A Latent Concept Topic Models for Robust Topic Inference Using Word Embeddings

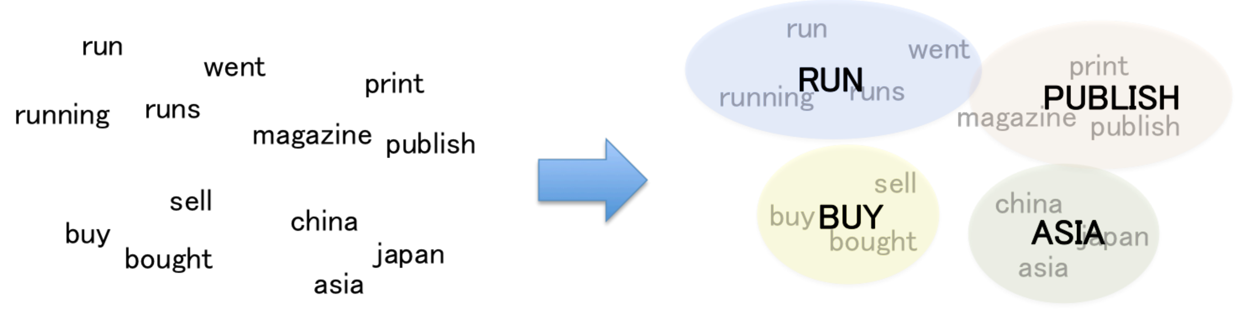
Weihua Hu (Utokyo) and Jun'ichi Tsujii (AIST)

Motivation

- **Document-level word co-occurrence is scarce when texts are short and vocabulary is diverse** (e.g. blog, SNS, newsgroup).
- **Probabilistic topic models** (e.g., LDA, pLSI) infers topics based on **document-level word co-occurrence.**

→ Conventional topic models are not effective.

→ Propose a novel topic model based on **co-occurrence statistics of latent concepts** to resolve the data sparsity.



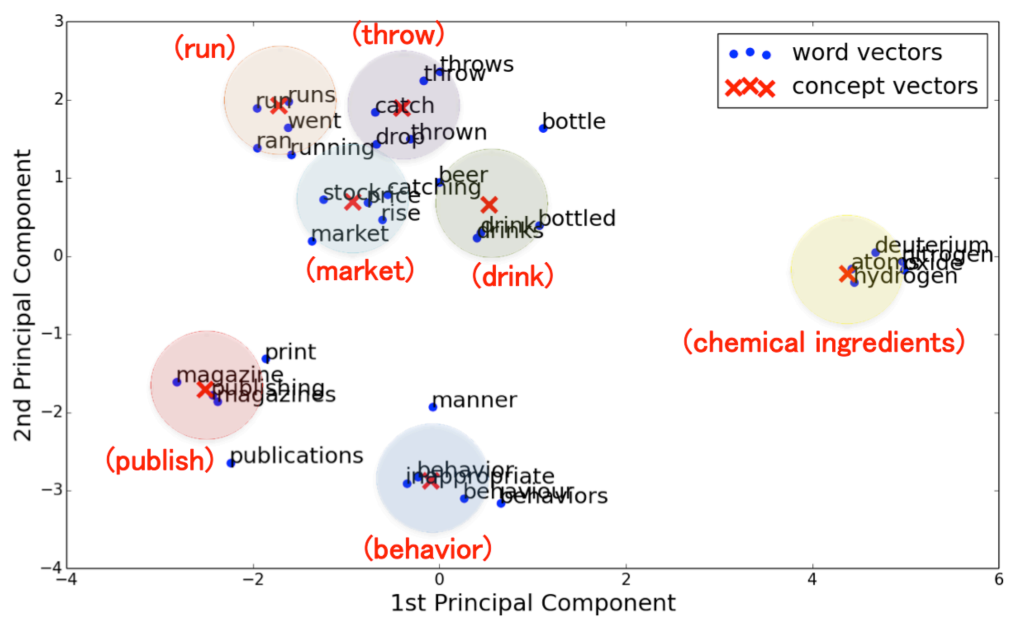
Surface words
Large #(vocabulary)
Small #(co-occurrence)

