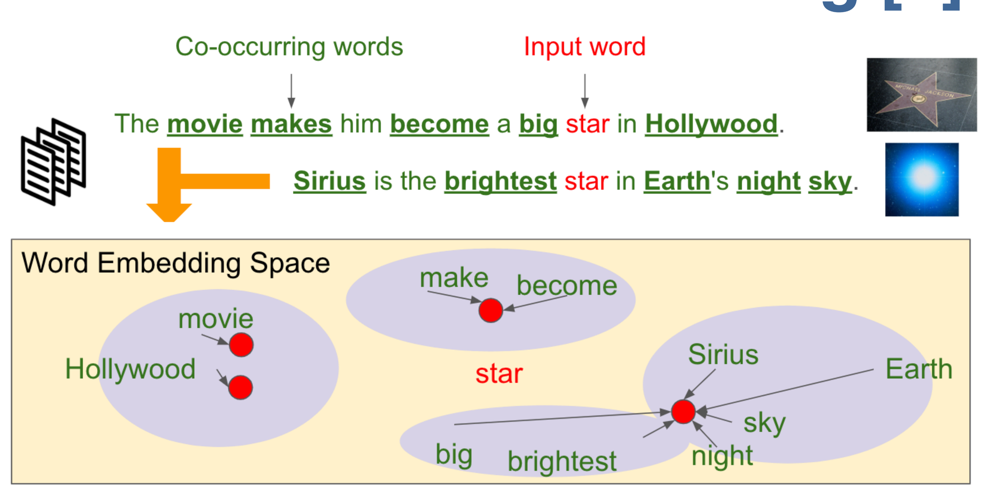
Extending Multi-sense Word Embedding to Phrases and Sentences for Unsupervised Semantic Applications

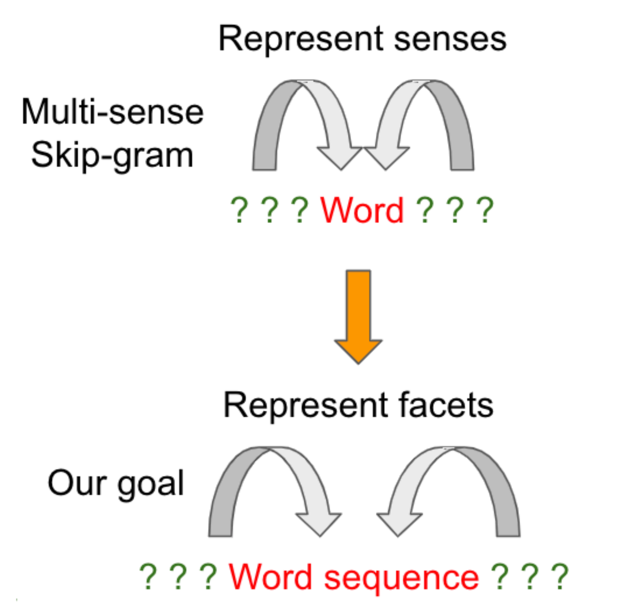
Haw-Shiuan Chang, Amol Agrawal, Andrew McCallum

Introduction

- Previous Work:
  - Word embedding represents the input word by a set co-occurring words
  - Co-occurring word distribution might have multiple modes
  - Multi-sense word embedding clusters the co-occurring words into centers
- Our goal:
  - Extending the methods to phrases and sentences
  - Do the similar thing but replacing the input word as a word sequence.
  - Senses of the input words -> Facets of the input sentence/phrases

