Part II: Text Representation Enhanced Topic Discovery

KDD 2023 Tutorial

Pretrained Language Representations for Text Understanding: A Weakly-Supervised Perspective

Yu Meng, Jiaxin Huang, Yu Zhang, Yunyi Zhang, Jiawei Han

Computer Science, University of Illinois Urbana-Champaign

Aug 9, 2023

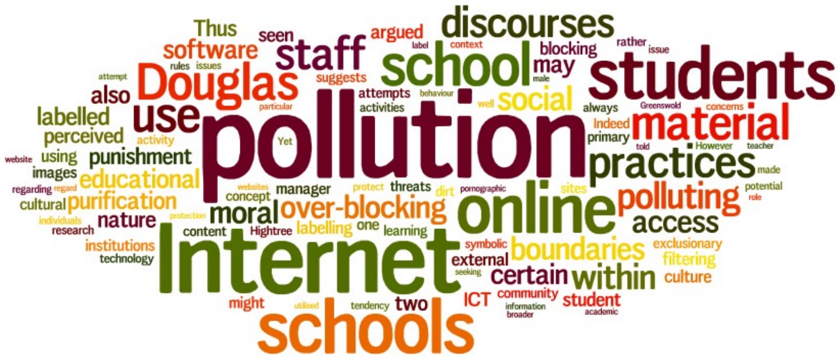
Tutorial Website:

Outline

- Traditional Topic Models
- Embedding-Based Discriminative Topic Mining
- Topic Discovery with PLMs

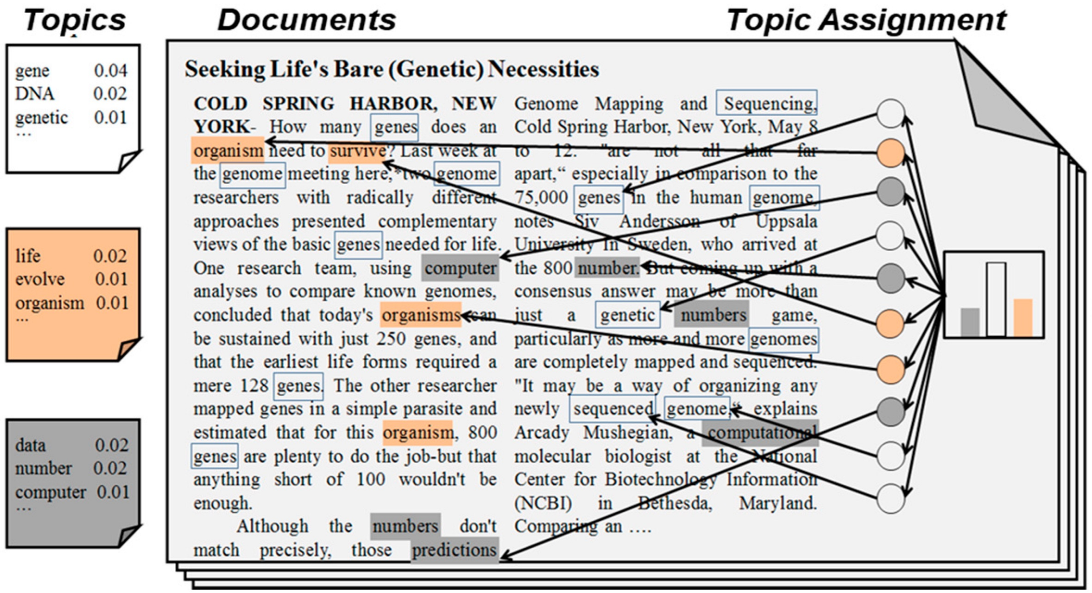
Topic Modeling: Introduction

- How to effectively & efficiently comprehend a large text corpus?
- Knowing what important topics are there is a good starting point!
- Topic discovery facilitates a wide spectrum of applications
  - Document classification/organization
  - Document retrieval/ranking
  - Text summarization



Topic Modeling: Overview

- How to discover topics automatically from the corpus?
- By modeling the corpus statistics!
  - Each document has a latent topic distribution
  - Each topic is described by a different word distribution



Word-Topic Matrix

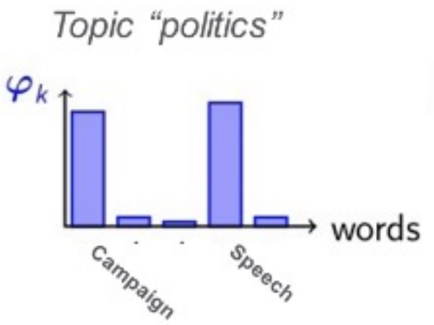
	Topic 1	…	Topic n
Word 1	0.09	…	0.06
…	….	…	…
Word n	0.08	…	0.01

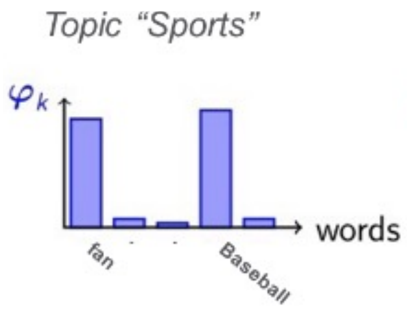
Document-Topic Matrix

	Topic 1	…	Topic n
Doc 1	0.23	…	0.33
…	….	…	…
Doc n	0.15	…	0.28

Latent Dirichlet Allocation (LDA): Overview

- Each document is represented as a mixture of various topics
  - Ex. A news document may be 40% on politics, 50% on economics, and 10% on sports
- Each topic is represented as a probability distribution over words
  - Ex. The distribution of "politics" vs. "sports" might be like:

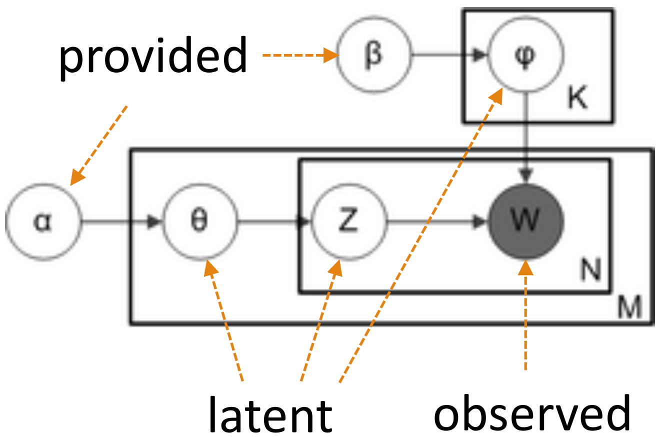




- Dirichlet priors are imposed to enforce sparse distributions:
  - Documents cover only a small set of topics (sparse document-topic distribution)
  - Topics use only a small set of words frequently (sparse topic-word distribution)

LDA: Inference

- Learning the LDA model (Inference)
- What need to be learned
  - Document-topic distribution θ (for assigning topics to documents)
  - Topic-word distribution φ (for topic interpretation)
  - Words' latent topic z
- How to learn the latent variables? – complicated due to intractable posterior
  - Monte Carlo simulation
  - Gibbs sampling
  - Variational inference
  - …



Issues with LDA

- LDA is completely unsupervised (i.e., users only input number of topics)
- Cannot take user supervision
  - Ex. What if a user is specifically interested in some topics but LDA doesn't discover them?

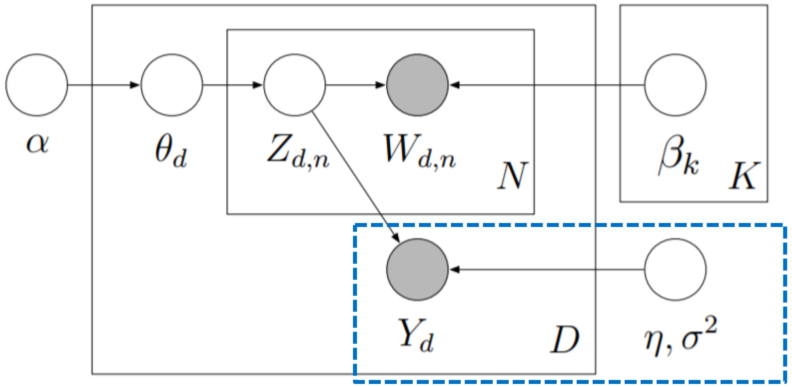
	Topic 1	Weight	Topic 2	Weight	Topic 3	Weight	Topic 4	Weight	Topic 5	Weight
0	life	0.018076	father	0.059603	official	0.017620	case	0.021908	art	0.010555
1	man	0.017714	graduate	0.048363	force	0.015388	law	0.020698	open	0.010413
2	woman	0.016657	son	0.042746	military	0.014587	court	0.019967	room	0.010363
3	book	0.010486	mrs	0.041379	war	0.011381	lawyer	0.016935	house	0.009002
4	family	0.010382	daughter	0.037156	government	0.010564	state	0.014501	building	0.008722
5	young	0.009896	mother	0.034542	troop	0.008949	judge	0.012487	artist	0.008264
6	write	0.009493	receive	0.029211	attack	0.008886	legal	0.011141	design	0.008162
7	child	0.009460	marry	0.029038	leader	0.008082	rule	0.009854	floor	0.008034
8	live	0.008819	yesterday	0.024107	peace	0.006835	decision	0.009261	museum	0.007917
9	love	0.007814	degree	0.022899	soldier	0.006562	file	0.008289	exhibition	0.007222

	Topic 6	Weight	Topic 7	Weight	Topic 8	Weight	Topic 9	Weight	Topic 10	Weight
0	group	0.051052	market	0.024976	serve	0.010918	change	0.007661	city	0.021776
1	member	0.040683	stock	0.024874	add	0.010185	system	0.007233	area	0.014865
2	meeting	0.016390	share	0.020583	minute	0.009301	problem	0.006835	build	0.014361
3	issue	0.014988	price	0.018141	pepper	0.009235	power	0.005400	building	0.014326
4	official	0.013069	sell	0.016564	oil	0.008976	create	0.005056	home	0.013632
5	support	0.011994	buy	0.015415	cook	0.008711	research	0.004712	resident	0.013483
6	leader	0.011799	company	0.015249	food	0.008689	produce	0.004574	community	0.012479
7	organization	0.011135	investor	0.015062	cup	0.008682	far	0.004447	local	0.010686
8	meet	0.010235	yesterday	0.012813	sauce	0.008209	result	0.004280	live	0.010661
9	effort	0.008479	analyst	0.010768	small	0.007864	kind	0.004166	project	0.010459

