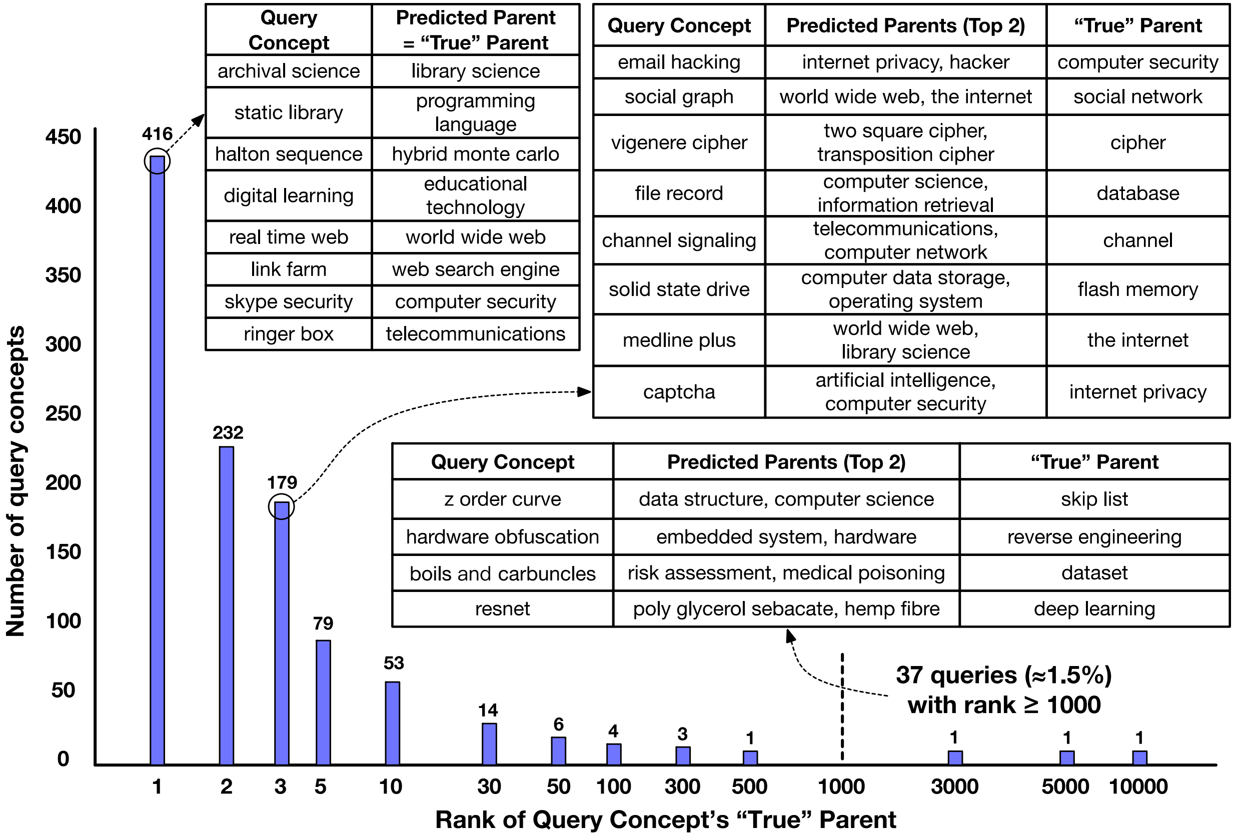
Leveraging Existing Taxonomy for Self-supervised Learning

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



TaxoExpan Framework Analysis

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

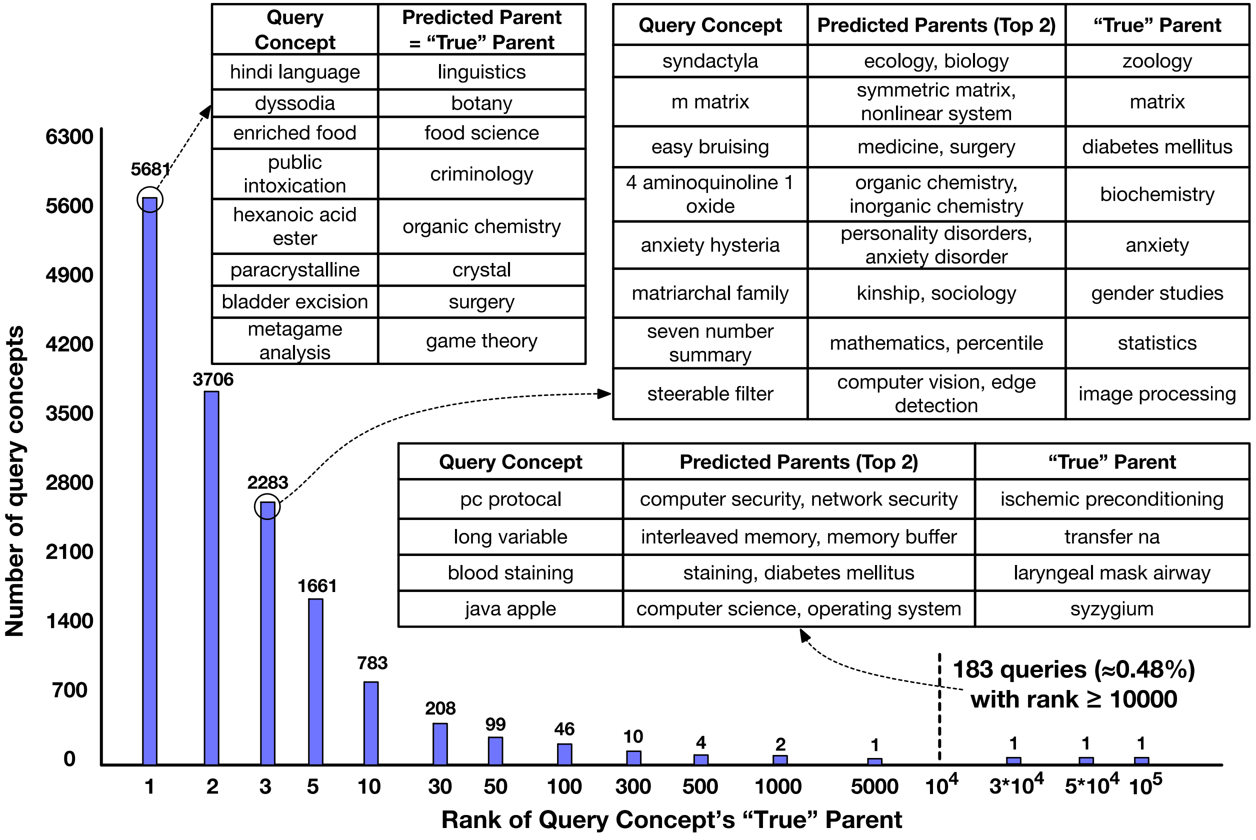


Query Concept	Predicted Parent = "True" Parent
archival science	library science
static library	programming language
halton sequence	hybrid monte carlo
digital learning	educational technology
real time web	world wide web
link farm	web search engine
skype security	computer security
ringer box	telecommunications

