Text as Data

Session 10: Embedding Regression

Mirko Wegemann

08 July 2026

Our meeting today

- Presentation by Beyza on 'supervised methods'
- From bags-of-words to a more complex representation of text
- We use embeddings to understand the context in which certain terms are used
- Using embedding regression, we can see how feature usage can change depending on the context

Why bags-of-words approaches aren't always enough...

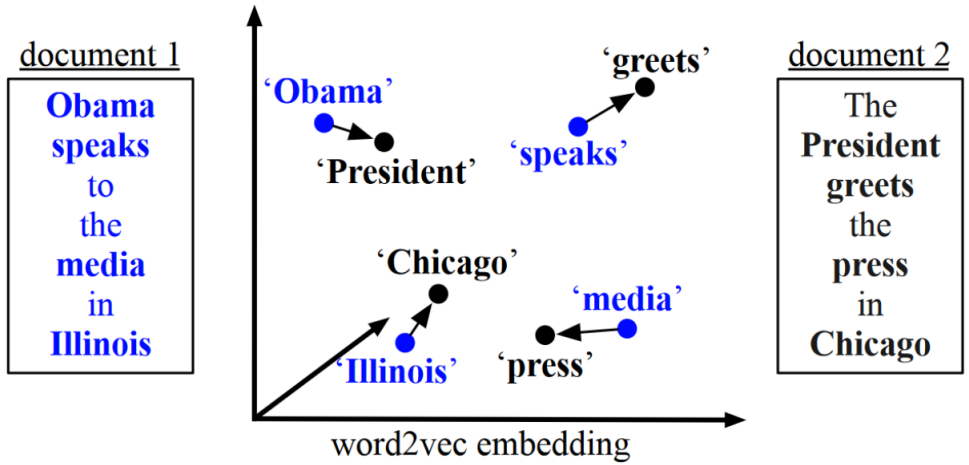
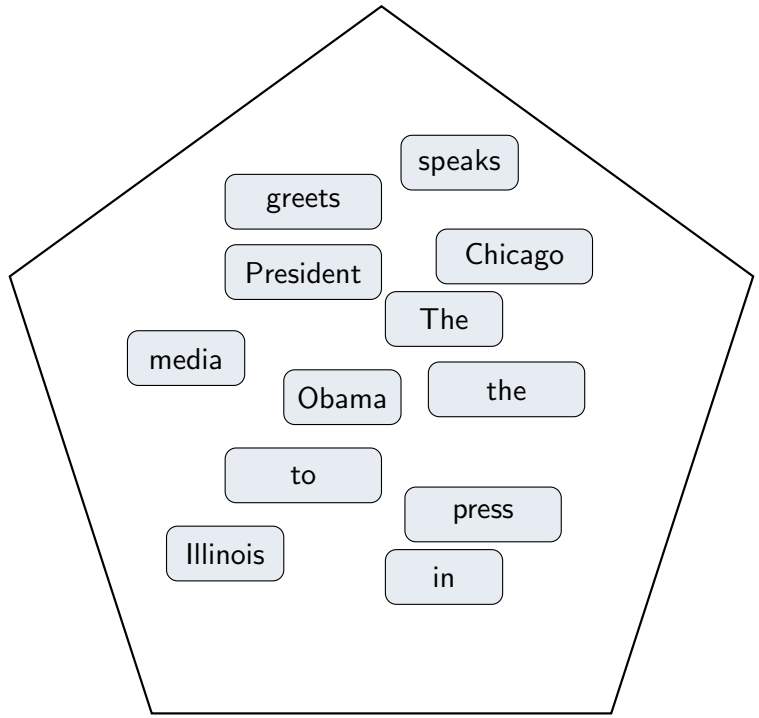


Figure 1. An illustration of the *word mover's distance*. All non-stop words (***bold***) of both documents are embedded into a *word2vec* space. The distance between the two documents is the minimum cumulative distance that all words in document 1 need to travel to exactly match document 2. (Best viewed in color.)

Kusner et al. (2015)

What bags-of-words approaches would do... I



What bags-of-words approaches would do... II

```
> dfm_example
Document-feature matrix of: 2 documents, 12 features (37.50% sparse) and 0 docvars.
       features
docs    obama speaks to the media in illinois . president greets press chicago
  text1     1      1  1   1     1  1        1 1         0      0     0       0
  text2     0      0  0   2     0  1        0 1         1      1     1       1
```

This is what our text looks like in a bags-of-words structure

What could be problematic here?

