Text reuse and Paraphrase Detection with Semantical Smith-Waterman Local Alignment

Alexander Furnas

Abstract

Coordination among political elites is an issue of substantive interest across political science. This paper details a new method for detecting text reuse and paraphrasing in political texts, a critical measurement task in leveraging new text data to observe patterns of coordination. The method proposed is an extension of the Smith-Waterman local alignment algorithm with semantically aware mismatch penalties. This modification enables detection of instances of text reuse in which words are changed to semantically similar alternatives to fit new contexts or disguise the source of the text. This method is applied to a corpus of tweets sent by Members of Congress and their electoral challengers during the 2016 election cycle.

Motivation

Studies of congressional behavior have long focused on the motivations of Members of Congress. Fenno (1978) famously argues that members are motivated by three primary goals: re-election, power within the institution, and good public policy. To achieve these goals they may collaborate or coordinate with other members. There has been substantial theoretical work exploring members voting decisions (Kingdon 1977), allocation of attention to issues (Hall 1998, Hall and Deardorff 2006), and how they communicate with their constituents (Miller and Stokes 1963, Mayhew 1974). Following Mayhew (1974), scholars have focused largely on how re-election motivation shapes legislators representation behavior. Recent work in using text analysis has looked at how members represent their work in Washington to their constituents (Grimmer 2010), and tested longstanding theories of distributive politics and credit claiming (Grimmer et al. 2012).

Members share the primary goals of re-election, institutional prestige and public policy outcomes with each other in varying degrees. While the between member coherence may vary just as their ideologies, constituencies and aspirations do, the existence of shared goals creates the opportunity for productive collaboration. Indeed, the Cox and McCubbins (2005) conceive of parties in the House as collaborative undertaking to protect a re-electable party brand by controlling the agenda. The bargaining within legislatures that enables logrolling and undergirds distributive theories of legislative organization is necessarily a collaborative and coordinated process (Baron and Ferejohn 1989, Shepsle and Weingast 1987, Berry and Fowler 2015).

Indeed, we may think of different types of collaboration and coordination that support members differing goals, ambitions and styles. A substantial body of work has explored how members collaborate by cosponsoring legislation (Zhang et al. 2008, Fowler 2006, Koger 2003).

Koger (2003) argues that cosponsorship serves as a form of position taking and as a signal to agenda-setters. Campbell (1982) finds that legislators cosponsor bills for ideological, partisan, and electoral reasons, suggesting that they see this form of cooperation as meaningful in achieving their goals.Fowler (2006) shows that centrality in the cosponsorship network is predictive of legislative effectiveness. Cosponsorship can send messages of bipartisanship to obstructionist minorities (Wilson and Young 1997). At the state level Bratton and Rouse (2011) find that ideology and homophily predict cosponsorship between legislators. Legislators use cosponsorship to signal to constituents, help pass policy they support and increase their standing within the institution. This form of coordination and collaboration between Members of Congress is an essential part of legislative behavior.

Recent work by Craig (2016) has explored between Member collaboration in Dear Colleague letters, a semi-formalized system of member-to-member communication used for information sharing, cosponsorship recruitment, issue advocacy or administrative or operational communication. Craig (2016) finds a significant degree of bipartisan collaboration in the Dear Colleague network and Box-Steffensmeier et al. (2015) finds that letters can have an impact on the legislative success of a bill. In this too, Members' are coordinating and collaborating with each other in pursuit of their goals.

In recent years, scholars in political science have begun using natural language processing techniques to analyze vast troves of politically relevant text (Grimmer and Stewart 2013). This work has applied techniques ranging from topic modeling (Roberts et al. 2014) to scaling to recover latent traits (Lowe and Benoit 2013).

Lin et al. (2014) use distinctive n-grams in member communications over time to detect

"semantic bursts" of coordinated among Senators, and find that these coordinated communication networks reflect underlying institutional and partisan factors. Lin et al. (2015) explore how differing the reuse and diffusion of different length n-grams reflect different behavioral processes within the chamber.

A recent body of work has begun explicitly to detect text reuse in political texts, particularly for modeling diffusion or policy (Smith et al. 2013, 2014, Wilkerson et al. 2015, Furnas and Shipan 2017, Linder et al. 2017). The detection of text-reuse offers an exciting new means for exploring the patterns of collaboration and coordination among political actors.

Wilkerson et al. (2015) detect reuse of legislative language from previously introduced bills in the Affordable Care Act. Their suggested re-framing of legislative behavior from bill to policy-idea as a meaningful unit of analysis suggests a greater degree of ideological diversity in ACA, and offers exciting new avenues for the study of legislative effectiveness. More recently, Burgess et al. (2016), Linder et al. (2017) look at how legislative text is shared and reused between states. Furnas and Shipan (2017) use the same instances of text reuse to construct new bridges for cross state ideal point estimation.

Wilkerson et al. (2015), Burgess et al. (2016), Linder et al. (2017) and Furnas and Shipan (2017) all use the same algorithm for identifying text reuse: the Smith-Waterman local alignment algorithm. The Smith-Waterman algorithm was originally developed in biostatistics to find the optimal local alignments of sequences of nucleic acids Smith and Waterman (1981). Work bySmith et al. (2013, 2014) has adapted the Smith-Waterman local alignment algorithm (SW, hereafter) for text purposes.

SW is an effective algorithm for finding examples of exact text reuse, or reuse of exact

text with additional text insertions. A global text alignment algorithm, like the Needleman-Wunsch, would do a poor job of detecting text from string A reused in string B, if B were identical to A but with a new appositional phrase inserted in the middleNeedleman and Wunsch (1970).

