清華大學
Tsinghua University

Deep Learning in Exploring Semantic Relatedness for Microblog Dimensionality Reduction

Lei Xu, Chunxiao Jiang, Yong Ren

Department of Electronic Engineering, Tsinghua University,
Beijing, 100084, China

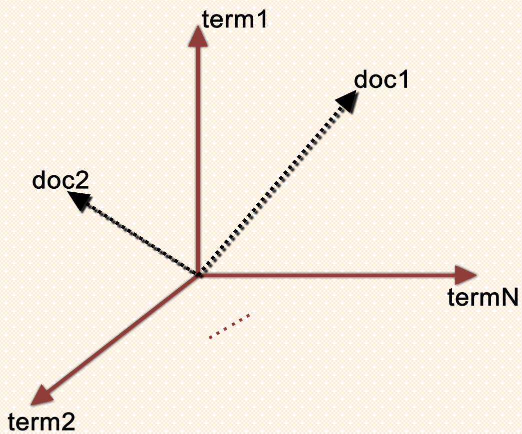
presented by ***Can Li***

Contents

- Introduction
- Basics of Deep Networks
- Tailor Deep Networks To Tweets
- Experiments
- Conclusion

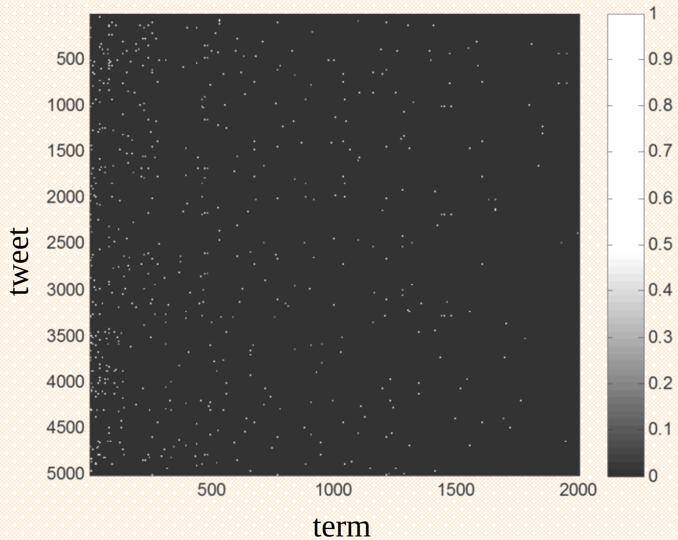
Introduction

- **Microblogging Services**: Twitter, Sina Weibo
- **Mining Microblog Text (Tweet)**
  - Text representation: vector space model[1]
  - Short length: data sparse problem



	term1	term2	term3	…	termN
doc1	1	0	5	…	3
doc2	0	2	4	…	0

vector space model



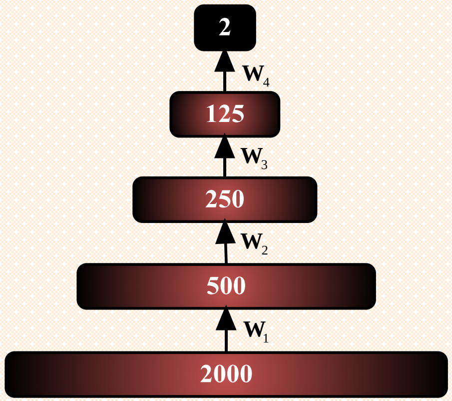
document-term matrix (normalized)

Introduction

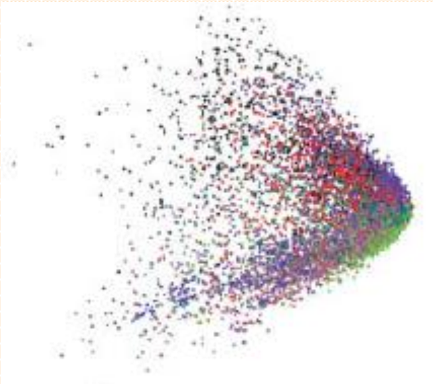
- **Solutions for Short Text**
  - **Expanding:** add semantically related terms [2][3]
  - **Dimension reduction**
    - Latent semantic analysis (LSA) [4]
    - Topic modeling
      - Low-dimensional representation: probability distribution over latent topics
      - Latent Dirichlet allocation (LDA) [11] and its variants [5][6]
      - Problem of topic-based representation: both the number of topics and the content of topics change frequently in microblog environment

