

Cross-lingual Paraphrase Detection

Supervisors: Prof. Vinay Namboodiri and Prof. B.V. Rathish Kumar

Md. Enayat Ullah

March 09, 2016

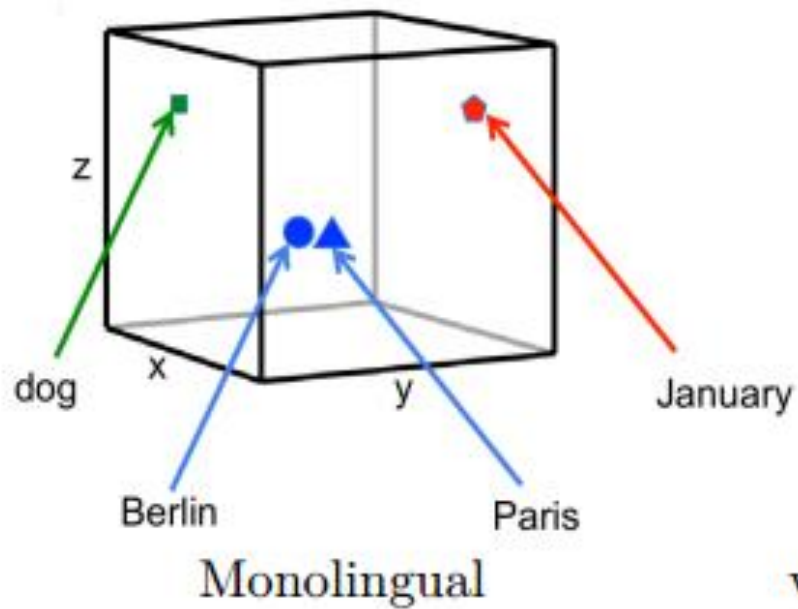
Outline

- Motivation
- Word Representations
- Bilingual Word Representations
- Paraphrase Detection
- Interim Results
- Future work: Seq2Seq Learning

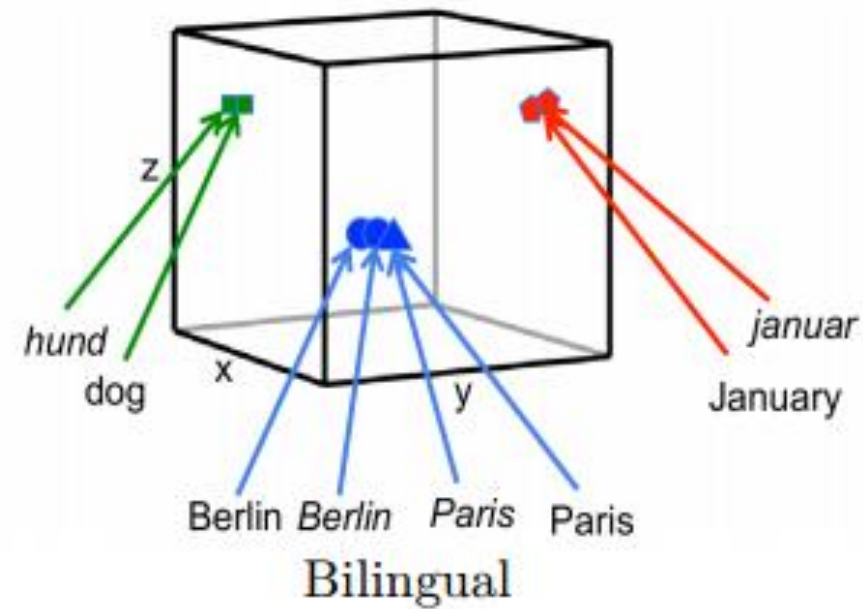
Objective

Statement	Paraphrase	Gold Truth
They had published an advertisement on the Internet on June 10, offering the cargo for sale	वे बिक्री के लिए माल की पेशकश, 10 जून को इंटरनेट पर एक विज्ञापन प्रकाशित किया था	YES
The initial report was made to New York Police department.	आरोप दिसंबर को किए गए कुछ पुलिस रिपोर्ट की वजह से उपजी	NO
बेटों एंथनी और केली, बेटियों लिंडा आशा और नोरा सोमर्स - और चार पोते वह अपने चार बच्चों के रूप में करते उसे जीवित रहते हैं।	Hope is survived by his wife; sons Anthony and Kelly; daughters Linda and Nora Somers; and four grandchildren.	YES
In response to sluggish sales cisco pared spending.	सिस्को सुस्त बिक्री के लिए क्षतिपूर्ति की तिमाही के दौरान खर्च मुकाबले।	NO

