中国科学技术大学

ACL2018

Incorporating Latent Meanings of Morphological Compositions to Enhance Word Embeddings

Yang Xu, Jiawei Liu, Wei Yang, and Liusheng Huang

School of Computer Science and Technology,
University of Science and Technology of China, Hefei, 230027, China

July 17th 2018

OUTLINE

01 Introduction

02 Latent Meaning Models

03 Experimental Setup

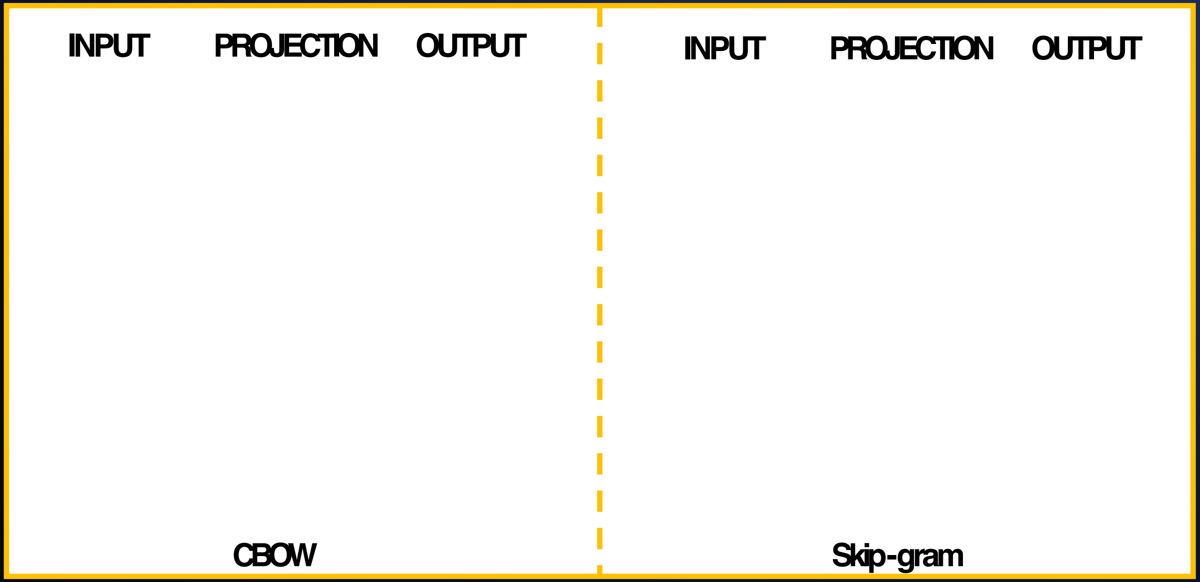
04 Experimental Results

05 Conclusion

01 Introduction

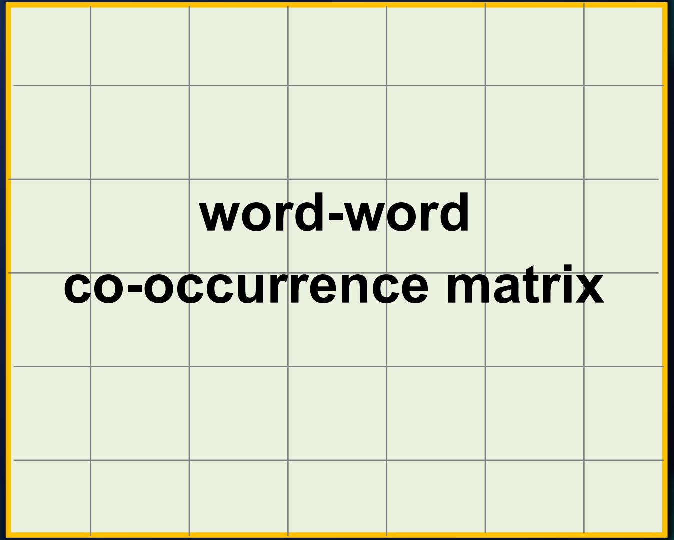
Word-level Word Embedding

01 Neural Network-Based



e.g., CBOW, Skip-gram (Mikolov et al.)

e.g., GloVe (Pennington et al.)



