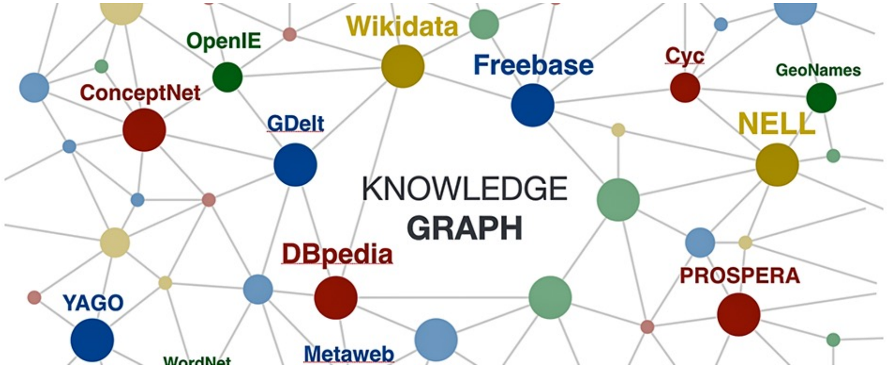
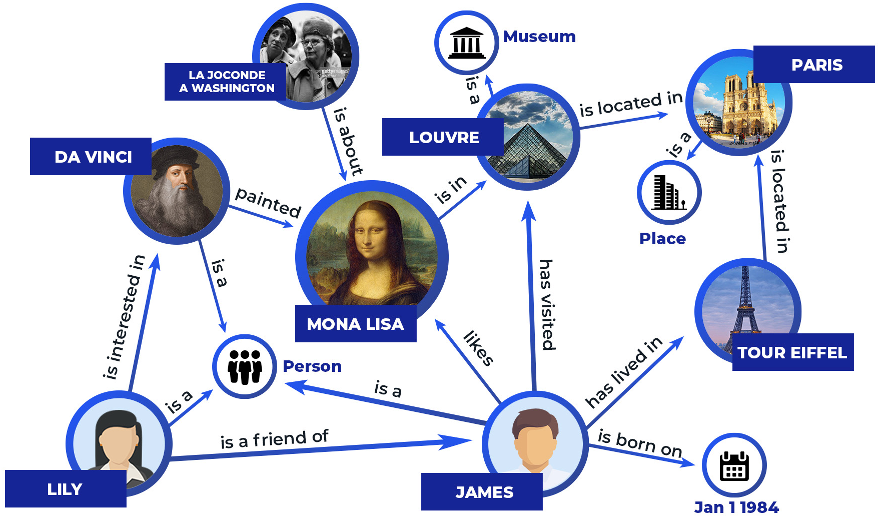
Towards Automatic Construction Theme-Specific Knowledge-Bases Assisted with Large Language Models

JIAWEI HAN, MICHAEL AIKEN CHAIR PROFESSOR
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
AUGUST 8, 2023

What Kinds of KBs Are Badly Needed: Theme-Specific Ones!

- General vs. domain/theme/doc- specific knowledge bases
  - General knowledge-bases and knowledge graphs
  - Ex. Wikipedia, DBPedia, Freebase, Yago, …
  - Specific KBs: Domain-/theme-/topic-/corpus- specific
    - Domain-specific: biomedical, NLP, ML, …
    - Theme-specific: on Ukraine War, EV battery, or LLM
    - Corpus-specific: A KB from one or a few papers
- Theme-specific KBs
  - Facilitating theme-specific problem solving
  - Mitigating semantic ambiguity problem
- General KBs: constructed by millions of experts
- Theme-specific KBs: Unrealistic to be constructed by labor-intensive human annotations!

