Part II: Revisiting Text Mining Fundamentals with Pretrained Language Models

KDD 2022 Tutorial

Adapting Pretrained Representations for Text Mining

Yu Meng, Jiaxin Huang, Yu Zhang, Jiawei Han

Computer Science, University of Illinois at Urbana-Champaign

Aug 14, 2022

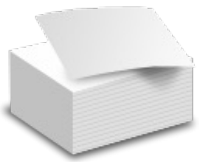
Outline

- Phrase Mining
  - Phrase Mining Introduction
  - UCPhrase: Unsupervised Context-aware Quality Phrase Tagging
- Constituency Parsing
- Named Entity Recognition
- Taxonomy Construction

Previous Phrase Mining/Chunking Models

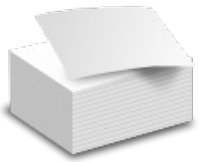
- Identifying and understanding quality phrases from context is a fundamental task in text mining.

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

Outline

- Phrase Mining
    - Phrase Mining Introduction
    - UCPhrase: Unsupervised Context-aware Quality Phrase Tagging [KDD’21]
- Constituency Parsing
- Named Entity Recognition
- Taxonomy Construction

Previous Phrase Mining/Chunking Models

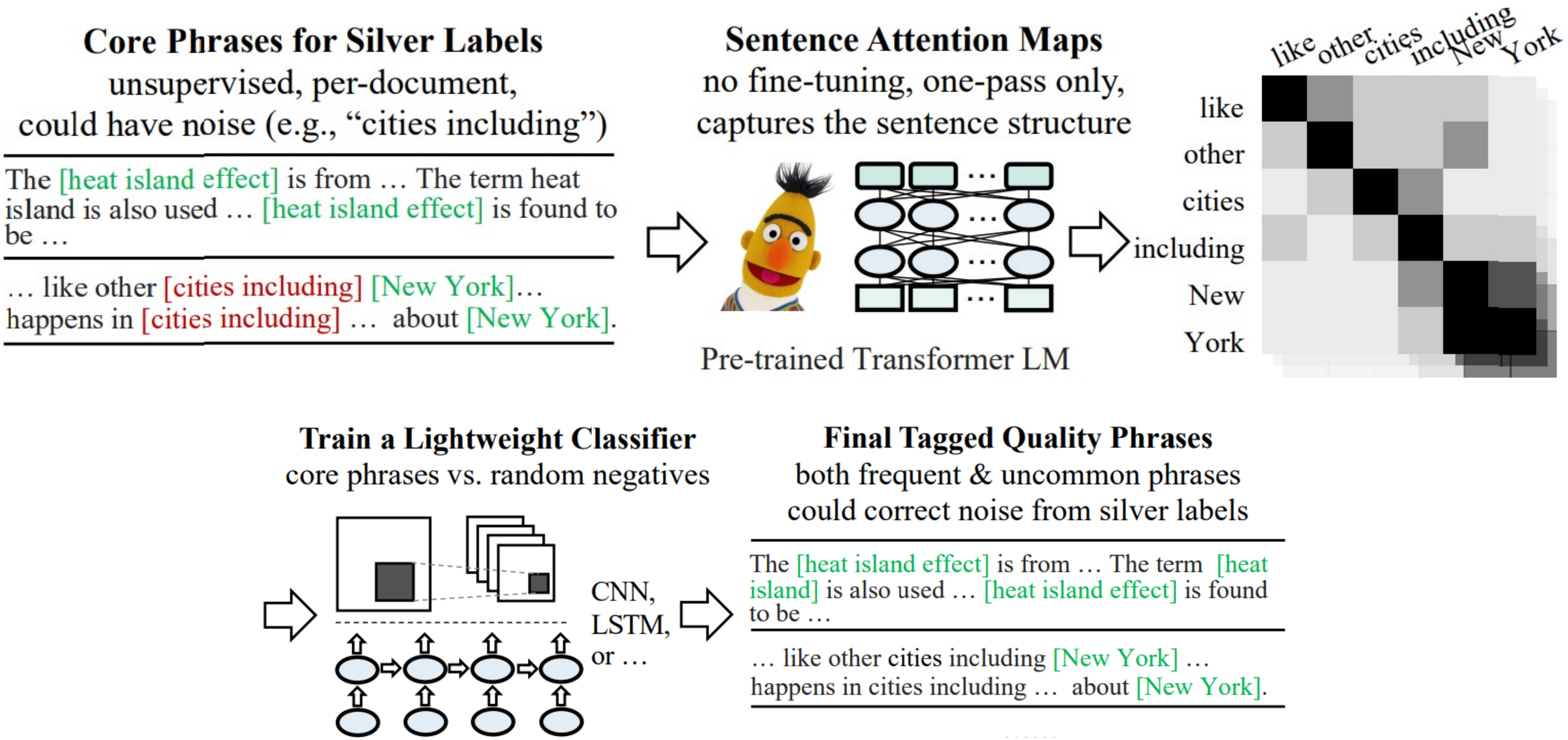
- Statistics-based models (*TopMine, SegPhrase, AutoPhrase)*
  - only work for frequent phrases, ignore valuable **infrequent / emerging phrases**
- Tagging-based models (*Spacy, StanfordNLP*)
  - do not have requirements for frequency
  - require **expensive and unscalable** sentence-level annotations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

Different Types of Supervisions

- Supervision
  - Human annotation
    - expensive, **hard to scale** to larger corpora and new domains
  - Distant supervision
    - tend to produce **incomplete labels** due to context-agnostic matching
      - e.g. "Heat [island effect] is found to be ..."
      - e.g. "Biomedical [data mining] is an important task where ..."
    - tend to match popular phrases, which form a small seen phrase vocabulary
      - easy for an embedding-based system to **memorize / overfit**

Framework of UCPhrase

- Silver Label Generation + Attention Map-based Span Prediction



Silver Label Generation

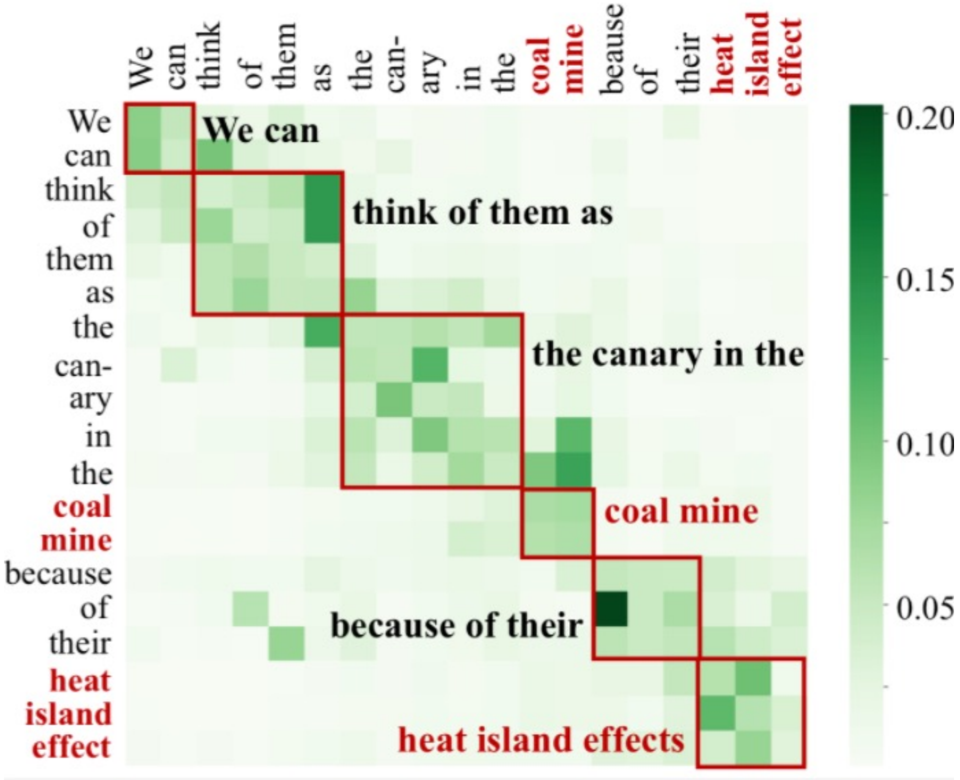
- How do human readers accumulate new phrases?
  - even without any prior knowledge we can recognize these consistently used patterns from a document
  - e.g., *task name, method name, dataset name, concepts* in a publication
  - e.g., *human name, organization, locations* in a news article
- Mining 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

Surface-Agnostic Feature Generation

- What's wrong with traditional embedding-based features?
  - embedding features are word identifiable -- it tells you which word you are looking at
  - easy to rigidly memorize all seen phrases / words in the training set / dictionary
  - fail to generalize to unseen phrases
- Good features for phrase recognition should be
  - agnostic to word **surface names** (so the model cannot rely on rigid memorization)
  - reveal the role that the span plays in the entire sentence (look at **sentence structure** rather than phrase names)

Attention Map

- 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

- Viewing the generated feature as a 144-channel image of size K*K
  - train a lightweight 2-layer CNN model for binary classification: is a phrase or not
  - why CNN: capture word interactions (attentions) from various ranges, also fast for training and inference
- Efficient implementation
  - only train the CNN module, without fine-tuning LM