02 Matrix Factorization-Based (Spectral Methods)

Morphology-based Word Embedding

01

Morpheme Embeddings

Prefix + Root + Suffix

$\vec{in-}$ $\vec{cred}$ $\vec{ible}$

+

Word Embeddings

Word

$\overrightarrow{\text{incredible}}$

Training Model

Generated Word Vectors

Generated Word

Generative Model

Morpheme Embeddings

Prefix Root Suffix

02

Our Original Intention

- Word-level models: InputWords;
  Output Word Embeddings

- Morphology-based models: Input Words + Morphemes
  Output Word Embeddings + Morpheme Embeddings

- Our Latent Meaning Models: InputWords + Latent Meanings of Morphemes
  Output Word Embeddings
  (no by-product, e.g., morpheme embedding)

PURPOSE: to not only encode morphological properties into words, but also enhance the semantic similarities among word embeddings

Explicit Models & Our Models

*Note: The lookup table can be derived from morphological lexicons.

02
Latent Meaning Models

CBOW with Negative Sampling

- Sequence of tokens

INPUT PROJECTION OUTPUT

- Objective Function:

$$L = \frac{1}{n}\sum_{i=1}^{n} \log p(t_i \mid Context(t_i))$$

rd)

- Negative Sampling:

(Context Words)

Three Specific Models

01

LMM-A
(Latent Meaning Model-Average)

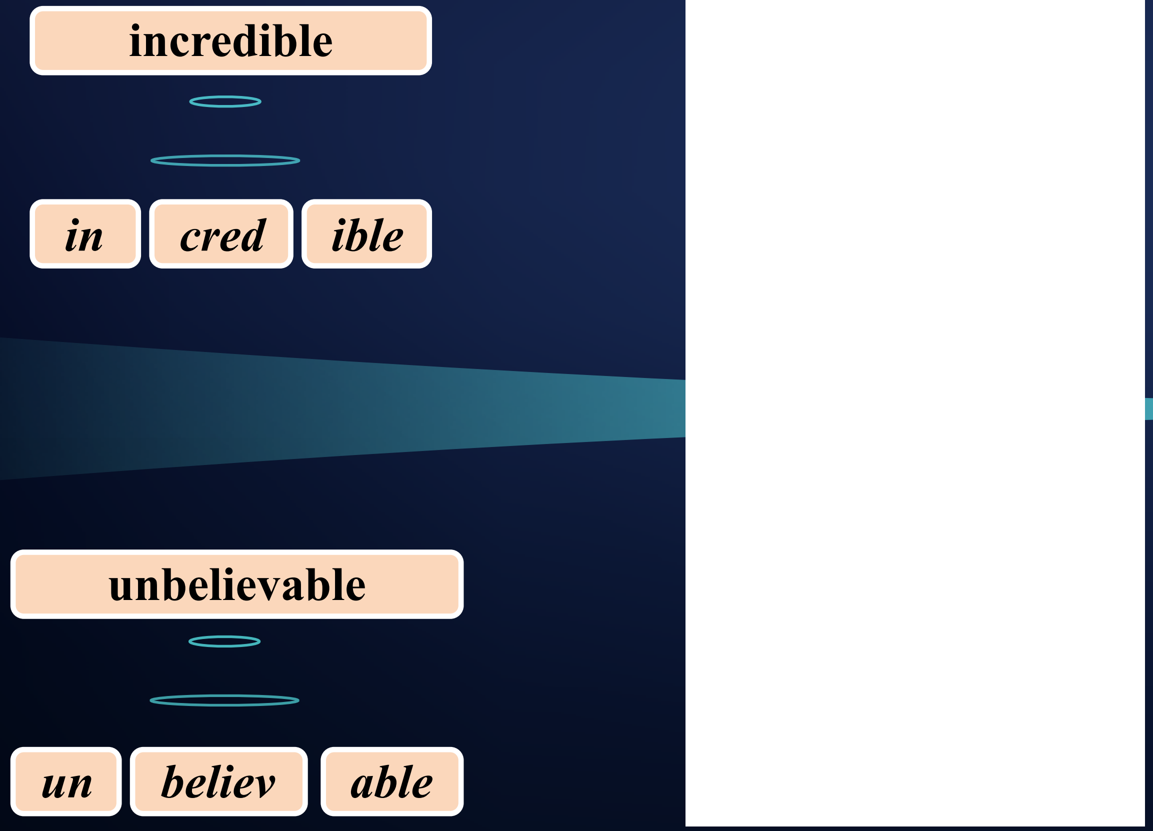
02

LMM-S
(Latent Meaning Model-Similarity)

03

LMM-M
(Latent Meaning Model-Max)

