UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining

Jiacheng Li, Jingbo Shang, Julian McAuley

Introduction

- High-quality phrase representations are essential to finding topics and related terms in documents (a.k.a. topic mining).
- Existing phrase representation learning methods either simply combine unigram representations in a context-free manner or rely on extensive annotations to learn context-aware knowledge.

Introduction

- **UCTopic,** a novel unsupervised contrastive learning framework for context-aware phrase representations and topic mining.
- **UCTopic** is pretrained in a large scale to distinguish if the contexts of two phrase mentions have the same semantics.
- However, traditional in-batch negatives cause performance decay when finetuning on a dataset with small topic numbers.
- Hence, cluster-assisted contrastive learning (CCL) is proposed which largely reduces noisy negatives.

Assumption

- The phrase semantics are determined by their context.
- Phrases that have the same mentions have the same semantics.

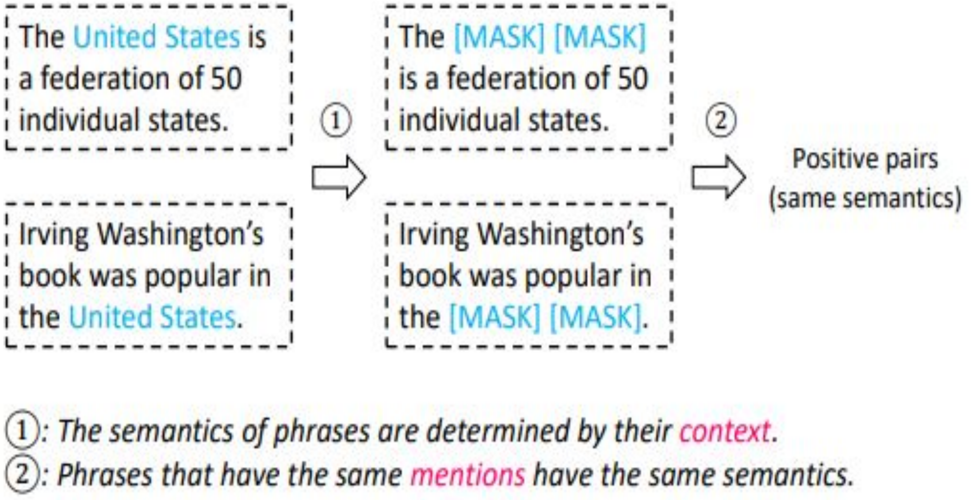
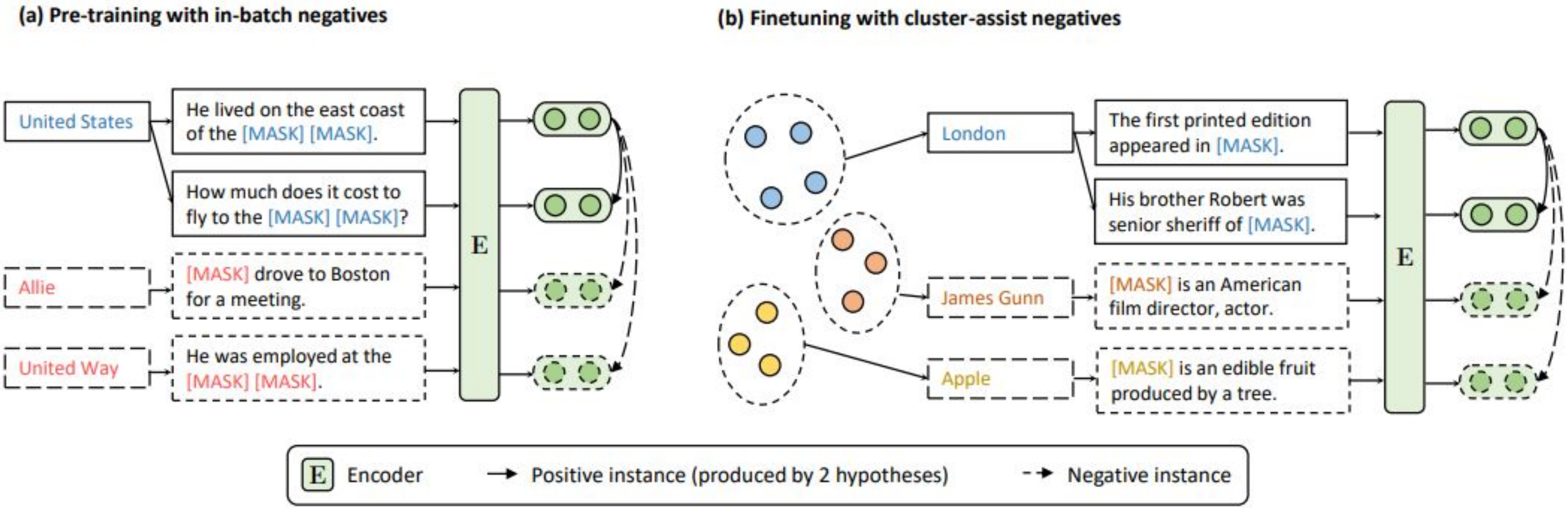