Word Embeddings



vs



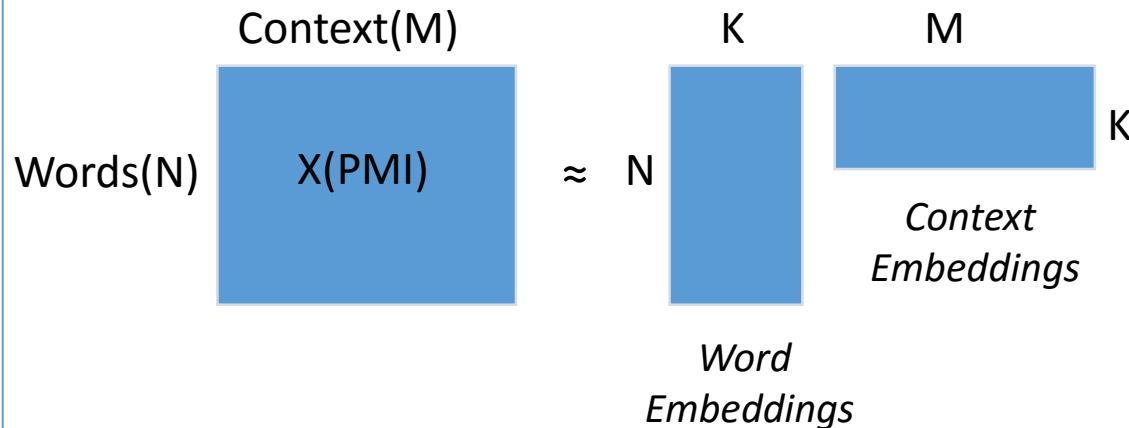
Word Embeddings

Distributional Semantics (*Count*)

Used since the 90's

Sparse word-context PMI/PPMI matrix

Decomposed with SVD/ Matrix Factorization

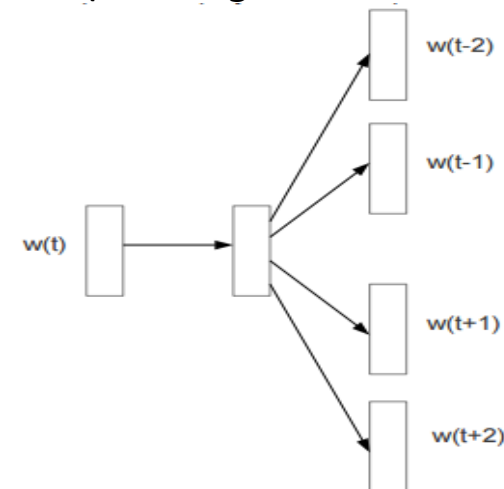


Word Embeddings (*Predict*)

Inspired by deep learning

`word2vec` (Mikolov et al., 2013)

GloVe (Pennington et al., 2014)




$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$$

Underlying Theory: **The Distributional Hypothesis** (Harris, '54; Firth, '57)

“Similar words occur in similar contexts”

Bilingual Word Embeddings

	Separate Training	Joint Training
Parallel Copora Limited Availability of data and bilingual lexicons!	Learning a Transformation matrix between Bilingual Lexicons	Learning Cross-lingual Word Embeddings via Matrix Co-factorization(<i>Shi et al</i>)
Comparable Copora Advantage: Lots of Data(Wikipedia)		<p>Bilingual Distributed Word Representations from Document-Aligned Comparable Data(<i>Vulic et al</i>)</p>  <pre>graph LR; A[Preprocessing] --> B[Merge and shuffle]; B --> C[Joint Training]</pre>

