Towards Automatic Construction of Knowledge Graphs from Unstructured Text

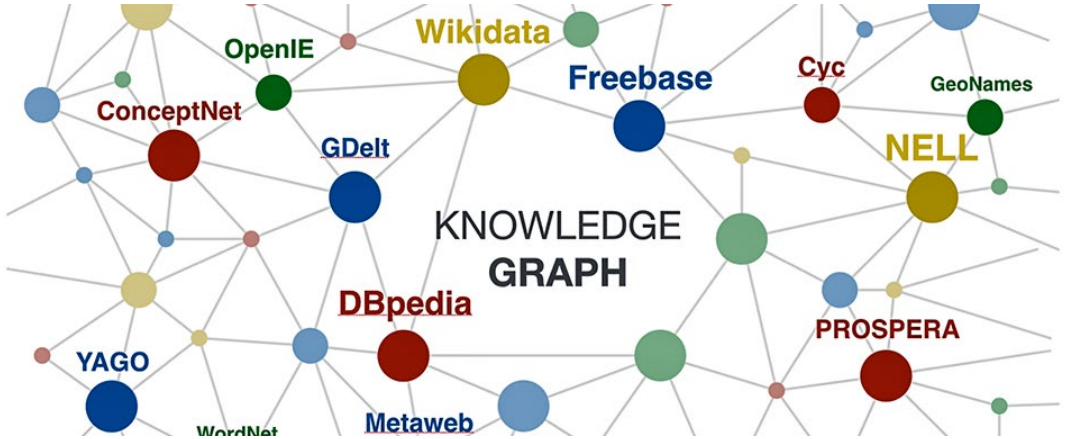
JIAWEI HAN
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
JULY 17, 2022

Outline

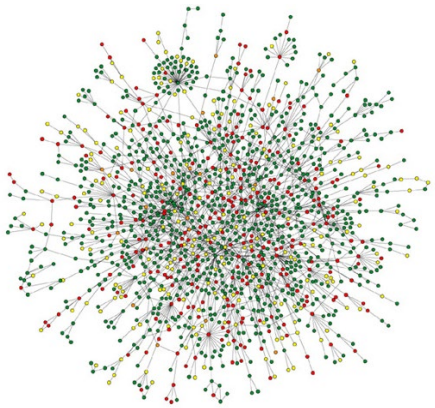
- What Kinds of Knowledge Graphs Do We Really Need?
- Key Issue: Construction of Theme-/Corpus-Based Knowledge Graphs
- The Role of Embedding and PLM in Knowledge Graph Construction
- Data Preparation: Taxonomy-Guided Text Classification
- Identifying Knowledge Graph Primitives: Entities, Properties and Relations
- Conclusion: Towards Theme/Corpus-Based Knowledge Graph Construction

Knowledge Graphs Are Ubiquitous

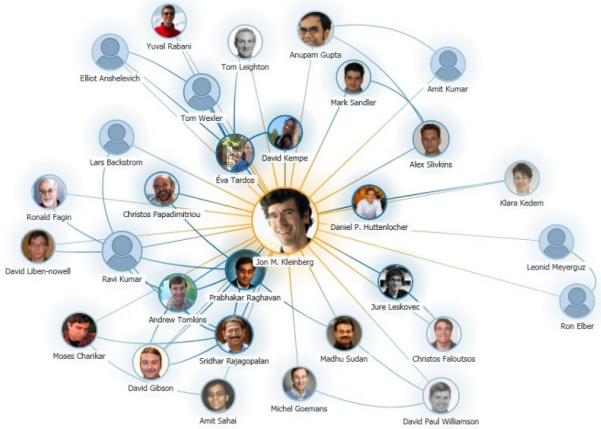
- Graphs and networks are ubiquitous
  - Web & Internet
  - Social networks
  - Biological networks
  - Research publication networks
  - ……
- Most would view knowledge graphs are graphs constructed from knowledge-bases



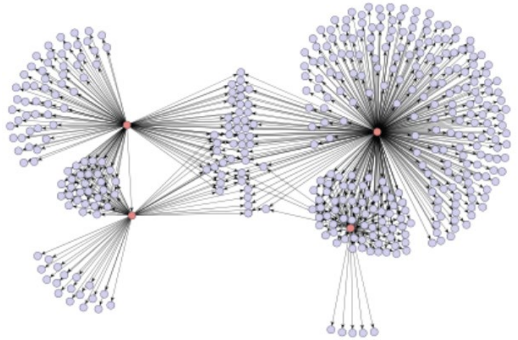
Social Networking Websites



Biological Network: Protein Interactions



Research Collaboration Network

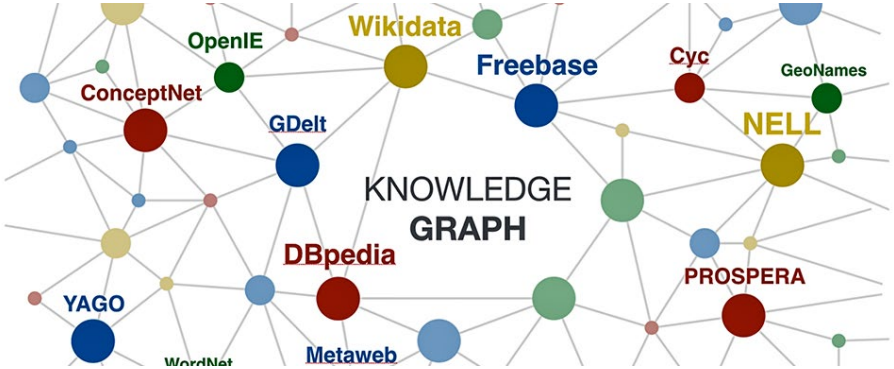


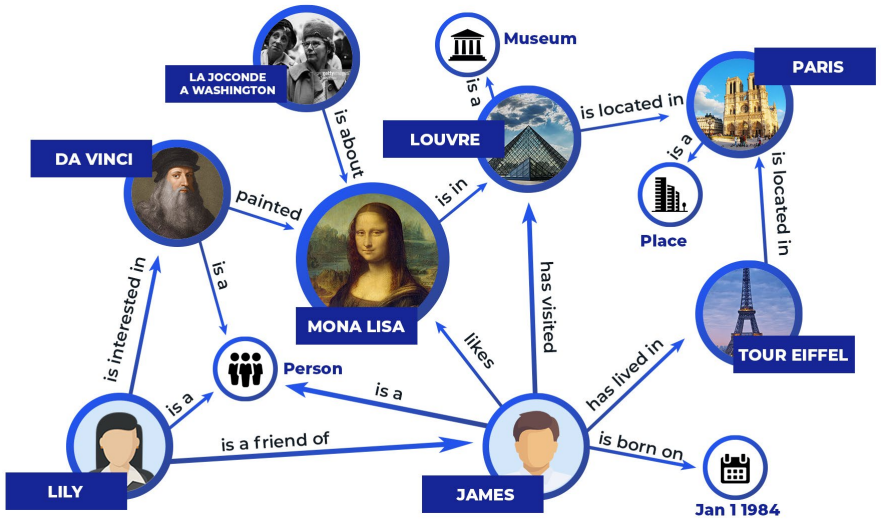
Product Recommendation Network

Ack. Figures are from Google images

What Kinds of Knowledge Graphs Do We Really Need?

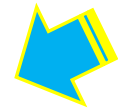
- What kinds of KGs could be most useful to us?
  - One gigantic knowledge graph vs. many small ones
  - From general knowledge-base vs. from a small text
  - Homogeneous (one-type) vs. heterogenous graphs
  - ……
- One general knowledge graph of world knowledge vs. theme/corpus-based small knowledge graphs
  - One big KG: ambiguous & dynamic entities/links
    - Ex. Thinking of Michael Jordan/Donald Trump
  - Theme-/corpus-based KG: Focused and accurate
    - Universally needed in applications from text
    - Challenges: human annotation vs. automation





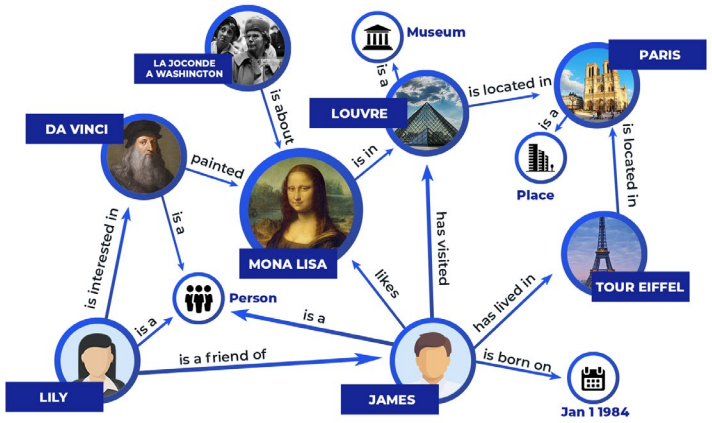
Ack. Figures are from Google images

Outline

- What Kinds of Knowledge Graphs Do We Really Need?
- Key Issue: Construction of Theme-/Corpus-Based Knowledge Graphs 
- The Role of Embedding and PLM in Knowledge Graph Construction
- Data Preparation: Taxonomy-Guided Text Classification
- Identifying Knowledge Graph Primitives: Entities, Properties and Relations
- Conclusion: Towards Theme/Corpus-Based Knowledge Graph Construction

Automated, Local Knowledge Graph Construction

- Local (theme-/corpus-based) knowledge graphs
  - Where are such local KGs?
    - None? – Scattered in local texts
      - Ex. Research papers, news, …
  - Human collection and annotation are unrealistic
    - Need automatic data collection and KG construction

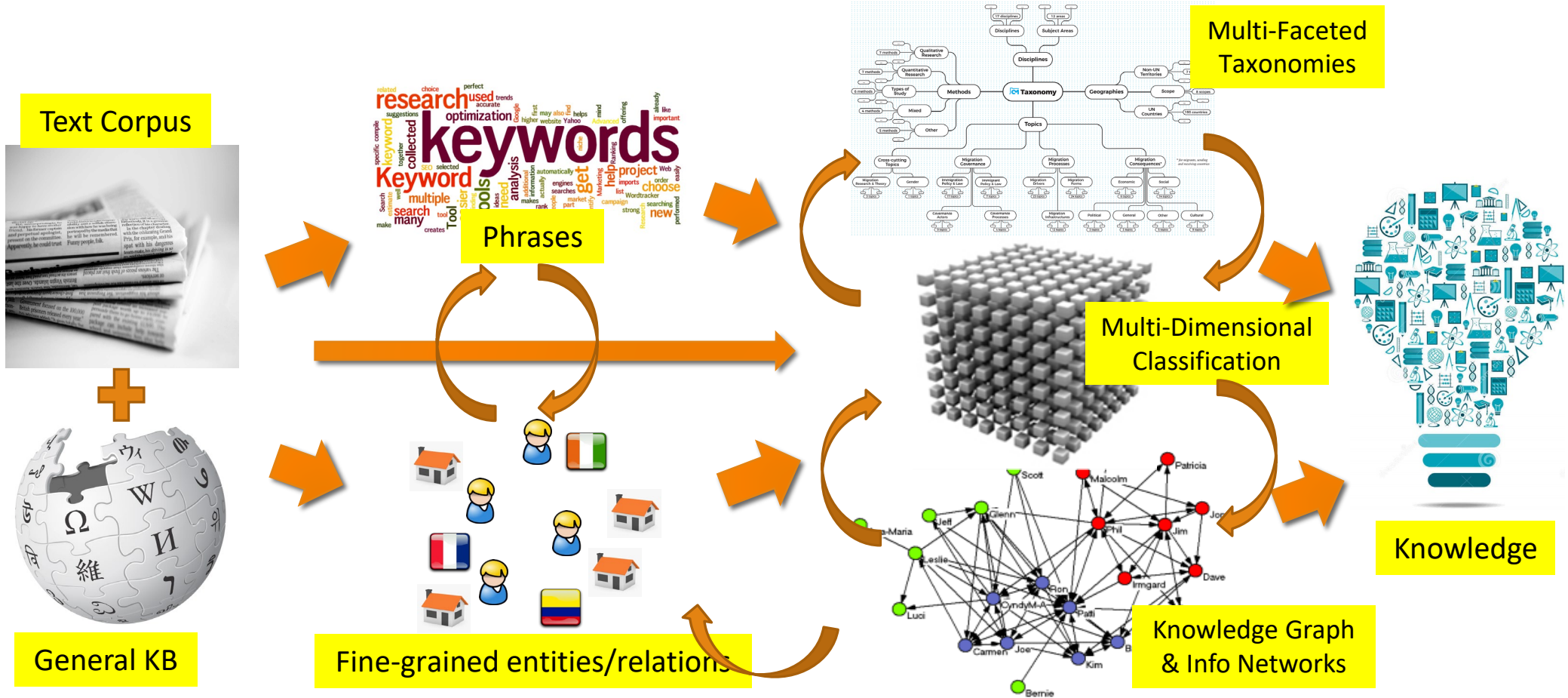


Ack. Figures are from Google images

- Automated, local (theme-/corpus-based) knowledge graph construction
  - An essential task for data mining and NLP (from text to knowledge graph: T2KG)
- Approaches to be explored
  - Embedding and pretrained language models (PLMs)
  - Data preparation and collection: taxonomy-guided text classification
  - Structured/typed information extraction for knowledge graph construction

