UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining

Anonymous ACL submission

Abstract

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. In this paper, we propose 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. The key to pretraining is positive pair construction from our phrase-oriented assumptions. However, we find traditional in-batch negatives cause performance decay when finetuning on a dataset with small topic numbers. Hence, we propose cluster-assisted contrastive learning (CCL) which largely reduces noisy negatives by selecting negatives from clusters and further improves phrase representations for topics accordingly. UCTopic outperforms the state-of-the-art phrase representation model by 38.2% NMI in average on four entity clustering tasks. Comprehensive evaluation on topic mining shows that UCTopic can extract coherent and diverse topical phrases.

1 Introduction

Topic modeling discovers abstract 'topics' in a collection of documents. A topic is typically modeled as a distribution over terms. High-quality phrase representations help topic models understand phrase semantics in order to find well-separated topics and extract coherent phrases. Some phrase representation methods (Wang et al., 2021; Yu and Dredze, 2015; Zhou et al., 2017) learn context-free representations by unigram embedding combination. Context-free representations tend to extract similar phrases mentions (e.g. "great food" and "good food", see Section 4.3). Context-aware methods such as DensePhrase (Lee et al., 2021) and LUKE (Yamada et al., 2020) need supervision from task-specific datasets or distant annotations with knowledge bases. Manual or distant supervision limits the ability to represent out-of-vocabulary phrases especially for domain-specific datasets. Recently, contrastive learning has shown effectiveness for unsupervised representation learning in visual (Chen et al., 2020) and textual (Gao et al., 2021) domains.

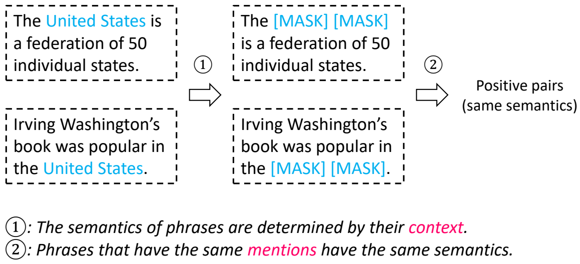

Figure 1: Two assumptions used in UCTopic to produce positive pairs for contrastive learning.

In this work, we seek to advance state-of-the-art phrase representation methods and demonstrate that a contrastive objective can be extremely effective at learning phrase semantics in sentences. We present UCTopic, an Unsupervised Contrastive learning framework for phrase representations and TOPIC mining, which can produce superior phrase embeddings and have topic-specific finetuning for topic mining. To conduct contrastive learning for phrase representations, we first seek to produce contrastive pairs. Existing data augmentation methods for natural language processing (NLP) such as back translation (Xie et al., 2020), synonym replacement (Zhang et al., 2015) and text mix up (Zhang et al., 2018) are not designed for phrase-oriented noise, and thus cannot produce training pairs for phrase representation learning. In UCTopic, we propose two assumptions about phrase semantics to obtain contrastive pairs:

1. The phrase semantics are determined by their context.

2. Phrases that have the same mentions have the same semantics.

As shown in Figure 1, given two sentences that contain the same phrase mentions (e.g., United States), we can mask the phrase mentions and the phrase semantics should stay the same based on assumption (1). Then, the phrase semantics from the two sentences are same as each other given assumption (2). Therefore, we can use the two masked sentences as positive pairs in contrastive learning. The intuition behind the two assumptions is that we expect the phrase representations from different sentences describing the same phrase should group together in the latent space. Masking the phrase mentions forces the model to learn representations from context which prevents overfitting and representation collapse (Gao et al., 2021). Based on the two assumptions, our context-aware phrase representations can be pre-trained on a large corpus via a contrastive objective without supervision.

For large-scale pre-training, we follow previous works (Chen et al., 2017; Henderson et al., 2017; Gao et al., 2021) and adopt in-batch negatives for training. However, we find in-batch negatives undermine the representation performance as finetuning (see Table 1). Because the number of topics is usually small in the finetuning dataset, examples in the same batch are likely to have the same topic. Hence, we cannot use in-batch negatives for data-specific finetuning. To solve this problem, we propose cluster-assisted contrastive learning (CCL) which leverages clustering results as pseudo-labels and sample negatives from highly confident examples in clusters. Cluster-assisted negative sampling has two advantages: (1) reducing potential positives from negative sampling compared to in-batch negatives; (2) the clusters are viewed as topics in documents, thus, cluster-assisted contrastive learning is a topic-specific finetuning process which pushes away instances from different topics in the latent space.

