Assessing Editing Patterns Across Document Versions

MATTHEW DENNY MONDAY 12TH DECEMBER, 2016

Many important political, social, and legal documents go through an extensive editing process before they are released: as laws, regulations, contracts, articles, books, etc. While the public often only sees the final version of these documents, many of the most important passages may not have been present in the original version, and were only introduced later in the process (Wilkerson et al., 2015). Understanding the editing process can therefore illuminate a great deal about the shifting priorities and power dynamics which shape the final versions of these documents. When available, first hand accounts of the bargaining process behind a particular set of edits can provide a rich window on these dynamics. However, such accounts are not widely available, and may not facilitate valid comparisons across cases. In this study, we introduce a set of computational measures which characterize the way documents are edited between versions, allowing a researcher to understand important dimensions of the editing process without having to read the document versions in detail.

There are a number of salient dimensions to editing that we seek to capture in this study. A researcher might want to know about the *granularity* of the edits. Where they fine grained edits such as changing individual words, striking or adding sentences, etc., or coarser edits such as the re-arranging of sections, of the addition or subtraction of sections/paragraphs or, some combination of both. Thy might also want to know about the overall *scope* of the edits —overall, how different are the two versions? The researcher might also like to know whether these edits were "on topic", or whether new topical content was added (adding to the vocabulary), or the topical content of the the bill was reduced (the vocabulary decreased in size). Along these lines, the researcher might like to know more generally whether the document grew or shrank in length, and what was added and removed. Finally, the researcher might want to know about the distribution of edits in the document: are they mostly at the beginning, the end, interspersed, and do they follow a consistent spatial pattern within the text? Together, a set of metrics that describe these dimensions of editing should provide a rich yet parsimonious characterization of the editing process.

There are essentially two ways to think about differences between document versions: one is in terms of overall text similarity (at multiple levels of granularity) which is described in Section 1. We approach this problem using ensembles of Dice coefficients (a measure of string similarity) to characterize the structure of edits made between two versions of a document. The second is to think in terms of the distribution of n-grams that match an n-gram across document versions (see Section 2). Each approach has its own advantages, but we illustrate how the sequence based approach provides a more natural and flexible characterization of editing process.

All analyses were performed on the corpus composed of all versions of all bills introduced in the United States Congress between 1993 and 2014 (Handler et al., 2016). The examples used in this study are currently being drawn from a subsample of bills introduced durring the 111th Congress (2009-2011), but will eventually be expanded to include all bills in the corpus. The plain text of each bill version was retrieved from the congress.gov website and minimal preprocessing was applied (some special characters and html markup were removed). All analyses were then performed in R, using the SpeedReader package.

1 Capturing Edit Patterns with an Ensemble of Dice Coefficients

A potentially valuable tool for characterizing the nature of edits between two version of a document is the Sørensen–Dice coefficient. The Dice coefficient is a measure of similarity between sets, and has been widely used in the NLP and information retrieval literature as a measure of string similarity (O'Connor, 2013; Manning and Schütze, 1999). Let n_t be the number of unique character bigrams (e.g. "bl", or "fa") shared in common by two strings x and y, and let n_x and n_y be the number of unique character bigrams in each of two strings. Then the Dice coefficient d, for these two strings is defined as:

$$d = \frac{2n_t}{n_x + n_y} \tag{1}$$

d takes it maximum value ($d = 1$) when the two strings share the same set of unique character bigrams, and its minimum value ($d = 0$) when those strings share no character bigrams in common. This measure can easily be generalized to operate on token bigrams (word pairs such as "national defense") by redefining n_t, n_x and n_y appropriately. d therefore provides a fast to calculate and relatively straight forward to evaluate measure of document similarity, when defined over token bigrams.

If we treat documents as long token vectors, we can readily calculate a Dice coefficient for any pair of documents in $\mathcal{O}(N_x N_y)$ time, where N_x, N_y are the number of tokens in documents x and y respectively. While such a measure is valuable for evaluating how a document changes from version to version, it only captures the nature of edits made at one level of granularity (token bigrams). It could be the case that a nefarious editor managed to rearrange all of the tokens in the document while preserving (close to) the same set of unique bigrams. Thus, our bigram based measure of document similarity might underestimate the magnitude of the edits made to the document. Therefore, a logical extension of the bigram Dice coefficient is to consider n-grams of varying length as the basis for our calculation.