Introduction

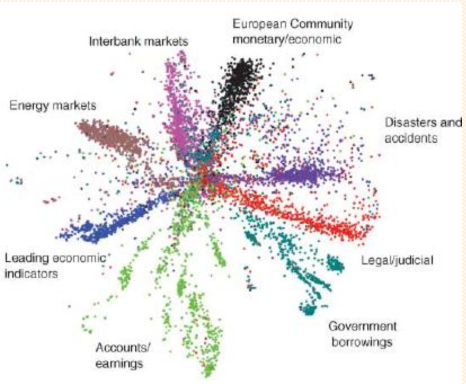
- **Deep Networks-based Dimensionality Reduction** [7~10]



deep netwroks



LSA



deep netwroks

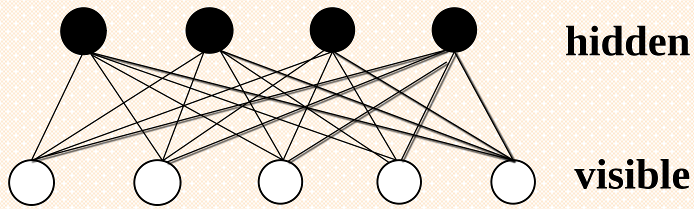
2-dimension embeddings of text data

- **Basic Idea of the Proposed Approach**: utilize the semantic relatedness derived from retweet and hashtags

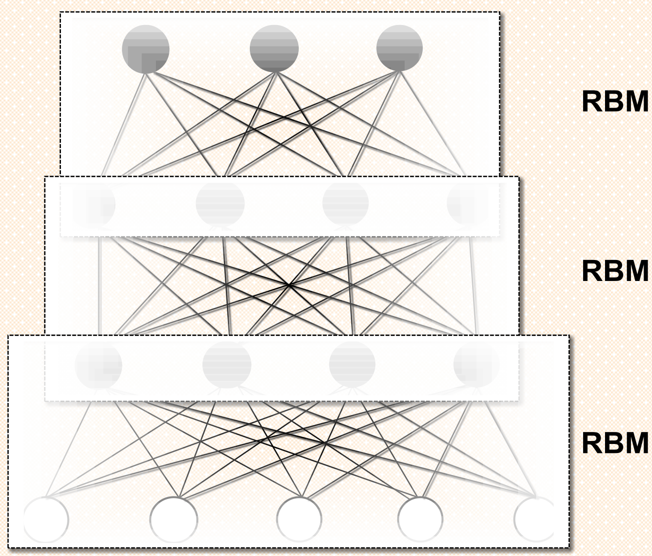
If one tweet is created by retweeting another tweet, or two tweets are labeled with the same hashtag, then the two tweets are semantically similar, or at least, related.

Basics of Deep Networks

- **Deep Belief Networks**
  - Restricted Boltzmann Machines

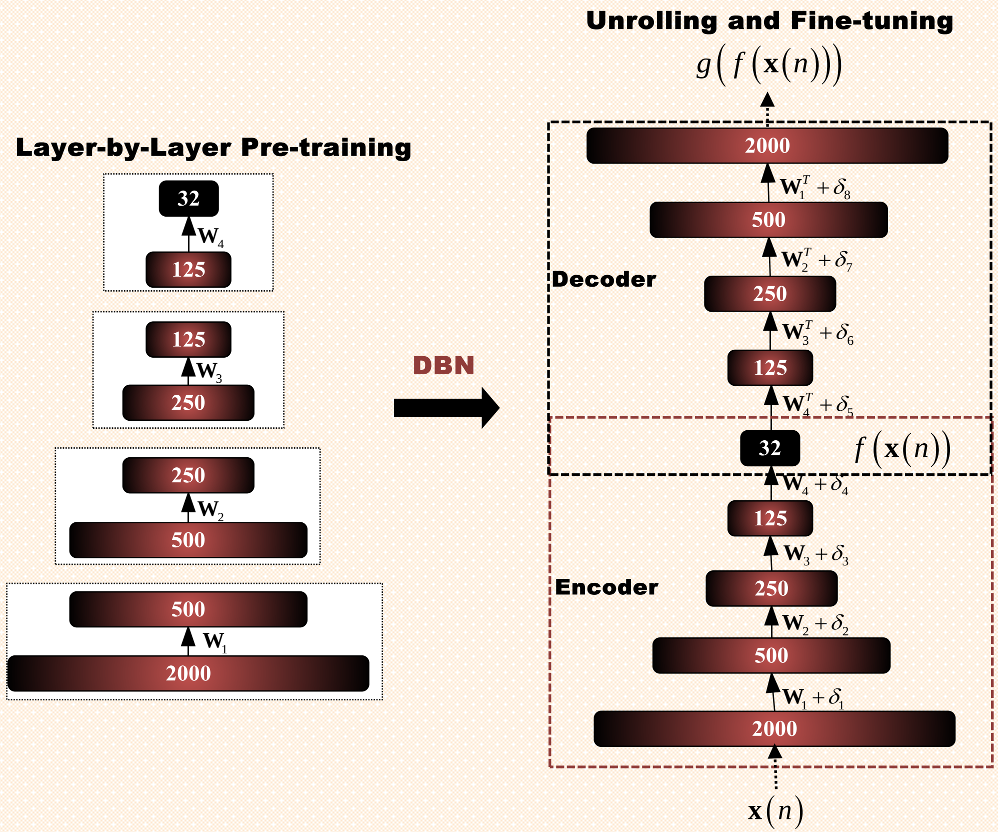


  - Stack of RBMs: layer-by-layer training



Basics of Deep Networks

- **Deep Autoencoder**
  - Pre-training: layer-by-layer
  - Fine-tuning: minimize the reconstruction error l_{AE}