Based on the two assumptions and cluster-assisted negative sampling introduced in this paper, we pre-train phrase representations on a large-scale dataset and then finetune on a specific dataset for topic mining in an unsupervised way. In our experiments, we select LUKE (Yamada et al., 2020) as our backbone phrase representation model and pre-train it on Wikipedia [1] English corpus. To evaluate the quality of phrase representations, we conduct entity clustering on four datasets and find that pre-trained UCTopic achieves 53.1% (NMI) improvement compared to LUKE. After learning data-specific features with CCL, UCTopic outperforms LUKE by 73.2% (NMI) in average. We perform topical phrase mining on three datasets and comprehensive evaluation indicates UCTopic extracts coherent and diverse topical phrases. Overall, our contributions are three-fold:

- We propose UCTopic which produces superior phrase representations by unsupervised contrastive learning based on positive pairs from our phrase-oriented assumptions.
- To finetune on topic mining datasets, we propose a cluster-assisted negative sampling method for contrastive learning. This method reduces false negative instances caused by in-batch negatives and further improves phrase representations for topics accordingly.
- We conduct extensive experiments on entity type clustering and topic mining. Objective metrics and a user study show that UCTopic can largely improve the phrase representations, then extracts more coherent and diverse topical phrases than existing topic mining methods.

2 Background

In this section, we introduce background knowledge about contrastive learning and our phrase encoder LUKE (Yamada et al., 2020).

2.1 Contrastive Learning

Contrastive learning aims to learn effective representations by pulling semantically close neighbors together and pushing apart non-neighbors in the latent space (Hadsell et al., 2006). Assume that we have a contrastive instance $\{x, x^+, x_1^-, \ldots, x_{N-1}^-\}$ including one positive and $N-1$ negative instances and their representations $\{\mathbf{h}, \mathbf{h}^+, \mathbf{h}_1^-, \ldots, \mathbf{h}_{N-1}^-\}$ from the encoder, we follow the contrastive learning framework (Sohn, 2016; Chen et al., 2020; Gao et al., 2021) and take cross-entropy as our objective function:

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

where τ is a temperature hyperparameter and $\mathrm{sim}(\mathbf{h}_1, \mathbf{h}_2)$ is the cosine similarity $\frac{\mathbf{h}_1^\top \mathbf{h}_2}{\|\mathbf{h}_1\| \cdot \|\mathbf{h}_2\|}$.

[1] https://dumps.wikimedia.org/

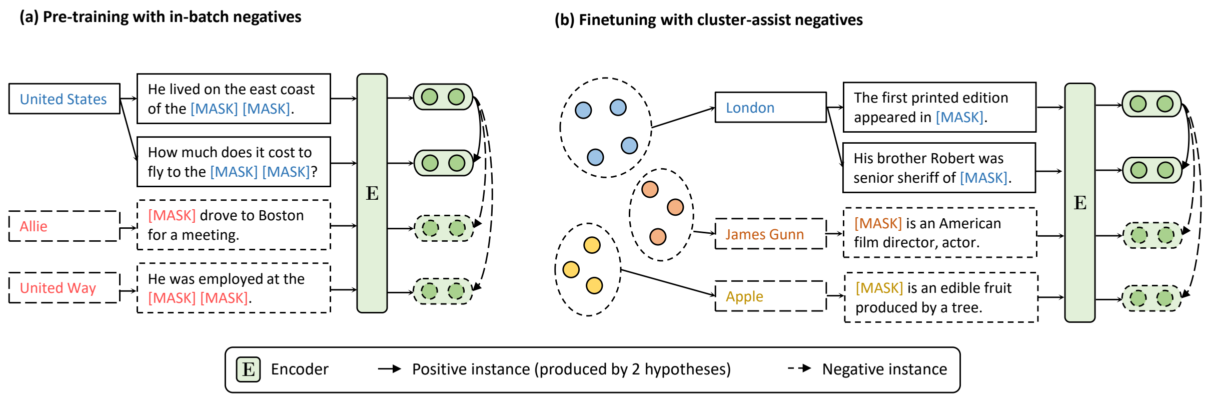

Figure 2: (a) Pre-training UCTopic on a large-scale dataset with positive instances from our two assumptions and in-batch negatives. (b) Finetuning UCTopic on a topic mining dataset with positive instances from our two assumptions and negatives from clustering.

2.2 Phrase Encoder

In this paper, our phrase encoder **E** is transformer-based model LUKE (Yamada et al., 2020). LUKE is a pre-trained language model that can directly output the representations of tokens and spans in sentences. Our 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). **E** encodes the phrase x and output the phrase representation $\mathbf{h} = \mathbf{E}(x) = \mathbf{E}(s, [l, r])$. Although LUKE can output span representations directly, we will show that span representations from LUKE are not able to represent phrases well (see Section 4.2). Different from LUKE, which is trained by predicting entities, UCTopic is trained by contrastive learning on phrase contexts. Hence, the phrase presentations from UCTopic are context-aware and robust to different domains.