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

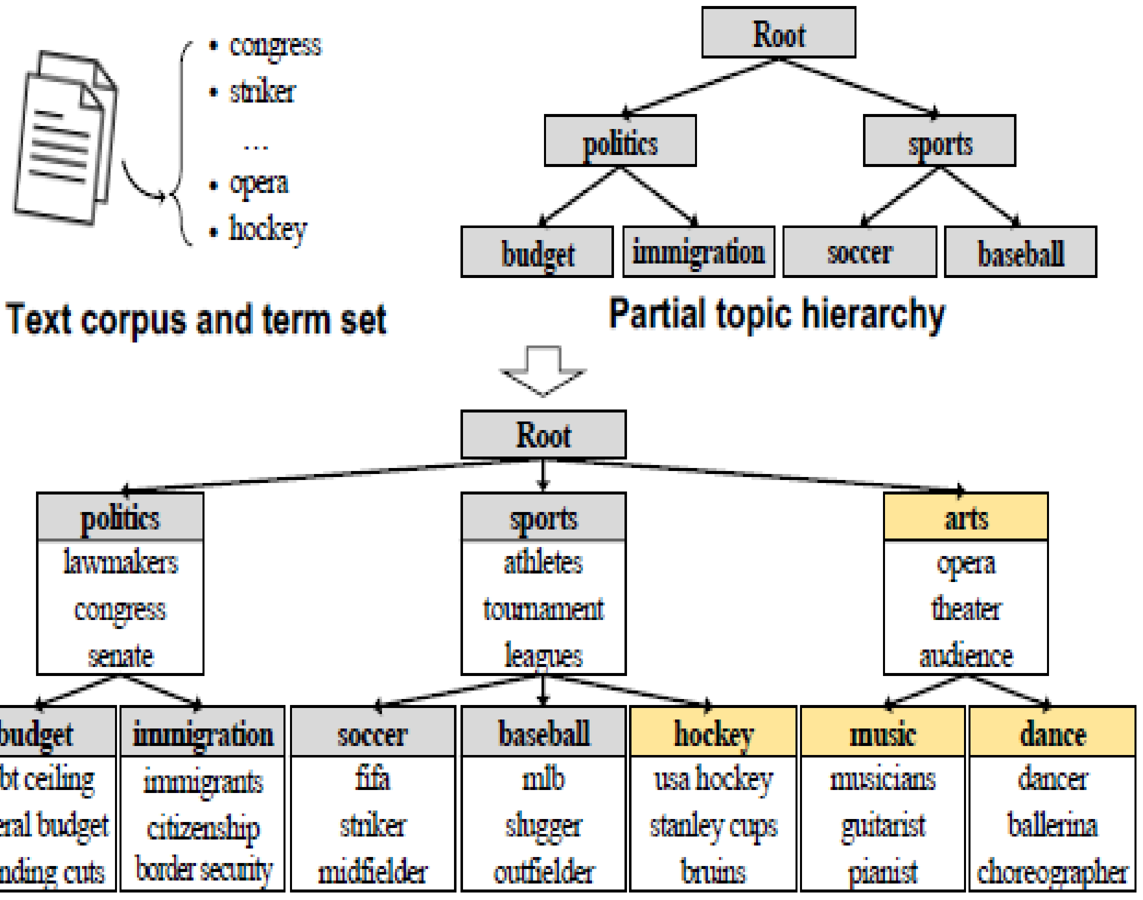
Query Concept	Predicted Parents (Top 2)	"True" Parent
syndactyla	ecology, biology	zoology
m matrix	symmetric matrix, nonlinear system	matrix
easy bruising	medicine, surgery	diabetes mellitus
4 aminoquinoline 1 oxide	organic chemistry, inorganic chemistry	biochemistry
anxiety hysteria	personality disorders, 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)

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

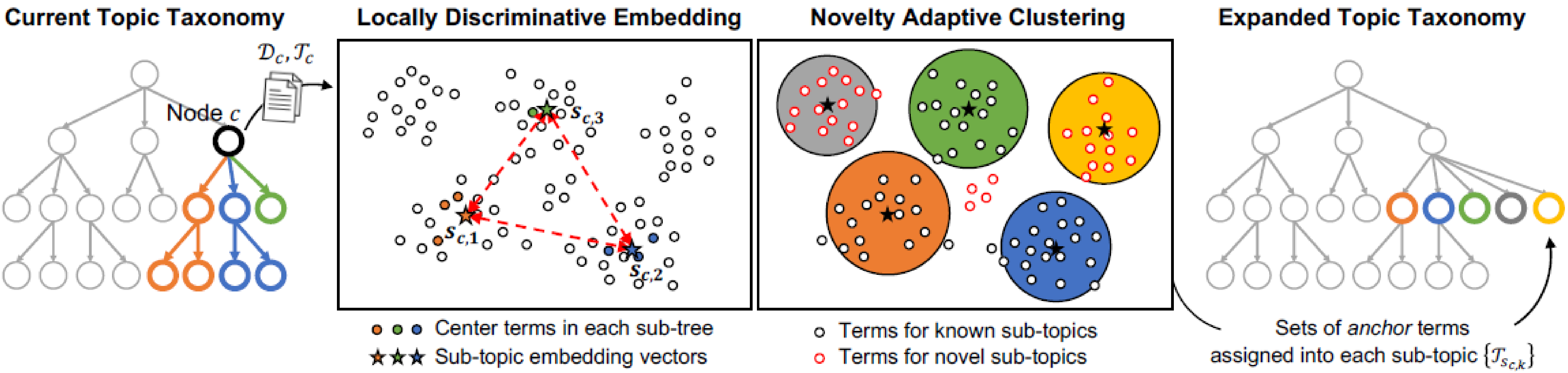
- Topic taxonomy completion: Task ≈ CoRel
- Results: Better quality than Corel
- Method:
  - Recursive expansion of a given topic hierarchy
  - Discovering novel sub-topic clusters of terms and documents



	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 surveillance eavesdropping spying national security agency phone records patriot act	number theory number theory modular form number fields iwasawa theory elliptic curves prime number theorem	accelerator physics 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

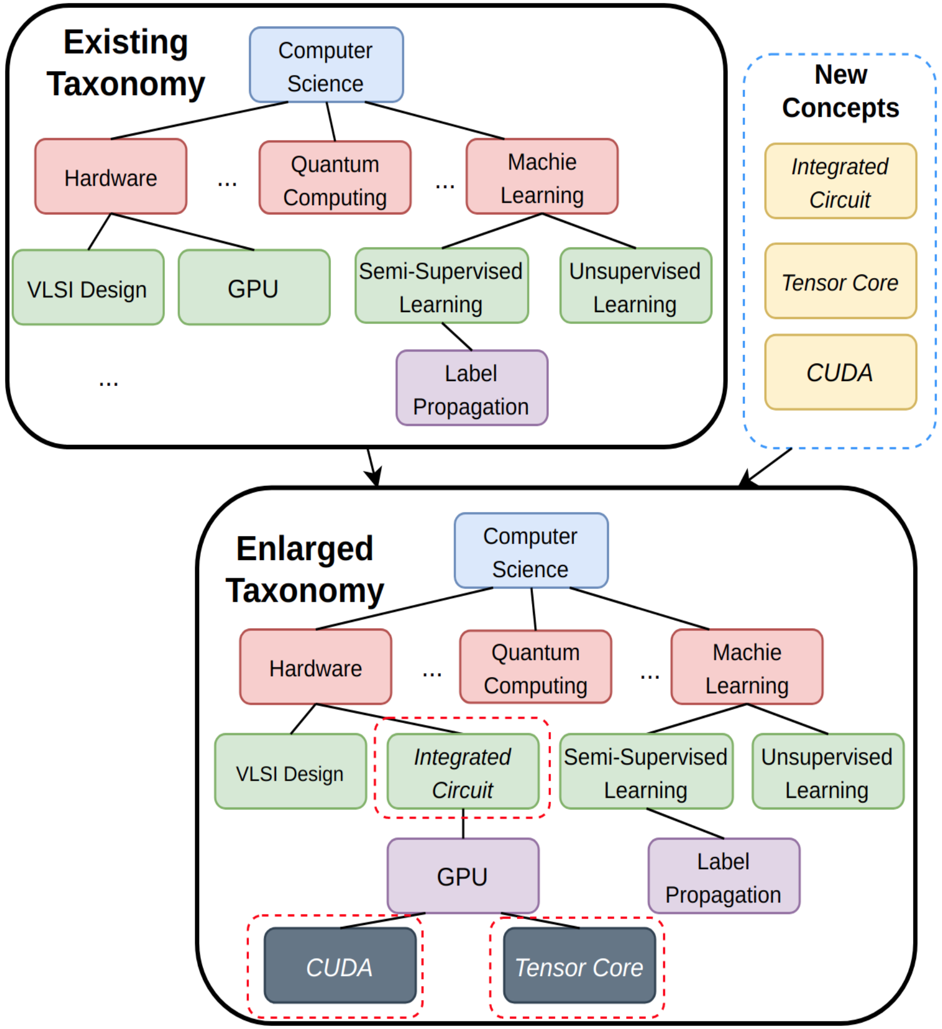
TaxoCom: Hierarchical Discovery of Novel Topic Clusters

