yin_2022_improving_deep_embedded_clustering_via_learning_cluster_level_representations

Year

2022

Author(s)

Yin, Qing and Wang, Zhihua and Song, Yunya and Xu, Yida and Niu, Shuai and Bai, Liang and Guo, Yike and Yang, Xian

Title

Improving Deep Embedded Clustering via Learning Cluster-level Representations

Venue

COLING

Topic labeling

Fully automated

Focus

Secondary

Type of contribution

Established approach

Underlying technique

Deep Embedding Clustering (DEC) model

Topic labeling parameters

See `Topic modeling parameters`

Label generation

The proposed model learns topics together with the clusters they belong to.
Clusters are associated with ground-truth labels provided by the dataset.
Cluster labels can be used to describe the topics belonging to it.

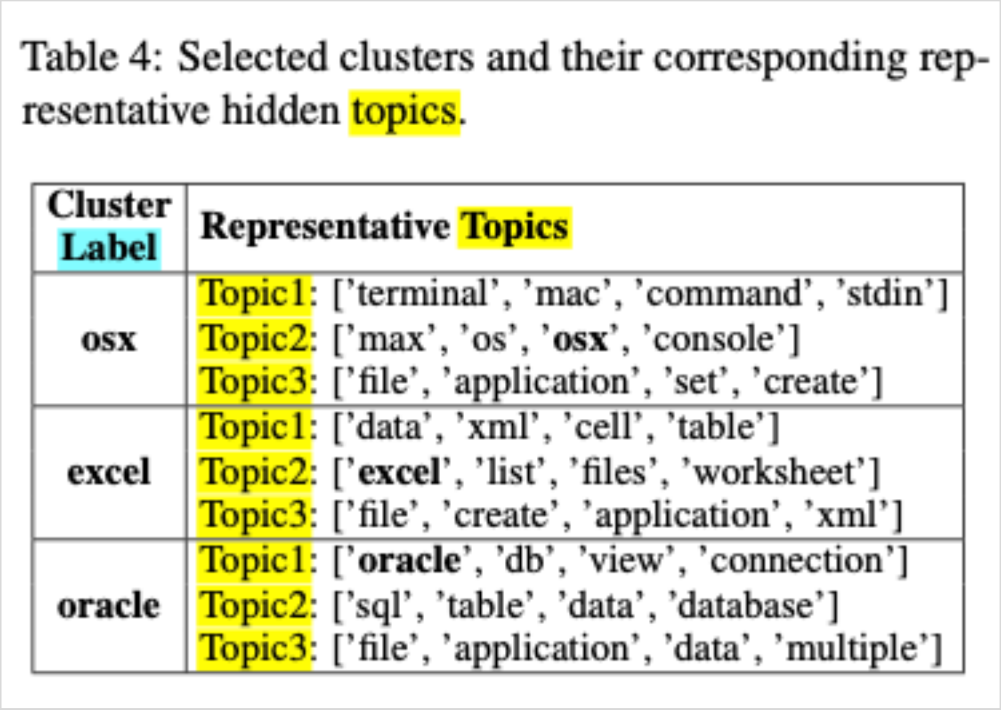


Table 4: Selected clusters and their corresponding representative hidden topics.

Cluster Label	Representative Topics
osx	Topic1: ['terminal', 'mac', 'command', 'stdin'] Topic2: ['max', 'os', '**osx**', 'console'] Topic3: ['file', 'application', 'set', 'create']
excel	Topic1: ['data', 'xml', 'cell', 'table'] Topic2: ['**excel**', 'list', 'files', 'worksheet'] Topic3: ['file', 'create', 'application', 'xml']
oracle	Topic1: ['**oracle**', 'db', 'view', 'connection'] Topic2: ['sql', 'table', 'data', 'database'] Topic3: ['file', 'application', 'data', 'multiple']

Motivation

\

Topic modeling

Embedded topic modelling (ETM) - Deep Embedding Clustering (DEC) (based short text clustering) model

Topic modeling parameters

Optimizer: Adam
Batch size: 200
SentenceBERT:

- Model: distilbert-base-nli-stsb-mean-tokens
- Max input length: 32

α: 10.0 for biomedical dataset and 1.0 for other datasets
Temperature parameter used in the contrasting module: 0.5

Nr. of topics

Biomedical dataset: 20
StackOverflow and AgNews: 5

Label

Datasets gold standard labels.

