#ODSC

@ODSC

OPEN DATA SCIENCE CONFERENCE

Boston | May 1 - 4 2018

@ODSC

word2vec for word & phrase translation

Outline: *word2vec* for word & phrase translation

- Word Vectors: Goals & Applications
- word2vec: Architecture & Refinements
- Implementation: keras, TensorFlow, gensim & command line
- word2vec for translation
  - From mono-lingual to bilingual word spaces
  - Learning a translation matrix using TensorFlow

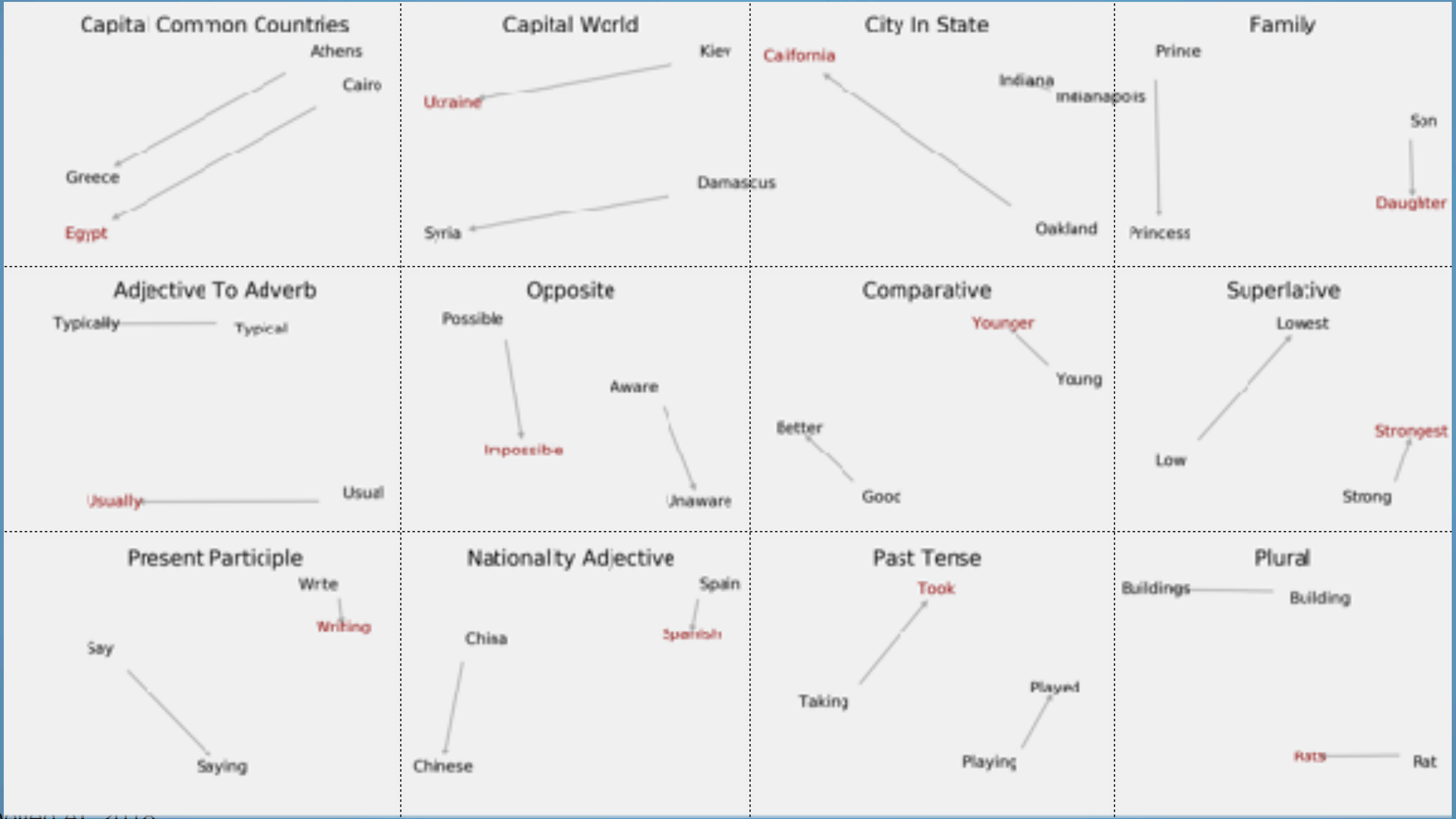
Motivation: *word2vec*

- Simple models + lots of data >> complex models + less data
  - e.g. *n-gram* models for statistical language modeling
- BUT
  - in-domain data for speech recognition is limited
  - Corpora for many languages have only a few billions words
- Complex model + lots of data >> simple model - if you can train it
- Mikolov et al (2013): architecture to scale word vector learning

