Partisan Representations: Measuring Linguistic Differences Using Embeddings

Pedro L. Rodríguez Sosa

@prodriguezsosa

Motivation

For **word embedding methods** (Mikolov et al., 2013a; Pennington et al., 2014) to gain traction in the social sciences there are a few open questions that must be addressed, including:

- Social scientists are interested in group differences:

1. How do we estimate/compare group-specific embeddings?

- Embedding models are usually trained on very large corpora:

2. Can we get reasonable results training on smaller corpora?

- Uncertainty measures are scarce in the embeddings literature:

3. How can we account for uncertainty in embeddings?

- Existing gold standard checks are specific to computer science:

4. How should we validate language differences (sub-models)?

In this project I take on these questions. I propose a general framework along with a series of novel methods and metrics.

Data

Corpus: Congressional Record 102nd-111th Congresses.
Key: use pre-processing of text to approximate ceteris paribus.

1. Pre-processing (kept to a minimum)**:**
- Remove non alpha-numeric characters.
- Lowercase.

2. Four Corpora:
- Split congressional corpus into 4 corpora.
- Corpora: Republican (R), Democrat (D), Female (F), Male (M).

3. Common Vocabulary:
- Intersect of top 10,000 for all four corpora.
- Vocabulary size: 8834 tokens.

4. Stratify:
- For party (gender) corpora guarantee same:
    - Female/male (republican/democrat) composition.
    - Proportion of corpus in vocabulary.

5. **Chunk & Fold:**
- Split corpora into equal-sized chunks (500 tokens each).
- Split each corpora's chunk set into 10 folds.

Parameters and Estimation

- Skip-Gram Wor2Vec model (Mikolov et al., 2013a).
- 10 fold CV for each group (40 sets of embeddings in total)
- Window: symmetric of size 6.
- Embedding size: 300
- Learning rate: 0.01
- Negative samples = positive samples.
- Epochs: 2 for Republican/Democrat corpus 3 for Female/Male.
- Loss: cross-entropy

Within Model vs. Cross Model

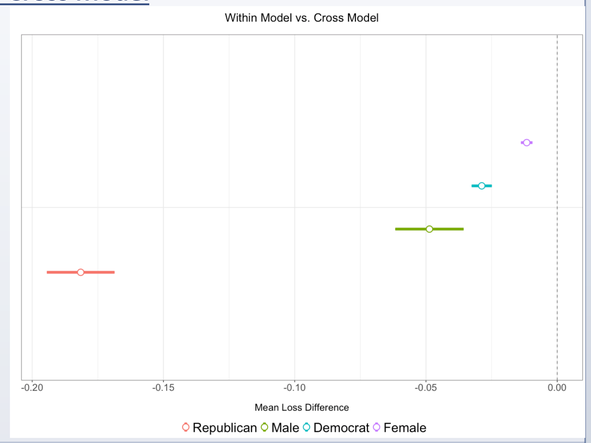
Methods

For each group (Republican, Democrat, Female, Male):

- Select model with the min loss on it's held-out fold (within-model).
- Evaluate opposite group's best model on the same fold (cross-model).
- Perform a means difference test of the two loss-histories.

Results

- All within-models significantly outperform cross-models.
- Within/cross performance differences vary by group.
- Difference is largest for Republicans (followed by Males).
- This suggests Republicans have particularly distinctive language.



Fighting Words

Data: MTurkers (self-identified as Republican/Democrat) provided lists of partisan words ("words likely to differ in meaning as a function of party"). From this set I select the top 10 most frequent.

Methods

For each group (Republican, Democrat, Female, Male):
- Compute it's distance matrix (using cosine distance).
- Estimate a significance threshold: lowest 1 percentile of distances.
- Identify set of significantly close neighbors for the 10 fighting words.

For each fighting word and each group comparison (F-M and R-D):
- Compute the **intersect-over-union** (IoU).
- Extract the set of non-overlapping tokens (Table 1 – Cols 2, 3 & 5, 6).
- Extract the set of overlapping tokens (Table 1 - Cols 1 & 4).

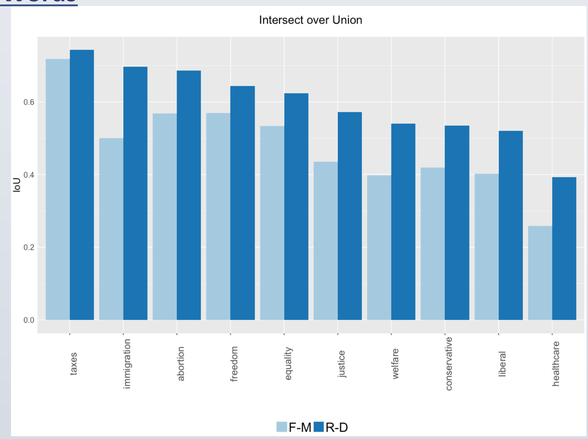


Table 1: Fighting Word: Healthcare

	Party			Gender	
Overlap	Republican	Democrat	Overlap	Female	Male
high-quality	physicians	improve	providers	nurses	physicians
uninsured	affordable	accessible	nurses	specialists	professionals
underserved	centers	provider	hospitals	underserved	affordable
quality	professionals	wellness	specialists	dental	care
providers	care		high-quality	nursing	patients

Human Validation

Lexical Decision Task

- Subjects are given a cue and 2 group-specific candidate words.
- They are asked to select the best "context word".

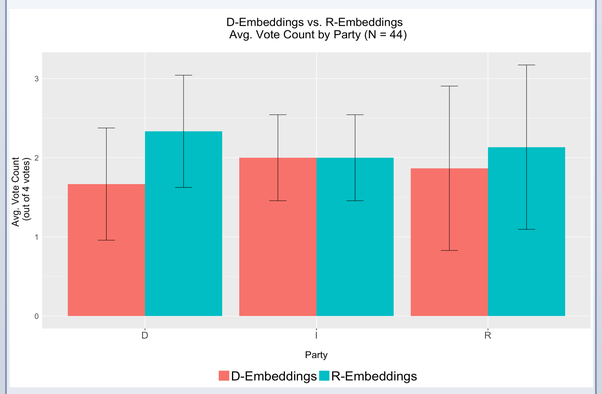
Hypothesis

- If language differences identified above are substantive then Republicans (Democrats) will select context words specific to Republicans (Democrats) with higher frequency.

Screenshot of LDT

Human Validation (results)

- Fighting word: "healthcare".



- Slight edge for Republican words for all groups.
- No significant differences with a sample of N = 44.

Main Takeaways

I'm sketching out a framework to:

- Estimate group differences in representations.
- Account for uncertainty in (a) differences (b) distance metrics.
- Validate differences using human input (gold standard).

Reasonable results are obtained with:

- Smaller corpora than is common in embeddings lit.
- No hyperparameter optimization.

Next Steps

- Extend loss-metric to specific words.
- Evaluate how local embeddings compare to global embeddings.
- Account for uncertainty in IoU metric.
- Hyperparameter optimization.

Literature (selection)

1. Mikolov, T., Chen, K., Corrado, G. & Dean, J (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

2. Data: Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts. Palo Alto, CA: Stanford Libraries [distributor], 2018-01-16. https://data.stanford.edu/congress_text.

Contact

- Code used in this poster is publicly available on my GitHub.
- To use LDT App and/or MTurk data contact me directly.
- **GitHub: @prodriguezsosa**
- Email: pedro.rodriguez@nyu.edu