Part II: Mining Topic Structures: Unsupervised and Seed-Guided Topic Discovery

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

Yu Zhang, Yunyi Zhang, Jiawei Han

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

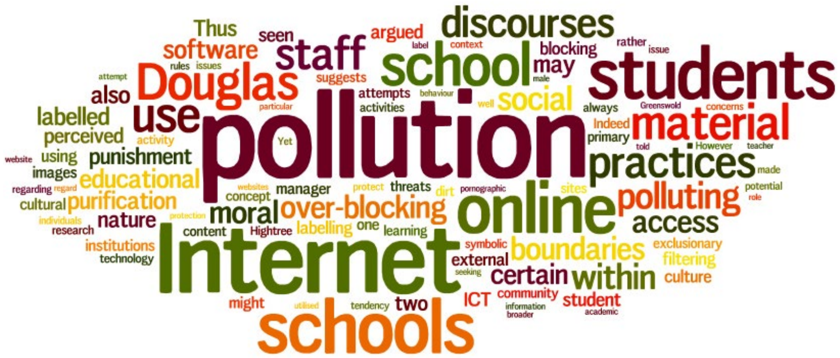
Mar 29, 2023

Outline

- Unsupervised Topic Discovery
  - Topic Modeling
  - Clustering-Based Topic Discovery
- Seed-Guided Topic Discovery

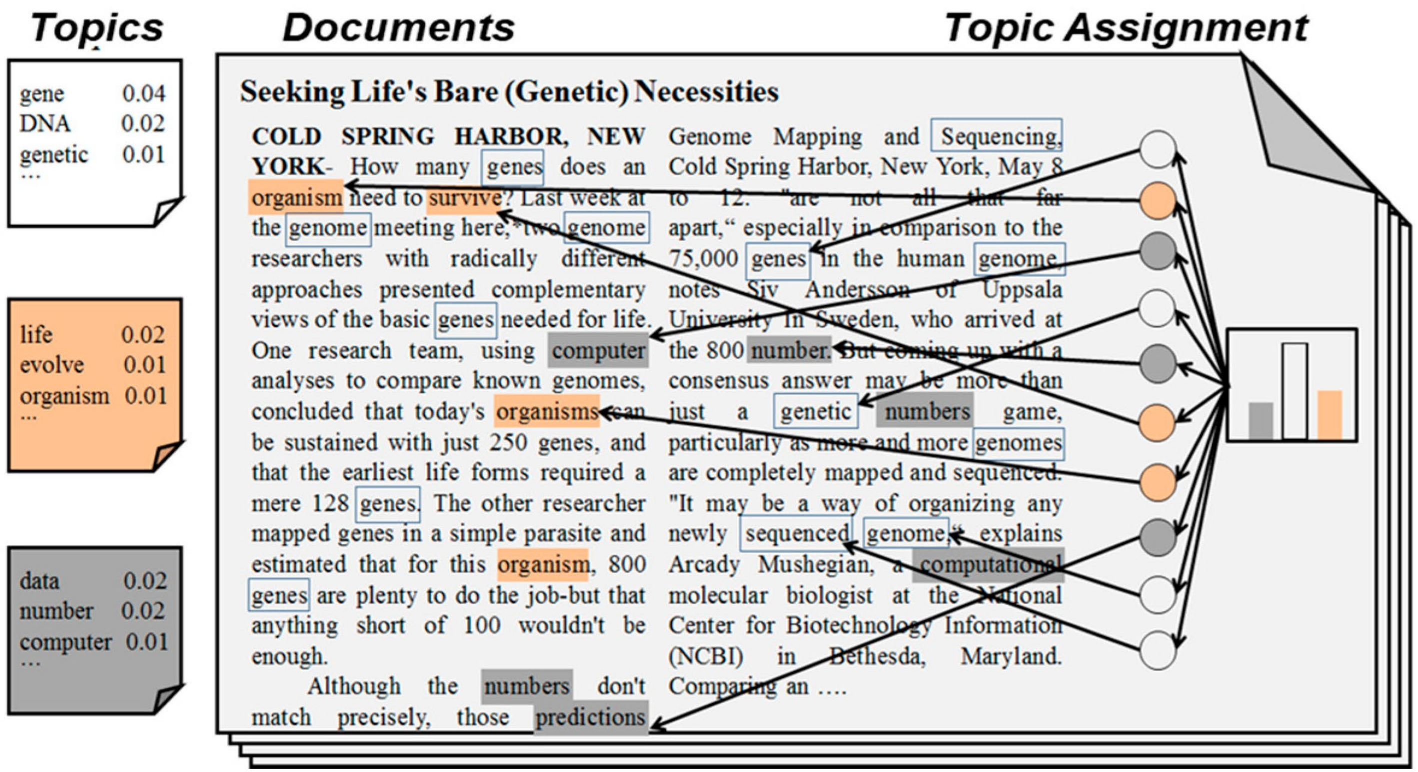
Topic Modeling: Introduction

- 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

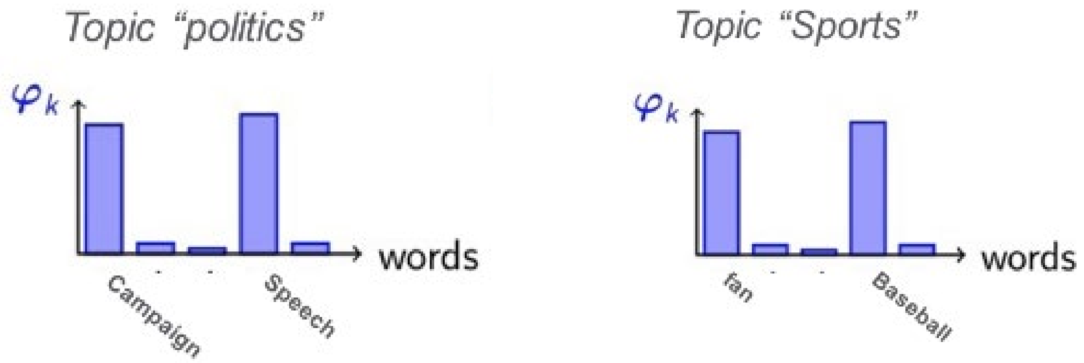
	Topic 1	…	Topic n
Word 1	0.09	…	0.06
…	….	…	…
Word n	0.08	…	0.01