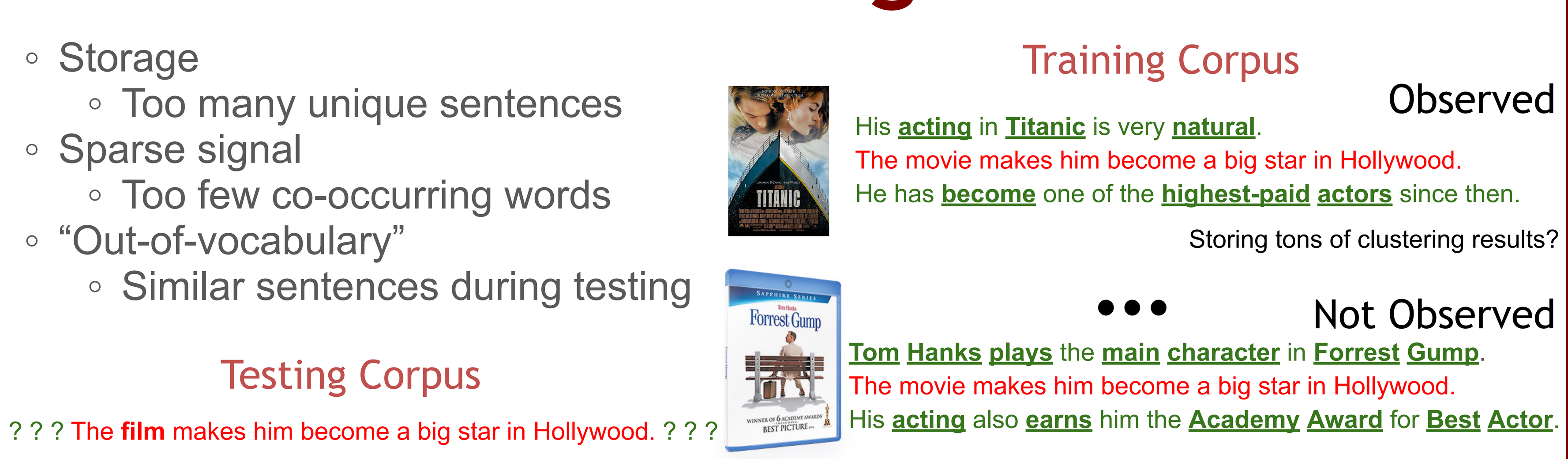
Multi-sense Embedding [1]





Challenges

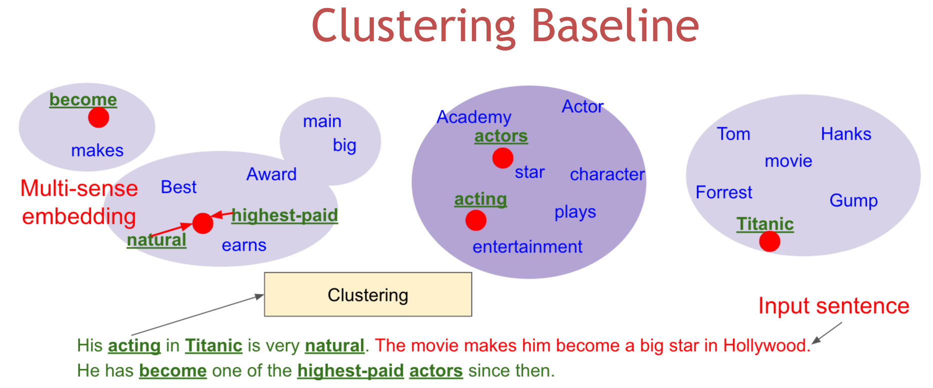
- Storage
  - Too many unique sentences
- Sparse signal
  - Too few co-occurring words
- "Out-of-vocabulary"
  - Similar sentences during testing



