How Well Sentence Embeddings Capture Meaning

Lyndon White Roberto Togneri Wei Liu
Mohammed Bennamoun

The University of Western Australia

SIGIR
Special Interest Group
on Information Retrieval

Sentence embeddings are vector representations

- Extension beyond word embeddings
- The various models produce embeddings as bi-products
- As they are bi-products what features are represented is not clear.

What could they represent?

Some combination of

- Word content
- Syntactic structure *(Parsing)*
- Subject *(Topic Classification)*
- **Semantic Polarity** *(Sentiment Analysis)*
- Meaning
- Noise
- other factors

Comparison to human similarity rankings

- Have humans rank, or score sample sentences on similarity.
- Evaluate embedding methods by looking for correlations.
- On current methods Gershman and Tenenbaum found generally poor results.

S. J. Gershman and J. B. Tenenbaum, "Phrase similarity in humans and machines," *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 2015, J. Mitchell and M. Lapata, "Vector-based models of semantic composition.," in *ACL*, 2008, pp. 236–244.

Classification by positional meaning

Category	Example
Adhesion to Vertical Surface	There is a magnet on the refrigerator.
Support by Horizontal Surface	There is an apple on the refrigerator.
Support from Above	There is an apple on the branch.
Full Containment	There is an apple in the refrigerator.
Partial Containment	There is an apple in the water.

- Categorise sentences based on the positional component of their meaning.
- Ritter et. al. found sum of word embeddings to outperform all more complex models.

S. Ritter, C. Long, D. Paperno, M. Baroni, M. Botvinick, and A. Goldberg, "Leveraging preposition ambiguity to assess compositional distributional models of semantics," *The Fourth Joint Conference on Lexical and Computational Semantics*, 2015.

Classification by positional meaning

Category	Example
Adhesion to Vertical Surface	There is a magnet on the refrigerator.
Support by Horizontal Surface	There is an apple on the refrigerator.
Support from Above	There is an apple on the branch.
Full Containment	There is an apple in the refrigerator.
Partial Containment	There is an apple in the water.

- Our work broadens this to classify on complete sentence meaning
- It also complements it by working with real world sentences – where overlap does matter.

S. Ritter, C. Long, D. Paperno, M. Baroni, M. Botvinick, and A. Goldberg, "Leveraging preposition ambiguity to assess compositional distributional models of semantics," *The Fourth Joint Conference on Lexical and Computational Semantics*, 2015.

When are sentences equivalent?

Bidirectional Entailment

A and B are semantically equivalent iff

$$A \models B \ \wedge \ B \models A$$

and we call A and B paraphrases

- The staff were rude.
- The service was impolite.
- The staff showed no respect.
- It could have been better.

When are sentences equivalent?

Bidirectional Entailment

A and B are semantically equivalent iff

$$A \models B \ \wedge \ B \models A$$

and we call A and B paraphrases

Equivelence Relation

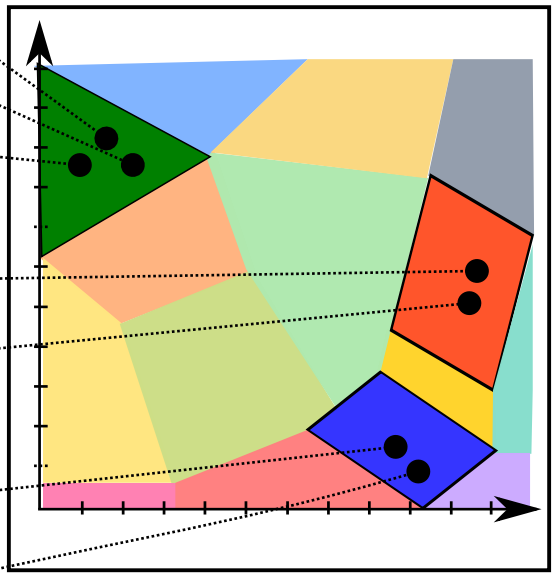
This is an equivalence relation, and so gives rise to a partitioning of all sentences by meaning.

Sentence partitions should map to vector partitions

The staff were rude.
The service was impolite.
The staff showed no respect.

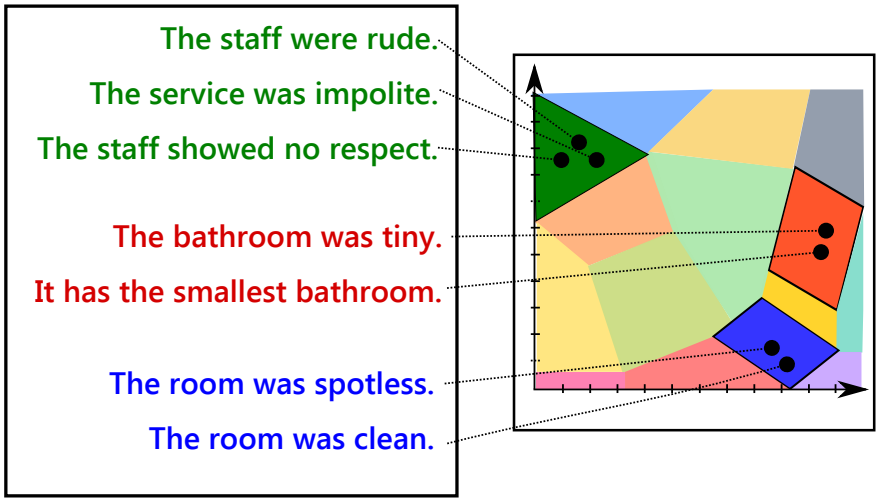
The bathroom was tiny.
It has the smallest bathroom.

The room was spotless.
The room was clean.



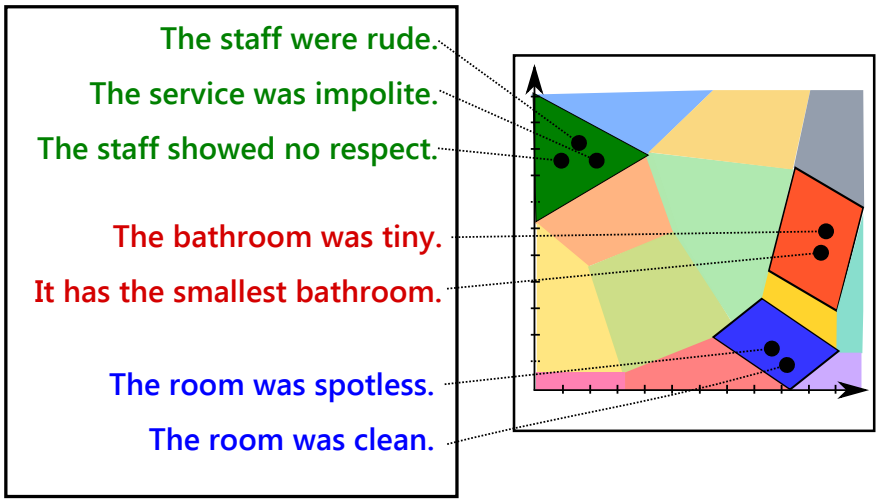
How to Evaluate this?

