Pseudo-supervised Approach for Text Clustering Based on Consensus Analysis

Peixin Chen, Wu Guo, Lirong Dai, Zhenhua Ling

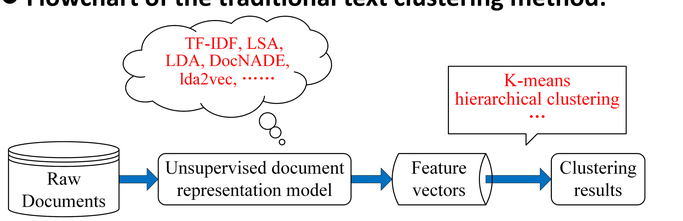
National Engineering Laboratory for Speech and Language Information Processing
University of Science and Technology of China, Hefei, China

Abstract

- Compared with unsupervised models, the supervised neural networks usually generate more discriminative features.
- In text clustering, a typical unsupervised task, there are no predefined labels for modeling training.
- To benefit from the discriminative ability of the supervised neural networks, we propose a pseudo-supervised approach for text clustering.

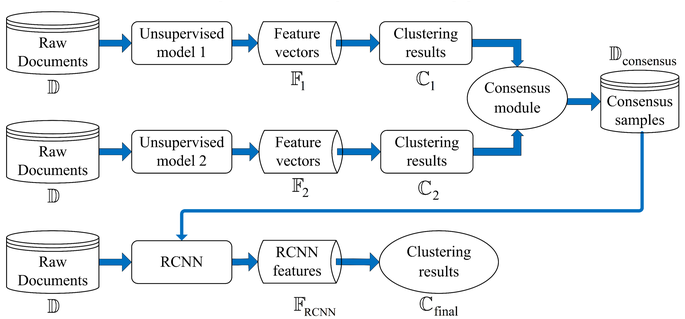
Background

- Flowchart of the traditional text clustering method:



Pseudo-supervised Approach

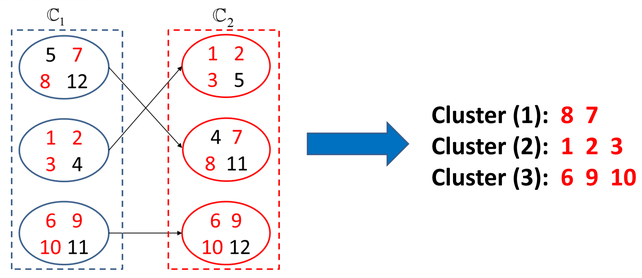
- Flowchart of the pseudo-supervised approach:



➢ **Step 1: Consensus Samples Generation:**

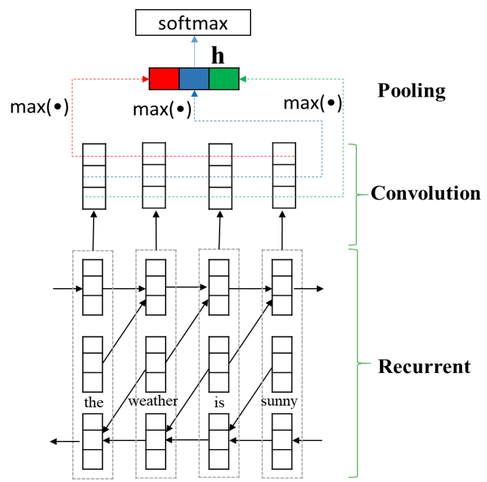
a) Align the cluster labels of two pre-clusterings $\mathbb{C}_1$ and $\mathbb{C}_2$ by Kuhn-Munkres algorithm. (Maximum Weighted Bipartite Matching)

b) The documents with consistent cluster labels in $\mathbb{C}_1$ and $\mathbb{C}_2$ are selected.



➢ **Step 2: RCNN training:**

The RCNN is trained on the consensus samples $\mathbb{D}_{consensus}$, using the cluster labels as pseudo-labels for training.



➢ **Step 3: Clustering:**

We utilize the trained RCNN to obtain the semantic features $\mathbf{h}$ for clustering.

Experiments

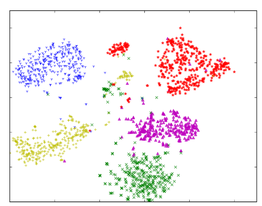
- Dataset: Fisher English corpus (released by LDC), which contains 11699 documents and involves 40 topics in total.
- Experimental results:

Table 1. ACC(%)/NMI(%) of baseline systems

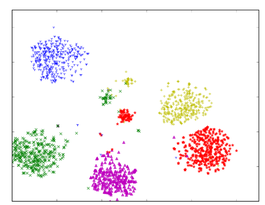
	30	40	50
LSA	69.08/78.29	80.43/81.29	76.39/80.14
LDA	72.34/78.95	**81.42**/81.97	74.85/80.13
DocNADE	**74.20/80.93**	80.70/**82.47**	**77.28/81.40**
lda2vec	72.74/79.77	80.37/80.91	74.95/79.73
TF-IDF	64.52/71.62	73.05/74.34	69.53/74.23

Table 2. ACC(%)/NMI(%) of pseudo-supervised frameworks

	30	40	50
LSA-LDA	76.39/84.14	89.19/89.65	84.80/**89.09**
LSA-lda2vec	76.61/85.12	**90.75/91.17**	83.20/88.81
LDA-DocNADE	76.66/84.37	89.64/90.09	**84.84**/88.18
DocNADE-lda2vec	**77.50/86.06**	87.43/89.36	84.24/88.39
LDA	72.33/78.96	81.76/82.38	75.18/80.48
DocNADE	74.28/81.07	80.81/82.65	77.41/81.73



(a) t-SNE of LDA features



(b) t-SNE of LSA-LDA pseudo-supervised features