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Domain (paper): Deep Embedded Clustering methods
Domain (corpus): News, online Q&A, Biomedical

Problem statement

Propose a novel Deep Embedding Clustering (DEC) (based short text clustering) model, which we named the deep embedded clustering model with cluster-level representation learning (DEC- CRL) to jointly learn cluster and instance level representations.
Extending the embedded topic modelling approach to introduce reconstruction constraints to help learn cluster-level representations.

Corpus

Origin: AgNews

Nr. of documents: 8,000 (6,400 training, 1,600 testing)

Details: Collection of news titles

Origin: StackOverflow

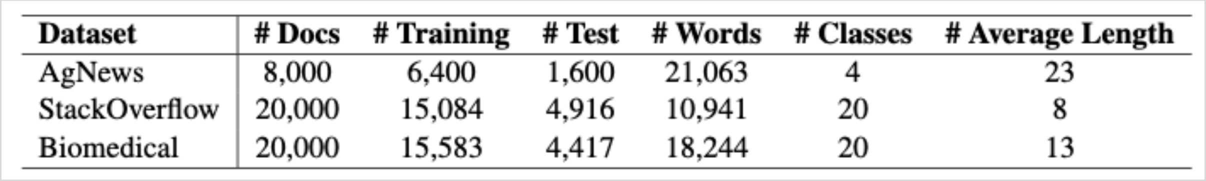
Nr. of documents: 20,000 (training and testing, 15,084 and 4,916 respectively)

Details: Challenge data released by Kaggle

Origin: Biomedical

Nr. of documents: 20,000

Details: Challenge data published in BioASQ Participants Area | BioASQ



Dataset	# Docs	# Training	# Test	# Words	# Classes	# Average Length
AgNews	8,000	6,400	1,600	21,063	4	23
StackOverflow	20,000	15,084	4,916	10,941	20	8
Biomedical	20,000	15,583	4,417	18,244	20	13

Document

Pre-processing

No mentioned pre-processing step.

```
@inproceedings{yin_2022_improving_deep_embedded_clustering_via_learning_cluster_level_representations,
    title = "Improving Deep Embedded Clustering via Learning Cluster-level Representations",
    author = "Yin, Qing  and
      Wang, Zhihua  and
      Song, Yunya  and
      Xu, Yida  and
      Niu, Shuai  and
      Bai, Liang  and
      Guo, Yike  and
      Yang, Xian",
    booktitle = "Proceedings of the 29th International Conference on
Computational Linguistics",
    month = oct,
    year = "2022",
    address = "Gyeongju, Republic of Korea",
    publisher = "International Committee on Computational Linguistics",
    url = "https://aclanthology.org/2022.coling-1.195",
    pages = "2226--2236",
    abstract = "Driven by recent advances in neural networks, various Deep
Embedding Clustering (DEC) based short text clustering models are being
developed. In these works, latent representation learning and text clustering
are performed simultaneously. Although these methods are becoming increasingly
popular, they use pure cluster-oriented objectives, which can produce
meaningless representations. To alleviate this problem, several improvements
have been developed to introduce additional learning objectives in the
clustering process, such as models based on contrastive learning. However,
existing efforts rely heavily on learning meaningful representations at the
instance level. They have limited focus on learning global representations,
which are necessary to capture the overall data structure at the cluster level.
In this paper, we propose a novel DEC model, which we named the deep embedded
clustering model with cluster-level representation learning (DECCRL) to jointly
learn cluster and instance level representations. Here, we extend the embedded
topic modelling approach to introduce reconstruction constraints to help learn
cluster-level representations. Experimental results on real-world short text
datasets demonstrate that our model produces meaningful clusters.",
}
```

#Thesis/Papers/Initial