Lecture 2: Word Embeddings in Social Science

Miriam Hurtado Bodell

LINKÖPING UNIVERSITY

Recap word embeddings

Word embedding models estimate numerical dense vector representation of words

$$\log p(\mathcal{D} \mid w, c) = \sum_{i=1}^{N} \left(\underbrace{\sum_{y \in C_i} \log \sigma(w_{x_i}^T c_y)}_{\text{positive samples}} + \underbrace{\sum_{y \in C_{ns}} \log(1 - \sigma(w_{x_i}^T c_y))}_{\text{negative samples}} \right)$$

where $\mathcal{D} = (x_1, \ldots, x_N)$ is the dataset indexed by $i \in \{1, \ldots, N\}$, C_i is the context window at i and C_{ns} is a randomized context window generated from the empirical distribution of words.

$\rightarrow$ unsupervised method

Word embeddings in the social sciences

What does embedding vectors represent?

- Technical answer: A word's position in a vector space
- Sociological answer: ?

Word embeddings in the social sciences

What does embedding vectors represent?

- Technical answer: A word's position in a vector space
- Sociological answer: The meaning of a word

Meaning of a word/concept/person is not inherent but emerge in relation to other words/concepts/people → maps to the foundations of word embeddings

Word embeddings in the social sciences

Usually not the explicit meaning of the word itself that is (sociological) interesting – but can be used a **sensor** of:

- Political ideology

  "Protecting our borders and upholding the Second Amendment are essential to preserving our American values and ensuring the safety and security of our citizens."

  "Ensuring that every American has access to affordable healthcare and quality education is fundamental to our nation's progress and prosperity"

Word embeddings in the social sciences

Usually not the meaning of the word itself that is (sociological) interesting – but can be used a **sensor** of:

- Political ideology
- Societal change (gay 1920 vs. 2020)
- Technical change (apple 1920 vs. 2020)

$\sim$ *culture*

Word embeddings in the social sciences

When does it make sense to use word embedding models?

- When theoretical construct of interest can be conceptualized as the relation between words or captured via word's meaning
- When polysemy is not important
- Corpus is a meaningful unit of analysis (NB perhaps not true!)

Pre-trained word embedding models

Many rely on pre-trained word embedding model (available in Python/R)
$\rightarrow$ the embedding reflects the meaning of words in training data

You must ask yourself: what is the overlap between inference target and training data?

- Can you make claims about different historical periods?
- Cultural differences between people who write online and classical literature?

Good when you might not have a lot of your own data!

Word embeddings in the social sciences

How can we use word embedding models?

Discovery

- Word embeddings estimates word vectors without input from researcher → can be used to explore cultural associations

Measurement

- Cosine similarity between word vectors = measure how semantically similar two words are
- Vector algebra create relevant measures of concepts (e.g. class = rich - poor)
- Input for classification tasks (fine-tuning, more next lecture)

Example **measurement**: Kozlowski et al. (2019)

Aim: Study how the markers of class has shifted over time

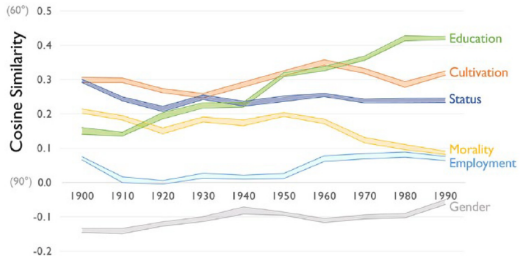
Cultural meaning = relation between word vectors

Method:

1. Manually created affluence, employment, status, education, and cultivation, gender and morality dimensions per decade using word vectors from 10 models trained on Google Ngrams corpus (5grams) by decade
2. **Validate** using survey data
3. Describe projections between dimensions using cosine similarity

Example **measurement**: Kozlowski et al. (2019)

Results: **Describe** how affluence and other dimensions of class relate at different time points – class a relatively stable concept over the last 100 years (*at least in printed books*)



Example **measurement/inference**: Best & Arseniev-Koehler (2023)

Aim: Answer the question of why are some diseases more stigmatized than others? And test theories of why stigma shifts over time.

Method:

1. Create two stigma dimensions using vector algebra: *judgement* (mean of immorality + negative traits) and *disgust*
2. **Validate**: (i) does "known" words place correctly, (ii) top words in dimension, (iii) compare with expert survey
3. Run regression with stigma dimensions as the dependent variable

Example **measurement/inference**: Best & Arseniev-Koehler (2023)

Results:

- Behavioral health conditions (+ STDs) attracts judgement than other health conditions, preventable deceases are more stigmatized
- Infections deceases more associated with disgust than other illnesses
- Medicalization increases stigma (opposite to the hypohesis) while advocacy lower stigma levels
- The stigma of chronic illnesses has decreased over time

→ *Nice example of how to move beyond purely descriptive analysis*!