- Know the partitions in sentence space
- Examine the partitions in the vector space
- See if they are good



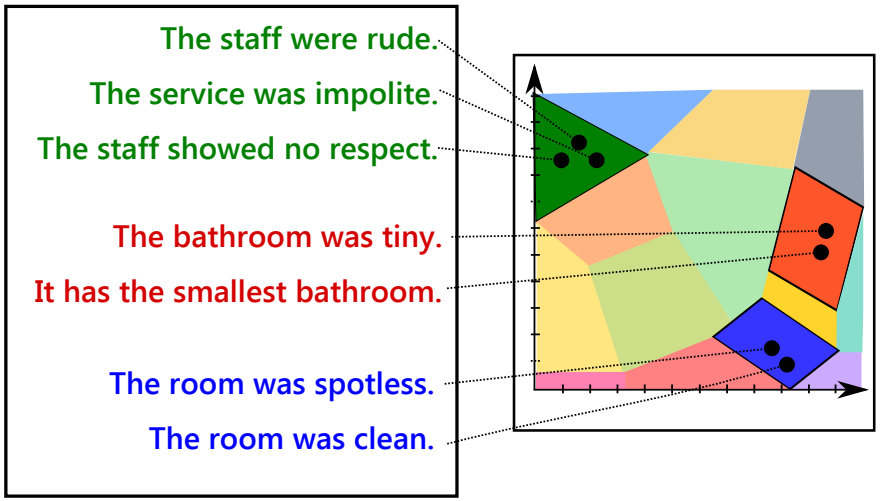
What is good?

- Concentrated
  - All sentences with same meaning go to small area.



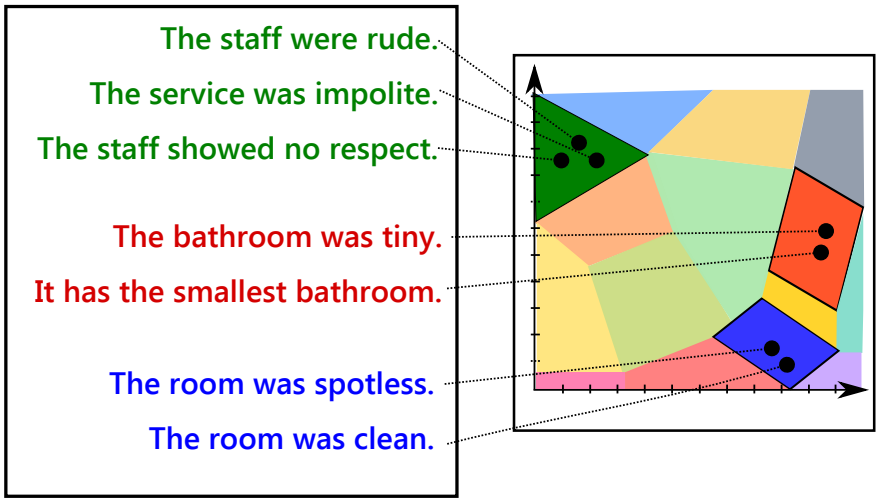
What is good?

- Concentrated
  - All sentences with same meaning go to small area.
- Distinct
  - Should not overlap, should be separate

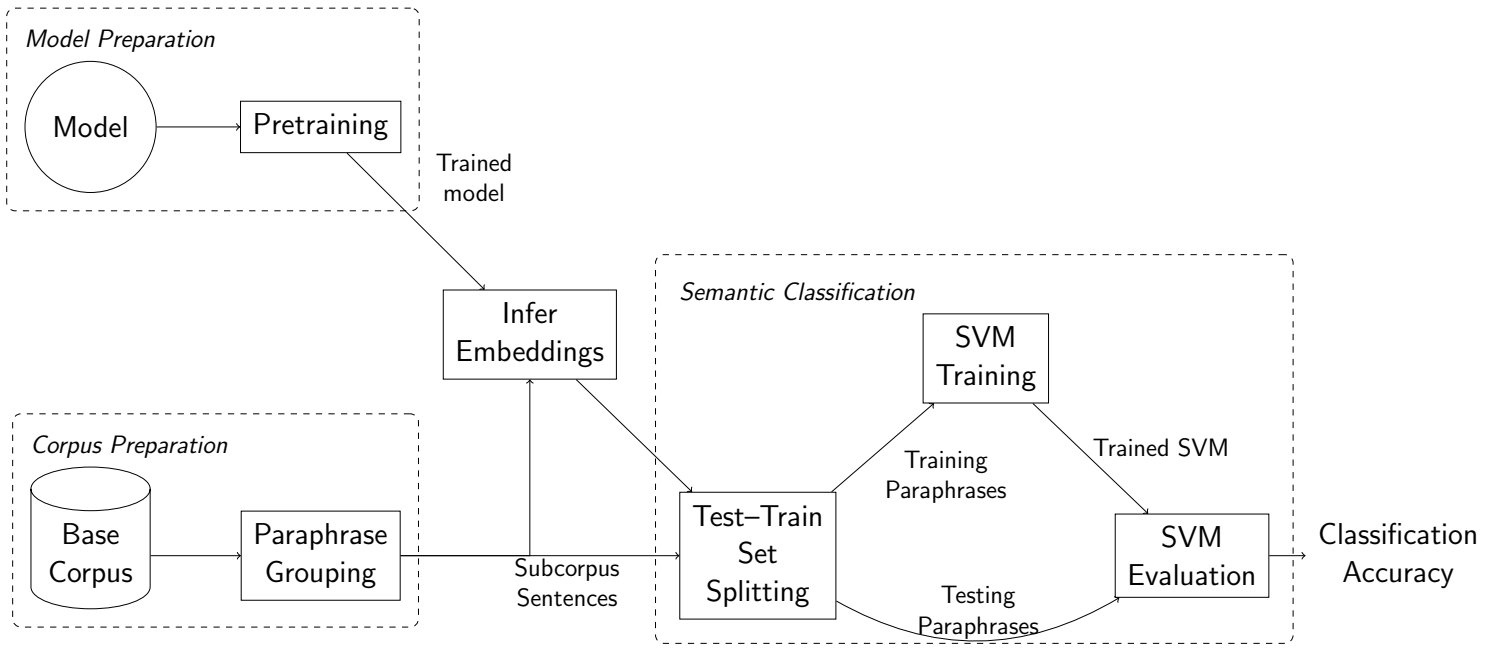


What is good?

