# Identifying Delegation in Congressional Bills

**Christian Fong***
Graduate School of Business
Stanford University
Stanford, CA 94305
christianfong@stanford.edu

## Abstract

Theories of delegation abound in political science, economics, and public administration, but there have been few empirical tests of these theories. One major roadblock is the lack of a comprehensive data set on acts of delegation. I take a necessary first step toward testing these theories by training a neural network to classify whether sections of bills are delegating public authority to the bureaucracy. Using a convolutional neural network architecture, I achieve excellent performance that vastly improves over the baseline.

## 1 Introduction

In the United States, government bureaucracies provide the services that are essential for economic growth and the maintenance of an orderly society. The United States Armed Forces defend the country from external threats, and the Federal Bureau of Investigation roots out internal threats. The Federal Reserve maintains economic stability. Agencies like the Food and Drug Administration, the Occupational Safety and Health Administration, and the Environmental Protection Agency promulgate and enforce regulatory standards to protect the public.

Yet there is a troubling tension in reliance on bureaucratic rule. On the one hand, there is a consensus in political science, economics, and public administration that autonomy is a key determinant of bureaucratic performance Wilson (1989). The bureaucrats who work in these agencies develop substantial expertise in their areas of specialization, and they are generally able to formulate more effective policies than their political superiors in Congress Gailmard and Patty (2012). However, these bureaucrats are not elected and often, due to civil service protections, cannot be fired for poor policy decisions. They are accountable to voters only insofar as Congress, which is composed of elected representatives, passes laws that specify what these bureaucrats can and cannot do. If Congress makes these laws too strict, then bureaucrats cannot bring their expertise to bear on pressing policy issues. If Congress makes these laws too lax, then bureaucrats may make policies that privilege their own predispositions over the wishes of the public Bawn (1995). Resolving how, if at all, Congress is able to find the happy middle ground is one of the most important challenges in political science.

Theories of Congressional delegation of public authority to federal agencies abound. However, empirical scrutiny of these theories has been hamstrung by the lack of a comprehensive data set on delegation Huber and Shipan (2000). This paper takes a first step toward remedying this lack of data by using a neural network to classify acts of delegation in Congressional bills. Using a convolutional neural network architecture, I am able to achieve performance that vastly exceeds the types of baseline models that would typically be employed in political science.

*https://people.stanford.edu/cjfong/

## 2 Data

My dataset is drawn from plain text versions of all bills passed by Congress between 1993 and 2015. This yields a corpus of 5,493 statues which collectively have over 30 million words. Apart from some front and end matter, these statutes are organized hierarchically into titles, subtitles, chapters, subchapters, parts, sections, subsections, subdivisions, paragraphs, subparagraphs, and sub-subparagraphs. Due to idiosyncracies arising from different types of bills, parsing this hierarchy is a non-trivial task. Additionally, the Stanford CoreNLP parser fails to recover sentences to an acceptable degree of accuracy. Although I plan to exploit the hierarchical structure and train a model to learn the divisions between the various components of this bill in the next iteration of this project, for now, I simply split bills by occurrence of a period followed by a newline. This splits bills so that each segment is, for example, the text uniquely associated with a subsection but with none of that parent subsection's child subdivisions, and so on for all levels of the hierarchy. I call these "statutory sentences" as a convenient shorthand. There are approximately 500,000 statutory sentences in the corpus.

The task is to classify whether each statutory sentence contains an act of delegation. An act of delegation is a mandate or permission for a federal agency or bureaucrat (including the President) to exercise public authority in some way. Allocating money for federal agencies to spend, instructing agencies to promulgate rules, granting agencies the ability to exempt themselves from pre-existing rules, requiring agencies to compile reports or commission pilot studies, and charging agencies with the enforcement of specific policies are all examples of acts of delegation.

**Positive Examples**

- (B) If a tribe chooses not to contract the development or implementation of the plan, the Secretary shall develop or implement, as appropriate, the plan in close consultation with the affected tribe.

- (2) Any civil aircraft, aircraft engine, propeller, appliance, or property on an aircraft involved in an accident in air commerce shall be preserved, and may be moved, only as provided by regulations of the Board.

**Negative Examples**

- (2) Any and all minerals reserved by paragraph (1) are hereby withdrawn from all forms of entry, appropriation, and patent under the mining, mineral leasing, and geothermal leasing laws of the United States.