SW is a dynamic programming algorithm that computes all of the possible alignments of two sequences, tabulating scores according to a bonus for aligned items, a penalty for mismatching items and a penalty for the insertion of gaps. Backpointers to previous parts of the alignment are stored in a table. The maximum possible score is found, and then backpointers are used to determine what path of the possible alignments yielded that optimal alignment. In this way it is similar to an edit distance calculation. All misalignments, that is cases where words in the two sequences do not match, are treated the same. Linder et al. (2017) provides an extensive explanation of the algorithm for political science audiences.

To date, SW has been by political scientists used largely to detect the reuse of legislative text. In a legislative context, where changing a single "shall" to a "may" can significantly alter the meaning of a provision, exact matches may be the appropriate form of text reuse to detect. However, in other forms of political text or speech we may be interested in a more flexible form of text reuse. SW is less effective for detecting instances of reuse in which moderate paraphrasing or adaptation has been done. This is especially relevant in the context of political speech, as two actors offering essentially the same message may need to adapt it to fit their circumstances.

As text reuse detection becomes an increasingly common tool to detect patterns of coordination and collaboration among political actors — beyond legislative language — I argue that a more flexible tool is necessary.

I propose a modification of the standard SW algorithm in to better detect this adaptive reuse. This modified version is called the Semantic Smith-Waterman (SSW. hereafter). Rather than treating all mismatches the same, the method proposed here scores each mismatch according to word similarity in a semantic space and weights these mismatches accordingly. The more similar the two words are the less the mismatch is penalized. While I explore only a simple semantic distance weighting here, this simple modification enables a variety of more complex and semantically meaningful penalization schemes.

Method

Consider a hypothetical legislator expressing one of the following sentiments:

A: *"My esteemed colleague is resident of the state of Michigan, and clearly proud of it."*

B: *"My esteemed colleague is resident of the state of denial, and clearly proud of it."*

In statement A, the legislator is making a straightforward statement about the state of residence of their colleague. In statement B, the legislator is making an ironic statement about their colleague's inability to confront facts. While the text in these two instances is almost exactly the same, their meanings are notably different.

Now consider a second hypothetical legislator, making the following statement:

C: *"My esteemed colleague is resident of the state of Ohio, and clearly proud of it."*

Is statement C more similar to A or B? The standard SW algorithm would say that C is equally as similar to A as it is to B. C aligns with A and B perfectly except for one mismatch

("Michigan" — "Ohio" and "denial" — "Ohio") which are penalized equally according to a constant.

The SSW scores these differently. The mismatch penalty applied depends on the relationships between the words in a semantic space. In the application of SSW presented here, words are compared in a google's pre-trained 300 dimension `word2vec` word embeddings. These vector space word representations are pre-trained according to the method described by Mikolov, Chen, Corrado and Dean (2013) using the google news corpus. However, a future more general application will present the possibility of using a variety of vector spaces. For example, it may be useful to construct a custom latent semantic space out of the corpus of documents being analyzed using Latent Semantic Analysis (Furnas et al. 1988, Deerwester et al. 1990). The algorithm presented here can be simply adapted for any custom vector space representation of words.

Algorithm 1 provides a detailed explication of the the SSW algorithm. The algorithm functions nearly identically to the standard SW, except that mismatches are scored according to a similarity function rather than penalized with a constant. The similarity score $s(a, b)$ for two words a and b is derived according to the scoring function described in Algorithm 2.

The SSW consists of the same four steps as SW:

1. Choice of scoring parameters.
2. The scoring and traceback matrices are initialized,
3. The scoring matrix is filled, and backpointers stored in the traceback matrix
4. Traceback starting from the highest scoring cell in the scoring matrix.

ALGORITHM 1: Semantic Smith Waterman [1]

Input: Two sequences of words, A and B, such that $A = a_1a_2...a_n$ and $B = b_1b_2...b_m$ where n and m are the lengths of A and B respectively.

Output: Optimal semantically aware local alignment of A and B with gaps, alignment score

initialize *Scoring matrix* **H** with dimensions $(n+1) \times (m+1)$,
$H_{k0} = H_{0l} = 0 \quad for \quad 0 \leq k \leq n \quad and \quad 0 \leq l \leq m$;
initialize *Traceback matrix* **T** with dimensions $(n) \times (m)$;

for *each node i in j in* $1 : n+1, 1 : m+1$ **do**

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + s(a_i, b_j), \\ H_{i-1,j} - W, \\ H_{i,j-1} - W, \\ 0 \end{cases}$$

where:
$H_{i-1,j-1} + s(a_i, b_j)$ is the semantically weighted score of aligning a_i and b_j,
$H_{i-1,j} - W_1$ is the score if a_i is at the end of a gap,
$H_{i,j-1} - W_1$ is the score if b_j is at the end of a gap,

$$T_{ij} = \begin{cases} (i-1, j-1), & \text{if } H_{ij} = H_{i-1,j-1} + s(a_i, b_j) \\ (i-1, j), & \text{if } H_{ij} = H_{i-k,j} - W \\ (i, j-1), & \text{if } H_{ij} = H_{i,j-l} - W \end{cases}$$

end

traceback *Starting at* $H_{ij} = max(\mathbf{H})$
alignment = [ ]
while $H_{ij} > 0$ **do**
append T_{ij} to alignment
set i,j to T_{ij}
end
return *score* $= max(\mathbf{H})$, *alignment = alignment*

Semantic mismatch scoring

At this point it some additional explanation of the semantic mismatch scoring used here is warranted at is the principal difference between SW and SSW. Standard SW generally imposes a constant penalty for a mismatch, and then either a linear penalty for gap scoring or an affine

[1] Notation and explication for the Semantic Smith-Waterman used here is adapted from the Wikipedia article for the standard Smith-Waterman.

gap penalty (where there is a greater penalty for creating than extending a gap). Both Wilkerson et al. (2015) and Linder et al. (2017) use affine gap scoring versions of SW. For the sake of simplicity the proof-of-concept implementation described here uses linear gap penalty, but there is no reason that SSW could not be extended to include affine gap scoring. A forthcoming python package that implements SSW will include an affine gap option.