Quantitative Evaluation

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

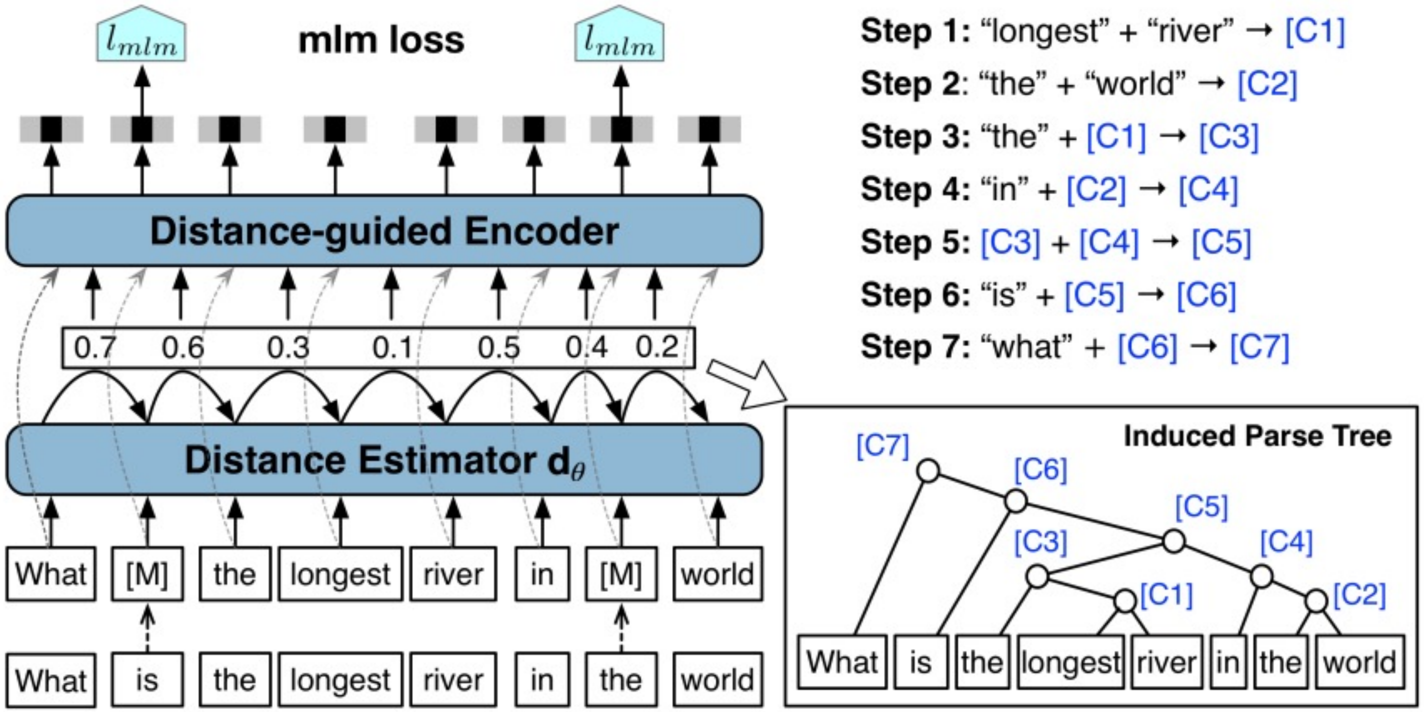
Method Type	Method Name	Task I: Phrase Ranking				Task II: KP Extract.				Task III: Phrase Tagging					
		KP20k		KPTimes		KP20K		KPTimes		KP20k			KPTimes		
		P@5K	P@50K	P@5K	P@50K	Rec.	F_1@10	Rec.	F_1@10	Prec.	Rec.	F_1	Prec.	Rec.	F_1
Pre-trained	PKE [3]	–	–	–	–	57.1	12.6	61.9	4.4	54.1	63.9	58.6	56.1	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
- Constituency Parsing
  - Phrase-aware Unsupervised Constituency Parsing [ACL'2022]
- Named Entity Recognition
- Taxonomy Construction

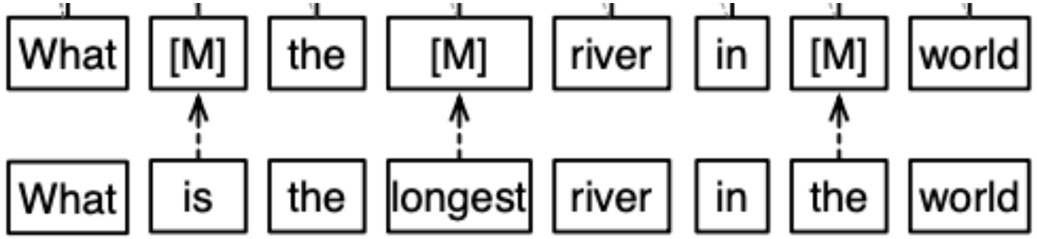
LM-based Unsupervised Constituency Parsing

- Represent discrete parsing tree as a distance sequence (given by a distance estimator)
- Distance information helps inject the parsing tree structure into encoder training via the MLM loss



Challenges With Current LM-Based Methods

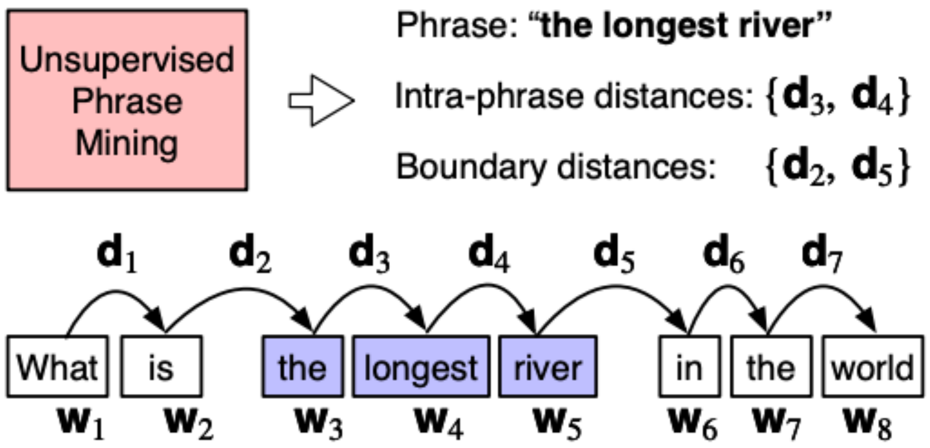
- The distance estimator is randomly initialized
  - yield suboptimal information for the encoder **in the cold start phase**
  - lead to suboptimal parsing accuracy due to **error accumulation**
- The token reconstruction task (MLM) mainly relies on the aggregation of **local information,** thus can hardly guide the model to manage **high-level structures across long distances**
  - Example: The prediction of “longest” mainly depends on its neighbor “river”



Phrase-Regularized Warm-Up

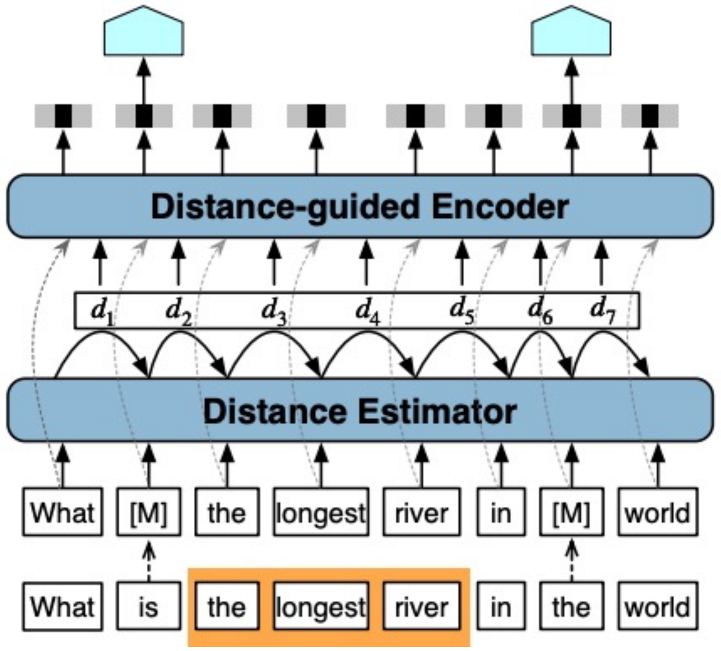
- Warm up the distance estimator via unsupervised extracted phrases
  - Can use any phrase tagger (e.g., UCPhrase)
- Encourage the average intra-phrase distance to be smaller than the average phrase boundary distance through a margin loss

$$\ell_{phrase} = \frac{1}{4} \cdot (max(0, \mathbf{d}_3 - \mathbf{d}_2) + max(0, \mathbf{d}_3 - \mathbf{d}_5) + max(0, \mathbf{d}_4 - \mathbf{d}_2) + max(0, \mathbf{d}_4 - \mathbf{d}_5))$$



Phrase-Guided Masked Language Modeling

- Given a sentence with tagged local phrases, sample a subset of them phrases to be excluded from being masked out
- By doing so, we try to push the model out of its comfort zone of local structure learning, and encourage it to focus more on how the local constituents are connected



Results

- Phrase-guided masked language modeling (PMLM) and phrase-regularized warm-up (PRW) both help improve the performance of existing LM-based parsers

Methods	F1 (%)
PRPN (Shen et al., 2018a)	37.4
ON-LSTM (Shen et al., 2018b)	47.7
URNNG (Kim et al., 2019c)	52.4
C-PCFG (Kim et al., 2019b)	55.2
Neural L-PCFGs (Zhu et al., 2020)	55.3
TreeTransformer (Wang et al., 2019)	47.9
+ PMLM	48.7
+ PRW	49.0
+ PRW + PMLM	49.3
StructFormer (Shen et al., 2020)	54.0
+ PMLM	54.1
+ PRW	55.3
+ PRW + PMLM	55.7

Table 1: Unlabeled F1 score (%) for unsupervised constituency parsing on WSJ test set.

Method	NP	VP	ADJ	ADV	SBA	PP
PRPN	59.2	46.7	44.3	32.8	50.0	57.2
ON-LSTM	64.5	41.0	38.1	31.6	52.5	54.4
C-PCFG	74.7	41.7	40.4	52.5	56.1	68.8
TreeTransformer	63.7	37.1	32.3	56.8	37.0	49.7
+ PMLM	63.5	37.9	31.7	56.8	38.0	50.4
+ PRW	64.2	36.3	27.9	53.8	36.2	53.0
+ PRW + PMLM	64.2	37.2	29.6	53.7	35.9	53.3
StructFormer	73.7	43.2	53.4	70.5	51.8	64.5
+ PMLM	73.6	43.7	53.4	69.3	51.9	64.6
+ PRW	74.0	44.9	52.9	69.9	52.7	69.4
+ PRW + PMLM	74.2	45.1	53.2	69.3	53.9	70.1

Table 2: Recall scores (%) of typed gold constituents.

Outline

- ❑ Phrase Mining
- ❑ Constituency Parsing
- ❑ Named Entity Recognition (NER)
  - ❑ Few-shot NER and Entity Typing
    - ❑ Few-Shot Named Entity Recognition: An Empirical Baseline Study [EMNLP'2021]
    - ❑ Few-Shot Fine-Grained Entity Typing with Automatic Label Interpretation and Instance Generation [KDD' 2022]
  - ❑ Distantly-supervised NER
- ❑ Taxonomy Construction

Motivation

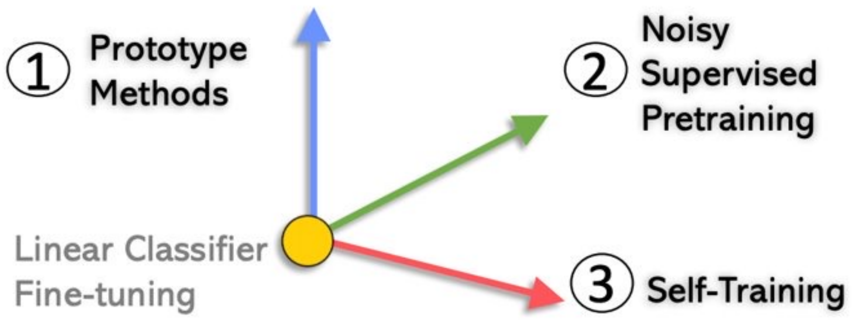
- Named entity recognition (NER) is a fundamental task in NLP with a wide spectrum of applications
  - question answering
  - knowledge base construction
  - dialog systems
  - …
- Deep neural models have achieved enormous success for NER
- However, a common bottleneck of training deep learning models is the acquisition of abundant high-quality human annotations (every entity in the sequence needs to be labeled!)