10 topics generated by LDA on The New York Times dataset

Supervised LDA (sLDA)

- Allow users to provide document annotations/labels
- Incorporate document labels into the generative process
  - For the ith document, choose $\theta_i \sim \mathrm{Dir}(\alpha)$ document's topic distribution
  - For the jth word in the ith document,
    - choose topic $z_{i,j} \sim \mathrm{Categorical}(\theta_i)$ word's topic
    - choose a word $w_{i,j} \sim \mathrm{Categorical}(\beta_{z_{i,j}})$
  - For the ith document, choose $y_i \sim N(\eta^\top \bar{z}_i, \sigma^2)$, $\bar{z}_i = \frac{1}{L}\sum_{j=1}^{L} z_{i,j}$



generate document's label

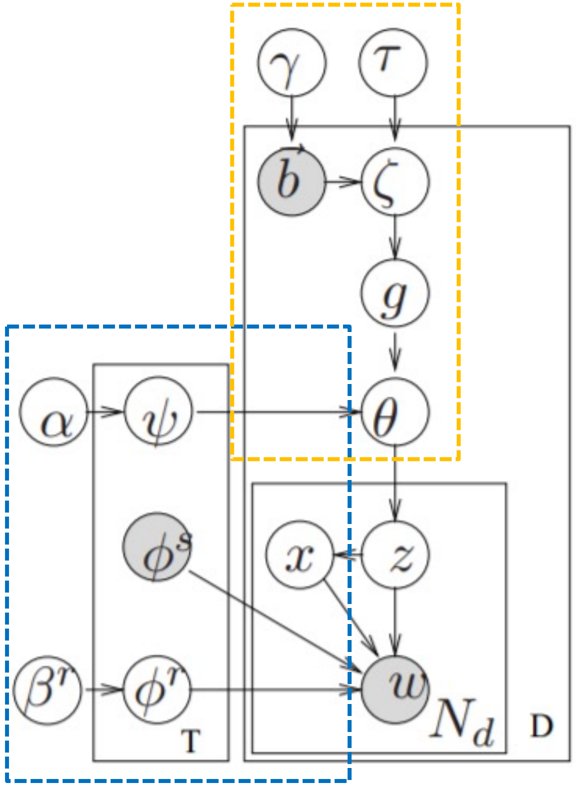
Seeded LDA: Guided Topic-Word Distribution

- Another form of user supervision: several seed words for each topic

1. For each k=1 $\cdots$ T,
   (a) Choose regular topic $\phi_k^r \sim \mathrm{Dir}(\beta_r)$.
   (b) Choose *seed* topic $\phi_k^s \sim \mathrm{Dir}(\beta_s)$.
   (c) Choose $\pi_k \sim \mathrm{Beta}(1,1)$.
2. For each seed set $s = 1 \cdots$ S,
   (a) Choose group-topic distribution $\psi_s \sim \mathrm{Dir}(\alpha)$.
3. For each document d,
   (a) Choose a binary vector $\vec{b}$ of length S.
   (b) Choose a document-group distribution $\zeta^d \sim \mathrm{Dir}(\tau\vec{b})$.
   (c) Choose a group variable $g \sim \mathrm{Mult}(\zeta^d)$.
   (d) Choose $\theta_d \sim \mathrm{Dir}(\psi_g)$. // of length T
   (e) For each token $i = 1 \cdots N_d$:
      i. Select a topic $z_i \sim \mathrm{Mult}(\theta_d)$.
      ii. Select an indicator $x_i \sim \mathrm{Bern}(\pi_{z_i})$.
      iii. if x_i is 0
         - Select a word $w_i \sim \mathrm{Mult}(\phi_{z_i}^r)$.
      iv. if x_i is 1
         - Select a word $w_i \sim \mathrm{Mult}(\phi_{z_i}^s)$.

Seed topics used to improve the document-topic distribution:
Group-topic distribution = seed set distribution over regular topics
Group-topic distribution used as prior to draw document-topic distribution

Seed topics used to improve the topic-word distribution:
Each word comes from either "regular topics" with a distribution over all word like in LDA, or "seed topics" which only generate words from the seed set



Outline

- Traditional Topic Models
- Embedding-Based Discriminative Topic Mining
  - Introduction of the Task
  - CatE: Discriminative Topic Mining via Category-Name Guided Text Embedding [WWW’20]
  - JoSH: Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding [KDD’20]
- Topic Discovery with PLMs

Limitations of Unsupervised Topic Discovery

- **Cannot incorporate user guidance:** Topic models tend to retrieve the most general and prominent topics from a text collection
  - may not be of a user's particular interest
  - provide a skewed and biased summarization of the corpus
- **Cannot enforce distinctiveness among retrieved topics:** Topic models do not impose discriminative constraints
  - E.g., three retrieved topics from the New York Times annotated corpus via LDA

Table 1: LDA retrieved topics on NYT dataset. The meanings of the retrieved topics have overlap with each other.

Topic 1	Topic 2	Topic 3
canada, united states canadian, economy	sports, united states olympic, games	united states, iraq government, president

Difficult to clearly define the meaning of the three topics due to an overlap of their semantics (e.g., the term "united states" appears in all 3 topics)

Seed-Guided, Discriminative Topic Mining

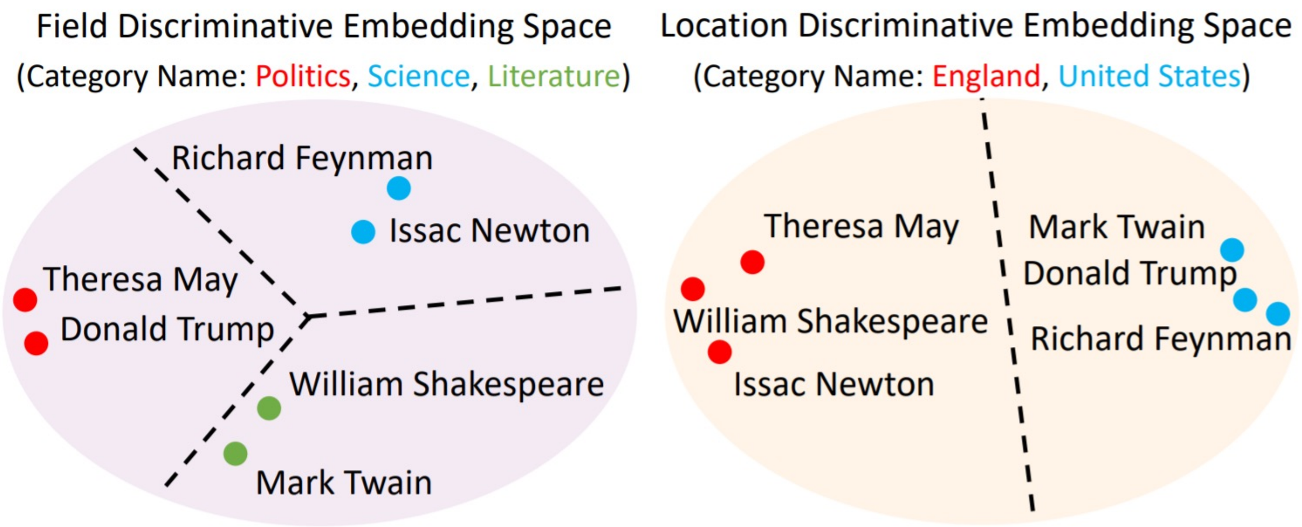
- **Discriminative Topic Mining:** Given a text corpus and a set of **category names**, retrieve a set of terms that **exclusively belong to** each category
  - E.g., given c_1: "The United States", c_2: "France", c_3: "Canada"
    - Yes to "Ontario" under c_3: (a province in Canada and exclusively belongs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

Outline

- Traditional Topic Models
- Embedding-Based Discriminative Topic Mining
  - Introduction of the Task
  - CatE: Discriminative Topic Mining via Category-Name Guided Text Embedding [WWW'20]
  - JoSH: Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding [KDD'20]
- Topic Discovery with PLMs

Discriminative Topic Mining via CatE

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



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

CatE Embedding: Text Generation Modeling

- Modeling text generation under user guidance
- A three-step process:
  1. A document d is generated conditioned on one of the n categories
  2. Each word w_i is generated conditioned on the semantics of the document d
  3. Surrounding words w_{i+j} in the local context window of w_i are generated conditioned on the semantics of the center word w_i
- Compute the likelihood of corpus generation conditioned on user-given categories

1. Topic assignment

2. Global context

3. Local context

CatE Embedding: Objective

- Objective: negative log-likelihood

$$P(\mathcal{D} \mid C) = \underbrace{\prod_{d \in \mathcal{D}} p(d \mid c_d)}_{\text{1. Topic assignment}} \underbrace{\prod_{w_i \in d} p(w_i \mid d)}_{\text{2. Global context}} \prod_{\substack{w_{i+j} \in d \\ -h \le j \le h, j \neq 0}} \underbrace{p(w_{i+j} \mid w_i)}_{\text{3. Local context}}$$

$$p(d \mid c_d) \propto p(c_d \mid d) p(d) \propto p(c_d \mid d) \propto \prod_{w \in d} p(c_d \mid w),$$ Decompose into word-topic distribution

- Introducing specificity

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

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

Category Representative Word Retrieval

- Ranking Measure for Selecting Class Representative Words:
- We find a representative word of category c_i and add it to the set S by

Prefer words having high embedding cosine similarity with the category name

Prefer words with low distributional specificity (more general)

$$w = arg\, min_w \text{rank}_{sim}(w, c_i) \cdot \text{rank}_{spec}(w)$$
$$s.t. \quad w \notin S \quad \text{and} \quad \kappa_w > \kappa_{c_i}.$$