In this study, we calculate Dice coefficients using n-grams of length $k \in \{1, ..., 50\}$ in order to capture how two documents differ at different scales. Thus for each document, we calculate $K = 50$ Dice coefficients, which we call $\mathbf{D}_K$, where each d_k is defined as:

$$d_k = \frac{2n_t^{(k)}}{n_x^{(k)} + n_y^{(k)}} \tag{2}$$

It is easy to verify that the Dice coefficient is non-increasing in the input n-gram length, but beyond this regularity, its behavior in n-gram size will be governed by the structure of edits between two document versions. Dice coefficients calculated on six congressional bill pairs using 1-50 grams as input are displayed in Figure 1. As we can see, there is a significant degree of variation in the edit structure.
More generally, we can find a correspondence between edit structure and the shapes of the curves in Figure 1. If the curve is flat and pegged at 1 (as with the two versions of 111-HR-1246), then we have clear evidence that the two bills are identical, as they share all 1-50 grams in common. Similarly, if a pair of bills was characterized by a flat curve pegged at 0, then it would be clear that the two versions were entirely unrelated (meaning they shared no vocabulary in common, and no higher order structures —as a logical consequence). If we now consider the case of 111-S-2812, we see that the curve is roughly linear and decreasing. This pattern corresponds to the insertion or deletion of a several smaller blocks of text, as we would expect to see such an action shift down the intercept, and have a roughly linear effect on the Dice coefficients as we increase the n-gram size.

Alternatively, we might observe a relatively low Dice coefficient that does not drop by much at all. This pattern would be associated with the insertion or deletion of one very large block of text that is mostly unrelated to the original substance of the bill. This follows from the low Dice coefficient for unigrams (a measure of vocabulary similarity), which does not drop off by much (indicating that the edits were

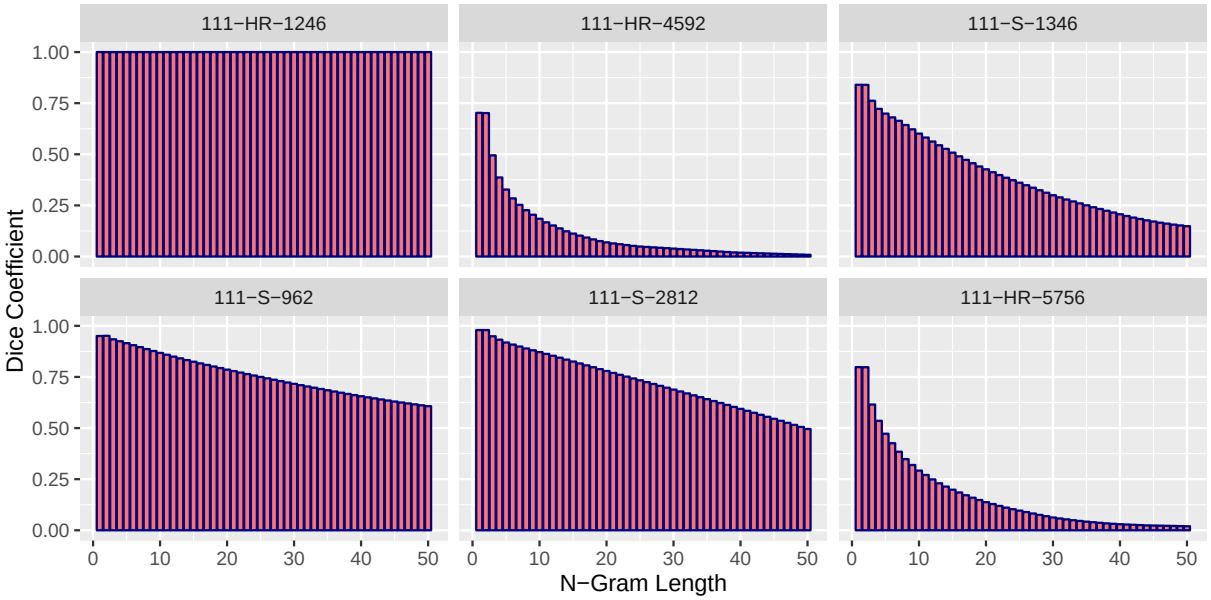

Figure 1: Example ensemble Dice coefficient plots for six pairs of congressional bills (two versions of the same bill for each plot).

concentrated in one area so their effects were roughly the same on smaller and larger n-grams). In contrast, we have bill pairs like 111-HR-4592, where we see a precipitous and non-linear drop off in the Dice coefficients as the n-gram size increases. This pattern is associated with finer grained edits that either replace most of the text wholesale with other text using the same unigram vocabulary, or alter the text at a token level.

