

Word & Sense Embedding and their Application to Word Sense Induction



Linfeng Song



Outline

- Word Embedding
- Sense Embedding
- Sense Embedding for Word Sense Induction
- Conclusion

Word Embedding

- Word Embedding is a set of language techniques in which words from the vocabulary are mapped to vectors of real numbers in a low-dimensional space.

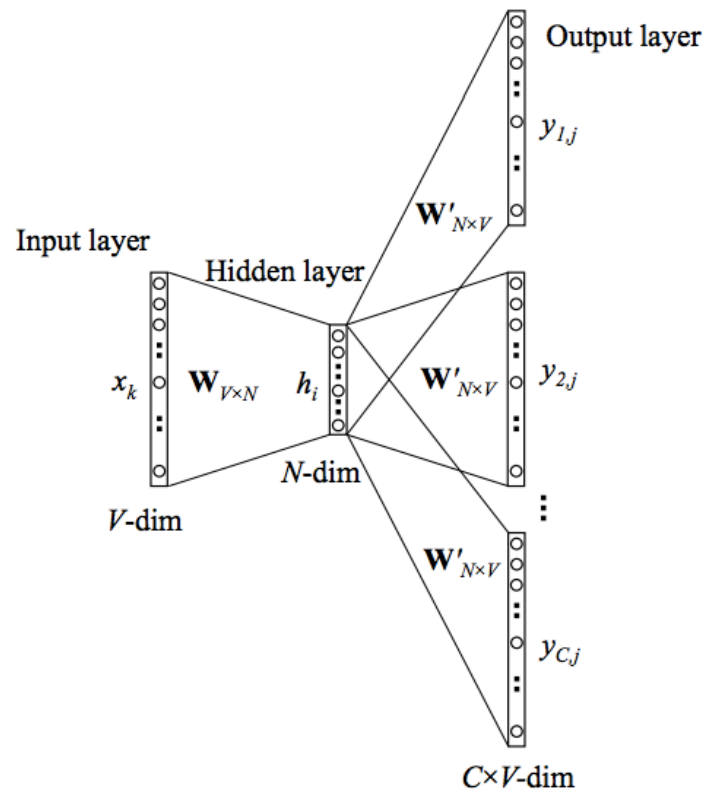
Word Embedding

- Word Embedding is a set of language techniques in which words from the vocabulary are mapped to vectors of real numbers in a low-dimensional space.
- Previous Methods
 - Build co-occurrence matrix from a corpus
 - Perform dimension reduction with PCA
 - Learn by counting

Word Embedding

- Word Embedding is a set of language techniques in which words from the vocabulary are mapped to vectors of real numbers in a low-dimensional space.
- Current methods
 - based on a neural network architecture
 - Learn by predicting