This implementation makes a few other simplifying assumptions that could be parameterized in a future implementation of SSW. These are as follows:

- Cosine similarity of word vectors in the`word2vec` semantic space is used as the similarity measure.
- Cosine similarity scores below .5 are considered a mismatch and standard constant mismatch penalty is applied.
- The constant mismatch penalty is set to be the same as the constant gap penalty W.
- Perfect matches are set given a positive score twice the magnitude of the gap/mismatch penalty.[1]

Algorithm 2 shows this implementation. Here, $\mathbf{V}$ is a vector space from the `word2vec` word embeddings.

Initialization

The scoring matrix is initialized to with the words from string A along the rows and from string B along the columns, but with a first row and column set to 0. The result is a matrix

ALGORITHM 2: Semantically Weighted Alignment Score

Input: Two words, a and b, semantic vector space $\mathbf{V}$, and baseline weight W.
Output: An alignment score, $s(a, b)$ to be used in the Semantic Smith-Waterman alignment
if $a = b$ **then**
    score $= 2W$
end
else if *a is a stop word or b is a stop word* **then**
    score $= -W$
end
else
    $s = \cos(\mathbf{V}[a], \mathbf{V}[b]) - .5$
    score$= 2Ws$
end

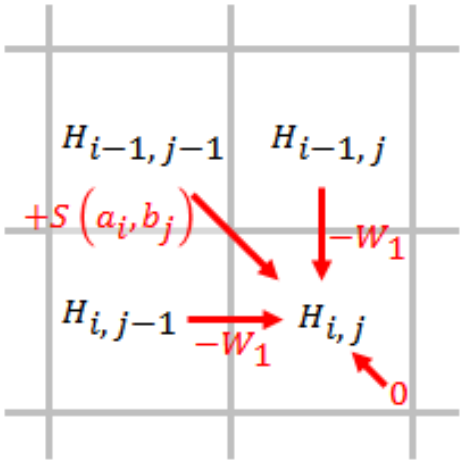

Figure 1: Scoring and Traceback. Source: Wikimedia Commons

with dimensions $(n + 1) \times (m + 1)$. Here, n is the number of words in A and m is the number of words in B. The traceback matrix is initialized to be size $n \times m$.

Scoring

As cells in the scoring matrix are filled,as shown in Figure 1, a pointer is stored in the traceback matrix indicating which of the neighboring cells provided the max value for $H_{i,j}$.

Traceback

The cell in the traceback matrix that corresponds to the cell in the scoring matrix which provides the maximum value is used as the starting position for traceback. Starting from this cell, the backpointers are used to determine the optimal local alignment of the two strings. Backtracing ends when the corresponding cell in the scoring matrix reaches 0.

Empirical application

To explore the utility of this new approach to text reuse detection, I applied both SW and SSW to a corpus of tweets of Members of Congress and congressional candidates during the 2016 election cycle.[2]

This corpus contains 893168 non-retweet tweets from 714 different accounts. Of these I took the first 100000 tweets and looked for optimal local alignments in the rest of the 893168 tweet corpus with both SW and SSW. These algorithms are computationally expensive, as they are $O(nm)$. Checking for each optimal alignment would require this be done $100000 \times 893168 \approx 8.93 \times 10^{10}$ times.

To shrink the search space, I compared all documents to each other in a simple vector space constructed using `nltk` and `gensim` in python (Bird 2006, Řehůřek and Sojka 2010). Because I wanted to find results with as similar exact words as possible to each other, I did no transformations on this vector space, and did not remove stop words. Each tweet was then

[1] Notation and explication for the Semantic Smith-Waterman used here is adapted from the Wikipedia article for the standard Smith-Waterman.

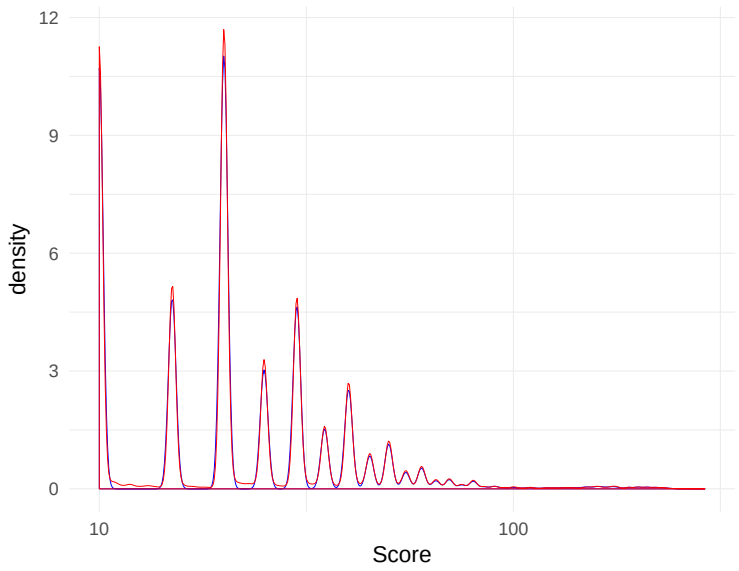

Figure 2: Smith-Waterman and Semantic Smith-Waterman Score densities

compared to the 100 closest tweets in this vector space using both SW and SSW. The scaling parameter used as the baseline gap penalty W was set to 10 for these trials.