3 UCTopic

UCTopic is an unsupervised contrastive learning method for phrase representations and topic mining. Our goal is to learn a phrase encoder as well as topic representations, so we can represent phrases effectively for general settings and find topics from documents in an unsupervised way. In this section, we introduce UCTopic from two aspects: (1) constructing positive pairs for phrases; (2) cluster-assisted contrastive learning.

3.1 Positive Instances

One critical problem in constrastive learning is to how to construct positive pairs (x, x^+). Previous works (Wu et al., 2020; Meng et al., 2021) apply augmentation techniques such as word deletion, reordering, and paraphrasing. However, these methods are not suitable for phrase representation learning. In this paper, we utilize the proposed assumptions introduced in Section 1 to construct positive instances for contrastive learning.

Consider an example to understand our positive instance generation process: In Figure 2 (a), phrase `United States` appears in two different sentences "*He lived on the east coast of the <u>United States</u>*" and "*How much does it cost to fly to the <u>United States</u>*". We expect the phrase (`United States`) representations from the two sentences to be similar to reflect phrase semantics. To encourage the model to learn phrase semantics from context and prevent the model from comparing phrase mentions in contrastive learning, we mask the phrase mentions with `[MASK]` token. The two masked sentences are used as positive instances. To decrease the inconsistency caused by masking between training and evaluation, in a positive pair, we keep one phrase mention unchanged in probability p.

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 **E** and apply in-batch negatives for pre-training. The training objective of UCTopic becomes:

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

for a mini-batch of N instances, where $\mathbf{h}_i$ is an instance in a batch.

3.2 Cluster-Assisted Contrastive Learning

We find that contrastive learning with in-batch negatives on small datasets can undermine the phrase representations (see Section 4.2). Different from pre-training on a large corpus, in-batch negatives usually contain instances that have similar semantics as positives. For example, one document has three topics and our batch size is 32. Thus, some instances in one batch are from the same topic but in-batch method views these instances as negatives with each other. In this case, contrastive learning has noisy training signals and then results in decreasing performance.

To reduce the noise in negatives while optimizing phrase representations according to topics in documents, we propose cluster-assisted contrastive learning (CCL). The basic idea is to utilize prior knowledge from pre-trained representations and clustering to reduce the noise existing in the negatives. Specifically, we first find the topics in documents with a clustering algorithm based on pre-trained phrase representations from UCTopic. The centroids of clusters are considered as topic representations for phrases. After computing the cosine distance between phrase instances and centroids, we select t percent of instances that are close to centroids and assign pseudo labels to them. Then, the label of a phrase mention p^m [2] is determined by the majority vote of instances $\{x_0^m, x_1^m, \ldots, x_n^m\}$ that contain p^m, where n is the number of sentences assigned pseudo labels. In this way, we get some prior knowledge of phrase mentions for the following contrastive learning. See Figure 2 (b); three phrase mentions (`London`, `James Gunn` and `Apple`) which belong to three different clusters are labeled by different topic categories.

Suppose we have a topic set $\mathcal{C}$ in our documents, with phrases and their pseudo labels, we construct positive pairs $(x_{c_i}, x_{c_i}^+)$ by method introduced in Section 3.1 for topic c_i where $c_i \in \mathcal{C}$. To have contrastive instances, we randomly select phrases $p_{c_j}^m$ and instances $x_{c_j}^m$ from topic c_j as negative instances $x_{c_j}^-$ in contrastive learning, where $c_j \in \mathcal{C} \wedge c_j \neq c_i$. As shown in Figure 2 (b), we construct positive pairs for phrase `London`, and use two phrases `James Gunn` and `Apple` from the other two clusters to randomly select negative instances. With pseudo labels, our method can avoid instances that have similar semantics as `London`.

[2]phrase mentions are extracted from sentence s, i.e., $p^m = s[l:r]$

The training objective of finetuning is:

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

As for the masking strategy in pre-training, we conduct masking for all training instances but keep $x_{c_i}^+$ and $x_{c_j}^-$ unchanged in probability p.

To infer the topic y of phrase instance x, we compute the cosine similarity between phrase representation $\mathbf{h}$ and topic representations $\tilde{\mathbf{h}}_{c_i}, c_i \in \mathcal{C}$. The nearest neighbor topic of x is used as phrase topic. Formally,

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

4 Experiments

In this section, we evaluate the effectiveness of contrastive learning. We start with entity clustering to compare the phrase representations from different methods. For topic modeling, we evaluate the topical phrases from three aspects and compare UCTopic to other topic modeling baselines.

