Deep Learning for NLP: Words as Vectors

School of Information Studies
Syracuse University

NLP Word Representations

Distributional similarity based representations

- Representing a word by means of its neighbors
  - "You shall know a word by the company it keeps." (Firth 1957)
  - Or linguistic items with similar distributions have similar meanings
  - This idea is also in similarity measures such as Mutual Information

One of the most successful ideas of modern statistical NLP

government debt problems turning into	banking	crises as has happened in
saying that Europe needs unified	banking	regulation to replace the hodgepodge

↖ These words will represent *banking* ↗

You can vary whether you use local or large context to get a more syntactic or semantic clustering

Diagram MS

NN Dense Word Vectors

Combine vector space semantics with probabilistic models to predict vectors of context words

- (Bengio et al 2003, Collobert & Weston 2008, Turian et al 2010)

A word is represented as a dense vector of numbers representing its context words

Older related ideas are

- SVD on term-context matrix
- Brown clusters

$$\textit{linguistics} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

Diagram MS

NN Learning Dense Word Vectors

Set up a classification task from unsupervised data where we have positive training examples directly from the data, and negative examples obtained by substituting a random word in the context (as described in Collobert et al JMLR 2011)

- Positive example:
  "cat sits on the mat"
- Negative example:
  "cat sits jeju the mat"

Classify which contexts are noise

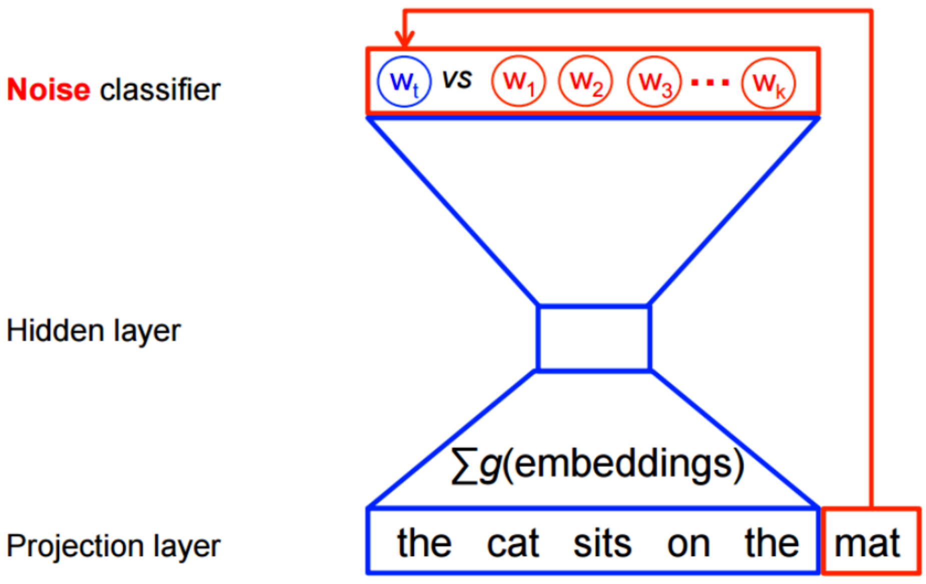


Diagram TensorFlow tutorial

School of Information Studies
Syracuse University

Word2Vec

The word vector classifier gives a simpler and faster implementation of a (shallow) RNN, (Mikolov 2013) with 2 algorithms

- CBOW (continuous bag of words) predicts the current word w, given the neighboring words in the window
- SkipGram predicts the neighboring words, given w

Allows the NN to be applied to large amounts of data

Hyperparameters

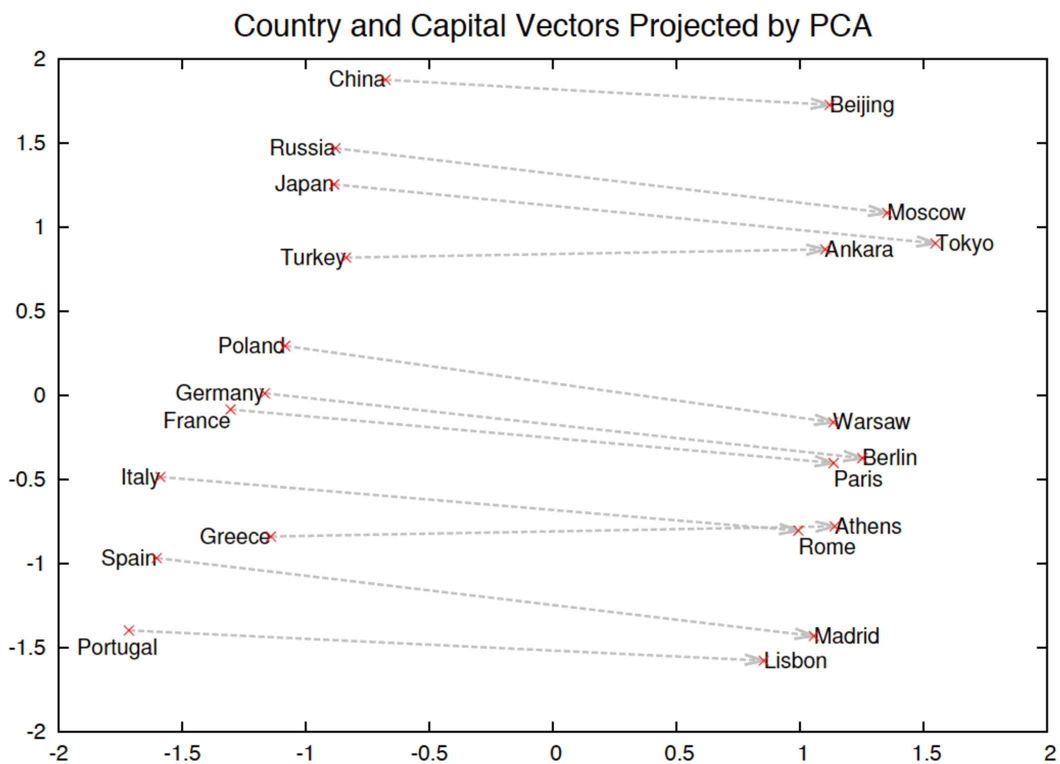
- Window size – the number of context words
- Network size – the number of neurons in the hidden layer
- Other parameters such as negative subsampling number