$$l_{AE}=\frac{1}{N}\sum_{n=1}^{N}\left\|\mathbf{x}(n)-g\left(f\left(\boldsymbol{x}(n)\right)\right)\right\|^2$$

Tailor Deep Networks to Tweets

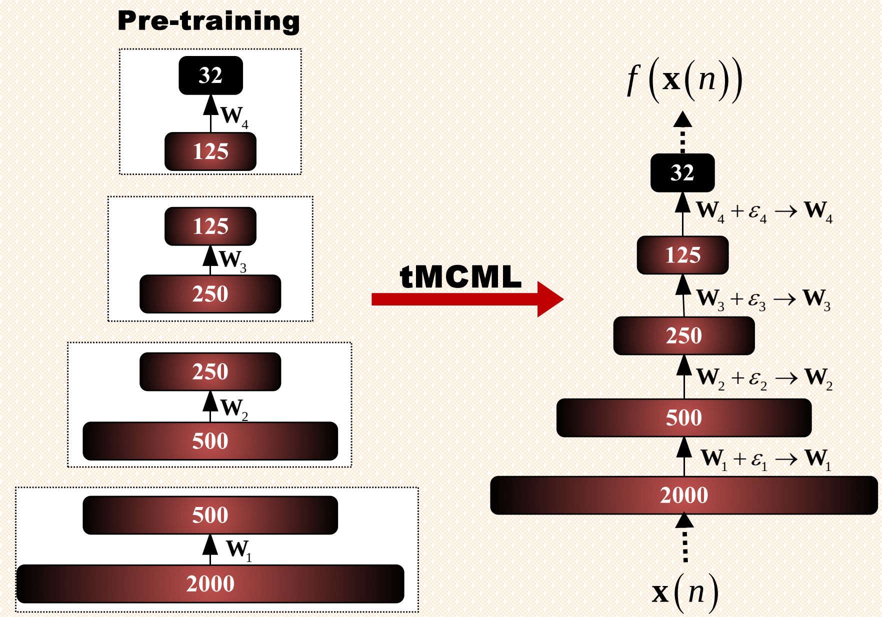
- **Basics of t-distributed Maximally Collapsing Metric Learning[12]**
  - Learns a mapping function $f(\cdot)$ from high-dimensional space to low-dimensional space
  - Supervised learning: (data, label)
  - Two probability distributions
    - $P = \{p_{ij}\}$: $p_{ij} > 0$ iff $\mathbf{x}(i)$ and $\mathbf{x}(j)$ belong to the same class
    - $Q = \{q_{ij}\}$: normalized t-distribution

$$q_{ij} = \frac{\left(1+\frac{d_{ij}^2}{\alpha}\right)^{-\frac{1+\alpha}{2}}}{\sum_{k,l:k\neq l}\left(1+\frac{d_{kl}^2}{\alpha}\right)^{-\frac{1+\alpha}{2}}}, \qquad q_{ii} = 0 \qquad d_{ij}^2 = \left\| f\left(\mathbf{x}(i)\right) - f\left(\mathbf{x}(j)\right)\right\|^2$$

    - q_{ij}: similarity in low-dimensional space
    - p_{ij}: ground truth of the similarity
  - Training objective: minimize $l_{tMCML} = KL(P \| Q) = \sum_i \sum_{j:j\neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$