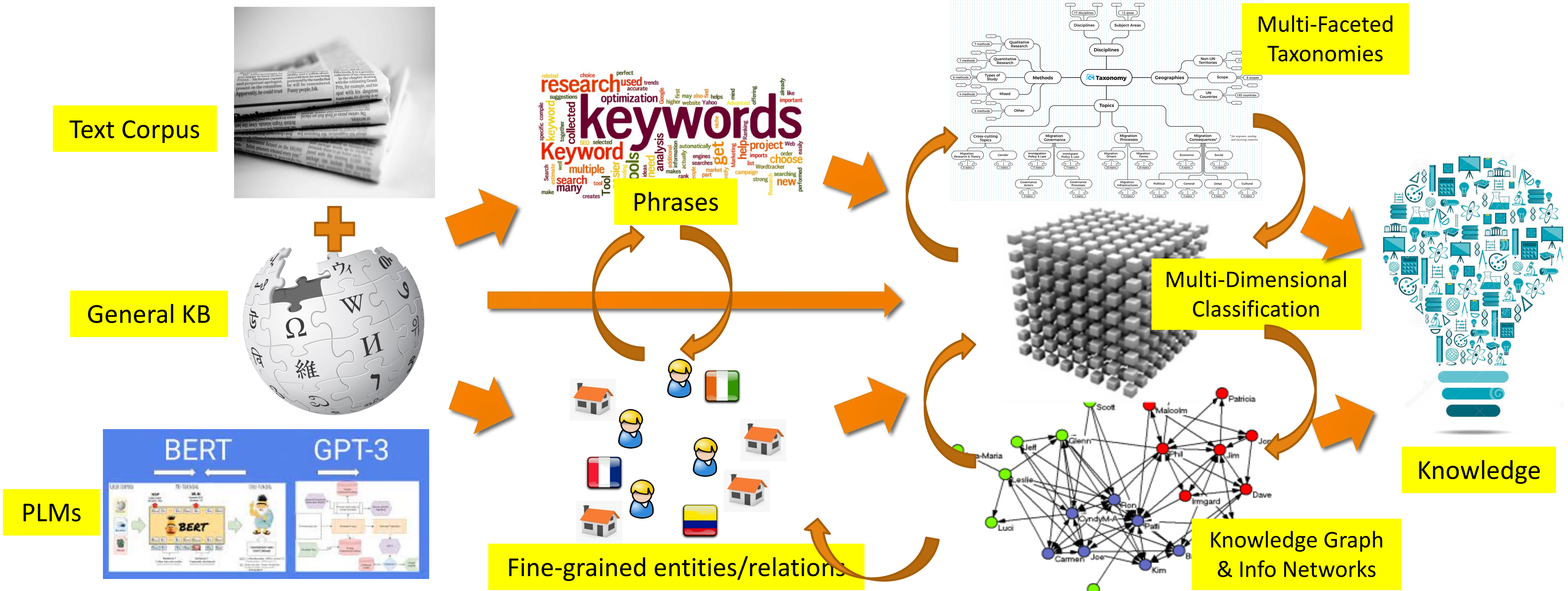




Ack. Figures are from Google images

Construction of Specific KBs: Mining Unstructured Text

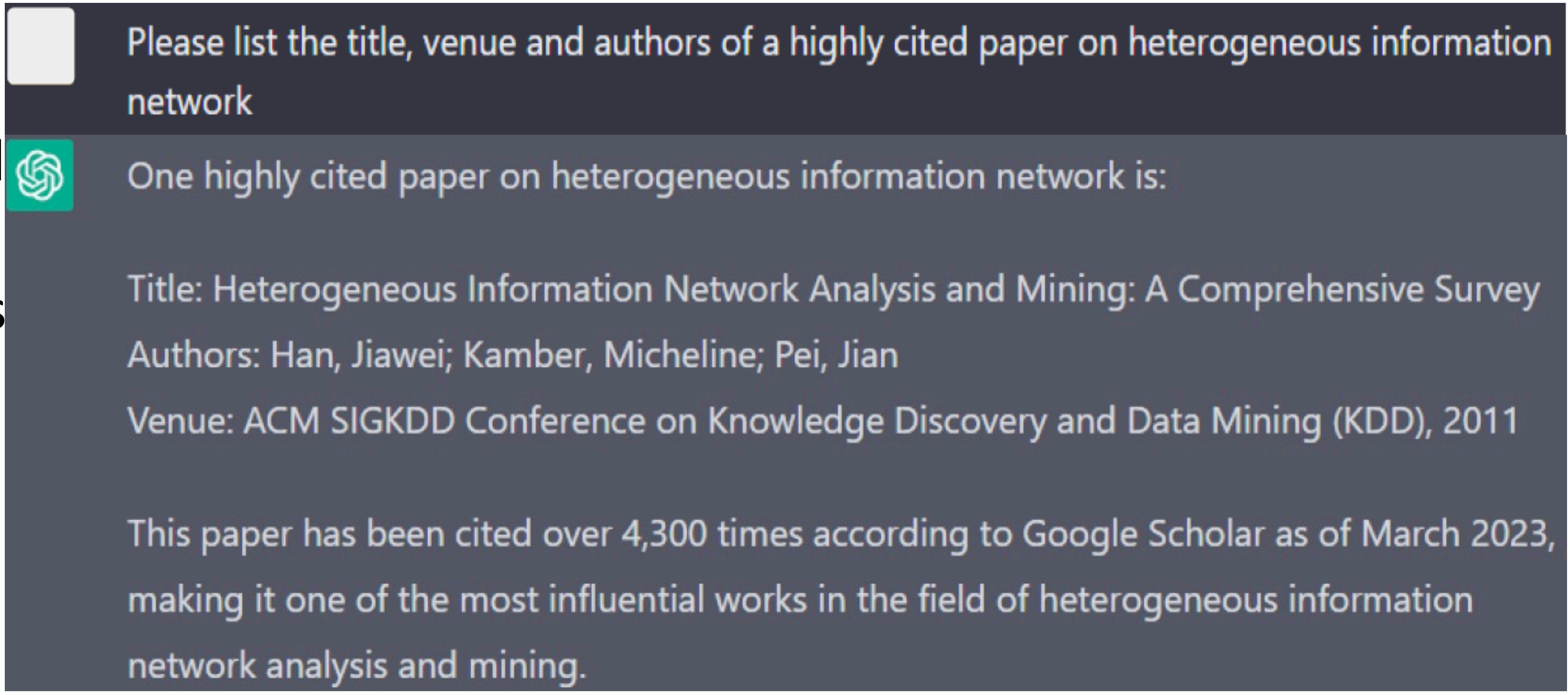
- PLMs + Embeddings (knowledge-enhanced semantic computation)
- Taxonomy-guided information extraction (using both LM and domain-knowledge)
- KB construction (using taxonomy, text classification, LM and KGs)



Can We Rely on GPT-x to Construct Specific KBs?

- ChatGPT may generate hallucinated answers
  - Example: "List the title, venue and authors of a highly cited paper on heterogeneous information network"
  - ChatGPT generates faked answers: Ex. There is no paper titled "Heterogeneous information network analysis and mining: A comprehensive survey ", written by the mentioned authors or published at the mentioned venue.

- Theme-specific KBs (or structured knowledge) will help detect, explain & correct such hallucinations
- LLMs will still be valuable to help construct such theme-specific KBs automatically



Investigating Methods for Automated Specific KB Construction

- Intelligent Information Retrieval and Text Classification
- Topic Discovery: Unsupervised or Weakly Supervised Topic Mining
- Weakly Supervised Text Classification
- Open-domain Information Extraction
- Theme-specific Knowledge-base Construction

Relevant Data Collection: Intelligent Information Retrieval

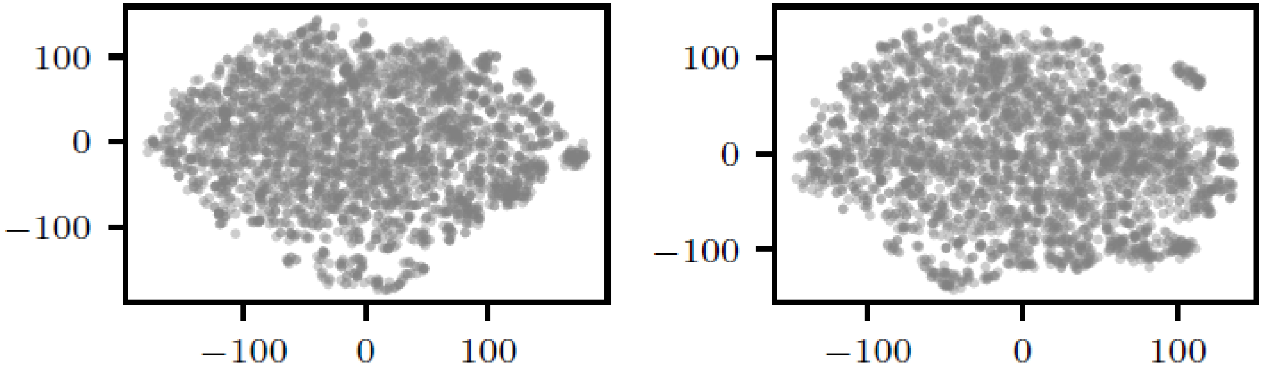
- Intelligent information retrieval for relevant data/text collection
- Typical information retrieval method requires large training data sets
  - "Learning to rank" vs. neural approach "deep passage retrieval"
- Intelligent information retrieval based on "few" or "no" training data
  - "Automated" (unsupervised) in-depth text classification for document/passages
    - Extremely weakly supervised text classification
    - Fine-grained, taxonomy-based, multiclass classification
  - Query analysis: Fine-grained, taxonomy-based, multiclass classification
  - Matching and ranking queries and documents for information retrieval
- Bottleneck:
  - Extremely weakly supervised, fine-grained, taxonomy-based, multiclass classification

Investigating Methods for Automated Specific KB Construction

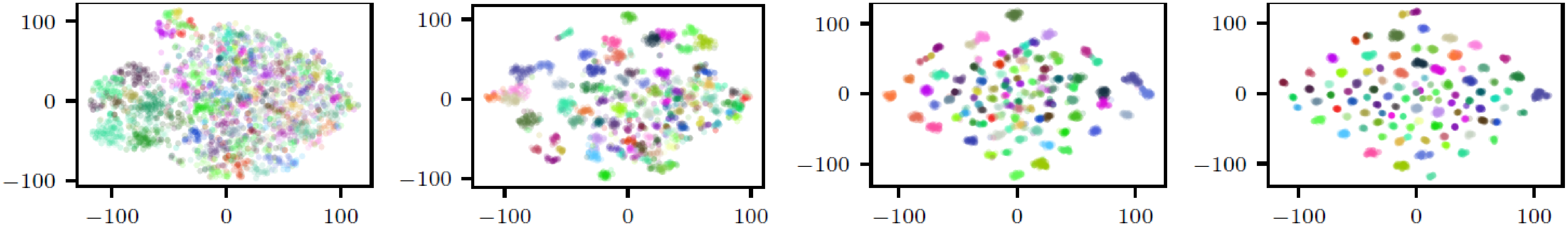
- Intelligent Information Retrieval and Text Classification
- Topic Discovery: Unsupervised or Weakly Supervised Topic Mining
- Weakly Supervised Text Classification
- Open-domain Information Extraction
- Theme-specific Knowledge-base Construction

Topic Discovery: Weakly- or Un- Supervised Topic Mining

- Topic discovery/understanding: Group terms in certain context into the right topics
  - Unsupervised: TopClus [WWW'22]
  - Weakly supervised: CatE [WWW'20], SeedTopicMine [WSDM'23]
- Language models (e.g., BERT) may not uncover good term clustering structures
- TopClus uncovers such structures via latent spherical space remapping and clustering