Below I present the distributions of these similarity scores. Note, of course, that the sample of similarity scores calculated is a non-random sample of similarities between these tweets because of how the search space was constrained. As Figure 3 shows, these scores are highly similar, of course, but the interesting results come in the instances where SSW is substantially different from SW.

Filtering the set of tweets to dyads with similarity scores above 90 — the equivalent of a 9 word reuse sequence — leaves a set of 144,031 tweets. We can observe substantial variation between the SSW and SW scores here. The two distributions converge above about 190, perfect matches at the tweet length limit.

SSW yields strictly higher scores than SW, as the semantic similarity score defaults to

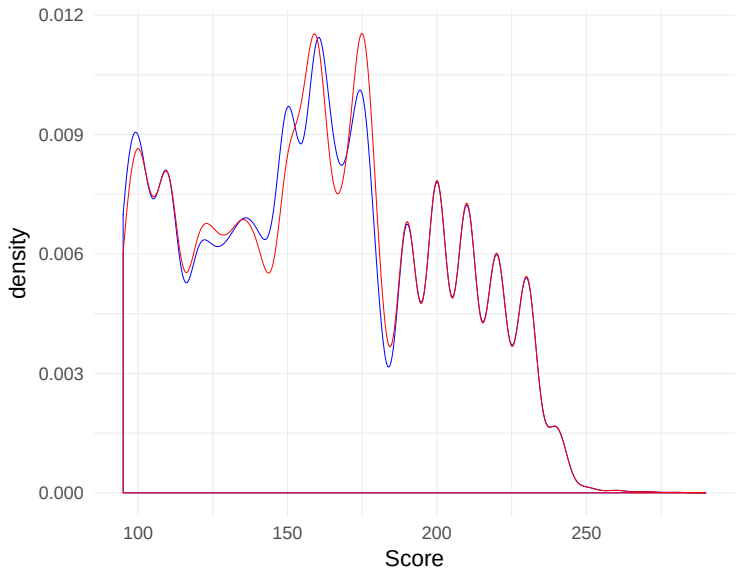

Figure 3: Smith-Waterman and Semantic Smith-Waterman Score densities above 90

the standard mismatch penalty if the semantic similarity score is negative. The distribution of these differences is shown in Figure 4 for all cases where that difference is non-zero.

Tables 1 and 2 provide a sample of tweet interesting tweet comparisons using SW and SSW. Those provided below are the top 50 tweets with SW scores above 90, sorted by the magnitude of the difference in their SW and SSW scores. They provide a useful illustration of the kinds of similarity that SSW can help detect that SW does not. For example, the tweet "Thank you to all brave men and women serving our nation and keeping us safe. #ArmedForcesDay https://t.co/xVswnTWMR0" and "Thank you to all of the brave men and women who serve our country and keep us safe. #ArmedForcesDay" express the same semantic content, but have been phrased slightly differently. SSW yields a score that is the equivalent of nearly 3 additional overlapping words.

Similarly, the tweet "My thoughts and prayers are with the victims, their families and all

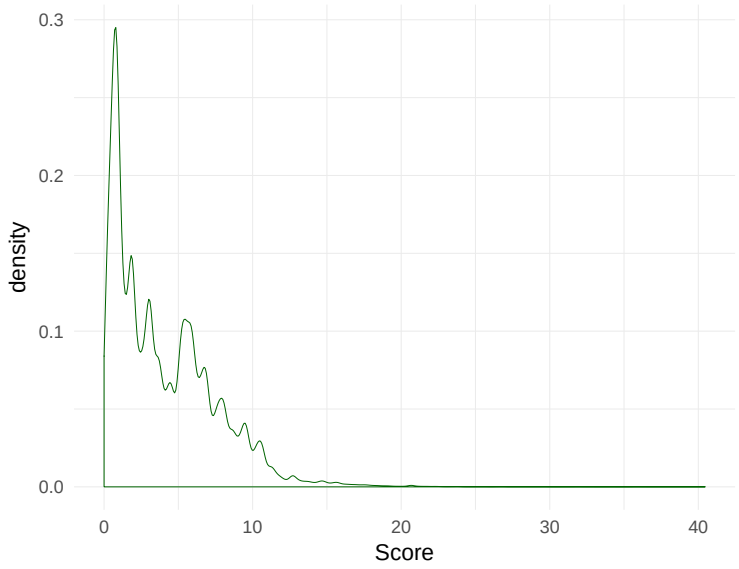

Figure 4: Density of difference between Smith-Waterman and Semantic Smith-Waterman Score densities

those affected by the attack in Orlando." contains the same meaningful semantic content as "My thoughts and prayers are with the victims, the injured, their families and all others impacted by today's senseless attacks in #Boston." but is adapted to fit a different context following a different attack. The virtue of SSW is in detecting the use of similar phrasing adapted to new contexts in this manner.

A closer examination of the tweet pair presented in tables 1 and 2 reveals some interesting themes. We see members responding with similar content to crises, breaking news, political and calendar events. In many cases, semantic meaning is identical but minor phrasing variations exist between members, often using different patterns of hash-tagging. SSW presents the possibility of detecting coordinated messaging despite the kinds of changes social media interns may introduce to fit with members goals and homestyles.

As a further demonstration of the utility of this tool, I have conducted a simple

descriptive network analysis of similar tweet dyads. Figure 5 shows the network of similar tweets between the Members of Congress and their challengers during the 2016 election cycle. The set of dyads which were compared using SSW was subset to include only dyads which scored above 105 (the minimal score included in tables 1 and 2). These dyads were considered to be an instance of shared messaging.