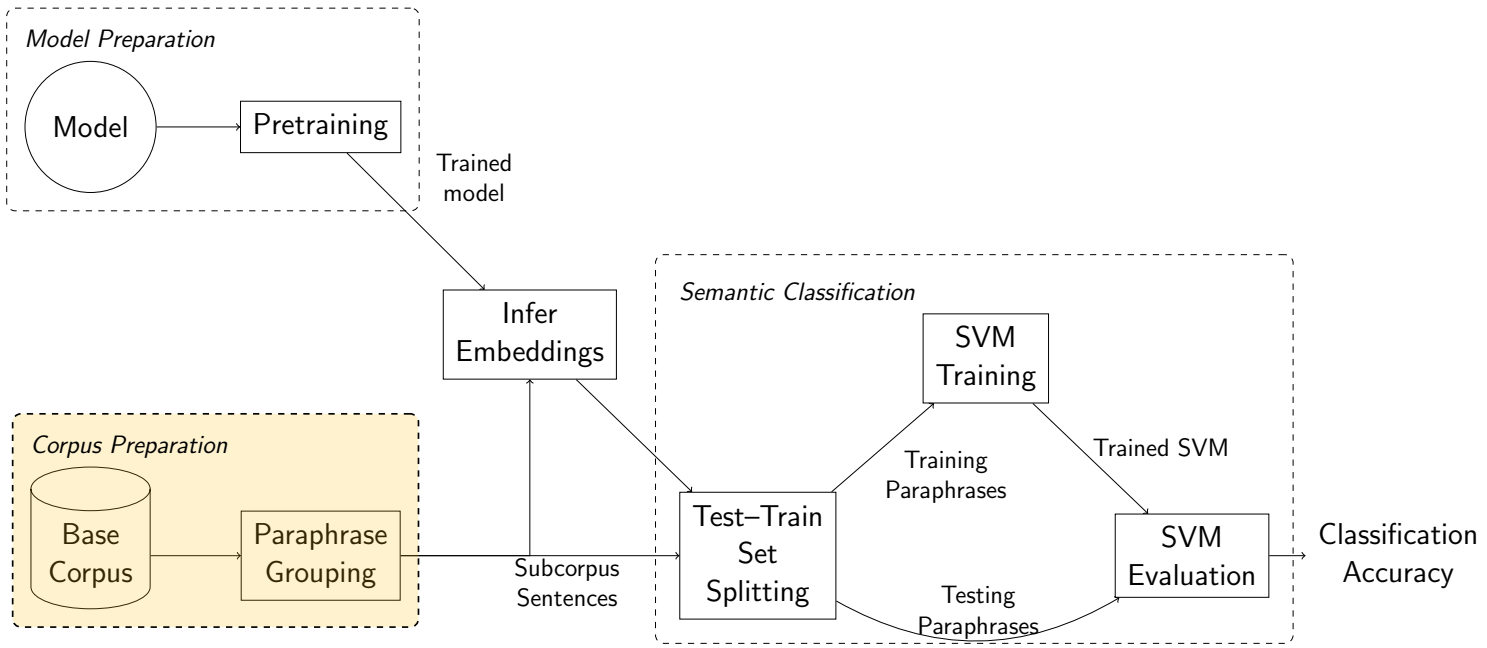
- Concentrated
  - All sentences with same meaning go to small area.
- Distinct
  - Should not overlap, should be separate
- Simple
  - No twists, bulges etc in vector spaces



Evaluate with a Semantic Classification Task



Corpus Preparation



The MSRP Corpus

- Microsoft Research Paraphrase Corpus
- gathered from online news sources
- Pairwise paraphrase detection tests
- Take transitive and symmetric closures

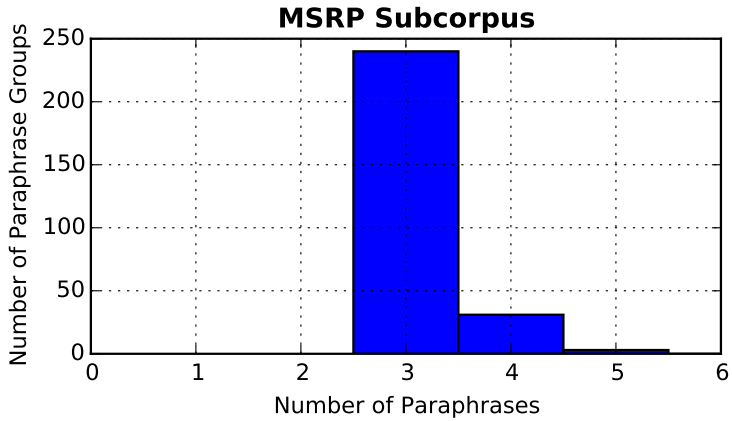
Grouping via closure

If A is a paraphrase of B, and B is a paraphrase of C, then the paraphrase group $\{A, B, C\}$ exists.

W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Third International Workshop on Paraphrasing (IWP2005)*, Asia Federation of Natural Language Processing, 2005.

The MSRP Corpus

- Microsoft Research Paraphrase Corpus
- gathered from online news sources
- Pairwise paraphrase detection tests
- Take transitive and symmetric closures



859 sentences, divided into 273 groups.

W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Third International Workshop on Paraphrasing (IWP2005)*, Asia Federation of Natural Language Processing, 2005.

Paraphrase groups from MSRP

- only intel corp. has a lower dividend yield .
- only intel 's 0.3 percent yield is lower .
- only the intel corporation has a lower dividend yield .
- only intel corp. 's 0.3 percent yield was lower .

Paraphrase groups from MSRP

- three such vigilante-style attacks forced the hacker organizer , who identified himself only as “ eleonora67 , ” to extend the contest until 6 p.m. edt sunday.
- three such vigilante-style attacks forced the hacker organizer , who identified himself only as “ eleonora [67] , ” to extend the contest until 7 p.m. est sunday.
- three such vigilante-style attacks forced the hacker organiser , who identified himself only as “ eleonora67] , ” to extend the contest until 8am (aest) today.
- three such vigilante-style attacks forced the hacker organizer , who identified himself only as “ eleonora [67] , ” to extend the contest until 3 p.m. arizona time sunday.

Paraphrase groups from MSRP

- belcher said the airport 's conference room was serving as a makeshift shelter for several area families who hiked up the wooded hillside in advance of rising water .
- belcher said the airport 's conference room became a shelter for several area families who had hiked up wooded hillsides in advance of rising water .
- belcher said the airport 's conference room became a shelter for several families who had hiked up hillsides ahead of rising water .

Issues with MSRP Corpus

Difficulty:

- Often intra-group differences only in single word or punctuation.
- Often inter-group sentences differs entirely in topic

W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Third International Workshop on Paraphrasing (IWP2005)*, Asia Federation of Natural Language Processing, 2005.

Issues with MSRP Corpus

Difficulty:

- Often intra-group differences only in single word or punctuation.
- Often inter-group sentences differs entirely in topic

Loose Semantic Equivalence

":.the majority of the equivalent pairs in this dataset exhibit 'mostly bidirectional entailments', with one sentence containing information 'that differs' from or is not contained in the other."

W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Third International Workshop on Paraphrasing (IWP2005)*, Asia Federation of Natural Language Processing, 2005.

The Opinosis Corpus

- Collection of highly redundant opinions
- Intended for evaluation of Summarisation algorithms
- Required manual partitioning

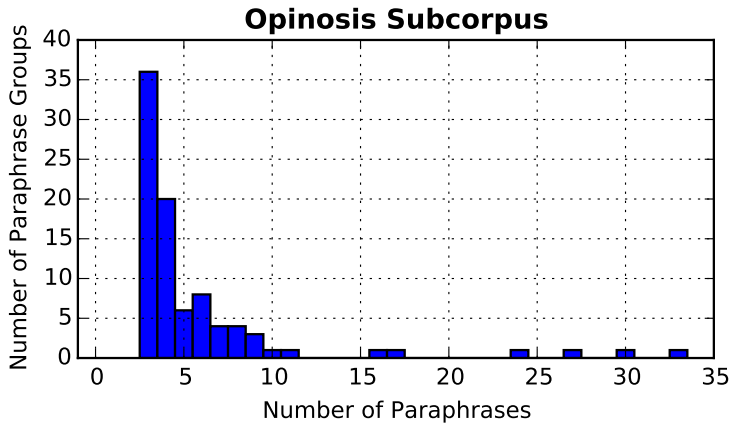
About:

- Electronics
- Hotels
- Auto-mobiles

K. Ganesan, C. Zhai, and J. Han, "Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions," in *Proceedings of the 23rd International Conference on Computational Linguistics*, Association for Computational Linguistics, 2010, pp. 340–348.

The Opinosis Corpus

- Collection of highly redundant opinions
- Intended for evaluation of Summarisation algorithms
- Required manual partitioning



521 Sentences, divided into 89 groups

K. Ganesan, C. Zhai, and J. Han, "Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions," in *Proceedings of the 23rd International Conference on Computational Linguistics*, Association for Computational Linguistics, 2010, pp. 340–348.