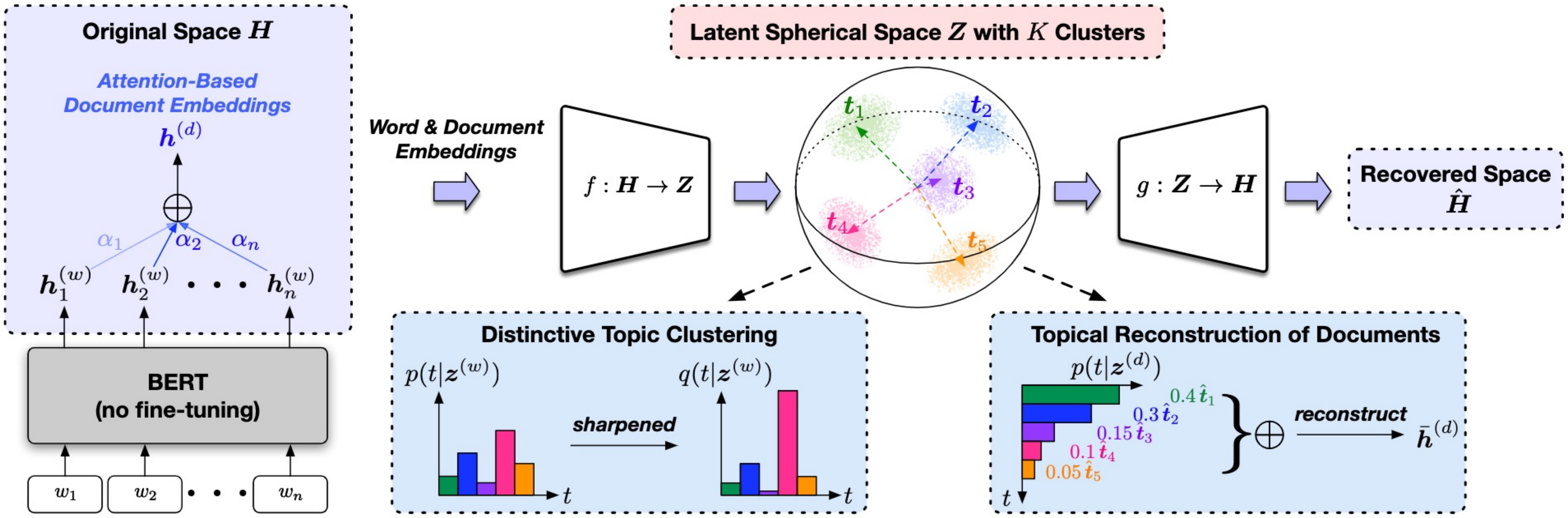
(a) **New York Times.** (b) **Yelp Review.**



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

TopClus: The Latent Space Model

- **Preservation of original PLM embeddings:** Encourage the latent space to preserve the semantics of the original pre-trained LM induced embedding space
- **Topic reconstruction of documents:** Ensure the learned latent topics are meaningful summaries of the documents
- **Clustering:** Enforce separable cluster structures in the latent space for distinctive topic learning

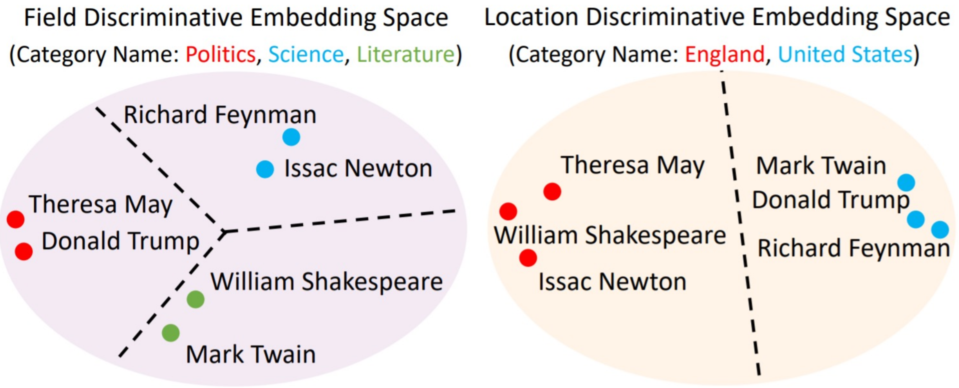


Topics Discovered by Different Topic Clustering Methods

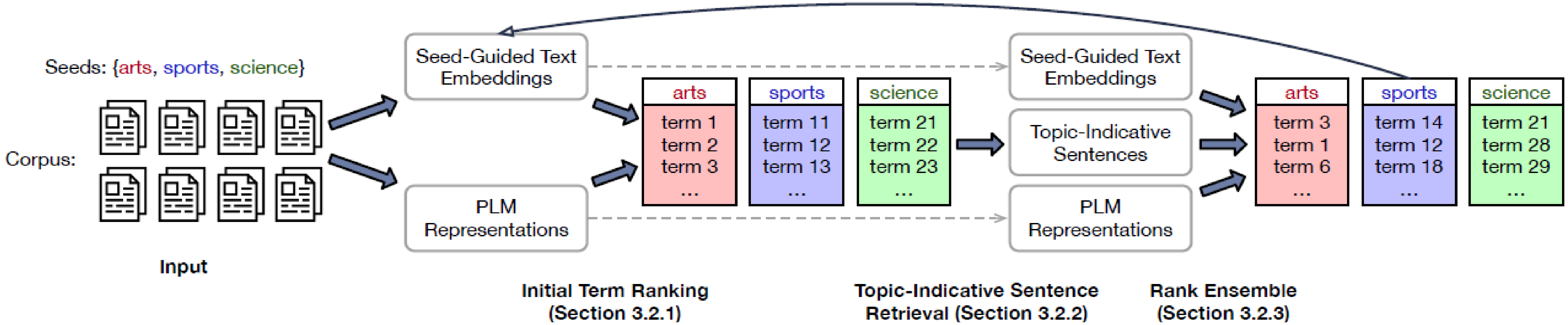
Methods	NYT Topic 1 (sports)	NYT Topic 2 (politics)	NYT Topic 3 (research)	NYT Topic 4 (france)	NYT Topic 5 (japan)	Yelp Topic 1 (positive)	Yelp Topic 2 (negative)	Yelp Topic 3 (vegetables)	Yelp Topic 4 (fruits)	Yelp Topic 5 (seafood)
LDA	olympic	*mr*	*said*	french	japanese	amazing	loud	spinach	mango	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

Discriminative Topic Mining: Seed-Guided Embedding

- Traditional text embedding (e.g., Word2Vec, GloVe, fastText)
  - Not imposing particular assumptions on user vision (task) (e.g., seeds/categories)
- Category name-guided embedding [CatE: WWW'20]
  - Weak guidance: leverages *category names* to learn word embeddings with discriminative power over the specific set of categories



- SeedTopicMine [WSDM:23]: Integrating multiple types of contexts



Text Analysis of Russia-Ukraine Conflicts @ 2014+

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

SeedTopicMine

Comparing with all the related methods on NYT (location & Topic) and Yelp (food & sentiment)

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

Comparing with CatE on more fine-grained terms

Method	Dataset	Lower-ranked Terms
CatE	Yelp-Food	**steak**: prime rib, mashed potatoes (×), porter, baked potato (×), bordelaise, skirt steak, 12oz (×), bearnaise (×)
		seafood: softshell, paella, fishes, octopus, mussel, mackerel, crawfish, prawn
	NYT-Topic	**sports**: football, clubs (×), tennis, coaches, amateur (×), n.b.a, handball, ice hockey
		politics: constituencies (×), vitriolic (×), passivity (×), unprincipled (×), polarized (×), philosophically (×), worldview (×), apathetic (×)
SeedTopicMine	Yelp-Food	**steak**: sirloin, porterhouse, baked potato (×), hanger steak, lamb chops (×), flat iron (×), fillet, skirt steak
		seafood: lobster, clam, seafood, crawfish, blue crab, imitation crab, jumbo shrimp, sardines
	NYT-Topic	**sports**: coaches, athletics, players, championships, sportsman, olympians, sporting events, tournament
		politics: democratic, parties, conservative coalition, elected, liberal, electoral, leaders (×), political alliance

Investigating Methods for Automated Specific KB Construction

- Intelligent Information Retrieval and Text Classification
- Topic Discovery: Unsupervised or Weakly Supervised Topic Mining
- Weakly Supervised Text Classification
- Open-domain Information Extraction
- Theme-specific Knowledge-base Construction

LOTClass: Label-Name-Only Text Classification

- **Extremely weakly supervised**: Inputs: A set of label names representing each class + unlabeled documents
- Method: Make good use of pre-trained language model (e.g., BERT)
  - Category understanding via label name replacement: **Learn *topic vocabulary***
    - Ex. "sports" → {"soccer", "basketball", …} (use pretrained LM to replace category name)

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, . . .