4.1 Implementation Details

For pre-training, we start from a pretrained LUKE-BASE model (Yamada et al., 2020). We follow previous works (Gao et al., 2021; Soares et al., 2019) and two losses are used concurrently: the masked language model loss and the contrastive learning loss with in-batch negatives. To generate the training corpus, we use English Wikipedia and extract text with hyper links as phrases. Phrases have the same entity ids from Wikidata [3] or have the same mentions are considered as the same phrases (i.e., phrases have the same semantics). We enumerate all sentence pairs containing the same phrase as positive pairs in contrastive learning. After processing, the pre-training dataset has 11.6 million sentences and 108.8 million training instances. Our pre-training learning rate is 5e-5, batch size is 100 and our model is optimized by AdamW in 1 epoch. The probability p of keeping phrase mentions unchanged is 0.5 and the temperature τ in the contrastive loss is set to 0.05.

4.2 Entity Clustering

To test the performance of phrase representations under objective tasks and metrics, we first apply

[3]https://www.wikidata.org/

UCTopic on entity clustering and compare to other representation learning methods.

Datasets. We conduct entity clustering on four datasets with annotated entities and their semantic categories are from general, review and biomedical domains: (1) CoNLL2003 (Sang and Meulder, 2003) consists of 20,744 sentences extracted from Reuters news articles. We use Person, Location, and Organization entities in our experiments.[4] (2) BC5CDR (Li et al., 2016) is the BioCreative V CDR task corpus. It contains 18,307 sentences from PubMed articles, with 15,953 chemical and 13,318 disease entities. (3) MIT Movie (MIT-M) (Liu et al., 2013) contains 12,218 sentences with Title and Person entities. (4) W-NUT 2017 (Derczynski et al., 2017) focuses on identifying unusual entities in the context of emerging discussions and contains 5,690 sentences and six kinds of entities [5].

Finetuning Setup. The learning rate for finetuning is 1e-5. We select t (percent of instances) from $\{5, 10, 20, 50\}$. The probability p of keeping phrase mentions unchanged and temperature τ in contrastive loss are the same as in pre-training settings. We apply K-Means to get pseudo labels for all experiments. Because UCTopic is an unsupervised method, we use all data to finetune and evaluate. All results for finetuning are the best results during training. We follow previous clustering works (Xu et al., 2017; Zhang et al., 2021) and adopt Accuracy (ACC) and Normalized Mutual Information (NMI) to evaluate different approaches.

Compared Baseline Methods. To demonstrate the effectiveness of our pre-training method and cluster-assisted contrastive learning (CCL), we compare baseline methods from two aspects:

(1) Pre-trained token or phrase representations:

- **Glove** (Pennington et al., 2014). Pre-trained word embeddings on 6B tokens and dimension is 300. We use averaging word embeddings as the representations of phrases.
- **BERT** (Devlin et al., 2019). Obtains phrase representations by averaging token representations (BERT-Ave.) or following CGExpan (Zhang et al., 2020) to substitute phrases with the `[MASK]` token, and use `[MASK]` representations as phrase embeddings (BERT-MASK).
- **LUKE** (Yamada et al., 2020). Use as backbone model to show the effectiveness of our contrastive learning for pre-training and finetuning.
- **DensePhrase** (Lee et al., 2021). Pre-trained phrase representation learning in a supervised way for question answering problem. We use a pre-trained model released from the authors to get phrase representations.
- **Phrase-BERT** (Wang et al., 2021). Context-agnostic phrase representations from pretraining. We use a pre-trained model from the authors and get representations by phrase mentions.
- **Ours w/o CCL**. Pre-trained phrase representations of UCTopic without cluster-assisted contrastive finetuning.

(2) Fine-tuning methods based on pre-trained representations of UCTopic.

- **Classifier**. We use pseudo labels as supervision to train a MLP layer and obtain a classifier of phrase categories.
- **In-Batch Contrastive Learning**. Same as contrastive learning for pre-training which uses in-batch negatives.
- **Autoencoder**. Widely used in previous neural topic and aspect extraction models (He et al., 2017; Iyyer et al., 2016; Tulkens and van Cranenburgh, 2020). We follow ABAE (He et al., 2017) to implement our autoencoder model for phrases.

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. *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.

Experimental Results. We report evaluation results of entity clustering in Table 1. Overall, UCTopic achieves the best results on all datasets and metrics. Specifically, UCTopic improves the state-of-the-art method (Phrase-BERT) by 38.2% NMI in average, and outperforms our backbone model (LUKE) by 73.2% NMI.