Paraphrase groups from Opinosis

- Perfect Location and Lovely Inn
- Great hotel, nice location for everything !
- Loved this hotel and location !
- A great hotel in a great location
- Great Hotel in a Great Location
- Overall though, the location was great and the hotel was nice .
- Lovely hotel and excellent location .

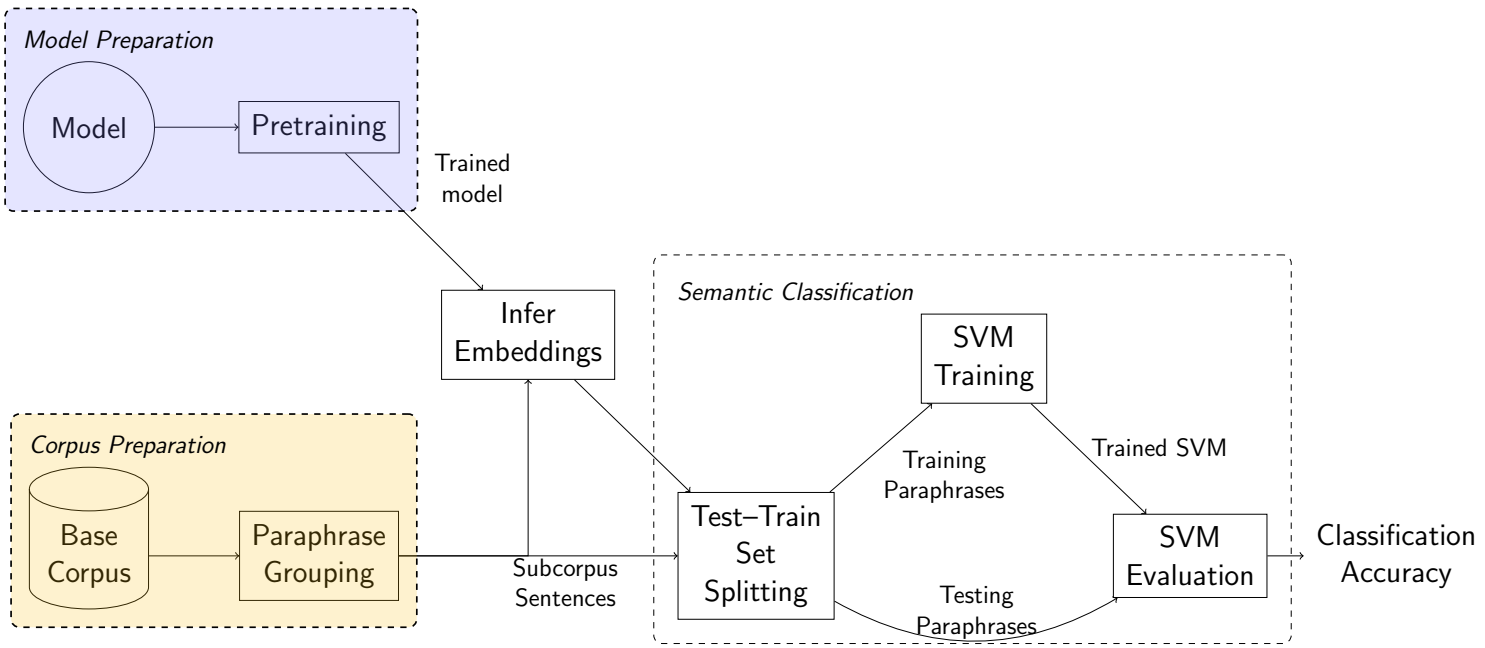
Paraphrase groups from Opinosis

- So, the location was awesome .
- Again, the location was great !
- You cannot beat the location !
- The location could not be any better .
- In addition, it is location, location, location .
- That said, the location is marvelous .

Paraphrase groups from Opinosis

- Best location in Fisherman's Wharf .
- Perfect location to Fishermans Wharf .
- Perfect Fisherman's Wharf Location
- Perfect location near Fisherman's Wharf .
- The location is great, only 2 blocks from Fisherman's Wharf .

Models and Model Preparation



Bag of Words (BOW / PCA BOW)

Bag of Words

A vector of word occurrence counts. It's length is equal to the vocabulary size.

H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

Bag of Words (BOW / PCA BOW)

Bag of Words

A vector of word occurrence counts. It's length is equal to the vocabulary size.

Why Principle Component Analysis?

Unsupervised dimensionality reduction, to test if BOW performance was due to is very large number of dimensions.