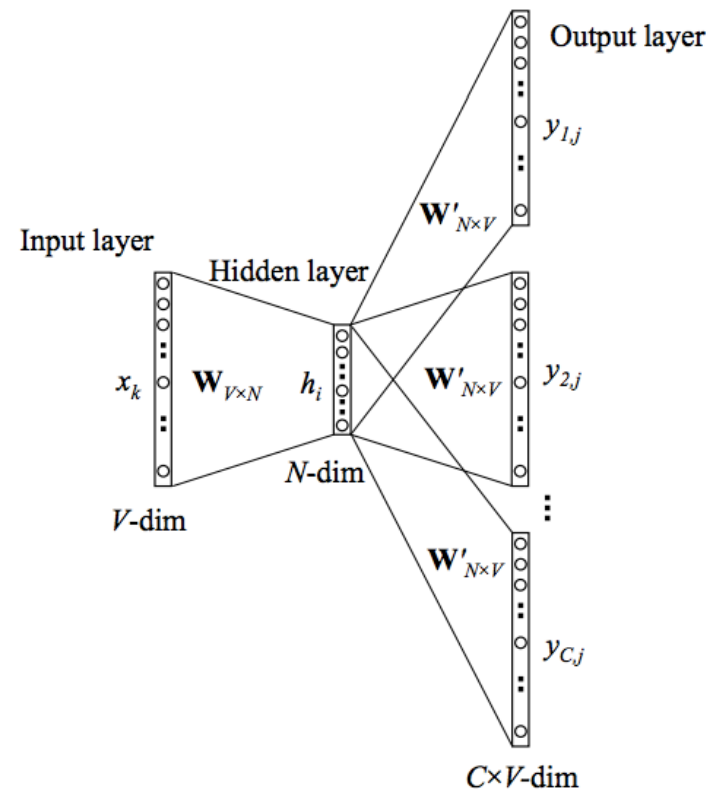
Skip-Gram Model



Skip-Gram Model

$$P(D = 1|w_i, w_j) = \frac{1}{1 + e^{-v_i^T v_j}}$$

$$P(D = 0|w_i, w_j) = 1 - P(D = 1|w_i, w_j)$$



Skip-Gram Model

➤ Given a document formalized as a list of (w_i, C_i)

$$J = \sum_{i=1}^T \left[\sum_{c \in C_i} P(D = 1 | v_i, v_c) + \sum_{c' \in V - C_i} P(D = 0 | v_i, v_{c'}) \right]$$

Sense Embedding

- Ubiquitous polysemous words harm the performance for most NLP systems
- Solution: learn a embedding for each sense instead

Sense Embedding

- Clustering-based
- Nonparametric
- Ontology-based

Sense Embedding

- Clustering-based
- Nonparametric
- Ontology-based

How sense is defined?

Sense Embedding

➤ Clustering-based

based on the distributional hypothesis of Harris, (1954):

➤ Nonparametric

a word sense is reflected by a set of contexts where it appears

➤ Ontology-based

based on the sense definition of a sense inventory

Reisinger and Mooney (2010)

clustering-based
non-parametric
ontology-based

➤ learn co-occurrence vector for each w_i

Reisinger and Mooney (2010)

clustering-based
non-parametric
ontology-based

- learn co-occurrence vector for each w_i
- cluster all tokens of w_i into K clusters
 - each token is represented by the context vector which is the average of word vectors in the context

Reisinger and Mooney (2010)

clustering-based
non-parametric
ontology-based

- learn co-occurrence vector for each w_i
- cluster all tokens of w_i into K clusters
 - each token is represented by the context vector which is the average of word vectors in the context
- learn one vector for each centroid of w_i
 - averaging all belonging context vectors

Reisinger and Mooney (2010)

clustering-based
non-parametric
ontology-based

➤ The similarity functions

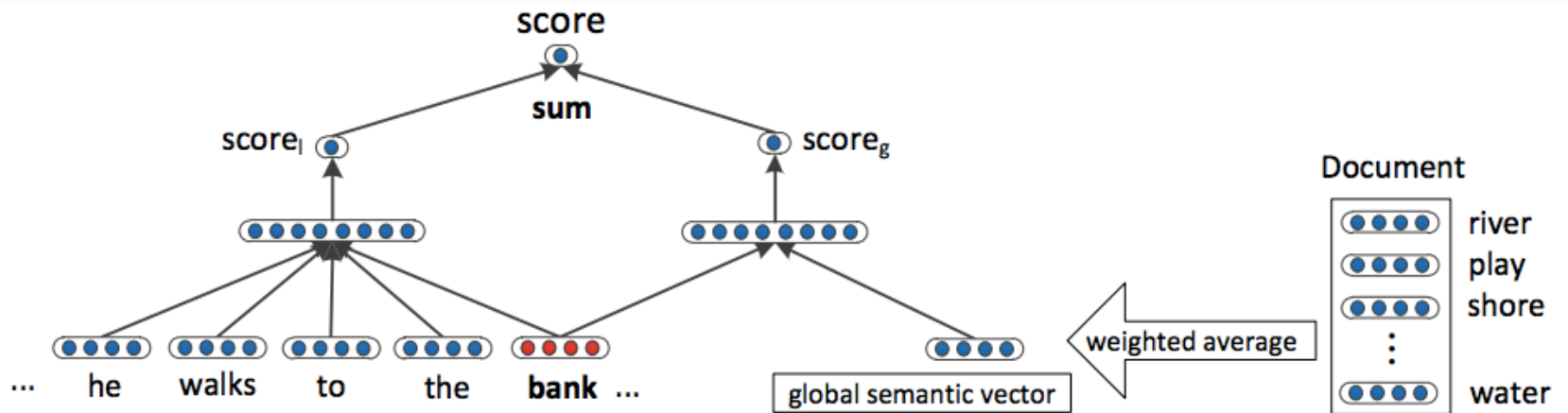
$$\text{MaxSim}(u, v) = \max_{1 \leq i \leq K, 1 \leq j \leq K} d(\pi_i(u), \pi_j(v))$$