[4] We do not evaluate on the Misc category because it does not represent a single semantic category.

[5] corporation, creative work, group, location, person, product

Model	UCTopic		LUKE	
Metric	ACC	NMI	ACC	NMI
Context+Mention	0.44	0.29	0.39	0.21
Mention	0.32 (-27%)	0.15 (-48%)	0.28 (-28%)	0.10 (-52%)
Context	0.43 (-3%)	0.16 (-44%)	0.27 (-31%)	0.07 (-67%)

Table 2: Ablation study on the input of phrase instances of W-NUT 2017. UCTopic here is pre-trained representations without CCL finetuning. Percentages in brackets are changes compared to Context+Mention.

When we compare different pre-trained representations, we find that our method (Ours w/o CCL) outperforms the other baselines on three datasets except MIT-M. There are two reasons: (1) All words in MIT-M are lower case which is inconsistent with our pretraining dataset. The inconsistency between training and test causes performance to decay. (2) Sentences from MIT-M are usually short (10.16 words in average) compared to other datasets (e.g., 17.9 words in W-NUT2017). UCTopic can obtain limited contextual information with short sentences. However, the performance decay caused by the two reasons can be eliminated by our CCL finetuning on dataset since on MIT-M UCTopic achieves better results (0.661 NMI) than Phrase-BERT (0.575 NMI) after CCL.

On the other hand, compared to other finetuning methods, our CCL finetuning can further improve the pre-trained phrase representations by capturing data-specific features. The improvement is up to 50% NMI on the MIT-M dataset. Ours w/ Class. performs worse than our pre-trained UCTopic in most cases which indicates that pseudo labels from clustering are noisy and cannot directly be used as supervision for representation learning. Ours w/ In-B. is similar as Ours w/ Class. which verifies our motivation on using CCL instead of in-batch negatives. An autoencoder can improve pre-trained representations on three datasets but the margins are limited and the performance even drops on W-NUT2017. Compared to other finetuning methods, our CCL finetuning consistently improves pre-trained phrase representations on different domains.

Context or Mentions. To investigate the source of UCTopic phrase semantics (i.e., phrase mentions or context), we conduct an ablation study on the type of input and compare UCTopic to LUKE. To eliminate the influence of repeated phrase mentions on clustering results, we use only one phrase instance (i.e., sentence and position of a phrase) for each phrase mention. As shown in Table 2, there are three types of inputs: (1) Context+Mention: The same input as experiments in Table 1 including the whole sentence that contains the phrase. (2) Mention: Use only phrase mentions as inputs of the two models. (3) Context: We mask the phrase mentions in sentences and models can only get information from the context. We can see that UCTopic gets more information from context (0.43 ACC, 0.16 NMI) than mentions (0.32 ACC, 0.15 NMI). Compared to LUKE, UCTopic is more robust to phrase mentions (when predicting on only context, UCTopic -3% ACC and -44% NMI vs. LUKE -31% ACC and -67% NMI).

4.3 Topical Phrase Mining

In this section, we apply UCTopic on topical phrase mining and conduct human evaluation to show our model outperforms previous baselines.

Experiment Setup. To find topical phrases in documents, we first extract noun phrases by spaCy [6] noun chunks and remove single pronoun words. Before CCL finetuning, we obtain the number of topics for each dataset by computing the Silhouette Coefficient (Rousseeuw, 1987) (details in Appendix A.1). Then, we conduct CCL on the dataset with the same settings as described in Section 4.2. Finally, after obtaining topic distribution $\mathbf{z}_x \in \mathbb{R}^{|\mathcal{C}|}$ for a phrase instance x in a sentence, we get context-agnostic phrase topics by using averaged topic distribution $\mathbf{z}_{p^m} = \frac{1}{n}\sum_{1\le i\le n}\mathbf{z}_{x_i^m}$, where phrase instances $\{x_i^m\}$ in different sentences have the same phrase mention p^m. The topic of a phrase mention has the highest probability in $\mathbf{z}_{p^m}$.

Dataset. We conduct topical phrase mining on three datasets from news, review and computer science domains.

- **Gest**. We collect restaurant reviews from Google Local[7] and use 100K reviews containing 143,969 sentences for topical phrase mining.
- **KP20k** (Meng et al., 2017) is a collection of titles and abstracts from computer science papers. 500K sentences are used in our experiments.
- **KPTimes** (Gallina et al., 2019) includes news articles from the New York Times from 2006 to 2017 and 10K news articles from the Japan Times. We use 500K sentences for topic mining.