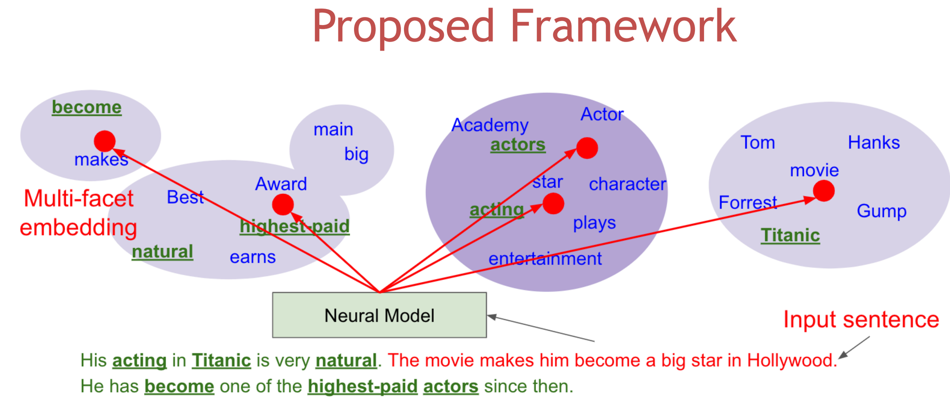
Main Idea

- Instead of clustering, we directly predict the cluster centers using a neural model
- Storage issue
  - Clusters are compressed in the parameters of the neural model
- Sparse signal issue
  - Clustering the co-occurring words of similar sentences
- "Out-of-vocabulary" issue
  - Can take any input sentence
- Model Design
  - What neural network architecture to use?
  - How to train end-to-end?
  - What clustering loss to use?