What bags-of-words approaches would do... III

The biggest disadvantage of bags-of-words approaches is that they do not capture the (1) **relation of a word to other words** or the (2) **position of a word** within a sentence.

Bags-of-words are context-blind.

What bags-of-words approaches would do... IV

We can distinguish between two types of embeddings that capture context differently:

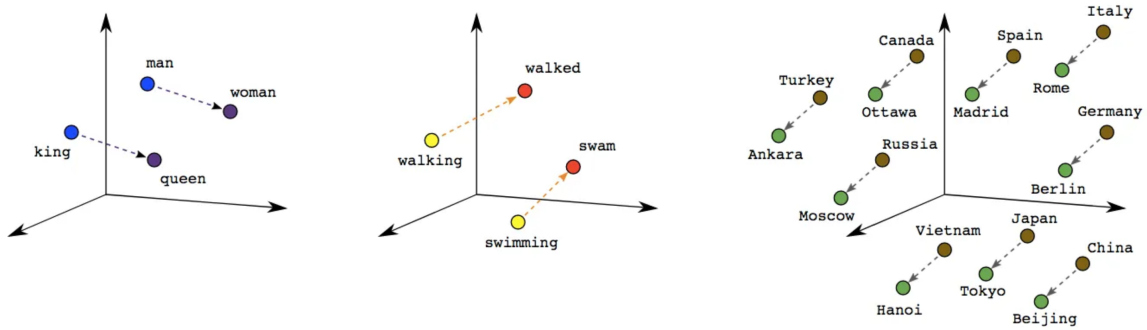
1. static embeddings, such as GloVe (Pennington et al. 2014)
2. contextual embeddings, such as BERT (Peters et al. 2018)

The idea behind Word Embeddings I

- Idea: Detecting Similarity of Words
- we can assign a **multidimensional vector** to a word that represents its relationship to other words
- "distances between such vectors are informative about the **semantic similarity** of the underlying concepts they connote for the corpus on which they were built" (P. Rodriguez and Spirling 2021)
- Word Embeddings "predict the occurrence of a word by the surrounding word in a text sequence" (Rheault and Cochrane 2020, p. 112)

The idea behind Word Embeddings II

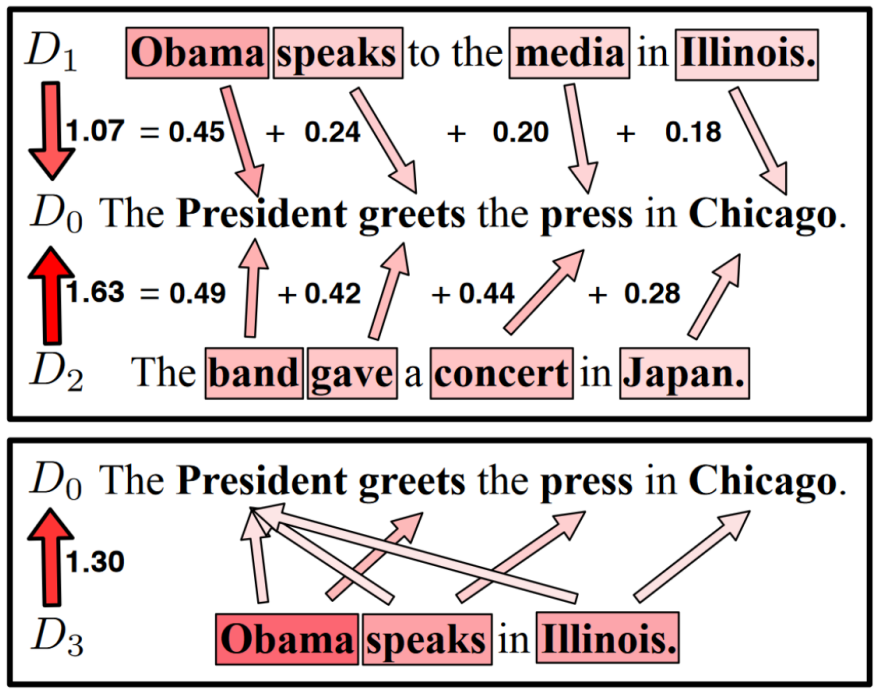
Embeddings transform features into a k-dimensional space (here simplified into three dimensions).