[6] https://spacy.io/

[7] https://www.google.com/maps

Datasets	Gest	KP20k	KPTimes
# of topics	22	10	16

Table 3: The numbers of topics in three datasets.

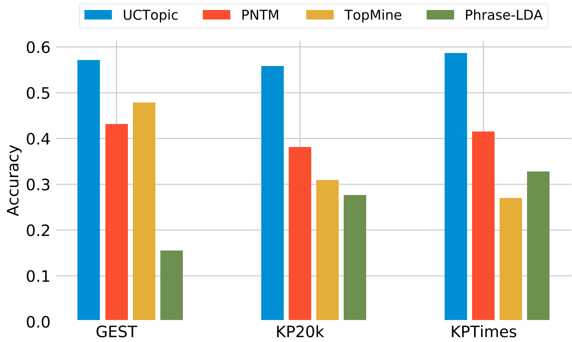

Figure 3: Results of phrase intrusion task.

The number of topics determined by Silhouette Coefficient is shown in Table 3.

Compared Baseline Methods. We compare UCTopic against three topic baselines:

- **Phrase-LDA** (Mimno, 2015). LDA model incorporates phrases by converting phrases into unigrams (e.g., "city view" to "city_view").
- **TopMine** (El-Kishky et al., 2014). A scalable pipeline that paritions a document into phrases, then uses phrases as constraints to ensure all words are placed in the same topic.
- **PNTM** (Wang et al., 2021). A topic model with Phrase-BERT by using an autoencoder that reconstructs a document representation. The model is the state-of-the-art topic model.

We do not include topic models such as LDA (Blei et al., 2003), PD-LDA (Lindsey et al., 2012), TNG (Wang et al., 2007), KERT (Danilevsky et al., 2014) as baselines, because these models are compared in TopMine and PNTM. For Phrase-LDA and PNTM, we use the same phrase list produced by UCTopic. TopMine produced phrases by itself.

	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.

Topical Phrase Evaluation. We evaluate the quality of topical phrases from three aspects: (1) *topical separation*; (2) *phrase coherence*; (3) *phrase informativeness and diversity*.

To evaluate *topical separation*, we perform the **phrase intrusion** task following previous work (El-Kishky et al., 2014; Chang et al., 2009). The phrase intrusion task involves a set of questions asking humans to discover the 'intruder' phrase from other phrases (details in Appendix B.1). Results of the task evaluate how well the phrases are separated by topics. The evaluation results are shown in Figure 3. UCTopic outperforms other baselines on three datasets, which means our model can find well-separated topics in documents.

To evaluate *phrase coherence* in one topic, we follow ABAE (He et al., 2017) and ask annotators to evaluate if the top 50 phrases from one topic are coherent (i.e., most phrases represent the same topic). 3 annotators evaluate four models on Gest and KP20k datasets. Numbers of coherent topics are shown in Table 4. We can see that UCTopic, PNTM and TopMine can recognize similar numbers of coherent topics, but the numbers of Phrase-LDA are less than the other three models. For a coherent topic, each of the top phrases will be labeled as correct if the phrase reflects the related topic. Same as ABAE, we adopt *precision@n* to evaluate the results. Figure 4 shows the results; we can see that UCTopic substantially outperforms other models and maintain high precision with a large n when the precision of other models decreases.

Finally, to evaluate *phrase informativeness and diversity*, we use tf-idf and word diversity (word-div.) to evaluate the top topical phrases (details in Appendix B.2). Results are shown in table 5. PNTM and UCTopic achieve similar tf-idf scores, because the two methods use the same phrase lists extracted from spaCy. UCTopic extracts the most diverse phrases in a topic, because our phrase representations are more context-aware. In contrast,

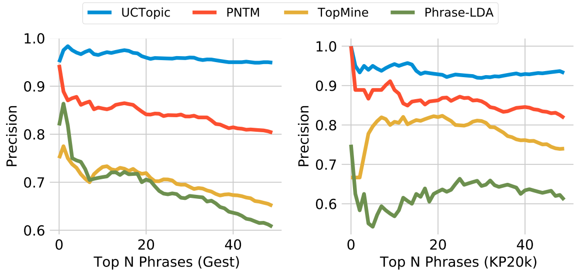

Figure 4: Results of top n precision.

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.

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.

since PNTM gets representations dependent on phrase mentions, the phrases from PNTM contain the same words and hence are less diverse.