Clustering Baseline

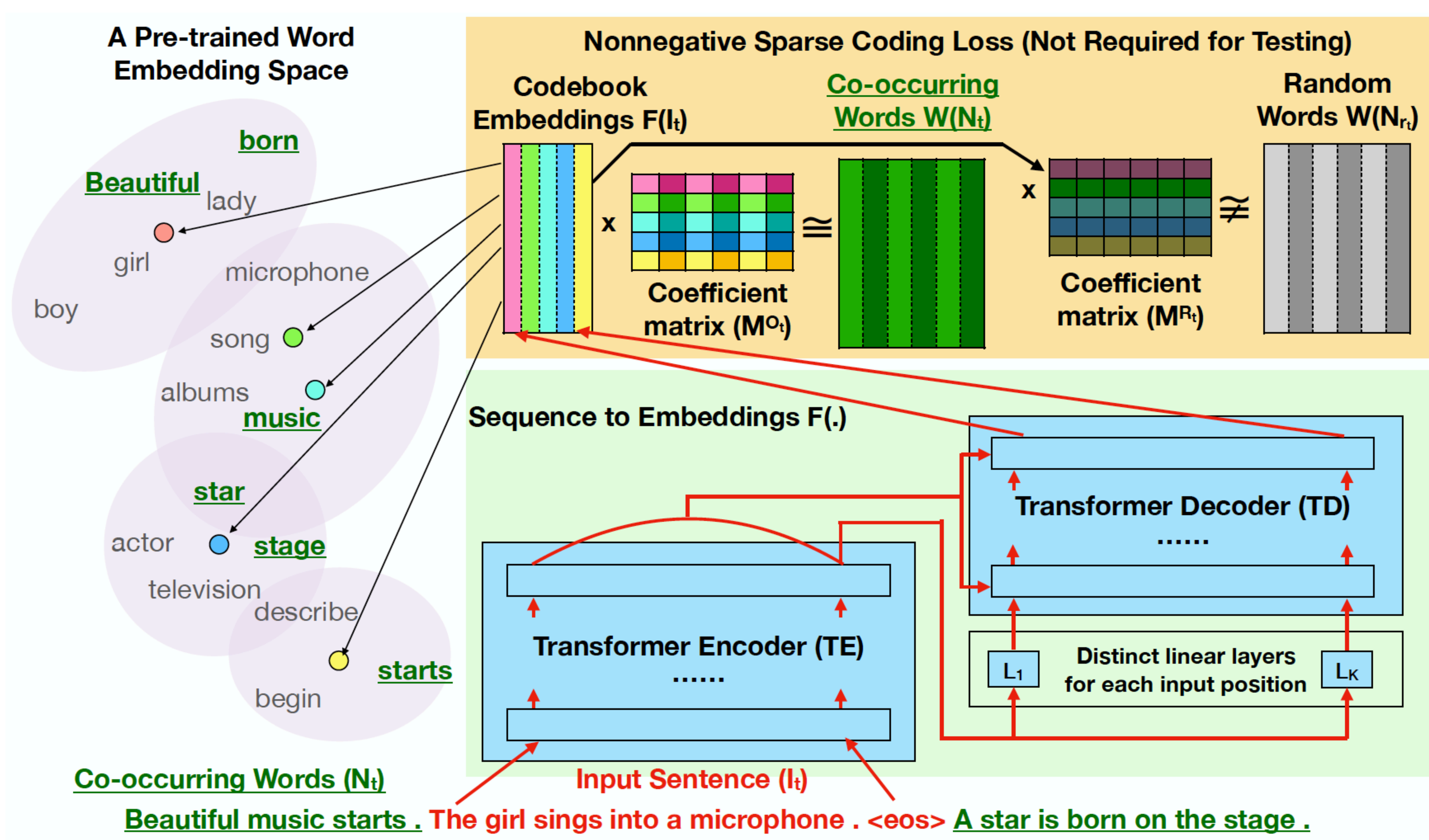


Proposed Framework



Our Method

Multi-facet Embedding



K-means Loss

Cluster center; Normalized word embedding

$$\sum_j ||(\sum_k M_{k,j}\underline{c}_k^t) - \underline{w}_j^t||^2$$

$M_{k,j} = 1$ if jth word belongs to the kth cluster.

$(L2)^2 = 2(1-\cos)$

$$||\underbrace{F(I_t)}_{[\underline{c}_1^t \dots \underline{c}_K^t]} M - \underbrace{W(N_t)}_{[\underline{w}_1^t \dots \underline{w}_j^t]}||^2$$

Non-Negative Sparse Coding (NNSC) Loss

$$L_t(\boldsymbol{F}) = Er(\boldsymbol{F}(\boldsymbol{I_t}), \boldsymbol{W}(\boldsymbol{N_t})) - Er(\boldsymbol{F}(\boldsymbol{I_t}), \boldsymbol{W}(\boldsymbol{N_{r_t}})), \quad (2)$$

$$Er(\boldsymbol{F}(\boldsymbol{I_t}), \boldsymbol{W}(\boldsymbol{N_t})) = ||\boldsymbol{F}(\boldsymbol{I_t})\boldsymbol{M}^{\boldsymbol{O_t}} - \boldsymbol{W}(\boldsymbol{N_t})||^2$$
$$s.t., \boldsymbol{M}^{\boldsymbol{O_t}} = \arg\min_{\boldsymbol{M}} ||\boldsymbol{F}(\boldsymbol{I_t})\boldsymbol{M} - \boldsymbol{W}(\boldsymbol{N_t})||^2 + \lambda||\boldsymbol{M}||_1,$$
$$\forall k, j,\ 0 \leq \boldsymbol{M}_{k,j} \leq 1, \quad (1)$$

- We use Transformer encoder and decoder to predict a set of centers
- NNSC loss is better because its gradient is more smooth
- We match the cluster centers and co-occurring words in each training iteration

Each Training Iteration

Step 1: Generate $\mathbf{F(I_t)}$
Step 2: Estimate $\mathbf{M^{Ot}}$ and $\mathbf{M^{Rt}}$
Step 3: Compute Loss $\mathbf{L_t(F)}$
Step 4: Fix $\mathbf{M^{Ot}}$ and $\mathbf{M^{Rt}}$ to do backprop

Experiments

- Multiple embeddings for sentence representation is much better than single embedding
  - similar for phrase representation
- Word importance estimation using the co-occurring distribution improves various scoring functions
- More facets are better in summarization

Visualizing Predicted Cluster Centers

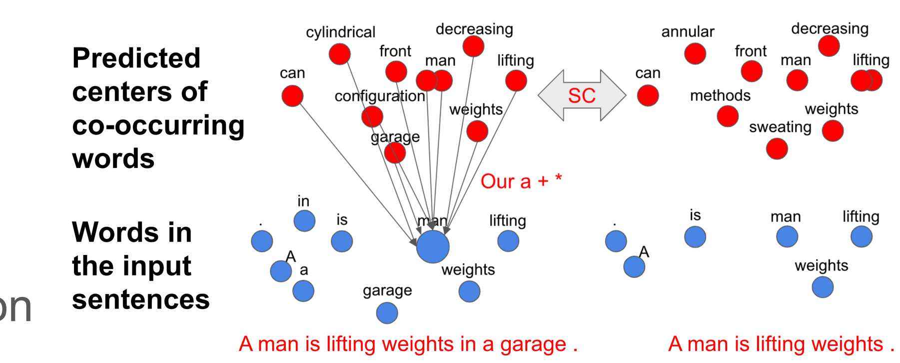
Input Phrase: civil order <eos>	
Output Embedding (K = 1):	
e1	government 0.817 civil 0.762 citizens 0.748
Output Embeddings (K = 3):	
e1	initiatives 0.736 organizations 0.725 efforts 0.725
e2	army 0.815 troops 0.804 soldiers 0.786
e3	court 0.758 federal 0.757 judicial 0.736