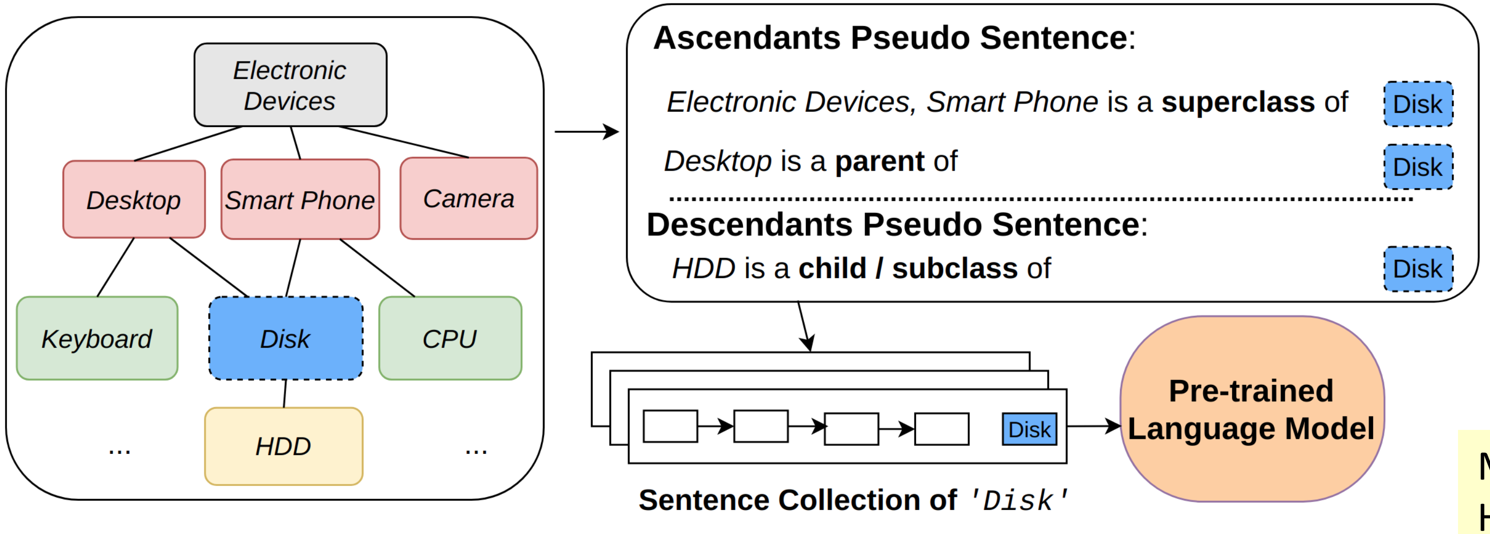


- Starting from the root node, it performs (i) locally discriminative embedding, and (ii) novelty adaptive clustering, to selectively assign the terms (of each node) into one of the child nodes
  - Locally discriminative embedding optimizes the text embedding space to be discriminative among known (i.e.,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

TaxoEnrich: Self-Supervised Taxonomy Completion via Structure-Semantic Representations [WWW'22]

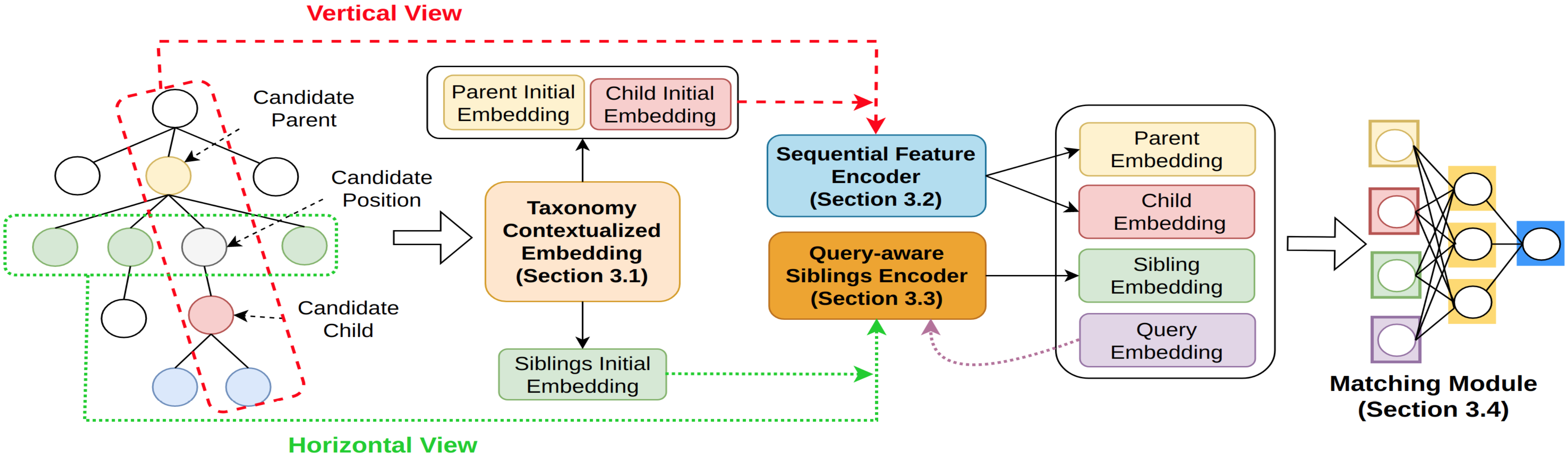
- Task: Inserting new concepts into an existing taxonomy
  - Find the relatedness between the concept and each candidate position
- How to capture extra semantic information?
  - Taxonomy-contextualized embedding
  - Layer-aware representation





Minhao Jiang, Xiangchen Song, Jieyu Zhang and Jiawei Han, "TaxoEnrich: Self-Supervised Taxonomy Completion via Structure-Semantic Representations" (WWW'22)