Our General Roadmap for Mining Unstructured Text

- Mining structuring from unstructured text
  - Embedding and PLM for mining semantics from text
  - Automated mining of phrases, topics, entities, links and types from text corpora

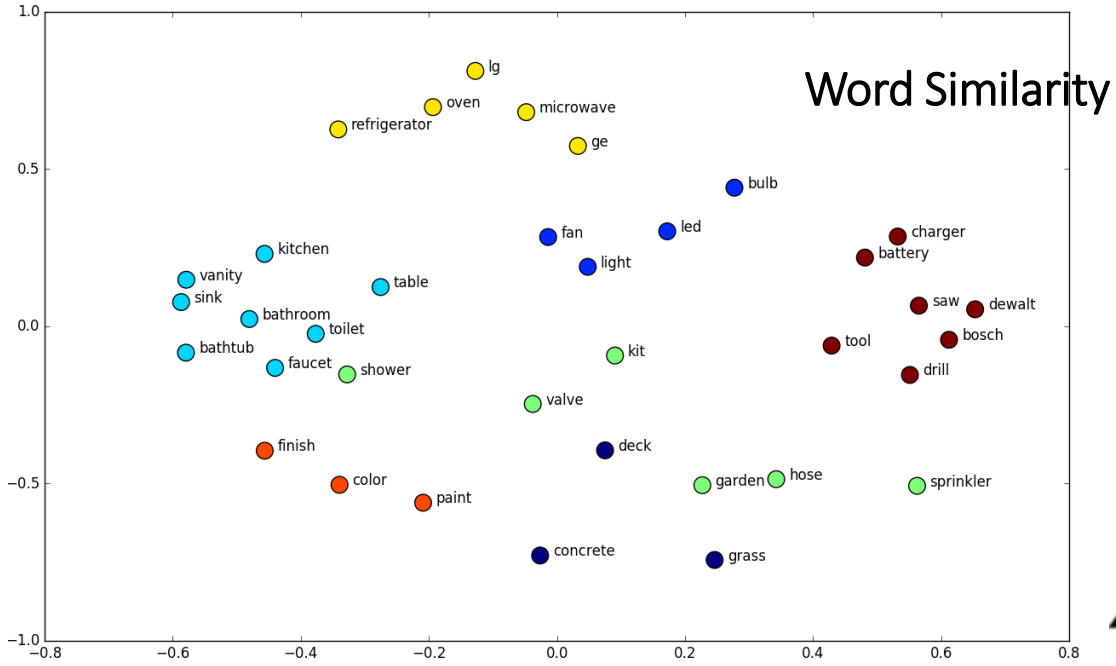


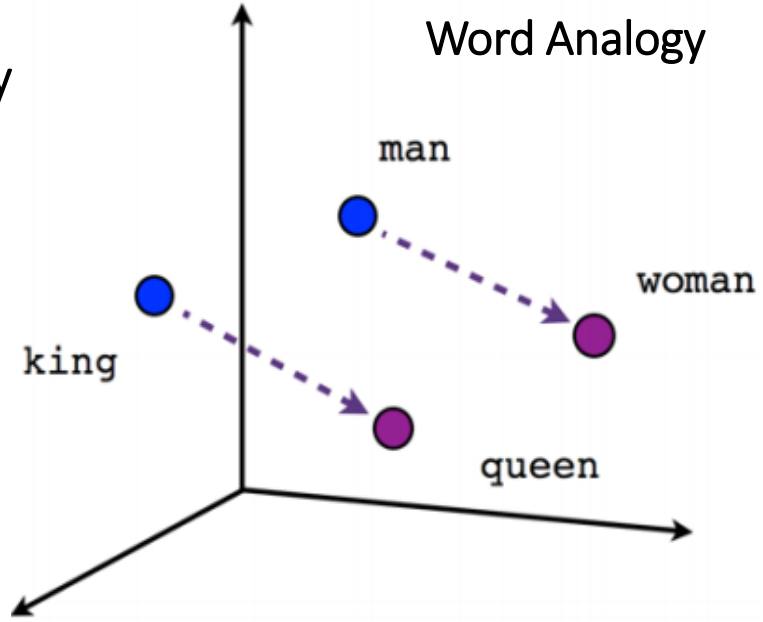
Outline

- What Kinds of Knowledge Graphs Do We Really Need?
- Key Issue: Construction of Theme-/Corpus-Based Knowledge Graphs
- The Role of Embedding and PLM in Knowledge Graph Construction 
- Data Preparation: Taxonomy-Guided Text Classification
- Identifying Knowledge Graph Primitives: Entities, Properties and Relations
- Conclusion: Towards Theme/Corpus-Based Knowledge Graph Construction

Representation Learning in Text: Text Embedding

- Symbolic representation (one-hot vector & bag of words) vs. distributive representation
- Embedding words in lower-dimension space: Handling sparsity & high dimensionality
- Unsupervised learning with distributive representation: A milestone in NLP and ML
- Key idea: Words with similar meanings are embedded closer
  - Word analogy: Linear relationships between words (e.g., king − queen = man−woman)



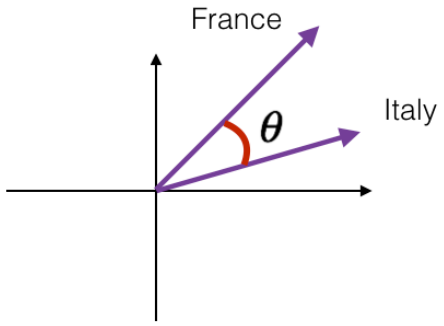


Typical embedding methods
- Word2Vec (Google)
- GloVe (Stanford)
- fastText (Facebook)

Trained in Euclidean space

Spherical Text Embedding [NeurIPS'19]

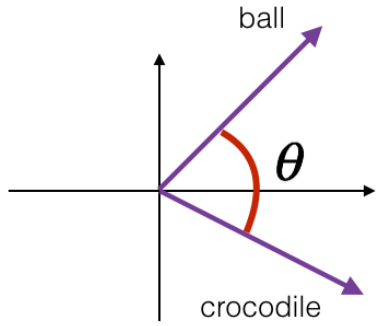
- Previous text embeddings (e.g., Word2Vec) are trained in the Euclidean space
  - But used on spherical space—Mostly directional similarity (i.e., cosine similarity)
  - Word similarity is derived using cosine similarity



France and Italy are quite similar

θ is close to 0°

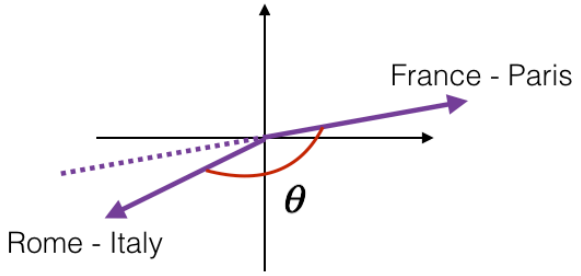
$\cos(\theta) \approx 1$



ball and crocodile are not similar

θ is close to 90°

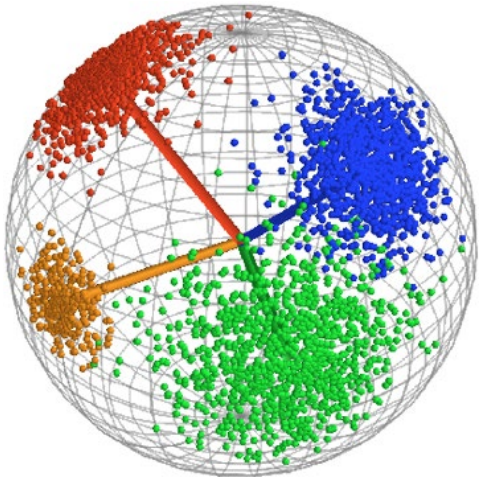
$\cos(\theta) \approx 0$



the two vectors are similar but opposite
the first one encodes (city - country)
while the second one encodes (country - city)

θ is close to 180°

$\cos(\theta) \approx -1$



  - Better document clustering performance when embeddings are normalized, and spherical clustering algorithms are used

Joint Embedding: Integrating Local and Global Contexts

- Local contexts can only partly define word semantics in unsupervised word embedding learning

Local contexts of "harmful"

If I hear someone screwing with my car (ie, setting off the **alarm**) and **taunting** me to come out, you can be very sure that my Colt Delta Elite will also be coming with me. It is not the screwing with the car that would get them **shot**, it is the potential physical **danger**. If they are **taunting** like that, it's very possible that they also intend to **rob** me and or do other physically *harmful* things. Here in Houston last year a woman heard the sound of someone …

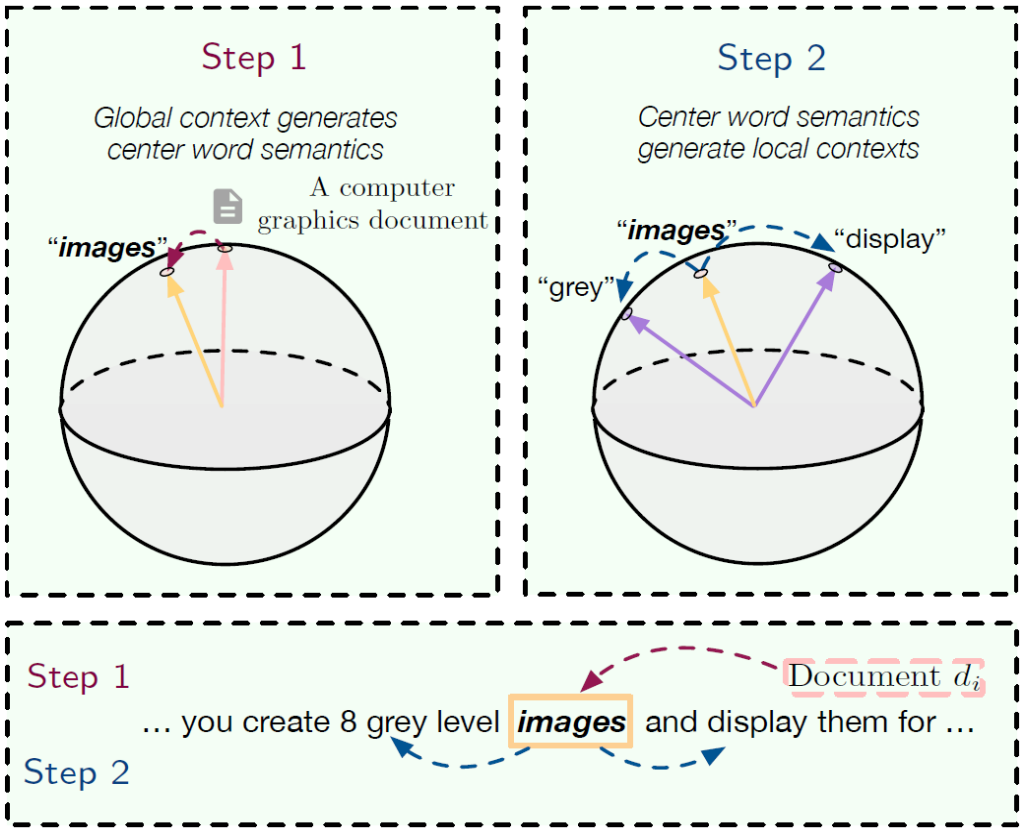
- Design a generative model on the sphere that follows how humans write articles:
  - First a general idea of the paragraph/doc, then start to write down each word in consistent with not only the paragraph/doc, but also the surrounding words

$$p(u \mid d) \propto \exp(\cos(\boldsymbol{u}, \boldsymbol{d})) \qquad p(v \mid u) \propto \exp(\cos(\boldsymbol{v}, \boldsymbol{u}))$$

Document/ Paragraph (d) → Center Word (u) → Surrounding Word (v)

Joint Spherical Embedding: Performance Comparison

Understanding the Spherical Generative Model



Word Similarity: Performance Comparison

Table 1: Spearman rank correlation on word similarity evaluation.

