Introduction to Word Embedding

What’s word embedding

- Word embedding = Word Vector = Word Representation

words → $\mathbb{R}^n$

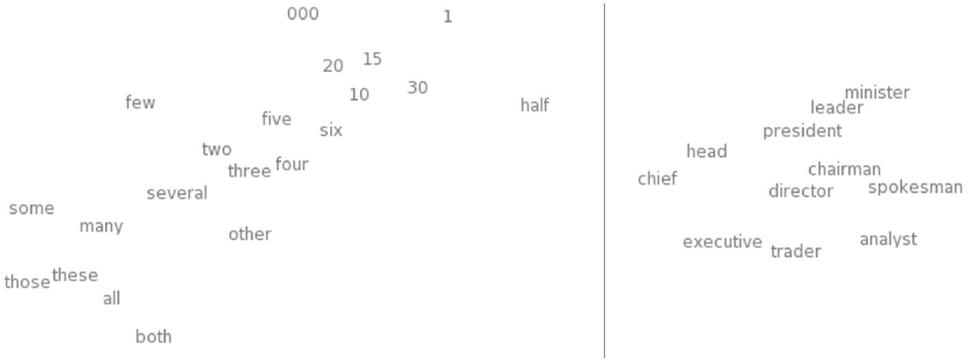
- Word embedding is a **parameterized function** mapping words in some language to high-dimensional vectors (perhaps 200 to 500 dimensions)

$$W(\text{“cat"}) = (0.2,\ -0.4,\ 0.7,\ \ldots)$$

$$W(\text{“mat"}) = (0.0,\ 0.6,\ -0.1,\ \ldots)$$

What’s word embedding[Cont.]

- Visualization of word embedding



- Analogies between words seem to be encoded in the difference vectors between words

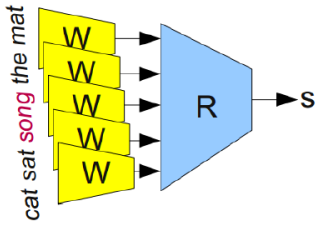
$$W(\text{“woman”}) - W(\text{“man”}) \simeq W(\text{“aunt”}) - W(\text{“uncle”})$$

$$W(\text{“woman”}) - W(\text{“man”}) \simeq W(\text{“queen”}) - W(\text{“king”})$$

- It turns out, though, that more sophisticated relationships are also encoded in this way. It seems almost miraculous!

An Application Example of Word Embedding

- one task we might train a network for is predicting whether a 5-gram (sequence of five words) is 'valid.'



Modular Network to determine if a 5-gram is 'valid' (From Bottou (2011) (http://arxiv.org/pdf/1102.1808v

$$R(W(\text{``cat''}), W(\text{``sat''}), W(\text{``on''}), W(\text{``the''}), W(\text{``mat''})) = 1$$

$$R(W(\text{``cat''}), W(\text{``sat''}), W(\text{``song''}), W(\text{``the''}), W(\text{``mat''})) = 0$$

- In order to predict these values accurately, the network needs to learn good parameters for both W and R

Model for W? [TBD]

- The skip-gram model and negative sampling

  paper: “Distributed Representations of Words and Phrases and their Compositionality” (Mikolov et al. 2013)

- A SIMPLE BUT TOUGH-TO-BEAT BASELINE FOR SENTENCE EMBEDDINGS (Sanjeev Arora, Yingyu Liang, Tengyu M
- Glove: https://nlp.stanford.edu/projects/glove/

Model for R? [TBD]

- As in Skip-Gram, the relationship is captured in word vector

More relationship are learned by Word Embedding

- The use of word representations… has become a key “secret sauce” for the success of many NLP systems in recent years, across tasks including named entity recognition, part-of-speech tagging, parsing, and semantic role labeling. (Luong et al. (2013)
- This general tactic – **learning a good representation on a task A and then using it on a task B** – is one of the major tricks in the Deep Learning toolbox.

  Pre-trained word vector ?

Word sense topic definition and thoughts

- P(sense of the word W|context word, ambiguity word W)

Related topics definition and thoughts

- In Natural Language, we learn the relationship of words in sentences. We can use the similar technique to learn relationships of entities (pages in Wikipedia). And predict/rank the related entities. (unsupervised learning?)
- Using vectors to represent entities
- 1). Preprocess huge data to identify entities (exception handling)
- 2). Find relationships
- 3). Model relationships
- Unsupervised ? Evaluation?