I then summed the number of tweet dyads above this threshold between each pair in the sample to derive a weighted edge between them. It is worth noting that while tweets from incumbent members sent 74.32 percent of the 893,168 tweets in the sample, they sent 97.05 percent of the tweets in the 13,489 dyads that surpass the 105 similarity threshold with other tweets.[3] Members of Congress appear to tweet much more similarly to each other than they do to their challengers. Figure 5 shows the network of these shared tweets, with actors colored by their party. Descriptively we can see that actors tend to share tweets with their co-partisans more frequently than out-partisans. It is interesting to note, however, that the degree of separation between these two communities is less than we tend to observe in other political networks like co-sponsorship, voting in Congress.

Of course much more work would be needed to test the association between tweet sharing and partisanship or other meaningful covariates like institutional position, geography, ideology, or issue interest rigorously. However this brief descriptive result serves to demonstrate the utility of SSW. This application of the technique produced a rich relational dataset that is well suited to testing further theories of congressional behavior.

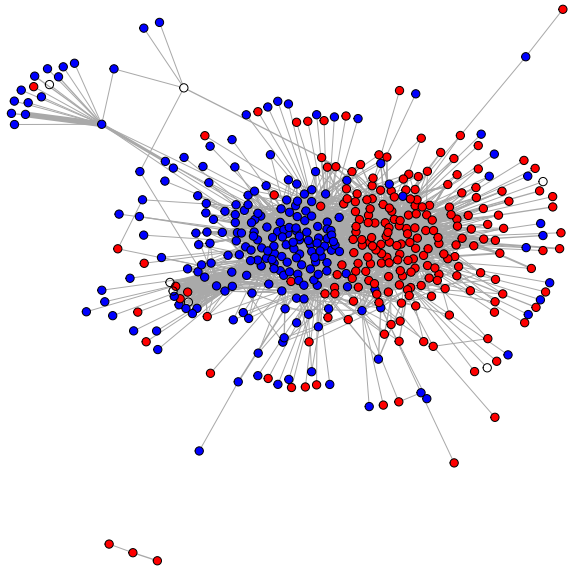

Figure 5: Network of similar tweeters

Conclusion and Future Work

This paper makes the case that coordination and cooperation is a central issue in the study of elite political actors generally and the study of congressional behavior in particular. In it I have presented a substantively meaningful extension of a text reuse detection tool that is better suited for detecting instances of reuse where the text has been adapted to suit new circumstances or new contexts.

The method details a manner for weighting misalignments in text strings according to semantically meaningful differences. This paper has demonstrated the usefulness of this technique by inferring a message-sharing network in a corpus of tweets by Congress Members and candidates. There is notably partisan, and there is much more apparent message overlap between members than candidates.

While the version presented here is a comparatively simple implementation, the method is highly extensible and can ultimately enable much more sophisticated and context sensitive weighting. For example, in the type of semantic word embeddings used here,`word2vec`, Mikolov, Yih and Zweig (2013) have observed semantically meaningful regularities in the vector space. That is, words pairs with the analogous semantic relationships (e.g. comparative::superlative) tend to have similar offsets as each other in vector the vectors space. It may be possible, then, to identify the offset relationship between compulsory and non-compulsory words, and weight the dimensions in which that offset occurs highly. This would allow for a version of SSW that could appropriately account for the small but extremely significant change of "shall" to "may" in a bill.

Notes

[1]This matches the most common or default ratio between gap and match penalties uses in most applications of the standard SW.

[2]These tweets were kindly shared with me by Joseph DiGrazia, and when he publishes something with this corpus, any work that I produce validating this measure using this corpus will cite him.

[3]It should be noted that some tweets are counted multiple times here, if they show up in multiple dyads.

References

Baron, D. P. and Ferejohn, J. A.: 1989, Bargaining in legislatures., *American political science review* **83**(04), 1181–1206.

Berry, C. R. and Fowler, A.: 2015, Cardinals or clerics? congressional committees and the distribution of pork, *American Journal of Political Science* .

Bird, S.: 2006, Nltk: the natural language toolkit, *Proceedings of the COLING/ACL on Interactive presentation sessions*, Association for Computational Linguistics, pp. 69–72.

Box-Steffensmeier, J. M., Christenson, D. P. and Craig, A. W.: 2015, Cue-taking in congress: Interest group signals from dear colleague letters. Presented at The Role of Special Interests in American Politics, University of Michigan.

Bratton, K. A. and Rouse, S. M.: 2011, Networks in the legislative arena: How group dynamics affect cosponsorship, *Legislative Studies Quarterly* **36**(3), 423–460.

Burgess, M., Giraudy, E., Katz-Samuels, J., Walsh, J., Willis, D., Haynes, L. and Ghani, R.: 2016, The legislative influence detector: Finding text reuse in state legislation, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 57–66.

Campbell, J. E.: 1982, Cosponsoring legislation in the us congress, *Legislative Studies Quarterly* pp. 415–422.

Cox, G. W. and McCubbins, M. D.: 2005, *Setting the agenda: Responsible party government in the US House of Representatives*, Cambridge University Press.

Craig, A.: 2016, The room where it happens: Collaborative strategies in the u.s. house of representatives. Presented at PolNet 2016.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R.: 1990, Indexing by latent semantic analysis, *Journal of the American society for information science* **41**(6), 391.

Fenno, R. F.: 1978, *Home style: House members in their districts*, Pearson College Division.

Fowler, J. H.: 2006, Connecting the congress: A study of cosponsorship networks, *Political Analysis* pp. 456–487.

Furnas, A. C. and Shipan, C.: 2017, Using model legislation to estimate ideology scores for state legislators. A paper prepared for presentation at the Midwest Political Science Association Annual Meeting, Chicago.

Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A. and Lochbaum, K. E.: 1988, Information retrieval using a singular value decomposition model of latent semantic structure, *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 465–480.