Male-Female Verb Tense Country-Capital

1

The idea behind Word Embeddings III



Kusner et al. (2015)

[1] Source: https://towardsdatascience.com/a-guide-to-word-embeddings-8a23817ab60f

What embeddings can be useful for...

P. Rodriguez and Spirling (2021) distinguish between two functions:

1. embeddings have an **intrinsic** value (word-level analysis)
   - , for example: if the distance between the term 'migrant' and 'hardworking' is closer for Greens than for conservatives, we could derive patterns of party communication from it
2. ...and an **instrumental** value: they capture more meaning than bags-of-words $\rightarrow$ They could improve our *classification tasks*

Embedding Usage Decisions I

- **word2vec** vs. **GloVe** (difference in training embeddings)?
- **pre-trained** vs. **locally-trained** Embeddings?
- **Context window** (window size): How much context is taken into account?
- **dimensionality**: How many dimensions do we consider?

Embedding Usage Decisions II

Pre-trained vs. local

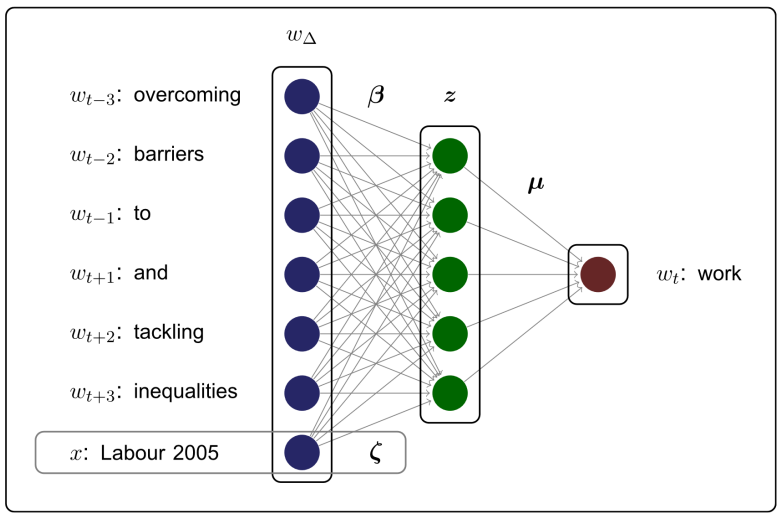
- Representation of a feature is usually learned through deep learning approaches (such as neural networks)
- We can either create corpus-specific embeddings that learn the domain-specific relationships
- ... or use pre-trained embeddings (which have been pre-trained using a large corpus such as Wikipedia)
- Decision depends on body size and specificity

Embedding Usage Decisions III

Context window

- The larger the context window (*window size*), the more distant words affect the embedding vector

We predict a word w_t through the surrounding words w_{t-1}, w_{t-2}, ..., w_{t-n}; $\rightarrow$ n is the context window



Rheault and Cochrane (2020, p. 116)

Embedding Usage Decisions IV

Dimensionality

- The number of dimensions can be freely selected
  - The term *dimensions* refers to the length of the vector representation of a feature
  - The more dimensions, the more information we capture, but the more computationally intensive it becomes (and the more likely models are to overfit)

Embedding Usage Decisions V

In practice, the choice of these *hyperparameters* has only marginal effects – pre-trained models work well (P. Rodriguez and Spirling 2021)

Using Word Embeddings I

There are three different methods to get embeddings

1. Training your own "local" embeddings with a neural network
2. Use of pre-trained embeddings (e.g. for GloVe or other multilingual embeddings)
3. fine-tuning of pre-trained embeddings

Using Word Embeddings II

Depending on the method, different steps are required. For **pre-trained embeddings**:

- Data Preparation
- Match between Input and Embeddings
- Data Analysis

For **locally trained embeddings**:

- Data Preparation
- Neural network to learn the embeddings
- Data Analysis

We're focusing on pre-trained embeddings here, but there's code at the end of the script for training your own embeddings.

Preparation in R

We don't necessarily have to do pre-processing, but we often take the following steps:

- Removing common words (or stop words) without much semantic meaning can improve our embeddings
- Removing punctuation and digits
- Creation of n-grams (e.g. European_Union) or lemmatization can also facilitate interpretation

... and now in `R`

Descriptive Analysis I

nearest neighbors (which words are close to each other?) An often used example (see before) is the equation

$$Berlin = Paris - France + Germany \tag{1}$$

Not only geographically, but also semantically, the distance between the word Berlin and Germany should be equal to that between Paris and France.

