Detecting Bias in Word Embedding Models

Professor Ameet Soni
Professor Krista Thomason
Swarthmore College

Lecture notes to introduce lab practicum

Background - NLP

Natural Language Processing (NLP)

- Goal: help computers understand human language
- Artificial Intelligence + Linguistics

General tasks:

- Generation language/communicate with humans
- Analyze/understand humans

Example NLP tasks

- Translation between two languages
- Extract information from text (“Find all relevant Title VII case law”)
- Sentiment analysis (“Is this review positive or negative?”)
- Personal Assistant
- Search
- Summarization
- Etc.

Difficulties

- Goal: map meaning to words
- Problem: this is very hard; language is
  - Ambiguous – same word can mean different concepts
  - Rich – the same concept can be said with many words
  - “Meaning” is never observed directly
- Representation: how should computers encode words?

Example

"John saw the woman with the telescope wrapped in paper"

- What's wrapped in paper?
- Who has the telescope?

Word Embeddings

- Words are not discrete concepts

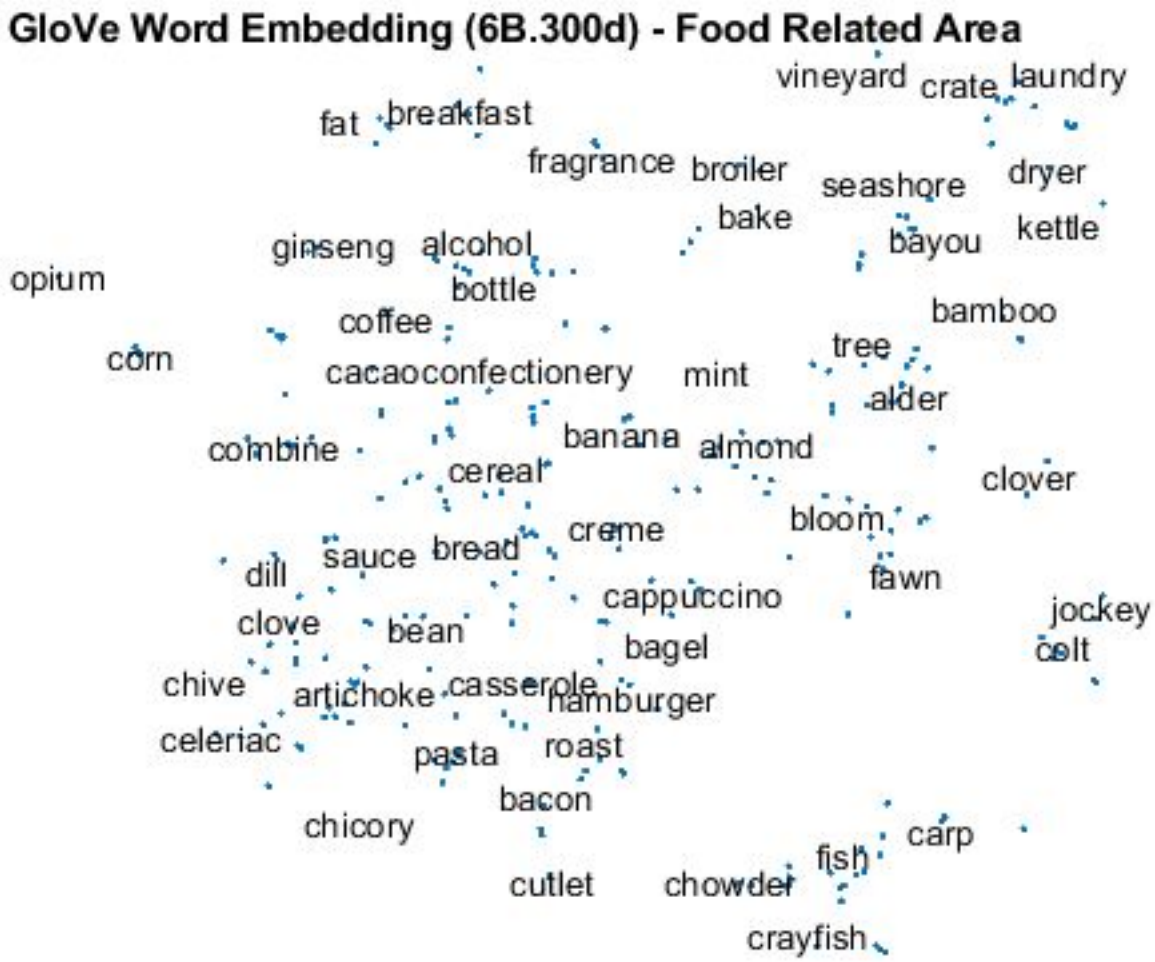
e.g., search “Seattle hotel” vs “Seattle motel” should be similar