Few-shot NER

- Current NER models are trained for a series of fixed categories (e.g., PERSON, LOCATION, etc.) using large amounts of labeled data.
- Few-shot NER learns to transfer to new domains/categories with only a few training examples.

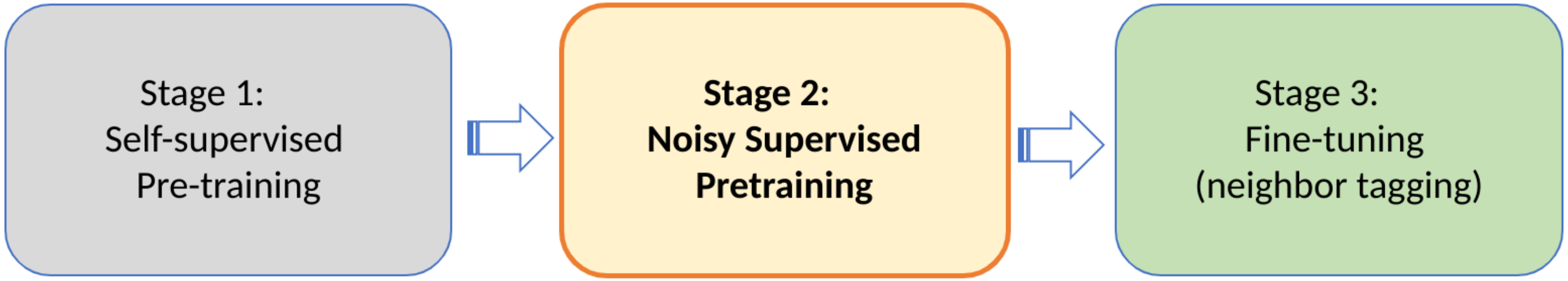
Our Empirical Study on Three Directions

- We explore three directions to improve the generalization ability of models in limited NER data settings.
- Prototype Methods (P) : A training objective typically used in few-shot learning setting to represent each class as a prototype
- Noisy Supervised Pretraining (NSP)
- Self-Training (ST)



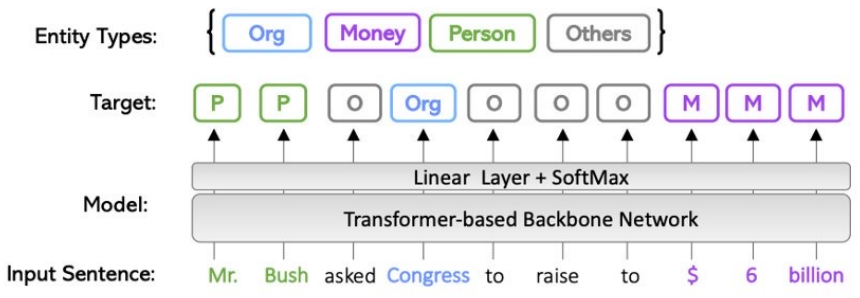
Noisy Supervised Pretraining

- Generic representations via self-supervised pre-trained language models are pre-trained with the task of randomly masked token prediction.
- The goal of NER: Identifying named entities as emphasized tokens and assigning labels to them. –> Outweigh the representations of entities for NER.
- Noisy Supervised Pretraining (NSP): Let the feature extractor model learn a discriminative NER space

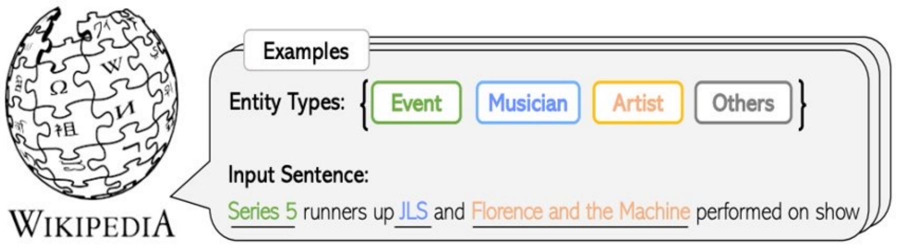


Noisy Supervised Pretraining

- The WiFine[1] dataset: 113 entity types; over 50 million sentences.



(a) Baseline: NER with a linear classifier



(c) Noisy supervised pre-training

	Wikipedia (6.8GB)	CONLL-2003	OntoNER	…
Research Topic	NER	NER	NER	
# Entity Types	113	4	18	
# Entity Instances	70,000,000+	23,499	11,066	
# Training Sent.	52,000,000+	14,041	8,528	
# Training Token.	1,300,000,000+	203,621	147,724	

Target

[1] Transforming Wikipedia into a Large-Scale Fine-Grained Entity Type Corpus. Abbas Ghaddar, Philippe Langlais, 2018

Self-Training

- Learn teacher model θ_tea via cross-entropy loss with labeled tokens.
- Generate soft labels using a teacher model on unlabeled tokens.

$$\tilde{\boldsymbol{y}}_i = f_{\boldsymbol{\theta}^{\texttt{tea}}}(\tilde{\boldsymbol{x}}_i), \forall \tilde{\boldsymbol{x}}_i \in \mathcal{D}^{\mathrm{U}}$$

- Learn a student model θ_stu via cross entropy loss on both labeled and unlabeled tokens.

$$\begin{aligned}\mathcal{L}_{\mathrm{ST}} = & \frac{1}{|\mathcal{D}^{\mathrm{L}}|} \sum_{\boldsymbol{x}_i \in \mathcal{D}^{\mathrm{L}}} \mathcal{L}(f_{\boldsymbol{\theta}^{\texttt{stu}}}(\boldsymbol{x}_i), \boldsymbol{y}_i) \\ & + \frac{\lambda_{\mathrm{U}}}{|\mathcal{D}^{\mathrm{U}}|} \sum_{\tilde{\boldsymbol{x}}_i \in \mathcal{D}^{\mathrm{U}}} \mathcal{L}(f_{\boldsymbol{\theta}^{\texttt{stu}}}(\tilde{\boldsymbol{x}}_i), \tilde{\boldsymbol{y}}_i)\end{aligned}$$

Experiments

- We collect 10 benchmark datasets for evaluating the model.
- The reason that we use multiple datasets across different domains is that they contain various entity types that could not be covered by the pretraining dataset.

Datasets	CoNLL	Onto	WikiGold	WNUT	Movie	Restaurant	SNIPS	ATIS	Multiwoz	I2B2
Domain	News	General	General	Social Media	Review	Review	Dialogue	Dialogue	Dialogue	Medical
#Train	14.0k	60.0k	1.0k	3.4k	7.8k	7.7k	13.6k	5.0k	20.3k	56.2k
#Test	3.5k	8.3k	339	1.3k	2.0k	1.5k	697	893	2.8k	51.7k
#Entity Types	4	18	4	6	12	8	53	79	14	23

Fine-tuning on Unseen Tasks

Datasets	Settings	① LC	② LC + NSP	③ P	④ P + NSP	⑤ LC + ST	⑥ LC + NSP + ST
CoNLL	5-shot	0.535	0.614	0.584	0.609	0.567	**0.654**
	10%	0.855	0.891	0.878	0.888	0.878	**0.895**
	100%	0.919	**0.920**	0.911	0.915	-	-
Onto	5-shot	0.577	0.688	0.533	0.570	0.605	**0.711**
	10%	0.861	**0.869**	0.854	0.846	0.867	0.867
	100%	0.892	**0.899**	0.886	0.883	-	-
WikiGold	5-shot	0.470	0.640	0.511	0.604	0.481	**0.684**
	10%	0.665	0.747	0.692	0.701	0.695	**0.759**
	100%	0.807	**0.839**	0.801	0.827	-	-
WNUT17	5-shot	0.257	0.342	0.295	0.359	0.300	**0.376**
	10%	0.483	0.492	0.485	0.478	0.490	**0.505**
	100%	0.489	0.520	0.552	**0.560**	-	-
MIT Movie	5-shot	0.513	0.531	0.380	0.438	0.541	**0.559**
	10%	0.651	0.657	0.563	0.583	0.659	**0.666**
	100%	**0.693**	0.692	0.632	0.641	-	-

Columns: Different Models

LC: Linear Classifier + PLM

NSP: Noisy Supervised Pretraining

P: Prototype-based Objective

ST: Self-Training

Rows: Different Tasks

5-shot: 5 example sentences for each entity type

10%: only use 10 percent of training data

100%: use all training data

Observations: 1. Noisy supervised pretraining creates a better discriminative NER space, leading to better results in most datasets.
2. Prototype-based methods can be better than linear classifier when the size of both labels and entity types are small.
3. Self-training methods that leverage unlabeled data constantly improve the results.