Document-Topic Matrix

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

Latent Dirichlet Allocation (LDA): Overview

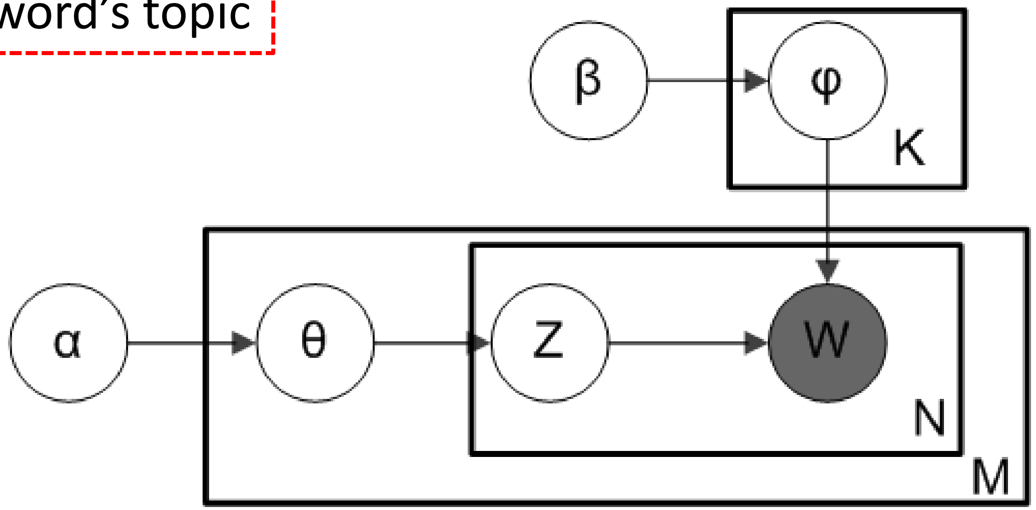
- Each document is represented as a mixture of various topics
  - E.g., a news document may be 40% on politics, 50% on economics, and 10% on sports
- Each topic is represented as a probability distribution over words
  - E.g., the distribution of “politics” vs. “sports” might be like:



- 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)

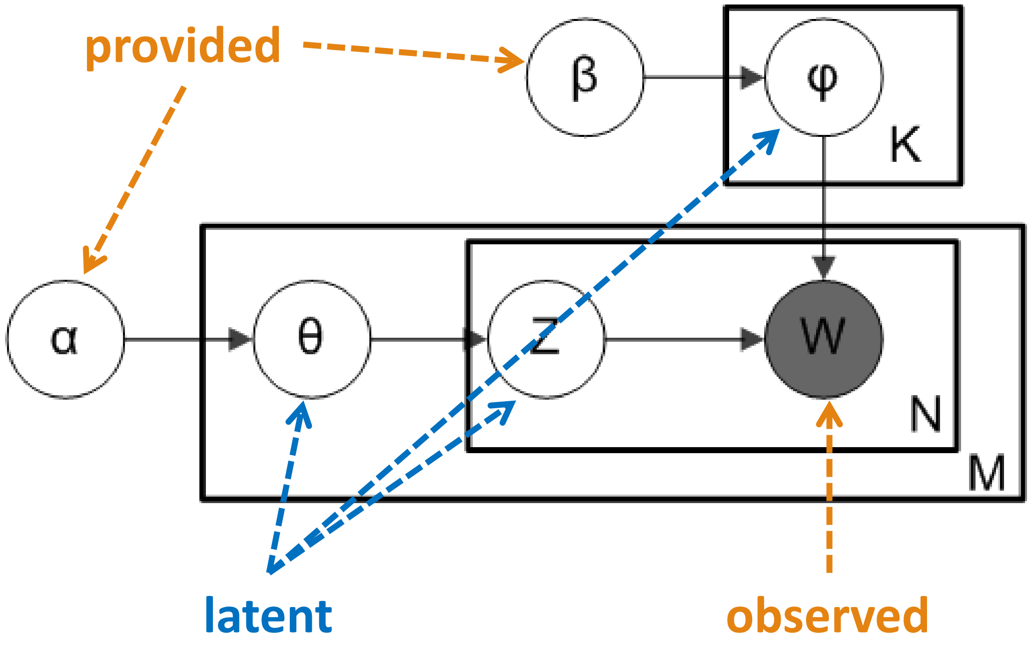
LDA: A Generative Model

- Formulating the statistical relationship between words, documents and latent topics as a generative process describing how documents are created:
  - For the i-th document, choose $\boxed{\theta_i} \sim \mathrm{Dir}(\alpha)$ document's topic distribution
  - For the k-th topic, choose $\boxed{\varphi_k} \sim \mathrm{Dir}(\beta)$ topic's word distribution
  - For the j-th word in the i-th document,
    - choose topic $\boxed{z_{i,j}} \sim \mathrm{Categorical}(\theta_i)$ word's topic
    - choose a word $w_{i,j} \sim \mathrm{Categorical}(\varphi_{z_{i,j}})$



LDA: Inference

- Learning the parameters of LDA
- 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
  - …



Outline

- Unsupervised Topic Discovery
  - Topic Modeling
  - Clustering-Based Topic Discovery
    - Directly Clustering of Text Embeddings [EMNLP'19]
    - TopClus: Latent Space Clustering of PLM Representations [WWW'22]
- Seed-Guided Topic Discovery

Clustering-Based Topic Discovery

- Topic modeling frameworks use **bag-of-words** features (i.e., only word counts in documents matter; word ordering is ignored)
- As we know, distributed text representations (text embeddings and language models) model better sequential information in text
- Can we take advantage of advanced text representations for topic discovery, as an alternative to topic modeling?   This leads to **Word Embedding + Clustering**
- **Word Embedding + Clustering:** Cast "topics" as clusters of word types — similar to taking the top-ranked words from each topic's distribution in topic modeling
  - How to obtain word clusters? Run clustering algorithms on word embeddings
  - Since the text embedding space captures word semantic similarity (i.e., high vector similarity implies high semantic similarity), using distance-based clustering algorithms (like K-means) will naturally group semantically similar words into the same cluster

Clustering-Based Topic Discovery: A benchmark study

- Clustering algorithms:
  - k-means (KM)
  - Gaussian Mixture Models (GMM)
- Embeddings:
  - Word2Vec
  - GloVe
  - fastText
  - Spherical text embedding
  - ELMo
  - BERT

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.

Clustering-Based Topic Discovery: Word Frequency

- One thing to consider is that text embeddings do not explicitly encode frequency information, which is important for topic discovery (i.e., more frequent words in the corpus may be more representative)
- Two ways to incorporate frequency information
  - Weighted clustering: Frequent words weigh more when computing cluster centroids
  - Rerank words in clusters: Rerank terms by frequency in each cluster when selecting representative terms