Figure 1: Two assumptions used in UCTopic to produce positive pairs for contrastive learning.

Background

- **Phrase Encoder (E):** Transformer-based model **LUKE** is used as the backbone encoder throughout this work.
- Let, phrase instance **x = (s, [l, r])** includes a sentence **s** and a character-level span **[l, r]** (l and r are left and right boundaries of a phrase). LUKE **(E)** encodes the phrase **x** and output the phrase representation **h = E(x) = E(s, [l, r])**.

UCTopic



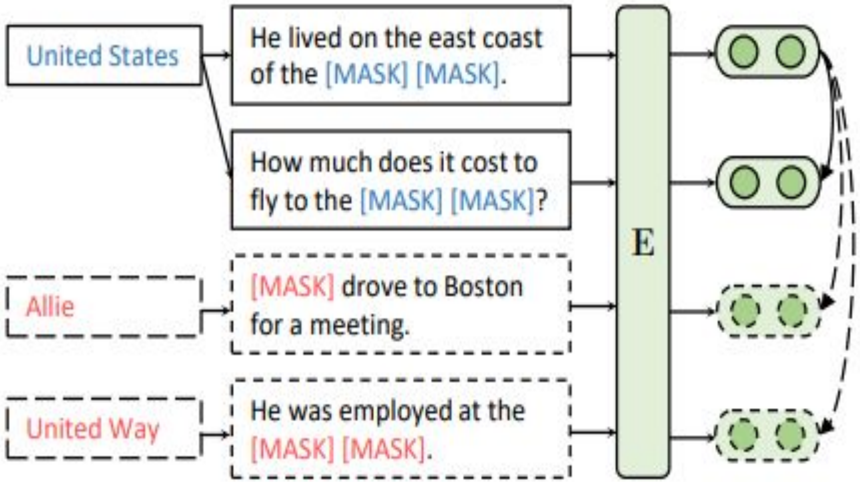
(a) Pre-training with in-batch negatives
(b) Finetuning with cluster-assist negatives
United States
He lived on the east coast of the [MASK] [MASK].
How much does it cost to fly to the [MASK] [MASK]?
Allie
[MASK] drove to Boston for a meeting.
United Way
He was employed at the [MASK] [MASK].
E
London
The first printed edition appeared in [MASK].
His brother Robert was senior sheriff of [MASK].
James Gunn
[MASK] is an American film director, actor.
Apple
[MASK] is an edible fruit produced by a tree.
E
E Encoder
Positive instance (produced by 2 hypotheses)
Negative instance

Positive Instances

Formally, suppose we have phrase instance $x = (s, [l, r])$ and its positive instance $x^+ = (s', [l', r'])$ where s denotes the sentence and $[l, r]$ are left and right boundaries of a phrase in s, we obtain the phrase representations $\mathbf{h}$ and $\mathbf{h}^+$ by encoder $\mathbf{E}$ and apply in-batch negatives for pre-training. The training objective of UCTOPIC becomes:

$$l = -\log \frac{e^{\mathrm{sim}(\mathbf{h},\mathbf{h}^+)/\tau}}{\sum_{i=1}^{N} e^{\mathrm{sim}(\mathbf{h},\mathbf{h}_i)/\tau}}, \qquad (2)$$