Embedding Space	Model	WordSim353	MEN	SimLex999
Euclidean	Word2Vec	0.711	0.726	0.311
	GloVe	0.598	0.690	0.321
	fastText	0.697	0.722	0.303
	BERT	0.477	0.594	0.287
Poincaré	Poincaré GloVe	0.623	0.652	0.321
Spherical	**JoSE**	**0.739**	**0.748**	**0.339**

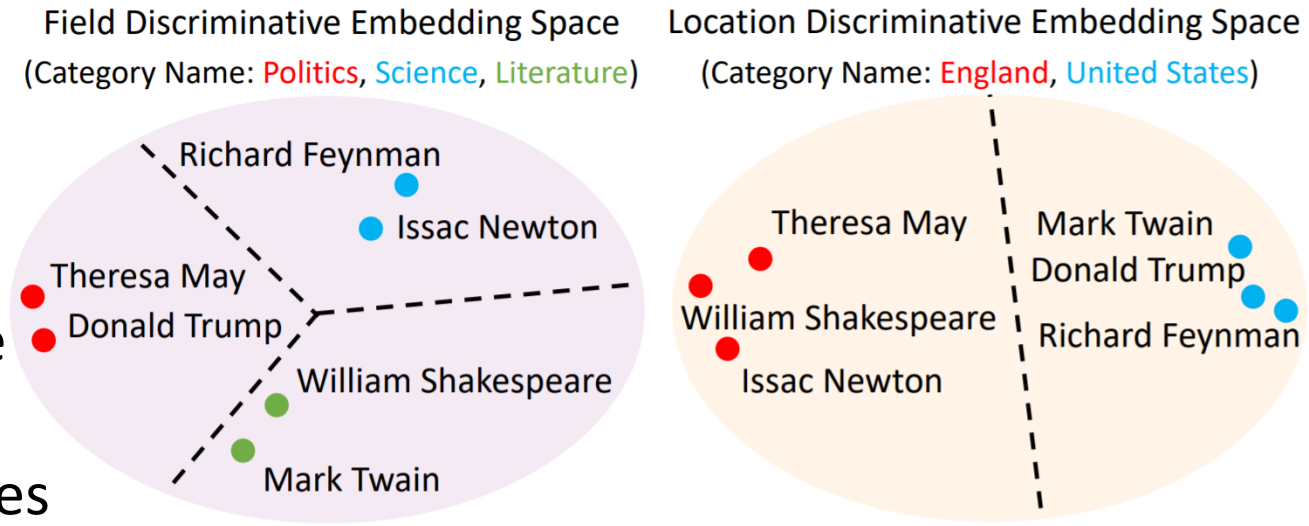
Global Context Helps Interpreting Acronyms

Table 5: Effect of Global Context on Interpreting Acronyms.

Acronyms	Global ($\lambda = \infty$)	Local ($\lambda = 0$)
CMU	**mellon**, **carnegie**, andrew, pa, pittsburgh	andrew, kfnjyea00uh, am2x, mr47, devineni
UIUC	**urbana**, **illinois**, uxa, **univ**, uchicago	uxa, ux4, ux1, mrcnext, cka52397
UNC	**chapel**, **carolina**, astro, images, usc	launchpad, gibbs, umr, lambada, jge
Caltech	**california**, gap, **institute**, keith, **technology**	juliet, jafoust, lmh, henling, bdunn
JHU	**johns**, camp, **hopkins**, nation, grand	pablo, hasch, iglesias, davidk, atlantis

Discriminative Topic Mining via Category Name-Guided Embedding

- Traditional text embedding (e.g., Word2Vec, GloVe, fastText, JoSE)
  - Mapping words with similar local contexts closer in the embedding space
  - Not imposing particular assumptions on the type of data distributions
- **CatE: Category Name-guided Embedding [WWW'20]**
  - Weak guidance: leverages *category names* to learn word embeddings with discriminative power over the specific set of categories
- CatE: Inputs
  - Category names + Corpus
- CatE: Outputs
  - The same set of celebrities are embedded differently given different sets of category names



Method of CatE: Category-name guided text Embedding

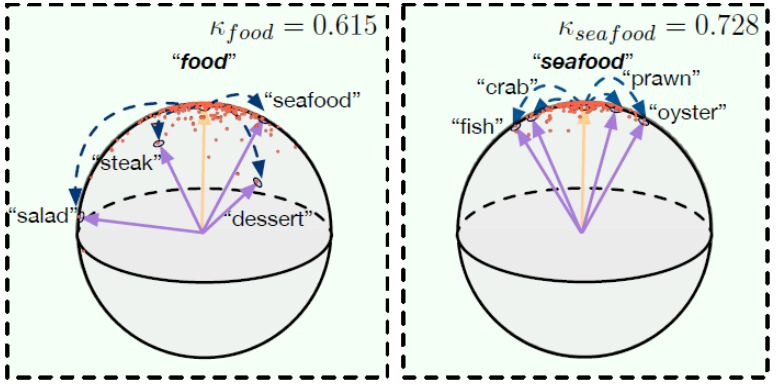
- A category-name guided text embedding learning module (E):
  - Takes a set of category names to learn category distinctive word embeddings by modeling the text generative process conditioned on the user provided categories

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

- A category representative words retrieval module (R):
  - Selects category representative words based on both word embedding similarity and word distributional specificity

The two modules (E + R) collaborate in an iterative way:

- E refines word embeddings and category embeddings
- R selects representative words that will be used by E in the next iteration



Performance Study on Discriminative Topic Mining

- Quantitative comparison
  - TC: topic coherence
  - MACC: Mean accuracy
- Qualitative Comparation of Discriminative Topic Mining

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

Methods	**NYT**-Location		**NYT**-Topic		**Yelp**-Food		**Yelp**-Sentiment	
	britain	canada	education	politics	burger	desserts	good	bad
LDA	company (×)	percent (×)	school	campaign	fatburger	ice cream	great	valet (×)
	companies (×)	economy (×)	students	clinton	dos (×)	chocolate	place (×)	peter (×)
	british	canadian	city (×)	mayor	liar (×)	gelato	love	aid (×)
	shares (×)	united states (×)	state (×)	election	cheeseburgers	tea (×)	friendly	relief (×)
	great britain	trade (×)	schools	political	bearing (×)	sweet	breakfast	rowdy
CatE	england	ontario	educational	political	burgers	dessert	delicious	sickening
	london	toronto	schools	international politics	cheeseburger	pastries	mindful	nasty
	britons	quebec	higher education	liberalism	hamburger	cheesecakes	excellent	dreadful
	scottish	montreal	secondary education	political philosophy	burger king	scones	wonderful	freaks
	great britain	ottawa	teachers	geopolitics	smash burger	ice cream	faithful	cheapskates

Text Analysis of Hong Kong Protests

Category representative phrases generated automatically

IT SHOWS RELEVANT WORDS OF DIFFERENT CATEGORIES;

category names and three examples from the experts

POLITICAL	POLICE	ECONOMIC	INFORMATION	INFRASTRUCTURE
pro democracy	tear gas	financial crisis	cbc news	hong kong university
pro beijing	hong kong police	economic downturn	cbs news	transportation
hong kong government	riot police	economic growth	fox news	international aiport
Chief executive	Water cannon	Infrastructure	Chinese state media	Mass transit railway
Mainland china	Pepper spray	Real estate	Bbc news	Lantau link
Pro establishment	Petrol bombs	Affordable housing	Global times	Flight cancellations
Mainland chinese	Hong kong government	Trade war	News media	Victoria harbour
Chief executive carrie lam	Beanbag rounds	The united states	Sina weibo	Rail operator
Carrie lam	Firing tear gas	Financial secretary	Internet censorship	Busiest airports
The chinese government	Tsuen wan	Global financial	Local media	Public transport

Text Analysis of Russia-Ukraine Conflicts

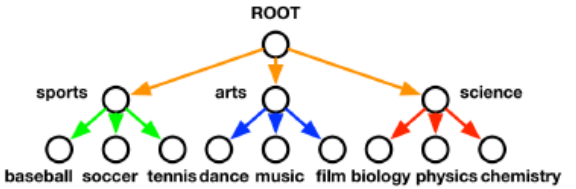
Category representative phrases generated automatically

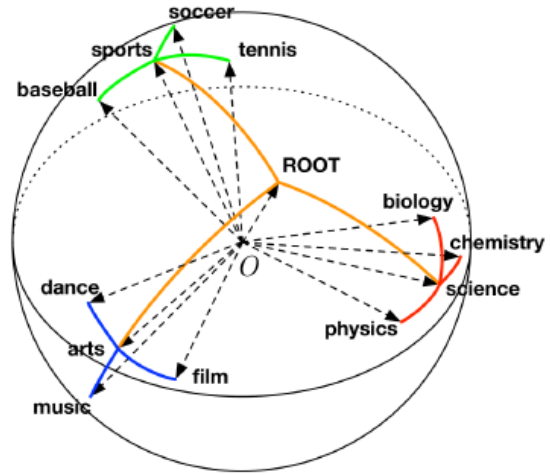
category names and three examples from the experts

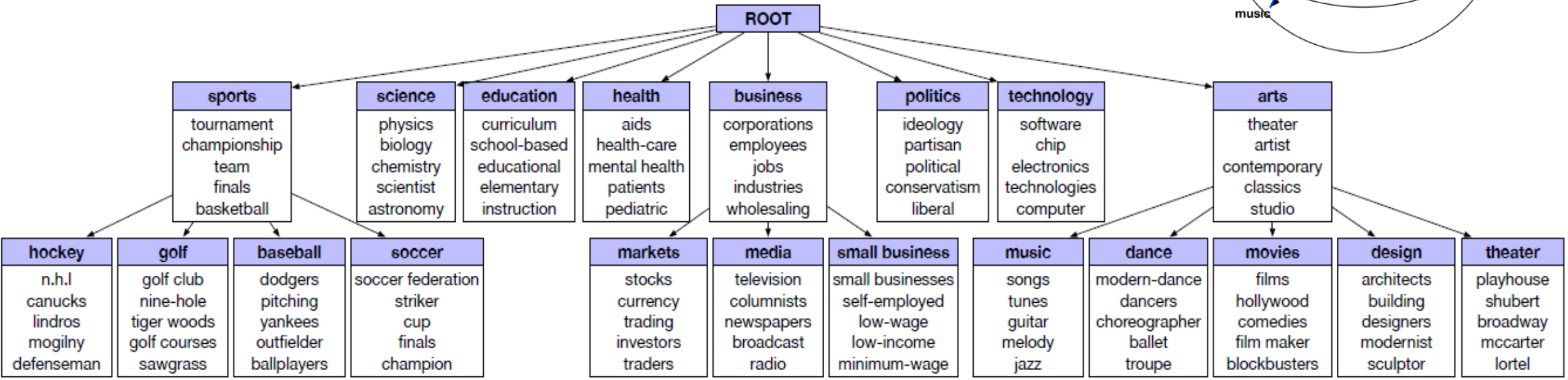
POLITICAL	MILITARY	ECONOMIC	SOCIAL	INFORMATION	CIVILIAN
Political power	Military forces	Employment	Demographic	Infowars	Urban areas
Dictator	Infantry	Economic activity	Ethnic	Information warfare	Residential area
Anarchy	Insurgents	Market	Population	Radio	Utilities
Pro government	Combatants	Finance	Language	Information security	Transportation
Neo nazi	National guard	European union	Ethnic russians	Ekho moskvy	Nuclear power plants
Viktor yanukovych	Armored vehicles	Foreign policy	Soviet union	Ukraine http empr	Power plants
Right sector	Special forces	Sergei ivanov	Western ukraine	Social media	Nuclear fuel
Pro russian	Self defense	Interior ministry	Russian language	News media	Crash site
Opposition politicians	Armored personnel	Economic sanctions	Police state	Novaya gazeta	Civil aviation
Maidan movement	Pro russian separatists	Rinat akhmetov	Anglo zionist empire	Ria novosti	Surface to air missile
Pro western	Donetsk oblast	Billion dollars	Maidan supporters	Rfe rl	Contaminated water
Kulikovo pole	Heavy fighting	Right sector	The vast majority	Mainstream media	Main entrance
Communist party	Peoples militia	Closer ties	Social media	Main stream	Emergency services
Civil war	Automatic rifles	Magnitsky act	Martial law	Intelligence community	Drinking water

Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding [KDD'20]

- JoSH: A joint tree and text embedding method
- Simultaneous modeling of the category tree structure in the spherical space
- Effective mining of category representative, hierarchical terms
  - Ex. In PubMed literature, finding distinct terms related to *hormones, enzymes, vitamins*, and *vaccines*