TaxoEnrich: The General Framework



- Taxonomy-contextualized embedding which incorporates both semantic meanings of concept and taxonomic relations based on powerful pretrained language models
- A taxonomy-aware sequential encoder which learns candidate position representations by encoding the structural information of taxonomy
- A query-aware sibling encoder which adaptively aggregates candidate siblings to augment candidate position representations based on their importance to the query-position matching
- A query-position matching model which extends existing work with new candidate position representations

Outline

- Phrase Mining
- Named Entity Recognition
- Taxonomy Construction
- Relation Extraction and Knowledge Graph Construction
  - Document-Based Relation Extraction
  - Automated Event Type Induction
  - Event Schema Discovery: Role Prediction

Document-Level Relation Extraction

- Document-level relation extraction (DocRE)
  - Extract semantic relations among entity pairs in a document
- Blindly considering the full document?
  - A subset of the sentences in the doc (“evidence”) should often be sufficient to identify the relation
- An evidence-enhanced DocRE framework: EIDER
  - Efficiently extracts evidence and effectively leverages the extracted evidence to improve DocRE
- Using a document-level relationship extraction dataset DocRED (2019)
- Relation extraction benefits natural language understanding in many ways
  - Ex. Knowledge graph construction

Head:Hero of the Day Tail:the United States Rel:[country of origin]
GT evidence sentences: [1,10] Extracted evidence: [1,10]

Original document as input: [1] Load is the sixth studio album by the American heavy metal band Metallica, released on June 4, 1996 by Elektra Records in **the United States** … [9] It was certified 5×platinum … for shipping five million copies in **the United States**. [10] Four singles—"**Hero of the Day**", "Until It Sleeps", "Mama Said", and "King Nothing" — were released as part of the marketing campaign for the album.
Prediction scores: NA: 17.63 **country of origin**: 14.79

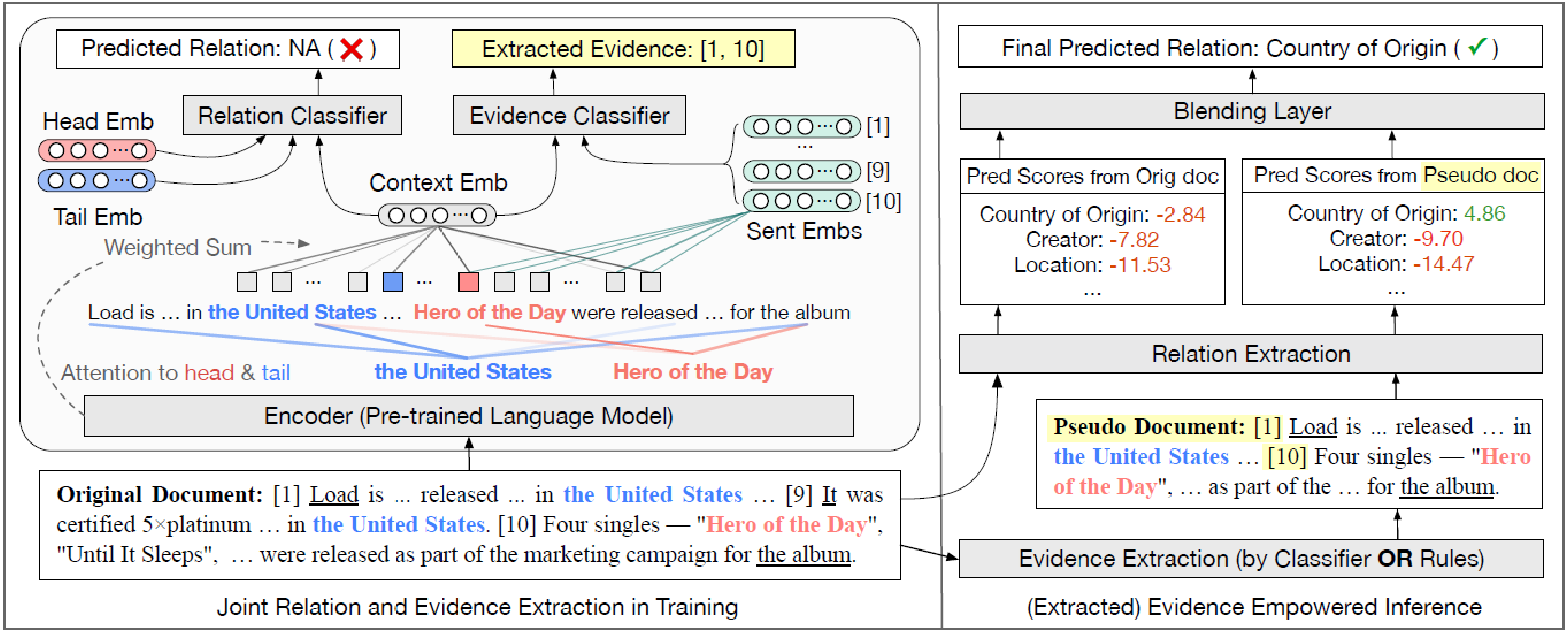
Extracted evidence as input: [1] Load is the sixth studio album … released … in **the United States** … [10] Four singles — "**Hero of the Day**", … were released … for the album.
Prediction scores: **country of origin**: 18.31 NA: 13.45

Final prediction of our model: **country of origin** (✓)

Only need [1]+[10] to identify [head, relation, tail]

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, Jiawei Han, “EIDER: Evidence-enhanced Document-level Relation Extraction”, ACL’22 Findings

EIDER Architecture



The left part (the training stage), we jointly extract relation and evidence using multi-task learning, where the two tasks have their own classifier and share the base encoder

The right part (the inference stage), we fuse the predictions on the original document and the extracted evidence using a blending layer

Outline

- Phrase Mining
- Named Entity Recognition
- Taxonomy Construction
- Relation Extraction and Knowledge Graph Construction
  - Document-Based Relation Extraction
  - Automated Event Type Induction
  - Event Schema Discovery: Role Prediction

New Event Type Representation

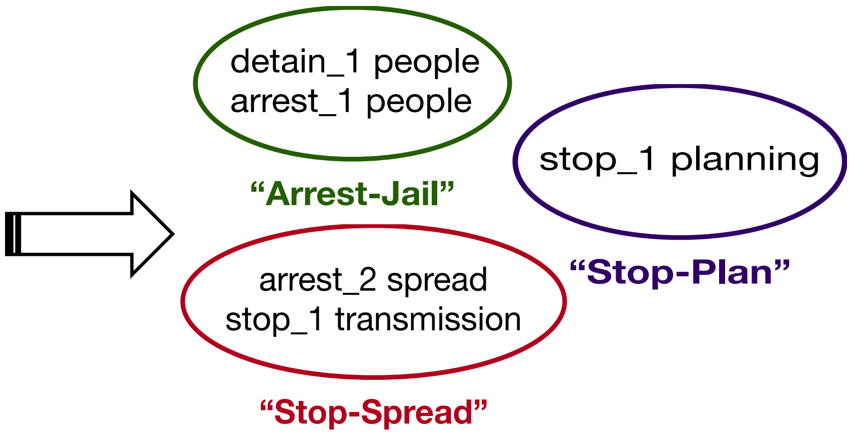
- About 90% of event types can be frequently triggered by a predicate verb
  - "frequently triggered": The event type is triggered by verbs more than five times

Datasets	ACE	ERE	RAMS
# of All Event Types	33	38	138
# of Verb Triggered Event Types	33	38	133
# of Verb Frequently Triggered Event Types	28	36	124