Word Map



Word Map

Word	Prefix		Root	Suffix	
incredible	*in*	*not*	*believe*	*able*	*capable*
unbelievable	*not*		*believe*	*able*	*capable*

#rows = |vocabulary|

*Note: The derivational morphemes, not the inflectional morphemes, are mainly concerned

Latent Meaning Model-Average (LMM-A)

- Sequence of tokens

- The latent meanings of ’s morphemes have **equal** contributions to

- The modified embedding of :

  : a set of latent meanings of ’s morphemes

  : the length of

- is utilized for training

A paradigm of LMM-A

Latent Meaning $Context(t_i)$

Latent Meaning Model-Similarity (LMM-S)

- Sequence of tokens

- The latent meanings of ’s morphemes are assigned with **different weights**:

$$\omega_{<t_j,\, w>} = \frac{\cos\left(v_{t_j}, v_w\right)}{\sum_{x \in M_j} \cos\left(v_{t_j}, v_x\right)}, w \in M_j$$

- The modified embedding of :

: a set of latent meanings of ’s morphemes

A paradigm of LMM-S

Latent Meaning

$Context\left(t_i\right)$

Latent Meaning Model-Max (LMM-M)

- Sequence of tokens

- Keep the latent meanings that have **maximum similarities** to :

$$P_{max}^{j} = \underset{w}{argmax} cos\left(v_{t_j}, v_w\right), w \in P_j$$

$$R_{max}^{j} = \underset{w}{argmax} cos\left(v_{t_j}, v_w\right), w \in R_j$$

$$S_{max}^{j} = \underset{w}{argmax} cos\left(v_{t_j}, v_w\right), w \in S_j$$

- The modified embedding of :

$$M_{max}^{j} = \left\{P_{max}^{j}, R_{max}^{j}, S_{max}^{j}\right\}$$

A paradigm of LMM-M

Latent Meaning $\quad Context(t_i)$

Update Rules for LMMs

- **New Objective Function (After modifying the input layer of CBOW):**

$$\hat{L} = \frac{1}{n}\sum_{i=1}^{n} \log p(v_{t_i} \mid \sum_{t_j \in Context(t_i)} \hat{v}_{t_j})$$

- **All parameters introduced by our models can be directly derived using the word map and word embeddings**

- **Update not just but the embeddings of the latent meanings with the same weights as they are assigned in the forward propagation period**

03
Experimental Setup

Corpus & Word Map

Corpus

- News corpus of 2009 (2013 ACL Eighth Workshop)
- Size: 1.7GB
- ~500 million tokens
- ~600,000 words
- Digits & punctuation marks are filtered

Word Map

- Morpheme segmentation using Morefessor (Creutz & Lagus, 2007)
- Assign latent meanings
- Lookup table
  - ▶ derived from the resources provided by Michigan State University*
  - ▶ 90 prefixes, 382 roots, 67 suffixes

*Resources web link:
https://msu.edu/~defores1/gre/roots/gre_rts_afx1.htm

Baselines & Parameter Settings

Baselines:

- Word-level models: CBOW, Skip-gram, GloVe
- Explicitly Morpheme-related Model (EMM)

A paradigm of EMM

Morphemes

Super-parameter Settings:

- Equal settings to all models
- Vector Dimension: 200
- Context window size: 5
- #Negative_Samples: 20

Evaluation Benchmarks (1/2)

Word Similarity:

Dataset

Name	#Pairs	Name	#Pairs	Name	#Pairs
RG-65	65	Rare-Word	2034	Men-3k	3000
Wordsim-353	353	SCWS	2003	WS-353-Related	252

Gold Standard Datasets — Widely-used Datasets

Syntactic Analogy:

- "a b as c _? (d)_" e.g., Queen King as Woman (Man)
- Microsoft Research Syntactic Analogies dataset (8000 items)

Evaluation Benchmarks (2/2)

Text Classification:

- 20 Newsgroups dataset (19000 documents of 20 different topics)
- 4 text classification tasks, each involves 10 topics
- Training/Validation/Test subsets (6:2:2)
- Feature vector: average word embedding of words in each document
- L2-regularized logistic regression classifier

04
Experimental Results

The Results on Word Similarity

ACL2018

	CBOW	Skip-gram	GloVe	EMM	LMM-A	LMM-S	LMM-M
Wordsim-353	58.77	61.94	49.40	60.01	62.05	**63.13**	61.54
Rare-Word	40.58	36.42	33.40	40.83	**43.12**	42.14	40.51
RG-65	56.50	62.81	59.92	60.85	62.51	62.49	**63.07**
SCWS	**63.13**	60.20	47.98	60.28	61.86	61.71	63.02
Men-3k	68.07	66.30	60.56	66.76	66.26	**68.36**	64.65
WS-353-Related	49.72	57.05	47.46	54.48	56.14	**58.47**	55.19