Case Study. We compare top phrases from UCTopic, PNTM and TopMine in Section 4.3. From examples, we can see the phrases are consistent with our user study and diversity evaluation. Although the phrases from PNTM are coherent, the diversity of phrases is less than others (e.g., "drinks", "bar drink", "just drink" from Gest) because context-agnostic representations let similar phrase mentions group together. The phrases from TopMine are diverse but are not coherent in some cases (e.g., "machine learning" and "support vector machine" in the programming topic). In contrast, UCTopic can extract coherent and diverse topical phrases from documents.

5 Related Work

Many attempts have been made to extract topical phrases via LDA (Blei et al., 2003). Wallach (2006) incorporated a bigram language model into LDA by a hierarchical dirichlet generative probabilistic model to share the topic across each word within a bigram. TNG (Wang et al., 2007) applied additional latent variables and word-specific multinomials to model bi-grams and combined bi-grams to form n-gram phrases. PD-LDA (Lindsey et al., 2012) used a hierarchical Pitman-Yor process to share the same topic among all words in a given n-gram. Danilevsky et al. (2014) ranked the resultant phrases based on four heuristic metrics. TOP-Mine (El-Kishky et al., 2014) proposed to restrict all constituent terms within a phrase to share the same latent topic and assign a phrase to the topic of its constituent words. Compared to previous topic mining methods, UCTopic builds on the success of pre-trained language models and unsupervised contrastive learning on a large-scale dataset. Therefore, UCTopic provides high-quality pre-trained phrase representations and state-of-the-art finetuning for topic mining.

Early works in phrase representation build upon a composition function that combines component word embeddings together into simple phrase embedding. Yu and Dredze (2015) implemented the function by rule-based composition over word vectors. Zhou et al. (2017) applied a pair-wise GRU model and datasets such as PPDB (Pavlick et al., 2015) to learn phrase representations. Phrase-BERT (Wang et al., 2021) composed token embeddings from BERT and pretrained on positive instances produced by GPT-2-based diverse paraphrasing model (Krishna et al., 2020). Lee et al. (2021) learned phrase representations from the supervision of reading comprehension tasks and applied representations on open-domain QA. Other works learned phrase embeddings for specific tasks such as semantic parsing (Socher et al., 2011) and machine translation (Bing et al., 2015). In this paper, we present unsupervised contrastive learning method for pre-training phrase representations of general purposes and for finetuning to topic-specific phrase representations.

6 Conclusion

In this paper, we propose UCTopic, a contrastive learning framework that can effectively learn phrase representations without supervision. To finetune on topic mining datasets, we propose cluster-assisted contrastive learning which reduces noise by selecting negatives from clusters. During finetuning, our phrase representations are optimized for topics in the document hence the representations are further improved. We conduct comprehensive experiments on entity clustering and topical phrase mining. Results show that UCTopic largely improves phrase representations. Objective metrics and a user study indicate UCTopic can extract coherent and diverse topical phrases.

7 Ethical Consideration

We do not anticipate any major ethical concerns; topic mining is a fundamental problem in natural language processing. A minor consideration is the potential for certain types of hidden biases to be introduced into our results (i.e., performance regressions for some subset of the data in spite of overall performance gains), for example by a biased selection of topical phrases. We did not observe any such issues in our experiments, and indeed these considerations seem low-risk for the specific datasets studied here.

References

Lidong Bing, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, and Rebecca J. Passonneau. 2015. Abstractive multi-document summarization via phrase selection and merging. In *ACL*.

David M. Blei, A. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Jonathan Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709.

Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On sampling strategies for neural network-based collaborative filtering. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Marina Danilevsky, Chi Wang, Nihit Desai, Xiang Ren, Jingyi Guo, and Jiawei Han. 2014. Automatic construction and ranking of topical keyphrases on collections of short documents. In *SDM*.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, and Jiawei Han. 2014. Scalable topical phrase mining from text corpora. *ArXiv*, abs/1406.6312.

Ygor Gallina, Florian Boudin, and Béatrice Daille. 2019. Kptimes: A large-scale dataset for keyphrase generation on news documents. In *INLG*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *ArXiv*, abs/2104.08821.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2:1735–1742.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *ACL*.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *ArXiv*, abs/1705.00652.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan L. Boyd-Graber, and Hal Daumé. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *NAACL*.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. *ArXiv*, abs/2010.05700.

Jinhyuk Lee, Mujeen Sung, Jaewoo Kang, and Danqi Chen. 2021. Learning dense representations of phrases at scale. In *ACL/IJCNLP*.

J. Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, A. P. Davis, C. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016.

Robert V. Lindsey, Will Headden, and Michael Stipicevic. 2012. A phrase-discovering topic model using hierarchical pitman-yor processes. In *EMNLP*.

