Tiny Word Embeddings Using Globally Informed Reconstruction

Sora Ohashi, Mao Isogawa, Tomoyuki Kajiwara, Yuki Arase (Osaka University) ohashi.sora@ist.osaka-u.ac.jp

Background: Word Embedding Reconstruction

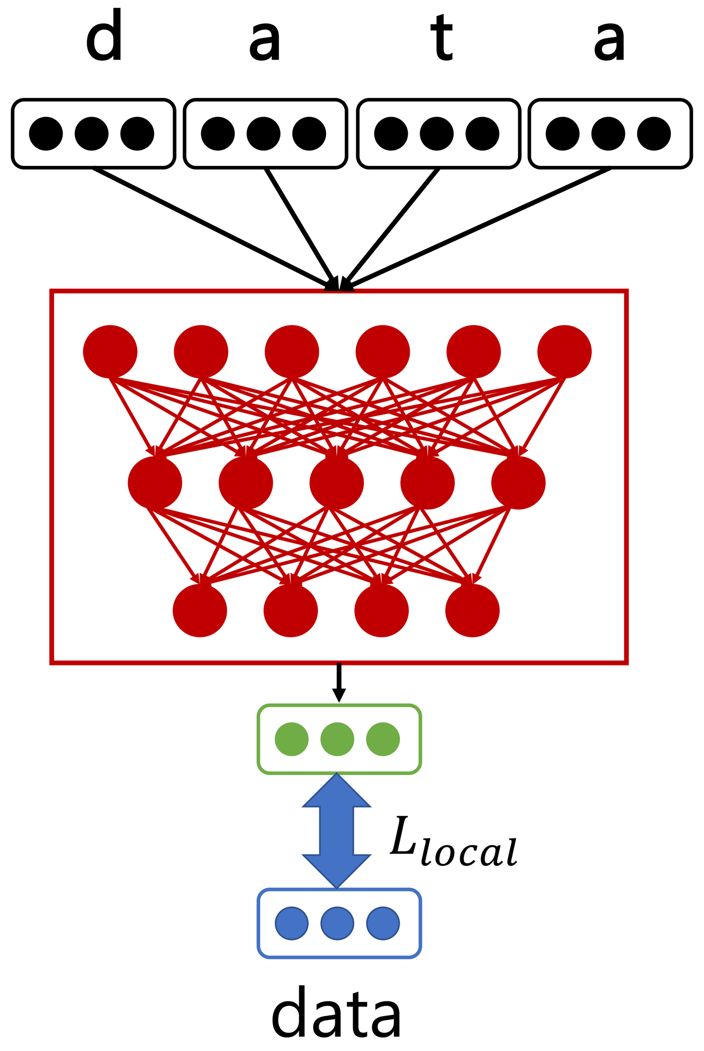
- Pre-trained word embeddings require a large memory space
- The technique of word embedding reconstruction makes the memory space smaller

A reconstruction model estimates an embedding of an input word according to subword embeddings. The model is trained with reconstruction loss to mimick the original word embeddings.

Local Reconstruction Loss

$$L_{local} = \frac{1}{d_w} \|\hat{\boldsymbol{e}}_w - \boldsymbol{e}_w\|^2$$

d_w : Dimension of pre-trained word embeddings
$\boldsymbol{e}_w$: Original word embedding
$\hat{\boldsymbol{e}}_w$: Reconstructed word embedding



1. Subword Tokenization
Tokenize the input word into subwords

2. Reconstruct
Generate word embedding using neural networks

3. Mimick
Train the model to generate a word embedding by mimicking the original embeddings

Fig 1: Reconstruction model

Approach: Globally Informed Reconstruction

- Existing reconstruction models only consider local information, i.e., the original embedding
- Our method consider the similarity among words as global information

We train a reconstruction model with both of the local reconstruction loss and globally informed reconstruction loss.

Globally Informed Reconstruction Loss

$$L = L_{local} + L_{global}$$

$$L_{global} = \frac{1}{|W|} \sum_{g \in W} \left(\cos(\hat{\boldsymbol{e}}_w, \boldsymbol{e}_g) - \cos(\boldsymbol{e}_w, \boldsymbol{e}_g)\right)^2$$

$\boldsymbol{e}_w$: Original word embedding
$\hat{\boldsymbol{e}}_w$: Reconstructed word embedding

- Sample 10 words from the training set to compute the global loss
- Half of the sampled words are the nearest neighbor of a target word