$$\text{AvgSim}(u, v) = \frac{1}{K^2} \sum_{i=1}^K \sum_{j=1}^K d(\pi_i(u), \pi_j(v))$$

$$\text{AvgSimC}(u, v) = \frac{1}{K^2} \sum_{i=1}^K \sum_{j=1}^K d(\text{vec}(c), \pi_i(u)) \times d(\text{vec}(c'), \pi_j(v)) \times d(\pi_i(u), \pi_j(v))$$

Huang et al. (2012b)

clustering-based
non-parametric
ontology-based



Huang et al. (2012b)

clustering-based
non-parametric
ontology-based

- learn word vectors
- re-label the data by clustering
- learn sense vectors via the same neural network

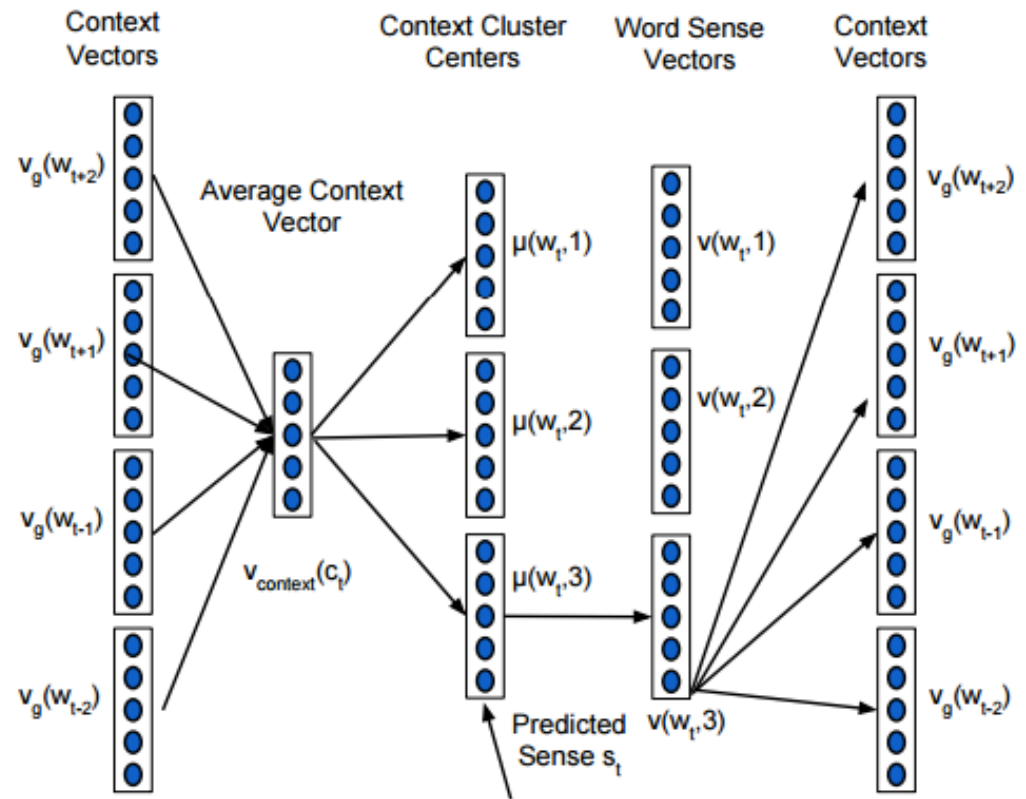
Huang et al. (2012b)

clustering-based
non-parametric
ontology-based

- learn word vectors
- re-label the data **problematic!**
The pipeline leads to error propagation!
- learn sense vectors via the same neural network