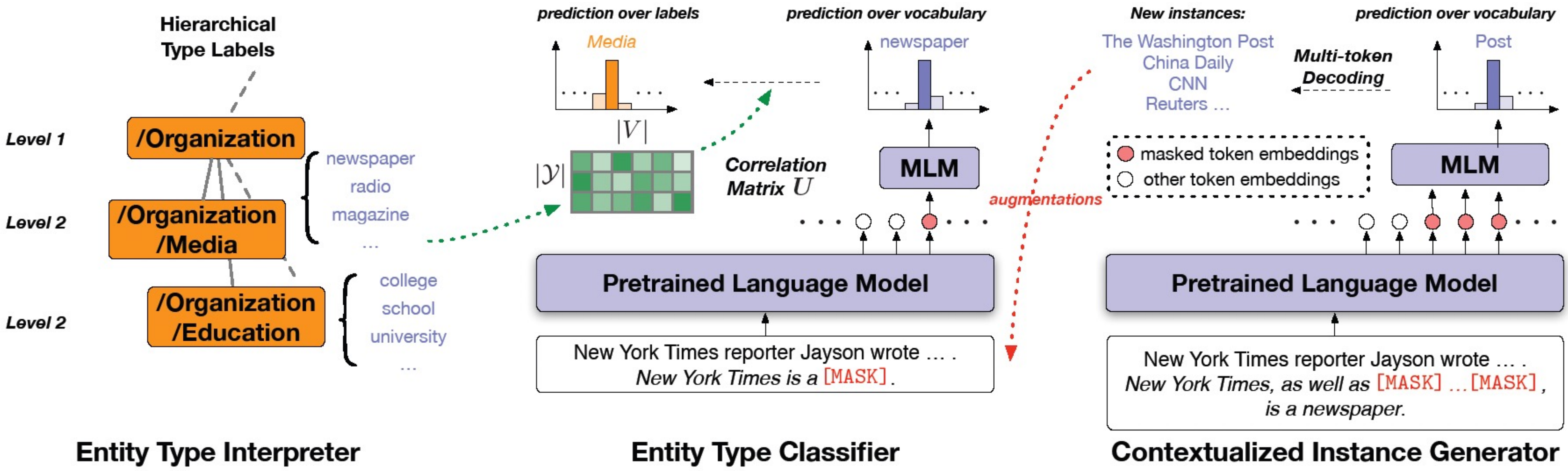
Outline

- Phrase Mining
- Constituency Parsing
- Named Entity Recognition (NER)
  - Few-shot NER and Entity Typing
    - Few-Shot Named Entity Recognition: An Empirical Baseline Study [EMNLP'2021]
    - Few-Shot Fine-Grained Entity Typing with Automatic Label Interpretation and Instance Generation [KDD' 2022]
  - Distantly-supervised NER
- Taxonomy Construction

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 :

[Context]. **New York Times**, as well as [MASK] [MASK] [MASK], is a *newspaper*.

Entity Mention

ranges from
1 to the length of original
entity mention

Predicted by
Entity Type
Classifier

Multi-Token Instance Generation

- We generate candidate instances by filling in one blank at each step (sampled from the output distribution), and recursively predict the other blanks conditioned on the already filled blanks.

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.

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

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

The conditional probability at each step

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			BBN			Few-NERD		
	(Acc.)	(Micro-F1)	(Macro-F1)	(Acc.)	(Micro-F1)	(Macro-F1)	(Acc.)	(Micro-F1)	(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
- Constituency Parsing
- 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

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

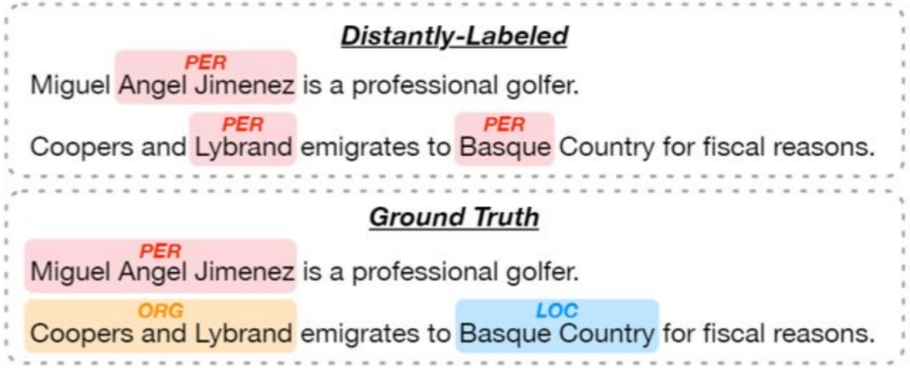
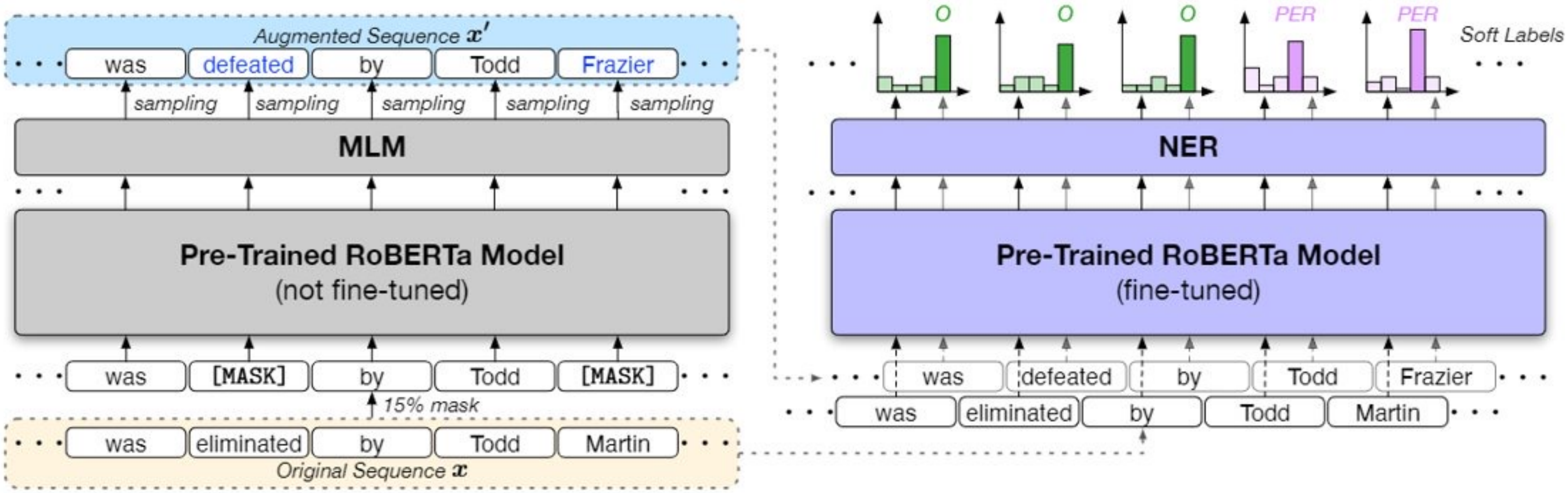


Figure 1: Distant labels obtained with knowledge bases may be incomplete and noisy, resulting in wrongly-labeled tokens.

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)

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

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

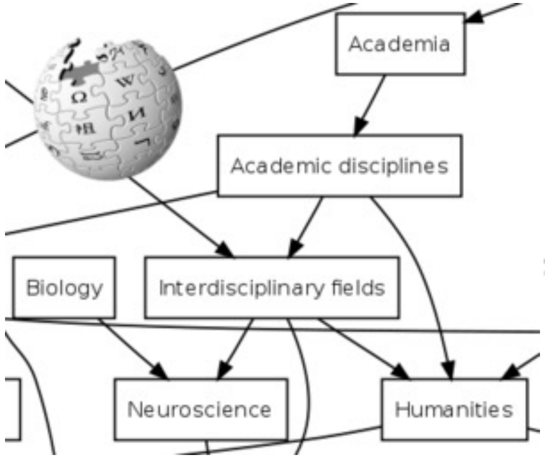
Table 2: Performance all methods on three datasets measured by precision (Pre.), recall (Rec.) and F1 scores.

Outline

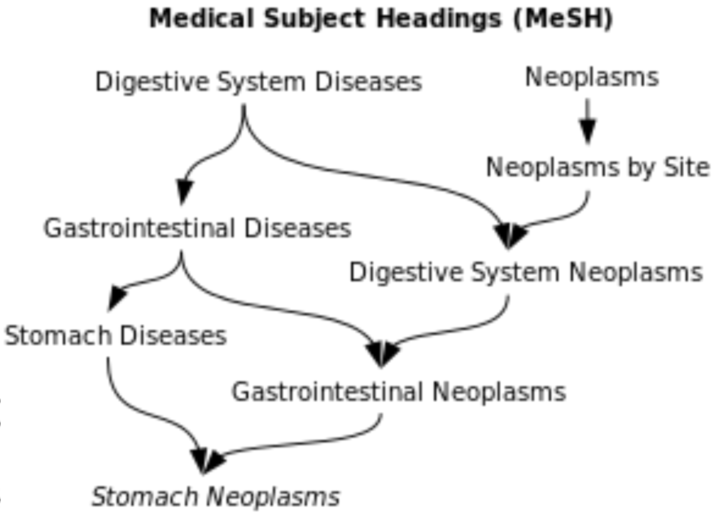
- Phrase Mining
- Constituency Parsing
- Named Entity Recognition
- Taxonomy Construction
  - Taxonomy Basics and Construction
  - Taxonomy Construction with Minimal User Guidance
  - Taxonomy Expansion

What is a Taxonomy?

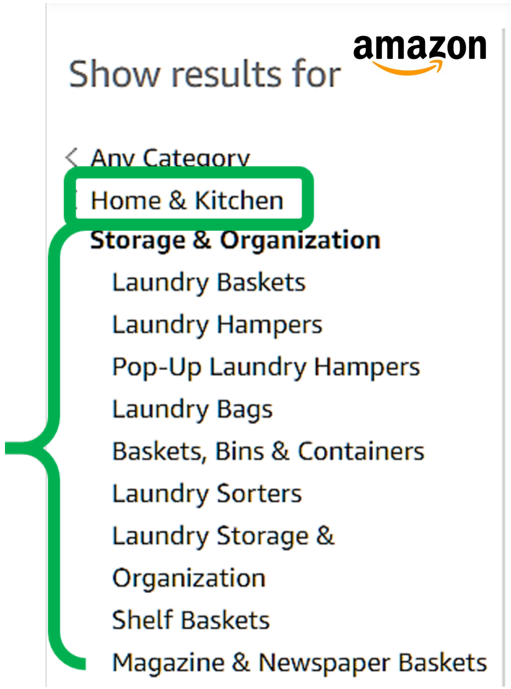
- Taxonomy is a hierarchical organization of concepts
- Taxonomy can benefit many knowledge-rich applications
  - Knowledge Organization, Document Categorization, Recommender System …