H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

Sum/Mean of Word Embeddings

- Simply adding up word embeddings
- We used pretrained Google News skip grams

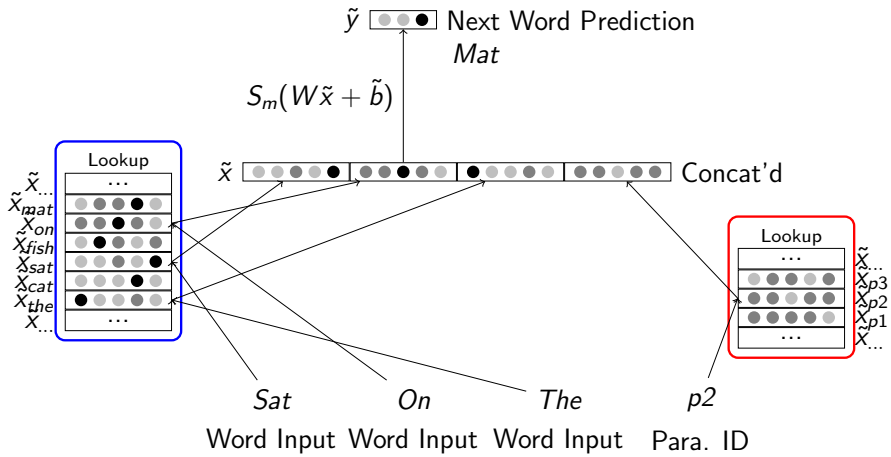
Identical Cosine Similarities

for representations of sentences A and B:
SOWE: s_A, s_B
MOWE:, m_A, m_B

$$d_{cos}(s_A, s_B) = d_{cos}(m_A, m_B)$$

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

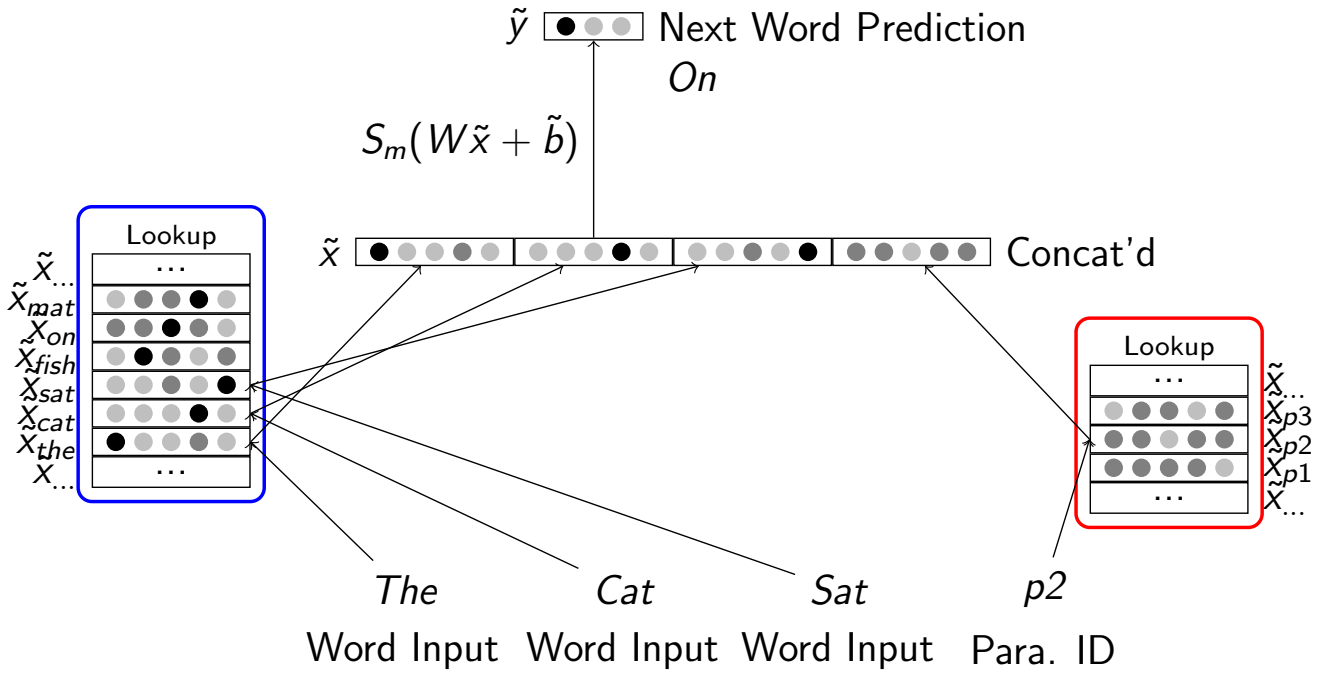
Paragraph Vector – Distributed Memory



- Similar to CBOW
- Uses paragraph vector as "memory" slot
- Giving an embedding for any word sequence

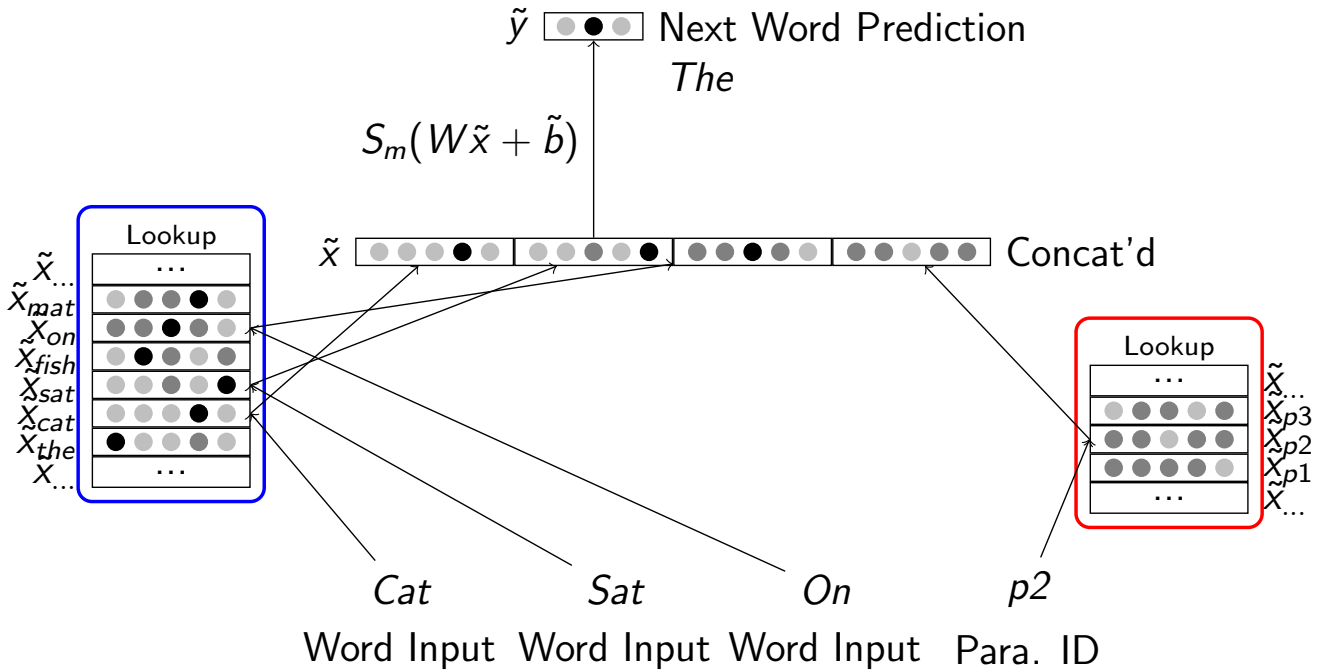
Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1188–1196.