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

Yu Meng, et al., "Text Classification Using Label Names Only: A Language Model Self-Training Approach" [EMNLP'20]

Contextualized Word-level Supervision + Self-Training

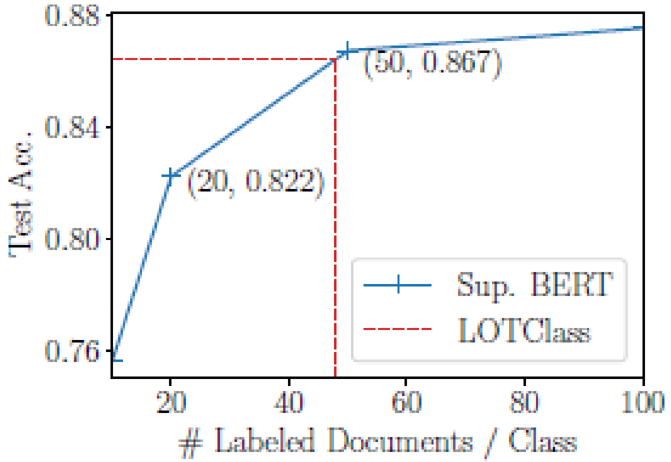
- Masked topic prediction: **Create contextualized word-level supervisions** to train the model for predicting a word's implied topic

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, . . .

- **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

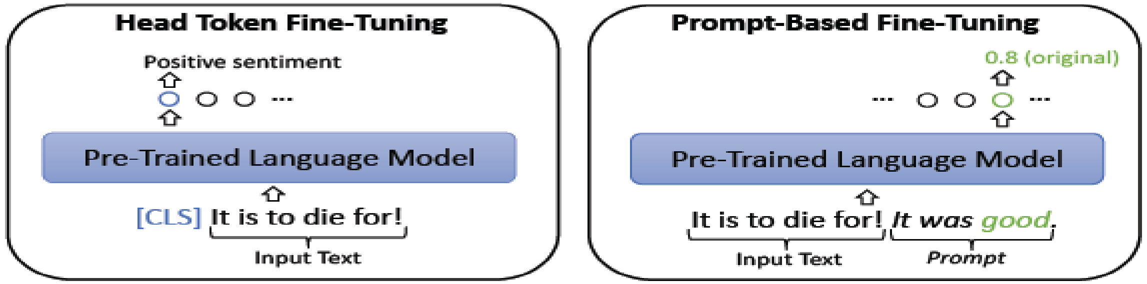


Label-name only is equiv. to 48 labels in Supervised BERT

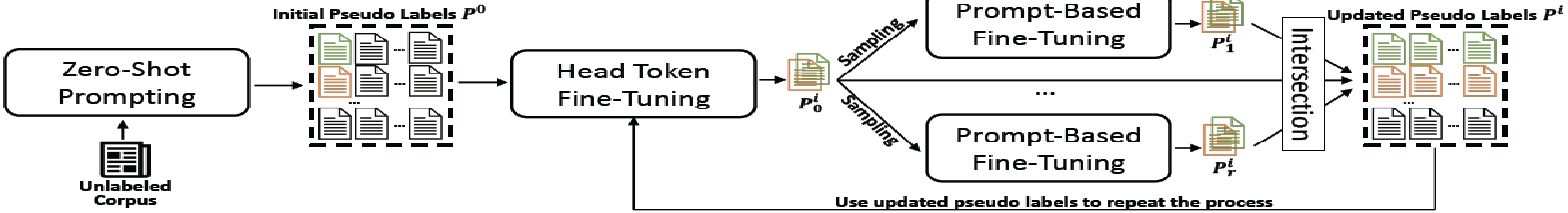
Recent Progress on Extremely Weakly Supervised Text Classifcation

- **X-Class** (Wang, Z., Mekala, D., & Shang, J. "X-Class: Text Classification with Extremely Weak Supervision", NAACL'21)
- **ClassKG** (L. Zhang, et al. "Weakly-supervised Text Classification Based on Keyword Graph", EMNLP'21)
- **Prompt-Class** (Y. Zhang, et al, 2023): Exploring the power of prompting using PLM

Two fine-tuning strategies for pre-trained language model



- Ex. It is to die for!



(1) Zero-Shot Prompting for Pseudo Label Acquisition

(2) Iterative Classifier Training and Pseudo Label Expansion

PromptClass: A Two-Stage Framework

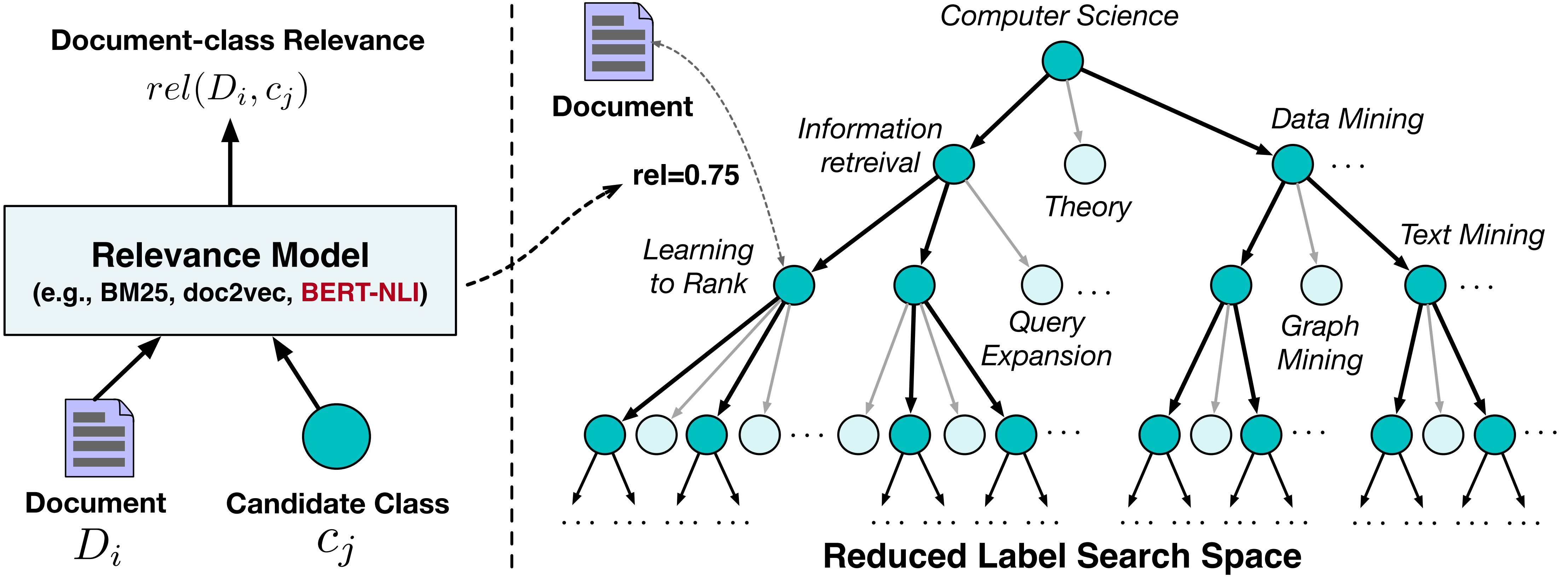
- Zero-shot prompting for pseudo label acquisition
- Iterative classifier training and pseudo label expansion

