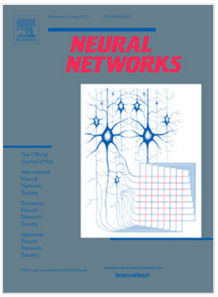
Accepted Manuscript



Self-Taught convolutional neural networks for short text clustering

Jiaming Xu, Bo Xu, Peng Wang, Suncong Zheng, Guanhua Tian, Jun Zhao, Bo Xu

PII: S0893-6080(16)30197-6
DOI: http://dx.doi.org/10.1016/j.neunet.2016.12.008
Reference: NN 3697

To appear in: *Neural Networks*

Received date: 25 November 2015
Revised date: 3 November 2016
Accepted date: 27 December 2016

Please cite this article as: Xu, J., Xu, B., Wang, P., Zheng, S., Tian, G., & Zhao, J., et al. Self-Taught convolutional neural networks for short text clustering. *Neural Networks* (2017), http://dx.doi.org/10.1016/j.neunet.2016.12.008

Self-Taught Convolutional Neural Networks for Short Text Clustering

Jiaming Xu[a], Bo Xu[a,*], Peng Wang[a], Suncong Zheng[a], Guanhua Tian[a], Jun Zhao[a,b], Bo Xu[a,c]

[a] *Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, P.R. China*
[b] *National Laboratory of Pattern Recognition (NLPR), Beijing, P.R. China*
[c] *Center for Excellence in Brain Science and Intelligence Technology, CAS. P.R. China*

Abstract

Short text clustering is a challenging problem due to its sparseness of text representation. Here we propose a flexible Self-Taught Convolutional neural network framework for Short Text Clustering (dubbed STC^2), which can flexibly and successfully incorporate more useful semantic features and learn non-biased deep text representation in an unsupervised manner. In our framework, the original raw text features are firstly embedded into compact binary codes by using one existing unsupervised dimensionality reduction methods. Then, word embeddings are explored and fed into convolutional neural networks to learn deep feature representations, meanwhile the output units are used to fit the pre-trained binary codes in the training process. Finally, we get the optimal clusters by employing K-means to cluster the learned representations. Extensive experimental results demonstrate that the proposed framework is effective, flexible and outperform several popular clustering methods when tested on three public short text datasets.

Keywords: Semantic Clustering, Neural Networks, Short Text, Unsupervised Learning

*Corresponding author. Email address: boxu@ia.ac.cn