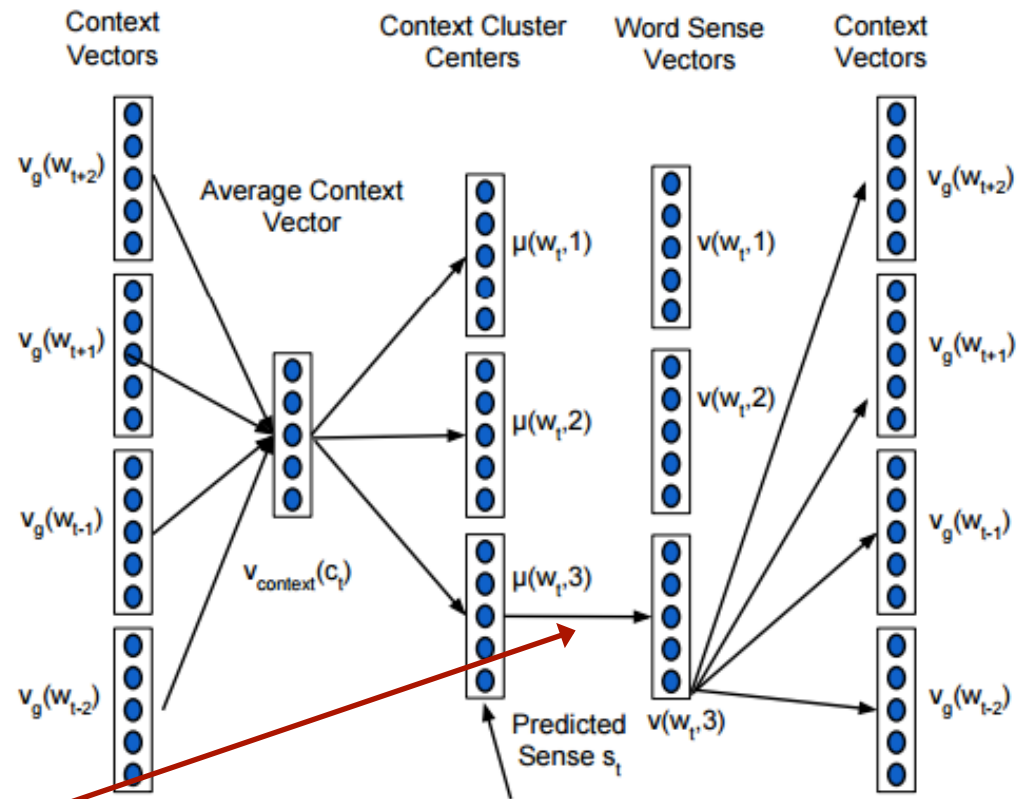
Neelakantan et al. (2014)

clustering-based
non-parametric
ontology-based



Neelakantan et al. (2014)