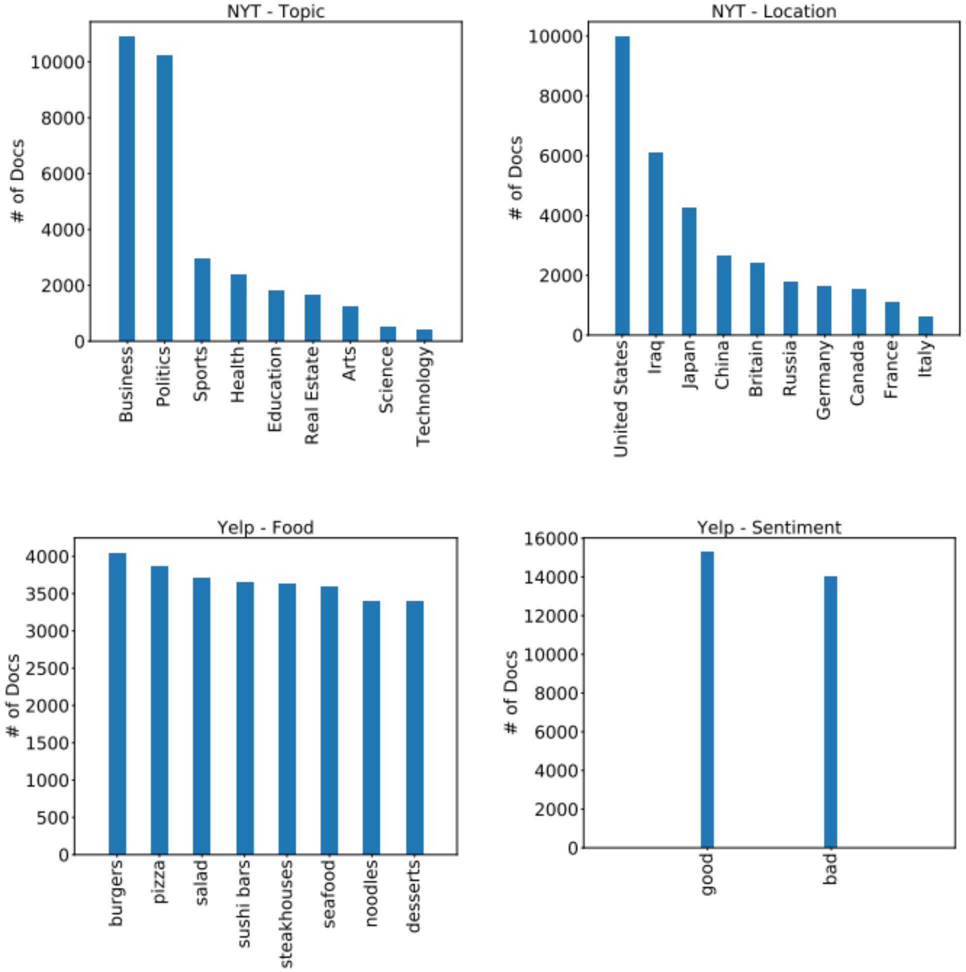
w hasn't been a representative word

w must be more specific than the category name

Quantitative Results

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

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



Dataset stat: # of docs by category name

Qualitative Results

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

Case Study: Effect of Distributional Specificity

- Coarse-to-fine topic presentation on NYT-Topic

Range of κ	Science ($\kappa_c = 0.539$)	Technology ($\kappa_c = 0.566$)	Health ($\kappa_c = 0.527$)
$\kappa_c < \kappa < 1.25\kappa_c$	scientist, academic, research, laboratory	machine, equipment, devices, engineering	medical, hospitals, patients, treatment
$1.25\kappa_c < \kappa < 1.5\kappa_c$	physics, sociology, biology, astronomy	information technology, computing, telecommunication, biotechnology	mental hygiene, infectious diseases, hospitalizations, immunizations
$1.5\kappa_c < \kappa < 1.75\kappa_c$	microbiology, anthropology, physiology, cosmology	wireless technology, nanotechnology, semiconductor industry, microelectronics	dental care, chronic illnesses, cardiovascular disease, diabetes
$\kappa > 1.75\kappa_c$	national science foundation, george washington university, hong kong university, american academy	integrated circuits, assemblers, circuit board, advanced micro devices	juvenile diabetes, high blood pressure, family violence, kidney failure

- The table lists the most similar words/phrases with each category (measured by embedding cosine similarity) from different ranges of distributional specificity
- When κ is smaller, the retrieved words have wider semantic coverage
- In our model design, if not imposing constraints on the κ, the retrieved words might be too general and do not belong to the category

Outline

- Traditional Topic Models
- Embedding-Based Discriminative Topic Mining
  - Introduction of the Task
  - CatE: Discriminative Topic Mining via Category-Name Guided Text Embedding [WWW’20]
  - JoSH: Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding [KDD’20]
- Topic Discovery with PLMs

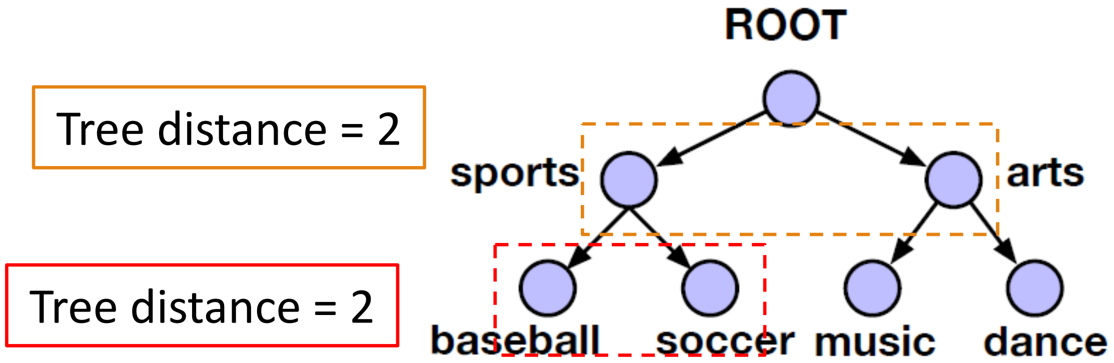
Motivation: Hierarchical Topic Mining

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

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

JoSH Embedding

- Difference from hyperbolic models (e.g., Poincare, Lorentz)
  - Hyperbolic embeddings preserve absolute tree distance (similar embedding distance => similar tree distance)
  - We do not aim to preserve the absolute tree distance, but rather use it as a relative measure



Although $d_{\text{tree}}(\text{sports}, \text{arts}) = d_{\text{tree}}(\text{baseball}, \text{soccer})$, "baseball" and "soccer" should be embedded closer than "sports" and "arts" to reflect semantic similarity.

Use tree distance in a relative manner: Since $d_{\text{tree}}(\text{sports}, \text{baseball}) < d_{\text{tree}}(\text{baseball}, \text{soccer})$, "baseball" and "sports" should be embedded closer than "baseball" and "soccer".

JoSH Text Embedding

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

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

1. Topic assignment

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

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

2. Global context

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

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

3. Local context

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

JoSH Tree Embedding

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

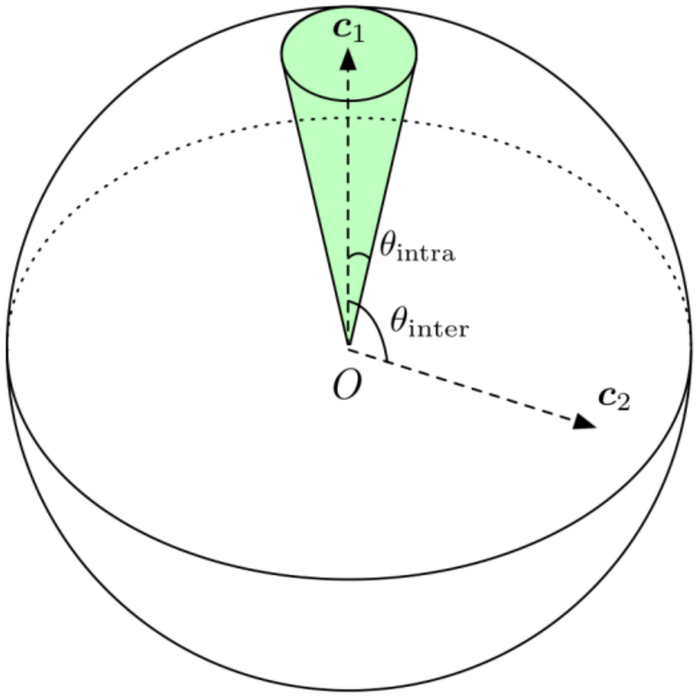
$$\mathcal{L}_{\text{intra}} = \sum_{c_i \in \mathcal{T}} \sum_{w_j \in C_i} \min(0, \boldsymbol{u}_{w_j}^\top \boldsymbol{c}_i - m_{\text{intra}}),$$

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

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

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

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



(a) Intra- & Inter-Category Configuration.

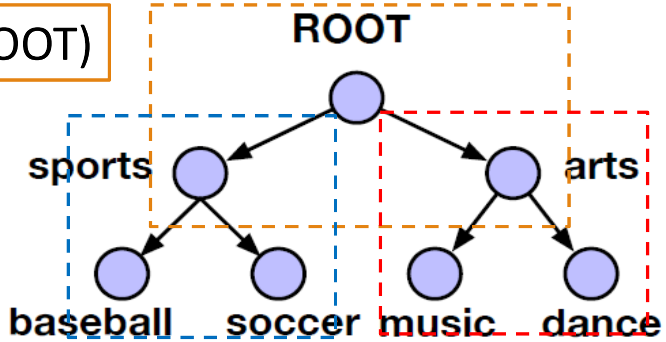
JoSH Tree Embedding

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

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

Local tree (ROOT)

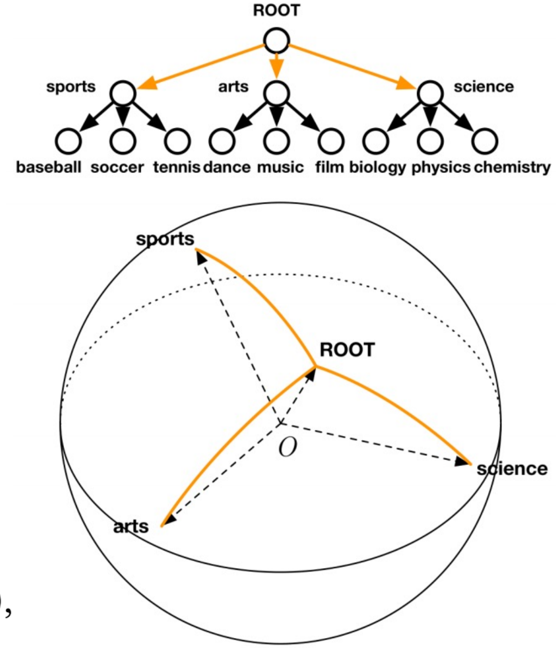
Local tree (sports)



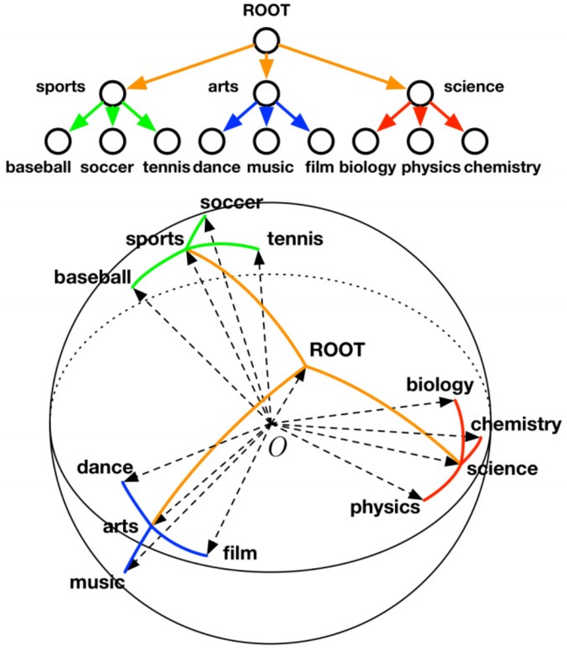
Local tree (arts)

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

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



(b) Embed First-Level Local Tree.