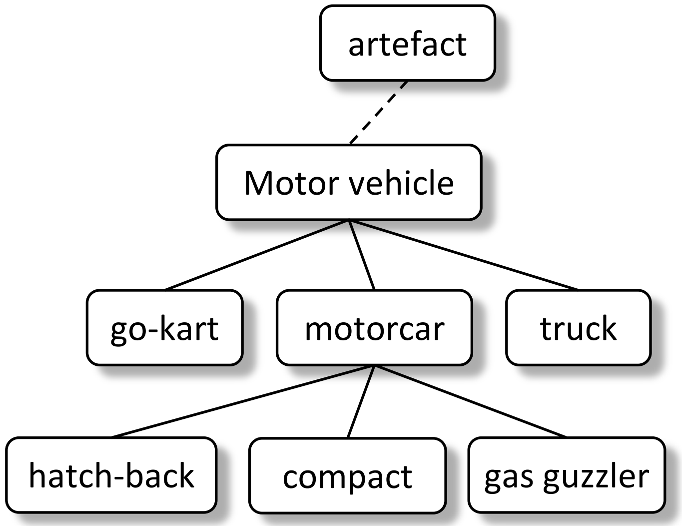
Wikipedia Category



MeSH



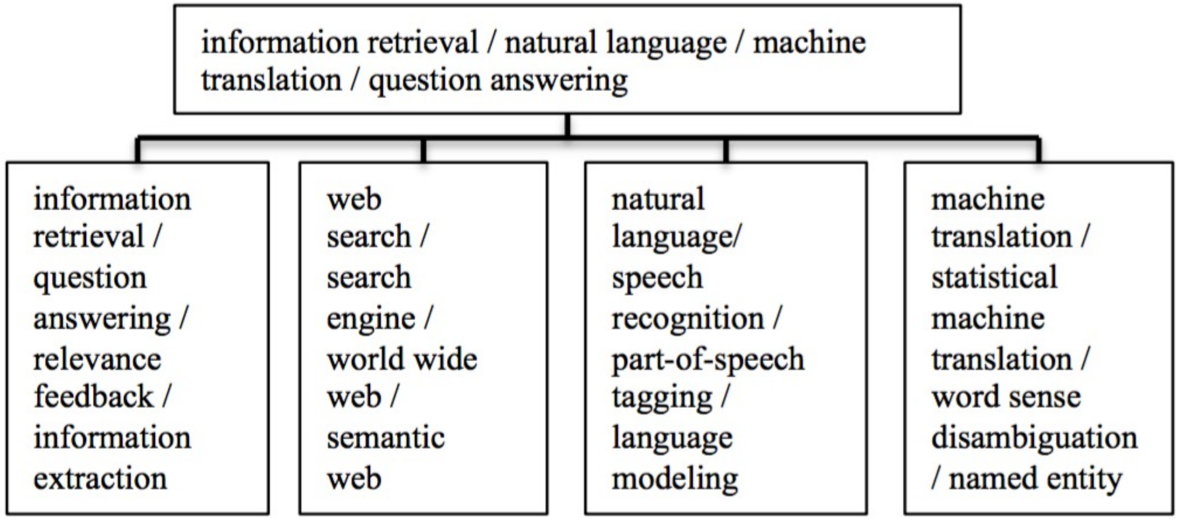
Amazon Product Category



WordNet

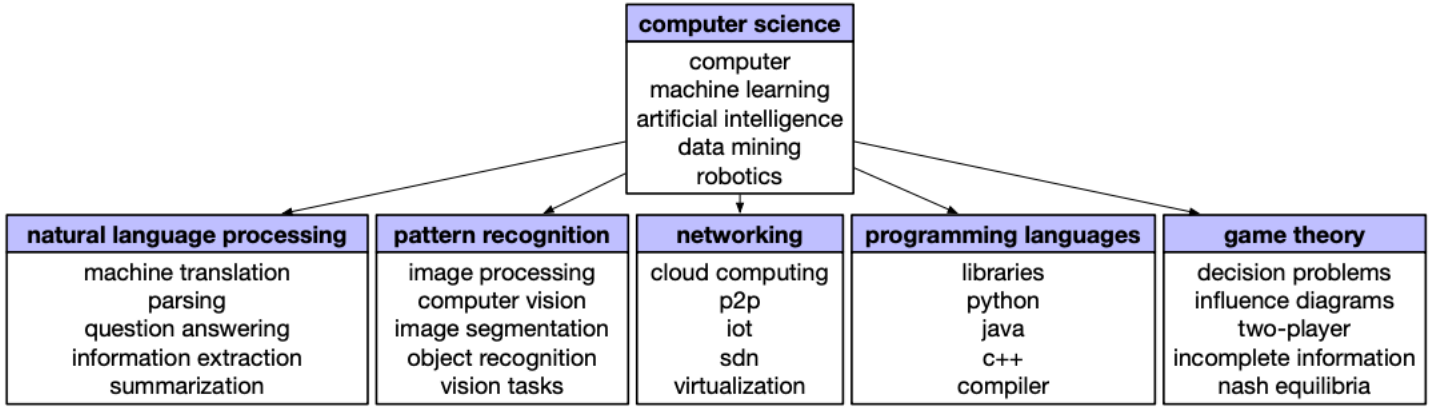
Clustering-based Taxonomy

- Compared to instance-based taxonomy (e.g., WordNet), clustering-based taxonomy has wider semantic coverage and facilitates clearer understanding of concepts.
- We focus on introducing clustering-based taxonomy construction in this tutorial.

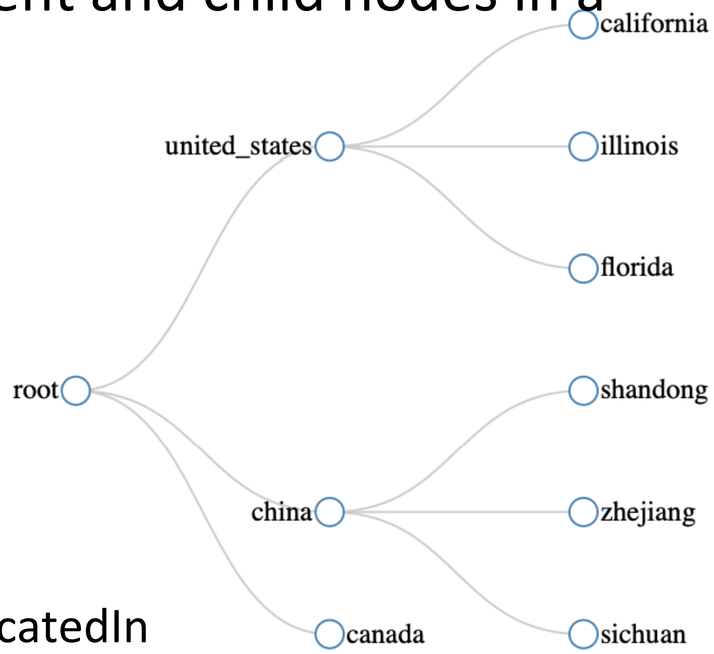


Multi-faceted Taxonomy Construction

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



Relation: IsSubfieldOf



Relation: IsLocatedIn

Two stages in constructing a complete taxonomy

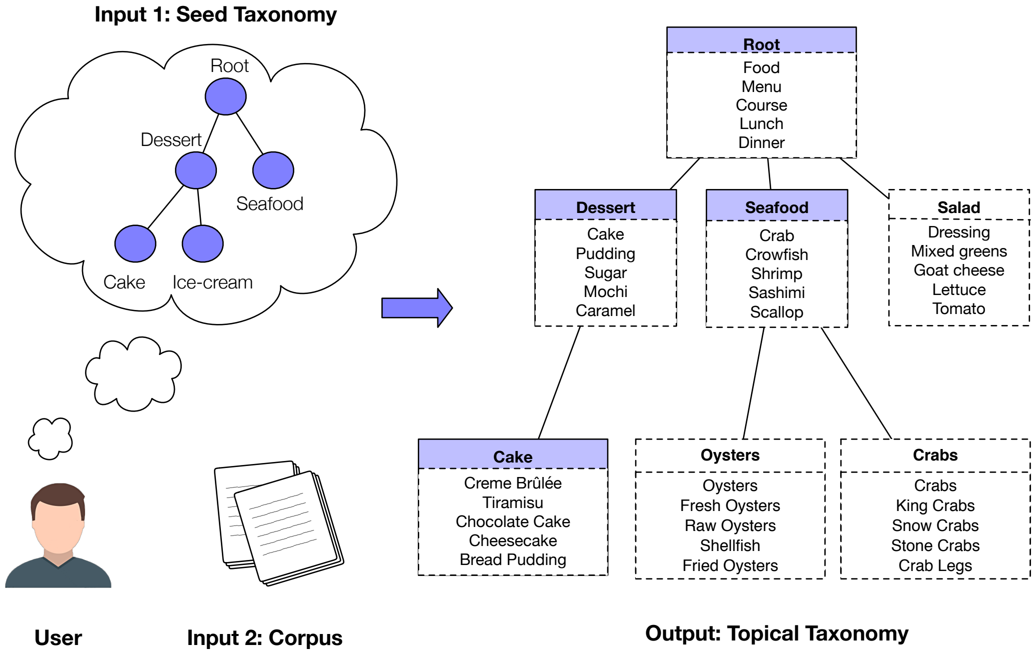
- Taxonomy Construction with Minimal User Guidance
  - Use a set of entities (possibly a seed taxonomy in a small scale) and unstructured text data to build a taxonomy organized by certain relations
- Taxonomy Expansion
  - Update an already constructed taxonomy by attaching new items to a suitable node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

Outline

- Phrase Mining
- Constituency Parsing
- Named Entity Recognition
- Taxonomy Construction
  - Taxonomy Basics and Construction
  - Taxonomy Construction with Minimal User Guidance
  - Taxonomy Expansion

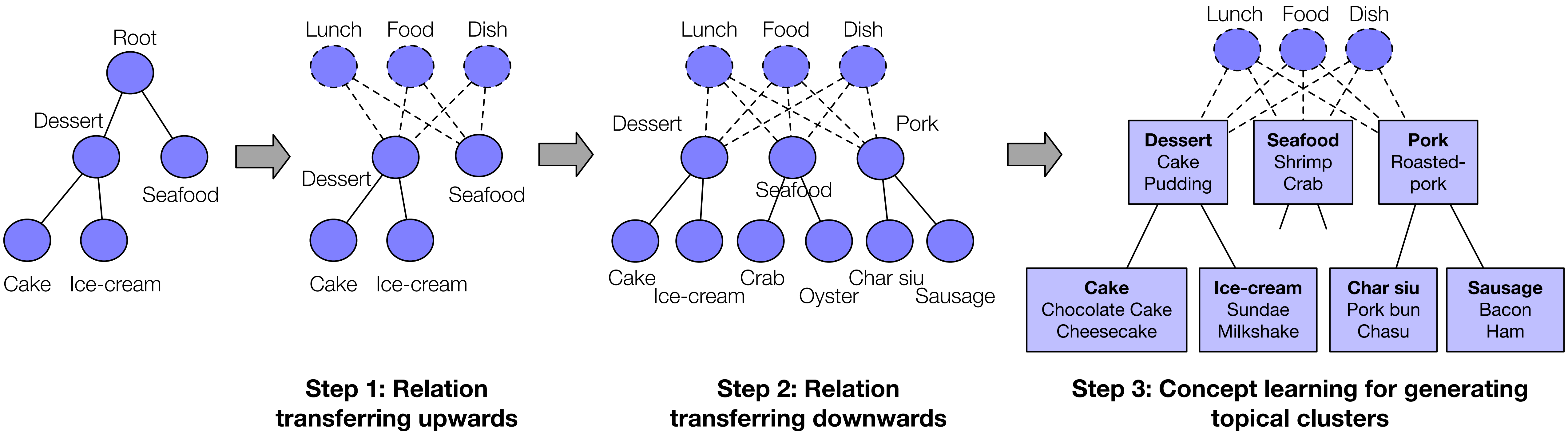
Seed-Guided Topical Taxonomy Construction