(a) Pre-training with in-batch negatives



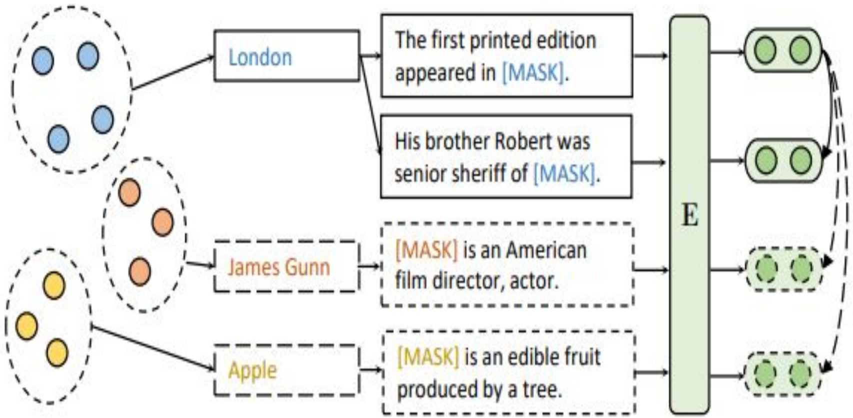
Cluster-Assisted Contrastive Learning

The training objective of finetuning is:

$$l = -\log \frac{e^{\mathrm{sim}(\mathbf{h}_{c_i}, \mathbf{h}_{c_i}^{+})/\tau}}{e^{\mathrm{sim}(\mathbf{h}_{c_i}, \mathbf{h}_{c_i}^{+})/\tau} + \sum_{c_j \in \mathcal{C}} e^{\mathrm{sim}(\mathbf{h}_{c_i}, \mathbf{h}_{c_j}^{-})/\tau}}. \quad (3)$$

$$y = \mathrm{argmax}_{c_i \in \mathcal{C}}(\mathrm{sim}(\mathbf{h}, \tilde{\mathbf{h}}_{c_i})) \quad (4)$$

(b) Finetuning with cluster-assist negatives



Experiments

- **Entity Clustering**

Datasets	CoNLL2003		BC5CDR		MIT-M		W-NUT2017	
Metrics	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
Pre-trained Representations								
Glove	0.528	0.166	0.587	0.026	0.880	0.434	0.368	0.188
BERT-Ave.	0.421	0.021	0.857	0.489	0.826	0.371	0.270	0.034
BERT-Mask	0.430	0.022	0.551	0.001	0.587	0.001	0.279	0.020
LUKE	0.590	0.281	0.794	0.411	0.831	0.432	0.434	0.205
DensePhrase	0.603	0.172	0.936	0.657	0.716	0.293	0.413	0.214
Phrase-BERT	0.643	0.297	0.918	0.617	<u>0.916</u>	<u>0.575</u>	0.452	0.241
Ours w/o CCL	<u>0.704</u>	<u>0.464</u>	<u>0.977</u>	<u>0.846</u>	0.845	0.439	<u>0.509</u>	<u>0.287</u>
Finetuning on Pre-trained UCTopic *Representations*								
Ours w/ Class.	0.703	0.458	0.972	0.827	0.738	0.323	0.482	0.283
Ours w/ In-B.	0.706	0.470	0.974	0.834	0.748	0.334	0.454	0.301
Ours w/ Auto.	0.717	0.492	0.979	0.857	0.858	0.458	0.402	0.282
UCTopic	**0.743**	**0.495**	**0.981**	**0.865**	**0.942**	**0.661**	**0.521**	**0.314**

Table 1: Performance of entity clustering on four datasets from different domains. *Class.* represents using a classifier on pseudo labels. *Auto.* represents Autoencoder. The best results among all methods are bolded and the best results of pre-trained representations are underlined. *In-B.* represents contrastive learning with in-batch negatives.

Experiments

- **Topical Phrase Mining**

Datasets	Gest	KP20k	KPTimes
# of topics	22	10	16

Table 3: The numbers of topics in three datasets.

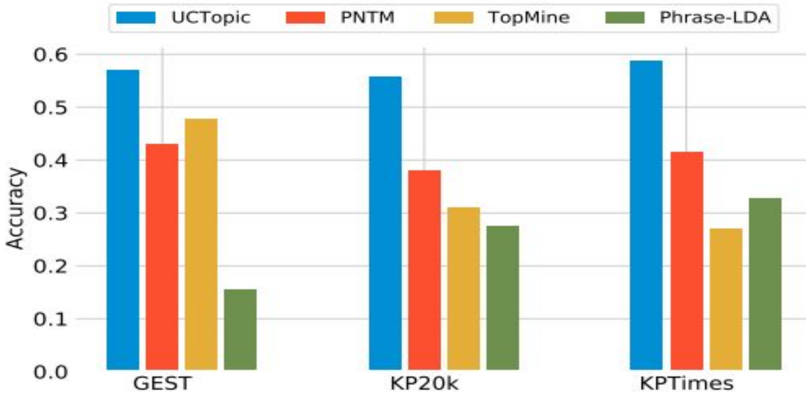


Figure 3: Results of phrase intrusion task.