Tailor Deep Networks to Tweets

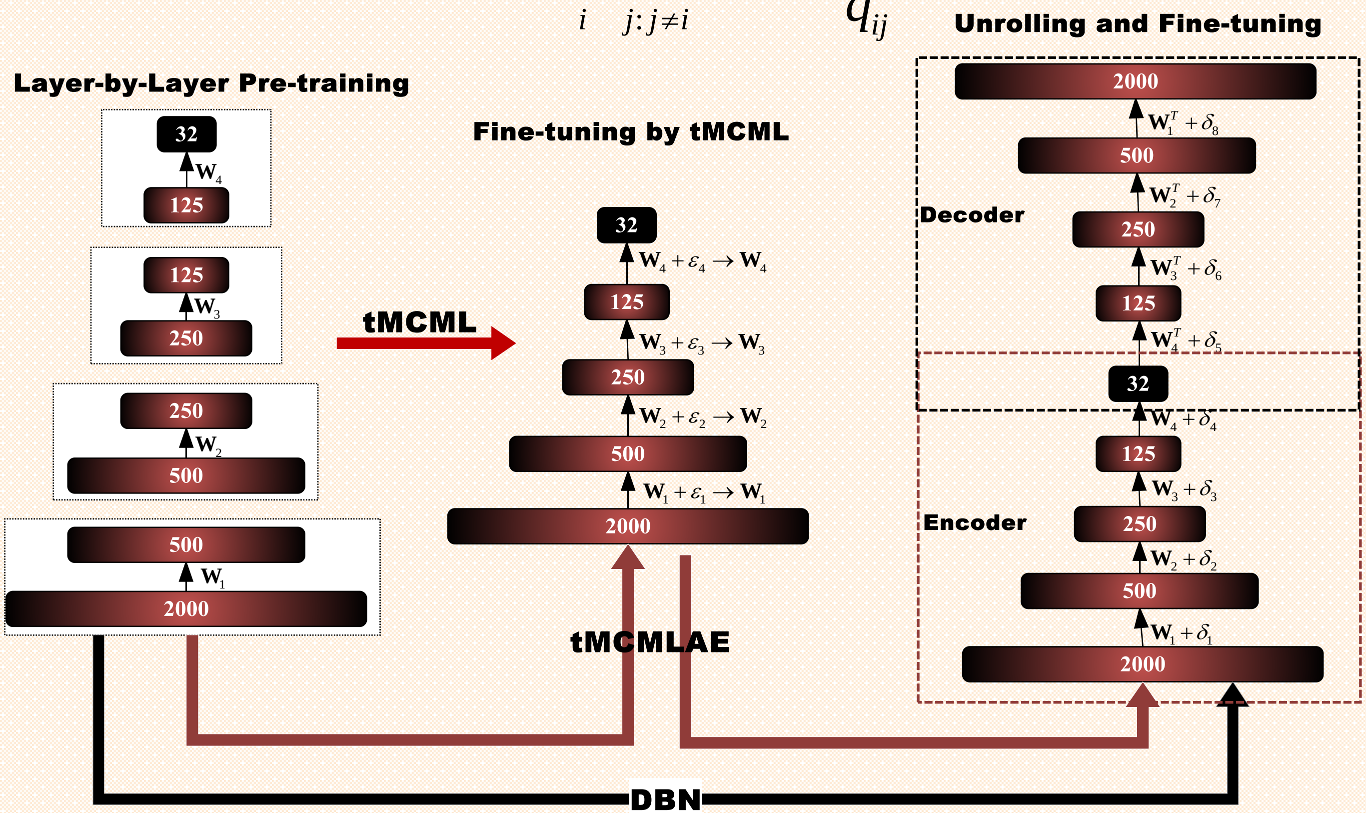
- **Apply tMCML to Tweets**
  - Supervised learning: (data, ~~label~~)
  - Define p_{ij}
    - Observation: two tweets that hold a *retweet* relationship or share the same *hashtag* are semantically similar
    - Indicator $\delta_{ij} = \begin{cases} 1, \mathbf{x}(i) \to \mathbf{x}(j) \vee \mathbf{x}(j) \to \mathbf{x}(i) \vee \#\mathbf{x}(i) = \#\mathbf{x}(j) \\ 0, \, else \end{cases}$
    - $p_{ij} = \frac{\delta_{ij}}{\sum_{kl:k\neq l} \delta_{kl}}$
  - Fine-tuning by tMCML



Tailor Deep Networks to Tweets

- **Double Fine-tuning**
  - What if only a small fraction of training samples are involved in a retweet relationship or labeled with hashtags?

$$l_{tMCML} = \sum_i \sum_{j:j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$



Experiments

- **DataSet**
  - Source: Sina Weibo
  - Original representation: term frequency vector, 2000 most frequent terms

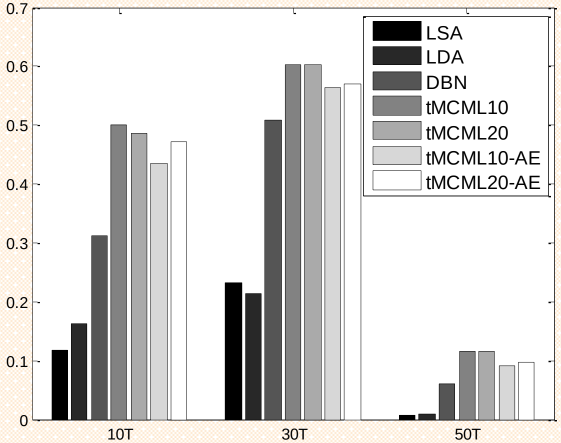
Test Set	Tweets	Topics	Avg Length of Tweets (Term)	Percentage of Non-zero Elements in the document-term matrix
10T	500	10	26.12	0.415%
30T	1500	30	27.32	0.416%
50T	2500	50	27.52	0.428%
Training Set	25750	~500	23.51	0.414%