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



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

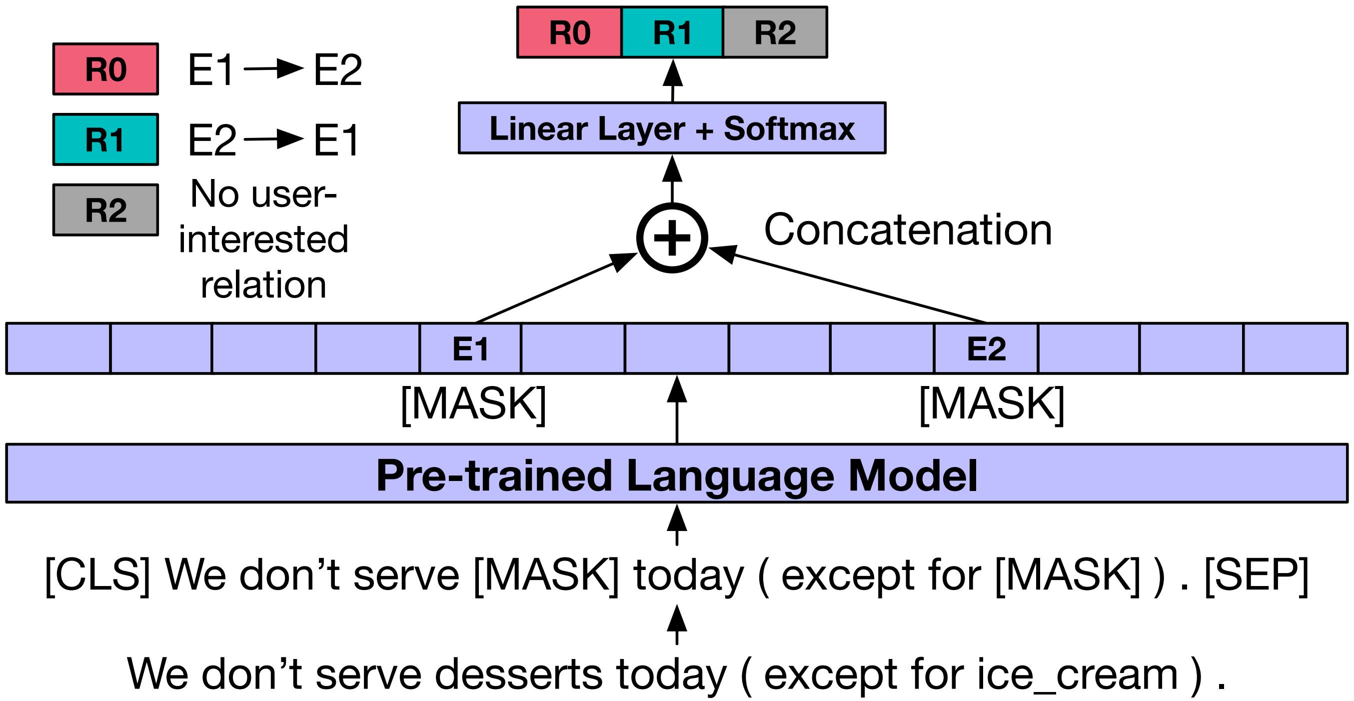
CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring [KDD'20]



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

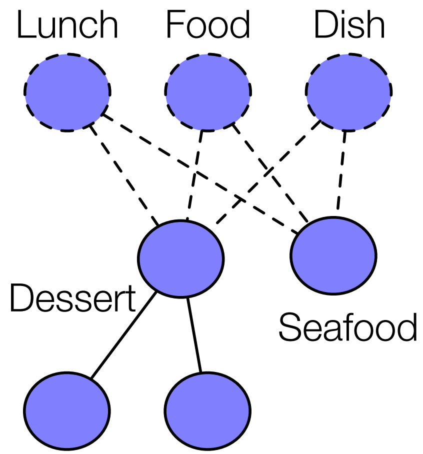
Relation Learning

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



Relation Transferring

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

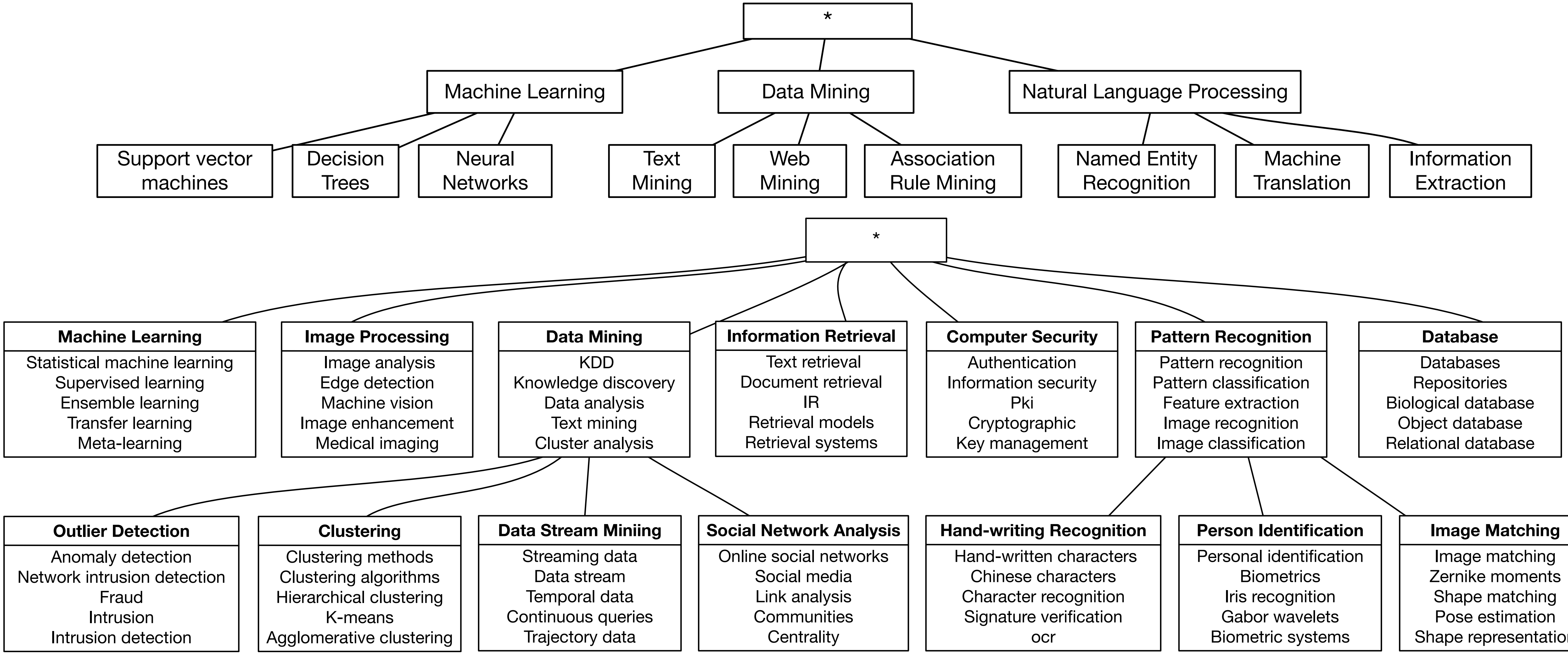


- We extract a list of parent nodes for each seed topic using the relation classifier. The common parent nodes shared by all user-given topics are treated as root nodes.
- To discover new topics (e.g, Pork), we transfer the relation downwards from these root nodes.

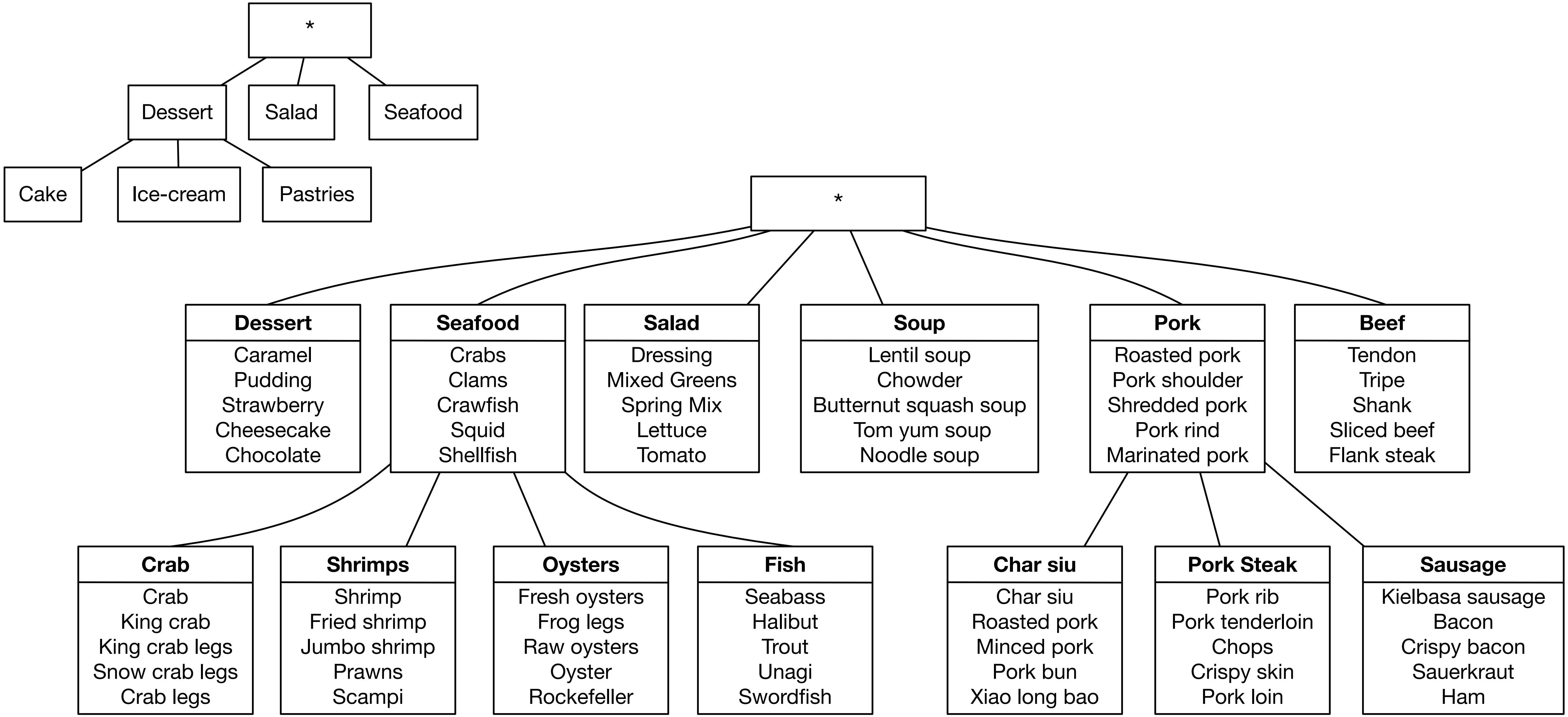
Concept Learning

- Subtopics should satisfy the following two constraints:
  - 1. must belong to representative words of that parent topic.
  - 2. must share parallel relations with given seed taxonomy.
- Learn a discriminative embedding space, so that each concept is surrounded by its representative terms.
- Therefore, we leverage a **weakly-supervised text embedding framework** to discriminate concepts in the embedding space, and this algorithm will be introduced in the next section.