What is the capital of Germany?

We can translate this equation into R.

```
> which_capital = embeddings["paris", , drop = FALSE] -
+ embeddings["france", , drop = FALSE] +
+ embeddings["germany", , drop = FALSE]
> capital_cos_sim = sim2(x = embeddings, y = which_capital, method = "cosine", norm = "L2")
> head(sort(capital_cos_sim[,1], decreasing = TRUE), 5)
Berlin Germany Frankfurt Hamburg Paris
0.7635345 0.7232592 0.6718858 0.6530555 0.6509711
```

More descriptive analyses

Terms near "migrant"

```
> find_nns(embeddings['migrant',], pre_trained =
    embeddings, N = 20)
[1] "migrant", "immigrant", "refugee", "worker",
    "undocumented", "farmworker", "indigenous"
[8] "unskilled" "migration" "expatriate"
    "immigration" "migratory" "resettlement" "labor"
[15] "employment", "unemployed", "population",
    "plight", "welfare", "labour"
```

Descriptive Analyses II

We can also use the `conText` package to look at similarities of features depending on the covariates (which words are used by which actors in the context of word *w* used?)

```
context_wv_parfam <- dem_group(context_dem, groups
    = context_dem@docvars$parfam)
dim(context_wv_parfam)

context_nns <- nns(context_wv_parfam, pre_trained =
    embeddings, N = 10, candidates =
    context_wv_parfam@features, as_list = TRUE)
context_nns
```

... and now in `R`

Embedding Regression I

- **contextual use** of a term (we are getting closer to contextual embeddings)
- for this we use covariates (as before in the descriptive analysis)
- we **convey an embedding** in different contexts and compare how similar it is (before we weight words that occur frequently in both contexts as less relevant)

P. L. Rodriguez et al. (2023) x

Embedding Regression in R

conText makes it possible to estimate similarities by groups for each target term and also provides standard errors

```
set.seed(451)
model2 <- conText(formula = trump ~ prepost,
data = toks,
pre_trained = embeddings,
transform = TRUE, transform_matrix = trans_mat,
bootstrap = TRUE,
num_bootstraps = 100,
permute = TRUE, num_permutations = 10,
window = 10,
verbose = T)
```

... and now in `R`

Evaluation of embeddings I

- for downstream tasks (e.g., classification with embeddings), we can use conventional machine learning metrics (see session 10: F1 score, accuracy, confusion matrix, etc.)
- more difficult for intrinsic tasks: What makes an embedding particularly good?

Evaluation of embeddings II

P. Rodriguez and Spirling (2021) propose 'Turing validation'

- Computers perform well when they are indistinguishable from human tasks
- in practice: People compare a list of word associations created by humans with a list of embeddings

$\rightarrow$ if people can't distinguish both lists, the performance is good

- other validations include correlation between models
- at *scaling tasks* like that of Rheault and Cochrane (2020): cross validation with other measures to party positions

What we learned today...

- ...what embeddings are and what distinguishes them from bags-of-words approaches
- ...how we can use them for our research
- ...how embedding's regression helps us capture additional context in the use of words

Next week...

- ...we have our last substantive meeting
- ...we deal with the instrumental purpose of word embeddings and build a neural network for classifying text
- ...if you want to follow the session, it is recommended to install Python and run through the preamble of the R script before the session

Literature I

Kusner, M. J., Sun, Y., Kolkin, N. I., & Weinberger, K. Q. (2015).From Word Embeddings To Document Distances.

Pennington, J., Socher, R., & Manning, C. D. (2014).Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. https://arxiv.org/abs/1802.05365

Rheault, L., & Cochrane, C. (2020).Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora. *Political Analysis*, *28*(1), 112–133. https://doi.org/10.1017/pan.2019.26

Literature II

Rodriguez, P., & Spirling, A. (2021).Word Embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics*. https://doi.org/10.1086/715162

Rodriguez, P. L., Spirling, A., & Stewart, B. M. (2023).Embedding Regression: Models for Context-Specific Description and Inference. *American Political Science Review*, *117*(4), 1255–1274. https://doi.org/10.1017/S0003055422001228