PV–DM



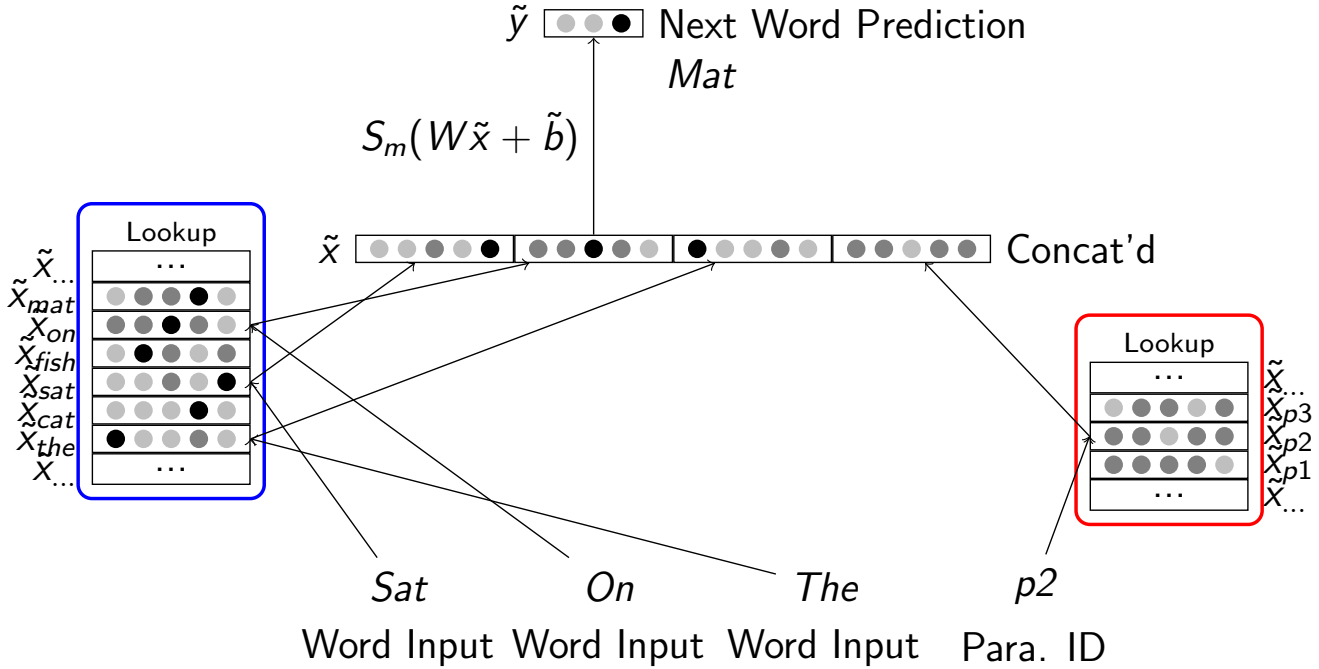
Process: The Cat Sat | On | The Mat

PV–DM



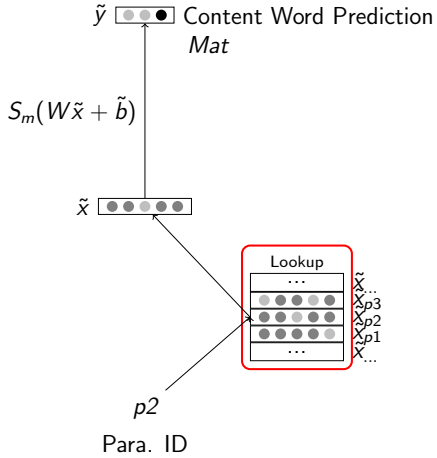
Process: The | Cat Sat On | The | Mat

PV–DM



Process: The Cat Sat On The Mat

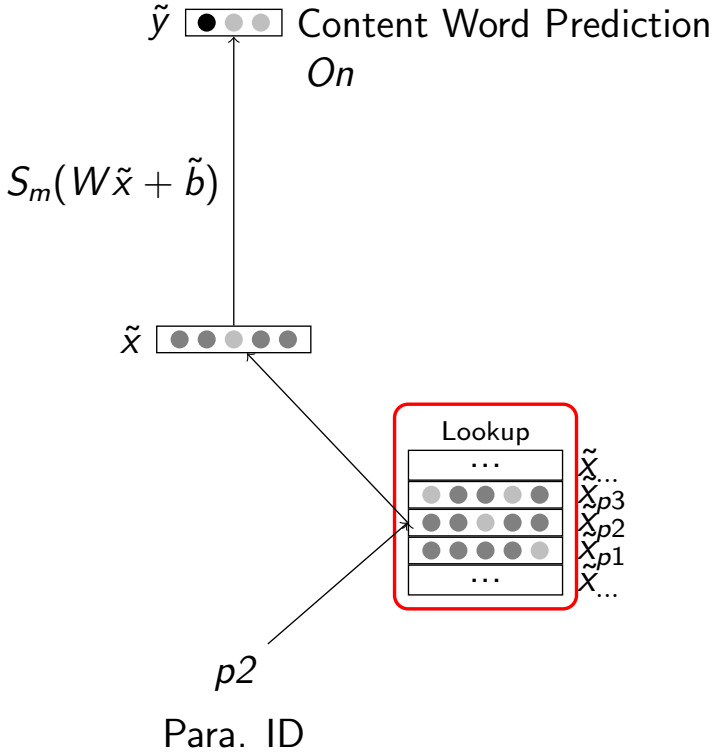
Paragraph Vector – Distributed Bag of Words



- Similar to skip grams
- Uses paragraph vector to predict word content
- Giving an embedding for any word sequence

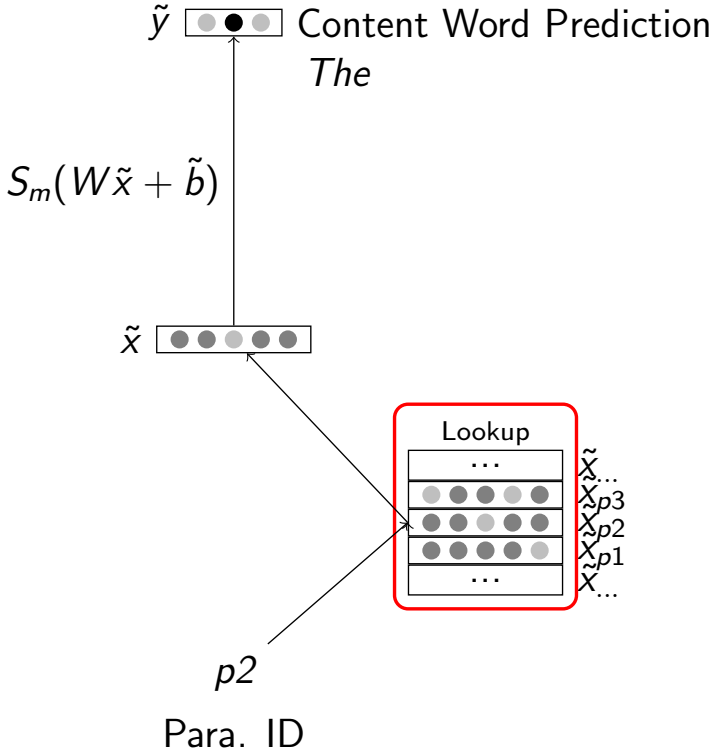
Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1188–1196.