Qualitative Results



Qualitative Results



Outline

- Phrase Mining
- Constituency Parsing
- Named Entity Recognition
- Taxonomy Construction
  - Taxonomy Basics and Construction
  - Taxonomy Construction with Minimal User Guidance
  - Taxonomy Expansion

Taxonomy Enrichment: Motivation

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

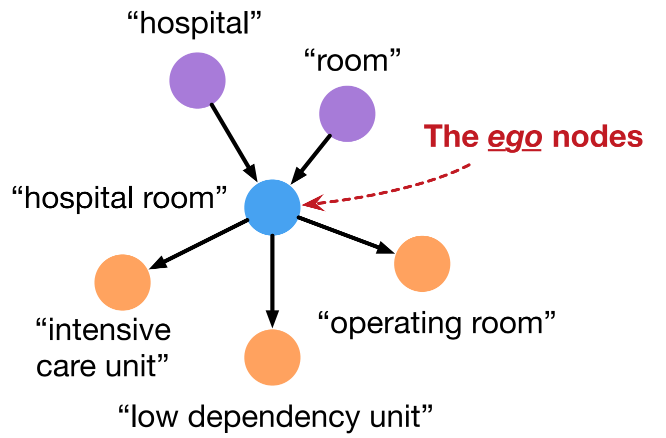
TaxoExpan: Self-supervised Taxonomy Expansion with Position-Enhanced Graph Neural Network [WWW' 20]

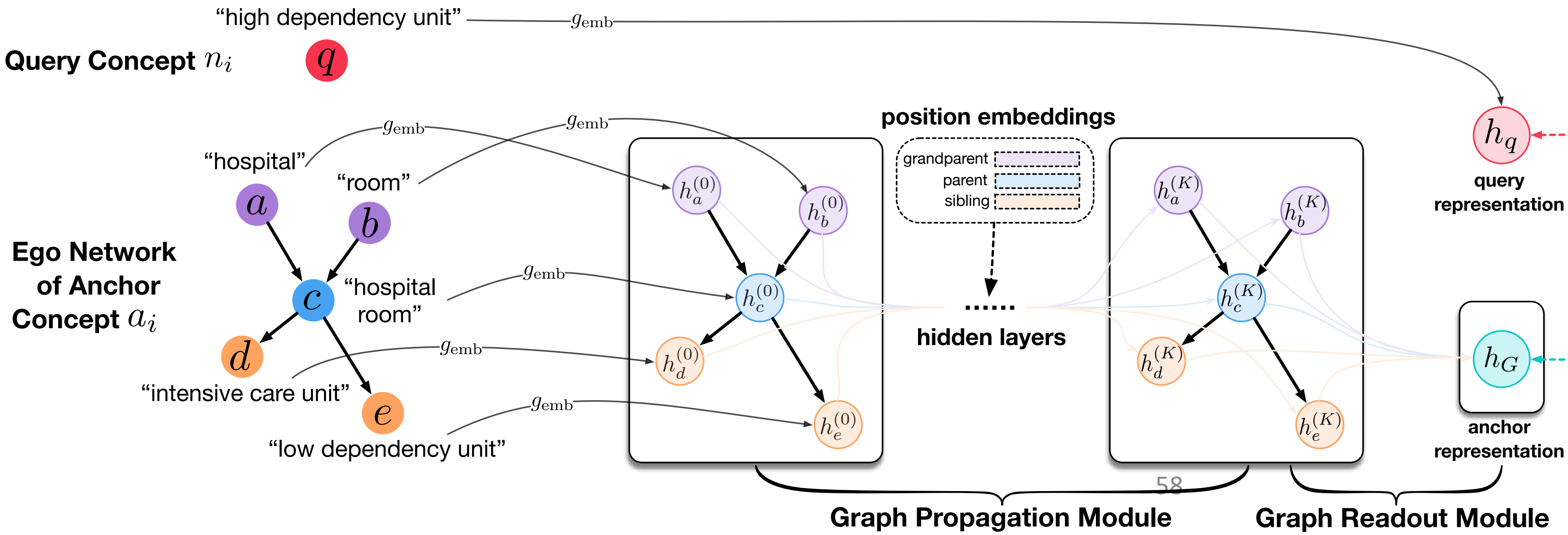
- **Two steps** in solving the problem:
  - Self-supervised term extraction
    - Automatically **extracts emerging terms** from a target domain
  - Self-supervised term attachment
    - A multi-class classification to match a new node to its potential parent
    - Heterogenous sources of information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

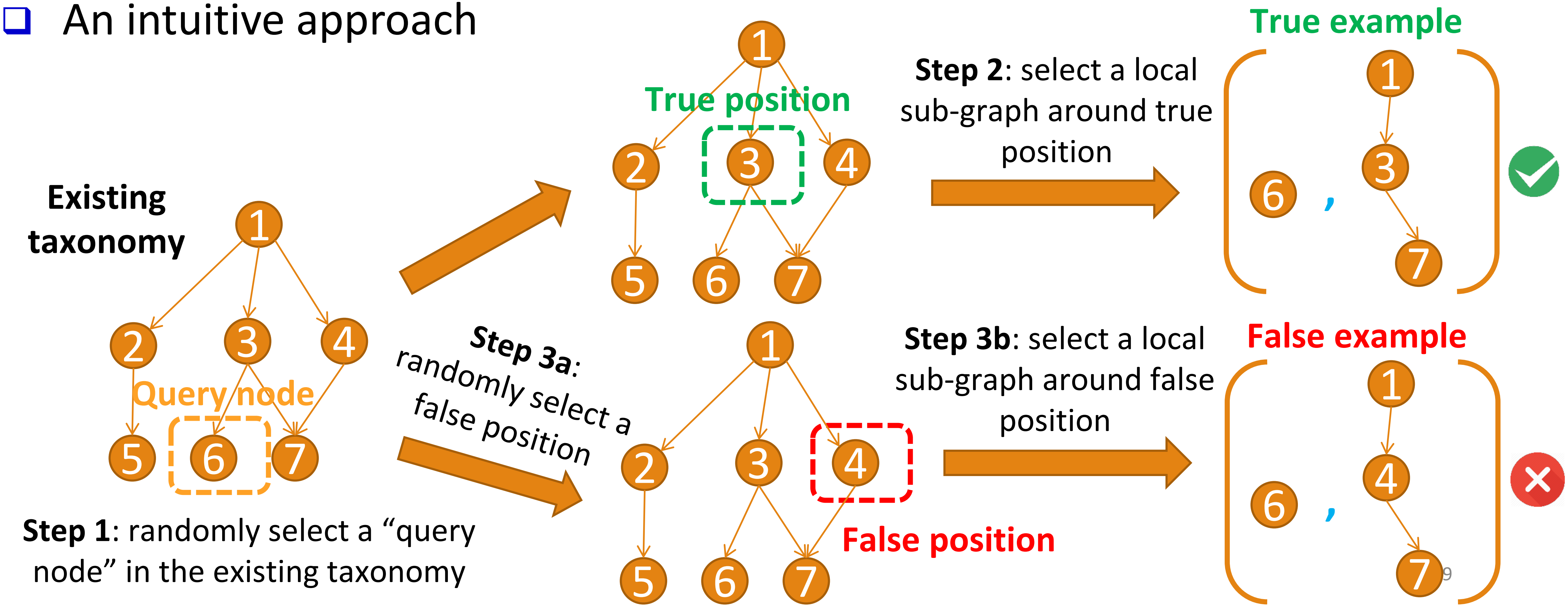
Query: “high dependency unit”