Jingjing Liu, Panupong Pasupat, Yining Wang, D. Scott Cyphers, and James R. Glass. 2013. Query understanding enhanced by hierarchical parsing structures. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 72–77.

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. In *ACL*.

Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul N. Bennett, Jiawei Han, and Xia Song. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *ArXiv*, abs/2102.08473.

David Mimno. 2015. Using phrases in mallet topic models. http://www.mimno.org/articles/phrases/.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *ACL*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Peter Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

E. T. K. Sang and F. D. Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *ArXiv*, cs.CL/0306050.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *ArXiv*, abs/1906.03158.

Richard Socher, Cliff Chiung-Yu Lin, A. Ng, and Christopher D. Manning. 2011. Parsing natural scenes and natural language with recursive neural networks. In *ICML*.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*.

Stéphan Tulkens and Andreas van Cranenburgh. 2020. Embarrassingly simple unsupervised aspect extraction. In *ACL*.

Hanna M. Wallach. 2006. Topic modeling: beyond bag-of-words. *Proceedings of the 23rd international conference on Machine learning*.

Shufan Wang, Laure Thompson, and Mohit Iyyer. 2021. Phrase-bert: Improved phrase embeddings from bert with an application to corpus exploration. *ArXiv*, abs/2109.06304.

Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 697–702.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *ArXiv*, abs/2012.15466.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised data augmentation for consistency training. *arXiv: Learning*.

Jiaming Xu, Bo Xu, Peng Wang, Suncong Zheng, Guanhua Tian, and Jun Zhao. 2017. Self-taught convolutional neural networks for short text clustering. *Neural networks : the official journal of the International Neural Network Society*, 88:22–31.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: Deep contextualized entity representations with entity-aware self-attention. In *EMNLP*.

Mo Yu and Mark Dredze. 2015. Learning composition models for phrase embeddings. *Transactions of the Association for Computational Linguistics*, 3:227–242.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Supporting clustering with contrastive learning. In *NAACL*.

Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412.

Xiang Zhang, Junbo Jake Zhao, and Yann André LeCun. 2015. Character-level convolutional networks for text classification. *ArXiv*, abs/1509.01626.

Yunyi Zhang, Jiaming Shen, Jingbo Shang, and Jiawei Han. 2020. Empower entity set expansion via language model probing. *ArXiv*, abs/2004.13897.

Zhihao Zhou, Lifu Huang, and Heng Ji. 2017. Learning phrase embeddings from paraphrases with grus. *ArXiv*, abs/1710.05094.

A Topical Phrase Mining

A.1 Find Numbers of Topics

We randomly sample 10K phrases from dataset and apply K-Means clustering on pre-trained UCTopic phrase representations with different cluster numbers. We compute Silhouette Coefficient score for different topic numbers and the number with the largest score will be used as the topic number in a dataset.

B User Study

B.1 Phrase Intrusion

In our experiments, each question has 6 phrases and 5 of them are randomly sampled from the top 50 phrases of one topic and the remaining phrase is randomly chosen from another topic (top 50 phrases). Annotators are asked to select the intruder phrase. We sample 50 questions for each method and each dataset (600 questions in total) and shuffle all questions. Because these questions are sampled independently, we asked 4 annotators to answer these questions and each annotator answers 150 questions in average.

B.2 Phrase Informativeness and Diversity

Informative phrases cannot be very common phrases in a corpus (e.g., "good food" in Gest) and we use tf-idf to evaluate the "importance" of a phrase. To eliminate the influence of phrase length, we use averaged word tf-idf in a phrase as the phrase tf-idf. Specifically, tf-idf$(p, d) = \frac{1}{m}\sum_{1\leq i\leq m}$ tf-idf(w_i^p), where d denotes the document and p is the phrase. In our experiments, a document is a sentence is a review.

Furthermore, we hope that our phrases are diverse enough in a topic instead of expressing the same meaning (e.g., "good food" and "great food"). To evaluate the diversity of the top phrases, we calculate the ratio of distinct words among all words. Formally, given a list of phrases $[p_1, p_2, \ldots, p_n]$, we tokenize the phrases into a word list $\mathbf{w} = [w_1^{p_1}, w_2^{p_1}, \ldots, w_m^{p_n}]$ and $\mathbf{w}'$ is the set of unique words in $\mathbf{w}$. The word diversity is computed by $\frac{|\mathbf{w}'|}{|\mathbf{w}|}$. We only evaluate coherent topics labeled in *phrase coherence* and the coherent topics numbers of Phrase-LDA are obviously smaller than others, hence we evaluate the other three models. We compute the tf-idf and word-div. on the top 10 phrases and use the averaged value on topics as final scores.