- (B) During the last 20 years, relatively little changed in how students were taught. Despite much research suggesting better alternatives, classrooms continue to be dominated by textbooks, teacher lectures, short-answer activity sheets, and unequal patterns of student attention.

There is no labeled set of statutory sentences. Accordingly, I have hand coded 3,600 statutory sentences. These hand codings feature 713 examples of delegation. Two features of this dataset are worrying. The first is that it is fairly small by the standards of deep learning. The second is that it is imbalanced; only a small proportion of all statutory sentences are delegating authority. In Section 5, I will show that the model performs well in spite of these limitations.

Statutory language often diverges considerably from ordinary language. Secretary refers not to a person who keeps appointments, but high level executives. Title refers not to a name for a document, but a particular structure in the statutory hierarchy. I am therefore pessimistic about the prospects for word embeddings trained on ordinary language such as Wikipedia. Using the implementation of GloVe provded by Pennington, Socher, and Manning, I train custom embeddings on the full corpus of 30 million words Pennington *et al.* (2014). I train 50, 100, and 200 dimensional embeddings on a window size of 10, using an $x_{\max} = 100$ and a minimum count of 5. I find that 50 iterations is more than sufficient for convergence.

Table 1: Results for Baseline LASSO

| | Accuracy | Recall | Precision |
|---|---|---|---|
| GloVe Averaging (200 dim) | 79.3% | 27.1% | 63.3% |
| Binary Indicators | 84.7% | 54.3% | 73.1% |

## 3 A LASSO Baseline

To illustrate the returns to deep learning approaches in this problem, I establish a baseline using the types of methods typical to political science. LASSO is mostly understood and widely used in political science, so I employ a $\ell_2$-regularized logistic regression as the baseline.

Topic models have been influential in political science, and reliance on the bag of words assumption has accompanied them Grimmer and Stewart (2013). As a first baseline, I use word counts as the features for the LASSO. I find that performance is far better with a binary bag of words (1 if the word ever occurs, 0 if it does not) than counts.

On the other hand, political scientists are unfamiliar with word embeddings. To show that the performance improvements from the model in Section 4 come from the convolutional neural network and not just the word embeddings, I also run a LASSO specification with an average of the word vectors for features. I find that the 200-dimensional embeddings have the best performance.

Using the cross-validation-based procedure in `glmnet` to obtain the regularization parameter, I fit LASSO to both feature sets Friedman *et al.* (2010). Table 1 demonstrates that the results are quite poor. With GloVe as the feature set, accuracy is little better than simply guessing the most common class. The binary features perform somewhat better, but applied researchers would not be willing to trust an analysis where half of the instances of delegation are missing and 26.9% of the supposed instances are false positives.

The baseline model is tricked by a number of phenomena that are subtly distinct from delegation. Most of these can be sorted into one of three categories. The first are restrictions placed on the agency. Recall that, in the definitions provided in Section 2, provisions which curtail the authority that can be exercised by the agency do not count as delegation. Here, the difficulty of dealing with negation when word order is discarded rears its head. The baseline incorrectly coded the following example as positive:

> (2) When approving an application under paragraph (1) of this subsection, the Secretary may not reduce the amount of operating assistance approved for another State or a local transportation authority within the affected urbanized areas.

The second is that the statutory sentence must explicitly specify the agency which is the recipient of the delegation of authority. In general, appropriations count as acts of delegation. However, if the appropriation does not explicitly specify the target agency, then it does not count. The baseline incorrectly coded the following example as a positive example:

> Section 394A of the Public Health Service Act, as redesignated by section 201(1) of this Act, is amended by striking "To carry out" and all that follows and inserting the following: "For the purpose of carrying out this part, there are authorized to be appropriated $50,000,000 for fiscal year 1994, and such sums as may be necessary for each of the fiscal years 1995 through 1998."

Finally, the recipient of the delegation must be a federal agency or an employee within one of these agencies. States, courts, private enterprises, and private citizens do not count. The baseline model incorrectly coded the following as a positive example:

> (j) State Contributions. –In enrolling a loan under the Program, the participating State shall contribute to the reserve fund an amount, as provided for in the participation agreement, which shall not be less than the sum of the amount of premium charges paid by the borrower and the participating financial institution