clustering-based
non-parametric
ontology-based

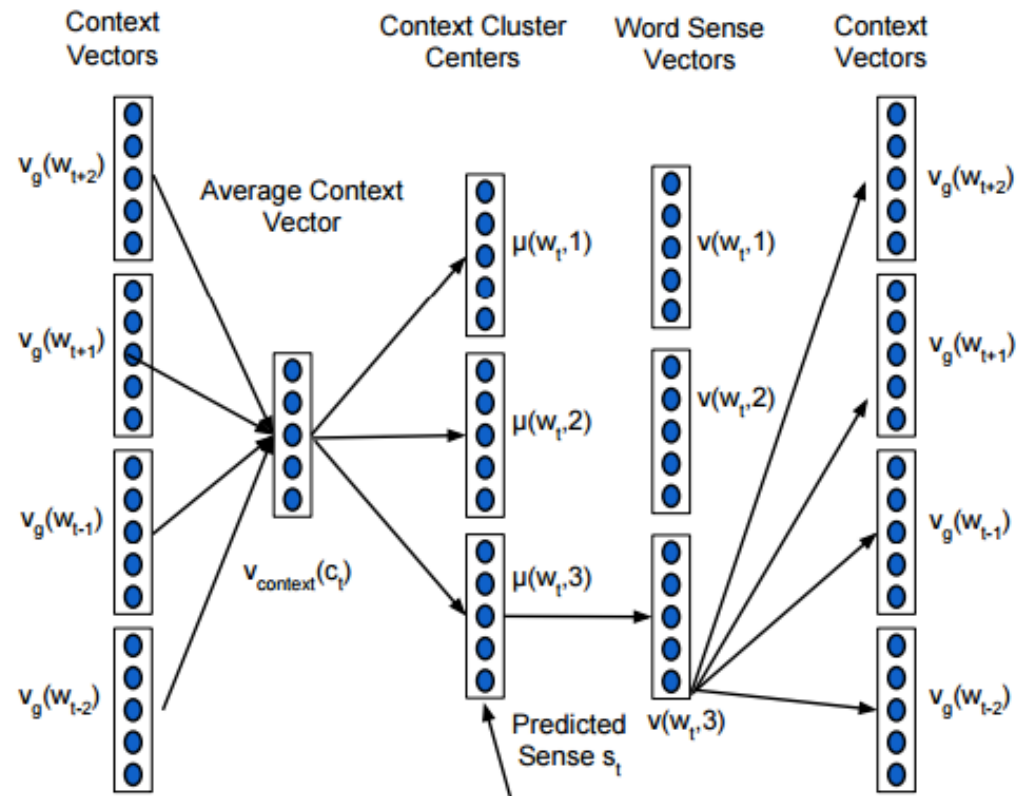


Use argmax to pick the cluster:

$$s_i =_k \text{cosine}(\mu(w_i, k), v_{context}(c_i)) \quad k \in [1, 2, \dots, K]$$

Neelakantan et al. (2014)

clustering-based
non-parametric
ontology-based



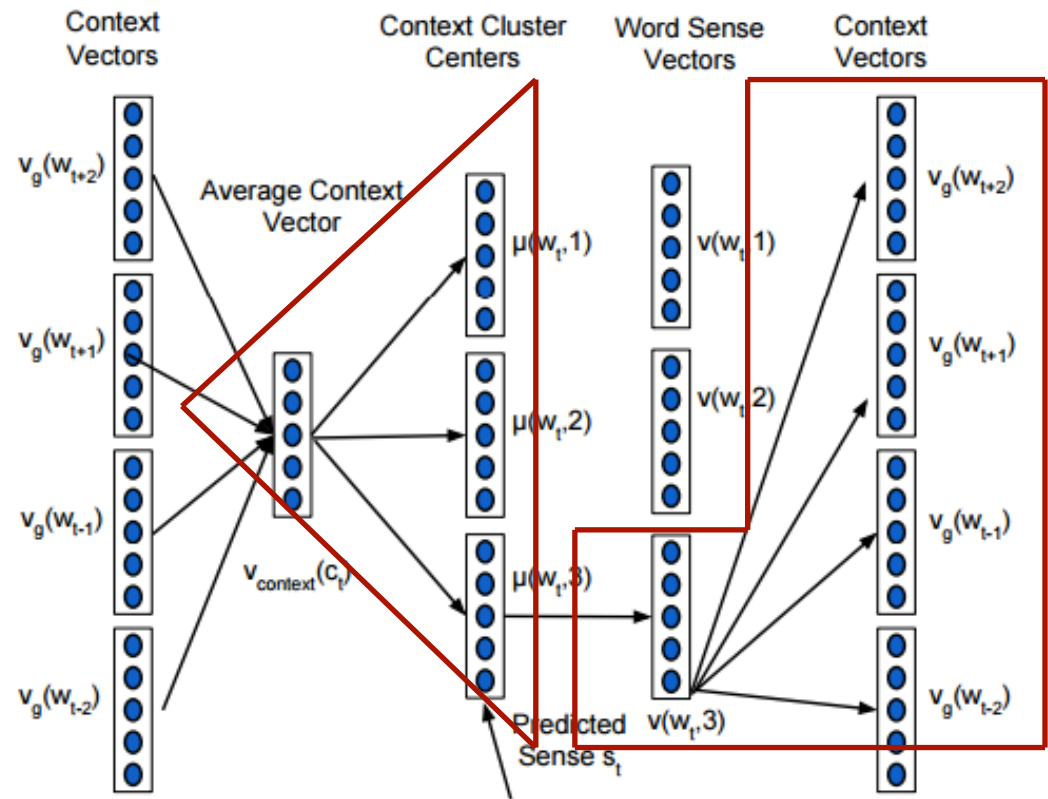
Probability for (not) observing words:

$$p(D = 1 | v_s(w_i, s_i), v_g(c)) = \frac{1}{1 + e^{-v_s(w_i, s_i)^T v_g(c)}}$$

$$p(D = 0 | v_s(w_i, s_i), v_g(c')) = 1 - p(D = 1 | v_s(w_i, s_i), v_g(c'))$$

Neelakantan et al. (2014)