Experiments

- **Phrase Intrusion**

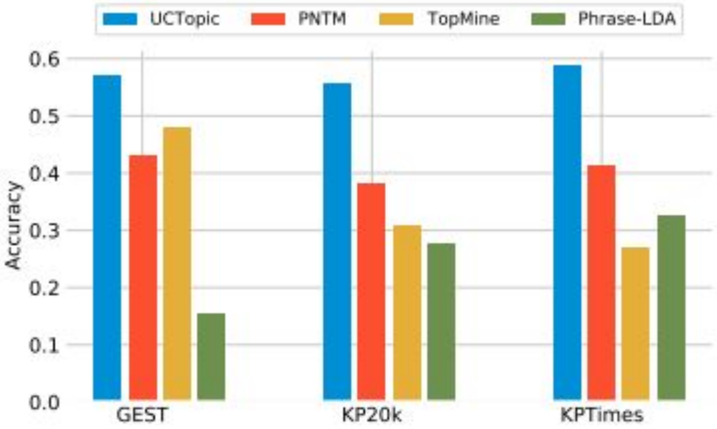


Figure 3: Results of phrase intrusion task.

Experiments

- **Phrase Coherence**

	UCTopic	PNTM	TopMine	P-LDA
Gest	20	18	20	11
KP20k	10	9	9	4

Table 4: Number of coherent topics on Gest and KP20k.

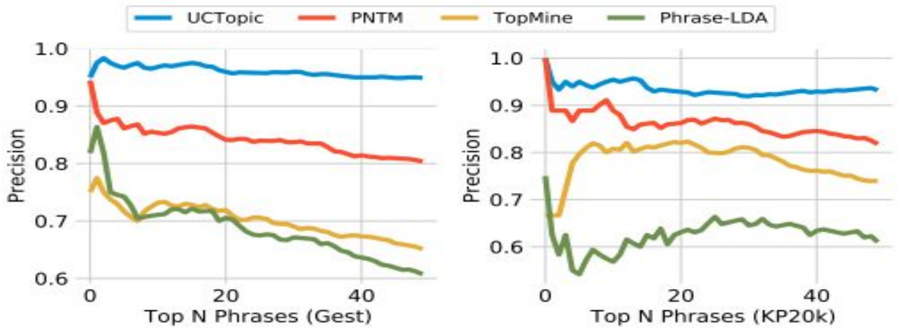


Figure 4: Results of top n precision.

Experiments

- **Phrase informativeness and diversity**

Datasets	**Gest**		**KP20k**	
Metrics	tf-idf	word-div.	tf-idf	word-div.
TopMine	**0.5379**	0.6101	0.2551	0.7288
PNTM	0.5152	0.5744	**0.3383**	0.6803
UCTopic	0.5186	**0.7486**	0.3311	**0.7600**

Table 5: Informativeness (tf-idf) and diversity (word-div.) of extracted topical phrases.

Experiments

- **Top topical phrases comparison**

Gest					KP20k	
Drinks		*Dishes*			*Programming*	
UCTopic	PNTM	UCTopic	PNTM	TopMine	UCTopic	TopMine
lager	drinks	cauliflower fried rice	great burger	mac cheese	markup language	software development
whisky	bar drink	chicken tortilla soup	great elk burger	ice cream	scripting language	software engineering
vodka	just drink	chicken burrito	great hamburger	potato salad	language construct	machine learning
whiskey	alcohol	fried calamari	good burger	french toast	java library	object oriented
rum	liquor	roast beef sandwich	good hamburger	chicken sandwich	programming structure	open source
own beer	booze	grill chicken sandwich	awesome steak	cream cheese	xml syntax	design process
ale	drink order	buffalo chicken sandwich	burger joint	fried chicken	module language	design implementation
craft cocktail	ok drink	pull pork sandwich	woody 's bbq	fried rice	programming framework	programming language
booze	alcoholic beverage	chicken biscuit	excellent burger	french fries	object-oriented language	source code
tap beer	beverage	tortilla soup	beef burger	bread pudding	python module	support vector machine

Table 6: Top topical phrases on Gest and KP20k and the minimum phrase frequency is 3.