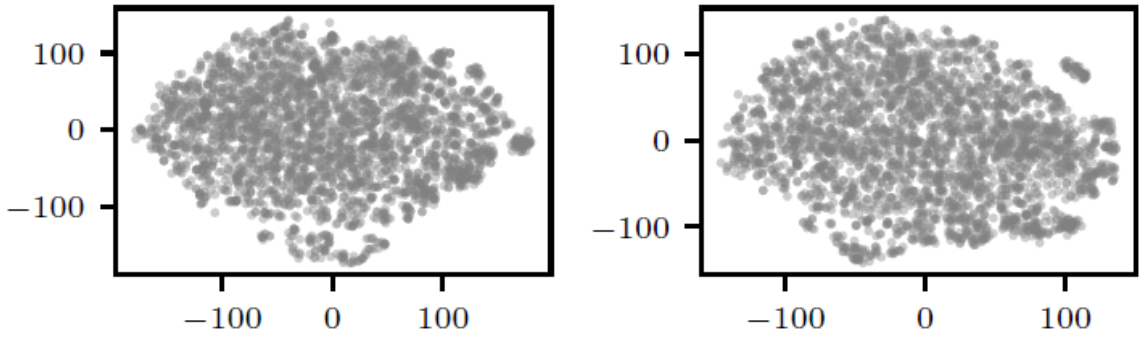






Topic Discovery via Latent Space Clustering of LM Embedding

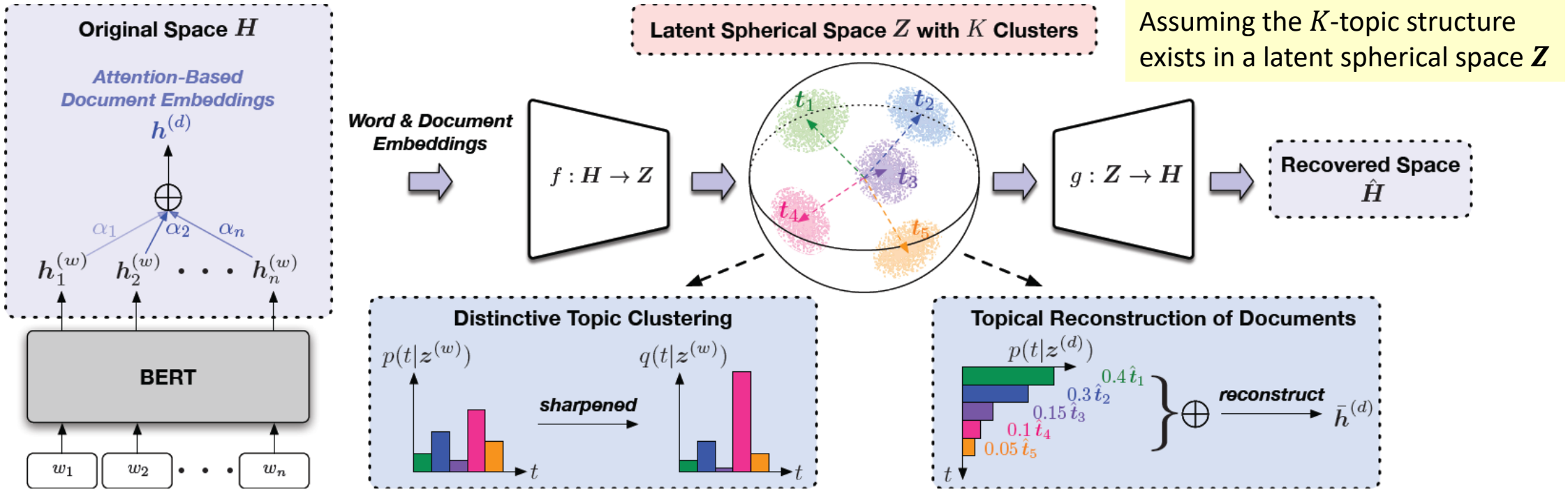
- Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang and Jiawei Han, "Topic Discovery via Latent Space Clustering of Language Model Embeddings", in WWW'22
- Task: Automatic discovery of coherent and meaningful topics from text corpora
- Limitations of topic modeling (a generative process)
  - Ignoring word ordering information in text (based on the "bag-of-words" assumption)
  - cannot leverage external knowledge to learn word semantics, and
  - Inducing an intractable posterior that requires approximation algorithms
- Why not directly deploy pre-trained language models (PLMs) for topic discovery?
  - The PLM embedding space is partitioned into extremely fine-grained clusters and lacks topic structures inherently
  - PLM embeddings are high-dimensional while distance functions can become meaningless
  - Lack of good document representations from PLMs



(a) New York Times. (b) Yelp Review.

Visualization of 3, 000 randomly sampled contextualized word embeddings of BERT: The embedding spaces do not have clearly separated clusters.

TopClus: Topic Discovery via Latent Space Clustering



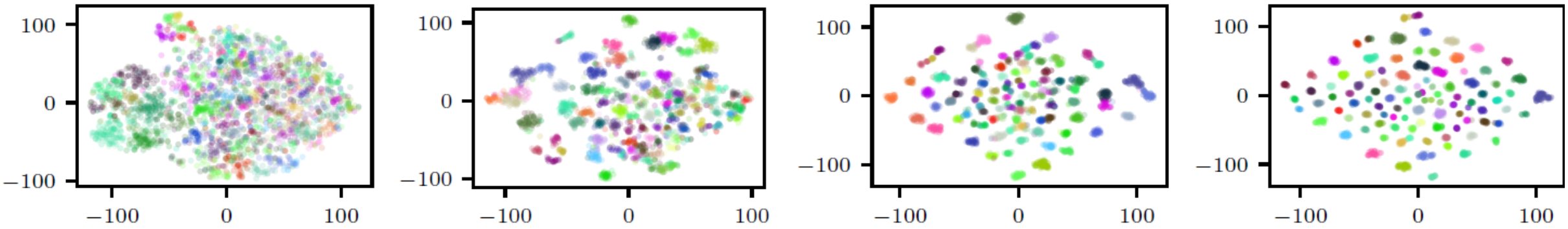
- Jointly learn the attention weights for document embeddings and the latent space generation model via three objectives
  - (1) a clustering loss that encourages distinctive topic learning in the latent space
  - (2) a topical reconstruction loss of documents that promotes meaningful topic representations for summarizing document semantics, and
  - (3) an embedding space preserving loss that maintains the semantics of the original embedding space

Qualitative Evaluation of Topic Discovery

Corpus	# documents	# words/doc.	Vocabulary
NYT	31,997	690	25,903
Yelp	29,280	114	11,419

Methods	NYT					Yelp				
	Topic 1 (sports)	Topic 2 (politics)	Topic 3 (research)	Topic 4 (france)	Topic 5 (japan)	Topic 1 (positive)	Topic 2 (negative)	Topic 3 (vegetables)	Topic 4 (fruits)	Topic 5 (seafood)
LDA	olympic	*mr*	*said*	french	japanese	amazing	loud	spinach	mango	fish
	year	bush	report	*union*	tokyo	*really*	awful	carrots	strawberry	*roll*
	said	president	evidence	*germany*	*year*	*place*	*sunday*	greens	*vanilla*	salmon
	games	white	findings	*workers*	matsui	phenomenal	*like*	salad	banana	*fresh*
	team	house	defense	paris	*said*	pleasant	slow	*dressing*	*peanut*	*good*
CorEx	baseball	house	possibility	french	japanese	great	*even*	garlic	strawberry	shrimp
	championship	white	challenge	*italy*	tokyo	friendly	bad	tomato	*caramel*	*beef*
	playing	support	reasons	paris	*index*	*atmosphere*	mean	onions	*sugar*	crab
	fans	*groups*	*give*	francs	osaka	love	cold	*toppings*	fruit	*dishes*
	league	*member*	planned	jacques	*electronics*	favorite	*literally*	*slices*	mango	*salt*
ETM	olympic	government	approach	french	japanese	nice	disappointed	avocado	strawberry	fish
	league	national	problems	*students*	*agreement*	worth	cold	*greek*	mango	shrimp
	national	*plan*	experts	paris	tokyo	*lunch*	*review*	salads	*sweet*	lobster
	basketball	public	*move*	*german*	*market*	recommend	*experience*	spinach	*soft*	crab
	athletes	support	*give*	*american*	*european*	friendly	bad	tomatoes	*flavors*	*chips*
BERTopic	swimming	bush	researchers	french	japanese	awesome	horrible	tomatoes	strawberry	lobster
	freestyle	democrats	scientists	paris	tokyo	*atmosphere*	*quality*	avocado	mango	crab
	popov	white	cases	lyon	ufj	friendly	disgusting	*soups*	*cup*	shrimp
	gold	bushs	*genetic*	*minister*	*company*	*night*	disappointing	kale	lemon	oysters
	olympic	house	study	*billion*	yen	good	*place*	cauliflower	banana	*amazing*
TopClus	athletes	government	hypothesis	french	japanese	good	tough	potatoes	strawberry	fish
	medalist	ministry	methodology	seine	tokyo	best	bad	onions	lemon	octopus
	olympics	bureaucracy	possibility	toulouse	osaka	friendly	painful	tomatoes	apples	shrimp
	tournaments	politicians	criteria	marseille	hokkaido	cozy	frustrating	cabbage	grape	lobster
	quarterfinal	electoral	assumptions	paris	yokohama	casual	brutal	mushrooms	peach	crab

TopClus: Performance Study

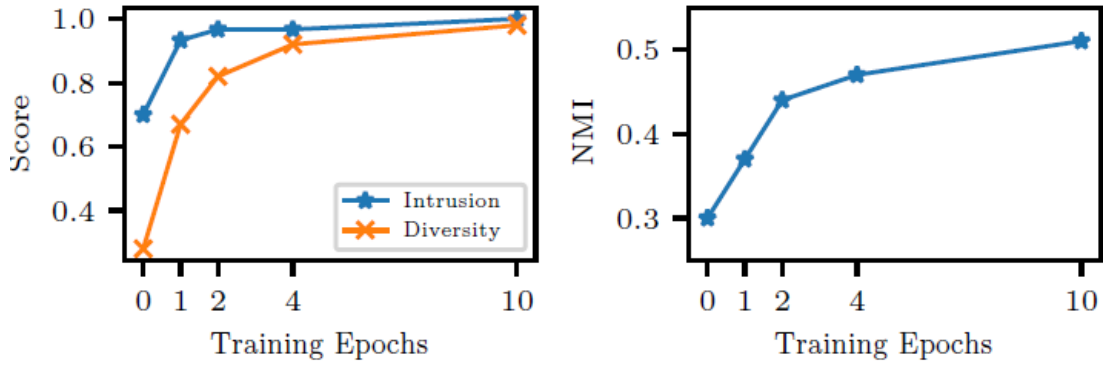


(a) Epoch 0. (b) Epoch 2. (c) Epoch 4. (d) Epoch 8.

Visualization using 3, 000 randomly sampled latent word embeddings during training. Embeddings assigned to the same cluster are in the same color. The latent space gradually exhibits distinctive and balanced cluster structure.

Methods	NYT		Yelp	
	Topic	Location	Food	Sentiment
LDA	0.39	0.20	0.09	0.07
CorEx	0.29	0.20	0.10	0.03
ETM	0.41	0.21	0.13	0.01
BERTopic	0.26	0.22	0.13	0.00
TopClus	**0.51**	**0.25**	**0.16**	**0.41**

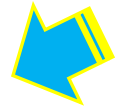
Document clustering NMI scores on two datasets under four sets of labels



(a) Topic Quality. (b) Document Clustering.

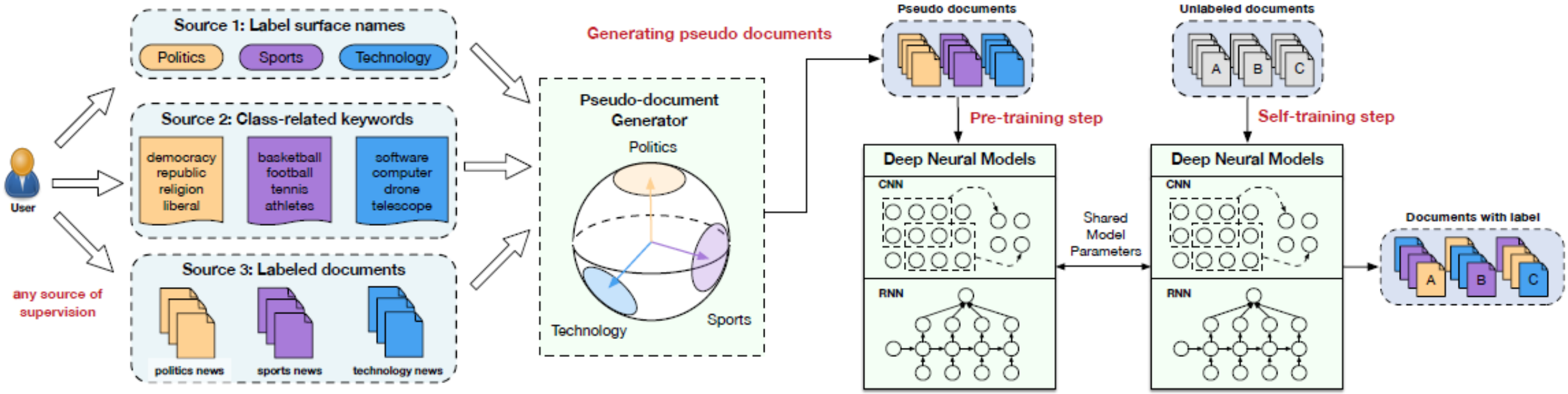
TopClus training on NYT: (a) topic coherence measured by intrusion test and topic diversity; (b) document clustering NMI scores over training