- While predicate verbs could be ambiguous, their word senses combined with object heads can clearly indicate the event types

Represent an event type as a cluster of <predicate sense, object head> (P-O) pairs

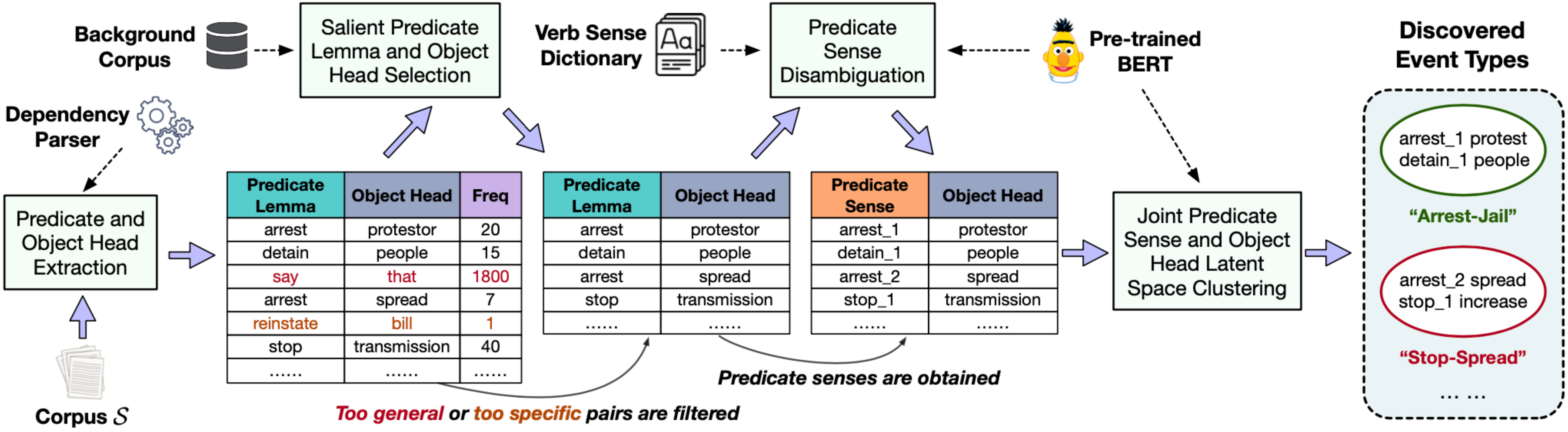
ID	Sentences
S1	Hundreds of *people* are **detained** for distributing purported false information online.
S2	The Zimbabwe CTU said 69 *people* were **arrested** during Wednesday's demonstrations.
S3	Researchers say that vaccinating 46 percent of Haitians could **arrest** the cholera *spread*.
S4	Collective efforts are needed by all nations to **stop** the COVID-19 *transmission*.
S5	More censorship of social media posts are enforced to **stop** protest *planning* online.



ETypeClus: Induce event types by finding those P-O pair clusters [EMNLP'21]

ETypeClus: Automated Event Type Induction

- **Step 1**: Extract predicates and object heads from corpus (Use a dependency parser + a set of linguistic rules)
- **Step 2**: Select salient predicate lemmas and object heads
- **Step 3**: Disambiguate predicate senses
- **Step 4**: Cluster <predicate sense, object head> pairs in a latent spherical space



Predicate Lemma	Object Head	Freq
arrest	protestor	20
detain	people	15
say	that	1800
arrest	spread	7
reinstate	bill	1
stop	transmission	40
……	……	……

Predicate Lemma	Object Head
arrest	protestor
detain	people
arrest	spread
stop	transmission
……	……

Predicate Sense	Object Head
arrest_1	protestor
detain_1	people
arrest_2	spread
stop_1	transmission
……	……

Predicate Sense Disambiguation

- Key idea: compare the usage of a predicate with each verb sense's example sentences in the verb sense dictionary
- How? Use the contextualization power of PLMs:
  - **Continuous representation**: hidden representation of the last layer
  - **Discrete features**: mask the target verb and let PLM predict the most possible replacements

Execute; 3 senses

Sense 1: Put to Death
Example 1: He was executed for murder.
Example 2: He is the first federal prisoner to be executed in 38 years.
Example 3: My dad's cousin was executed by the mafia for collaborating with the police.

Sense 2: Do, Put to Effect
Example 1: We will see the deal executed as planned.
Example 2: The whole play was executed with great precision.
Example 3:I executed a program I had written many times and got valid output.

Sense 3: Sign a legal document before witnesses
Example 1: The president executed the treaty.

Step 3.1a: Obtain BERT embedding

My dad's cousin was *executed* by the mafia for collaborating …

↓

[-0.234, 0.165, 1.564, -0.234, -0.557, 0.413, 0.165, 0.234…]

Step 3.1b: Obtain BERT masked prediction results

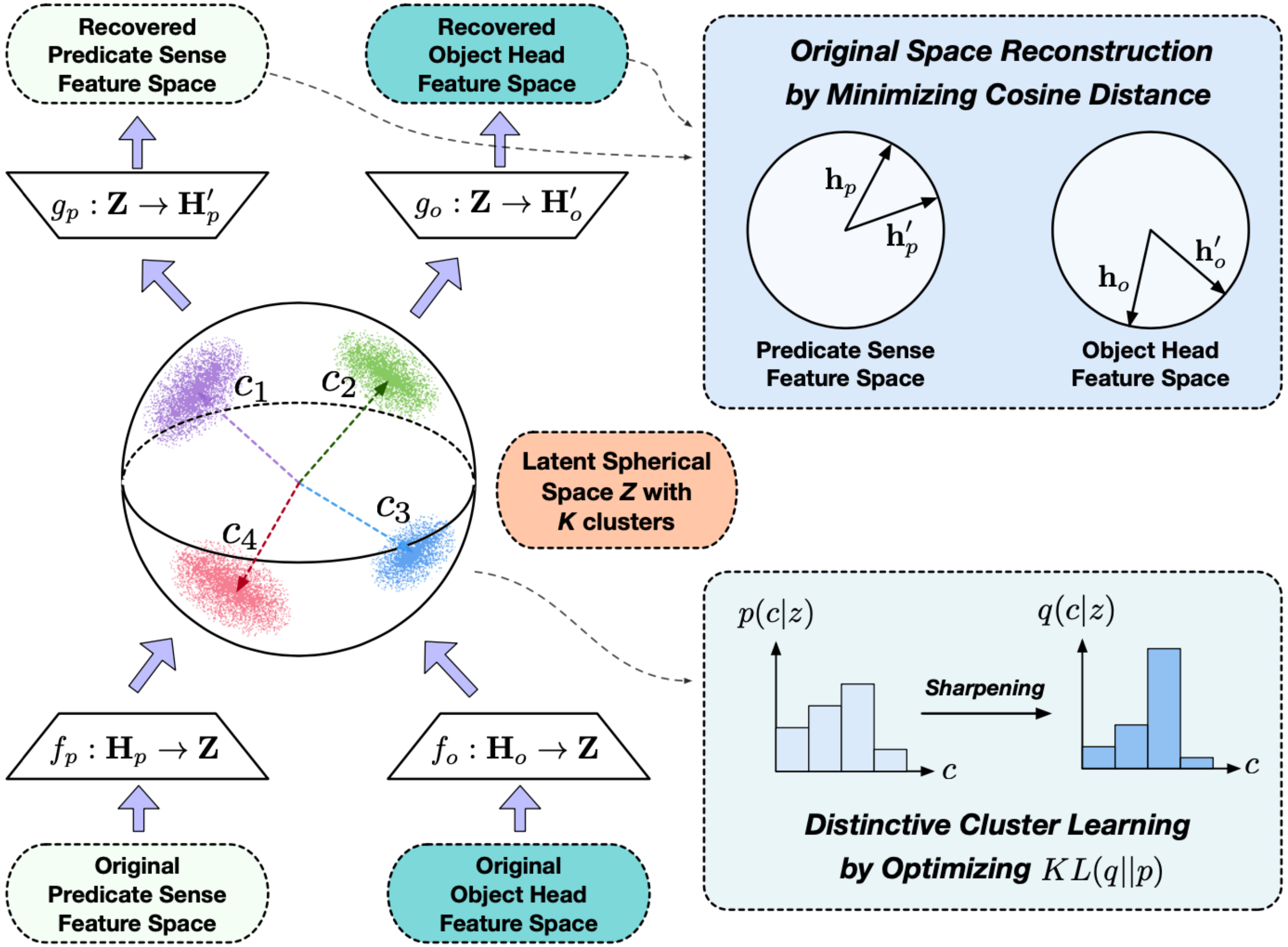
My dad's cousin was [*MASK*] by the mafia for collaborating …