Clustering-Based Topic Discovery: Results

- Use k-means (KM)/Gaussian Mixture Models (GMM) as clustering algorithm and use Spherical text embedding/BERT as representations leads to comparable results with LDA

weighted clustering + reranking

	Reuters								20 Newsgroups							
	$\diamond$		$\diamond^w$		$\diamond_r$		$\diamond_r^w$		$\diamond$		$\diamond^w$		$\diamond_r$		$\diamond_r^w$	
	KM	GMM	KM	GMM	KM	GMM	KM	GMM	KM	GMM	KM	GMM	KM	GMM	KM	GMM
Word2vec	-0.39	-0.47	-0.21	-0.09	0.02	0.01	0.03	0.08	-0.21	-0.10	-0.11	0.13	0.18	0.16	0.19	0.20
ELMo	-0.73	-0.55	-0.43	0.00	-0.10	-0.08	-0.02	0.06	-0.56	-0.13	-0.38	0.18	0.13	0.14	0.16	0.19
GloVe	-0.67	-0.59	-0.04	0.01	-0.27	-0.03	0.01	0.05	-0.18	-0.12	0.06	0.24	0.22	0.23	0.23	0.23
Fasttext	-0.68	-0.70	-0.46	-0.08	0.00	0.00	0.06	0.11	-0.32	-0.20	-0.18	0.21	0.24	0.23	0.25	0.24
Spherical	-0.53	-0.65	-0.07	0.09	0.01	-0.05	0.10	0.12	-0.05	-0.24	0.24	0.23	0.25	0.22	0.26	0.24
BERT	-0.43	-0.19	-0.07	0.12	0.00	-0.01	0.12	0.15	0.04	0.14	0.25	0.25	0.17	0.19	0.25	0.25
average	-0.57	-0.52	-0.21	0.01	-0.06	-0.03	0.05	0.10	-0.21	-0.11	-0.02	0.21	0.20	0.20	0.23	0.23
std. dev.	0.14	0.18	0.19	0.09	0.12	0.03	0.05	0.04	0.21	0.13	0.25	0.05	0.04	0.04	0.04	0.02

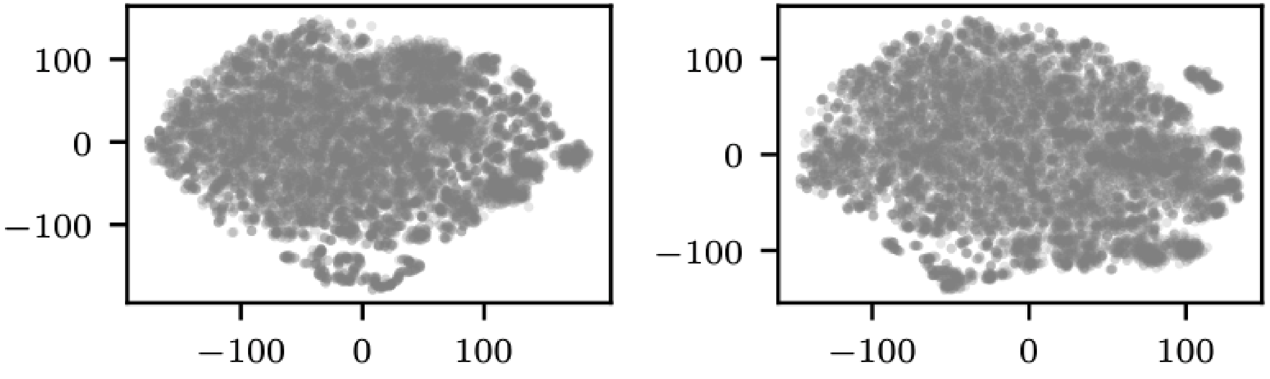
Table 1: NPMI Results (higher is better) for pre-trained word embeddings and k-means (KM), and Gaussian Mixture Models (GMM). $\diamond^w$ indicates weighted and $\diamond_r$ indicates reranking of top words. For Reuters (left table), LDA has an NPMI score of 0.12, while GMM_r^w BERT achieves 0.15. For 20NG (right), both LDA and KM_r^w Spherical achieve a score of 0.26. All results are averaged across 5 random seeds.

Exploring Pre-Trained Language Models

- 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

Naively Clustering Pre-trained Embeddings?

- 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



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

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.

Meng, Y., Zhang, Y., Huang, J., Zhang, Y., & Han, J. (2022). Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations. WWW.

Root of the Challenges: Too Many Clusters

- 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 ($K << |V|$) 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

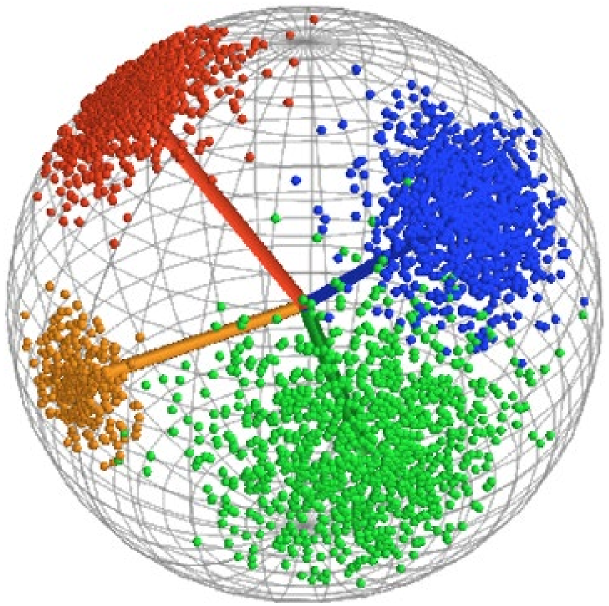
- We propose to project the original embedding space into a latent space with K clusters of words corresponding to K latent topics
- We assume that the latent space is **lower-dimensional** and **spherical**, with the following preferable properties:
  - **Spherical latent space** employs angular similarity 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

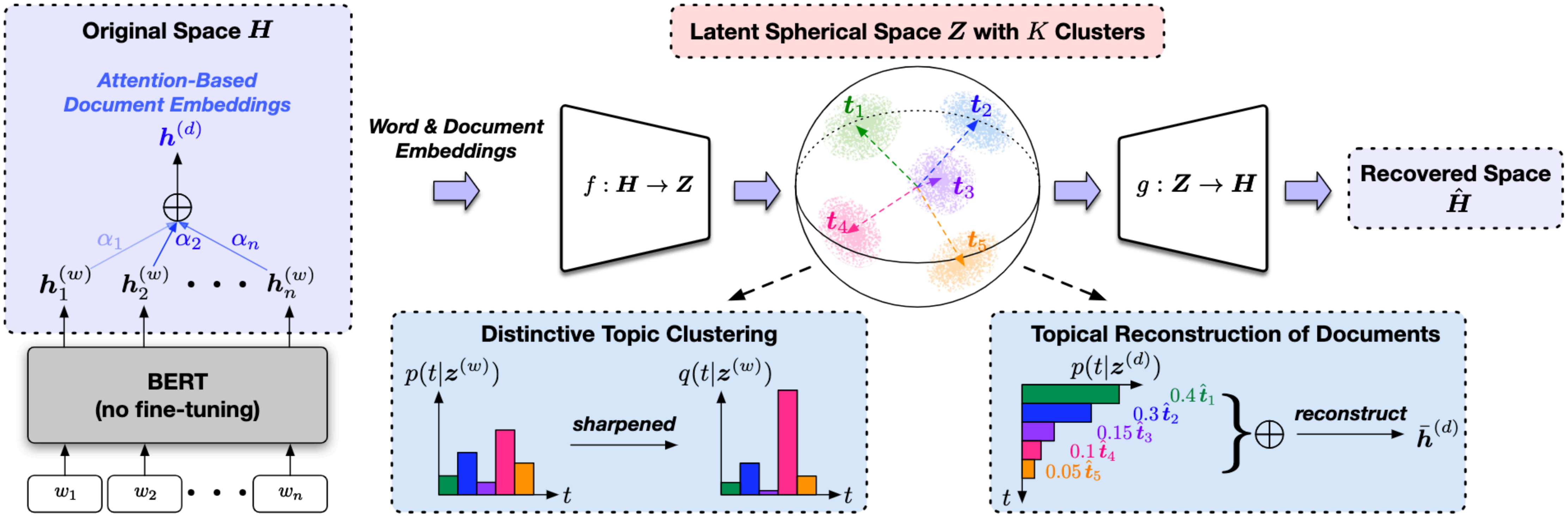
$$t_k \sim \text{Uniform}(K),\ \boldsymbol{z}_i \sim \text{vMF}_{d'}(\boldsymbol{t}_k, \kappa),\ \boldsymbol{h}_i = g(\boldsymbol{z}_i).$$