Experiments

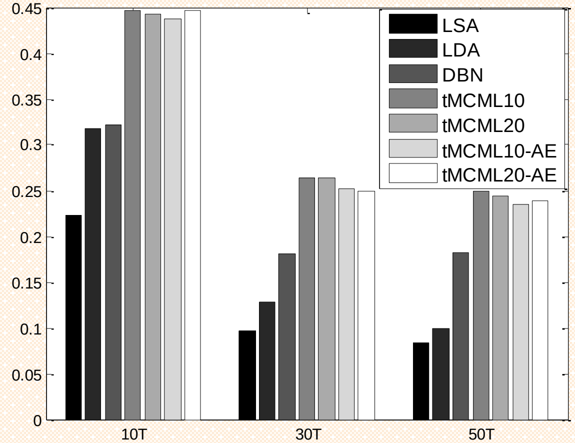
- **Experiment Setup**
  - **Deep Models**
    - Architecture: $2000\text{-}\overbrace{500\text{-}250\text{-}125}^{\text{logistic}}\text{-}\overset{\text{linear}}{\uparrow 32}$
    - **DBN**: pre-training 10 epochs, fine-tuning 20 epochs
    - **tMCML10/tMCML20:** tMCML-based fine-tuning 10/20 epochs
    - **tMCML10-AE/tMCML20-AE:** fine-tuning tMCML10/tMCML20 for 20 epochs
  - **Reference Models**
    - **LSA** (latent semantic analysis): 32 latent concepts
    - **LDA** (latent Dirichlet allocation): 32 latent topics

Experiments

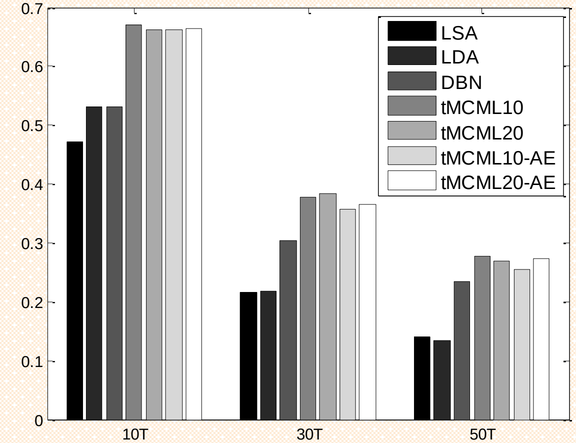
- **Evaluation Metrics**
  - Cluster analysis on low-dimensional representations: *k*-means
  - Cluster evaluation indices [13][14]
    - Adjust Rand Index (ARI)
    - Joint Normalized Mutual Information (NMI)
    - Set Matching F1-measure(SM-f1)
- **Results**



ARI



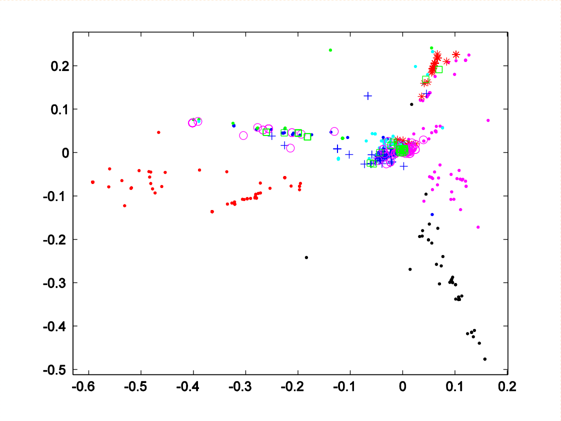
NMI



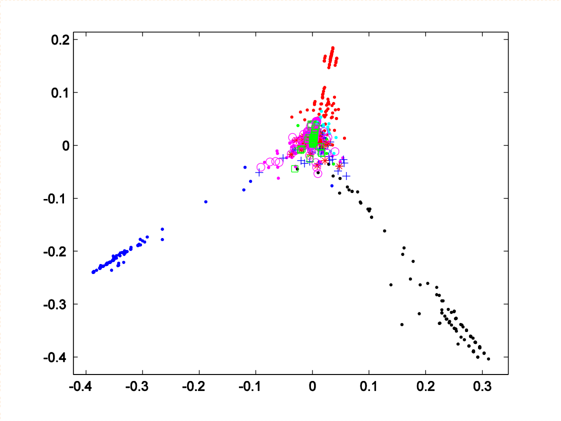
SF-M1

Experiments

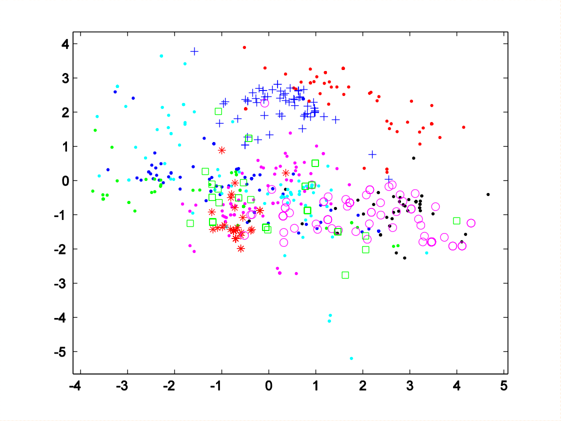
- **Discussion: Advantages of Deep Models**