(c) Embed Second-Level Local Trees.

Experiments: Qualitative Results on NYT

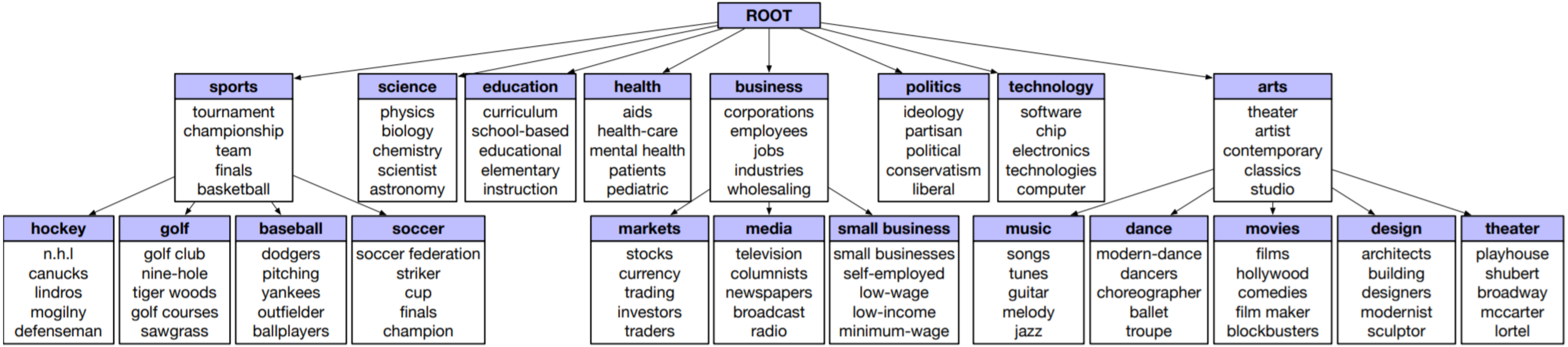
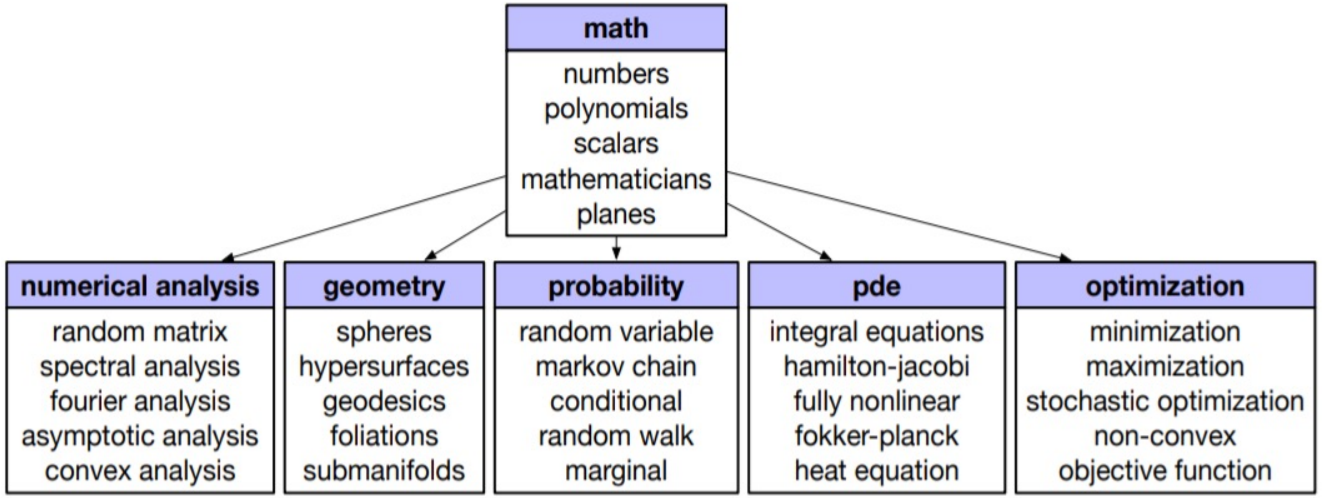
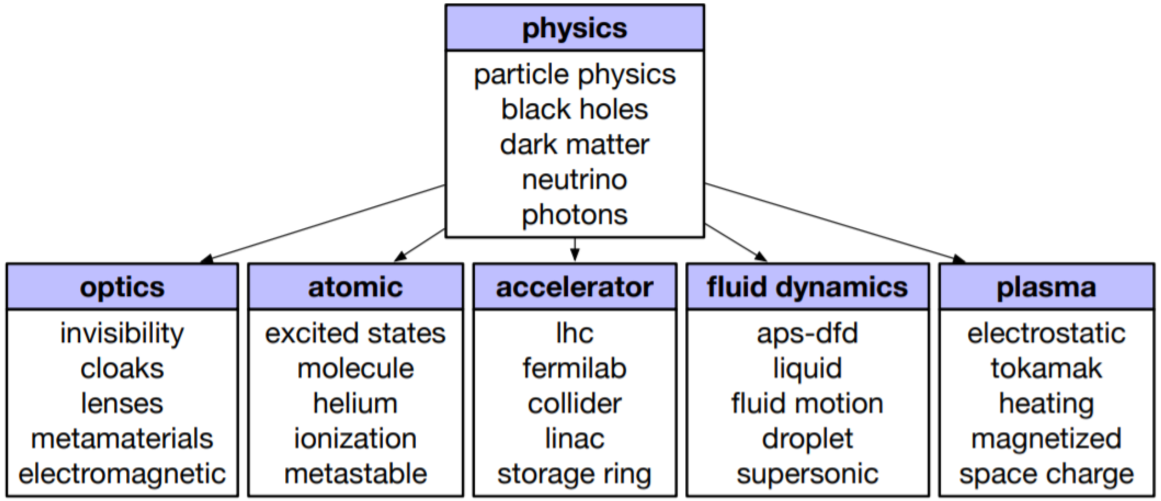


Figure 3: Hierarchical Topic Mining results on NYT.

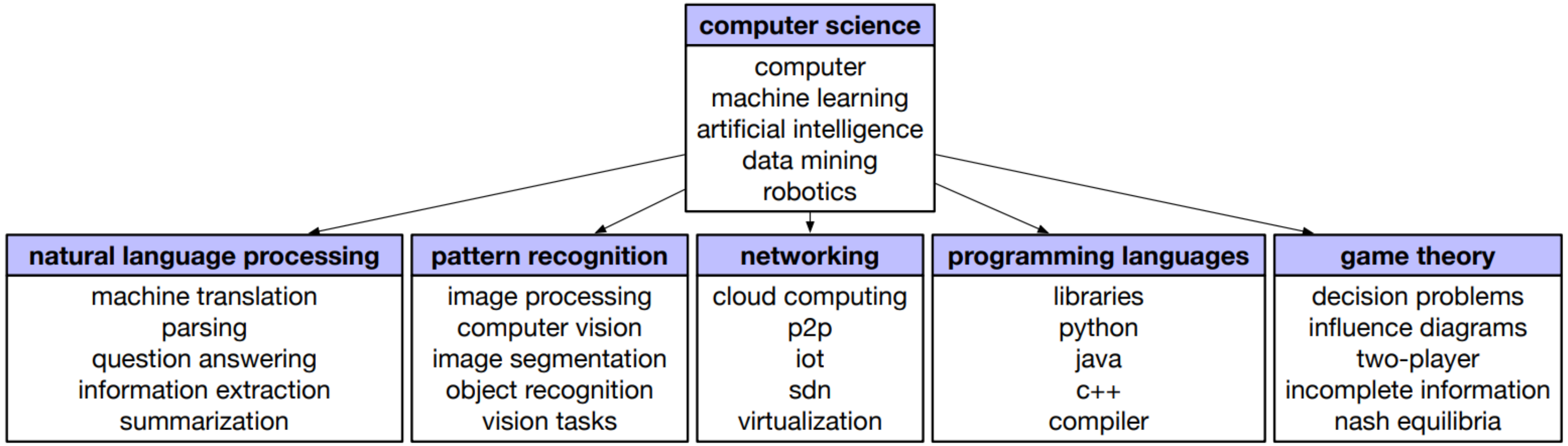
Experiments: Qualitative Results on ArXiv and Quantitative Results



(a) "Math" subtree.



(b) "Physics" subtree.



(c) "Computer Science" subtree.