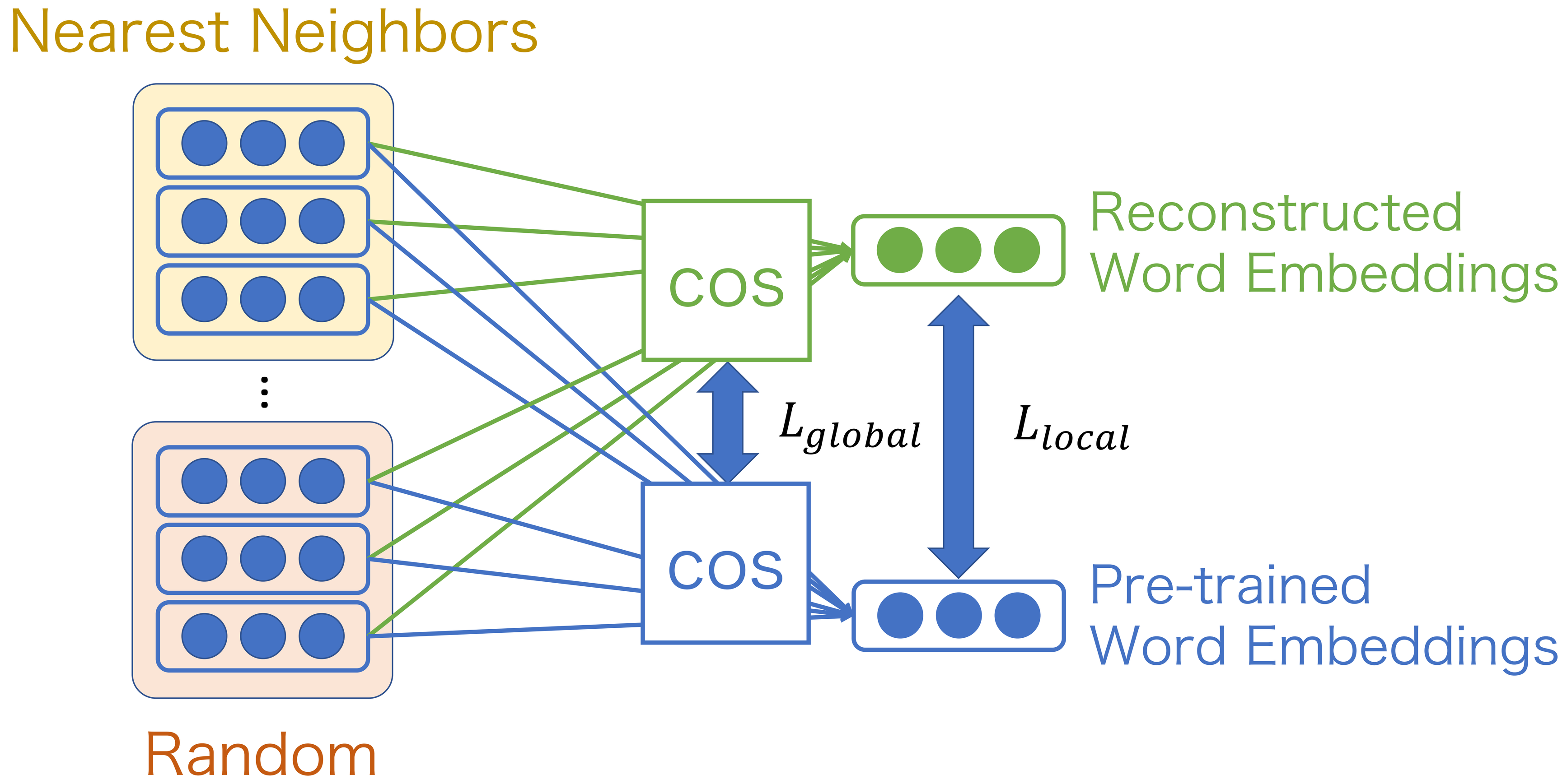


Fig 2: Globally informed reconstruction loss

Evaluation: Word Similarity Estimation Task

Datasets

	# of word-pairs
Rubenstein-Goodenough	65
Miller-Charles	30
WordSim-353	353
MEN	3,000
Stanford Rare Word Similarity	2,034

The memory space was reduced to 0.5% while 86% of the quality was preserved

	Spearman's ρ	Size (MB)
Character-RNN	0.534	14
+ Global Loss	**0.540**	
Character-CNN	0.594	25
+ Global Loss	**0.602**	
Bag of N-gram (Small)	0.191	12
+ Global Loss	**0.210**	
N-gram SAM (Small)	0.494	12
+ Global Loss	**0.618**	
fastText	0.719	2230

Table 1: Experimental results

fastText	glasgow	edinburgh	birmingham
N-gram SAM	lon	lond	canton
+ Global Loss	glasgow	chicago	edinburgh
fastText	influenza	pneumonia	bronchitis
N-gram SAM	litis	lam	tis
+ Global Loss	influenza	pneumonia	pneumonias

Table 2: Nearest Neighbors of the word "london" (upper) and "flu" (lower)