- Idea: words with similar meaning occur in similar contexts
  - “hotel” and “motel” are used similarly in sentences
- Word embeddings: represented words by what they *co-occur* with
  - “book”, “room”, “rate” are commonly used with “hotel” and “motel”

Fast Forward

- Add theory, math, advances in computing, lots of data…

Embeddings Visualized

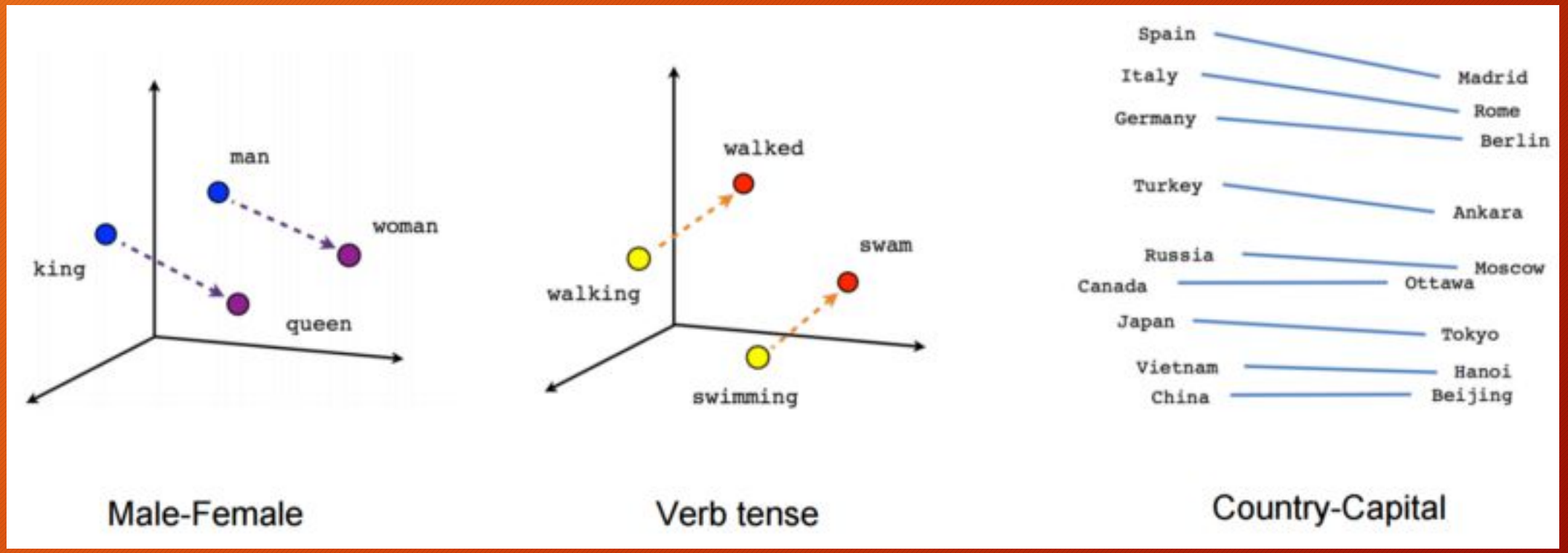


Similar items cluster together

Note: these are 2D projections of the original embeddings using principal component analysis (PCA)