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

Outline

- Traditional Topic Models
- Embedding-Based Discriminative Topic Mining
- Topic Discovery with PLMs
  - TopClus: Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations [WWW'22]
  - SeedTopicMine: Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts [WSDM'23]
  - EvMine: Unsupervised Key Event Detection from Massive Text Corpora [KDD'22]

Motivation

- Recently, pre-trained language models (LMs) have achieved enormous success in lots of tasks
  - They employ Transformer as the backbone architecture for capturing the **long-range, high-order** semantic dependency in text sequences, yielding superior representations
  - They are pre-trained on large-scale text corpora like Wikipedia, they carry **generic linguistic features** that can be generalized to almost any text-related applications
- Given the strong representation power of the contextualized embeddings, it is natural to consider simply **clustering** them as an alternative to topic models
- Topics are essentially interpreted via clusters of semantically coherent and meaningful words
- Interestingly, such an attempt has not been reported successful yet

The Challenges

- Why not naively cluster pre-trained embeddings?
- Visualization: The embedding spaces do not exhibit clearly separated clusters
- Applying K-means with a typical K (e.g., K=100) to these spaces leads to low-quality and unstable clusters

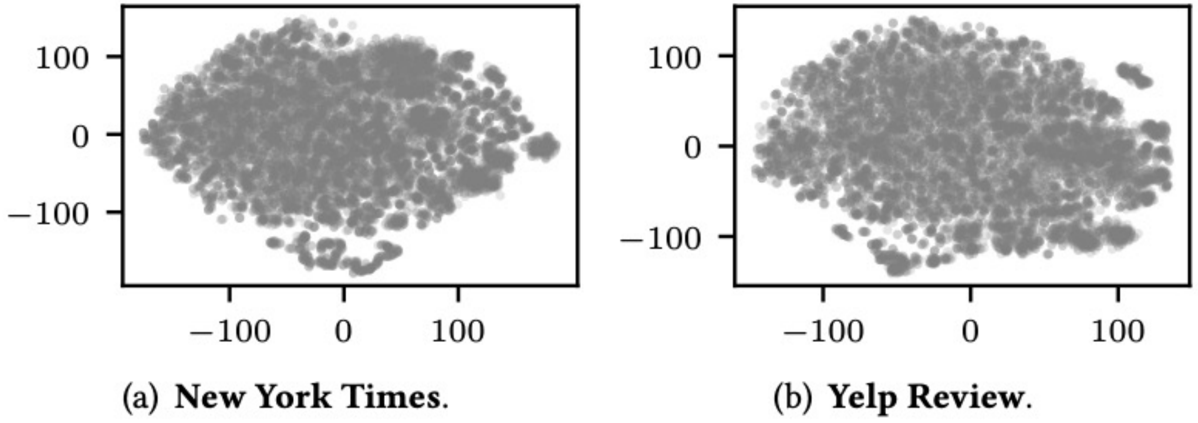


Figure 1: Visualization using t-SNE of 10,000 randomly sampled contextualized word embeddings of BERT on (a) NYT and (b) Yelp datasets, respectively. The embedding spaces do not have clearly separated clusters.

The Challenges

- Theoretically, such embedding space structure is due to **too many clusters**
- **Theorem**: The MLM pre-training objective of BERT assumes that the learned contextualized embeddings are generated from a Gaussian Mixture Model (GMM) with |V| mixture components where |V| is the vocabulary size of BERT.
- **Mismatch** between the number of clusters in the pre-trained LM embedding space and the number of topics to be discovered
  - If a smaller K ($K \ll |V|$) is used, the resulting partition will not fit the original data well, resulting in unstable and low-quality clusters
  - If a bigger K ($K \approx |V|$) is used, most clusters will contain only one unique term, which is meaningless for topic discovery

The Latent Space Model

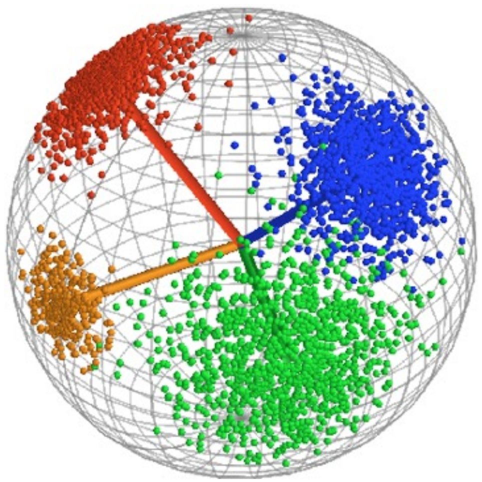
- We propose to project the original embedding space into a latent space with K clusters of words corresponding to K latent topics
- We assume that the latent space is **lower-dimensional** and **spherical**, with the following preferable properties:
  - **Spherical latent space** employs angular similarity between vectors to capture word semantic correlations, which works better than Euclidean metrics
  - **Lower-dimensional space** mitigates the “curse of dimensionality”
  - Projection from high-dimension to lower-dimension space forces the model to discard the information that is not helpful for forming topic clusters (e.g., syntactic features, “play”, “plays” and “playing” should not represent different topics)

Latent Topic Space

- We propose a generative model for the joint learning

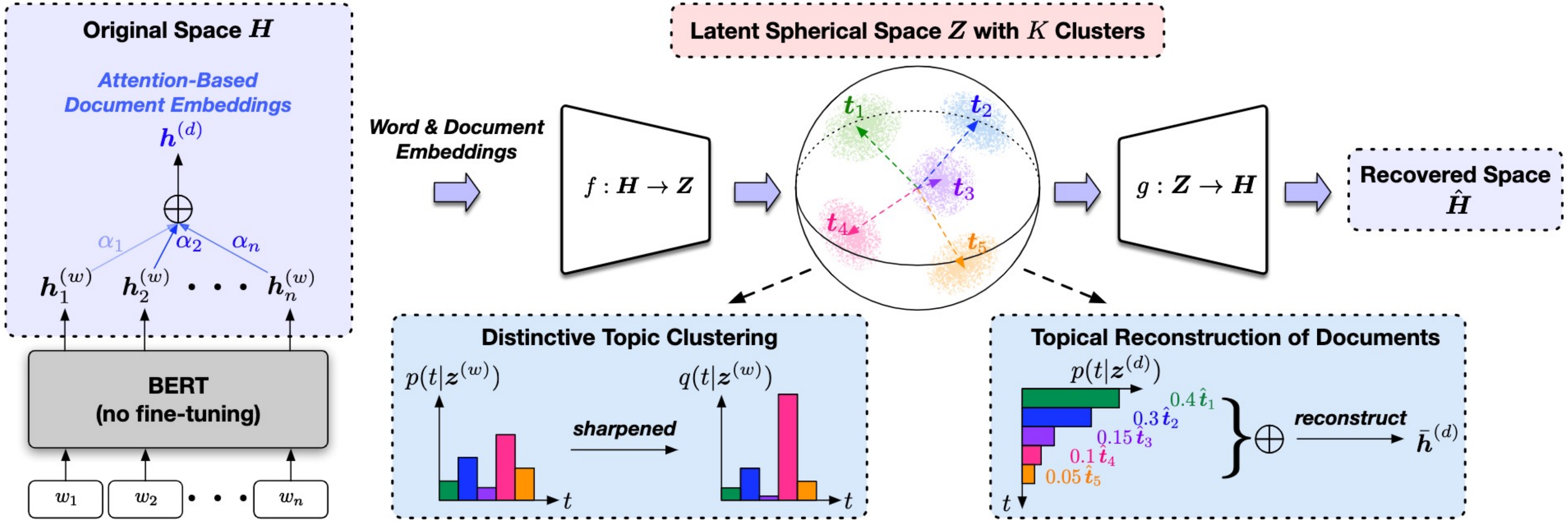
$$t_k \sim \mathrm{Uniform}(K),\ \boldsymbol{z}_i \sim \mathrm{vMF}_{d'}(\boldsymbol{t}_k, \kappa),\ \boldsymbol{h}_i = g(\boldsymbol{z}_i).$$

- A topic t is sampled from a uniform distribution over the K topics
- A latent embedding z is generated from the vMF distribution associated with topic t



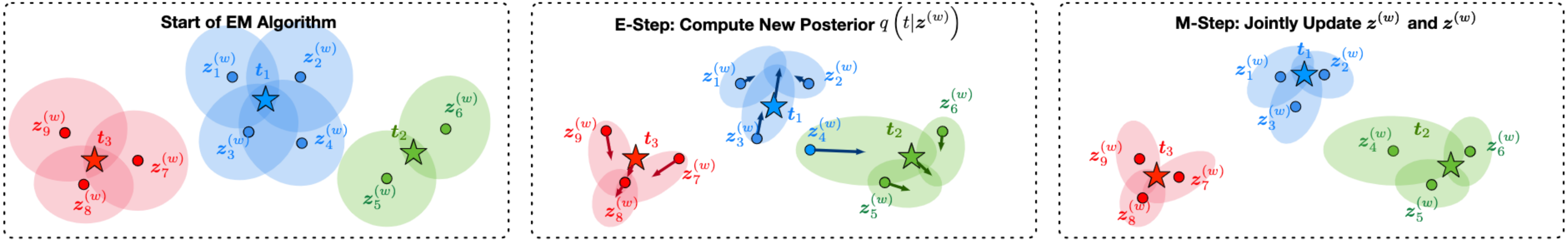
The Latent Space Model

- How to train the generative model?
  - A preservation loss that encourages the latent space to preserve the semantics of the original pre-trained LM induced embedding space **(preservation of original PLM embeddings)**
  - A reconstruction loss to ensure the learned latent topics are meaningful summaries of the documents **(Topic reconstruction of documents)**
  - A clustering loss that enforces separable cluster structures in the latent space for distinctive topic learning **(clustering)**



The Clustering Loss

- An EM algorithm, analogous to K-means
  - The E-step estimates a new cluster assignment of each word based on the current parameters
  - The M-step updates the model parameters given the cluster assignments



(a) Start of EM Algorithm. (b) E-Step. (c) M-Step.