↓

{killed: 0.66, wanted: 0.09, murdered: 0.04, executed: 0.02, …}

Cluster <predicate sense, object head> pairs in a latent spherical space

- Joint Embedding and Clustering
  - We propose to **jointly** embed and cluster P-O pairs in a latent **spherical** space
  - The P-O pair embedding learning is guided by the clustering objective
  - The clustering quality is improved with the good structure of the latent space



Experiments on ACE and ERE Datasets

Recover human-labeled event types

Identify new types and finer-grained types compared with human labeled ones

- Run ETypeClus to generate 100 candidate clusters
  - On ACE dataset, we recover 24 out 33 types (19 out of 20 most frequent types)
  - On ERE dataset, we recover 28 out 38 types (18 out of 20 most frequent types)

Event Type	Top Ranked P-O Pairs	Example Sentences in Corpus
Arrest-Jail	⟨arrest_0, protester⟩ ⟨arrest_0, militant⟩ ⟨arrest_0, suspect⟩	• For the most part the marches went off peacefully, but in New York a small group of *protesters* were **arrested** after they refused to go home at the end of their rally, police sources said. • On Tuesday, Saudi security officials said three suspected al-Qaida *militants* were **arrested** in Jiddah, Saudi Arabia.
Build$^{\triangledown}$	⟨build_0, facility⟩ ⟨build_0, center⟩ ⟨build_0, housing⟩	• Plans were underway to **build** destruction *facilities* at all other locations but now the Bush junta has removed from its proposed defense budget for fiscal year 2006 all but the minimum funding. • Virginia is apparently going to be **build** a data *center* in Richmond, a back-up data center, and a help desk/call center as a follow-on to the creation of VITA, the Virginia Information Technology Agency.
Transfer-Money	⟨fund_0, activity⟩ ⟨fund_0, operation⟩ ⟨fund_0, people⟩	• The grants will **fund** advisory *activities*, including local capacity building, infrastructure development and product development. • The White House had hoped to hold off asking for more money to **fund** military *operations* in Iraq and Afghanistan until after the election, but with costs rising faster than expected, it sent a request for an early installment of $25 billion to Congress this week.
Bombing$^{\triangledown}$	⟨bomb_0, factory⟩ ⟨bomb_0, checkpoint⟩ ⟨bomb_0, base⟩	• He **bombed** the Aspirin *factory* in 1998 (which turned out to have nothing to do with Bin Laden) the week he revealed he had been lying to us for eight months about Lewinsky. • Prosecutors then also pointed to the men's suicide bomber training in 2011 in Somalia and association with Beledi, who prosecutors said **bombed** a government *checkpoint* in Mogadishu that year.

Experiments on Pandemic Dataset

Human Intrusion Test of P-O Pair Cluster Quality

Methods	K-Menas	AggClus	JCSC	ETypeClus
Accuracy	86.7	64.4	54.4	**91.1**

Interesting event types

Examples sentences for identified event types

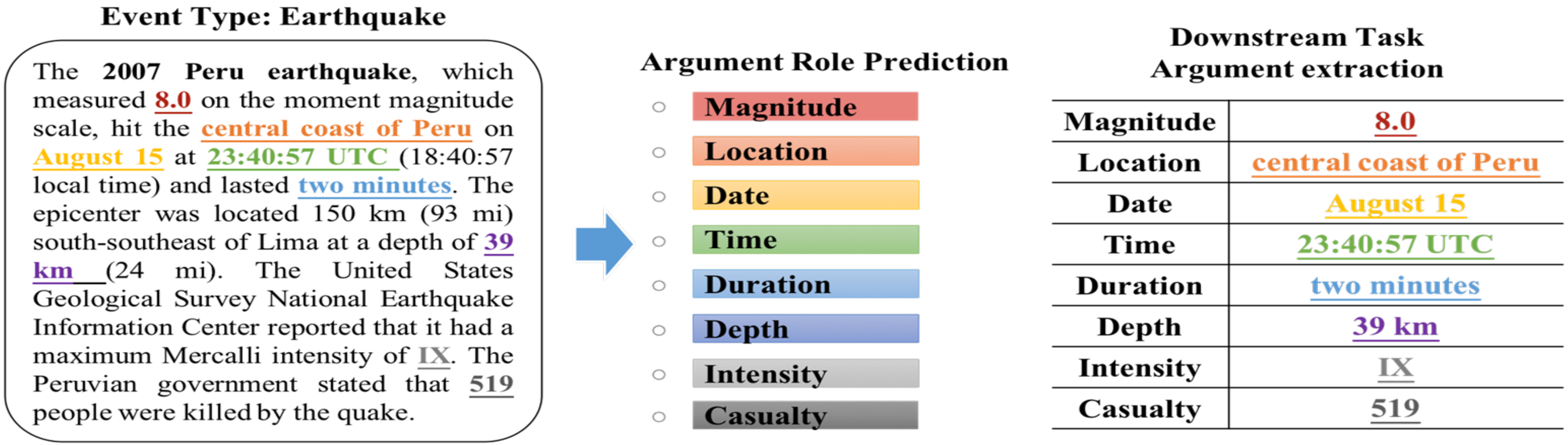
Event Type	Top Ranked P-O Pairs	Example Sentences in Corpus
Spread Virus	⟨spread_2, virus⟩ ⟨spread_2, disease⟩ ⟨spread_2, coronavirus⟩	• What is the best way to keep from **spreading** the *virus* through coughing or sneezing? • Farmers quickly mobilized to fight the misperceptions that pigs could **spread** the *disease*. • In the UK, Asians have been punched in the face, accused of **spreading** *coronavirus*.
Prevent Spread	⟨prevent_1, spread⟩ ⟨mitigate_1, spread⟩ ⟨mitigate_1, transmission⟩	• Infection prevention and control measures are critical to **prevent** the possible *spread* of MERS-CoV. • A vaccine can **mitigate** *spread*, but not fully prevent the virus circulating. • Asymptomatic infection could also potentially be directly harnessed to **mitigate** *transmission*.
Vaccinate People	⟨vaccinate_0, person⟩ ⟨immunize_0, people⟩ ⟨vaccinate_0, family⟩	• All *persons* in a recommended vaccination target group should be **vaccinated** with the 2009 H1N1 monovalent vaccine and the seasonal influenza vaccine. • U.K. Will Start **Immunizing** *People* Against COVID-19 On Tuesday, Officials Say. • "..." says Henrietta Aviga, a nurse travelling around villages to **vaccinate** and educate *families*.