clustering-based
non-parametric
ontology-based



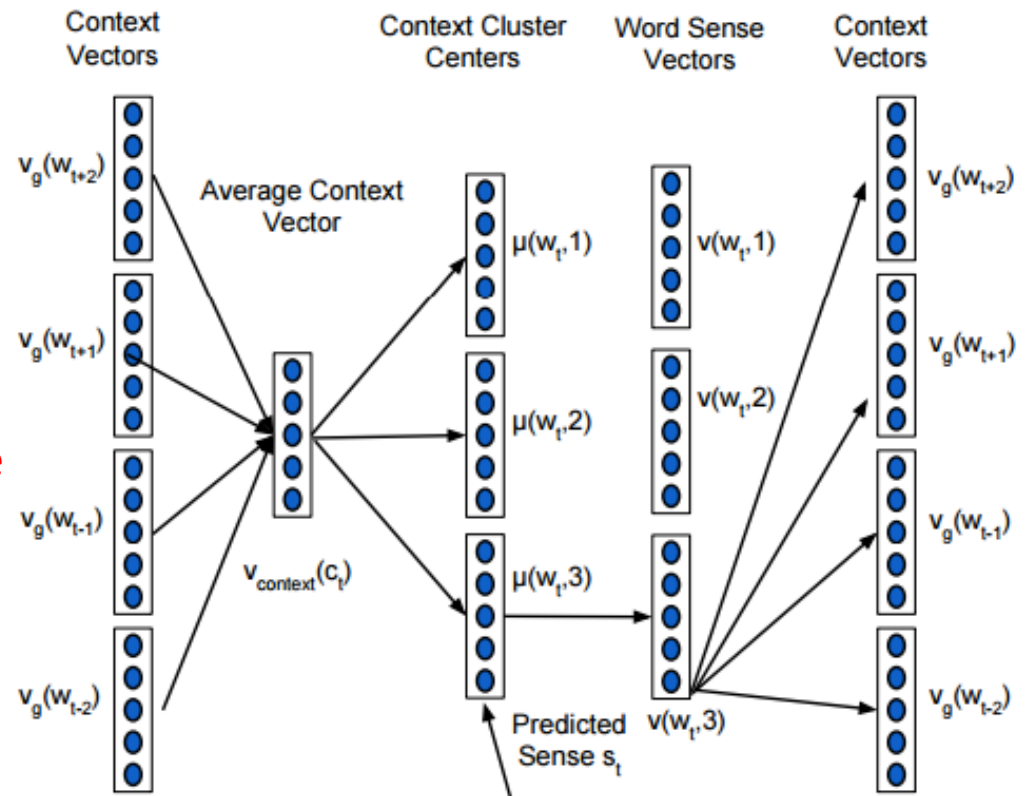
Update as Kmeans

Update as Skip-gram

Neelakantan et al. (2014)

clustering-based
non-parametric
ontology-based

Problematic!
Intuitively, different words should have
different number of senses



Li and Jurafsky (2015)

clustering-based
non-parametric
ontology-based

- Create a Chinese Restaurant Process for each word
 - a sense corresponds to a table
 - a data point is a customer

Li and Jurafsky (2015)

clustering-based
non-parametric
ontology-based

- Create a Chinese Restaurant Process for each word
 - a sense corresponds to a table
 - a data point is a customer
- Probability for choosing a sense is defined as:

$$p(s_i = k_t) \propto \begin{cases} N_t p(k_t | c_i), & \text{if } k_t \text{ already exists} \\ \gamma, & \text{if } k_t \text{ is new} \end{cases}$$