Grimmer, J.: 2010, A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases, *Political Analysis* pp. 1–35.

Grimmer, J., Messing, S. and Westwood, S. J.: 2012, How words and money cultivate a personal vote: The effect of legislator credit claiming on constituent credit allocation, *American Political Science Review* **106**(04), 703–719.

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

Hall, R. L.: 1998, *Participation in congress*, Yale Univ Pr.

Hall, R. L. and Deardorff, A. V.: 2006, Lobbying as legislative subsidy, *American Political Science Review* **100**(01), 69–84.

Kingdon, J. W.: 1977, Models of legislative voting, *The Journal of Politics* **39**(3), 563–595.

Koger, G.: 2003, Position taking and cosponsorship in the us house, *Legislative Studies Quarterly* **28**(2), 225–246.

Lin, Y.-R., Margolin, D. and Lazer, D.: 2014, Tracing coordination and cooperation structures via semantic burst detection, *EAI Endorsed Transactions on Collaborative Computing* **14**(2).

Lin, Y.-R., Margolin, D. and Lazer, D.: 2015, Uncovering social semantics from textual traces: A theory-driven approach and evidence from public statements of us members of congress, *Journal of the Association for Information Science and Technology* .

Linder, F., Desmarais, B., Burgess, M. and Giraudy, E.: 2017, Text as policy: Measuring policy similarity through bill text reuse. Working paper.

Lowe, W. and Benoit, K.: 2013, Validating estimates of latent traits from textual data using human judgment as a benchmark, *Political Analysis* pp. 298–313.

Mayhew, D. R.: 1974, *Congress: The electoral connection*, Yale University Press.

Mikolov, T., Chen, K., Corrado, G. and Dean, J.: 2013, Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* .

Mikolov, T., Yih, W.-t. and Zweig, G.: 2013, Linguistic regularities in continuous space word representations., *Hlt-naacl*, Vol. 13, pp. 746–751.

Miller, W. E. and Stokes, D. E.: 1963, Constituency influence in congress, *American political science review* **57**(01), 45–56.

Needleman, S. B. and Wunsch, C. D.: 1970, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *Journal of molecular biology* **48**(3), 443–453.

Řehůřek, R. and Sojka, P.: 2010, Software Framework for Topic Modelling with Large Corpora, *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta, pp. 45–50.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B. and Rand, D. G.: 2014, Structural topic models for open-ended survey responses, *American Journal of Political Science* **58**(4), 1064–1082.

Shepsle, K. A. and Weingast, B. R.: 1987, The institutional foundations of committee power, *American Political Science Review* **81**(01), 85–104.

Smith, D. A., Cordell, R. and Dillon, E. M.: 2013, Infectious texts: Modeling text reuse in nineteenth-century newspapers, *Big Data, 2013 IEEE International Conference on*, IEEE, pp. 86–94.

Smith, D. A., Cordell, R., Dillon, E. M., Stramp, N. and Wilkerson, J.: 2014, Detecting and modeling local text reuse, *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, IEEE Press, pp. 183–192.

Smith, T. F. and Waterman, M. S.: 1981, Identification of common molecular subsequences, *Journal of molecular biology* **147**(1), 195–197.

Wilkerson, J., Smith, D. and Stramp, N.: 2015, Tracing the flow of policy ideas in legislatures: A text reuse approach, *American Journal of Political Science* **59**(4), 943–956.

Wilson, R. K. and Young, C. D.: 1997, Cosponsorship in the us congress, *Legislative Studies Quarterly* pp. 25–43.

Zhang, Y., Friend, A. J., Traud, A. L., Porter, M. A., Fowler, J. H. and Mucha, P. J.: 2008, Community structure in congressional cosponsorship networks, *Physica A: Statistical Mechanics and its Applications* **387**(7), 1705–1712.