PV–DBOW



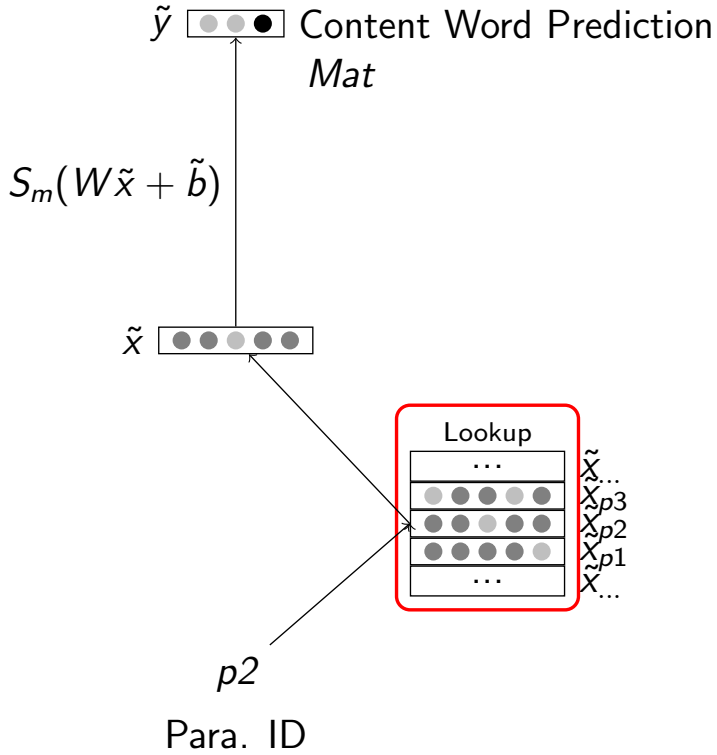
Process: The Cat Sat On The Mat

PV–DBOW



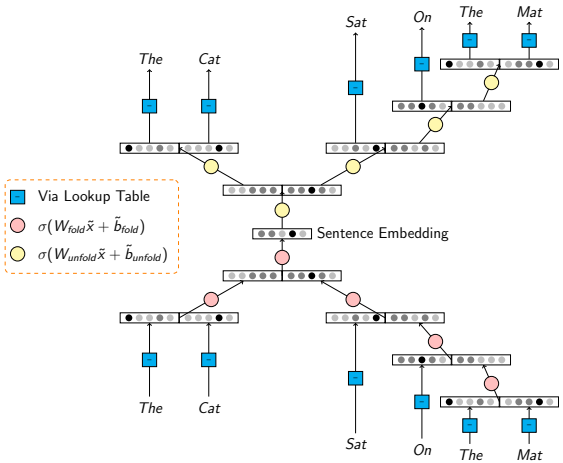
Process: The Cat Sat On The Mat

PV–DBOW



Process: The Cat Sat On The Mat

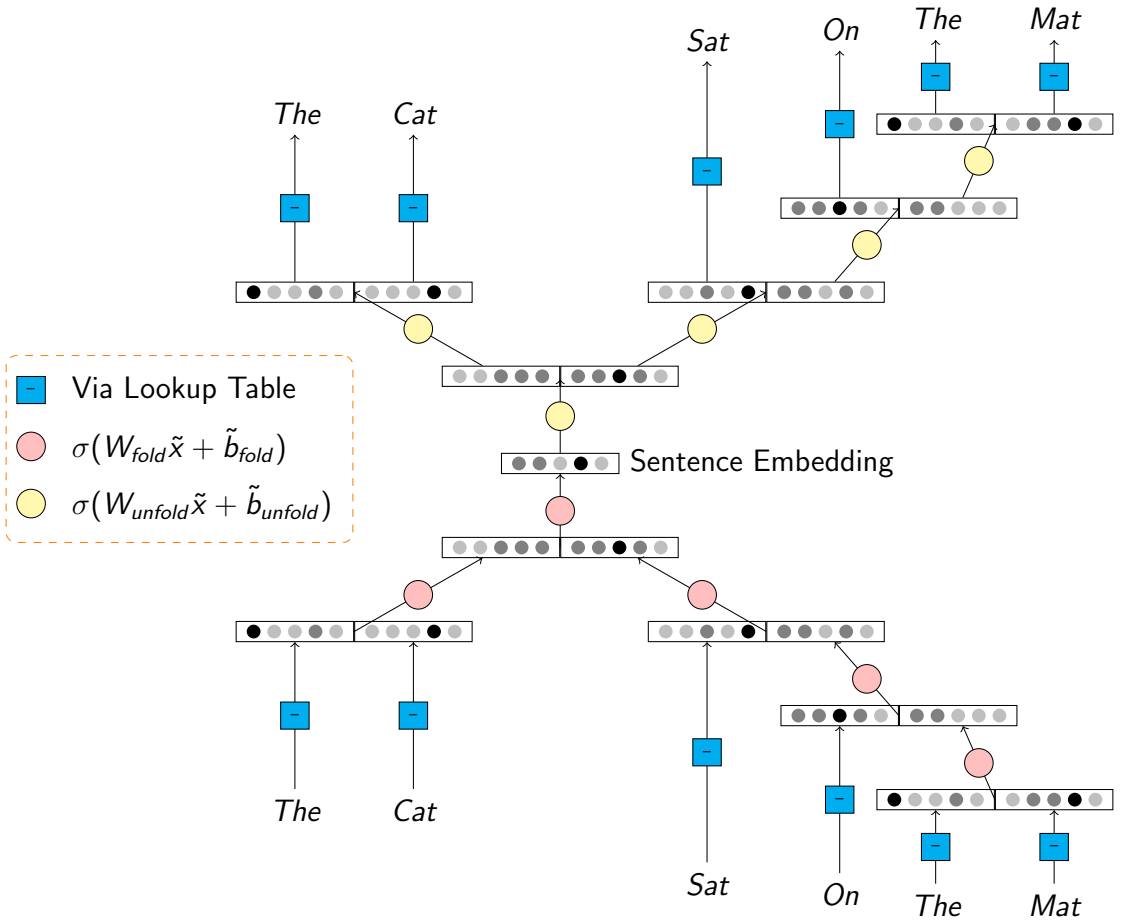
Unfolding Recursive AutoEncoder



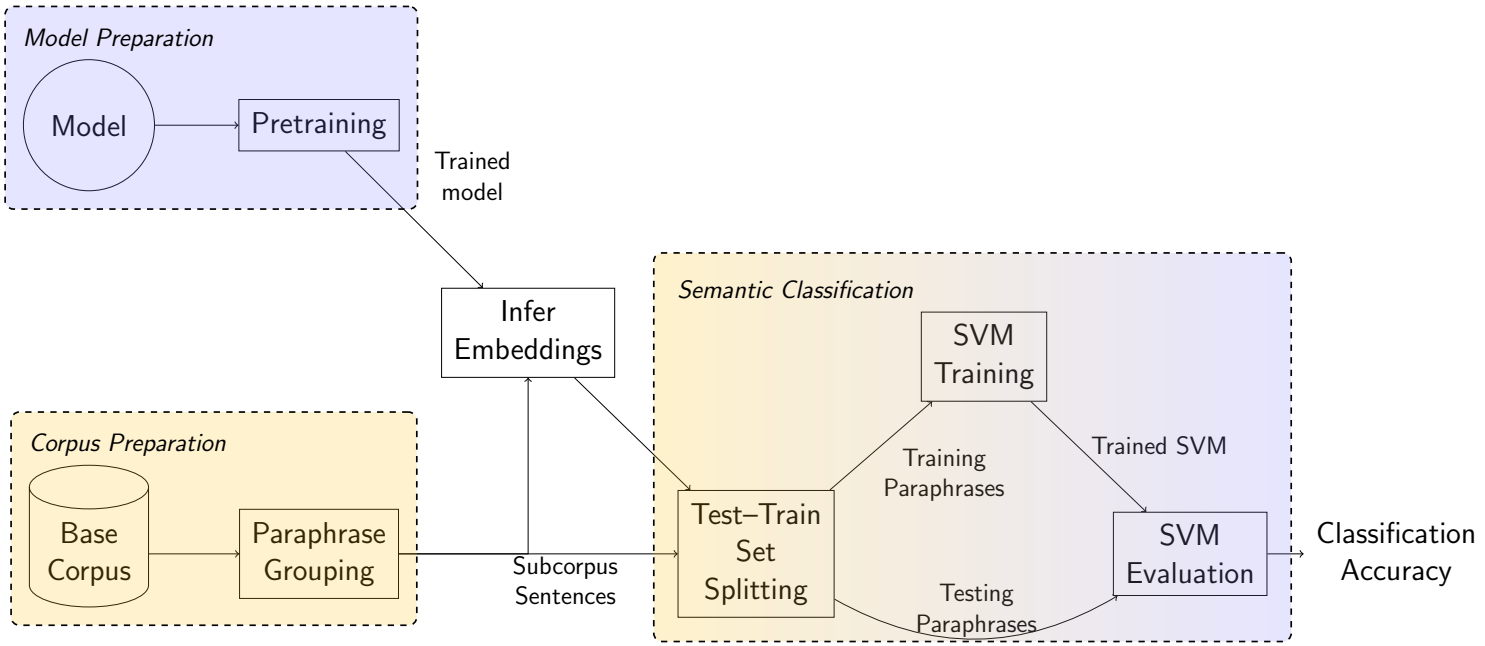
- Compositional Model
- Recursive pairwise merging
- Trained to unfold to original.

R. Socher, E. H. Huang, J. Pennington, A. Y. Ng, and C. D. Manning, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection," in *Advances in Neural Information Processing Systems 24*, 2011.