School of Information Studies
Syracuse University

Dense Word Vector Space

In the resulting space, similar words should be closer together

- Syntactic similarities, such as word tense or plurals
- Semantic similarities



Length 1000 vectors projected to 2D, diagram from Mikolov et al 2013 (NIPS)

Dense Word Vector Space

Showing some of the nearest words in the vector space (Mikolov 2013)

target:	Redmond	Havel	ninjutsu	graffiti	capitulate
	Redmond Wash.	Vaclav Havel	ninja	spray paint	capitulation
	Redmond Washington	president Vaclav Havel	martial arts	grafitti	capitulated
	Microsoft	Velvet Revolution	swordsmanship	taggers	capitulating

Analogies Task

How can we evaluate whether the dense word vectors represent good word similarities?

Solve problems of the type:

- "a is to b as c is to __"

Mikolov et al (HLT 2013) constructed a test set of 8k syntactic relations

- Noun plurals and possessives, verb tenses, adjectival comparitives and superlatives

Semantic test set from Semeval-2012 Task 2

Word Relationships

Mikolov's results are that analogies testing dimensions of similarity can do quite well just by doing vector subtractions

- Syntactically – plurals, verb tenses, adjective forms

$$x_{apple} - x_{apples} \approx x_{car} - x_{cars} \approx x_{family} - x_{families}$$

- Semantically (analogies from Semeval 2012 task 2)

$$x_{shirt} - x_{clothing} \approx x_{chair} - x_{furniture}$$

$$x_{king} - x_{man} \approx x_{queen} - x_{woman}$$

Diagram RS

Word Analogies

Results from Mikolov et al 2013 (HLT) using word2vec

- Trained on 320M words of broadcast news data
- With 82k word vocabulary

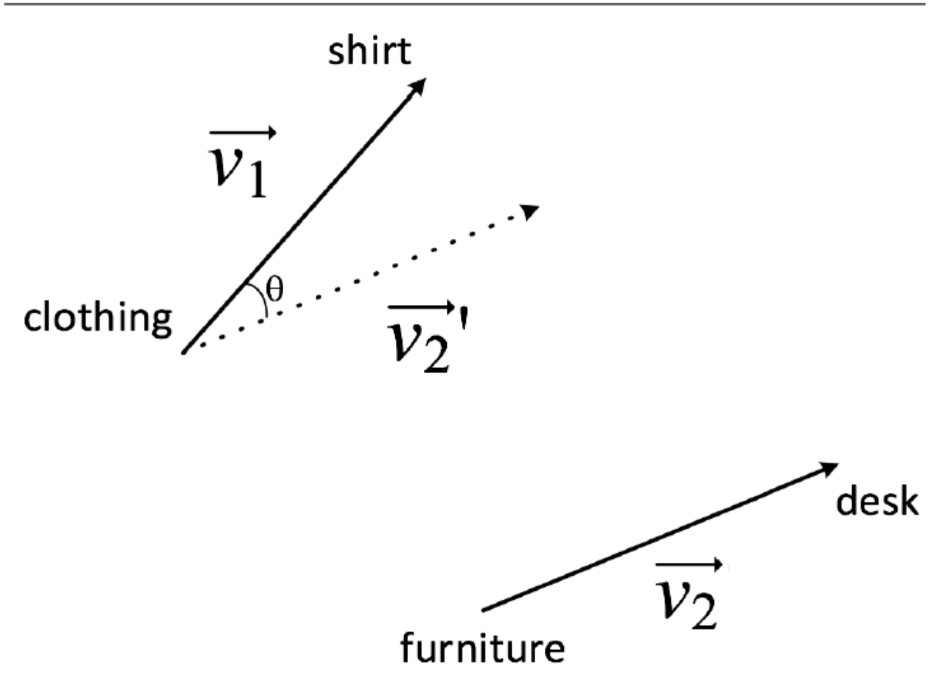


Diagram MS

Method	Syntax % correct
LSA 320 dim	16.5 [best]
RNN 80 dim	16.2
RNN 320 dim	28.5
RNN 1600 dim	39.6
Method	**Semantics Spearm ρ**
UTD-NB (Rink & H. 2012)	0.230 [Semeval win]
LSA 640	0.149
RNN 80	0.211
RNN 1600	0.275 [new SOTA]