- 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



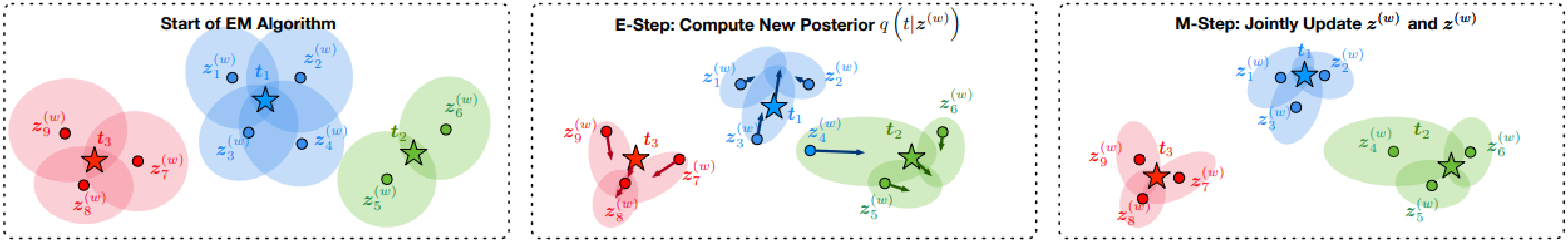
The Latent Space Model

- How to train the generative model?
  - **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



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



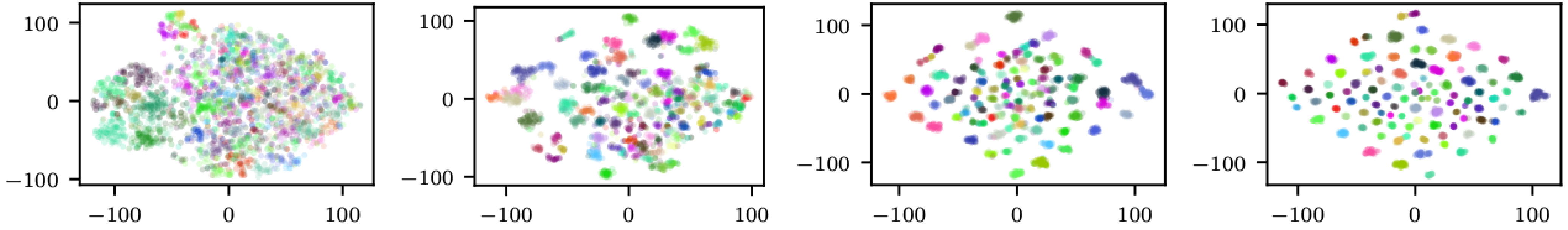
(a) Start of EM Algorithm. (b) E-Step. (c) M-Step.

Quantitative Results and Visualization

- Performance comparison

Methods	NYT				Yelp			
	UMass	UCI	Int.	Div.	UMass	UCI	Int.	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**

- Visualization



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

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.

Qualitative Results

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

Outline

- Unsupervised Topic Discovery
- Seed-Guided Topic Discovery
  - 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]
  - SeedTopicMine: Integrating Multiple Types of Contexts [WSDM'23]

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

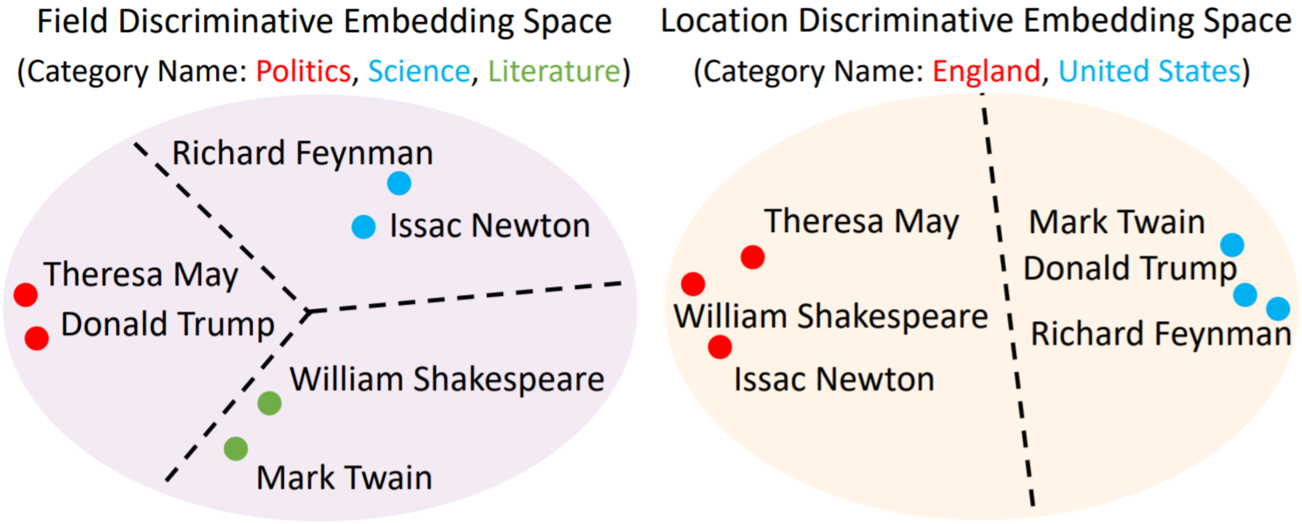
- **Discriminative Topic Mining:** Given a text corpus and a set of **category names**, retrieve a set of terms that **exclusively belong to** each category
  - E.g., given c_1: "The United States", c_2: "France", c_3: "Canada"
    - Yes to "Ontario" under c_3: (a province in Canada and exclusively belongs 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

- Unsupervised Topic Discovery
- Seed-Guided Topic Discovery
  - 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]
  - SeedTopicMine: Integrating Multiple Types of Contexts [WSDM'23]

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

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

$$p(d \mid c_d) \propto p(c_d \mid d)p(d) \propto p(c_d \mid d) \propto \prod_{w \in d} p(c_d \mid w),$$