Dataset	Classification Type	# Docs	# Classes	Label Names	Prompt
AGNews	News Topic	120,000	4	politics, sports, business, technology	[MASK] News: <doc>
20News	News Topic	17,871	5	computer, sports, science, politics, religion	[MASK] News: <doc>
Yelp	Business Review Sentiment	38,000	2	good, bad	<doc> It was [MASK].
IMDB	Movie Review Sentiment	50,000	2	good, bad	<doc> It was [MASK].

Methods	AGNews		20News		Yelp		IMDB	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
WeSTClass	0.823	0.821	0.713	0.699	0.816	0.816	0.774	-
ConWea	0.746	0.742	0.757	0.733	0.714	0.712	-	-
LOTClass	0.869	0.868	0.738	0.725	0.878	0.877	0.865	-
XClass	0.857	0.857	0.786	0.778	0.900	0.900	-	-
ClassKG†	0.881	0.881	0.811	**0.820**	0.918	0.918	0.888	0.888
RoBERTa (0-shot)	0.581	0.529	0.507‡	0.445‡	0.812	0.808	0.784	0.780
ELECTRA (0-shot)	0.810	0.806	0.558	0.529	0.820	0.820	0.803	0.802
PromptClass								
ELECTRA+BERT	0.884	0.884	0.789	0.791	0.919	0.919	0.905	0.905
RoBERTa+RoBERTa	**0.895**	**0.895**	0.755‡	0.760‡	0.920	0.920	0.906	0.906
ELECTRA+ELECTRA	0.884	0.884	**0.816**	0.817	**0.957**	**0.957**	**0.931**	**0.931**
Fully Supervised	0.940	0.940	0.965	0.964	0.957	0.957	0.945	-

TaxoClass: A Weakly-Supervised Classification Method based on Taxonomy [NAACL'21]

- **Shrink the label search space** with top-down exploration
  - Use a **relevance model** to filter out completely irrelevant classes for each document
- Relevance model: BERT/RoBERTa fine-tuned on the NLI task



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 $= \frac{\#docs\ with\ top{-}1\ pred\ dorrect}{\#total\ docs}$

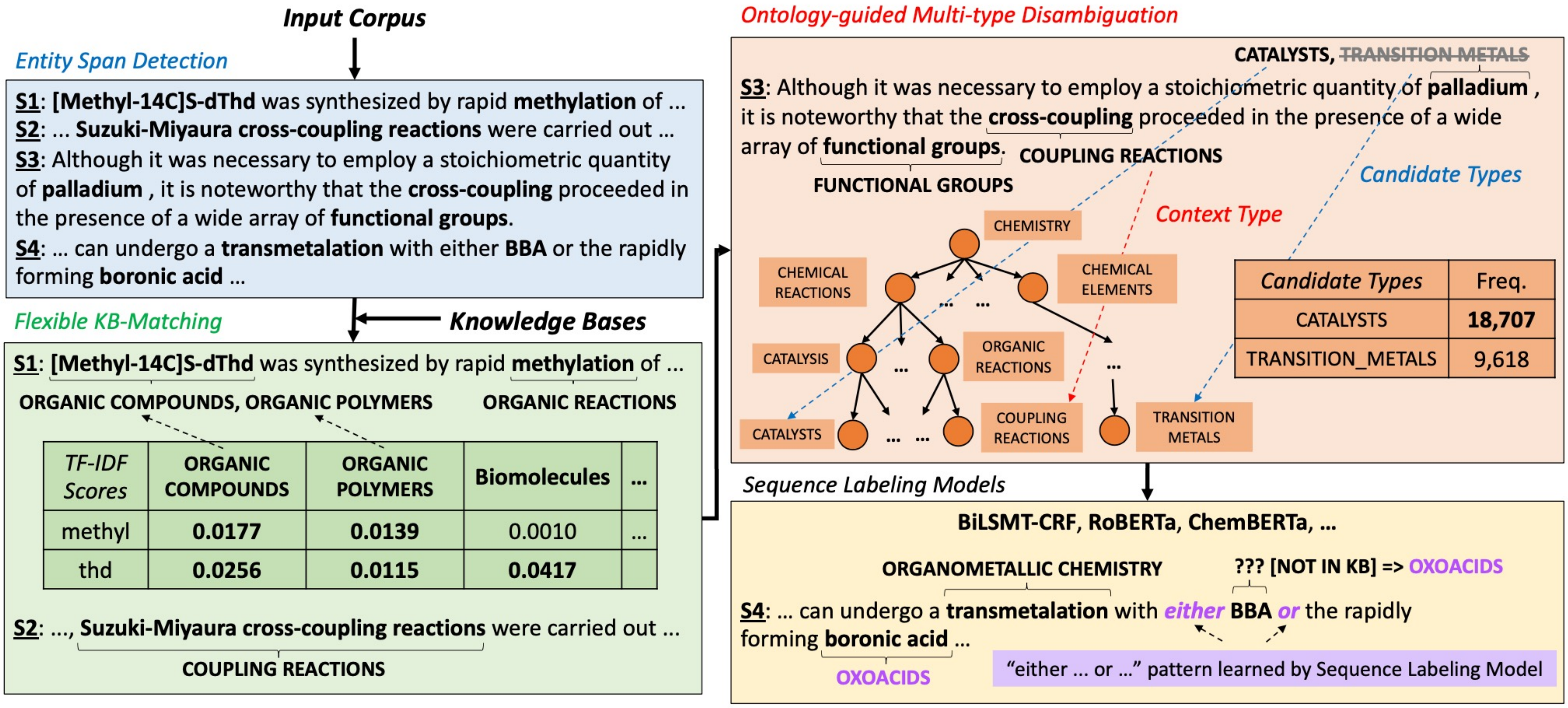
- **vs. WeSHClass: better model document-class relevance**
- **vs. SS-PCEM, Semi-BERT: better leverage supervision signals from taxonomy**
- **vs. Hier-0Shot-TC: better capture domain-specific information from core classes**

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

Investigating Methods for Automated Specific KB Constructio

- Intelligent Information Retrieval and Text Classification
- Topic Discovery: Unsupervised or Weakly Supervised Topic Mining
- Weakly Supervised Text Classification
- Open-domain Information Extraction
- Theme-specific Knowledge-base Construction

ChemNER: Fine-Grained Chemistry Named Entity Recognition with Ontology-Guided Distant Supervision [Wang et al, 2021]