Outline

- Phrase Mining
- Named Entity Recognition
- Taxonomy Construction
- Relation Extraction and Knowledge Graph Construction
  - Document-Based Relation Extraction
  - Automated Event Type Induction
  - Event Schema Discovery: Role Prediction

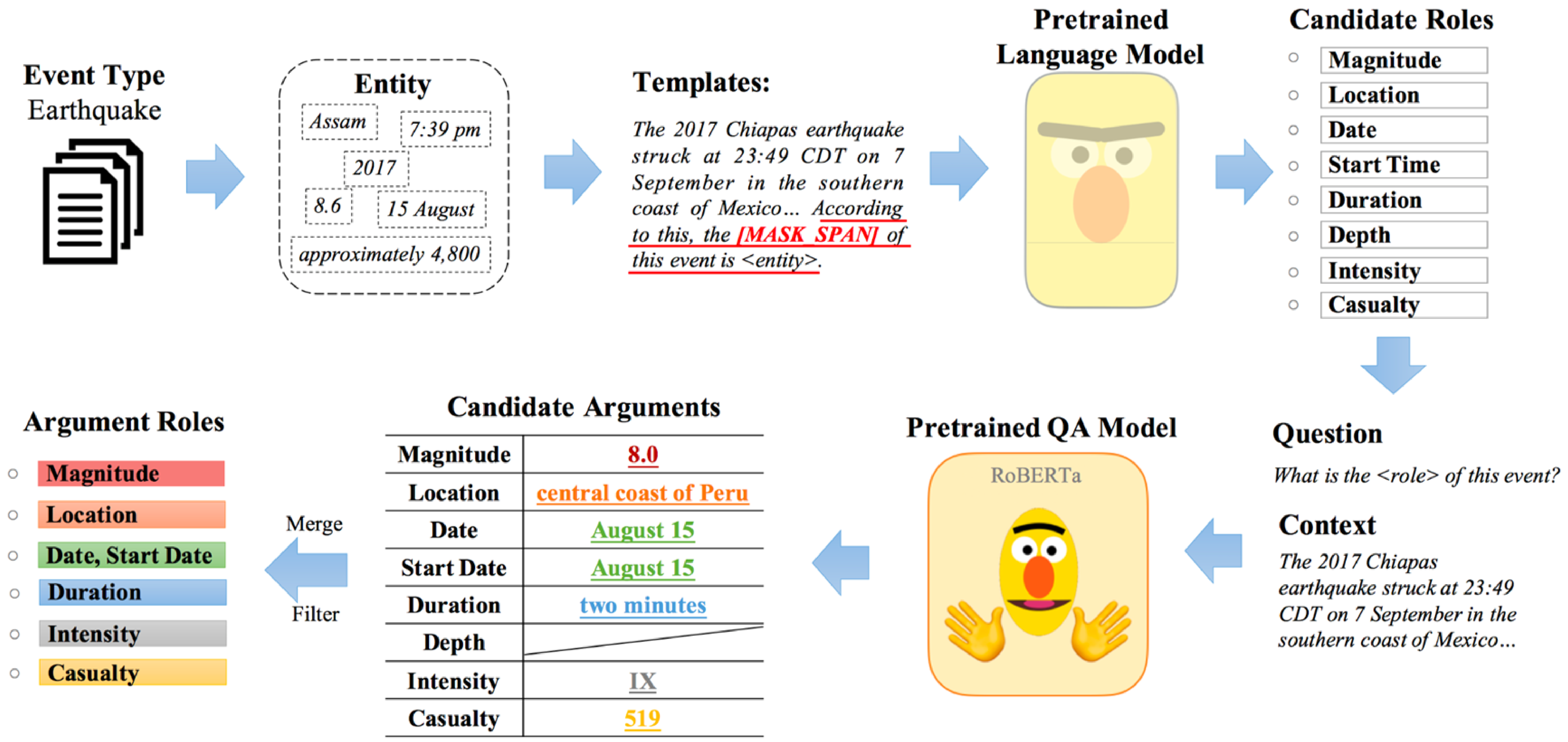
Open-Vocabulary Argument Role Prediction

- Related Work:
  - Most of existing studies rely on hand-crafted ontologies (costly, cannot generalize)
  - A few studies try to automatically induce argument roles (limited pre-defined glossary)
- **New Task**: Infer a set of argument role names for a given event type to describe the crucial relations between the event type and its arguments



Yizhu Jiao, Sha Li, Yiqing Xie, Ming Zhong, Heng Ji and Jiawei Han "Open-Vocabulary Argument Role Prediction for Event Extraction", EMNLP'22

Framework for RolePred (Argument Role Prediction)



RolePred 1: Candidate Role Generation

- Predict candidate role names for named entities by casting it as a prompt-based in-filling task
  - Prompt Construction: (using Generation Model : T5)
    - *Context*. According to this, the ⟨MASK SPAN⟩ of this Event Type is Entity.
  - Ex. *The 1964 Alaskan earthquake, also known as the Great Alaskan earthquake, occurred at 5:36 PM AKST on Good Friday, March 27.* According to this, the ⟨MASK SPAN⟩ of this earthquake is 5:36 PM.
    - ⟨MASK SPAN⟩ is expected to be filled with *time* (or *start time*) as the argument role
  - Considering the entity's general semantic type: person, location, number, etc., we slightly alter the prompt to fluently and naturally support the unmasking argument roles

Prompt design for different entities

Entity Type	Prompt
PERSON	*According to this, Entity play the role of ⟨MASK SPAN⟩ in this Event Type.*
LOCATION	*According to this, the ⟨MASK SPAN⟩ is Entity in this Event Type.*
NUMBER	*According to this, the number of ⟨MASK SPAN⟩ of this Event Type is Entity.*
OTHER TYPES	*According to this, the ⟨MASK SPAN⟩ of this Event Type is Entity.*

RolePred 2: Candidate Argument Extraction

- Formulate the argument extraction problem into question-answering task
  - Input: follow a standard BERT-style format (Model: BERT based pretrained QA model)
    - [CLS] What is the Event Role in this Event Type event? [SEP] Document [SEP]
  - Ex. [CLS] What is the *casualty* in this *pandemic* event? [SEP] *The COVID-19 pandemic is an ongoing global pandemic of coronavirus disease. It's estimated that the worldwide total number of deaths has exceeded five million ...* [SEP]
    - The argument is expected to be five million
    - Note that, for some roles, a given document may not mention its argument. That is, the above-constructed question can be unanswerable. Thus, for each extracted answer, we set a threshold on its probability from the QA model to filter out some unreliable results.
  - Benefit
    - Widely adaptable to any argument role or event type
    - Judge if some arguments exist
    - Search for arguments in a document (not within a sentence)

RolePred 3: Argument Role Selection

- Role Filtering
  - Judge the salience of an argument role by involving multiple event instances of the same type
    - Ex. *intensity* of the *earthquake* events; *host* for the *award ceremony* events
  - A role name belongs to the event type only if most of the event instances have their associated argument
- Role Merging
  - Different roles can represent similar semantics and share the same arguments in an event
    - Ex. The *date*, *official date*, and *original date* may refer to the same day for a firework event
  - The semantic similarity of two roles is determined by the frequency that they share the same argument in the event instances
    - Ex. Given 10 instances of the firework event, if two roles, *date*, and *official date*, have the same day as their arguments in 5 instances, their similarity is 0.5