Decompose into word-topic distribution

- Introducing specificity

> **Definition 2** (Word Distributional Specificity). We assume there is a scalar $\kappa_w \ge 0$ 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)

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

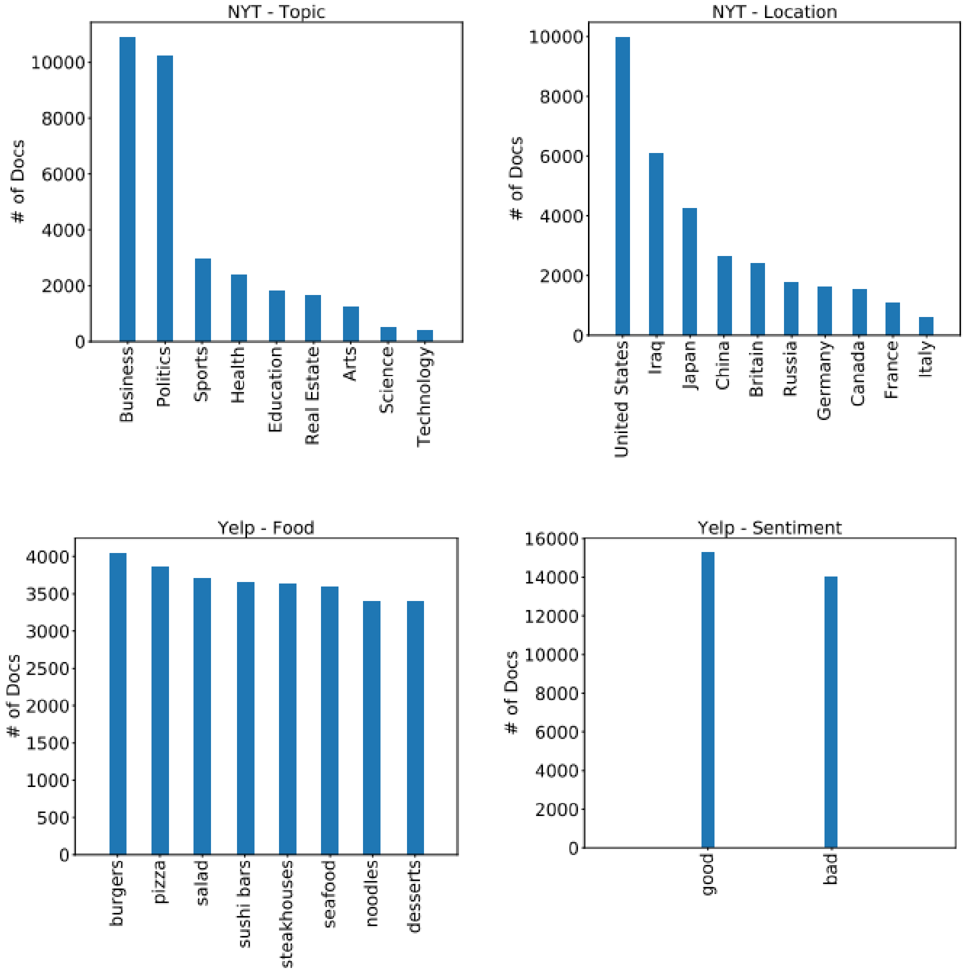
w hasn't been a representative word

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

- Unsupervised Topic Discovery
- Seed-Guided Topic Discovery
  - 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]
  - SeedTopicMine: Integrating Multiple Types of Contexts [WSDM'23]

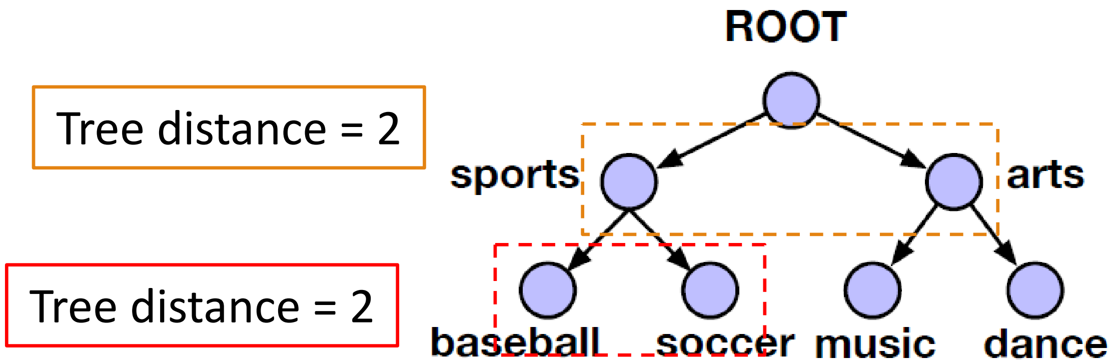
Motivation: Hierarchical Topic Mining

- 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

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

  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

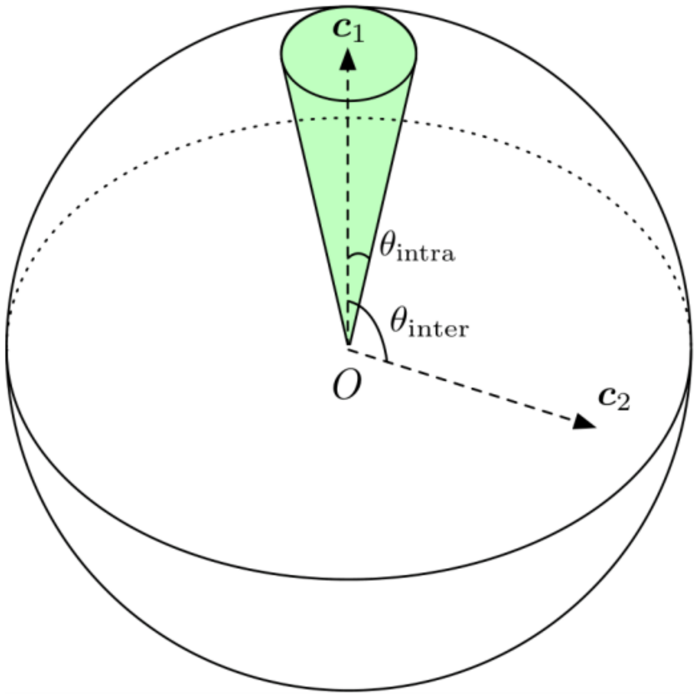
$$\mathcal{L}_{\text{intra}} = \sum_{c_i \in \mathcal{T}} \sum_{w_j \in C_i} \min(0, \boldsymbol{u}_{w_j}^{\top} \boldsymbol{c}_i - m_{\text{intra}}),$$

- **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.