(Given different models) Spearman's rank correlation (%) on different datasets

The Results on Syntactic Analogy

Question: "a b as c __(d)__"

Answer:

	CBOW	Skip-gram	GloVe	EMM	LMM-A	LMM-S	LMM-M
Syntactic Analogy	13.46	13.14	13.94	17.34	**20.38**	17.59	18.30

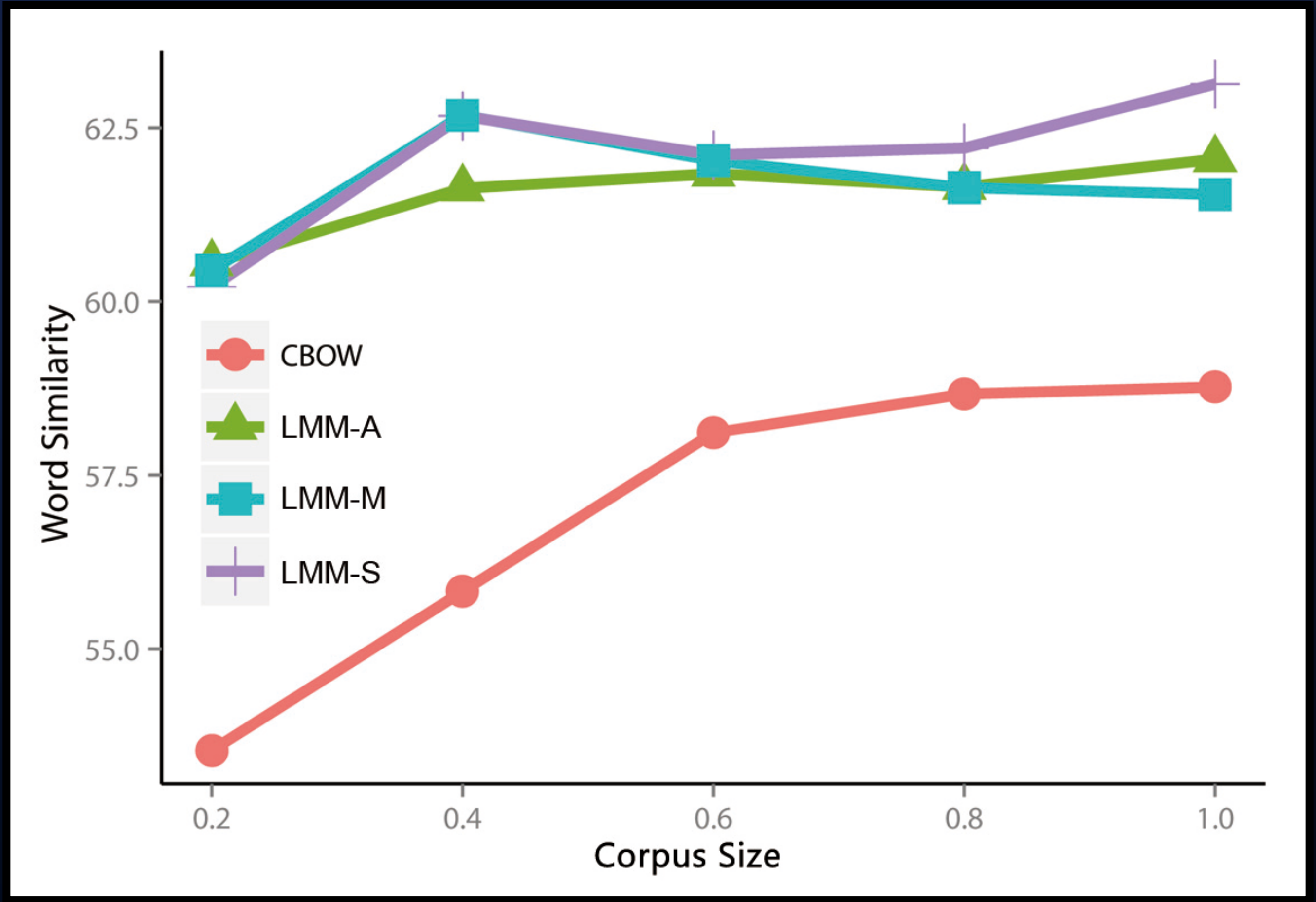
Syntactic analogy performance (%)