Matlab.com

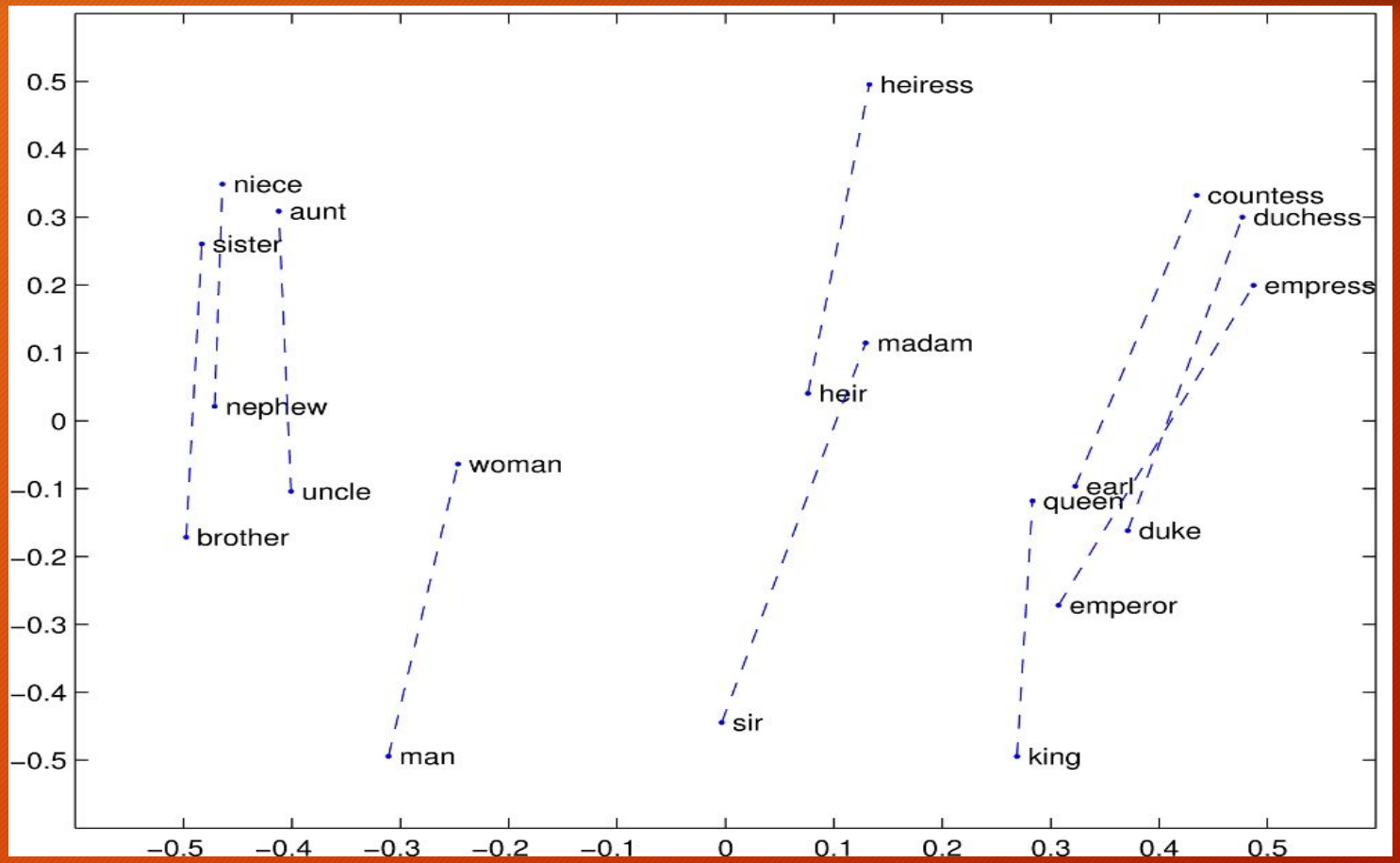
Learning Semantics



https://www.tensorflow.org/tutorials/representation/word2vec

Note: these are 2D or 3D projections of the original embeddings using principal component analysis (PCA)