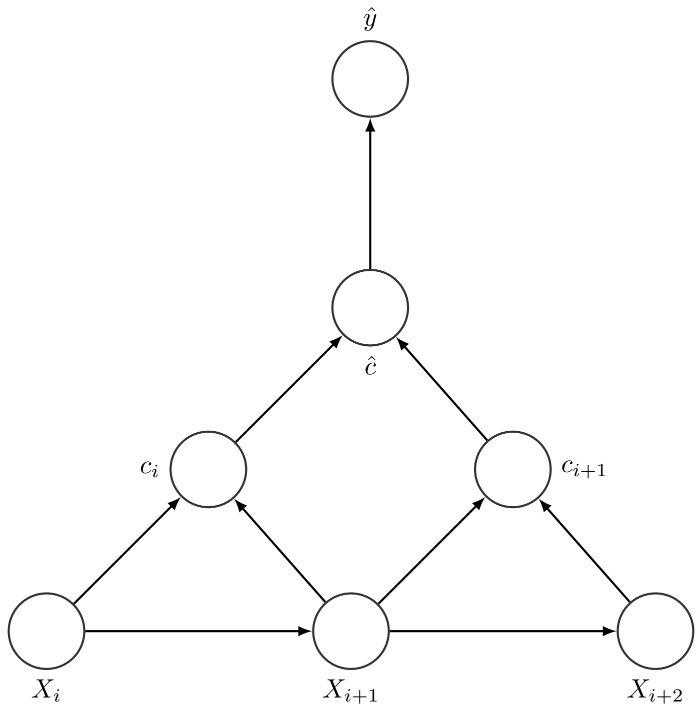


Figure 1: Toy architecture with a window size 2 and a sentence of length 3.

These account for the unacceptably low precision. The low recall follows from the objective function. The objective was to maximize the log-likelihood, and the majority of the examples were negative. This encouraged conservatism on the part of the model. The baseline may be able to achieve better performance with a modified objective function. However, in the succeeding sections, I will show that minimizing a very similar objective function with a convolutional neural network will achieve far better performance.

## 4 A Convolutional Neural Network Approach

Several features of the data and task recommend the convolutional neural network (CNN) architecture proposed by Kim Kim (2014). First, delegation is often given away by a small, variable-length phrases in much longer texts; for example, "the Secretary shall" or, "the Department, in consultation with various tribes, may." Second, delegation is a local feature. This justifies the use of filtering windows. If it occurs in one part of a statutory sentence delegates authority, it does not matter what the rest of the statutory sentence does. This justifies the use of max pooling.'

Figure 4 provides an intuition-building toy example of the architecture employed here, which is a stripped down version of Kim's CNN. The example gives a sentence of length 3 with a window size of 2. The bottom layer consists of static word embeddings. The second layer filters these word embeddings into features, the third player pools these features into a single feature, and the final layer makes a prediction from the pooled feature.

Formally, let $H$ be the set of windows under consideration. Let $\boldsymbol{X}$ be a $n \times d$ matrix, where d is the dimension of the word embeddings and $n$ is the length of the sentence. Let $\boldsymbol{W}$ be an $h \times 1$ matrix, $\boldsymbol{b}$ be a scalar, $\hat{c}^{(h)}$ be a scalar, $\hat{\boldsymbol{c}}$ be a $|H|$-vector, and let $\boldsymbol{h}^{(s)}$ be a length a scalar. Finally, let $\boldsymbol{y}$ and $\boldsymbol{haty}$ be $m$-vectors, where $m$ is the number of statutory sentences available to th estimation. $\lambda$ is a scalar regularization parameter.

$$
\begin{aligned}
c_i^{(h)} &= \text{ReLU}(\boldsymbol{X}_{i:i+h-1}\boldsymbol{W} + \boldsymbol{b}) \\
\hat{c}^{(h)} &= \max_i c_i^{(h)} \\
\hat{\boldsymbol{c}} &= [\hat{c}^{(h)}]_{h \in H} \\
\hat{y} &= \text{Sigmoid}(\hat{c}\boldsymbol{w}^{(s)} + \boldsymbol{h}^{(s)}) \\
J &= \text{CE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \lambda(||\boldsymbol{W}||_2^2 + ||\boldsymbol{w}^{(s)}||_2^2)
\end{aligned}
$$

With binary labels, minimizing cross entropy is equivalent to minimizing log-likelihood. This makes the objective function minimized here comparable to the objective function used by the baseline LASSO.

Although the architecture is a feature-sparse version of Kim's, I implement the model from scratch with TensorFlow. I use the Adam Optimizer with a learning rate of 0.001 and run for a maximum of 16 epochs. If the validation error does not improve for two consecutive epochs, the model fitting stops early.