The Results on Text Classification

	CBOW	Skip-gram	GloVe	EMM	LMM-A	LMM-S	LMM-M
Text Classification	78.26	79.40	77.01	80.00	80.67	80.59	**81.28**

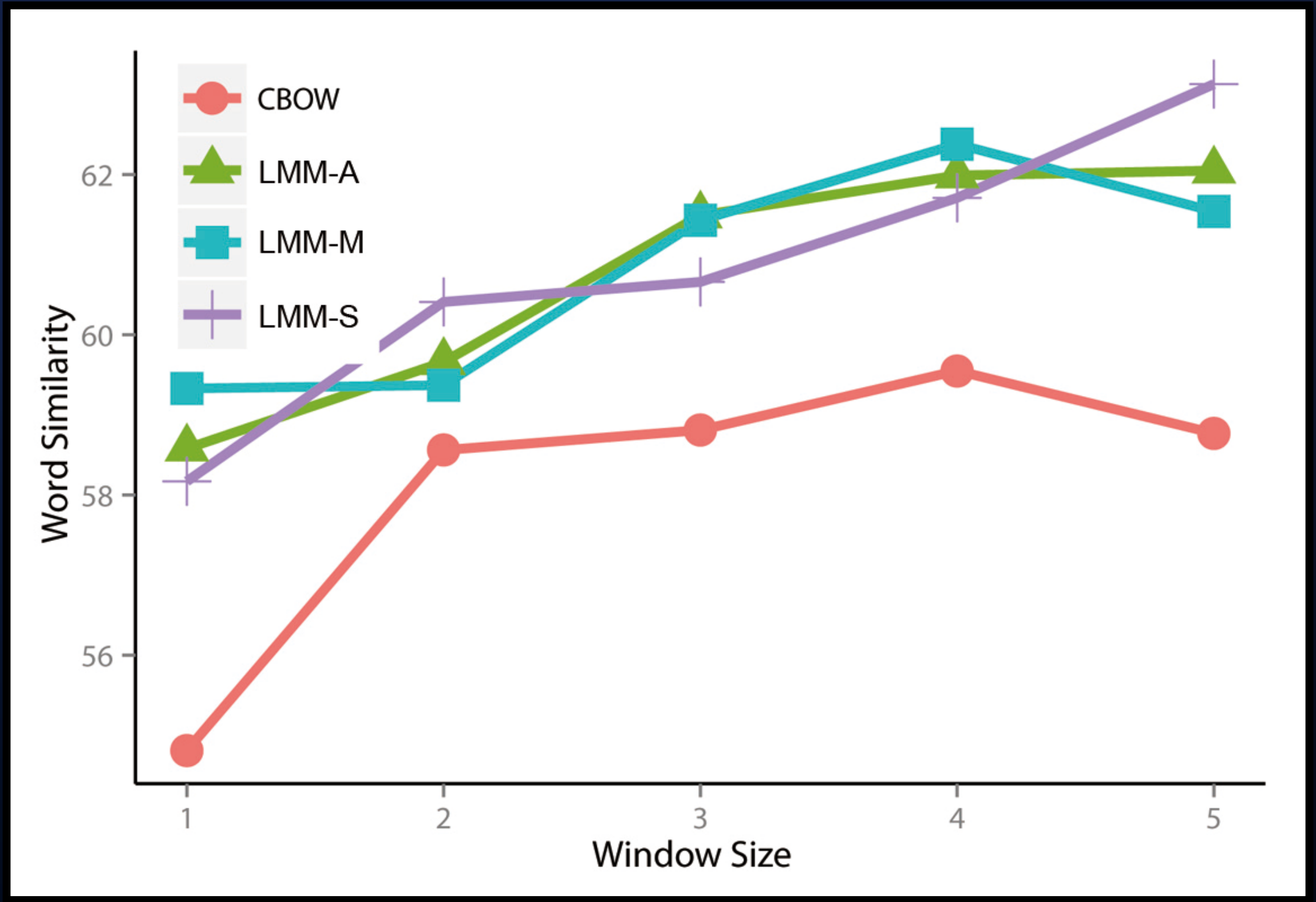
Average text classification accuracy across the 4 tasks (%)