Outline

- What Kinds of Knowledge Graphs Do We Really Need?
- Key Issue: Construction of Theme-/Corpus-Based Knowledge Graphs
- The Role of Embedding and PLM in Knowledge Graph Construction
- Data Preparation: Taxonomy-Guided Text Classification 
- Identifying Knowledge Graph Primitives: Entities, Properties and Relations
- Conclusion: Towards Theme/Corpus-Based Knowledge Graph Construction

WeSTClass: Weakly Supervised Text Classification

- Modeling class distribution in word2vec embedding space
  - Word2vec embedding captures **skip-gram (local) similarity** (i.e., words with similar local context windows are expected to have similar meanings)



WeSTClass (Weakly Supervised Text Classification): CIKM'18
WeSHClass (Weakly Supervised Hierarchical Text Classification): AAAI'19

LOTClass: Label-Name-Only Text Classification [EMNLP'20]

- Yu Meng, et al., "Text Classification Using Label Names Only: A Language Model Self-Training Approach" [EMNLP'20]
- Inputs: A set of label names representing each class + unlabeled documents
- Method (3 steps): Make good use of pre-trained language model (e.g., BERT)
  - Step 1. Category understanding via label name replacement (learn *topic vocabulary*)
    - Ex. "sports" → {"soccer", "basketball", …} (use pretrained LM to replace category name)

- Learn topic vocabulary using label name only
- Make good use of pretrained LM (e.g., BERT)
- Result from AGNews dataset

Label Name	Category Vocabulary
politics	politics, political, politicians, government, elections, politician, democracy, democratic, governing, party, leadership, state, election, politically, affairs, issues, governments, voters, debate, cabinet, congress, democrat, president, religion, . . .
sports	sports, games, sporting, game, athletics, national, athletic, espn, soccer, basketball, stadium, arts, racing, baseball, tv, hockey, pro, press, team, red, home, bay, kings, city, legends, winning, miracle, olympic, ball, giants, players, champions, boxing, . . .
business	business, trade, commercial, enterprise, shop, money, market, commerce, corporate, global, future, sales, general, international, group, retail, management, companies, operations, operation, store, corporation, venture, economic, division, firm, . . .
technology	technology, tech, software, technological, device, equipment, hardware, devices, infrastructure, system, knowledge, technique, digital, technical, concept, systems, gear, techniques, functionality, process, material, facility, feature, method, . . .

LOTClass: Label-Name-Only Text Classification

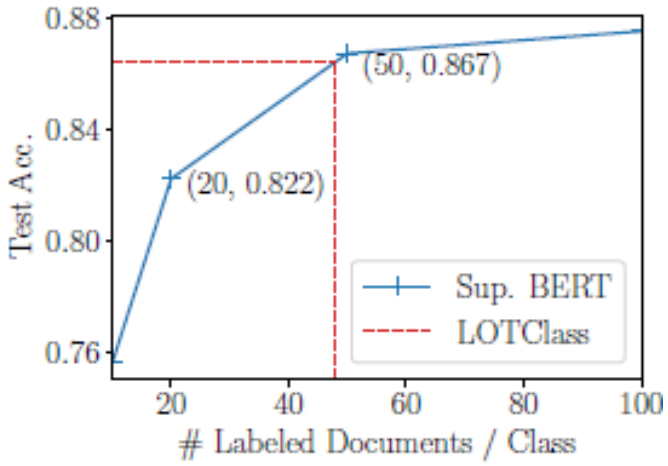
- Step 2: Masked topic prediction: Create contextualized word-level supervisions to train the model for predicting a word's implied topic

Different contexts leads to different BERT language model prediction

Sentence	Language Model Prediction
The oldest annual US team **sports** competition that includes professionals is not in baseball, or football or basketball or hockey. It's in soccer.	sports, baseball, handball, soccer, basketball, football, tennis, sport, championship, hockey, . . .
Samsung's new SPH-V5400 mobile phone **sports** a built-in 1-inch, 1.5-gigabyte hard disk that can store about 15 times more data than conventional handsets, Samsung said.	has, with, features, uses, includes, had, is, contains, featured, have, incorporates, requires, offers, . . .

- Step 3: Self-training: Generalize the model via self-training on abundant unlabeled data to make document-level topic prediction

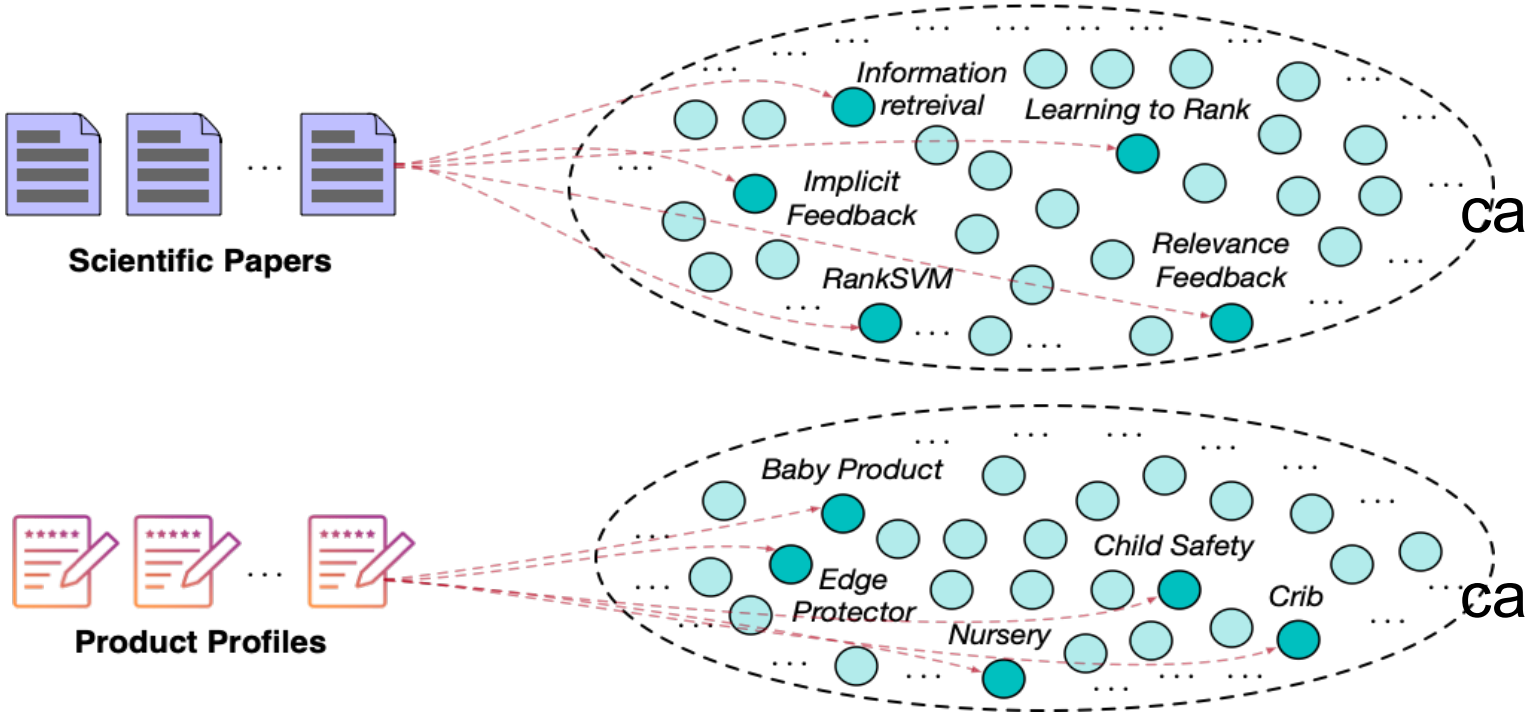
Supervision Type	Methods	AG News	DBPedia	IMDB	Amazon
Weakly-Sup.	**Dataless** (Chang et al., 2008)	0.696	0.634	0.505	0.501
	WeSTClass (Meng et al., 2018)	0.823	0.811	0.774	0.753
	BERT w. simple match	0.752	0.722	0.677	0.654
	LOTClass w/o. self train	0.822	0.860	0.802	0.853
	LOTClass	**0.864**	**0.911**	**0.865**	**0.916**
Semi-Sup.	**UDA** (Xie et al., 2019)	0.869	0.986	0.887	0.960
Supervised	**char-CNN** (Zhang et al., 2015)	0.872	0.983	0.853	0.945
	BERT (Devlin et al., 2019)	0.944	0.993	0.945	0.972



Label-name only is equiv. to 48 labels in Supervised BERT

Need: "Structuring"/Tagging Unstructured Documents

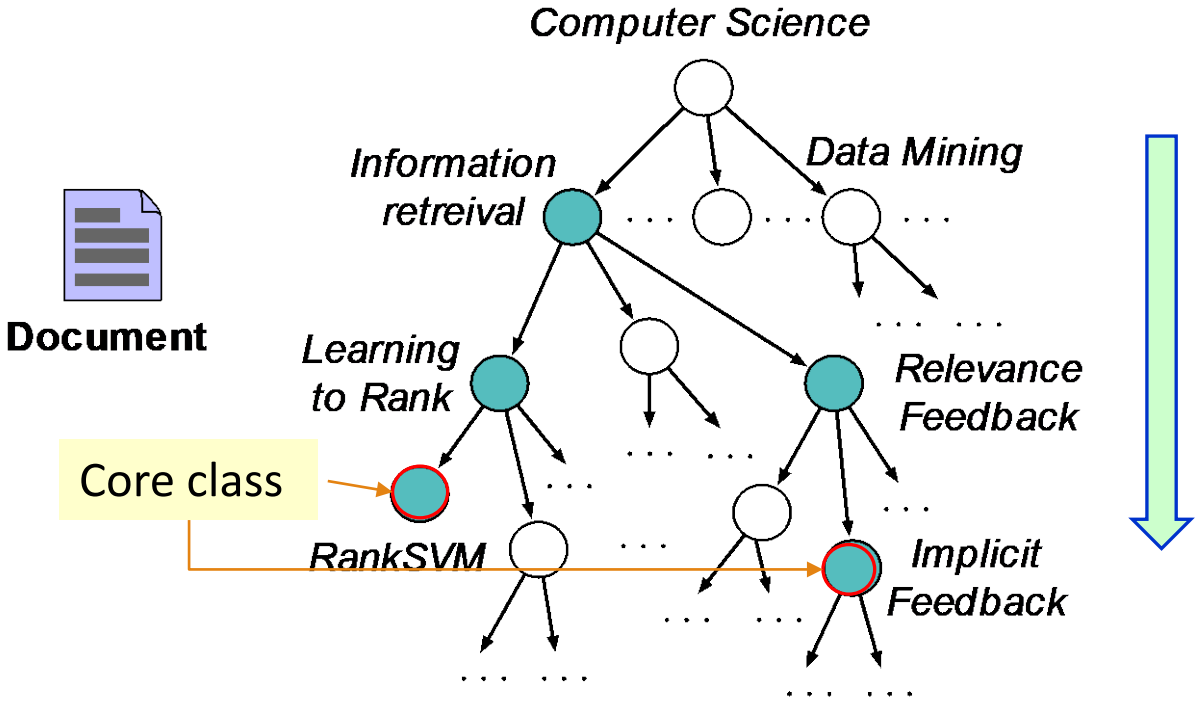
- Task: Tag each doc. with **a set of relevant classes** from **a huge candidate pool**

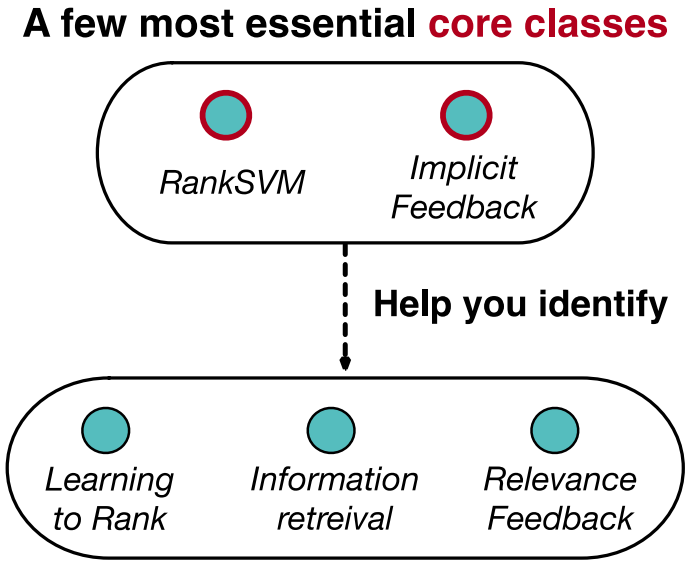


- Challenges:
  - Huge label space, multi-label tagging
  - Limited labeled data— hard for supervised models