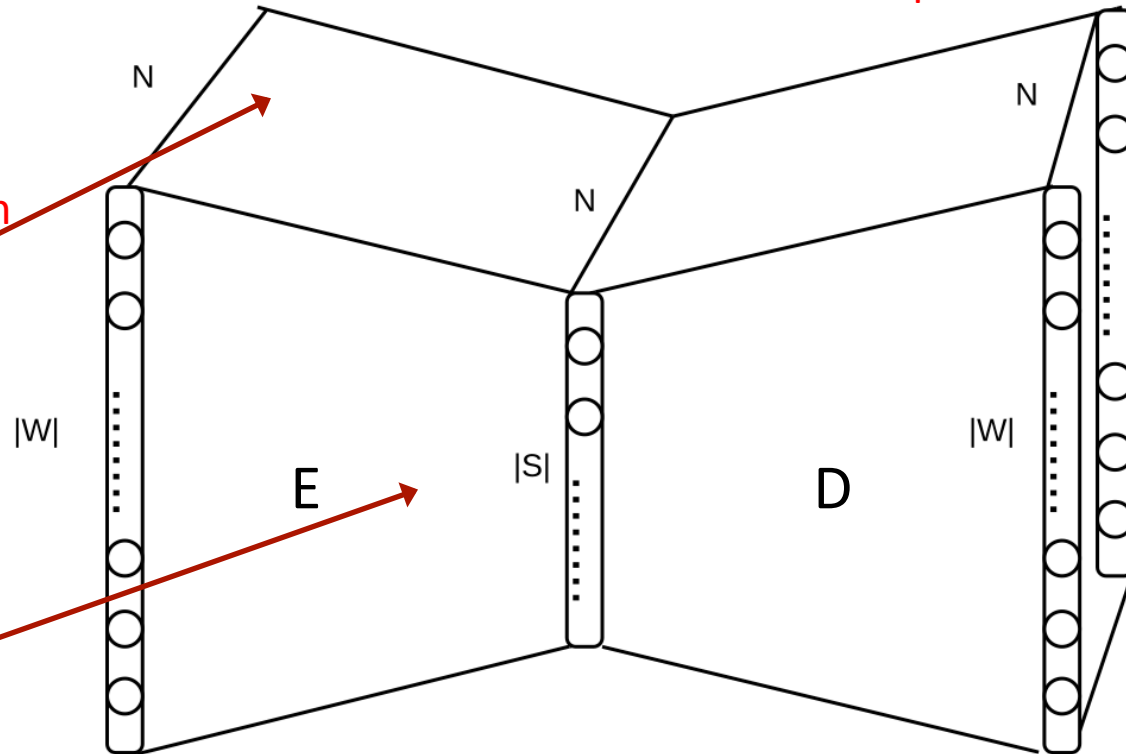
Rothe and Schutze (2015)

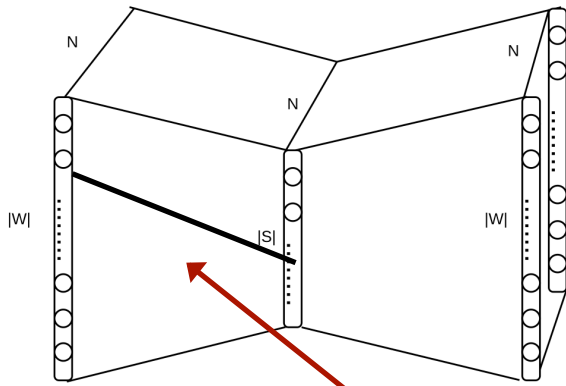
clustering-based
non-parametric
ontology-based

Best Student Paper of ACL 2015

Independent between
different dimensions

Each dimension is
an AutoEncoder





Rothe and Schutze (2015)

clustering-based
non-parametric
ontology-based

	Synset ₁	...	Synset _i	...	Synset _n
people	1 lexeme	
...		
dog		...	1	...	
...		
hound		...	1	...	
...		

Rothe and Schütze (2015)

clustering-based
non-parametric
ontology-based

➤ Objectives

$$\operatorname{argmin}_{D^{(d)}, E^{(d)}} \|D^{(d)} E^{(d)} w^{(d)} - w^{(d)}\| \quad \forall d$$

$$\operatorname{argmin}_{D^{(d)}, E^{(d)}} \|E^{(d)} \operatorname{diag}(w^{(d)}) - D^{(d)} \operatorname{diag}(s^{(d)})\| \quad \forall d$$

$$\operatorname{argmin}_{E^{(d)}} \|R E^{(d)} w^{(d)}\| \quad \forall d$$

Evaluating on Word Similarity task

Model	MaxSim	AvgSim	AvgSimC
Huang	26.1	62.8	65.7
MSSG	57.26	67.2	69.3
MSSG-NP	59.80	67.3	69.1
CRP	66.4	-	67.0
Retro	-	-	41.7
EM	-	-	61.3
Retro+EM	-	-	58.7
AutoExtend	-	68.9	69.8

Huang et al. (2012b)

Neelakantan et al. (2014)

Li and Jurafsky (2015)

Jauhar et al. (2015)


Rothe and Schutze (2015)