The Impact of Corpus Size



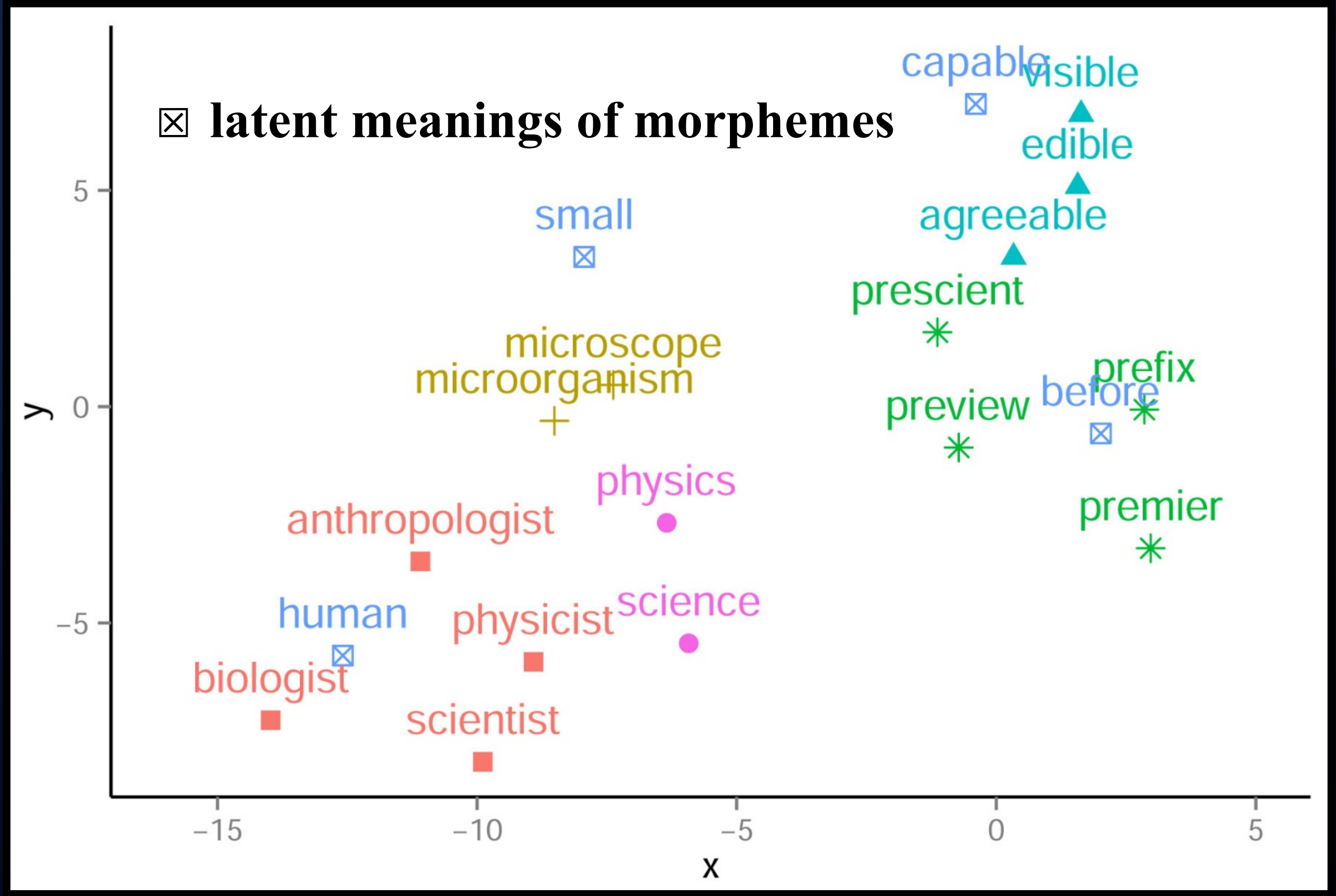
Results on Wordsim-353 task with different corpus size

The Impact of Context Window Size



Results on Wordsim-353 task with different context window size

Word Embedding Visualization



Visualization of word embeddings based on PCA

05 Conclusions

Conclusions

- Employ latent meanings of morphemes rather than the internal compositions themselves to train word embeddings
- By modifying the input layer and update rules of CBOW, we proposed three latent meaning models (LMM-A, LMM-S, LMM-M)
- The comprehensive quality of word embedings are enhanced by incorporating latent meanings of morphemes
- In the future, we intend to evaluate our models for some morpheme-rich languages like Russian, German, etc.

ACL2018

Thank you!

Questions?