## 5 Results

There are several parameter tuning decisions required even for this stripped down model. The set of window sizes in $H$ and the embedding dimension determine the richness of the feature space. The value of $\lambda$ determines the amount of regularization. I also ran experiments for a dropout parameter and the type of nonlinearity in the filtering layer, but these are not reported. Dropout performed poorly relative to regularizing with $\lambda$, and neither the sigmoid nor the hyperbolic tangent performed as well as rectified linear units.

The results of the parameter tuning experiments are reported in Table 2. The training set had 2,700 examples, the validation set had 300 examples, and the test set had 600 examples. Because of the smallness of the dataset, I gave very high priority to ensuring that the model did not overfit. I considered the model to have overfit if, for at least three of the four final epochs, training accuracy went up while validation accuracy went down. This is an aggressive definition of overfitting, but I felt it was warranted to ensure good performance on the test set. For each window size, I tried all embedding dimensions. If $\ell_2 = 0.02$ overfit, I increased $\ell_2$ by increments of .02 until the model did not overfit.

Using all window sizes between 2 and 6 with an embedding dimension of 200 and an $\ell_2$ of 0.06 gives the best validation accuracy on the held out test set among those specifications which were not overfit. Table 3 shows that this achieved excellent results. The test accuracy is lower than the validation accuracy and the recall is lower than might be hoped, but the precision is phenomenal. The relatively low test recall is likely a result of the imbalance in the labels; maximizing the F1 score directly or increasing the weights of the positive examples might achieve even better performance.

Encouragingly, the CNN was able to resolve the issues with negation, unspecified targets of delegation, and non-federal agency delegations that plagued the baseline model. The examples that are misclassified by the CNN tend to be quite challenging. It struggles when the distance between the subject and predicate is longer than the maximum window size. For example, the following example was a false negative:

> (e) Requirement for Clarity.–Officers and employees of the Federal Government who prescribe regulations to implement this Act and the amendments made by this Act shall make every effort practicable to ensure that the regulations are concise and are easily understandable by potential offerors as well as by Government officials.

Semantically subtle sentences are also challenging. Consider the following example, which is a false positive:

> "(b) Penalties. – "(1) Criminal.–Any person who knowingly violates any regulation promulgated by the Secretary pursuant to this section, or any provision of

Table 2: Parameter Tuning of the CNN

| Windows | Embed Dim | $\ell_2$ | Train. Acc | Val. Acc | Overfit? |
|---|---|---|---|---|---|
| 2, 4, 6 | 200 | .02 | 92.4 | 90.7 | No |
| 2, 4, 6 | 100 | .02 | 91.0 | 91.3 | Yes |
| 2, 4, 6 | 50 | .02 | 76.3 | 76.7 | No |
| 2, 3, 4, 5, 6 | 200 | .02 | 94.0 | 93.7 | Yes |
| 2, 3, 4, 5, 6 | 200 | .04 | 91.9 | 91.3 | Yes |
| **2, 3, 4, 5, 6** | **200** | **.06** | **89.3** | **91.7** | **No** |
| 2, 3, 4, 5, 6 | 100 | .02 | 91.0 | 89.7 | Yes |
| 2, 3, 4, 5, 6 | 100 | .04 | 91.0 | 91.3 | Yes |
| 2, 3, 4, 5, 6 | 100 | .06 | 90.1 | 90.3 | No |
| 2, 3, 4, 5, 6 | 50 | .02 | 90.0 | 91.3 | Yes |
| 2, 4, 6, 8 | 200 | .02 | 88.4 | 89.3 | Yes |
| 2, 4, 6, 8 | 100 | .02 | 90.3 | 92.0 | Yes |
| 2, 4, 6, 8 | 50 | .02 | 90.7 | 90.3 | Yes |
| 2-8 | 200 | .02 | 91.0 | 91.0 | Yes |
| 2-8 | 100 | .02 | 92.5 | 91.3 | Yes |
| 2-8 | 50 | .02 | 94.4 | 92.0 | No |
| 2, 4 ,6, 8, 10, 12 | 200 | .02 | 89.7 | 89.7 | Yes |
| 2, 4 ,6, 8, 10, 12 | 200 | .06 | 88.1 | 88.0 | Yes |
| 2, 4 ,6, 8, 10, 12 | 200 | .08 | 90.0 | 91.0 | No |
| 2, 4, 6, 8, 10, 12 | 100 | .02 | 92.5 | 91.3 | No |
| 2, 4, 6, 8, 10, 12 | 100 | .04 | 90.9 | 90.3 | Yes |
| 2, 4, 6, 8, 10, 12 | 100 | .06 | 91.8 | 88.7 | No |
| 2, 4, 6, 8, 10, 12 | 50 | .02 | 90.1 | 87.7 | Yes |
| 2-12 | 100 | .02 | 93.7 | 91.0 | Yes |
| 2-12 | 100 | .04 | 95.4 | 91.3 | Yes |
| 2-12 | 100 | .06 | 93.5 | 91.7 | No |
| 2-12 | 50 | .02 | 91.6 | 93.0 | Yes |
| 2-12 | 50 | .04 | 97.0 | 90.3 | Yes |
| 2-12 | 50 | .06 | 93.4 | 91.7 | Yes |
| 2-12 | 50 | .08 | 92.1 | 90.7 | No |
| 3, 5, 7, 9, 11 | 200 | .02 | 90.8 | 90.7 | No |
| 3, 5, 7, 9, 11 | 100 | .02 | 91.6 | 88.3 | No |
| 3, 5, 7, 9, 11 | 50 | .02 | 78.1 | 76.0 | No |