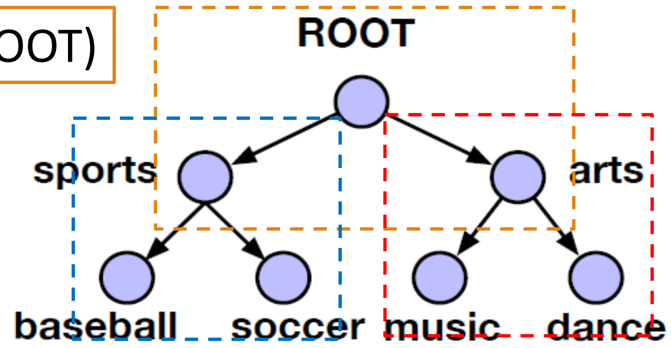
JoSH Tree Embedding

- **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)

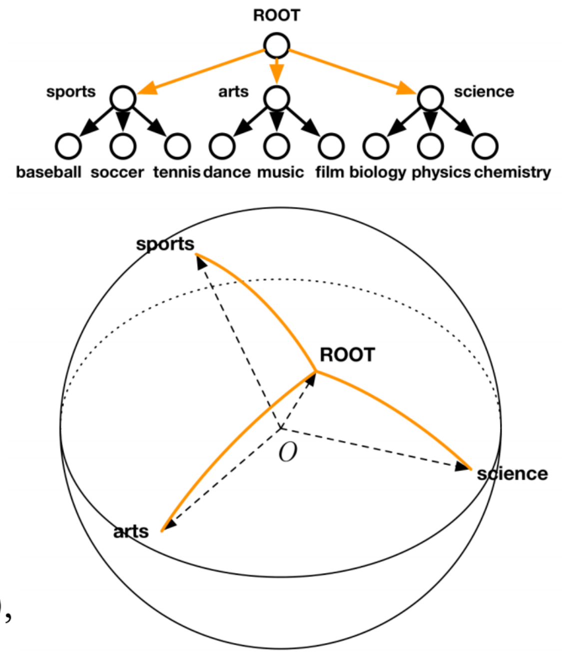
Local tree (sports)



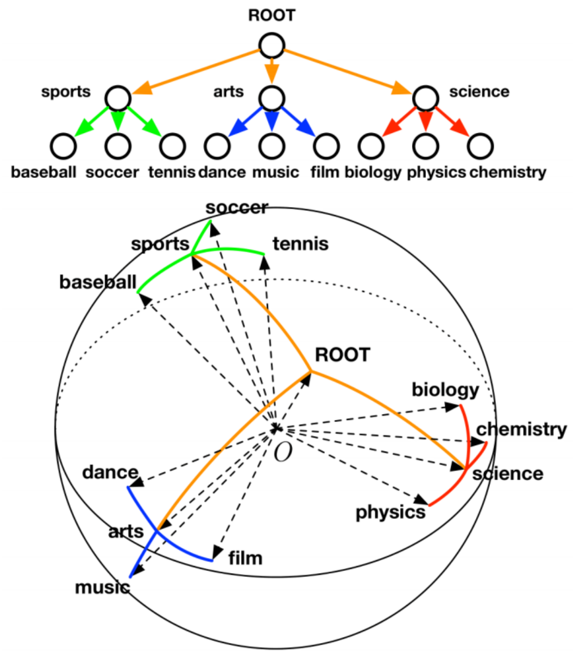
Local tree (arts)

- **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

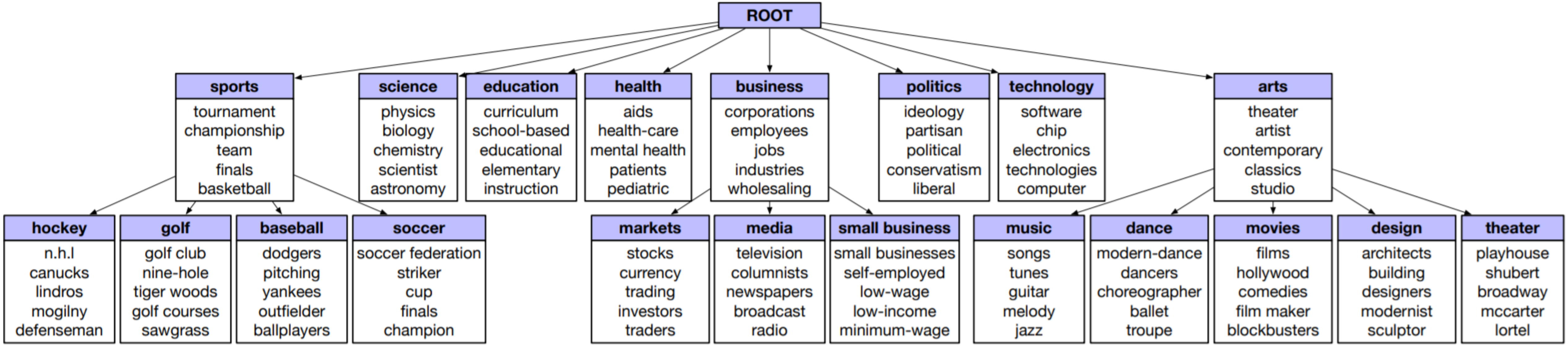
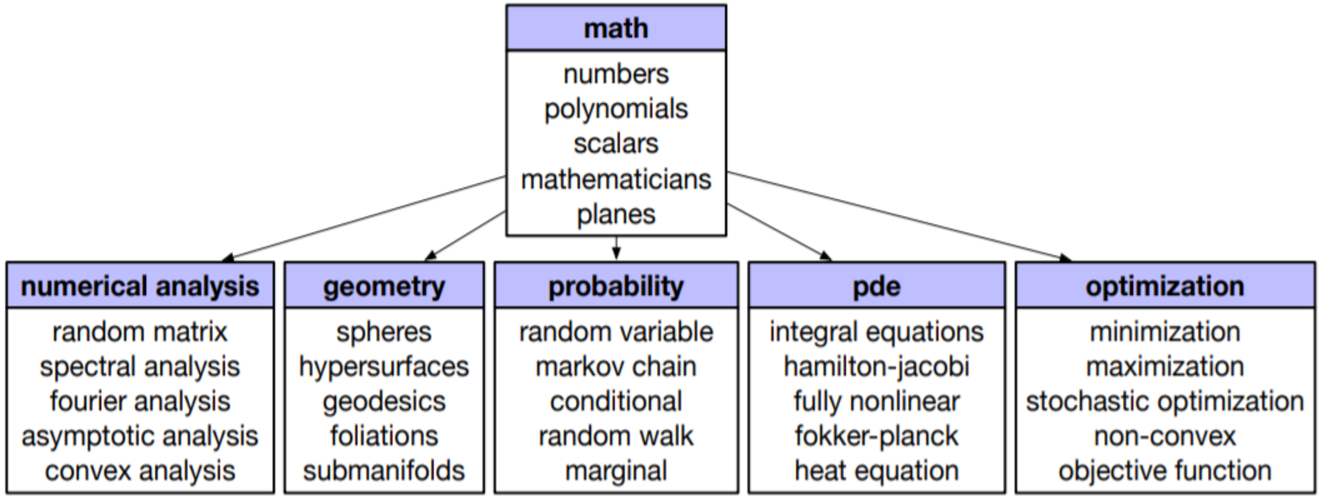
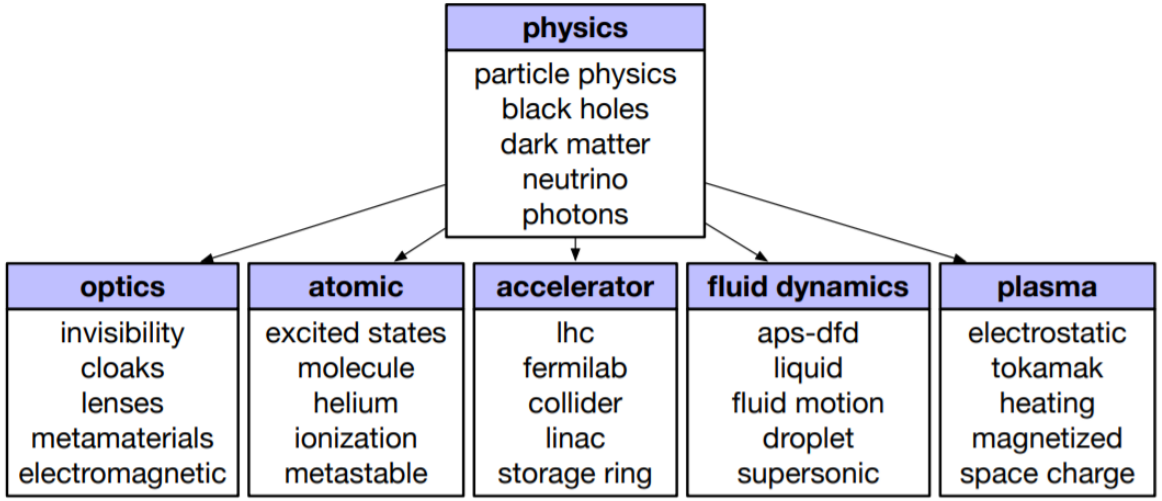