Experiments

- Topic Discovery

Quantitative

Methods	NYT UMass	NYT UCI	NYT Int.	NYT Div.	Yelp UMass	Yelp UCI	Yelp Int.	Yelp Div.
LDA	-3.75	-1.76	0.53	0.78	-4.71	-2.47	0.47	0.65
CorEx	-3.83	-0.96	0.77	-	-4.75	-1.91	0.43	-
ETM	-2.98	-0.98	0.67	0.30	-3.04	-0.33	0.47	0.16
BERTopic	-3.78	-0.51	0.70	0.61	-6.37	-2.05	0.73	0.36
TopClus	**-2.67**	**-0.45**	**0.93**	**0.99**	**-1.35**	**-0.27**	**0.87**	**0.96**

Qualitative

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

Experiments

- Visualization

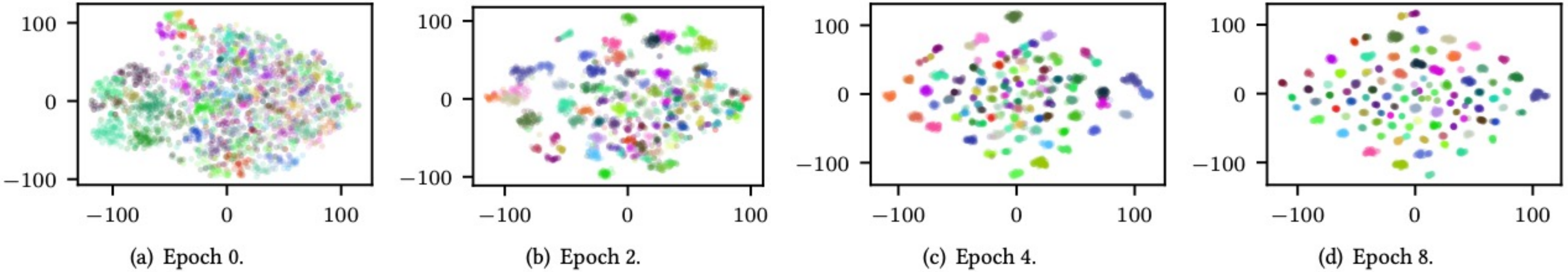


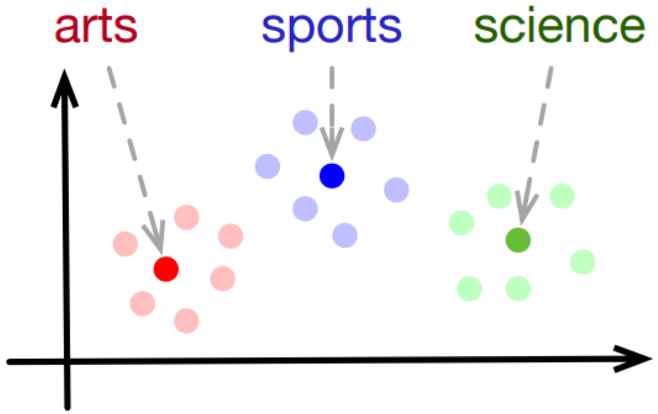
Figure 5: Visualization using t-SNE of 10, 000 randomly sampled latent embeddings during the course of TopClus training. Embeddings assigned to the same cluster are denoted with the same color. The latent space gradually exhibits distinctive and balanced cluster structure.

Outline

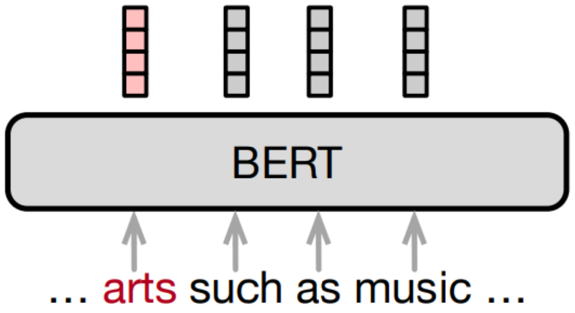
- Traditional Topic Models
- Embedding-Based Discriminative Topic Mining
- Topic Discovery with PLMs
  - TopClus: Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations [WWW’22]
  - SeedTopicMine: Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts [WSDM’23]
  - EvMine: Unsupervised Key Event Detection from Massive Text Corpora [KDD’22]

Commonly Used Context Information

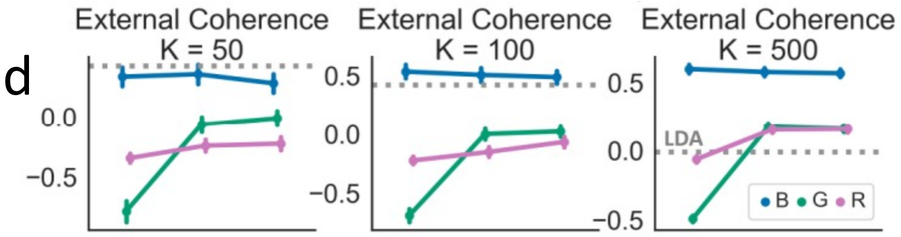
- **Context Type I - Skip-Gram Embeddings**
  - Previous slides have shown that clustering skip-gram embeddings underperforms clustering output representations of contextualized language models such as BERT in unsupervised topic modeling.



- **Context Type II - Pre-trained Language Model Representations**
  - Previous slides have shown that BERT representations suffer from the curse of dimensionality and may not form clearly separated clusters



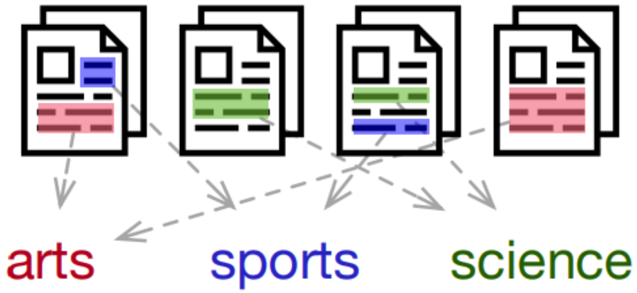
  - Thompson and Mimno [1] find that GPT-2 representations work well only if the outputs of certain layers are taken, and RoBERTa-induced topics are consistently of poor quality.



[1] Thompson, L., and Mimno, D. (2020). Topic modeling with contextualized word representation clusters. arXiv.

Commonly Used Context Information

- **Context Type III - Topic-Indicative Documents**
  - Supervised topic models [1] propose to leverage document-level training data. However, such information relies on **massive human annotation**, which is not available under the seed-guided setting.
  - A document may be **too broad** to be viewed as a context unit because each document can be relevant to multiple topics simultaneously.



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

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

SeedTopicMine: Overview

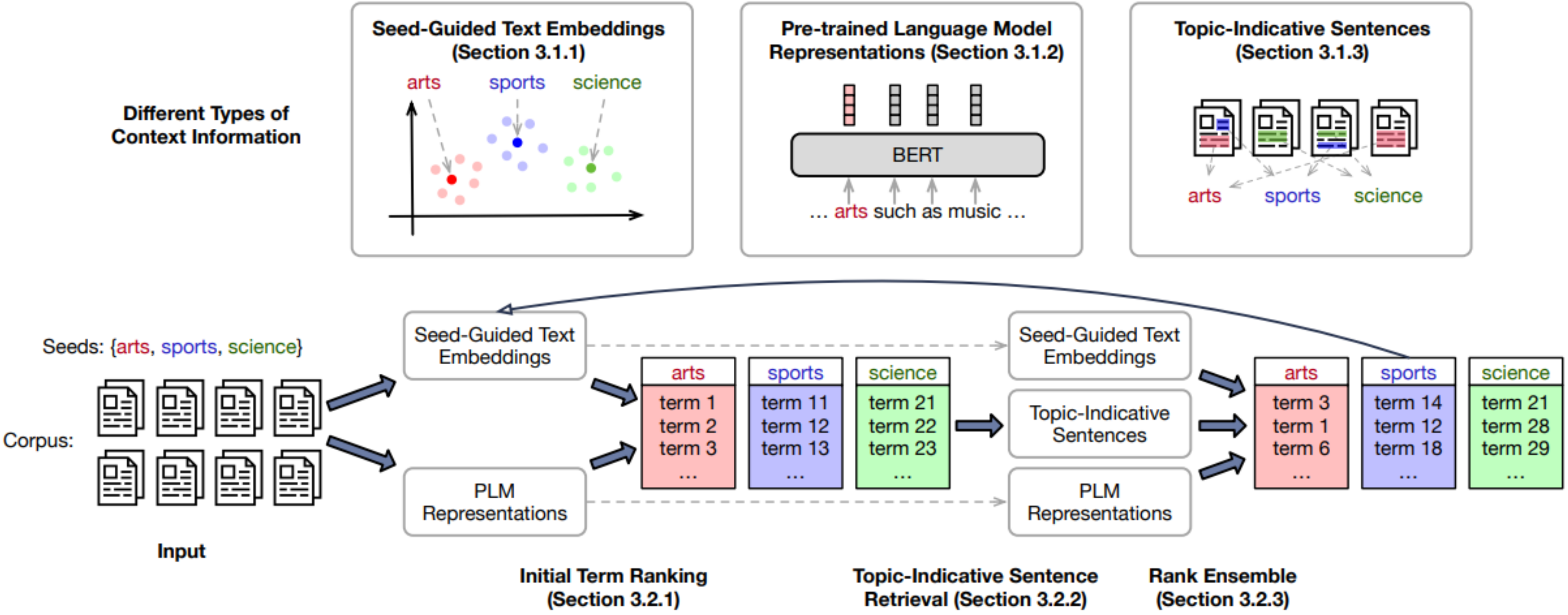
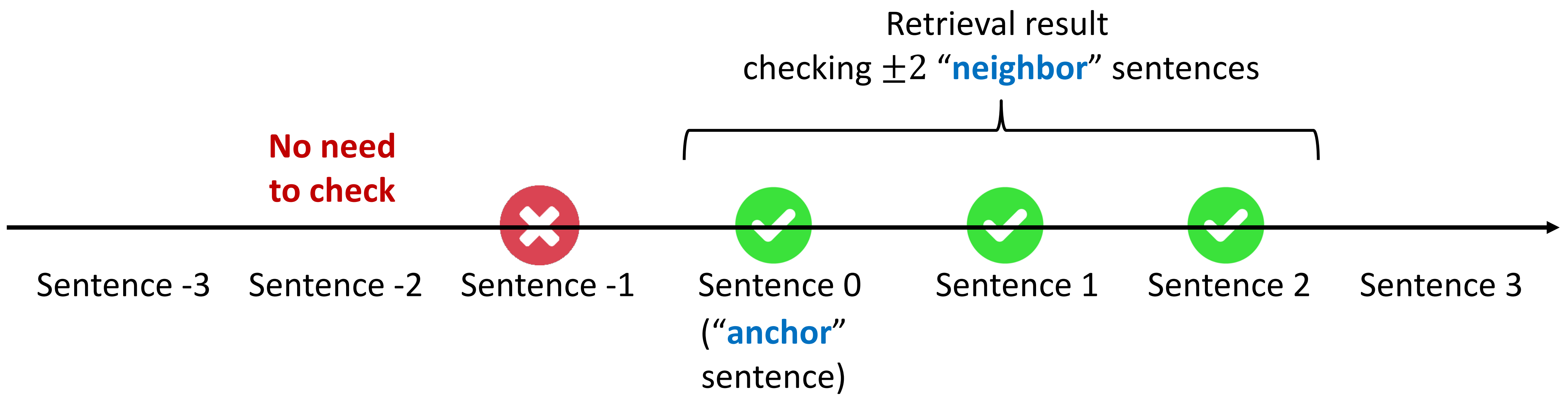


Figure 1: Overview of the SeedTopicMine framework.

