Short Text Clustering via Convolutional Neural Networks

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, Hongwei Hao

Introduction

What we have:

- A corpus of short texts.

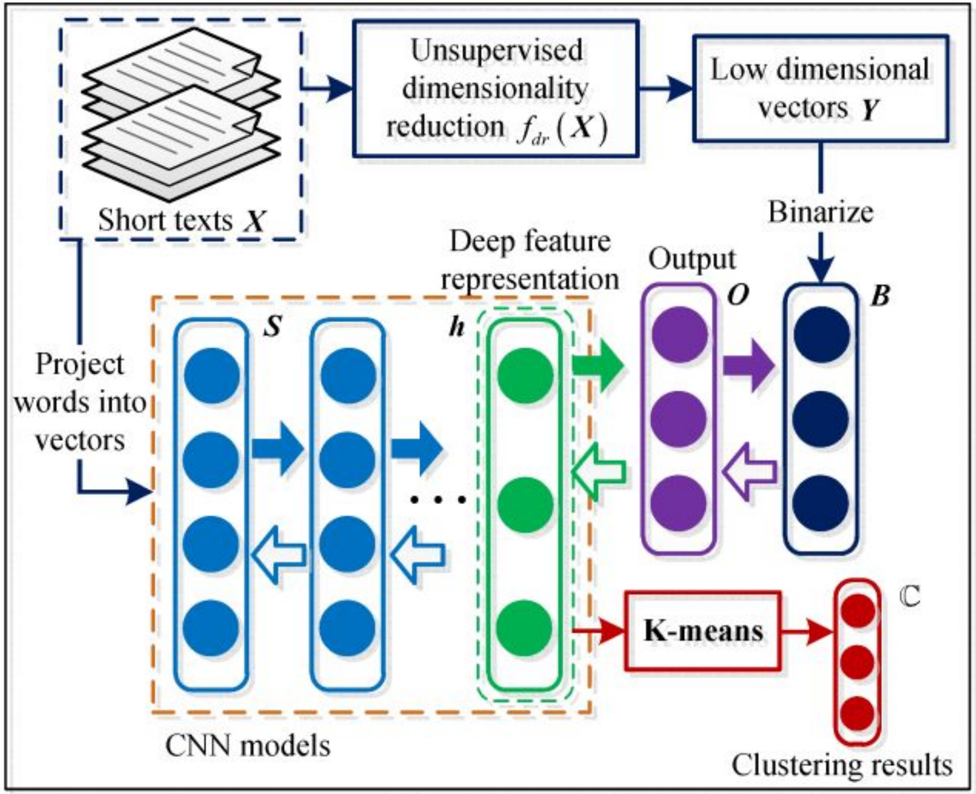
What we need:

- To make clustering of this corpus based on semantics.

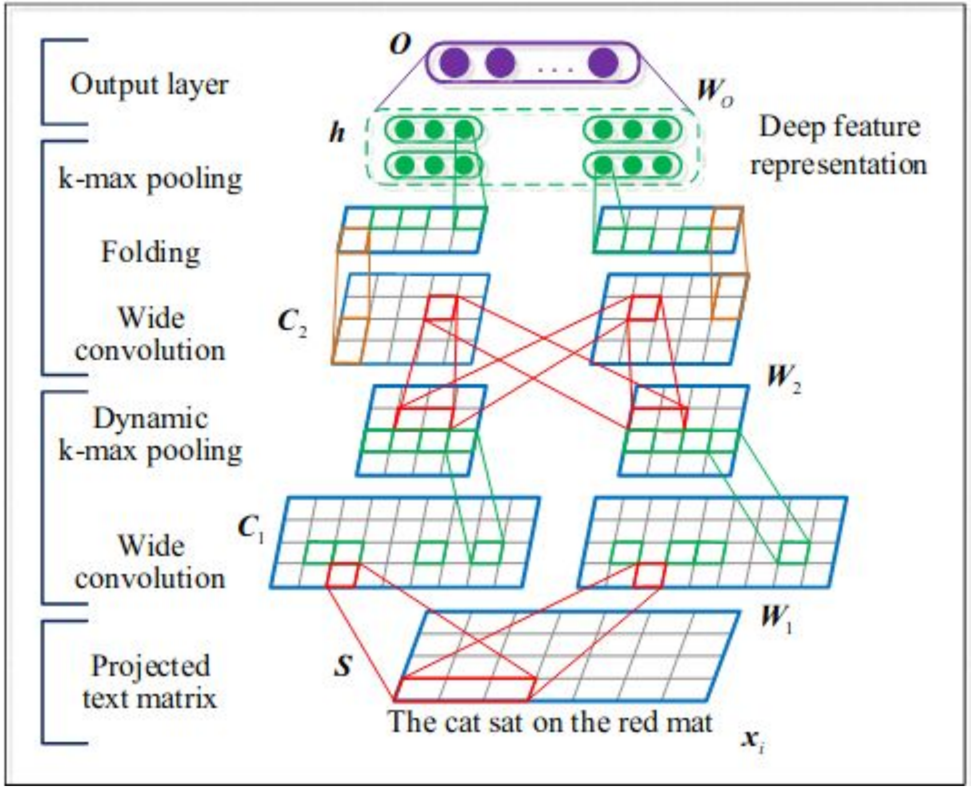
What problems we have:

- Due to sparseness of text we cannot use traditional approaches like TF-IDF.

Proposed architecture



CNN architecture



Unsupervised dimensionality reduction and binarization

Dimensionality reduction function is defined as follows:

$$\mathbf{Y} = f_{dr}(\mathbf{X}), \tag{2}$$

where, $\mathbf{Y} \in \mathbb{R}^{q \times n}$ are the q-dimensional reduced latent space representations.

We consider the following methods:

- Average Embedding (AE)
- Latent Semantic Analysis (LSA)
- Laplacian Eigenmaps (LE)
- Locality Preserving Indexing (LPI)

Experiments

Datasets we use:

Dataset	C	Num.	Len.	$\lvert V \rvert$
SearchSnippets	8	12,340	17.88/38	30,642
StackOverflow	20	20,000	8.31/34	22,956
Biomedical	20	20,000	12.88/53	18,888

Table 1: Statistics for the text datasets. C: the number of classes; Num: the dataset size; Len.: the mean/max length of texts and $\lvert V \rvert$: the vocabulary size.

Dataset topics

SearchSnippets: 8 different domains			
business	computers	health	education
culture	engineering	sports	politics
StackOverflow: 20 semantic tags			
svn	oracle	bash	apache
excel	matlab	cocoa	visual-studio
osx	wordpress	spring	hibernate
scala	sharepoint	ajax	drupal
qt	haskell	linq	magento
Biomedical: 20 MeSH major topics			
aging	chemistry	cats	erythrocytes
glucose	potassium	lung	lymphocytes
spleen	mutation	skin	norepinephrine
insulin	prognosis	risk	myocardium
sodium	mathematics	swine	temperature

Table 2: Description of semantic topics (that is, tags/labels) from the three text datasets used in our experiments.

Results

	SearchSnippets	StackOverflow	Biomedical
Method	ACC (%)	ACC (%)	ACC (%)
K-means (TF)	24.75±2.22	13.51±2.18	15.18±1.78
K-means (TF-IDF)	33.77±3.92	20.31±3.95	27.99±2.83
SkipVec (Uni)	28.23±1.08	08.79±0.19	16.44±0.50
SkipVec (Bi)	29.24±1.57	09.59±0.15	16.11±0.60
SkipVec (Combine)	33.58±1.95	09.34±0.24	16.27±0.33
RecNN (Top)	21.21±1.62	13.13±0.80	13.73±0.67
RecNN (Ave.)	65.59±5.35	40.79±1.38	37.05±1.27
RecNN (Top+Ave.)	65.53±5.64	40.45±1.60	36.68±1.29
Para2vec	69.07±2.53	32.55±0.89	41.26±1.22
STC^2-AE	68.34±2.51	40.05±1.77	37.44±1.19
STC^2-LSA	73.09±1.45	35.81±1.80	38.47±1.55
STC^2-LE	**77.09±3.99**	**51.13±2.80**	**43.62±1.00**
STC^2-LPI	**77.01±4.13**	**51.14±2.92**	43.00±1.25