TaxoClass [NAACL'21]: Taxonomy Comes to Rescue

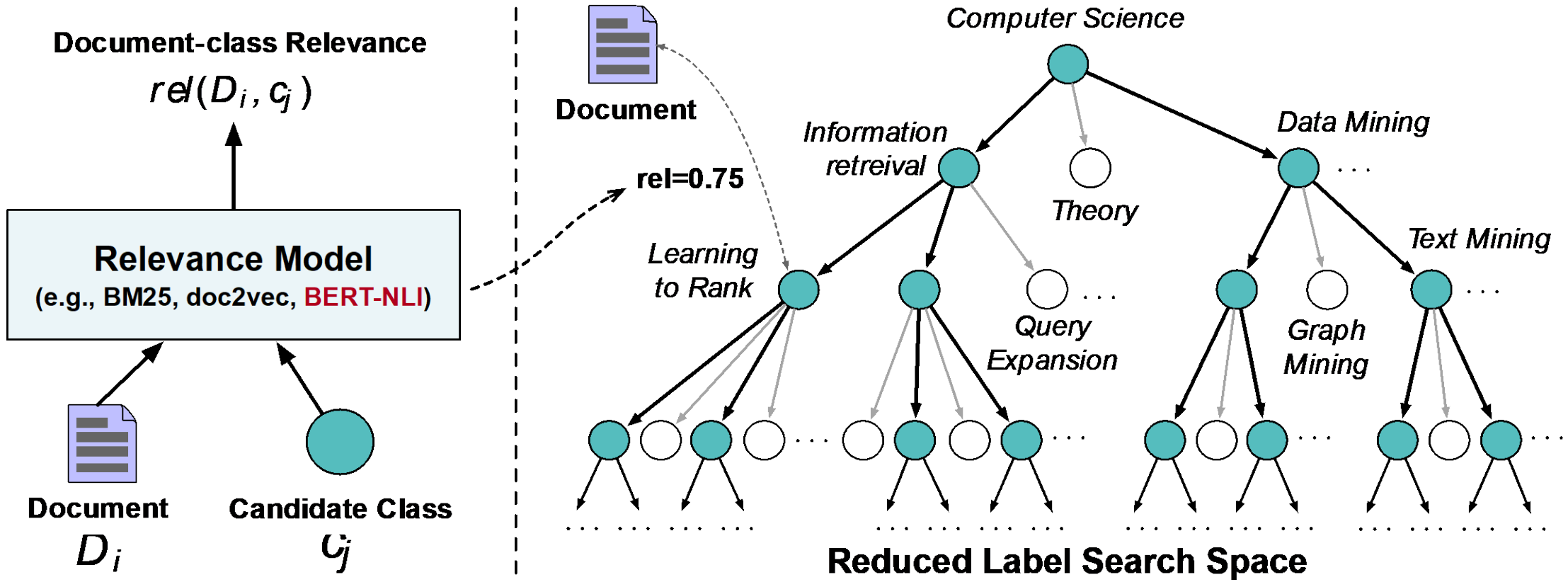
- J. Shen, et al. "TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names", NAACL'21
- Taxonomy!— Structure the huge label space by organizing classes hierarchically
  - Enable fast label space exploration in a top-down way
- Facilitate multi-label tagging by capturing class relations





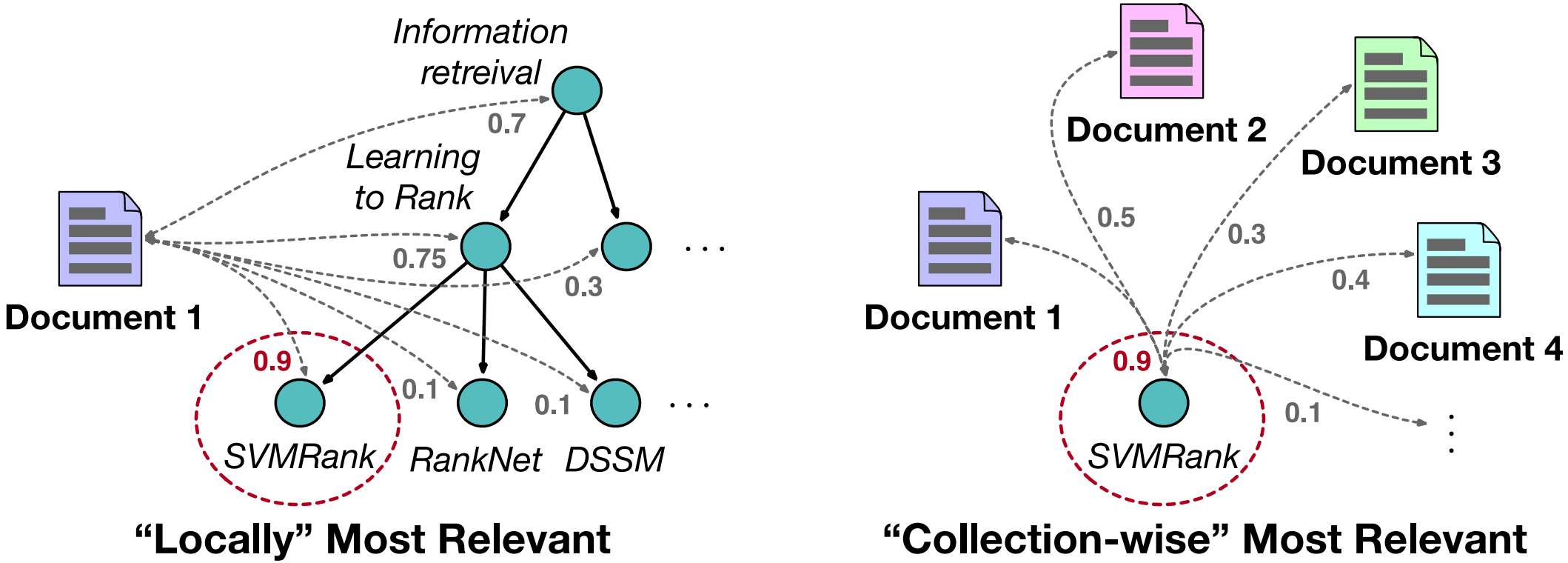
TaxoClass: A Weakly-Supervised Classification Method based on Taxonomy

- **Shrink the label search space** with top-down exploration
  - Use a **relevance model** to filter out completely irrelevant classes for each document



TaxoClass: A Weakly-Supervised Classification Method based on Taxonomy

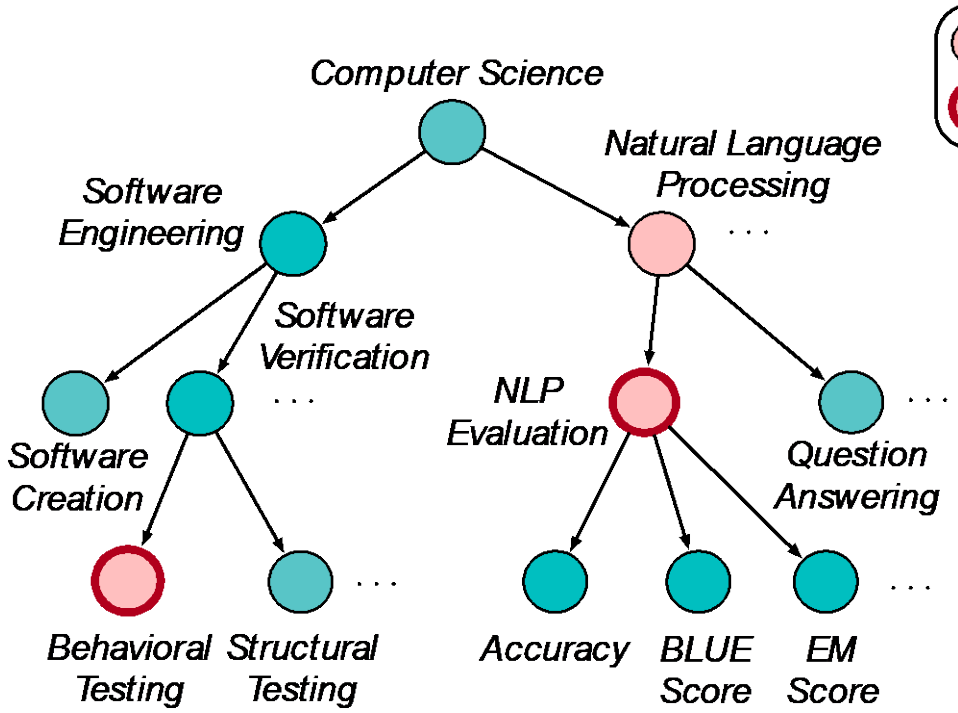
- **Shrink the label search space** with top-down exploration
- **Identify document core classes** in reduced label search space
- **Generalize from core classes** with bootstrapping and self-training

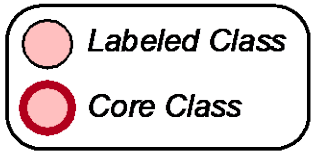


TaxoClass: Case Studies

dblp

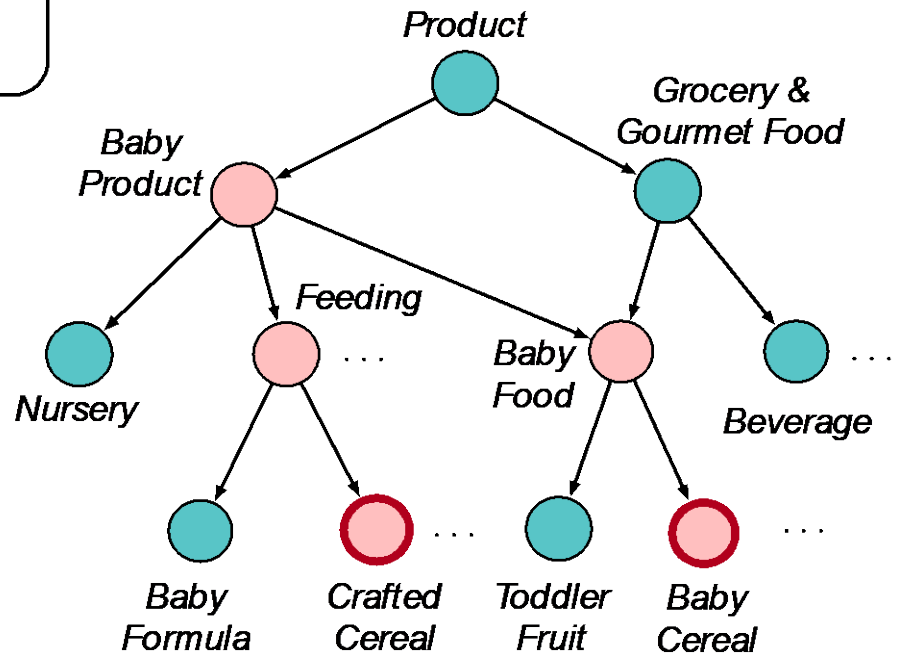
Document
Inspired by principles of **behavioral testing** in software engineering, we introduce CheckList, a task-agnostic methodology for **testing NLP models**…





amazon

Document
When our **son** was about **4 months old**, doctor said we could give him **crafted cereal** so we bought it. It digests well and doesn't lock up his bowels at all ...



TaxoClass: Performance Comparison

	Methods	Amazon		DBPedia	
		Example-F1	P@1	Example-F1	P@1
Weakly-supervised multi-class classification method	**WeSHClass** (Meng et al., AAAI'19)	0.246	0.577	0.305	0.536
Semi-supervised methods using 30% of training set	**SS-PCEM** (Xiao et al., WebConf'19)	0.292	0.537	0.385	0.742
	Semi-BERT (Devlin et al., NAACL'19)	0.339	0.592	0.428	0.761
Zero-shot method	**Hier-0Shot-TC** (Yin et al., EMNLP'19)	0.474	0.714	0.677	0.787
	TaxoClass (NAACL'21)	0.593	0.812	0.816	0.894

Example-F1 $= \frac{1}{N}\sum_{i=1}^{N} \frac{2|true_i \cap pred_i|}{|true_i|+|pred_i|}$, P@1 $= \frac{\#docs\ with\ top-1\ pred\ dorrect}{\#total\ docs}$

- **vs. WeSHClass**: better model document-class relevance
- **vs. SS-PCEM, Semi-BERT**: better leverage supervision signals from taxonomy
- **vs. Hier-0Shot-TC**: better capture domain-specific information from core classes

Amazon: 49K product reviews (29.5K training + 19.7K testing), 531 classes
DBPedia: 245K Wiki articles (196K training + 49K testing), 298 classes

Outline

- What Kinds of Knowledge Graphs Do We Really Need?
- Key Issue: Construction of Theme-/Corpus-Based Knowledge Graphs
- The Role of Embedding and PLM in Knowledge Graph Construction
- Data Preparation: Taxonomy-Guided Text Classification
- Identifying Knowledge Graph Primitives: Entities, Properties and Relations
- Conclusion: Towards Theme/Corpus-Based Knowledge Graph Construction

Automatic Extraction of Knowledge Graph Primitives

- Automatic extraction of knowledge graph primitives: Entities, properties and relations
- X. Wang et al., "ChemNER: Fine-Grained Chemistry Named Entity Recognition with Ontology-Guided Distant Supervision", EMNLP'21
- ChemNER: Fine-grained NER for scientific literature
  - Assign the **most accurate fine-grained type** to each mention under certain context with **ontology-guided multi-type disambiguation**
- Address the **multi-type annotation** problem
  - Example text: "Although it was necessary to employ a stoichiometric quantity of [**palladium**]$_{\text{CATALYST, TRANSITION METAL}}$, it is … "
    - **> 60% chemical entities:** can be matched to more than one entity type in the *knowledge bases* provided by experts (e.g., Chem DB, MESH ontology)
    - On average, each chemical entity can be matched to **~4 types** in the knowledge bases

The ChemNER Framework

Input Corpus

Entity Span Detection

S1: **[Methyl-14C]S-dThd** was synthesized by rapid **methylation** of ...
S2: ... **Suzuki-Miyaura cross-coupling reactions** were carried out ...
S3: Although it was necessary to employ a stoichiometric quantity of **palladium** , it is noteworthy that the **cross-coupling** proceeded in the presence of a wide array of **functional groups**.
S4: ... can undergo a **transmetalation** with either **BBA** or the rapidly forming **boronic acid** ...