Table 3: Test Performance

| | |
|---|---|
| Accuracy | 90.0% |
| Recall | 70.0% |
| Precision | 93.0% |

> sections 7 through 10 or any regulation promulgated by the Secretary pursuant to such sections, shall be fined under title 18, United States Code, or imprisoned not more than 1 year, or both.

The statutory sentence is not delegating public authority to a federal agency; rather, it is defining punishments for defying a public authority that has been delegated to the Secretary earlier in the statute.

Finally, and perhaps most correctably, the CNN also misclassified jurisdictional transfers. These are acts of delegation, but they have a very different syntactic form from most other acts of delegation. For example,

> (b) Additional FAA References.–(1) Except as provided in paragraphs (2) and (3), the Foreign Assistance Act of 1961 is amended by striking "agency primarily responsible for administering this part", "agency primarily responsible for administering part I", "agency primarily responsible for administering part I of this Act" each place such phrase appears and inserting "Department of State"

## 6 Future Directions

The precision of this model would likely be satisfactory to applied researchers who wanted to use a comprehensive dataset constructed via deep learning, but the recall would not be. Some of the issues, such as the misclassification of jurisdictional transfers, could probably be solved with more training data. The semantically subtle sentences will likely stubborn holdouts even for extremely good models.

The problem of distant subjects and predicates is at the boundary of problems that are unlikely to be solved by throwing more data at the current setup but seem like they should be tractable for a good model. Richard Socher suggested trying an LSTM with max pooling as an alternative model at the poster session Gers *et al.* (2000). This is an excellent suggestion that I regrettably was unable to implement in time to include in this report. Such a model could plausibly store information from a distant subject to inform the evaluation of the predicate. I will be presenting a revision of this work at the Society for Political Methodology Annual Meeting in July, and I will certainly implement an LSTM for comparison by then.

In future iterations, I will also move beyond merely classifying the presence or absence of delegation. In other work, I have proposed hypotheses about the relationship between the type of authority being delegated and the level of discretion that will accompany the delegation. I would like to have a richer class structure, splitting delegation into appropriations, regulations, policy exploration, cutbacks, and provision of services. I would also like to classify the amount of discretion that accompanies an act of delegation. This can be achieved either by classifying the presence of certain features that accompany delegation (a hand-coded version of this was performed by Selin on a small amount of state-level legislation) as well as qualitative assessments of the degree of discretion Selin (2015).

## References

Bawn, K. (1995). Political control versus expertise: Congressional choices about administrative procedures. *American Political Science Review* **89**, 01, 62–73.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33**, 1, 1.

Gailmard, S. and Patty, J. W. (2012). *Learning While Governing: Expertise and Accountability in the Executive Branch*. University of Chicago Press.

Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: Continual prediction with lstm. *Neural computation* **12**, 10, 2451–2471.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* **21**, 3, 267–297.

Huber, J. D. and Shipan, C. R. (2000). The costs of control: Legislators, agencies, and transaction costs. *Legislative Studies Quarterly* 25–52.

Kim, Y. (2014). Convolutional neural networks for sentence classification **14**, 1746–1751.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. **14**, 1532–1543.

Selin, J. L. (2015). What makes an agency independent? *American Journal of Political Science* **59**, 4, 971–987.

Wilson, J. Q. (1989). *Bureaucracy: What government agencies do and why they do it*. Basic Books.