Input Sentence: SMS messages are used in some countries as reminders of hospital appointments . <eos>	
Output Embedding (K = 1):	
e1	information 0.702, use 0.701, specific 0.700
Output Embeddings (K = 3):	
e1	can 0.769, possible 0.767, specific 0.767
e2	hospital 0.857, medical 0.780, hospitals 0.739
e3	SMS 0.791, Mobile 0.635, Messaging 0.631
Output Embeddings (K = 10):	
e1	can 0.854, should 0.834, either 0.821
e2	hospital 0.886, medical 0.771, hospitals 0.745
e3	services 0.768, service 0.749, web 0.722
e4	SMS 0.891, sms 0.745, messaging 0.686
e5	messages 0.891, message 0.801, emails 0.679
e6	systems 0.728, technologies 0.725, integrated 0.723
e7	appointments 0.791, appointment 0.735, duties 0.613
e8	confirmation 0.590, request 0.568, receipt 0.563
e9	countries 0.855, nations 0.737, Europe 0.732
e10	Implementation 0.613, Application 0.610, Programs 0.603

Unsupervised Phrase Similarity

Method		SemEval 2013		Turney (5)	Turney (10)
Model	Score	AUC	F1	Accuracy	Accuracy
BERT	CLS	54.7	66.7	29.2	15.5
	Avg	66.5	67.1	43.4	24.3
GloVe	Avg	79.5	73.7	25.9	12.9
FCT LM†	Emb	-	67.2	42.6	27.6
Ours (K=10)	SC	80.3	72.8	45.6	28.8
	Emb	85.6	77.1	49.4	31.8
Ours (K=1)	SC	81.1	72.7	45.3	28.4
	Emb	**87.8**	**78.6**	**50.3**	**32.5**

Unsupervised Sentence Similarity



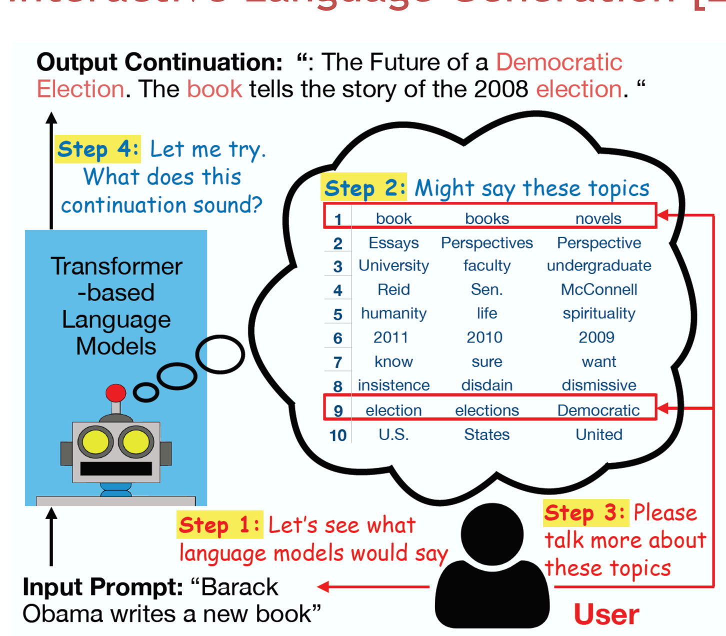
Method		Dev		Test	
Score	Model	All	Low	All	Low
Cosine	Skip-thought	43.2	28.1	30.4	21.2
CLS	BERT	9.6	-0.4	4.1	0.2
Avg	BERT	62.3	42.1	51.2	39.1
SC	Our c K1	55.7	43.7	47.6	45.4
	Our c K10	63.0	51.8	52.6	47.8
WMD	GloVe	58.8	35.3	40.9	25.4
	Our a K1	63.1	43.3	47.5	34.8
	Our a K10	66.7	47.4	52.6	39.8
Prob_WMD	GloVe	75.1	59.6	63.1	52.5
	Our a K1	74.4	60.8	62.9	54.4
	Our a K10	**76.2**	**62.6**	**66.1**	58.1
Avg	GloVe	51.7	32.8	36.6	30.9
	Our a K1	54.5	40.2	44.1	40.6
	Our a K10	61.7	47.1	50.0	46.5
Prob_avg	GloVe	70.7	56.6	59.2	54.8
	Our a K1	68.5	56.0	58.1	55.2
	Our a K10	72.0	60.5	61.4	**59.3**
SIF†	GloVe	75.1	65.7	63.2	58.1
	Our a K1	72.5	64.0	61.7	58.5
	Our a K10	**75.2**	**67.6**	**64.6**	**62.2**
	Our a (k-means) K10	71.5	62.3	61.5	57.2
sentence-BERT (100 pairs)*		71.2	55.5	64.5	58.2