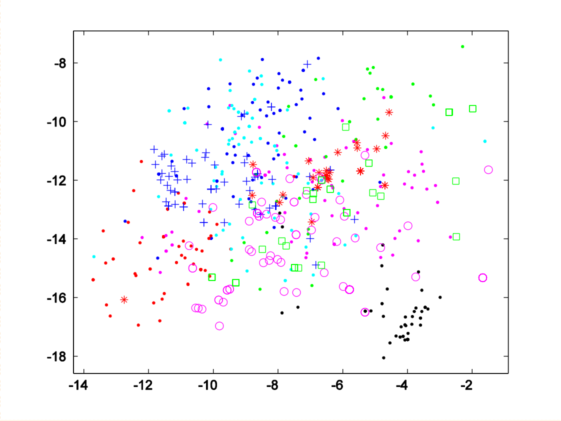
LSA



LDA



DBN



tMCML20

LSA: linear dimension reduction

LDA: fixed topics

Deep Networks: less insensitive to the change of topics

Experiments

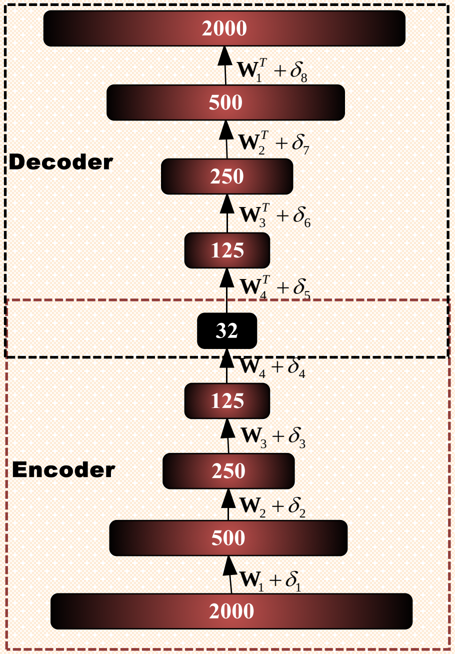
- **Discussion: Advantages of tMCML**

unsupervised

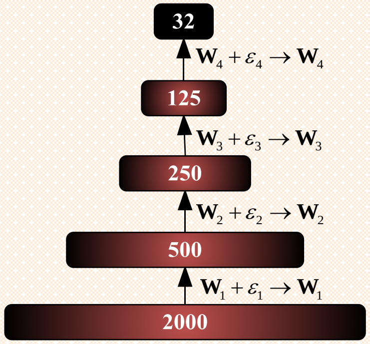
$$l_{AE} = \frac{1}{N}\sum_{n=1}^{N}\left\|\mathbf{x}(n) - g\left(f\left(\mathbf{x}(n)\right)\right)\right\|^2$$

semi-supervised

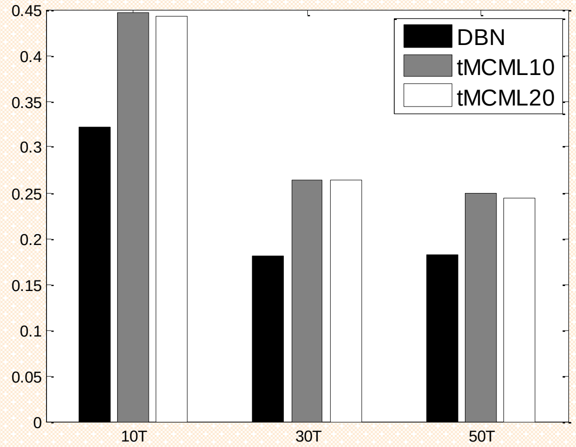
$$l_{tMCML} = \sum_{i}\sum_{j:j\neq i} p_{ij}\log\frac{p_{ij}}{q_{ij}}$$