Table 1: Tweet pairs with SW above 90 and SSW above 105

SW Score	SSW Score	Query Tweet	Response Tweet
100	129.17	Thank you to all brave men and women serving our nation and keeping us safe. #ArmedForcesDay https://t.co/xVswnTWMR0	Thank you to all of the brave men and women who serve our country and keep us safe. #ArmedForcesDay
110	134.63	My thoughts and prayers are with the victims, their families and all those affected by the attack in Orlando.	My thoughts and prayers are with the victims, the injured, their families and all others impacted by today's senseless attacks in #Boston.
95	118.66	My staff and I are safe and praying for anyone who may be injured. Thankful for the U.S. Capitol Police and their swift actions today. #GA08	My staff and I are safe and accounted for. Praying for those who may have been hurt and grateful to the Capitol Police.
105	126.79	Today is #equalpayday-50 years after the enactment of the Equal Pay Act, women still earn just 77 cents to every dollar a man earns.	51 years after passage of the Equal Pay Act, a woman still earns 77 cents for every dollar a man earns. This is unacceptable #NoMadMenPay
110	131.26	#LWCF has preserved our natural & cultural heritage for 50yrs. We must reauthorize this essential tool for conserving iconic #publiclands	#LWCF has preserved our natural & cultural heritage for 50 yrs. We should embrace this tool to conserve iconic #publiclands
110	131.26	#LWCF has preserved our natural & cultural heritage for 50yrs. We must reauthorize this essential tool for conserving iconic #publiclands	#LWCF has preserved our natural & cultural heritage for 50 yrs. We should embrace this tool to conserve iconic #publiclands
110	131.26	#LWCF has preserved our natural & cultural heritage for 50yrs. We must reauthorize this essential tool for conserving iconic #publiclands	#LWCF has preserved our natural & cultural heritage for 50 yrs. We should embrace this tool to conserve iconic #publiclands
110	131.26	#LWCF has preserved our natural & cultural heritage for 50yrs. We must reauthorize this essential tool for conserving iconic #publiclands	#LWCF has preserved our natural & cultural heritage for 50 yrs. We should embrace this tool to conserve iconic #publiclands
100	121.15	My thoughts and prayers are with the victims of this unspeakable tragedy in Boston and their families.	Thoughts and prayers are with the victims of the horrific attack in Orlando and their families.
95	115.97	I am shocked to hear of the horrific explosions in Boston today. My thoughts are with the victims & families & the brave first responders.	I am deeply saddened to hear about the tragedy in #Orlando. My thoughts are with the victims, their families, and the city.
105	125.52	#ObamaBudget by the numbers: $964 billion in new spending, $1.1 trillion in new taxes & $8.2 trillion in new debt. Americans deserve better	President Obamas #budget by the numbers: $8.2 trillion in new debt, $1.1 trillion in new taxes, & $964 billion in new spending.
105	125.52	#ObamaBudget by the numbers: $964 billion in new spending, $1.1 trillion in new taxes & $8.2 trillion in new debt. Americans deserve better	President Obamas #budget by the numbers: $8.2 trillion in new debt, $1.1 trillion in new taxes, & $964 billion in new spending.
95	114.95	Wishing all of the moms out there a very happy Mother's Day. Thank you for all you do! #happymothersday https://t.co/DqhlqdtDq4	I want to wish all of the dads a very Happy Fathers Day and thank you for everything you do. https://t.co/In5sXa8bi6
115	134.77	My thoughts and prayers are with the families of the two officers who were slain in Des Moines, Iowa.	My thoughts and prayers are with the families of the five warriors who were killed in today's helicopter crash in... http://t.co/Zu0OiOAff3
95	114.16	#Friedrichs decision is a victory for fair access to union representation, but we must stay vigilant against attacks on organized labor.	SCOTUS #Friedrichs tie decision is a win for union representation & workers. We must remain strong against attacks on organized labor.
95	113.48	My thoughts and prayers are with everyone impacted by the tornadoes in Oklahoma.	My thoughts and prayers are with all those impacted by the tornado in Oklahoma.
110	128.06	This morning my thoughts and prayers are with the people of Oklahoma and all those affected by yesterday's tornado. #PrayforOklahoma	RT@SpeakerBoehner: Our thoughts and prayers are with the people of Japan and all those impacted by this mornings disaster.
110	127.88	Today is the last day to register to vote in the November 3rd elections in Virginia. Here's more information: https://t.co/DZ5I9uQbuJ	Today is the last day to register to vote in the Nov election in GA. Don't be left out or left behind, register today http://t.co/1G4uOQdd20
120	137.76	#Energy & #Manufacturing go hand in hand. Lower energy costs keep #USA manufacturers competitive, and create #jobs. — #NationOfBuilders	#Energy & #manufacturing go hand in hand. Lower energy costs keeps U.S. manufacturers competitive & creates #jobs #NationOfBuilders
120	137.76	#Energy & #Manufacturing go hand in hand. Lower energy costs keep #USA manufacturers competitive, and create #jobs. — #NationOfBuilders	#Energy & #manufacturing go hand in hand. Lower energy costs keeps U.S. manufacturers competitive & creates #jobs #NationOfBuilders
100	117.36	#Startups are creating #jobs & supporting #innovation. RT to give them the recognition they deserve! #StartupDay http://t.co/P9cdZlkEkL	#Startups create #jobs here in #MA03 & support #innovation. RT to give them recognition they deserve! #StartupDay http://t.co/7TF0fIyi5a
95	112.36	Start-up companies are creating jobs & supporting innovation. RT to give them recognition they deserve! #StartupDay http://t.co/ttZ7TtAhqi	Happy #StartupDay! Startup companies create #jobs & support #innovation. RT to give them recognition they deserve! https://t.co/8fKmjO0RPv
115	132.24	I believe Congress should do the job it was elected to do, or they should not be paid. #NE02 https://t.co/vty3zwlpho	Members of Congress must do the job they were elected to do, or they should not be paid. RT if you agree. https://t.co/8Wtsq9MYwg

