Finding Similar Bills. Assessing the Validity of Text Reuse to Measure Policy Similarity

Much political science research relies on identifying the substantive similarity of bills. Especially in the study of policy diffusion identification of which bills are describing equivalent policy is crucial. Scholars of policy diffusion mostly rely on hand-coded datasets of equivalent bills. However, in order to study policy diffusion more broadly, for example by extending the study to multiple policy areas, large amounts of legislative text have would have to be analyzed manually. To find bills that describe the same or very similar policies very fine-grained analysis of the bills is necessary, because not only the exact topic, but also the same provisions have to be identified. Common text analysis algorithms such as topic models or bag-of-words based text similarity measures are therefore too coarse to identify policy similarity. However, legislators often directly use text from bills introduced in other legislatures or bill text provided by interest groups (model legislation) if they are drafting bills on policies that have been introduced in other states already. We use this fact, and propose to use text-sequencing algorithms to find such matching text between bills. We describe a new approach of how to apply sequence alignment algorithms to large amounts of legislative text and assess the validity of text reuse as a measure of substantive bill similarity. We demonstrate the validity of text reuse as a measure of policy similarity on a database of ca. 500,000 bills from US-state legislatures.

Sequencing algorithms namely the Smith-Waterman local alignment algorithm originate from DNA sequence matching and has been applied to text in the past, mostly for plagiarism detection (Su et al., 2008; Irving, 2004). In political science, it has been used to track the legislative history of congressional bills (Wilkerson et al., 2015). There are two major challenges for the application of the local alignment algorithm to detect policy similarity: First, the algorithm does an exhaustive search for all potential matches between two text sequences and it is therefore computationally unfeasible to calculate all pairwise similarities for large collections of bills. Second, a lot of text that is reused in bills is not of substantive interest. This boilerplate text, bill headers and footers and other procedural text, as well as reoccurring definitions or policy area specific phrases, is not indicative of bills having similar content. Wilkerson et al. (2015) tackled these problems by restrictively preselecting text that is analyzed with the local alignment algorithm and through selection of non-boilerplate text using supervised machine learning.

We propose a novel way to approach these problems. Our approach does not require the manual coding of text for the supervised classification of boilerplate text and is less restrictive in the pre-selection of bills to analyze in depth. Specifically, we apply a 3-step procedure that reduces the computational load as well as reducing the amount of boilerplate text that is detected by that algorithm. First, we preselect bills that are analyzed with the alignment algorithm with a measure related to cosine similarity. We remove all text with a tf-idf score lower than a certain threshold to avoid analyzing clear boilerplate text and then apply the algorithm to this reduced data.

The second part of our analysis then focuses on how well the alignment between bills captures substantive policy similarity. In a first step we apply topic models to the identified text alignments in order to check whether the reused text captures substantively relevant policy language, or boilerplate text. If a clear boilerplate topic emerges we can quantify how large a portion of the identified alignments is policy relevant compared to boilerplate.

Furthermore, we use a hand-coded dataset of equivalent bills, to investigate how well the local alignment algorithm predicts actual policy equivalence. In order to evaluate the validity of the text reuse measure on a basis broader then would be possible with hand-coded ground truth datasets, we assess the predictive validity of the measure. We do so in two ways: First, we asses how the measure performs in predicting policy diffusion ties between states as established by Desmarais et al. (2015). If two states have a strong diffusion tie, higher numbers of aligned bills would be expected. And second, we assess how well the similarity measure on the bill level predicts the ideological distance between the bills' sponsors using state legislators' ideal points (Shor and McCarty, 2011). That is, if two bills are sponsored by ideologically close legislators, we expect more text alignment than between bills of ideologically close legislators.

Our analyses show that the local alignment algorithm in combination with our pre and post processing procedure produces valid measures of policy similarity between bills. We aim to publish a dyadic dataset of similarity measures in order to allow a broad range of political scientists to use them for substantive research on US state politics.

References

Desmarais, B. a., J. J. Harden, and F. J. Boehmke (2015, 5). Persistent policy pathways: Inferring diffusion networks in the american states. *American Political Science Review 109*, 392–406.

Irving, R. (2004). Plagiarism and collusion detection using the smith-waterman algorithm. Technical report, University of Glasgow, Department of Computing Science.

Shor, B. and N. McCarty (2011, 8). The ideological mapping of american legislatures. *American Political Science Review 105*, 530–551.

Su, Z., B.-R. Ahn, K.-Y. Eom, M.-K. Kang, J.-P. Kim, and M.-K. Kim (2008). Plagiarism detection using the levenshtein distance and smith-waterman algorithm. In *Innovative Computing Information and Control, 2008. ICICIC'08. 3rd International Conference on*, pp. 569–569. IEEE.

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