Latent concepts
Small #(concepts)
Large #(co-occurrence)

Proposal

- Use **Neural word embedding** (e.g., word2vec, Glove) to capture **conceptual similarity of words.**

→ Each cluster corresponds to one **latent concept.**



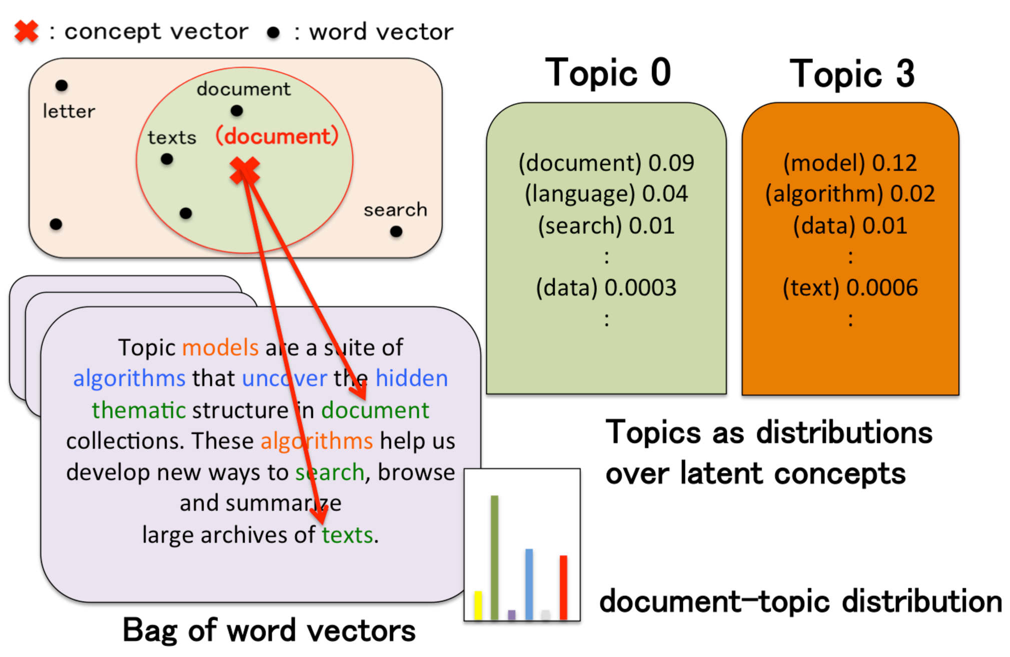
GloVe vectors (Pennington+ 2014)