Table 2: More Tweet pairs with SW above 90 and SSW above 105

SW Score	SSW Score	Query Tweet	Response Tweet
105	122.23	Sad to hear of the terror attack against innocent civilians in #TelAviv #Israel. My thoughts & prayers are with the victims & their families	I condemn the vicious terrorist attack against innocent worshipers in #Jerusalem. My thoughts & prayers are with the victims' families.
95	112.17	We can invest in young girls by addressing reproductive health, education, livelihoods, & civic engagement. #YouthDay #SRHR	We must invest in girls w/ approaches that address sexual & reproductive health, education, livelihoods, & civic engagement! #YouthDay #SRHR
100	117.14	I stand with #LGBT Ukrainians celebrating #OdessaPride2016. No one should be targeted because of who they are or who they love.	As we mark LGBT Pride Month, we are reminded that no one should ever be a target because of who they are or whom they love.
110	126.97	Today is the 79th anniversary of #SocialSecurity, a crucial part of #SocialSafetyNet for many deserving Americans that we must protect!	Today is 79th anniv of #SocialSecurity, a critical part of the #SocialSafetyNet for millions of hard-working Americans that we must protect!
130	146.97	Today is the 79th anniversary of #SocialSecurity, a crucial part of #SocialSafetyNet for many deserving Americans that we must protect!	Today is the 79th anniv of #SocialSecurity, a critical part of #SocialSafetyNet for millions of hard-working Americans that we must protect!
100	116.51	It's past time to #DisarmHate. Tomorrow as LGBTQ & gun violence prevention groups unite to demand action, I stand with them #DisarmHateRally	Tomorrow LGBTQ & gun violence prevention advocates are uniting in DC to demand action. I stand with them #DisarmHate https://t.co/M1bibV69SV
100	116.51	It's past time to #DisarmHate. Tomorrow as LGBTQ & gun violence prevention groups unite to demand action, I stand with them #DisarmHateRally	Tomorrow LGBTQ & gun violence prevention advocates are uniting in DC to demand action. I stand with them #DisarmHate https://t.co/adHJL27B0Z
100	116.51	It's past time to #DisarmHate. Tomorrow as LGBTQ & gun violence prevention groups unite to demand action, I stand with them #DisarmHateRally	Tomorrow LGBTQ & gun violence prevention advocates are uniting in DC to demand action. I stand with them #DisarmHate https://t.co/VGO27MOimg
105	121.40	On Pearl Harbor Remembrance Day, we honor those who have served and sacrificed for our country. http://t.co/1uzTNA2j6K	As the son of a WWII vet, on Pearl Harbor Remembrance Day we honor all who served & sacrificed for our great nation. http://t.co/f9kscuCdW1
120	136.19	Of course, we still have a lot more work to do to make sure that veterans get the care they deserve, and I'll continue to stay on it.	The VA still has a lot of work to do to make sure that our Veterans are getting the care they've earned and deserve! https://t.co/Rj3o9iU2jh
95	110.94	Did you know the govt paid $1.2 million to pay people to play World of Warcraft? http://t.co/Nrg000gx Instead of tax hikes #CutWaste	Tax dollars at work: govt paid $1.2 mil to study seniors playing World of Warcraft http://t.co/YVJvGUA1 Instead of tax hikes #CutWaste
125	140.81	For 40 years the Hyde Amendment has told women that USA laws & rights don't apply equally to us. We must #BeBoldEndHyde	For 4 decades, the Hyde Amendment has told women that our laws + rights don't apply equally to them. It is time to #BeBoldEndHyde.
95	110.60	Our thoughts and prayers go out to the families and victims of the devastating shooting yesterday in Kansas City,... http://t.co/mSarFd3zJ4	My thoughts and prayers go out to the family and friends of the victims of the tragic school shooting in Connecticut this morning.
105	120.50	More than 88,000 people in Illinois lost unemployment insurance. Its unacceptable that House GOP refuse to allow a vote to #RenewUI.	Since Dec. 62,915 people in #Massachusetts have lost unemployment insurance because House GOP refuses to allow a vote to #RenewUI.
110	125.47	Our thoughts & prayers are with the friends & family of Officer Jacai Colson in the midst of this tragedy. https://t.co/fTpPx92ON7	My thoughts & prayers are with the family & friends of Officer Jacai Colson, who was killed in the line of duty today, & @PGPDNews family.
105	120.40	My deepest condolences go out to the families and friends of the victims of this mornings tragic shooting at the #DCNavyYard.	My thoughts and prayers go out to the family and friends of the victims of the tragic school shooting in Connecticut this morning.
105	120.34	You have 9 days left to get your submissions in for this years Congressional Art Competition. Details here –> http://t.co/nHS1IEkHZq	#NY21 High School Students: You have just 3 days left to get your entries in for the Congressional Art Competition https://t.co/Fa6lNmF1Xs
170	185.26	A veto on the #FY16NDAA withholds support and resources our troops and our nation needs. Mr. President, it's up to you to #SignTheBill.	A veto on this bill withholds support and resources that our troops and our nation need. Mr. President, it's up to you to #SignTheBill
95	110.19	May is Military Appreciation Month. Thank you to our men and women in uniform who defend freedom each day. VIDEO: http://t.co/xlEfFYhuzm	Thank you to all of our men and women in uniform who protect our freedom every day #ArmedForcesDay https://t.co/aBEK4xXho7
95	110.16	Happy Mothers Day to my mom Carmen and all mothers. Thank you for everything you do! https://t.co/HiHmUvZYdy	Wishing a Happy Mother's Day to my Mom and all of the amazing Moms today. Thank you for everything you do for us! https://t.co/LlEuEafpKM
115	130.04	My thoughts and prayers are with the victims and loved ones of this morning's tragic shooting in Orlando	My thoughts and prayers are with the victims and their loved ones of the horrific attack in Orlando.
110	124.89	My thoughts and prayers are with the people of Japan and those affected by the earthquake and tsunami. If you have... http://fb.me/TlBlz6hF	RT@SpeakerBoehner: Our thoughts and prayers are with the people of Japan and all those impacted by this mornings disaster.
110	124.89	My thoughts and prayers are with the people of Japan and those affected by the earthquake and tsunami. If you have... http://fb.me/TlBlz6hF	RT@SpeakerBoehner: Our thoughts and prayers are with the people of Japan and all those impacted by this mornings disaster.
105	119.72	The @HouseGOP decided to let @ExImBankUS shut down even though it puts American jobs at risk. RT to show your support for #ExIm4Jobs	.@HouseGOP decided to #EndExIm even though shutting down @ExImBankUS will put American jobs at risk. RT to show your support for #ExIm4Jobs.
105	119.72	The @HouseGOP decided to let @ExImBankUS shut down even though it puts American jobs at risk. RT to show your support for #ExIm4Jobs	.@HouseGOP decided to #EndExIm even though shutting down @ExImBankUS will put American jobs at risk. RT to show your support for #ExIm4Jobs.
150	164.68	From literacy tests to cutting early voting. From poll taxes to unnecessary voter IDs. Our county has more to do. #RestoreTheVRA #VRA50	From literacy tests to cutting early voting, poll taxes to unnecessary voter IDs We have more to do. #RestoreTheVRA http://t.co/KLt7RJjZ6F