An Embedding Model for Predicting Roll-Call Votes

Peter E. Kraft Hirsh Jain Alexander M. Rush
Harvard University, Cambridge, MA

HARVARD
John A. Paulson School of Engineering and Applied Sciences

Task

- Given a congressperson and the text of a bill, can we predict how that congressperson will vote on the bill?
- Provides method for quantifying relationships between congresspeople and bills, topics, and ideas.

Ideal Vectors

- Most previous work represented congresspeople as ideal points.
- Assumed all legislators and bills are single points in one-dimensional "political space."
- First prior attempt at prediction task made by Gerrish and Blei (2011).
- They developed *ideal point topic model*, integrates topic model similar to LDA for bill text with ideal point model for congresspeople. Used variational inference to approximate posterior distribution of topics.
- Our model represents legislators as *ideal vectors* in higher-dimensional space.
- Ideal vectors are easy-to-train, multidimensional representation of legislator ideology.

Acknowledgements

We would like to thank the Harvard Natural Language Processing Group for its support, as well as Sam Wiseman and Saketh Rama for useful guidance.

References

[1] Joshua D Clinton.
Using roll call estimates to test models of politics.
Annual Review of Political Science, 15:79–99, 2012.

[2] Sean Gerrish and David M Blei.
Predicting legislative roll calls from text.
In *Proceedings of the 28th international conference on machine learning (icml-11)*, pages 489–496, 2011.

[3] Jeffrey Pennington, Richard Socher, and Christopher D Manning.
Glove: Global vectors for word representation.
In *EMNLP*, volume 14, pages 1532–1543, 2014.

Model

- Simple bilinear model that uses low-dimensional embeddings to model each word in our dictionary and each congressperson.
- Models the probability of congressperson c voting "yes" on a bill containing words $\mathcal{B}$:

$$p(y = yea \,|\mathcal{B}, c) = \sigma((\mathbf{W}\left(\sum_{w\in\mathcal{B}} \mathbf{e}_w/|\mathcal{B}|\right) + \mathbf{b}) \cdot \mathbf{v}_c)$$

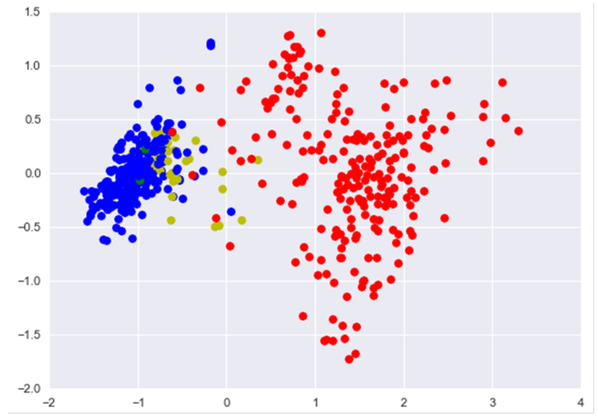
- Bills are represented with word embeddings ($\mathbf{e}_w \in \mathbb{R}^{d_{word}}$ for word w) to capture multivariate relationships between words and their meanings.
- Word embeddings are initialized with GLoVe.
- Congresspeople are represented by ideal vectors, $\mathbf{v}_c \in \mathbb{R}^{d_{emb}}$ for congressperson c (with $d_{emb} = 10$).
- We train $\mathbf{W} \in \mathbb{R}^{d_{emb} \times d_{word}}$ and bias $\mathbf{b} \in \mathbb{R}^{d_{emb}}$.

Data

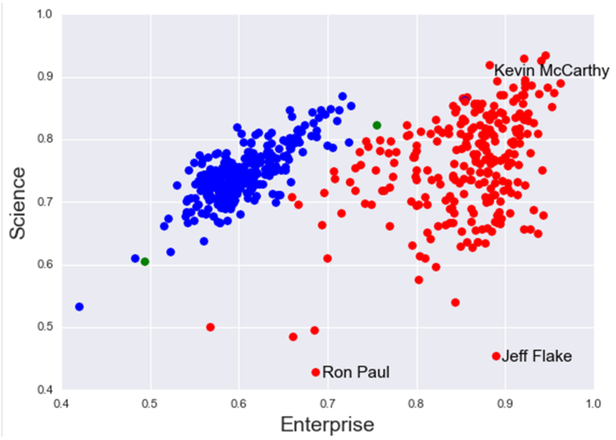
- Derived from the GovTrack database.
- Contains all votes on the full text (not amendments) of bills or resolutions.
- Uses data from the 106th to 111th Congress.
- Only contains yes-or-no votes, omitting abstentions and "present" votes (in accordance with previous work).
- Consists of 4067 bills and over a million unique yes-or-no votes.

Congress	# Bills	House	Senate	Pres
106	557	R	R	Clinton
107	505	R	D	Bush
108	607	R	R	Bush
109	579	R	R	Bush
110	854	D	D	Bush
111	965	D	D	Obama

Analysis



- PCA projection of ideal vectors for both houses of 111th Congress.
- Republicans red, Democrats who voted for Affordable Care Act (ACA) blue, Democrats who voted against ACA yellow, independents green.
- Model learns how (majority) Democrats much more unified than (minority) Republicans.
- Model also learns how conservative Democrats (who vote against ACA) closer to Republicans than other Democrats are.



- Relative favorability of congresspeople towards "Enterprise" versus "Science" in 110th Congress.
- Coordinates are sigmoids of dot products of ideal vectors with normalized word vectors.
- GOP red, Democrats blue, independents green.
- Model learns how both parties broadly support science, but Republicans are more pro-business.
- Model learns stances of individuals: Ron Paul (Libertarian), Kevin McCarthy (mainstream Republican), Jeff Flake (budget hawk).

Lexical Properties

Democrats	Republicans
economic	veterans
exchange	head
state	opportunities
carrying	provided
government	promote

- Top five words by cosine similarity for each party in the 110th Congress with stop words removed.
- Democratic words are mostly words of budget and government as Democrats were majority party
- Republican words mostly emphasize Republican themes and values

Final Results

Congress	YEA	GB	IDP	EMB
106	83.0	-	79.5	84.9
107	85.9	-	85.8	89.7
108	87.1	-	85.0	91.9
109	83.5	-	81.5	88.4
110	82.7	-	80.8	92.1
111	86.0	-	85.7	93.4
Avg	84.5	89	83.1	90.6

- Calculated accuracies on our model and baselines.
- YEA is a majority class baseline that assumes every legislator votes `yea`.
- GB is from Gerrish and Blei (2011)'s ideal point topic model.
- IDP is our model with $d_{emb} = 1$ to simulate a simple ideal point model.
- EMB is our model.

Conclusion

- We developed a novel model for predicting congressional roll-call votes from bill text.
- Our model outperforms any previous model while being extremely simple.
- We introduce ideal vectors as a fast, simple, multidimensional alternative to ideal points.

{pkraft@college,hirshjain@college,srush@seas}.harvard.edu