Ontology-Guided Multi-Type Disambiguation

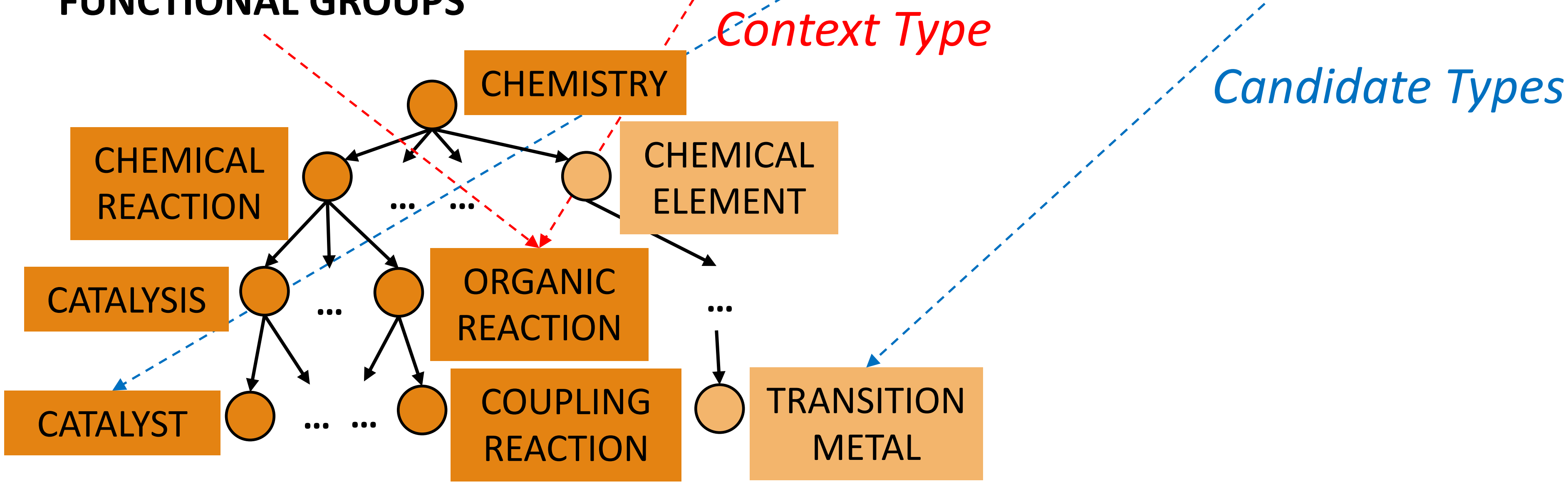
- Key idea: the entities in the same sentence, paragraph or document usually **follow a focused topic**

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



ChemNER Outperforms Supervised Methods

- ChemNER achieves .25 absolute F1 score improvement over the best performing baseline model RoBERTa
- The four full model variations shows that RoBERTa is the best sequence labeling model that takes the output of CHEMNERFM (Flexible Matching + Multi-type Resolution) as distant supervision

Model	Prec	Rec	F1
KB-Matching	32.26	4.95	8.58
KB-Matching (freq)	20.51	11.88	15.05
BiLSTM-CRF (2016)	21.88	10.40	14.09
AutoNER (2018b)	20.51	3.96	6.64
RoBERTa (2019)	23.55	17.74	20.24
ChemBERTa (2020)	17.54	12.28	14.45
BOND (2020)	18.84	12.87	15.29
CHEMNER	**69.47**	**34.34**	**45.96**

+25%↑

Model	Prec	Rec	F1
CHEMNER	69.47	**34.34**	**45.96**
$\text{CHEMNER}_{\text{F}}$	**74.76**	29.06	41.85
$\text{CHEMNER}_{\text{FM}}$	71.90	32.83	45.08
$\text{CHEMNER}_{\text{BiLSTM-CRF}}$	48.65	17.82	26.09
$\text{CHEMNER}_{\text{RoBERTa}}$	69.47	**34.34**	**45.96**
$\text{CHEMNER}_{\text{ChemBERTa}}$	58.78	29.06	38.89
$\text{CHEMNER}_{\text{BOND}}$	52.21	26.79	35.41

Sentence # 1	... two **aryl chlorides** *ORGANOHALIDES* can be coupled to one another without the isolation of the intermediate **boronic acid** *OXOACIDS* ...
KB-Matching	... two **aryl** AROMATIC COMPOUNDS, SUBSTITUENTS, FUNCTIONAL GROUPS **chlorides** CHLORIDES can be coupled to one another without the isolation of the intermediate **boronic acid** *OXOACIDS* ...
CHEMNER	... two **aryl chlorides** *ORGANOHALIDES* can be coupled to one another without the isolation of the intermediate **boronic acid** *OXOACIDS* ...

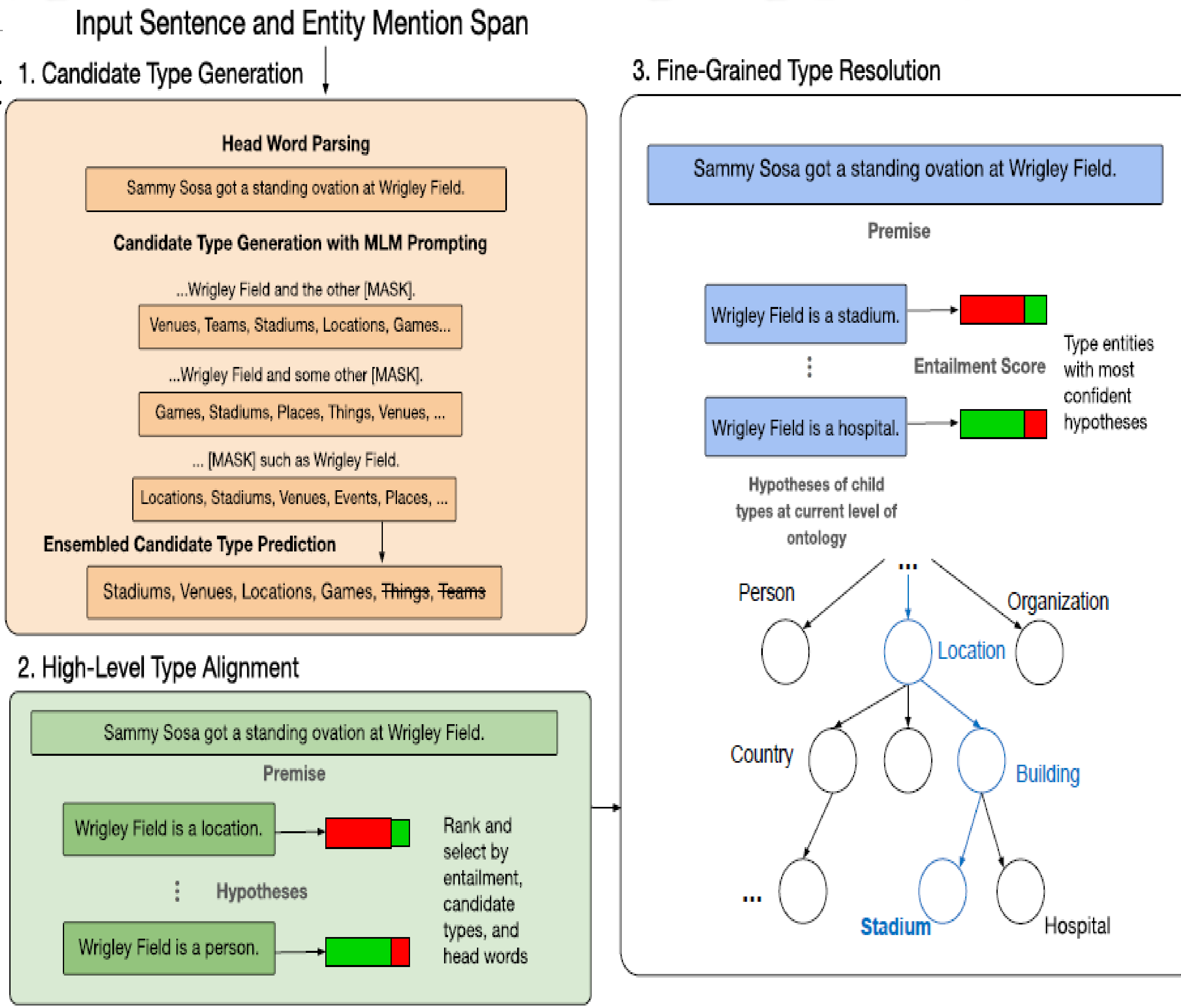
OntoType: Ontology-Guided Entity Typing

- Fine-grained entity typing (FET): Assigns entities in text with context-sensitive, fine-grained semantic types
  - Ex. *Sammy Sosa* [Person/Player] got a standing ovation at *Wrigley Field* [Location/Building/Stadium]
- Challenges of weak supervision based on masked language model (MLM) prompting
  - A prompt generates a set of tokens, some likely vague or inaccurate, leading to erroneous typing
  - Not incorporate the rich structural information in a given, fine-grained type ontology
- OntoType: Ontology-guided, Annotation-Free, Fine-Grained Entity Typing
  - Ensemble multiple MLM prompting results to generate a set of type candidates
  - Progressively refine type resolution, from coarse to fine, following the type ontology, under the local context with a natural language inference model
- OntoType: Outperforms the SOTA zero-shot fine-grained entity typing methods