Qualitative Evaluation

Bilingual Lexicon Extraction

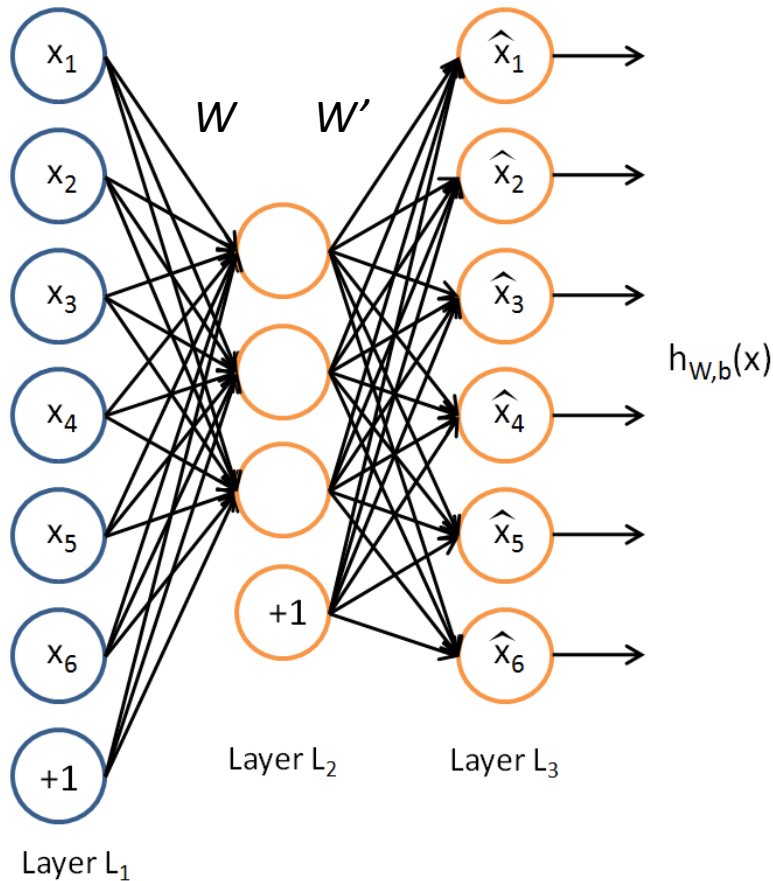
Word	Nearest Neighbours in Sorted Order
knowledge	श्रीमद्भगवद्गीता doer vedas ignorance ज्ञान
father	पिता mother wife माता child
बाज़ार	price शेयर बाजार market markets
बेहतर	better अच्छी अगर improve निर्धारित
sudden	cardiac पूर्णहृदरोध defibrillation ऊष्माघात अतिताप
in	a the में to की
भेजना	भेजने ईमेल प्रेषण email bes
पीना	रिसेप्टरों मूत्रवर्धक सेवन drinks पेय
run	spielen away travel fahrt athleten
vater	father bruder wife son eltern
send	भेजने protocol BDR SSH NSSA

Qualitative Evaluation

Suggested Word Translation in Context (SWTC)

Sentence	Possible Translations	Gold Truth	Model's Prediction
He is engaged to that foreign actress.	सगाई व्यस्त बंधना संलग्न लगना	सगाई	सगाई
Why are you engaging me in useless conversation?	सगाई व्यस्त बंधना संलग्न लगना	संलग्न लगना	लगना
The match was well played.	मैच दियासलाई विवाह जोड़ा मुक्काबला होड़ मेल खेल प्रतियोगिता माचिस	मैच मुक्काबला खेल	मेल
The couple looks like a perfect match.	मैच दियासलाई विवाह जोड़ा मुक्काबला होड़ मेल खेल प्रतियोगिता माचिस	विवाह जोड़ा मेल	मुक्काबला