Flexible KB-Matching ← ***Knowledge Bases***

S1: **[Methyl-14C]S-dThd** was synthesized by rapid **methylation** of ...

ORGANIC COMPOUNDS, ORGANIC POLYMERS **ORGANIC REACTIONS**

TF-IDF Scores	**ORGANIC COMPOUNDS**	**ORGANIC POLYMERS**	**Biomolecules**	...
methyl	**0.0177**	**0.0139**	0.0010	...
thd	**0.0256**	**0.0115**	**0.0417**	

S2: ..., **Suzuki-Miyaura cross-coupling reactions** were carried out ...

COUPLING REACTIONS

Ontology-guided Multi-type Disambiguation

CATALYSTS, ~~TRANSITION METALS~~

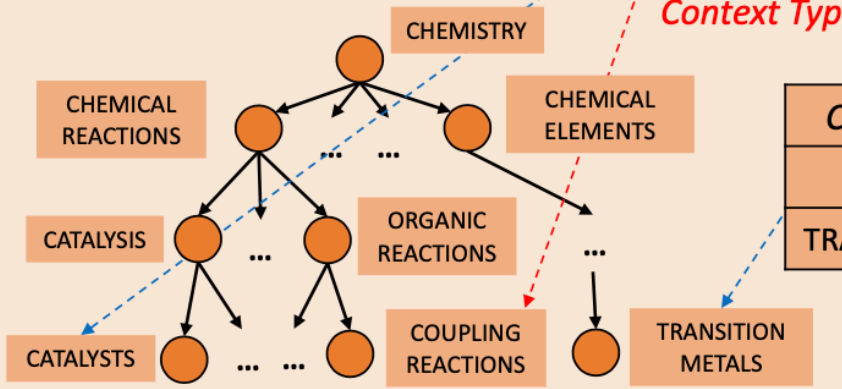
S3: Although it was necessary to employ a stoichiometric quantity of **palladium** , it is noteworthy that the **cross-coupling** proceeded in the presence of a wide array of **functional groups**.

COUPLING REACTIONS

FUNCTIONAL GROUPS

Candidate Types

Context Type



Candidate Types	Freq.
CATALYSTS	**18,707**
TRANSITION_METALS	9,618

Sequence Labeling Models

BiLSMT-CRF, RoBERTa, ChemBERTa, ...

ORGANOMETALLIC CHEMISTRY **??? [NOT IN KB] => OXOACIDS**

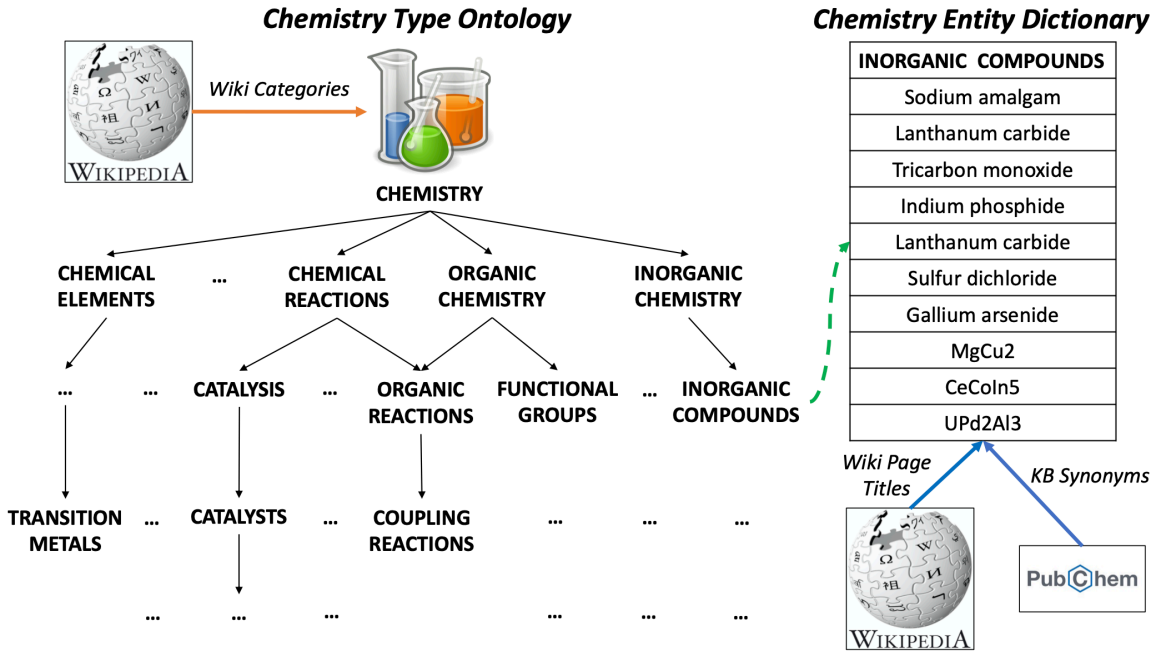
S4: ... can undergo a **transmetalation** with ***either*** **BBA** ***or*** the rapidly forming **boronic acid** ...

OXOACIDS

"either ... or ..." pattern learned by Sequence Labeling Model

Chemistry Ontology

- Fine-grained chemistry type ontology:
  - **Wikipedia categories** rooted under *Chemistry*
  - Categories => Entity Types
  - Associated Page Titles => Entity Dictionaries
  - **Expert proved 62 fine-grained types**



Category:Chemistry

From Wikipedia, the free encyclopedia

Subcategories

This category has the following 73 subcategories, out of 73 total.

.
- ▶ Chemists (12 C, 3 P)

*
- ▶ Chemistry set index pages (1 C, 655 P)

+
- ▶ Chemical elements (132 C, 127 P)

A
- ▶ Acid–base chemistry (5 C, 49 P)
- ▶ Analytical chemistry (19 C, 222 P)
- ▶ Astrochemistry (1 C, 38 P)
- ▶ Atmospheric chemistry (24 P)
- ▶ Chemistry literature (3 C, 2 P)

M
- ▶ Materials science (35 C, 400 P)
- ▶ Medicinal chemistry (8 C, 77 P, 10 F)
- ▶ Metallurgy (14 C, 161 P)
- ▶ Microwave chemistry (4 P)
- ▶ Chemical mixtures (6 C, 44 P)
- ▶ Molecular physics (10 C, 79 P)
- ▶ Molecules (10 C, 20 P)

N
- ▶ Chemical nomenclature (4 C, 84 P)
- ▶ Nuclear chemistry (8 C, 59 P)

Pages in category "Chemistry"

The following 132 pages are in this category, out of 132 total. This list may not reflect recent changes (lea more).

- Chemistry

*
- Portal:Chemistry

0–9
- 2-Hexoxyethanol

A
- Acid–base reaction
- Actinide chemistry
- Allotropy
- Alloy
- Fluorine cycle
- Forensic chemistry
- Free element

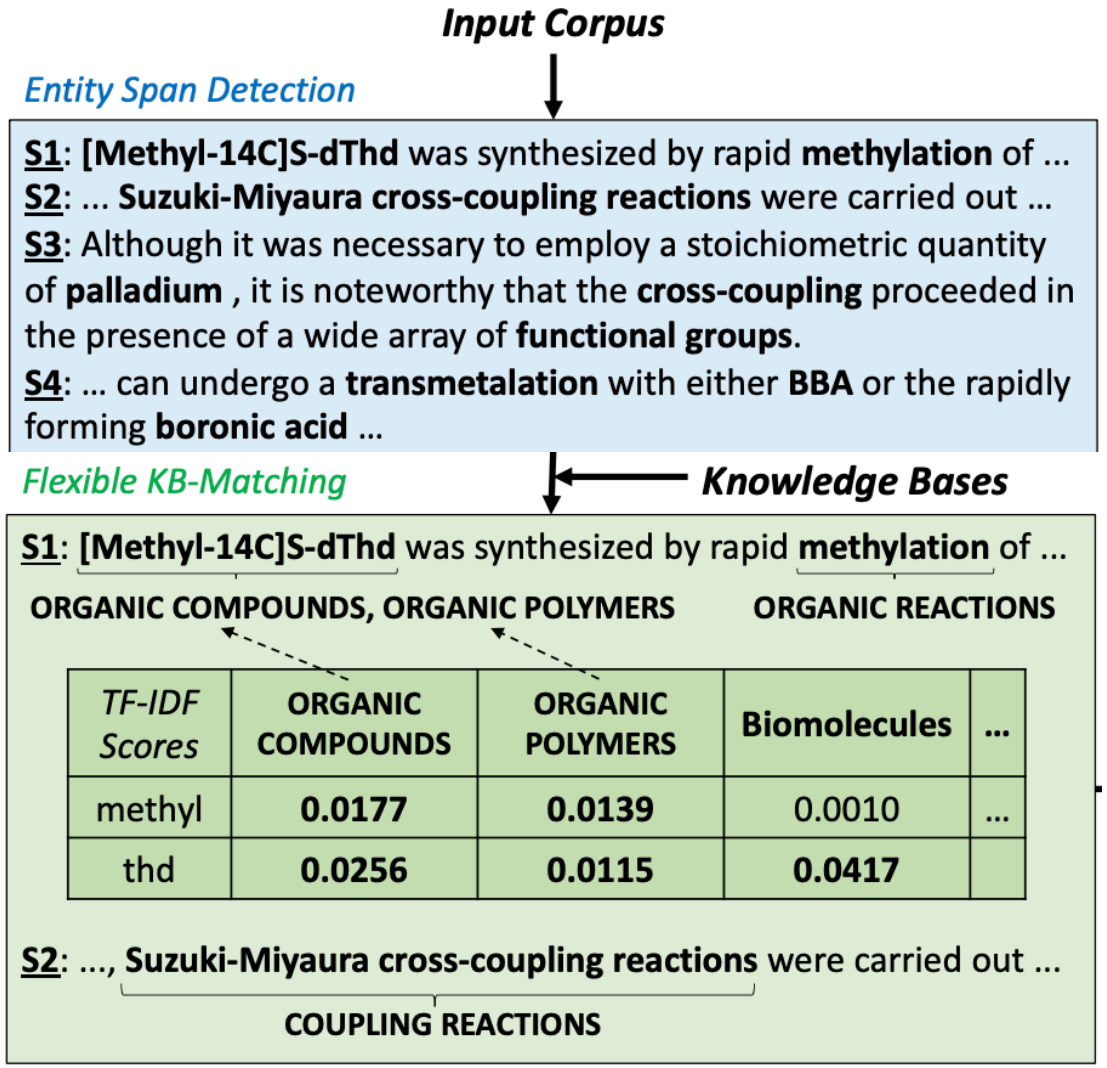
G
- Geometry index
- Glossary of chemistry terms
- Gold cycle
- Green chemistry

H
- Harbi al-Himyari

I

Entity Span Detection & Flexible KB Matching

- Entity Span Detection
  - Chemical phrase chunking
  - **ChemDataExtractor** (Swain and Cole, 2016) and **GeniaTagger** (Tsuruoka and Tsujii, 2005)
- Flexible KB Matching
  - **TF-IDF-based majority voting**
  - Match long and complex chemistry entities (e.g., chemical compounds) that does not exist in the KBs