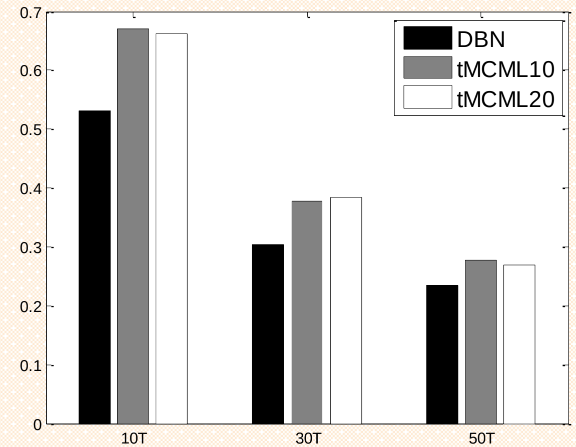
autoencoder



tMCML



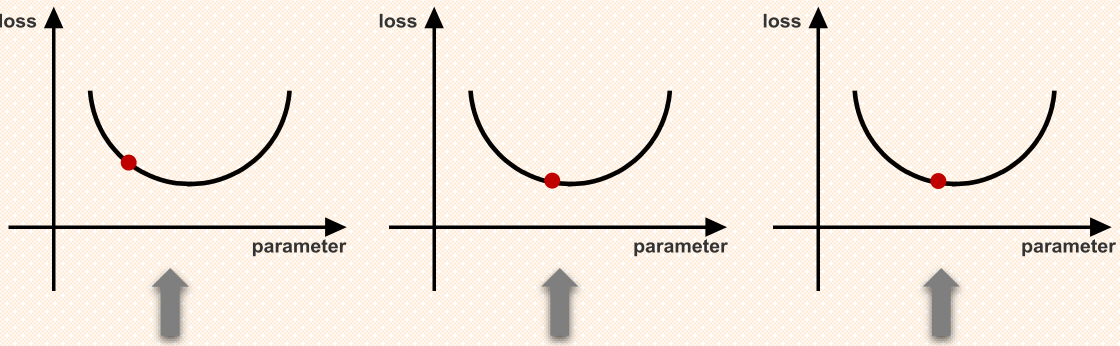
NMI

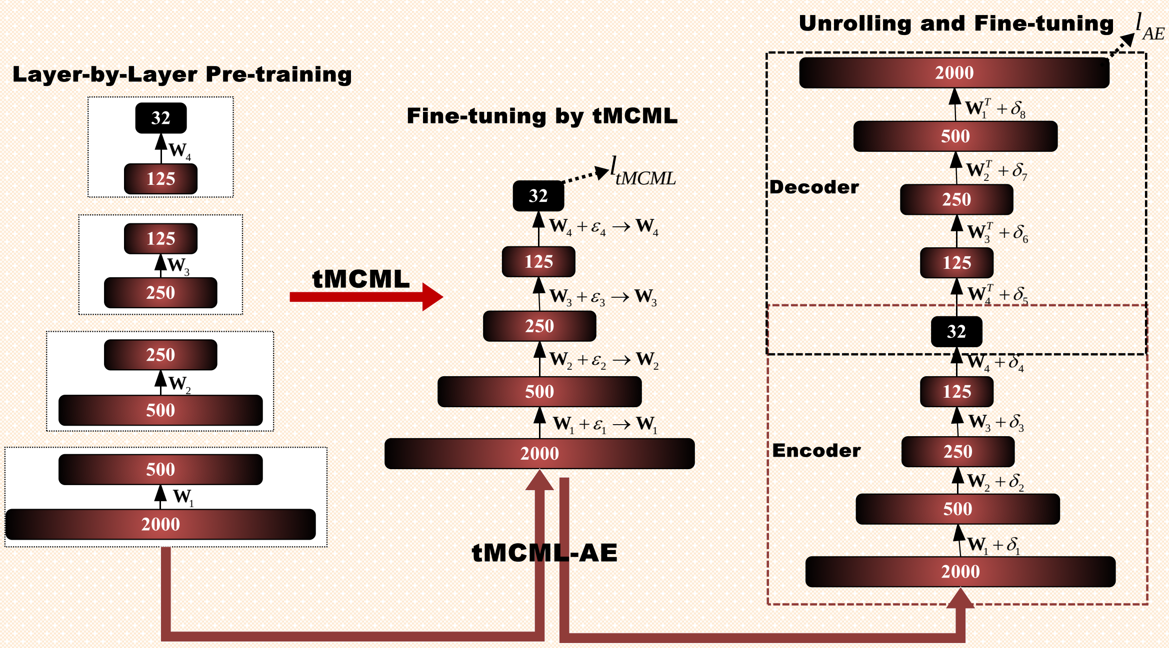


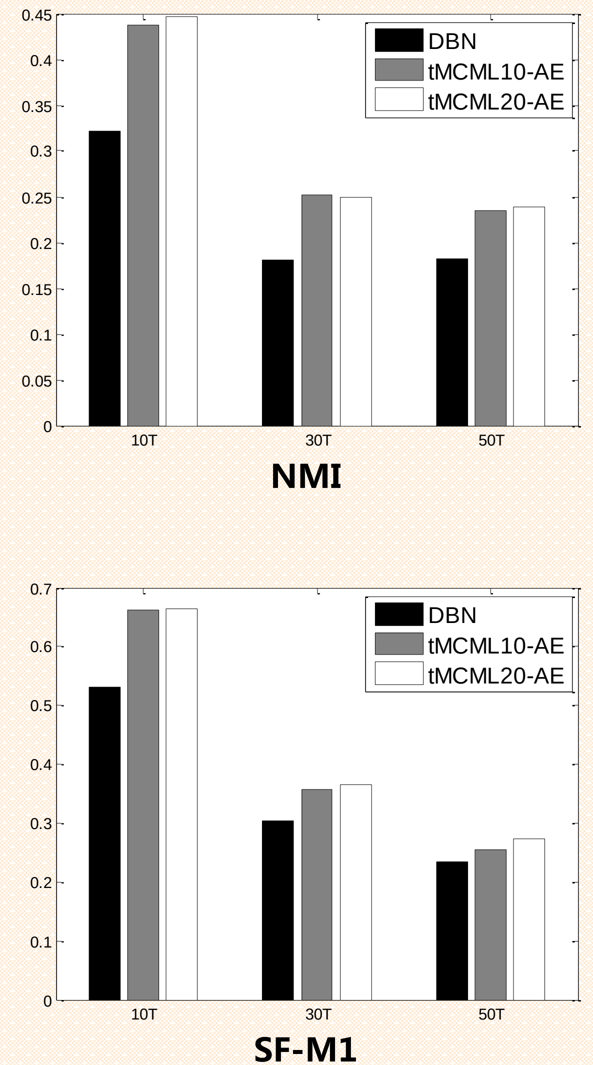
SF-M1

Experiments

- **Discussion: Importance of Pre-training**







Conclusion

- **Microblog Dimensionality Reduction**
  - Deep networks-based model
  - Semantic relatedness: *retweet*, *#hashtags*

- **Future Work**
  - Representations towards specific microblog mining tasks (e.g. sentiment classification)
  - Other types of meta-information in microblogs (e.g. embedded links)

Thanks a lot for your attention!

We'd like to thank the committees of GlobalSIP 2015 for providing us the great opportunity to share our study with professional colleagues !

References

[1] G. Salton, A. Wong, and C.-S. Yang, “A vector space model for automatic indexing,” Communications of the ACM, vol. 18, no. 11, pp. 613–620, 1975.

[2] Y. Xi-Wei, “Feature extension for short text,” in Proceedings of the Third International Symposium on Computer Science and Computational Technology, 2010, pp. 338–341.

[3] X. Hu, L. Tang, and H. Liu, “Enhancing accessibility of microblogging messages using semantic knowledge,” in Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011, pp. 2465–2468.

[4] X. Yan and H. Zhao, “Chinese microblog topic detection based on the latent semantic analysis and structural property,” Journal of Networks, vol. 8, no. 4, pp. 917–923, 2013.

[5] D. Ramage, S. T. Dumais, and D. J. Liebling, “Characterizing microblogs with topic models,” in Proceedings of the 4th International Conference on Weblogs and Social Media, 2010, pp. 130–137.