Learning Relationships



"Vector differences capture as much as possible the meaning specified by the juxtaposition of two words"

https://nlp.stanford.edu/projects/glove/images/man_woman.jpg

Word Embedding Today

- Many approaches
- Many success stories

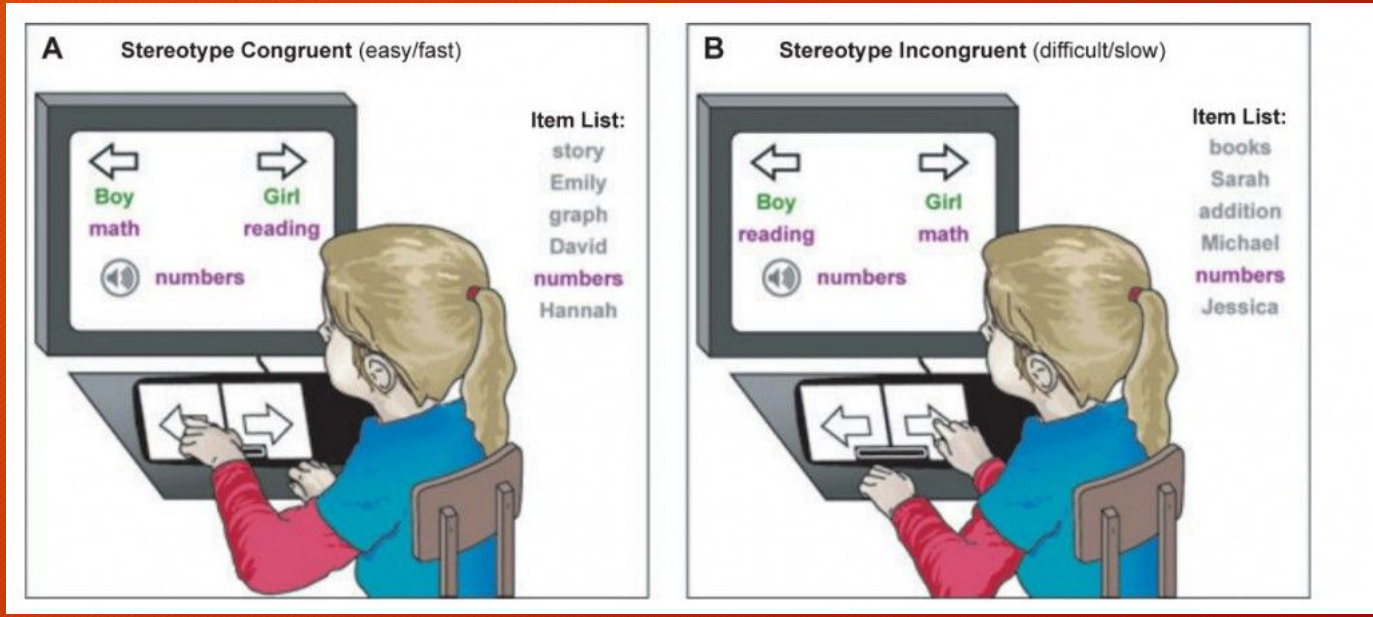
- Data driven – learned through provided text e.g., Wikipedia
- Dense and distributed representations

Lab Practicum Part 1: Learning Embeddings

- Given: learned word embeddings
  - GloVe algorithm https://nlp.stanford.edu/projects/glove/
  - Training corpus: twitter, wikipedia, web
- Goal: validate the usefulness of word embeddings
- Read handout and README.md to run findSimilarWords.py

Lab Practicum Part 2: Word Embedding Association Tests

- From Caliskan et al., “Semantics derived automatically from language corpora contain human-like biases”
- Modeled after Implicit Association Test
- Read documentation and run weatTest.py
- Complete Lab Practicum Assignment



Real-World Example:
Gender Bias in Google Translate

In Turkish, *o* is a gender neutral pronoun (*he*, *she*, or *it*)

Turkish - detected

o bir aşçı
o bir mühendis
o bir doktor
o bir hemşire
o bir temizlikçi
o bir polis
o bir asker
o bir öğretmen

Real-World Example: Gender Bias in Google Translate

In Turkish, *o* is a gender neutral pronoun (*he*, *she*, or *it*)

Turkish - detected	English
o bir aşçı	she is a cook
o bir mühendis	he is an engineer
o bir doktor	he is a doctor
o bir hemşire	she is a nurse
o bir temizlikçi	he is a cleaner
o bir polis	He-she is a police
o bir asker	he is a soldier
o bir öğretmen	She's a teacher