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

SeedTopicMine: Topic-Indicative Sentence Retrieval

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



Quantitative Results

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

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

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

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

Qualitative Results

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

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

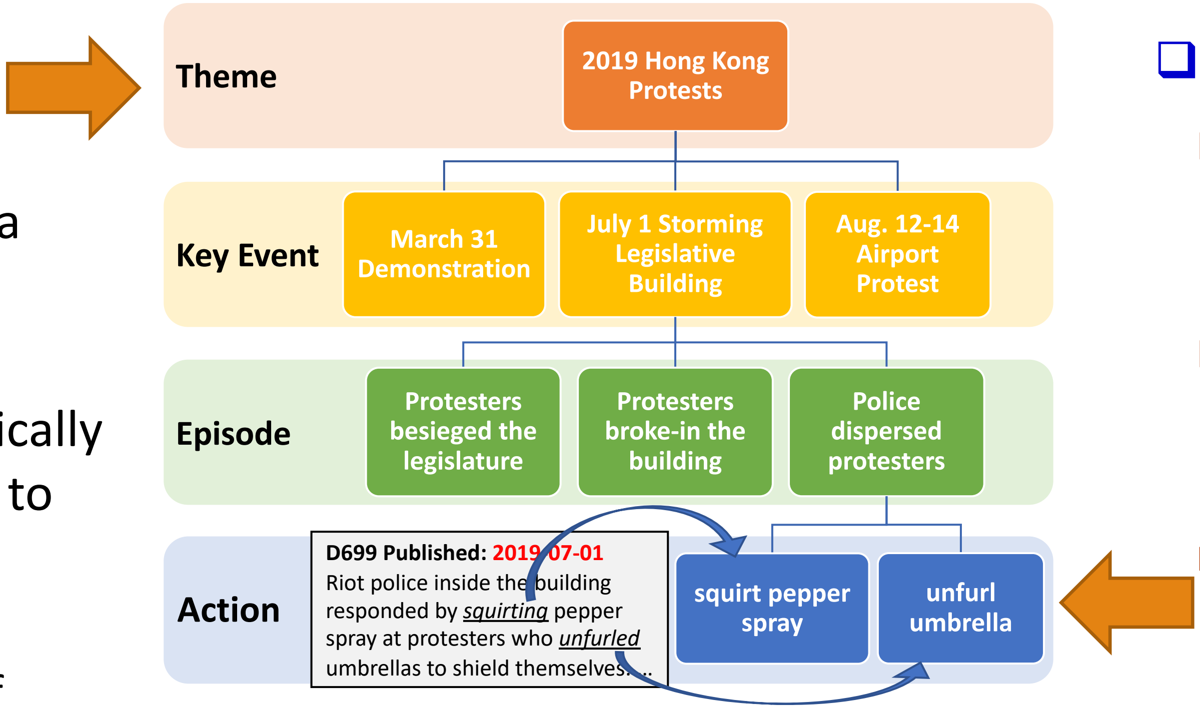
Outline

- Traditional Topic Models
- Embedding-Based Discriminative Topic Mining
- Topic Discovery with PLMs
  - TopClus: Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations [WWW'22]
  - SeedTopicMine: Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts [WSDM'23]
  - EvMine: Unsupervised Key Event Detection from Massive Text Corpora [KDD'22]

Motivation & Related Works

- Real-world events are naturally organized in a hierarchical way
  - "Big events" have more general themes and may last longer
  - "Small events" have more concrete topics and may last shorter

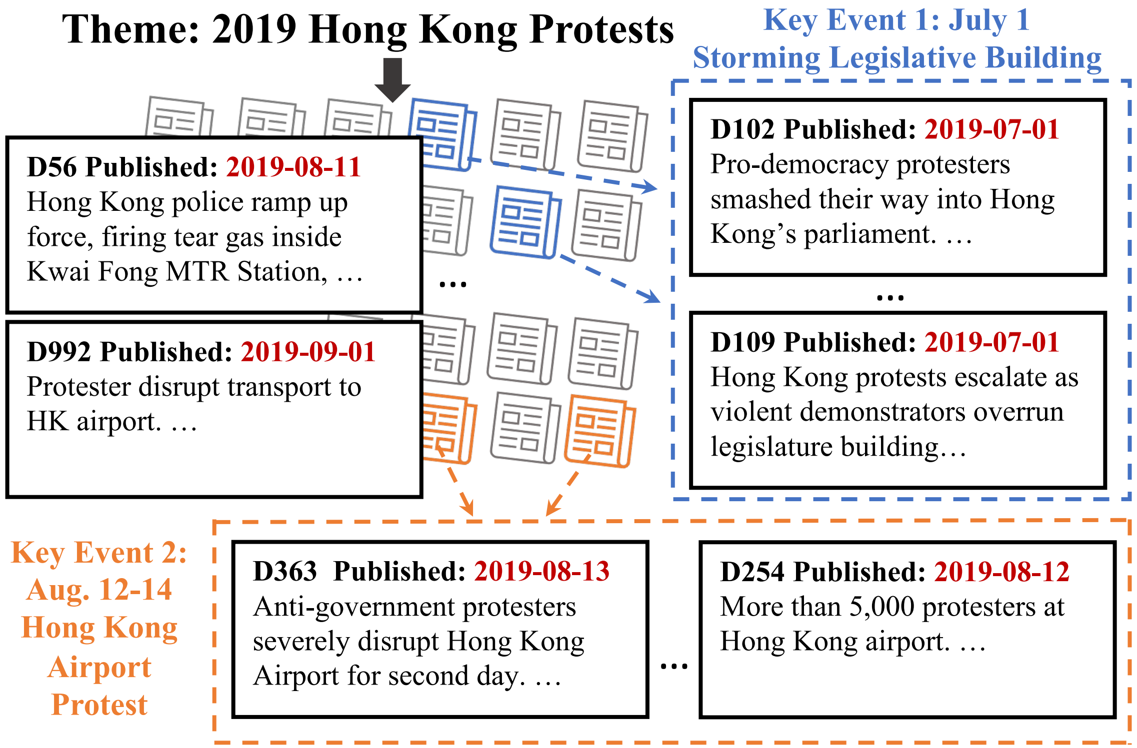
- Topic Detection and Tracking
  - Detects themes from a corpus as document clusters
  - Themes are often topically distinct and thus easy to separate
  - These methods cannot distinguish key events of similar theme



- Action Extraction
  - Extracts mention-level actions with triggers and arguments
  - Rely on human curated schema and labeled training data
  - Too fine-grained to get an overall picture of an event

A New task: Key Event Detection

- Goal:
  - Detects key events given a news corpus about one general theme
  - Key events: non-overlapping documents clusters that not necessarily exhaust the corpus
- Challenges:
  - Key events are thematically similar and temporally closer
  - Impractical to label documents for model training



Zhang, Y., Guo, F., Shen, J., & Han, J. (2022). Unsupervised Key Event Detection from Massive Text Corpora. KDD.

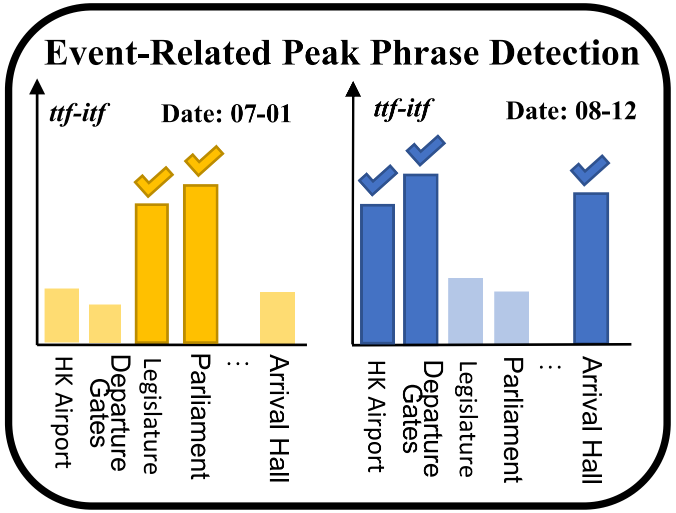
EvMine: Event-related Peak Phrase Detection

- We introduce the idea of <u>temporal term frequency – inverse time frequency</u>
- Temporal term frequency (**ttf**):
  - Measures how frequent a phrase is on a day
  - Aggregates frequencies from later days with decreasing weights to for delays and back referencing in news articles

$$\text{ttf}(p,t) = \frac{1}{n_t} \sum_{i=0}^{n_t - 1} \left(1 - \frac{i}{n_t}\right) freq_{t+i}(p),$$

- Inverse time frequency (**itf**):
  - An event-indicative phrase will only be mentioned frequently around the event happening time

$$\text{itf}(p) = \frac{\max \mathcal{T} - \min \mathcal{T} + 1}{|\{t \in \mathcal{T} | freq_t(p) > 0\}|},$$