[6] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang, “Transferring topical knowledge from auxiliary long texts for short text clustering,” in Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ser. CIKM ’11, New York, NY, USA, 2011, pp. 775–784.

[7] L. Deng, “A tutorial survey of architectures, algorithms, and applications for deep learning,” APSIPA Transactions on Signal and Information Processing, vol. 3, p. e2, 2014.

[8] G. Hinton and R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science, vol. 313, no. 5786, pp. 504–507, Jul 2006.

[9] R. Salakhutdinov and G. Hinton, “Semantic hashing,” Int. J. Approx. Reasoning, vol. 50, no. 7, pp. 969–978, Jul 2009.

[10] S. Zhou, Q. Chen, and X. Wang, “Active deep learning method for semi-supervised sentiment classification,” Neurocomputing, vol. 120, pp. 536–546, 2013.

[11] D. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2001.

[12] M. R. Min, L. Maaten, Z. Yuan, A. J. Bonner, and Z. Zhang, “Deep supervised t-distributed embedding,” in Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 791–798.

[13] N. X. Vinh, J. Epps, and J. Bailey, “Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance.” Journal of Machine Learning Research, vol. 11, pp. 2837–2854, 2010.

[14] E. Amig´ o, J. Gonzalo, J. Artiles, and F. Verdejo, “A comparison of extrinsic clustering evaluation metrics based on formal constraints,” Inf. Retr., vol. 12, no. 4, pp. 461–486, Aug 2009.

[15] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, “Why does unsupervised pre-training help deep learning?” J. Mach. Learn. Res., vol. 11, pp. 625–660, Mar 2010.