Goals & Applications

- Embed words in continuous vector space to better encode text
- Distributional hypothesis: similar distribution <> similar meaning
- Word vectors capture semantic meaning
    1. Similar words will be close to each other
    2. Words have multiple degrees of similarity

word2vec Evaluation based on Analogies



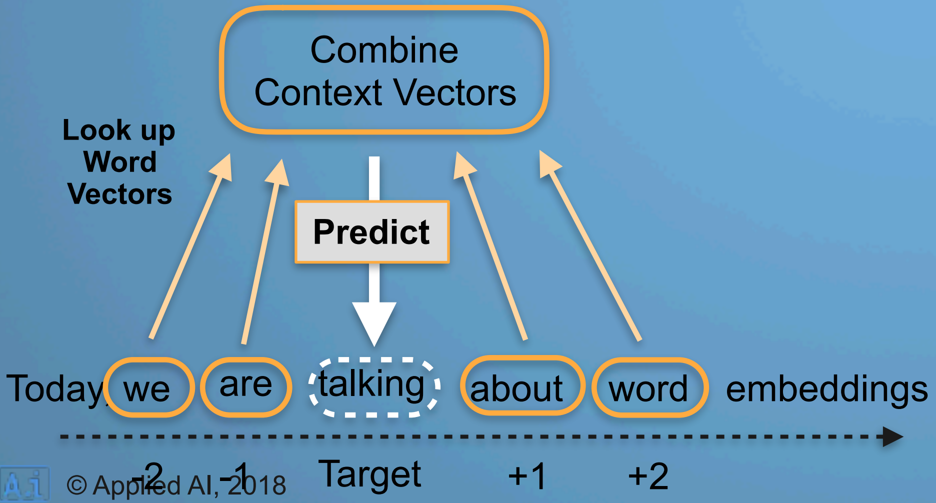
2D Projection of 300D vectors(using Incremental PCA)

Learn Word Embeddings by Relating Words to Context

- word2vec comes in two predictive flavors

Continuous Bag of Words:

Context => Word



Combine Context Vectors

Look up Word Vectors

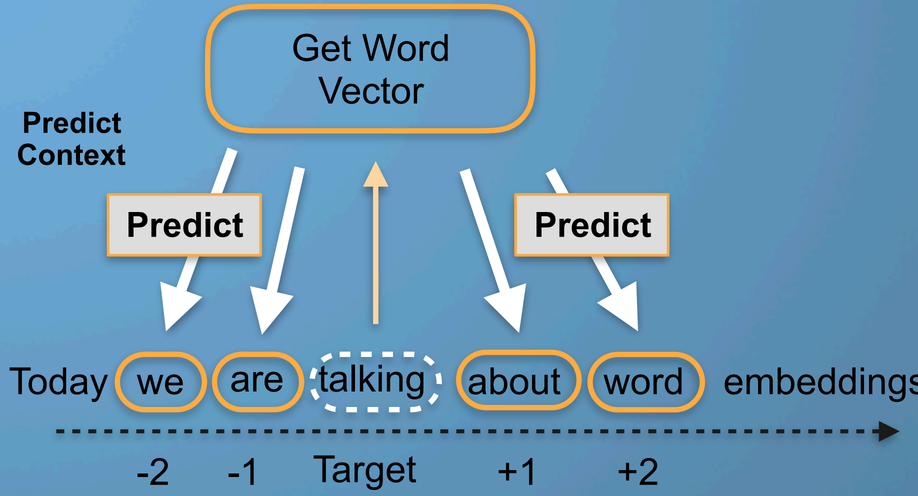
Predict

Today we are talking about word embeddings

-2 -1 Target +1 +2

Skip-Gram:

Word => Context



Get Word Vector

Predict Context

Predict

Predict

Today we are talking about word embeddings

-2 -1 Target +1 +2

From Softmax to Noise Contrastive Estimation

- Neural probabilistic language models maximize the likelihood using the (expensive) softmax objective
- word2vec: binary classification of the true word vs k random ‘noise’ words
- Scales with the number of noise words, not with the vocabulary
- Approximates the softmax result in the limit

Preprocessing

- Input data in text form.
  - Detect sentence boundaries
  - Tokenize
  - Remove punctuation
  - Create n-grams
- We’ll use TED 2013 text that is already sentence-aligned

The material for today’s workshop

- Presentation, Data & Notebooks:
    - https://github.com/stefan-jansen/word2vec-translation