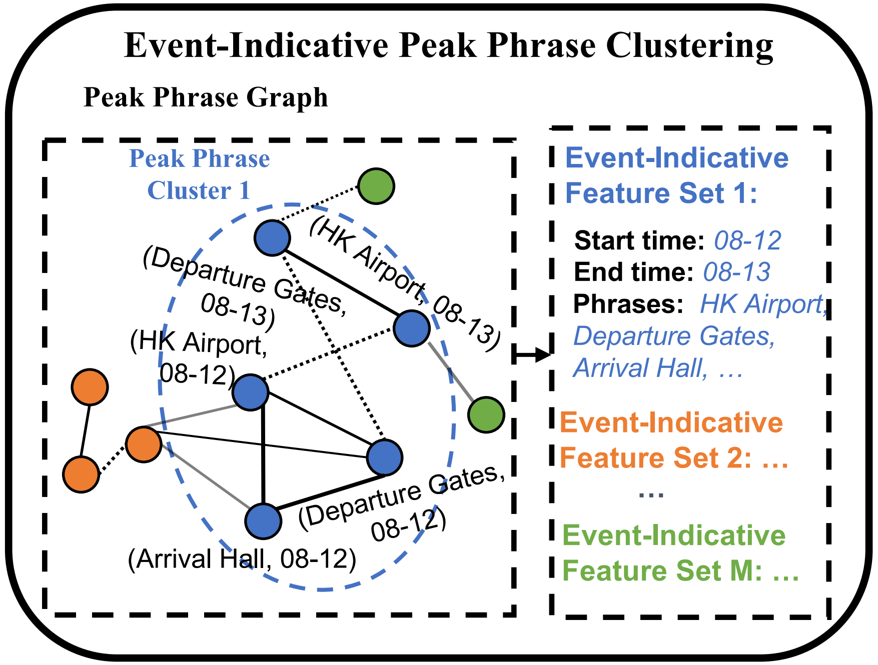
Top-ranked Peak Phrases

Phrase	Time
Hong Kong airport	2019-08-12
Victoria park	2019-08-18
legislative council	2019-07-01
Hong Kong airport	2019-08-13
…	…

49

EvMine: Event-indicative Peak Phrase Clustering

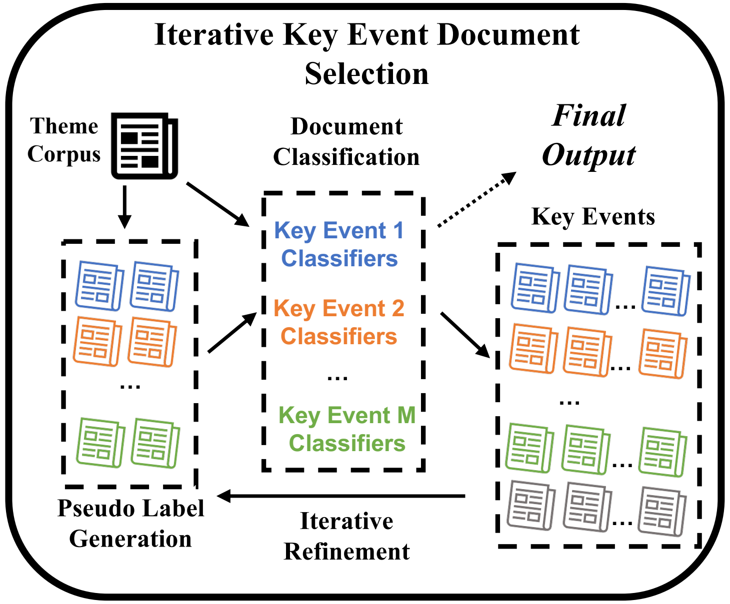
- A graph-based method to combine textual and temporal information
- Peak phrase graph construction:
  - Each node is a peak phrase $n_i = (p_i, t_i)$
  - Two types of edges:
    - Same-day peak phrases: edge weights are combination of NPMI for document-level thematic similarity and PLM-based phrase embedding similarity for semantic closeness.
    - Same-phrase consecutive-day peak phrases: connected with a constant edge weight (> 1)
- Form event-indicative feature sets with Louvain community detection algorithm



50

EvMine: Iterative Key Event Document Selection

- Pseudo Label Generation:
  - Select top-ranked documents by their number of times matched with event-indicative phrases
- Classifier Training: sampling and ensemble
  - Observation: much more negatives than positives in the corpus for each key event
  - For each key event, train multiple binary SVM classifiers by each time randomly sampling negative documents from the corpus
- Pseudo label refinement:
  - Remove current pseudo labels whose prediction score is negative
  - Enrich pseudo labels with top-n selected documents



Experiments: Quantitative Results

- Datasets:
  - **HK Protest**: We retrieve news articles about the theme “2019 Hong Kong protest”.
  - **Ebola**: We collect from English part of a multilingual news clustering dataset that about the theme “2014 Ebola Outbreak”.

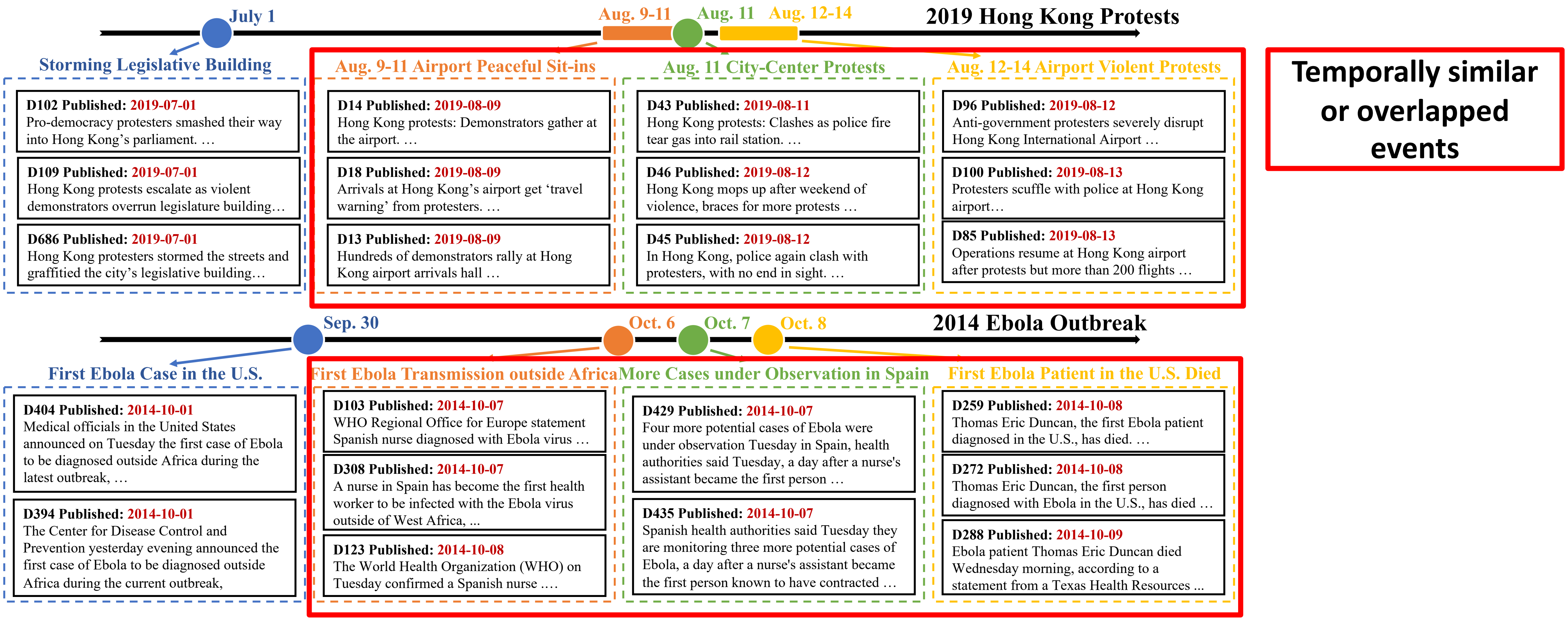
Table 1: Datasets statistics.

Dataset	# Docs	# Sents/Doc	# Words/Doc	# Events	# Docs/Event
HK Protest	1675	32.8	653.4	36	14.0
Ebola	741	25.2	554.4	17	43.6

- Quantitative results: key event detection ability based on top-k documents of each event

Methods	Ebola						HK Protest					
	5-prec	5-recall	5-F1	10-prec	10-recall	10-F1	5-prec	5-recall	5-F1	10-prec	10-recall	10-F1
newsLens [19]	0.481	**0.765**	0.591	0.524	0.647	0.579	0.352	**0.886**	0.504	0.571	0.343	0.429
Miranda et al. [21]	0.444	0.706	0.545	0.733	0.647	0.688	0.481	0.371	0.419	0.286	0.057	0.095
Staykovski et al. [35]	0.414	0.706	0.522	0.688	0.647	0.667	0.442	0.657	0.529	0.444	0.114	0.182
S-BERT	0.545	0.706	0.615	0.833	0.588	0.689	0.522	0.657	0.582	0.500	0.257	0.340
EvMine-NoClass	0.799	0.612	0.693	0.764	0.494	0.600	0.750	0.583	0.656	0.750	0.417	0.536
EvMine-COOC	**0.846**	0.647	0.733	**0.909**	0.588	0.714	0.815	0.611	0.698	0.807	0.431	0.561
EvMine-NoLM	0.784	0.659	0.715	0.865	0.635	0.732	0.905	0.608	0.728	0.942	0.453	0.612
EvMine-Single	0.814	0.671	0.735	0.872	0.635	0.734	0.916	0.636	0.751	0.958	0.458	0.620
EvMine	0.829	0.682	**0.748**	0.883	**0.653**	**0.751**	**0.934**	0.664	**0.776**	**0.960**	**0.464**	**0.625**

Experiments: Qualitative Results



Experiments: Qualitative Results

Thematically similar events: Airport Protest

Thematically similar events: First Case in the U.S.

54

References

- Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2003). Hierarchical topic models and the nested Chinese restaurant process. NIPS.
- Blei, D. M., & McAuliffe, J. D. (2007). Supervised topic models. NIPS.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research.
- Mimno, D., Li, W., & McCallum, A. (2007). Mixtures of hierarchical topics with pachinko allocation. ICML.
- Jagarlamudi, J., Daumé III, H., & Udupa, R. (2012). Incorporating lexical priors into topic models. EACL.
- Meng, Y., Huang, J., Wang, G., Wang, Z., Zhang, C., Zhang, Y., & Han, J. (2020). Discriminative topic mining via category-name guided text embedding. WWW.
- Meng, Y., Zhang, Y., Huang, J., Zhang, Y., Zhang, C., & Han, J. (2020). Hierarchical topic mining via joint spherical tree and text embedding. KDD.
- Meng, Y., Zhang, Y., Huang, J., Zhang, Y., & Han, J. (2022). Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations. WWW.
- Sia, S., Dalmia, A., & Mielke, S. J. (2020). Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! EMNLP.
- Zhang, Y., Guo, F., Shen, J., & Han, J. (2022). Unsupervised Key Event Detection from Massive Text Corpora. KDD.
- Zhang, Y., Zhang, Y., Michalski, M., Jiang, Y., Meng, Y., & Han, J. (2023). Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts. WSDM.