Figure 3: Hierarchical Topic Mining results on NYT.

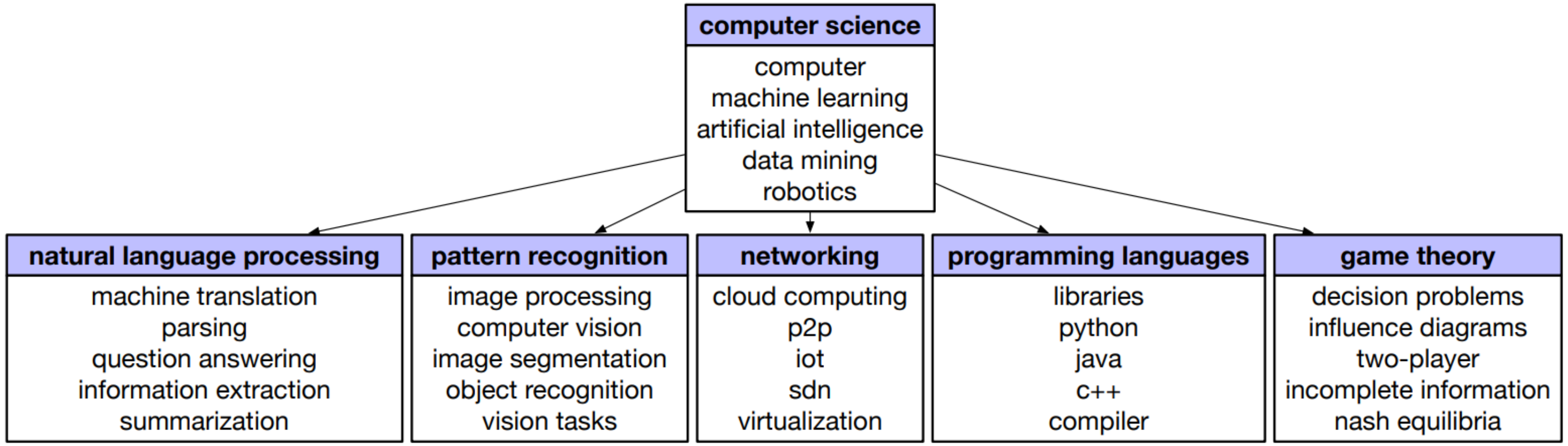
Experiments: Qualitative Results on ArXiv and Quantitative Results



(a) “Math” subtree.



(b) “Physics” subtree.



(c) “Computer Science” subtree.

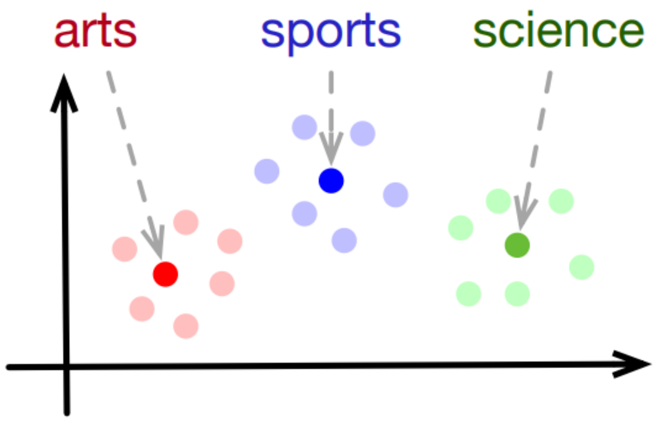
Models	NYT		arXiv	
	TC	MACC	TC	MACC
hLDA	-0.0070	0.1636	-0.0124	0.1471
hPAM	0.0074	0.3091	0.0037	0.1824
JoSE	0.0140	0.6818	0.0051	0.7412
Poincaré GloVe	0.0092	0.6182	-0.0050	0.5588
Anchored CorEx	0.0117	0.3909	0.0060	0.4941
CatE	0.0149	0.9000	0.0066	0.8176
JoSH	**0.0166**	**0.9091**	**0.0074**	**0.8324**

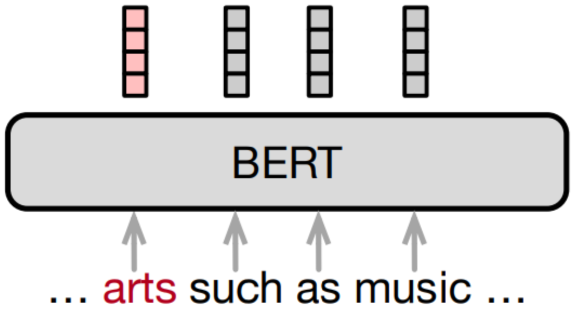
Outline

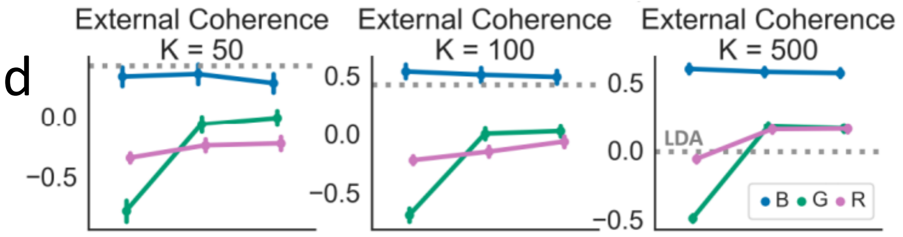
- Unsupervised Topic Discovery
- Seed-Guided Topic Discovery
  - CatE: Discriminative Topic Mining via Category-Name Guided Text Embedding [WWW'20]
  - JoSH: Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding [KDD'20]
  - SeedTopicMine: Integrating Multiple Types of Contexts [WSDM'23]

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.