Leveraging Existing Taxonomy for Self-supervised Learning

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



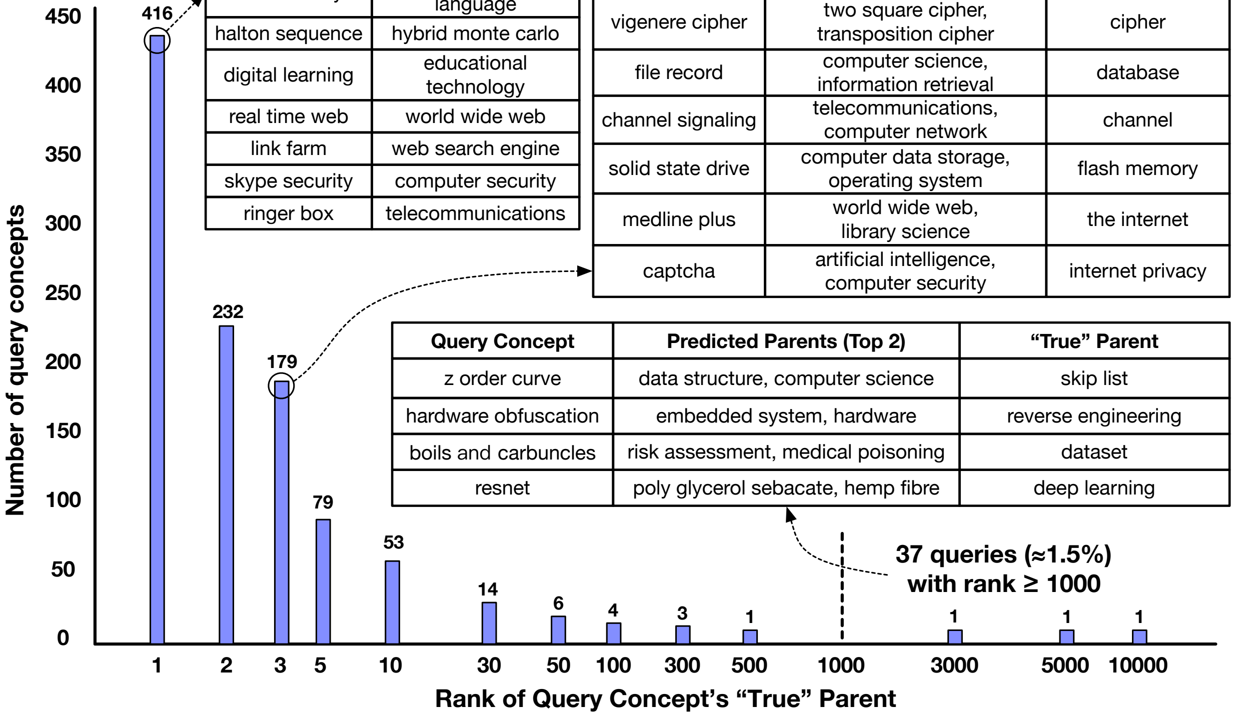
TaxoExpan Framework Analysis

- 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

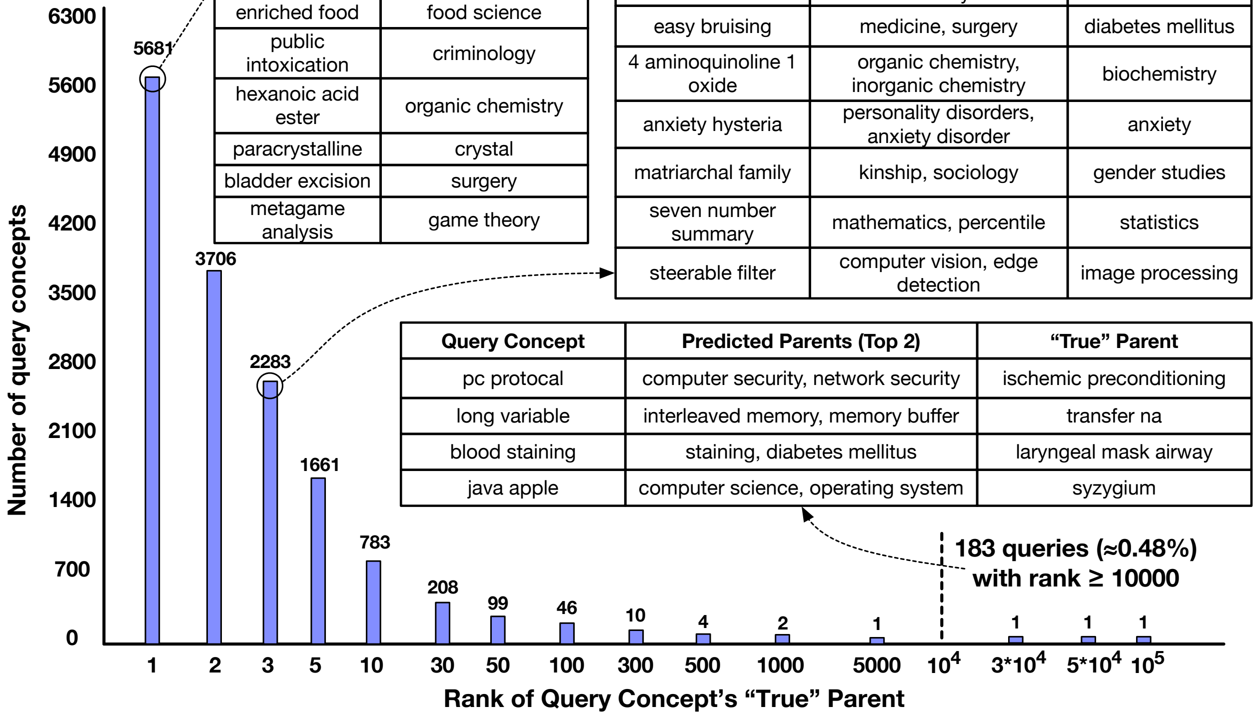


(a) MAG-CS Dataset (totally 2450 query concepts)

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

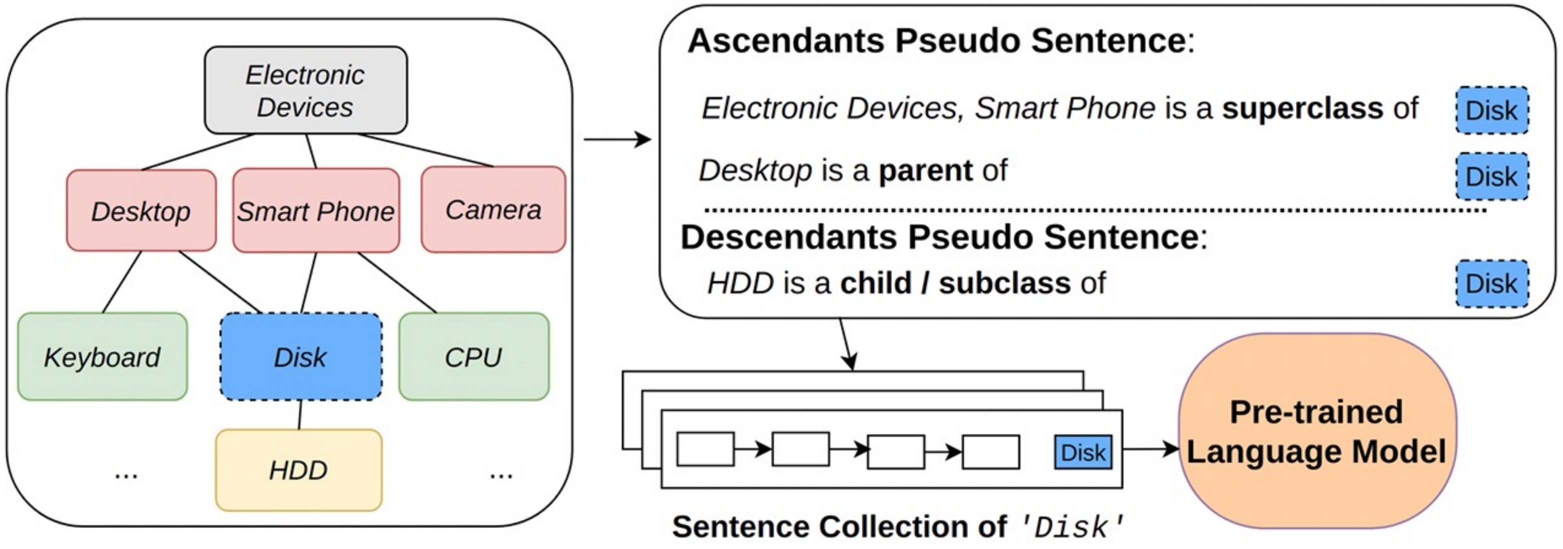


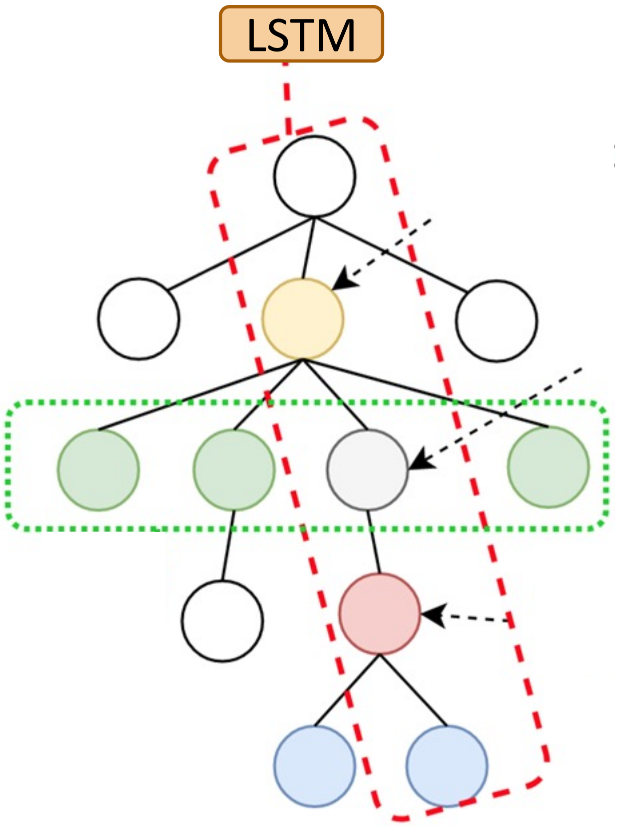
(b) MAG-Full Dataset (totally 37804 query concepts)

60

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

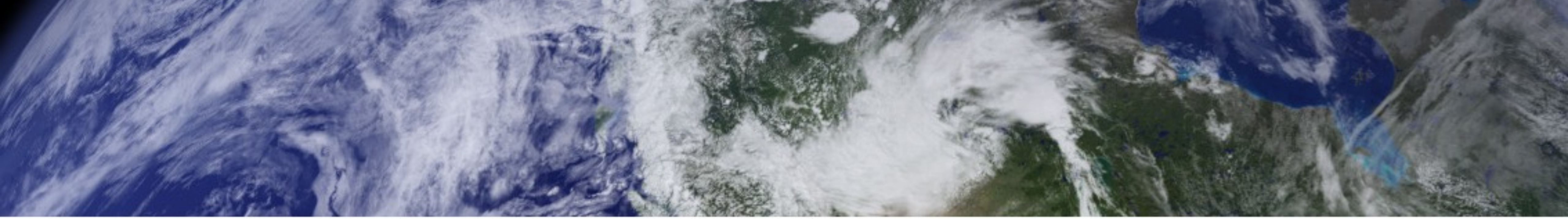
- Extra semantic information
  - Taxonomy-contextualized embedding
  - Layer-aware representation





References

- Xiaotao Gu , Zihan Wang , Zhenyu Bi , Yu Meng, Liyuan Liu, Jiawei Han, Jingbo Shang. “UCPhrase: Unsupervised Context-aware Quality Phrase Tagging.” (KDD’21)
- Xiaotao Gu, Yikang Shen, Jiaming Shen, Jingbo Shang, Jiawei Han, “Phrase-aware Unsupervised Constituency Parsing” (ACL’22)
- Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R Voss, and Jiawei Han. 2014. Scalable Topical Phrase Mining from Text Corpora. Proceedings of the VLDB Endowment 8, 3 (2014).
- Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. IEEE Transactions on Knowledge and Data Engineering 30, 10 (2018), 1825–1837.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.
- Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. 55–60.
- 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)
- 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)
- Minhao Jiang, Xiangchen Song, Jieyu Zhang and Jiawei Han, “TaxoEnrich: Self-Supervised Taxonomy Completion via Structure-Semantic Representations” (WWW’22)



Q&A