Word Embeddings for Political Science Text Analysis

What are Word Embeddings?

D1: The cat jumps over the dog
D2: The kitten hops ovqer the hound
D3: Die Katze springt über den Hund

	the	cat	jump	over	dog	kitten	ovqer	hound	die	katze	springt	über	den	hund
D1	2	1	1	1	1	0	0	0	0	0	0	0	0	0
D2	2	0	0	0	0	1	1	1	0	0	0	0	0	0
D3	0	0	0	0	0	0	0	0	1	1	1	1	1	1

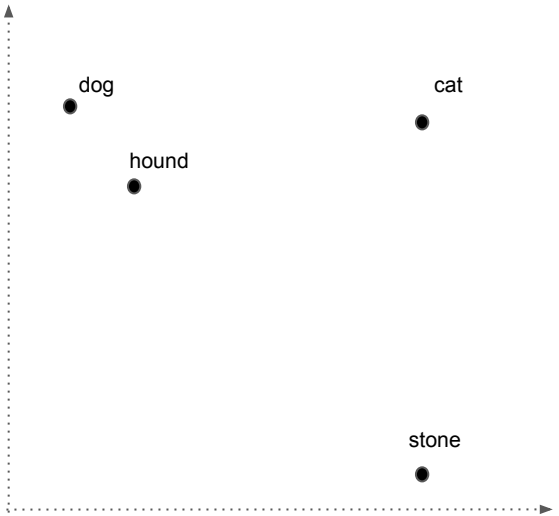
Word Representations

- Each word is represented by its unique character sequence
- Spatial: dist(dog, hound) = dist(dog, cat) = dist(dog, stone) = ?

Word Representations

- Each word is represented by its unique character sequence
- Spatial: dist(dog, hound) = dist(dog, cat) = dist(dog, stone)
- Embeddings give each word a location in a space:

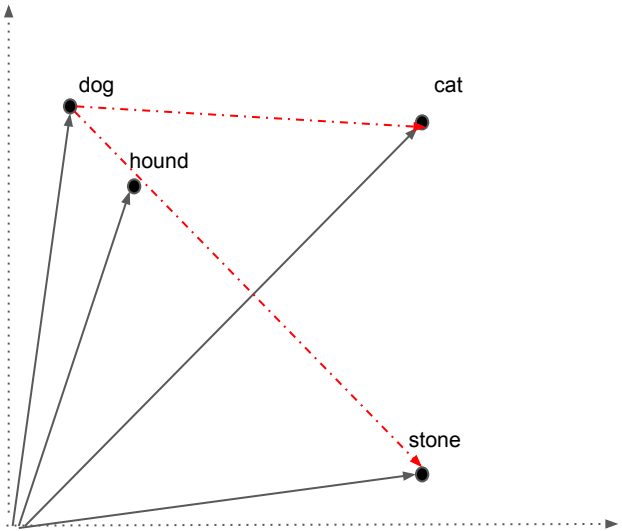
dog = [x, y]
hound = [z, w]
....



Word Representations

- Each word is represented by its unique character sequence
- Spatial: dist(dog, hound) = dist(dog, cat) = dist(dog, stone)
- Embeddings give each word a location in a space:

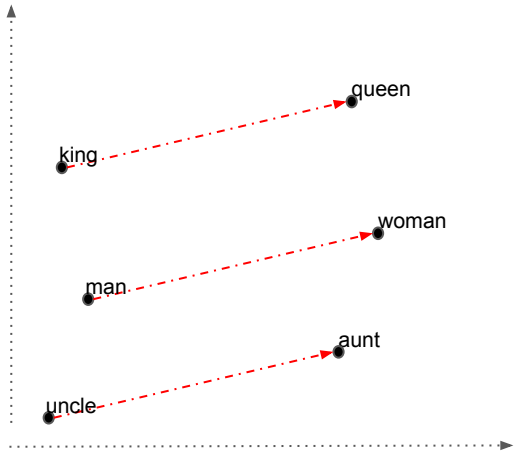
dog = [x, y]
hound = [z, w]
....



||dog - cat|| < ||dog - stone||

How to get them

- “A word is characterized by the company it keeps” (Firth 1957)
  - Factor analysis (documents)
  - Topic Models (documents)
  - Neural Networks (sliding window)
  - Other Dimension reduction techniques

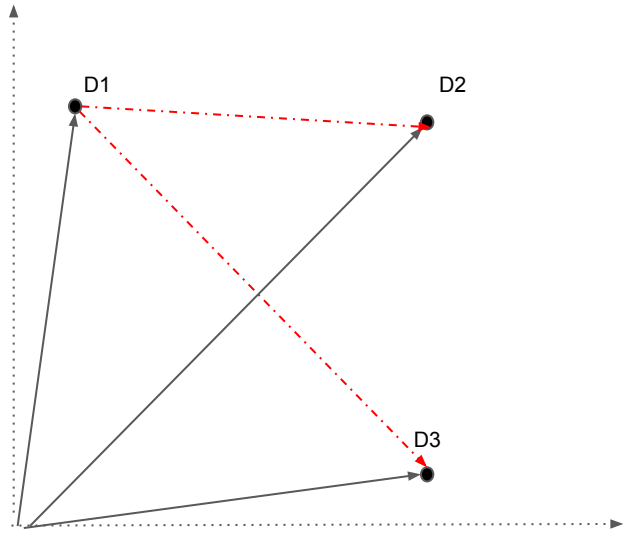


Shallow neural networks allow for training on gigantic corpora. E.g. all of Google News or all of Wikipedia -> Good representation of ‘general meaning’ of words (Mikolov et al. 2013).

Document Embeddings

D1: The cat jumped over the dog
D2: The kitten hopped across the hound
D3: Die Katze springt über den Hund

	x	y	...
D1	1.34	5	...
D2	4.34	4.5	...
D3	4.2	1.6	...



Political Science Text Analysis

What we could use it for right away:

- Short documents (open ended survey answers, twitter posts, sentences, etc.)
- Transfer trained models between languages (e.g. comparative: use models trained on rich american politics data for other countries)
- Improve text sequence alignments (e.g. Smith Waterman Algorithm)

What is lacking:

- Statistical modeling of location in semantic spaces (e.g. identify significant change in location, model trajectories of words and documents, etc.)

Dissertation Outline

1. Introducing Word Embeddings
    a. Introduction
    b. Innovations/Addons
        i. (Visual) Interpretation (unsupervised)
        ii. Machine learning interpretation (supervised)
2. Political Science Application Application
    a. Short document classification (CAPR Poll, Twitter, Comparative Manifesto Data)
    b. Automatic text processing in multiple languages (Comparative Manifesto Project, Congressional Bill Project)
3. Methodological Development
    a. Spatial statistics of semantic spaces
        i. Detect significant spatial differences between documents
        ii. Detect movement of words