Experiment: Argument Role Prediction

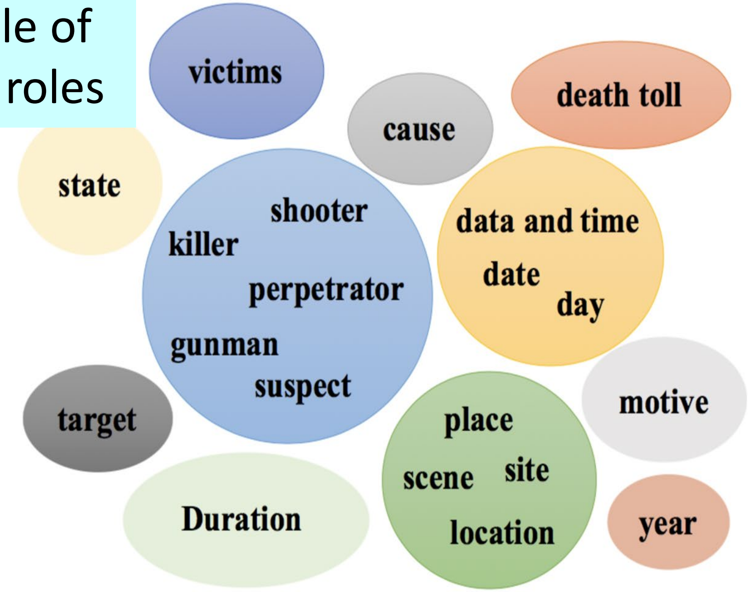
Argument Role Prediction

Models	Hard Matching Precision	Hard Matching Recall	Hard Matching F1	Soft Matching Precision	Soft Matching Recall	Soft Matching F1
LiberalEE	0.1342	0.2613	0.1773	0.3474	0.5340	0.4209
VASE	0.0926	0.1436	0.1125	0.2581	0.4274	0.3218
ODEE	0.1241	0.3076	0.1768	0.3204	0.4862	0.3862
CLEVE	0.1363	0.2716	0.1815	0.3599	0.5712	0.4415
ROLEPRED (BERT)	0.2128	0.4582	0.2906	0.4188	0.6896	0.5211
ROLEPRED (T5)	**0.2552**	**0.6461**	**0.3659**	**0.4591**	**0.7079**	**0.5570**
- RoleMerge	0.2233	0.6962	0.3381	0.4234	0.7677	0.5457
- RoleMerge - RoleFilter	0.1928	0.6582	0.2983	0.4188	0.7084	0.5264
Human	0.6098	0.8270	0.7020	0.7365	0.8732	0.7990

Argument Extraction w/o Golden Roles

Models	P	R	F1
LiberalEE	0.2009	0.2941	0.2387
VASE	0.2123	0.3257	0.2570
ODEE	0.2402	0.3712	0.2917
CLEVE	0.3529	0.3890	0.3701
ROLEPRED (BERT)	0.4170	0.4333	0.4250
ROLEPRED (Roberta)	**0.4131**	**0.5774**	**0.4817**
- RoleMerge	0.3855	0.6187	0.4750
- RoleMerge - RoleFilter	0.4397	0.5001	0.4679
ROLEPRED (Gold Roles)	0.6664	0.4948	0.5679

An example of generated roles



Extracted events by RolePred and baselines

Output of RolePred

Victims	Maura Binkley and Nancy Van Vessem
State	Florida
Date	November 2, 2018
Killer	Scott Paul Beierle
Place	The yoga studio
Time	5:37 p.m. EDT
Duration	three and a half minutes
Motive	hatred of women
Target	Tallahassee Hot Yoga, a yoga studio
Year	2018

Output of ODEE

Agent	The gunman
Patient	six women

Output of CLEVE

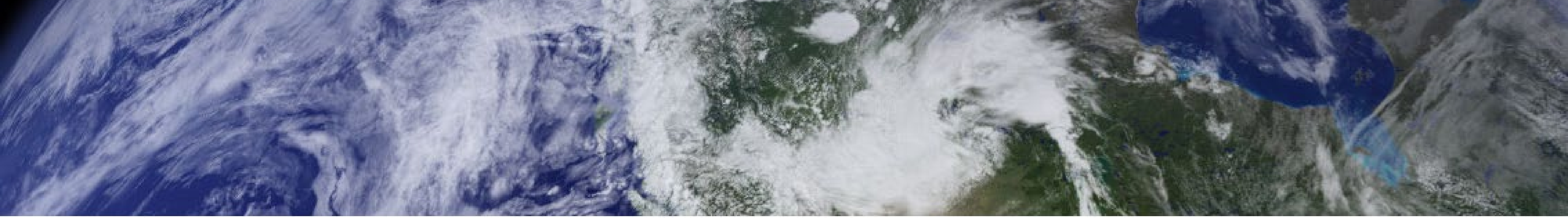
Agent	Scott Paul Beierle
Patient	six women
Time	2018

References I

- Xiaotao Gu , Zihan Wang , Zhenyu Bi , Yu Meng, Liyuan Liu, Jiawei Han, Jingbo Shang. "UCPhrase: Unsupervised Context-aware Quality Phrase Tagging" (KDD'21)
- Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. "Few-Shot Named Entity Recognition: An Empirical Baseline Study" (EMNLP'21)
- Jiaxin Huang, Yu Meng, and Jiawei Han. "Few-Shot Fine-Grained Entity Typing with Automatic Label Interpretation and Instance Generation" (KDD'22)
- 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)
- Minhao Jiang, Xiangchen Song, Jieyu Zhang and Jiawei Han, "TaxoEnrich: Self-Supervised Taxonomy Completion via Structure-Semantic Representations" (WWW'22)
- Yizhu Jiao, Sha Li, Yiqing Xie, Ming Zhong, Heng Ji, and Jiawei Han. "Open-Vocabulary Argument Role Prediction for Event Extraction" (EMNLP'22)
- Dongha Lee, Jiaming Shen, SeongKu Kang, Susik Yoon, Jiawei Han, and Hwanjo Yu. "TaxoCom: Topic Taxonomy Completion with Hierarchical Discovery of Novel Topic Clusters" (WWW'22)

References II

- Yu Meng, Yunyi Zhang, Jiaxin Huang, Xuan Wang, Yu Zhang, Heng Ji, and Jiawei Han. "Distantly-Supervised Named Entity Recognition with Noise-Robust Learning and Language Model Augmented Self-Training" (EMNLP'21)
- 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)
- 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)
- Jiaming Shen, Yunyi Zhang, Heng Ji, and Jiawei Han. "Corpus-based Open-Domain Event Type Induction" (EMNLP'21)
- Jinfeng Xiao, Mohab Elkaref, Nathan Herr, Geeth De Mel, and Jiawei Han. "Taxonomy-Guided Fine-Grained Entity Set Expansion" (SDM'23)
- Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. "EIDER: Evidence-enhanced Document-level Relation Extraction" (ACL'22)
- Yunyi Zhang, Jiaming Shen, Jingbo Shang, and Jiawei Han. "Empower Entity Set Expansion via Language Model Probing" (ACL'20)



Q&A