Unsupervised Extractive Summarization

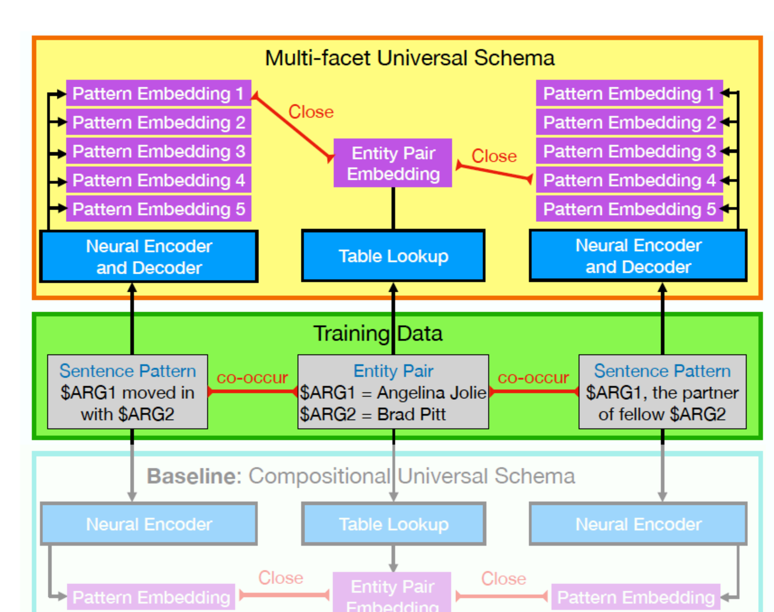
Setting	Method	R-1	R-2	Len
Unsup, No Sent Order	Random	28.1	8.0	68.7
	Textgraph (tfidf)†	33.2	11.8	-
	Textgraph (BERT)†	30.8	9.6	-
	W Emb (GloVe)	26.6	8.8	37.0
	Sent Emb (GloVe)	32.6	10.7	78.2
	W Emb (BERT)	31.3	11.2	45.0
	Sent Emb (BERT)	32.3	10.6	91.2
	Our c (K=3)	32.2	10.1	75.4
	Our c (K=10)	34.0	11.6	81.3
	Our c (K=100)	**35.0**	**12.8**	92.9
Unsup	Lead-3	40.3	17.6	87.0
	PACSUM (BERT)†	**40.7**	**17.8**	-
Sup	RL*	**41.7**	**19.5**	-

Other Applications

Interactive Language Generation [2]



Distantly Supervised Relation Extraction [3]



Conclusion

- We propose a framework for learning the cooccurring distribution of the words beside a sentence or a phrase.
- Even though there are usually only a few words that co-occur with each sentence, we demonstrate that the proposed models can learn to predict interpretable cluster centers conditioned on an (unseen) sentence.

References

[1] Neelakantan, A., Shankar, J., Passos, A., & McCallum, A. (2014). Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. In *EMNLP*.
[2] Chang, H-S, Yuan, J., Iyyer, M., & McCallum, A. (2021). Changing the Mind of Transformers for Topically-Controllable Language Generation. In *EACL*.
[3] Paul, R*., Chang, H-S*, & McCallum, A. (2021). Multi-facet Universal Schema. In *EACL*.