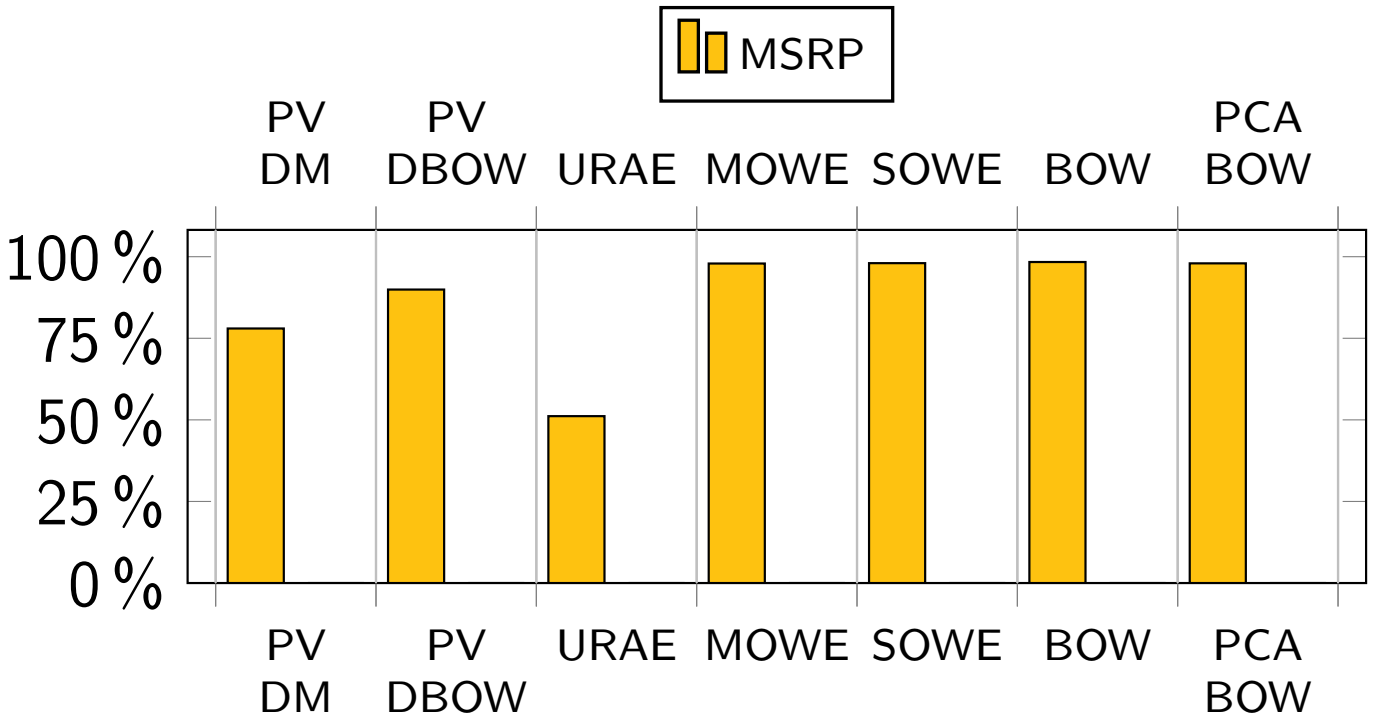
URAE


The
Cat
Sat
On
The
Mat
Via Lookup Table
$\sigma(W_{fold}\tilde{x} + \tilde{b}_{fold})$
$\sigma(W_{unfold}\tilde{x} + \tilde{b}_{unfold})$
Sentence Embedding
The
Cat
Sat
On
The
Mat

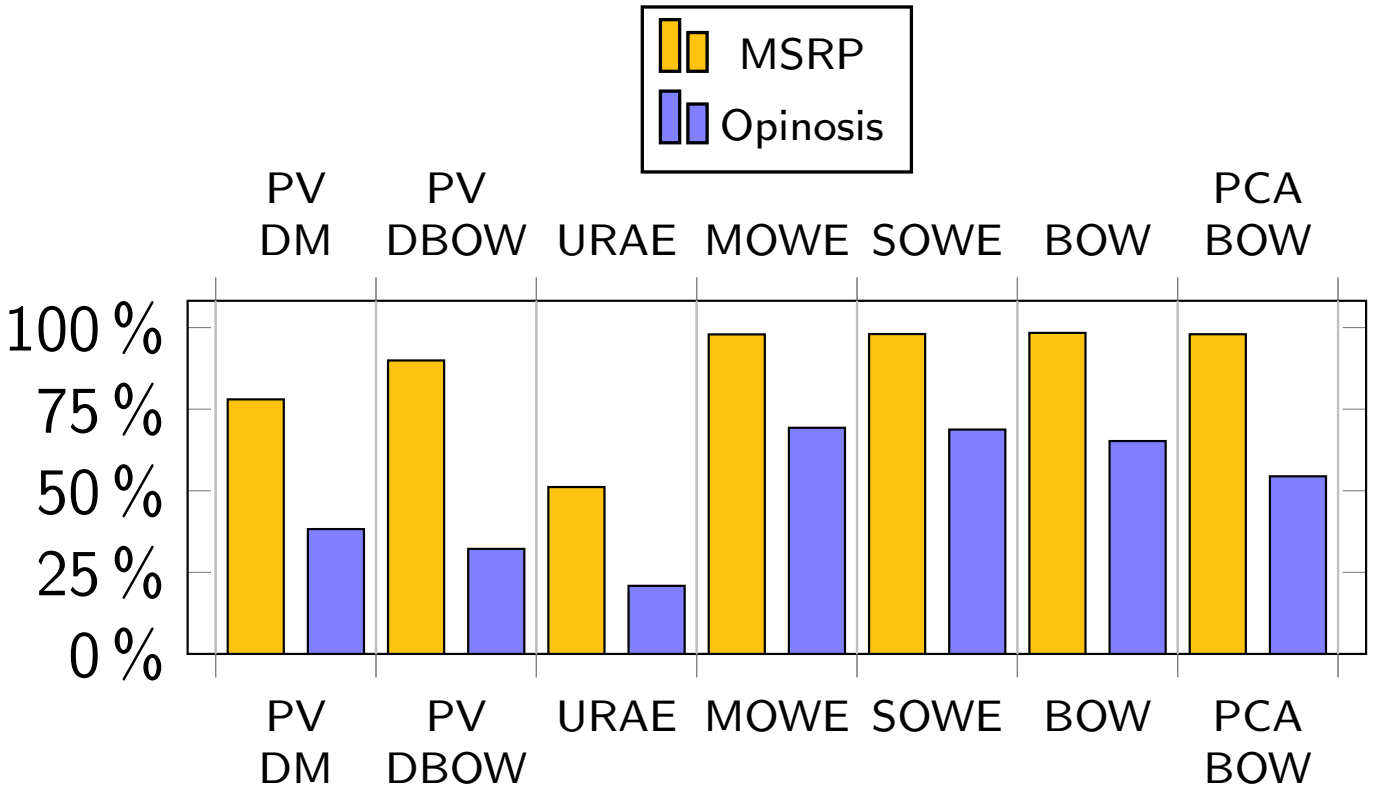
Experimental Evaluation and Results



Semantic Classification Success Rate



Semantic Classification Success Rate



Limitations

- Embedding Training corpus distinct from Evaluation Corpus
- URAE playing at disadvantage
  - 200 dim to others 300 dim
- Only linearly separable partitioning considered.

Conclusion

- Evaluate the vector space partitioning, by semantic equivalence.
- using linear SVM for semantic classification task
- BOW, MOWE and SOWE unexpectedly strong
- Word content matters, and must be preserved

Annex: Exact Results

Name	MSRP	Opiniosis
PV–DM	78	38.26
PV–DBOW	89.93	32.19
URAE	51.14	20.87
MOWE	97.91	69.3
SOWE	98.02	68.75
BOW	98.37	65.23
PCA–BOW	97.96	54.43