Validation

Typically, validation consists of (combination);

- Comparing found associations with survey data (conduct new survey or use old survey)
- Explore the vector space; do you find "known associations"
- How robust are created dimensions to specification (e.g. words used to create them)

Robustness: run multiple models and take the average over model runs (uncertainty estimates)

Limitations of word embedding models in social science

Validation is not straight forward – especially when dealing with historical corpora where comparison to human judgement is not possible

Potentially riddled with (unknown) **bias**, especially problematic when using pre-trained embeddings

How do move beyond measures related to **binary** dimensions?

Cannot deal with polysemy

Extensions of word embedding models in social science

Dynamic word embeddings: can compare how words evolve over time

Example studies above train separate models per time slice (perhaps not comparable), other align different vectors post hoc

Learning between time slices help model performance (non-probabilistic, Bamler & Mandt 2017, Rodman 2020; probabilistic, Rudolph & Blei 2018)

Interpretable dynamic word embeddings (Hurtado Bodell et al. 2019); follow binary dimension over time

Dynamic word embeddings + issue with polysemy: example from my own research

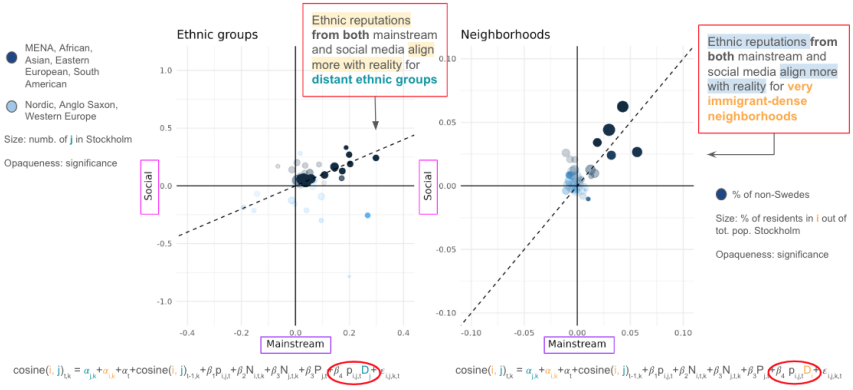
RQs: How well does the ethnic reputation of neighborhoods and ethnic realities align? How does it differ between mainstream and social media?

Ethnic reputation = the imagined association of particular places with particular ethnic groups

I don't want to live in Rinkeby, it is dangerous
One really beautiful place is Djursholm, I would love it there!
I don't want to live with Somalis, they are dangerous

$cosine(Rinkeby, Somlis) = 0.9$ vs. actual proportions of Somalis in Rinkeby at year t

Dynamic word embeddings + issue with polysemy: example from my own research



$$\text{cosine}(i,j)_{i,k} = \alpha_{j,k} + \alpha_{i,k} + \alpha_t + \text{cosine}(i,j)_{t-1,k} + \beta_1 p_{i,j,t} + \beta_2 N_{i,j,k} + \beta_3 N_{j,i,k} + \beta_4 P_{j,i,t} + \beta_5 p_{i,j,t} D + \varepsilon_{i,j,k,t}$$

$$\text{cosine}(i,j)_{i,k} = \alpha_{j,k} + \alpha_{i,k} + \alpha_t + \text{cosine}(i,j)_{t-1,k} + \beta_1 p_{i,j,t} + \beta_2 N_{i,j,k} + \beta_3 N_{j,i,k} + \beta_4 P_{j,i,t} + \beta_5 p_{i,j,t} D + \varepsilon_{i,j,k,t}$$

Extensions of word embedding models in social science

Embedding regression (Rodriguez et al., 2023): how word meanings differ between document-level properties

Example RQs:

- Do men and women use the word "relationship" differently? How do they differ?
- Did the meaning of "refugee" shift after the "Refugee crisis" 2015?

Seek to find the difference between pre-trained word vectors dependent on context (on Monday: decoders)

Alternative: group-based embeddings (Rudolph et al. 2017)

Summary

- Social scientists use estimated word vectors from word embedding models to study shifts in word meaning as a sensor of a shift in cultural meaning
- So far only used to study shifts in public culture (i.e. macro-level feature not micro-level/meso-level features)
- Been criticized for being difficulty to validate and poor performance on polysemic words

Thank you!

www.liu.se

LINKÖPING UNIVERSITY