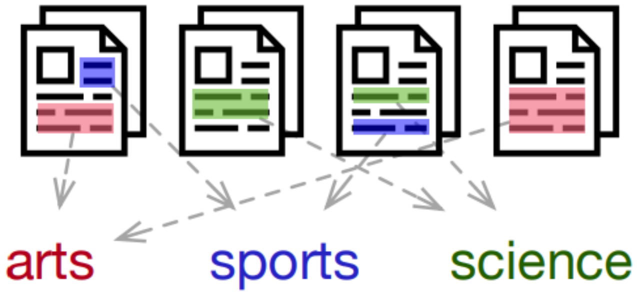




[1] Thompson, L., and Mimno, D. (2020). Topic modeling with contextualized word representation clusters. arXiv.

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

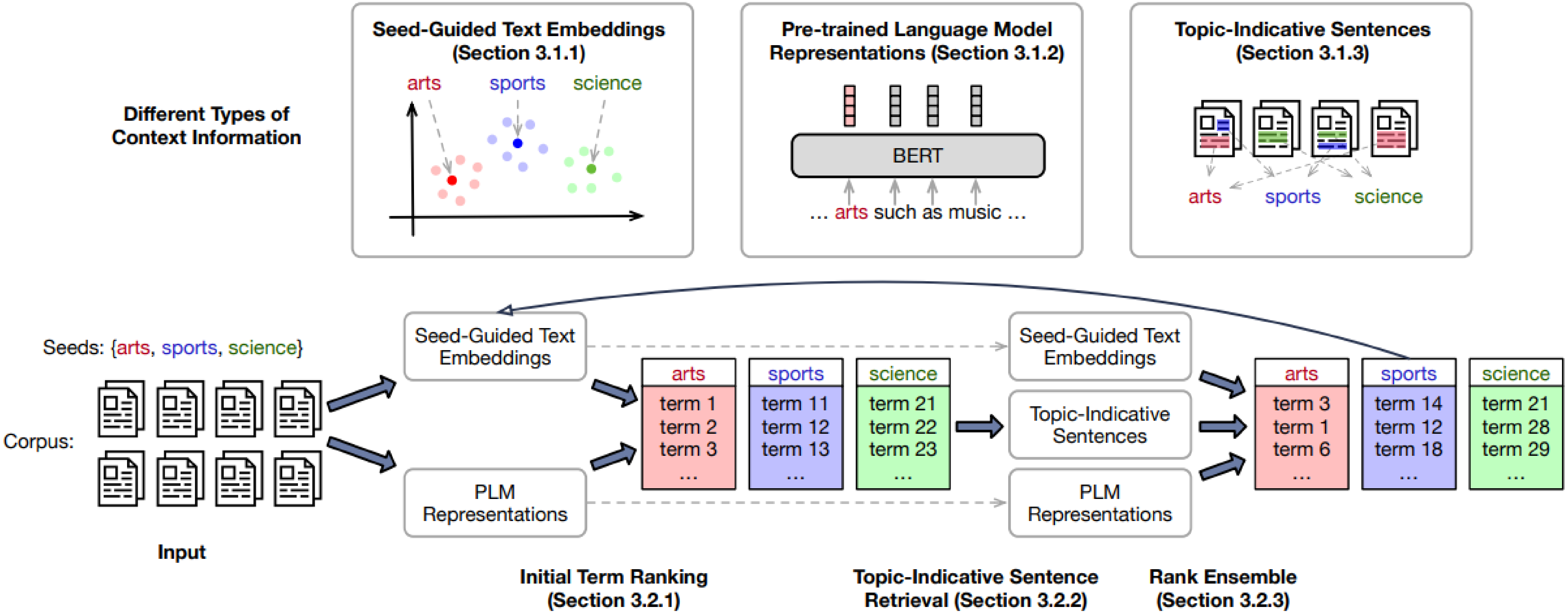
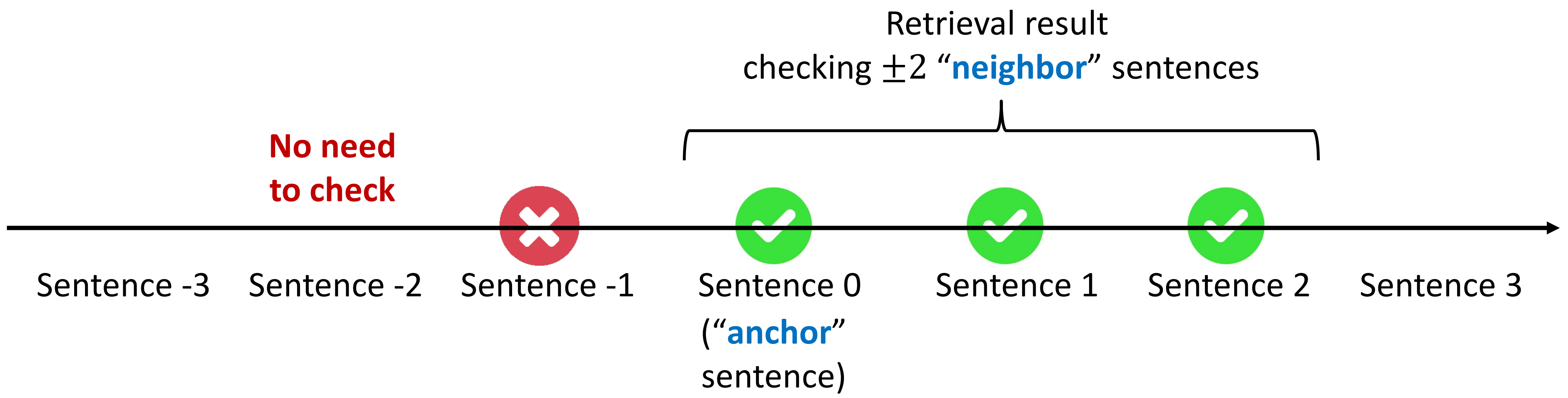


Figure 1: Overview of the SEEDTOPICMINE framework.

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 (×)	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

References

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research.
- 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., Zhang, Y., Michalski, M., Jiang, Y., Meng, Y., & Han, J. (2023). Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts. WSDM.