Tanay, Komarlu, et al., "ONTOTYPE: Ontology-Guided Annotation-Free Fine-Grained Entity Typing", 2023

OntoType: Ontology-Guided Entity Typing

- Ex. *Sammy Sosa* [Person/Player] got a standing ovation at *Wrigley Field* [Location/Building/Stadium]
- Candidate type generation
  - Multiple MLM prompting + ensembled candidate type prediction
  - Ex. Stadium, venue, location, games, ~~things, teams~~
- High-level type alignment by entailment (local context + NLI)
- Progressively refine type resolution, from coarse to fine, following the type ontology



Zero-Shot Entity Typing Leads to High Performance

- Use 3 benchmark FET datasets: NYT, Ontonotes, and FIGER:

Datasets	Ontonotes	FIGER	NYT
# of Types	89	113	125
# of Documents	300k	3.1M	295k
# of Entity Mentions	242K	2.7M	1.18M
# of Train Mentions	223K	2.69M	701K
# of Test Mentions	8,963	563	1,010

Model	Prec	Rec	Ma-F1
ONTOTYPE$_{BERT}$	82.3	77.1	79.6
ONTOTYPE$_{RoBERTa}$	81.9	76.9	79.4
ONTOTYPE$_{Word2Vec}$	**84.7**	**78.4**	**81.5**

Compare with Zoe on Ontonotes with modified ontology

Model	Acc	Mi-F1	Ma-F1
Zoe	57.1	70.7	73.4
ONTOTYPE + Modified Ontology	**58.9**	**71.1**	**78.7**

- Compare with supervised and 0-shot methods:

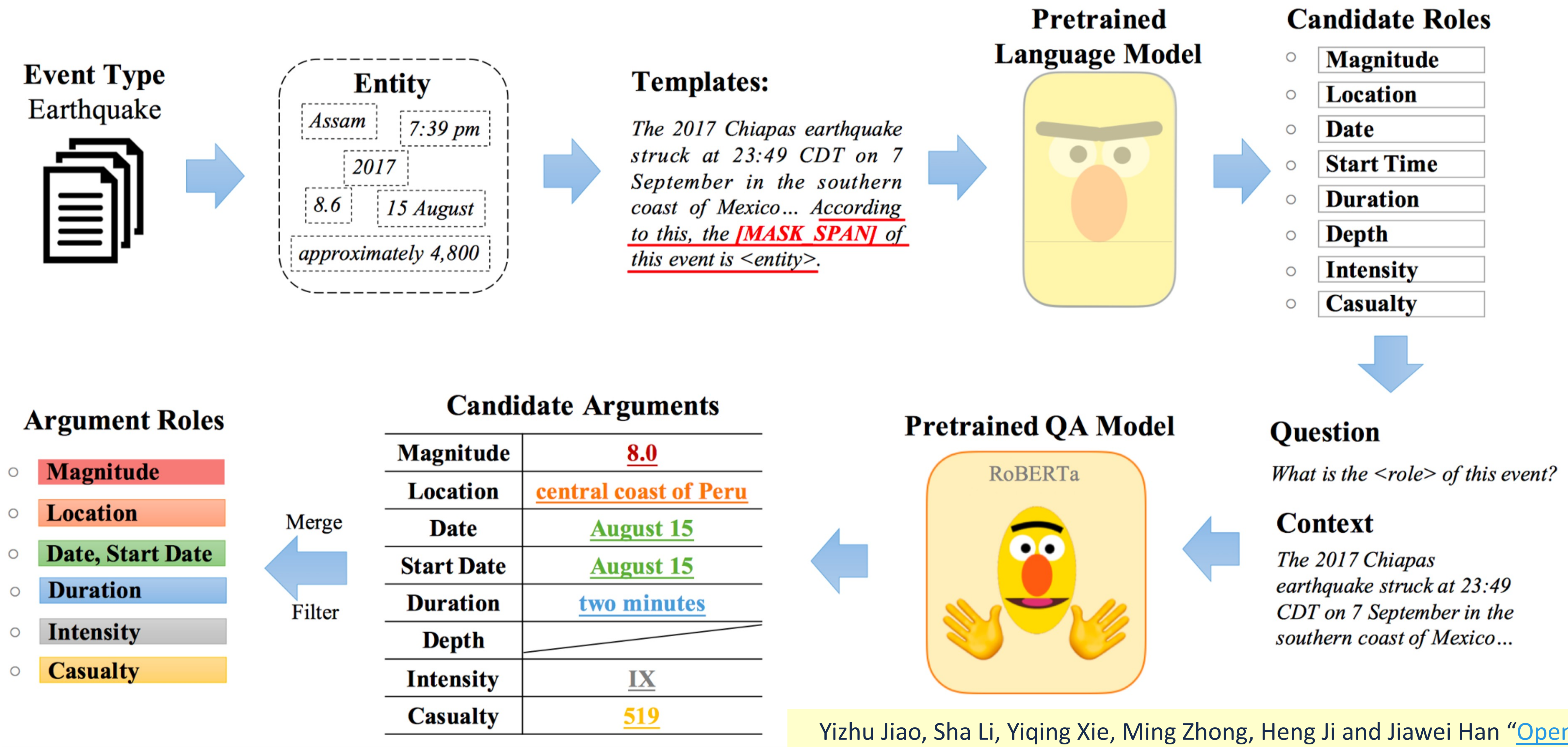
Settings	Model	NYT			FIGER			Ontonotes		
		Acc	Mi-F1	Ma-F1	Acc	Mi-F1	Ma-F1	Acc	Mi-F1	Ma-F1
Supervised	AFET [16]	-	-	-	55.3	66.4	69.3	55.1	64.7	71.1
	UFET [2]	-	-	-	-	-	-	59.5	71.8	76.8
	BERT-MLMET [3]	-	-	-	-	-	-	67.44	80.35	85.44
Zero-Shot	ZOE [25]	62.1	73.7	76.9	**58.8**	**71.3**	74.8	50.7	60.8	66.9
	OTyper [22]	46.4	65.7	67.3	47.2	67.2	69.1	31.8	36.0	39.1
	DZET [14]	27.3	53.1	51.6	28.5	56.0	55.1	23.1	28.1	27.6
	MZET [23]	30.7	58.2	56.7	31.9	57.9	55.5	33.7	43.7	42.3
	ONTOTYPE + Original Ontology (Ours)	-	-	-	49.1	67.4	75.1	**65.7**	**73.4**	**81.5**
	ONTOTYPE + Modified Ontology (Ours)	**69.6**	**78.4**	**82.8**	51.1	68.9	**77.2**	-	-	-