- Define **topics as distributions over latent concepts.**

→ **Resolve data sparsity in short texts.**

- Model the **generative process of word embeddings.**

→ LCTM can naturally **handle Out of Vocabulary (OOV) words.**

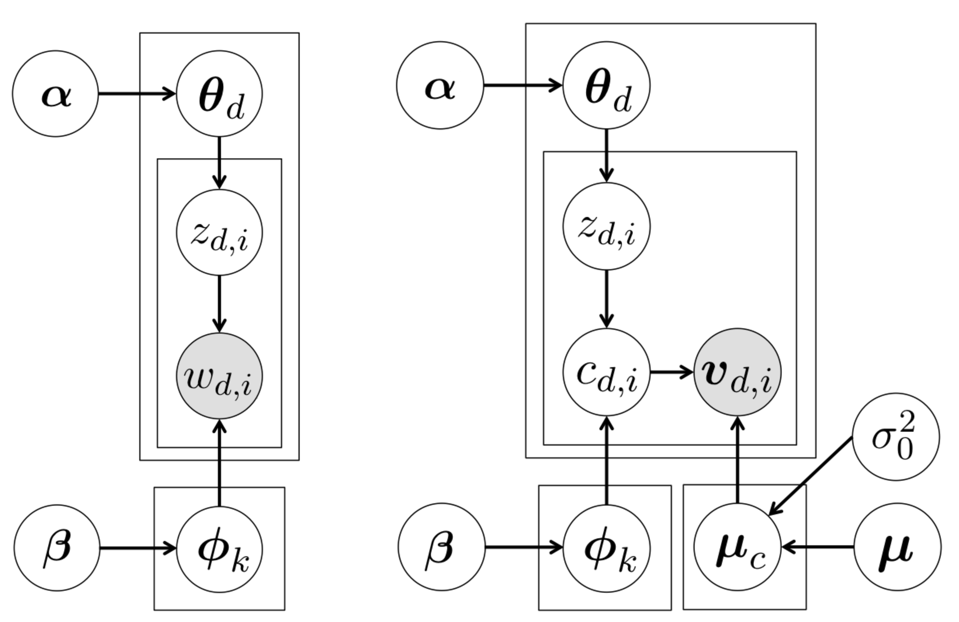


Latent Concept Topic Model (LCTM)

Gaussian variance parameter σ^2 **controls the range of the emission.**

Graphical Models

Add another layer of latent variables (**latent concepts**) to **mediate data sparsity**.



(a) LDA. (b) LCTM.

Notations

- α β : Dirichlet prior parameters
- μ σ_0^2 : Gaussian prior parameters
- θ_d : document-topic distribution
- ϕ_k : topic-concept (word) distribution
- $w_{d,i}$: word type
- $v_{d,i}$: word vector
- $z_{d,i}$: latent topic
- $c_{d,i}$: latent concept
- μ_c : concept vector

Overview of topic inference

- **Collapsed Gibbs sampler** for the approximate inference.
- Sample **latent concepts** in addition to **topics**.

・Sampling of a topic assignment

$$p(z_{d,i}=k \mid c_{d,i}=c, \boldsymbol{z}^{-d,i}, \boldsymbol{c}^{-d,i}, \boldsymbol{v}) \propto \left(n_{d,k}^{-d,i}+\alpha_k\right) \cdot \frac{n_{k,c}^{-d,i}+\beta_c}{n_{k,\cdot}^{-d,i}+\sum_{c'}\beta_{c'}}$$

Prop of topic k in the same doc — Prob of topic k generating concept c

・Sampling of a concept assignment

$$p(c_{d,i}=c \mid z_{d,i}=k, \boldsymbol{v}_{d,i}, \boldsymbol{z}^{-d,i}, \boldsymbol{c}^{-d,i}, \boldsymbol{v}^{-d,i}) \propto \left(n_{k,c}^{-d,i}+\beta_c\right) \cdot \mathcal{N}(\boldsymbol{v}_{d,i} \mid \overline{\boldsymbol{\mu}}_c, \sigma_c^2 \boldsymbol{I})$$

Prob of topic k generating concept c — Prob of concept c generating word vec v

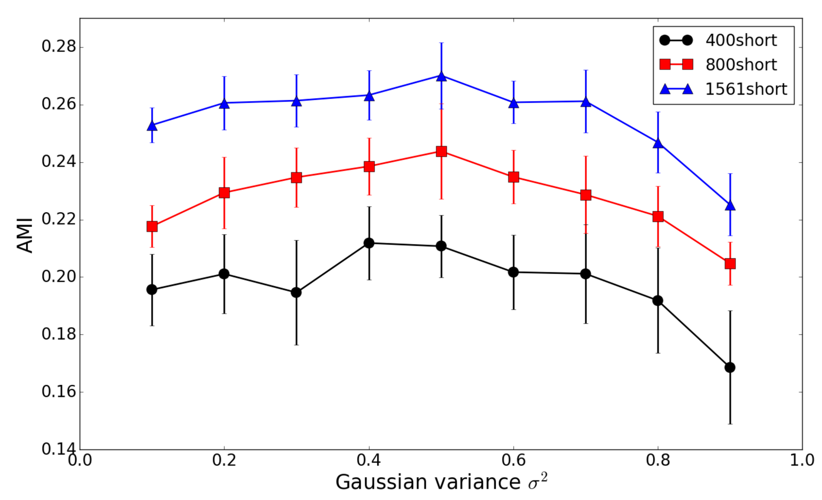
$\mathcal{N}(\cdot \mid \overline{\boldsymbol{\mu}}_c, \sigma_c^2 \boldsymbol{I})$: Gaussian distribution corresponding to latent concept c