Ontology-Guided Multi-Type Disambiguation

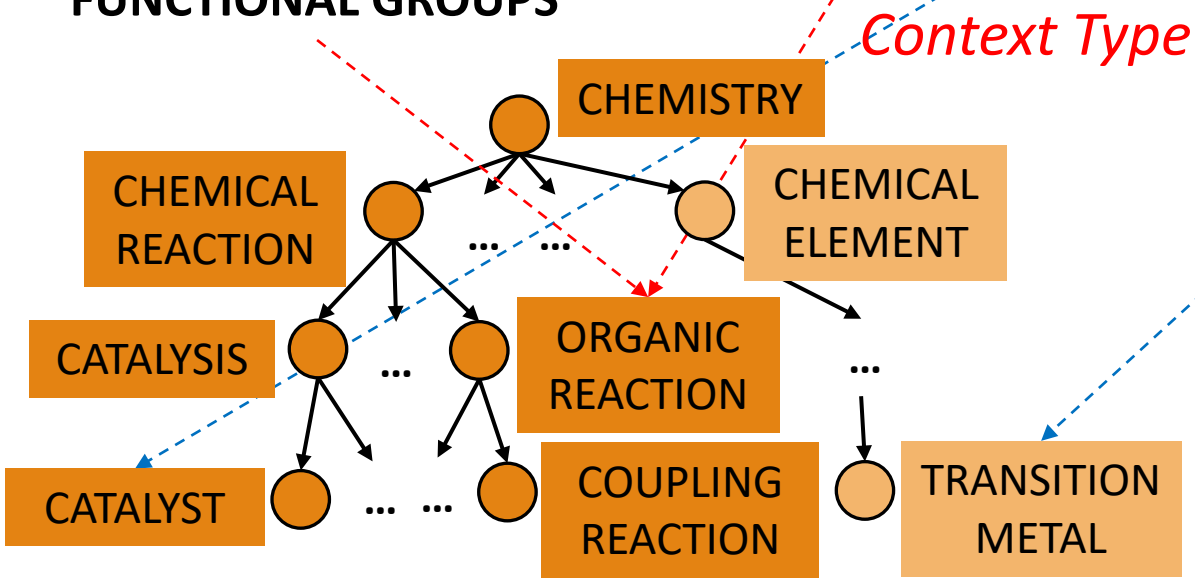
- Key idea: the entities in the same sentence, paragraph or document usually **follow a focused topic.**

CATALYST, ~~TRANSITION METAL~~

Although it was necessary to employ a stoichiometric quantity of **palladium** , it is noteworthy that the **cross-coupling** proceeded in the presence of a wide array of **functional groups**.

COUPLING REACTIONS

FUNCTIONAL GROUPS



Ontology-Guided Multi-Type Disambiguation

- The disambiguation score:

How close are the candidate and context types on the ontology

$$S_d(t_{e_i}^j) = \frac{\sum_{k\in[1,\dots,n],k\neq i,|T_{e_k}|=1} dep(lca(t_{e_i}^j, t_{e_k}))}{n * dep(t_{e_i}^j)}$$

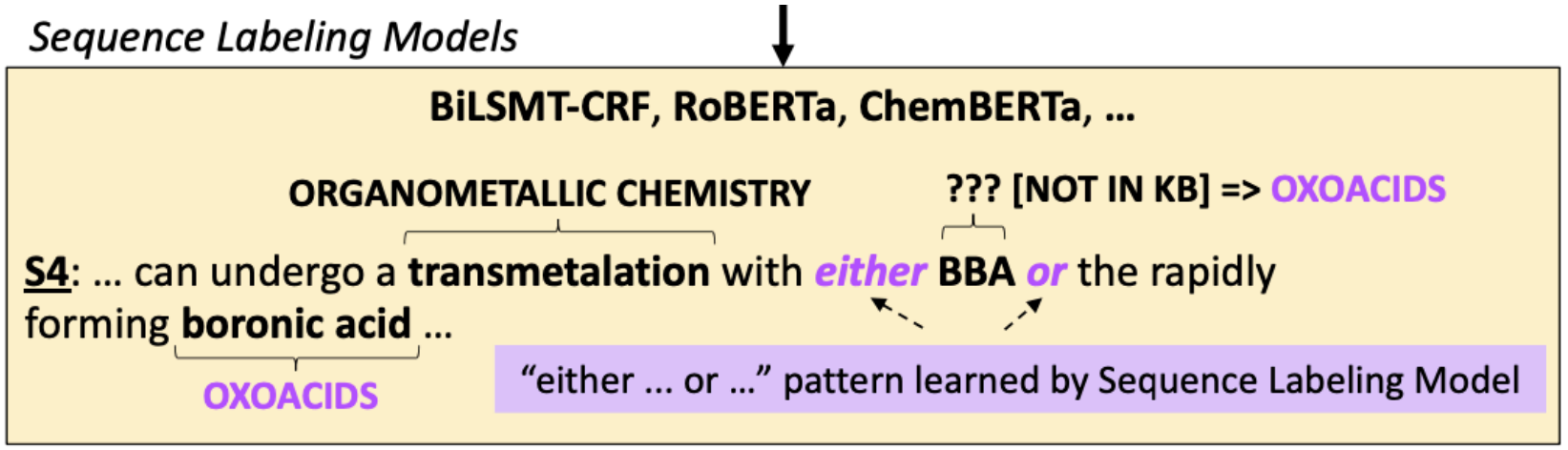
How fine-grained is the candidate type on the ontology

- $lca(.,.)$ - the lowest common ancestor of two nodes on the type ontology
- $dep(.)$ - the depth of the type node on the type ontology

A **larger S_d score** → Candidate type is more likely to be **correct**

Exploring PLM: Sequence Labeling Model

- The flexible KB-matching and multi-type disambiguation cannot cover all the new entities in the corpus
- Train a sequence labeling model to further improve the recall
  - **RoBERTa** (Liu *et al*., 2019), **ChemBERTa** (Chithrananda *et al*., 2020)



Chem NER: Performance Comparison

- **Dataset**:
  - Training: **85,702** unlabeled sentences + **62** fine-grained chemistry types
  - Test: **3,000** expert-annotated sentences

	Method	Precision	Recall	F1 Score
	KB-Matching	0.21	0.12	0.15
Supervised NER	BiLSTM-CRF (2016)	0.22	0.10	0.14
Supervised NER	RoBERTa (2019)	0.24	0.18	0.20
Supervised NER	ChemBERTa (2020)	0.18	0.12	0.14
Distant NER	AutoNER (2018)	0.21	0.04	0.06
Distant NER	BOND (2020)	0.19	0.13	0.15
	ChemNER (2021)	**0.69**	**0.34**	**0.46**

$$Precision\ (P) = \frac{\#Truth\ Positive}{\#Prediction}$$

$$Recall(R) = \frac{\#True\ Positive}{\#Ground - Truth}$$

$$F1\ Score = \frac{2 \times P \times R}{P + R}$$

+0.26 absolution F1 ↑

Recently Published Efforts Related To KG Construction

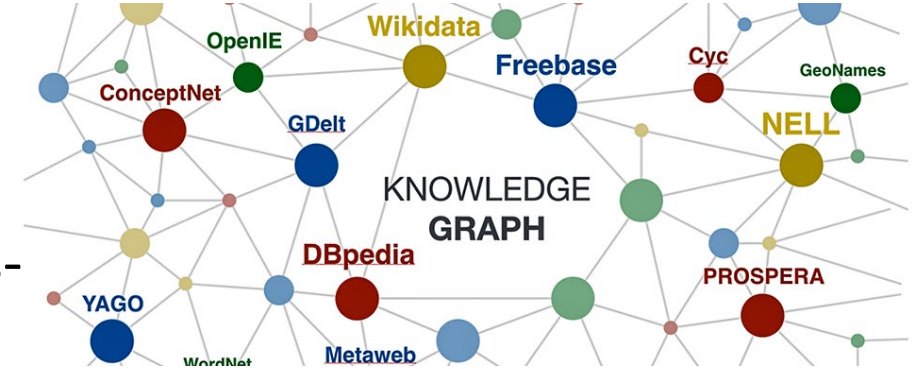
- Jiaxin Huang, Yu Meng, and Jiawei Han, "Few-Shot Fine-Grained Entity Typing with Automatic Label Interpretation and Instance Generation", KDD'22
- Yunyi Zhang, Fang Guo, Jiaming Shen, and Jiawei Han., "Unsupervised Key Event Detection from Massive Text Corpus", KDD'22
- Yu Zhang, Yu Meng, Xuan Wang, Sheng Wang, Jiawei Han, "Seed-Guided Topic Discovery with Out-of-Vocabulary Seeds", NAACL'22
- Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, Jiawei Han, "SAIS: Supervising and Augmenting Intermediate Steps for Document-Level Relation Extraction ", NAACL'22
- Xiaotao Gu, Yikang Shen, Jiaming Shen, Jingbo Shang, Jiawei Han, "Phrase-aware Unsupervised Constituency Parsing", NAACL'22
- Minhao Jiang, Xiangchen Song, Jieyu Zhang and Jiawei Han, "TaxoEnrich: Self-Supervised Taxonomy Completion via Structure-Semantic Representations", WWW'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
- Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang and Jiawei Han, "Topic Discovery via Latent Space Clustering of Language Model Embeddings", WWW'22
- Xinyang Zhang, Chenwei Zhang, Xian Li, Xin Dong, Jingbo Shang, Christos Faloutsos and Jiawei Han, "OA-Mine: Open-World Attribute Mining for E-Commerce Products with Weak Supervision", WWW'22
- Yu Zhang, Zhihong Shen, Chieh-Han Wu, Boya Xie, Junheng Hao, Ye-Yi Wang, Kuansan Wang and Jiawei Han, "Metadata-Induced Contrastive Learning for Zero-Shot Multi-Label Text Classification", WWW'22

Outline

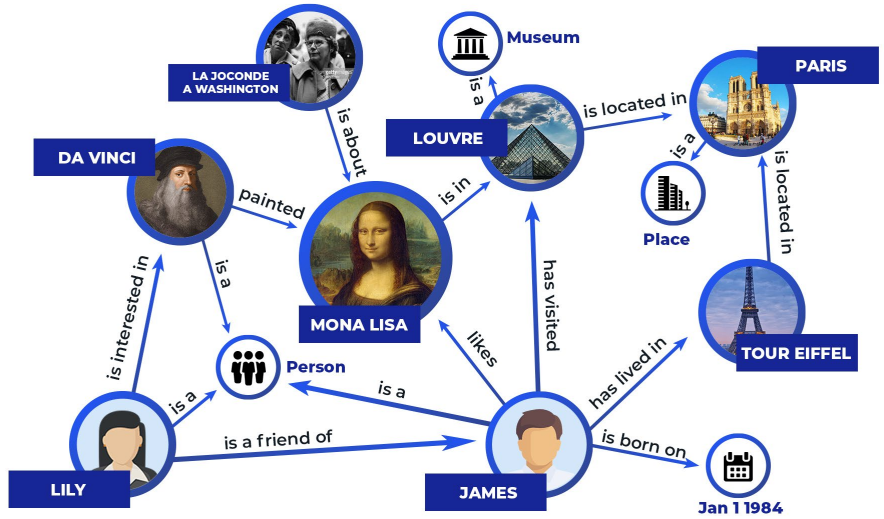
- What Kinds of Knowledge Graphs Do We Really Need?
- Key Issue: Construction of Theme-/Corpus-Based Knowledge Graphs
- The Role of Embedding and PLM in Knowledge Graph Construction
- Data Preparation: Taxonomy-Guided Text Classification
- Identifying Knowledge Graph Primitives: Entities, Properties and Relations
- Conclusion: Towards Theme/Corpus-Based Knowledge Graph Construction

Conclusions

- What kinds of knowledge graphs do we really need?
  - Theme-/corpus-based knowledge graphs
- Key issue: Automated construction of theme-/corpus-based knowledge graphs from text
  - Exploring the power of embedding and Pre-tained Language Models (PLMs)
  - Collecting and preparing data using taxonomy-guided text classification
  - Identifying knowledge graph primitives: entities, properties and relations
- Towards theme/corpus-based knowledge graph construction



Typical KGs from Knowledge-Bases



Typed Entity-Relation-Property Graphs from Text

Ack. Figures are from Google images

Jiaming Shen joins Google Research in 2021!

References

Yu Meng received 2021 Google PhD Fellowship!

Jiaxin Huang received 2021 MSR PhD Fellowship!

- Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance Kaplan and Jiawei Han, "Spherical Text Embedding", in Proc. 2019 Conf. on Neural Information Processing Systems (NeurIPS'19), Dec. 2019
- Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, Chao Zhang and Jiawei Han, "Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding", in Proc. of 2020 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'20), Aug. 2020
- Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang and Jiawei Han, "Discriminative Topic Mining via Category-Name Guided Text Embedding", in Proc. 2020 The Wide Web Conf. (WWW'20), Apr. 2020
- Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang and Jiawei Han, "Text Classification Using Label Names Only: A Language Model Self-Training Approach", in Proc. 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP'20), Nov. 2020
- Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang and Jiawei Han, "Topic Discovery via Latent Space Clustering of Language Model Embeddings", in WWW'22
- Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren and Jiawei Han, "TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names", NAACL'21
- Xuan Wang, Vivian Hu, Xiangchen Song, Shweta Garg, Jinfeng Xiao and Jiawei Han, "ChemNER: Fine-Grained Chemistry Named Entity Recognition with Ontology-guided Distant Supervision", EMNLP'21