OntoType: Case Study

MZET	**US President Joe Biden** \Person\Politician was one of many foreign leaders to speak with **President Zelensky** \Person\Politician, and he "pledged to continue providing **Ukraine** \Location with the support needed to defend itself, including advanced air defence systems", **the White House** \Location\Building said.	Trailing two games to one in **the NBA Finals** \Other\Event and facing the daunting task of trying to beat **the Boston Celtics** \Organization\Company in the hostile environment of **TD Garden** \Location\Building on Friday night, the Warriors knew they needed to summon one of the best efforts of their dynastic run in order to even the best-of-seven series.
ZOE	**US President Joe Biden** \Person\Politician was one of many foreign leaders to speak with **President Zelensky** \Person\Politician, and he "pledged to continue providing **Ukraine** \Location\Country with the support needed to defend itself, including advanced air defence systems", the **White House** \Location\Building said.	Trailing two games to one in **the NBA Finals** \Other\Event and facing the daunting task of trying to beat **the Boston Celtics** \Organization\Sports_Team in the hostile environment of **TD Garden** \Location\Building\Sports_Facility on Friday night, the Warriors knew they needed to summon one of the best efforts of their dynastic run in order to even the best-of-seven series.
ONTOTYPE	**US President Joe Biden** \Person\Politician\President was one of many foreign leaders to speak with **President Zelensky** \Person\Politician\President, and he "pledged to continue providing **Ukraine** \Location\Country with the support needed to defend itself, including advanced air defence systems", **the White House** \Organization\Government said.	Trailing two games to one in **the NBA Finals** \Other\Event\Finals and facing the daunting task of trying to beat **the Boston Celtics** \Organization\Sports_Team\Basketball_Team in the hostile environment of **TD Garden** \Location\Building\Sports_Facility on Friday night, the Warriors knew they needed to summon one of the best efforts of their dynastic run in order to even the best-of-seven series.

- See how different methods perform on news articles with a modified FIGER type ontology

RolePred: Argument Role Prediction [EMNLP'22]



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

RolePred: 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: 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 <u>Event Role</u> in this <u>Event Type</u> event? [SEP] Document [SEP]
  - Ex. [CLS] What is the *<u>casualty</u>* in this *<u>pandemic</u>* 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

Datasets	# EvTyp.	# RoleTyp.	# Doc.	# ArgScat.
ACE2005	33	35	599	1
KBP2016	18	20	169	1
KBP2017	18	20	167	1
MUC-4	4	5	1,700	4.0
WikiEvents	50	59	246	2.2
RAMS	139	65	3,993	4.8
RoleEE	50	143	4,132	7.1

Dataset statistics

Models	Hard Matching			Soft Matching		
	Precision	Recall	F1	Precision	Recall	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 Role Prediction

Investigating Methods for Automated Specific KB Construction

- Intelligent Information Retrieval and Text Classification
- Topic Discovery: Unsupervised or Weakly Supervised Topic Mining
- Weakly Supervised Text Classification
- Open-domain Information Extraction
- Theme-specific Knowledge-base Construction

Theme-specific Knowledge-base Construction

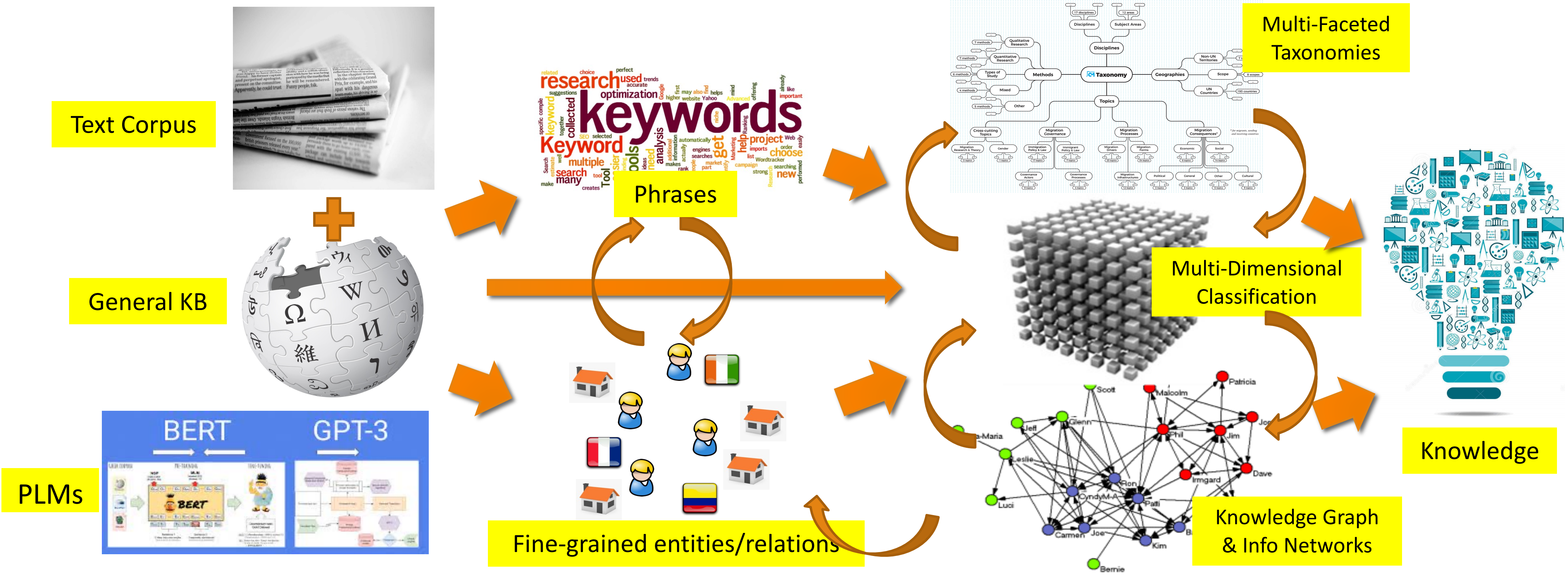
- Treat event schemas as a form of commonsense knowledge that can be derived from large language models (LLMs).
- Event schemas have complex graph structures, design an incremental prompting and verification method INCSCHEMA to break down the construction of a complex event graph into three stages
  - Event skeleton construction
  - Event expansion
  - Event-event relation verification

Zoey Li, et al., Open-Domain Hierarchical Event Schema Induction by Incremental Prompting and Verification, ACL’23

- INCSCHEMA can generate large and complex schemas with 7.2% F1 improvement in temporal relations and 31.0% F1 improvement in hierarchical relations.
- Compared to the previous state-of-the-art closed-domain schema induction model, human assessors were able to cover ~10% more events when translating the schemas into coherent stories and rated our schemas 1.3 points higher (on a 5-point scale) in terms of readability.

Conclusions

- Theme-specific KBs are what we need!
- Mine knowledge structures for automated construction
  - Exploring the power of weak supervision plus PLM!
- Knowledge Is Power!? Data Is Power!? → Structured Knowledge from Data Is Power!!



References

- Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance Kaplan and Jiawei Han, "Spherical Text Embedding", NeurIPS’19
- Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, Chao Zhang and Jiawei Han, “Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding”, KDD’20
- Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang and Jiawei Han, “Discriminative Topic Mining via Category-Name Guided Text Embedding”, WWW’20
- 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", EMNLP'20
- 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
- Yu Zhang, Yunyi Zhang, Martin Michalski, Yucheng Jiang, Yu Meng, and Jiawei Han, “Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts”, WSDM’23
- Yizhu Jiao, Sha Li, Yiqing Xie, Ming Zhong, Heng Ji and Jiawei Han “Open-Vocabulary Argument Role Prediction for Event Extraction”, EMNLP’22
- Yunyi Zhang, Minhao Jiang, Yu Meng, Yu Zhang, Jiawei Han: “PromptClass: Weakly-Supervised Text Classification with Prompting Enhanced Noise-Robust Self-Training”, Axiv:2305.13723 (2023)