Evaluating on Word Similarity task

Model	MaxSim	AvgSim	AvgSimC
Huang	26.1	62.8	65.7
MSSG	57.26	67.2	69.3
MSSG-NP	59.80	67.3	69.1
CRP	66.4	-	67.0
Retro	-	-	41.7
EM	-	-	61.3
Retro+EM	-	-	58.7
AutoExtend	-	68.9	69.8

Evaluating on Word Similarity task

Model	MaxSim	AvgSim	AvgSimC
Huang	26.1	62.8	65.7
MSSG	57.26	67.2	69.3
MSSG-NP	59.80	67.3	69.1
CRP	66.4	-	67.0
Retro	-	-	41.7
EM	-	-	61.3
Retro+EM	-	-	58.7
AutoExtend	-	68.9	69.8



Sense Embedding for Word Sense Induction

- Word Sense Induction (WSI)
 - automatically discover senses from unlabeled data without referring to any sense inventory

Sense Embedding for Word Sense Induction

- Word Sense Induction (WSI)
 - automatically discover senses from unlabeled data without referring to any sense inventory
- Previous methods on WSI
 - learn co-occurrence vectors by counting
 - learn centroids by clustering

Sense Embedding for Word Sense Induction

- Word Sense Induction (WSI)
 - automatically discover senses from unlabeled data without referring to any sense inventory
- Previous methods on WSI
 - learn co-occurrence vectors by counting
 - learn centroids by clustering
 - **problematic: have to learn a model for each word impractical for real applications**

Sense Embedding for Word Sense Induction

- Compare with existing methods, Sense Embedding:
 - perform joint learning for multiple words
 - learn by predicting
 - learn by predicting >> learn by counting

Sense Embedding for Word Sense Induction

- Compare with existing methods, Sense Embedding:
 - perform joint learning for multiple words
 - learn by predicting
 - **Promising for this task!**

Sense Embedding for Word Sense Induction

System	SemEval-2010 WSI
UoY (2010)	62.4
NMF _{lib} (2011)	62.6
NB (2013)	65.4
Spectral (2014)	60.7
SE-WSI-fix	66.3
SE-WSI-CRP	61.2
CRP-PPMI	59.2
WE-Kmeans	58.6

Best result of the task

By Charniak @Brown U

By CMU

Sense Embedding for Word Sense Induction

System	SemEval-2010 WSI
UoY (2010)	62.4
NMF _{lib} (2011)	62.6
NB (2013)	65.4
Spectral (2014)	60.7
SE-WSI-fix	66.3
SE-WSI-CRP	61.2
CRP-PPMI	59.2
WE-Kmeans	58.6

Joint learning is better!

Neelakantan et al. (2014)

word2vec + Kmeans

Sense Embedding for Word Sense Induction

System	SemEval-2010 WSI
UoY (2010)	62.4
NMF _{lib} (2011)	62.6
NB (2013)	65.4
Spectral (2014)	60.7
SE-WSI-fix	66.3
SE-WSI-CRP	61.2
CRP-PPMI	59.2
WE-Kmeans	58.6

Learn by predicting
is better!

Li and Jurafsky (2015)

Co-occur+ CRP

Conclusion

- Introduced previous and current techniques for Word Embedding
 - Skip-gram
- Describe 3 directions for Sense Embedding
 - Clustering-based
 - Nonparametric
 - Ontology-based
- Sense Embedding for Word Sense Induction
 - **Best performance right now!**

Recent Publications

- Sense Embedding Learning for Word Sense Induction. Linfeng Song, Zhiguo Wang and Daniel Gildea. In submission.
- A Synchronous Hyperedge Replacement Grammar based approach for AMR parsing. Xiaochang Peng, Linfeng Song and Daniel Gildea. In Proceedings of CoNLL 2015, Beijing, China, 2015.
- Joint Morphological Generation and Syntactic Linearization. Linfeng Song, Yue Zhang, Kai Song and Qun Liu. In Proceedings of AACL 2014, Quebec City, Canada, July 27-31, 2014.
- Syntactic SMT Using a Discriminative Text Generation Model. Yue Zhang, Kai Song, Linfeng Song, Jingbo Zhu and Qun Liu. In Proceedings of EMNLP 2014, Doha, Qatar, 2014.

➤ Thank you for listening

➤ Questions?