Table 4: Comparison of ACC of our proposed methods and three clustering methods on three datasets. For RecNN (Top), K-means is conducted on the learned vectors of the top tree node. For RecNN (Ave.), K-means is conducted on the average of all vectors in the tree. More details about the baseline setting are described in Section 4.3

	SearchSnippets	StackOverflow	Biomedical
Method	ACC (%)	ACC (%)	ACC (%)
bi-LSTM (last)	64.50±3.18	46.83±1.79	36.50±1.08
bi-LSTM (mean)	65.85±4.18	44.93±1.83	35.60±1.21
bi-LSTM (max)	61.70±5.10	38.74±1.62	32.83±0.73
bi-GRU (last)	70.18±2.62	43.36±1.46	35.19±0.78
bi-GRU (mean)	70.29±2.61	44.53±1.81	36.75±1.21
bi-GRU (max)	65.69±1.02	**54.40±2.07**	37.23±1.19
LPI (best)	47.11±2.91	38.04±1.72	37.15±1.16
STC^2-LPI	**77.01±4.13**	51.14±2.92	**43.00±1.25**

Table 6: Comparison of ACC of our proposed methods and some other non-biased models on three datasets. For LPI, we project the text under the best dimension as described in Section 4.3. For both bi-LSTM and bi-GRU based clustering methods, the binary codes generated from LPI are used to guide the learning of bi-LSTM/bi-GRU models.

Results

	SearchSnippets	StackOverflow	Biomedical
Method	NMI (%)	NMI (%)	NMI (%)
K-means (TF)	09.03±2.30	07.81±2.56	09.36±2.04
K-means (TF-IDF)	21.40±4.35	15.64±4.68	25.43±3.23
SkipVec (Uni)	10.98±0.93	02.24±0.13	10.52±0.41
SkipVec (Bi)	09.27±0.29	02.89±0.20	10.15±0.59
SkipVec (Combine)	13.85±0.78	02.72±0.34	10.72±0.46
RecNN (Top)	04.04±0.74	09.90±0.96	08.87±0.53
RecNN (Ave.)	50.55±1.71	40.58±0.91	33.85±0.50
RecNN (Top+Ave.)	50.44±1.84	40.21±1.18	33.75±0.50
Para2vec	50.51±0.86	27.86±0.56	34.83±0.43
STC^2-AE	54.01±1.55	38.22±1.31	33.58±0.48
STC^2-LSA	54.53±1.47	34.38±1.12	33.90±0.67
STC^2-LE	**63.16±1.56**	**49.03±1.46**	38.05±0.48
STC^2-LPI	62.94±1.65	**49.08±1.49**	**38.18±0.47**

Table 5: Comparison of NMI of our proposed methods and three clustering methods on three datasets. For RecNN (Top), K-means is conducted on the learned vectors of the top tree node. For RecNN (Ave.), K-means is conducted on the average of all vectors in the tree. More details about the baseline setting are described in Section 4.3

	SearchSnippets	StackOverflow	Biomedical
Method	NMI (%)	NMI (%)	NMI (%)
bi-LSTM (last)	50.32±1.15	41.89±0.90	34.51±0.34
bi-LSTM (mean)	52.11±1.69	40.93±0.91	34.03±0.28
bi-LSTM (max)	46.81±2.38	36.73±0.56	31.90±0.23
bi-GRU (last)	56.00±0.75	38.73±0.78	32.91±0.40
bi-GRU (mean)	55.76±0.85	39.84±0.94	34.27±0.27
bi-GRU (max)	51.11±1.06	**51.10±1.31**	32.74±0.34
LPI (best)	38.48±2.39	27.21±0.88	29.73±0.30
STC^2-LPI	**62.94±1.65**	49.08±1.49	**38.18±0.47**

Table 7: Comparison of NMI of our proposed methods and some other non-biased models on three datasets. For LPI, we project the text under the best dimension as described in Section 4.3. For both bi-LSTM and bi-GRU based clustering methods, the binary codes generated from LPI are used to guide the learning of bi-LSTM/bi-GRU models.

Conclusion

With the emergence of social media, short text clustering has become an increasing important task. This paper explores a new perspective to cluster short texts based on deep feature representation learned from the proposed self-taught convolutional neural networks. Our extensive experimental study on three short text datasets shows that our approach can achieve a significantly better performance.

http://lvdmaaten.github.io/tsne/