Paraphrase Detection

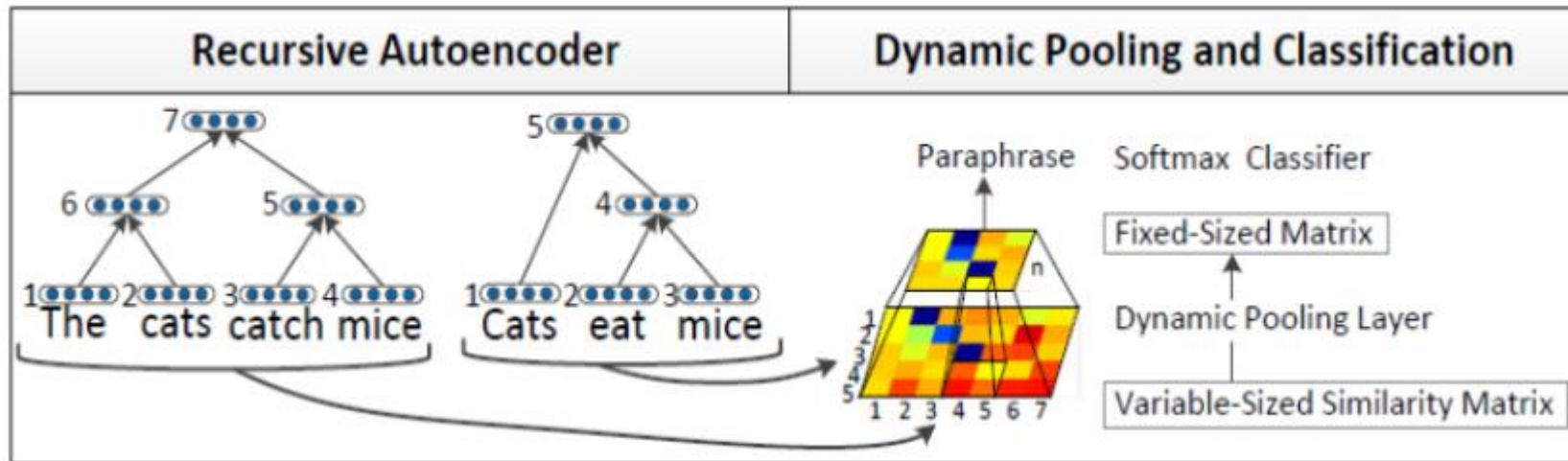


Autoencoders:

- Learn lower-dimensional embeddings
- Reconstruction loss as the objective

$$L(x, x') = ||x - x'||^2 = ||x - \sigma_2(W' \sigma_1(Wx))||^2$$

Paraphrase Detection



Dynamic Pooling and Recursive Autoencoders for Paraphrase Detection
(Socher et al.)

- Pooling to get a fixed-size representation
- Train any classifier like Softmax/SVM on pooled similarity matrices

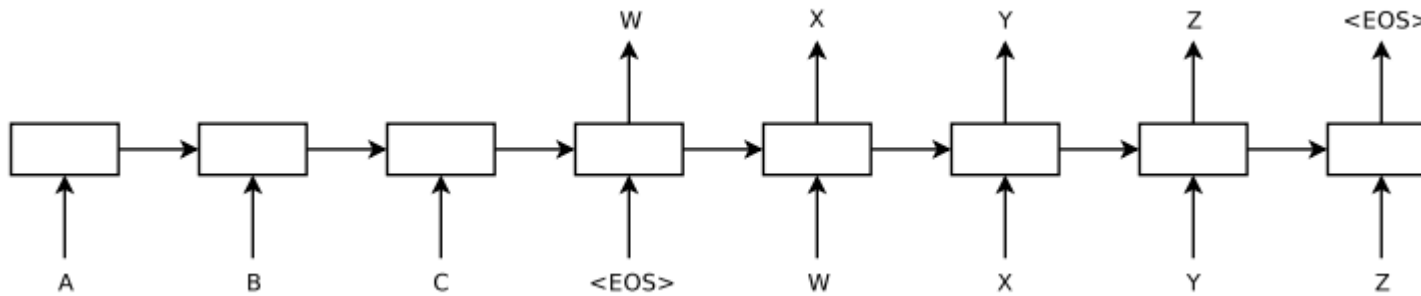
Interim Result

Evaluation on MSR paraphrase data

Languages	Softmax	Soft-max+features	Linear SVM+features	RBF SVM+features
EN to EN	66.21	68.14	68.81	70.15
EN to HI	65.53	66.55	63.23	66.55
HI to EN	64.64	65.98	63.86	67.21
HI to HI	60.78	60.34	62.45	64.67

Future Work

- Sequence2Sequence Learning with Neural Networks (*Sutskever et al, '14*)



- Excellent results in Machine Translation(MT)
- Adapting the model to paraphrase detection module

References

1. Ivan Vulic and Marie-Francine Moens. Bilingual distributed word representations from document aligned comparable data, 2015
2. Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. 2011.
3. Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. Learning cross-lingual word embeddings via matrix co-factorization. Volume 2: Short Papers, page 567, 2015.
4. Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems*. 2014.