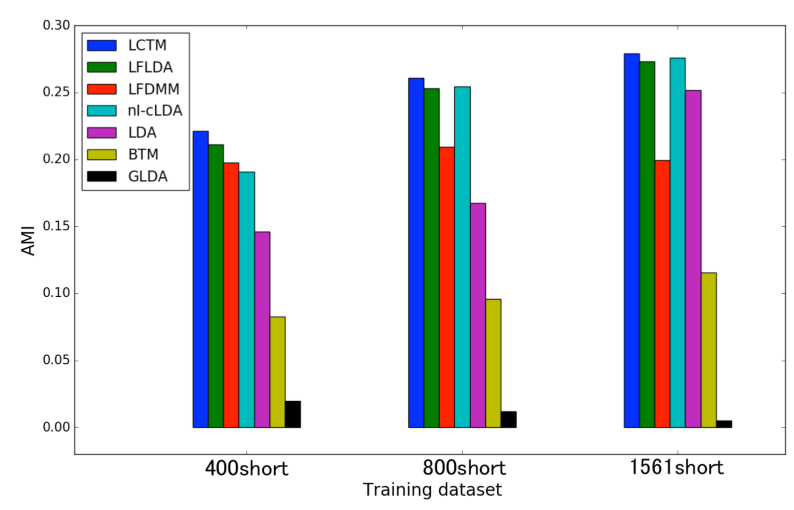
Experimental Results

Dataset: Short posts (less than 50 words) of 20Newsgroup.

1. Performance on document clustering

- Gaussian variance with $\sigma^2 = 0.5$ consistently performs well.
- LCTM outperforms TM w/o word embeddings.
- LCTM performs comparable to TM w/ word embeddings.



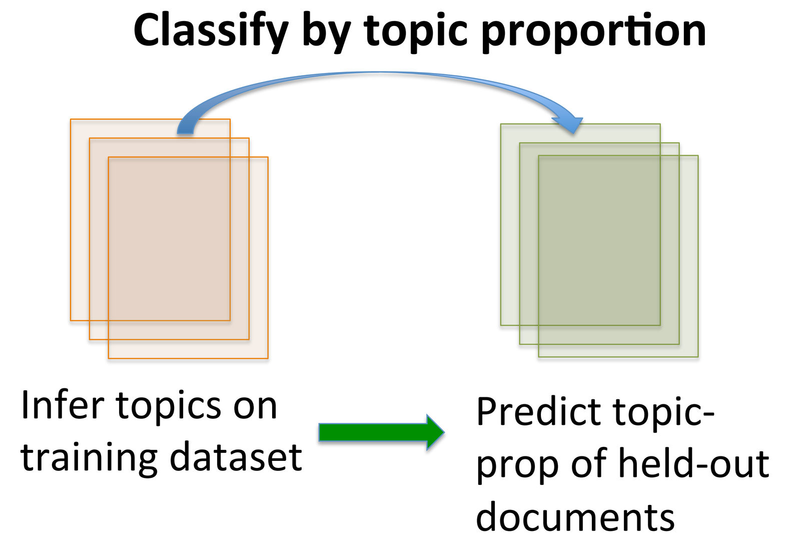


Clustering performance measured by Adjusted Mutual Information (AMI)

2. Performance on handling a high degree of OOV words

- LCTM-UNK (LCTM that ignores OOV) outperforms other TMs.
- LCTM further improves performance of LCTM-UNK.

→ **LCTM effectively incorporates OOV words** in held-out documents.



Experimental setting

Training Set	400short	800short	1561short
OOV prop	0.348	0.253	0.181
Method	Classification Accuracy		
LCTM	**0.302**	**0.367**	**0.416**
LCTM-UNK	0.262	0.340	0.406
LFLDA	0.253	0.333	0.410
nI-cLDA	0.261	0.333	0.412
LDA	0.215	0.293	0.382
GLDA	0.0527	0.0529	0.0529
Chance Rate	0.0539	0.0539	0.0539

Classification accuracy on held-out documents

Conclusion

- Introduced LCTM that infers topics based on document-level **co-occurrence of latent concepts**.
- Showed that LCTM can effectively **handle OOV words in held-out documents.**
- The same method can be readily applied to topic models that extend LDA.