Thus we can see that roughly two dimensions characterize these curves. The first is the general height, which is controlled at a high level by vocabulary similarity. We will revisit this issue in the next section. The second is the granularity of the edits.

1.1 Edit Granularity

We measure how granular the edits in a document are by looking at how quickly the Dice coefficients drop as we increase the n-gram size. The intuition for this approach is that the similarity (Dice coefficient) between a pair of document versions will decrease more rapidly (in increasing n-gram size) if the edits are more fine grained. This will happen because editing every other word in the document (for example) could theoretically preserve the vocabulary, but leave the second version with no 2-K-grams in common with the first. On the other end of the spectrum, if a block of text were inserted at the end of the document, it would have a much smoother effect on the Dice coefficients as we increase the n-gram size on which they are based. Let $\widetilde{\mathbf{M}} = \{\widetilde{m_1}, \widetilde{m_k}\}$ be a linear interpolation of d_1 and d_K, approximating the maximum value d_k could attain. It is still an open problem to calculate the exact theoretical maximum for d_k, but in our experiments we find $\widetilde{\mathbf{M}}$ be a very close approximation. We can then define the edit *granularity* between a pair of document versions, $G \in [0, 1]$ as:

$$G = \begin{cases} \sum_{k=2}^{K-1} \frac{\max(\widetilde{m_k} - d_k, 0)}{(K-2)(\widetilde{m_k} - d_K)} & \text{if } d_1 - d_K > 0 \\ 0 & \text{if } d_1 - d_K = 0 \end{cases} \tag{3}$$

This statistic takes on its maximum value for documents like 111-HR-4592, where there is a very precipitous drop off in Dice coefficients as n-gram length increases (reaching their minimum value quickly). In the case were $G = 1$, the Dice coefficients for 2 through K-grams would be identical (and less than the

unigram case), indicating a very fine-grained edit structure (changing individual words evenly spread throughout the document). On the other end of the spectrum $G = 0$ indicates that edits were only made in a single block (insertion, deletion, or replacement of a contiguous section of the document). Thus G provides a smooth continuum on which to evaluate the granularity of the edits made between two versions of a document.

1.2 Edit Scope

We measure the scope of the edits to the document as the additive inverse of the average of $\mathbf{D}_K$. The intuition here is simple. If all $d_k = 1$, then no editing has occurred, and thus the scope of the edits is zero. Conversely, if all $d_k = 0$, then the document has essentially been gutted and replaced with completely different text (the maximum possbile scope for an edit). Thus we define the edit *scope* between a pair of document versions, $S \in [0, 1]$ as:

$$S = 1 - \frac{\sum_{k=1}^{K} d_k}{K} \tag{4}$$

This measure is related to the edit granularity, except it does not account for the rate of decline in the dice coefficients, and instead captures the overall magnitude of the edits to the document. One of the key weaknesses of this measure is that it does not distinguish between edits due to additions, subtractions, or replacements of text blocks. Therefore, we introduce two measures of edit topicality in the next section which are designed to distinguish between these types of changes. However, edit scope does provide a simple measure to complement granularity in capturing the patterns of editing between two documents.

1.3 Edit Topicality: Additions, Subtractions and Replacements

While capturing the scope and granularity of edits made between two versions of a document is important to assessing the overall effects of these edits, these metrics provide very limited leverage on an equally important question: to what degree was the editor removing content, adding content, or simply replacing content?

Generally, we want to find the proportion of terms appearing in each version of the document that do not appear in the other version. Intuitively, the proportion of terms that appear in the original document, but do not appear in the edited version provides a measure of how much content has been removed from the original document. Similarly, the proportion of terms that appear in the edited version, but do not appear in the original document provides a measure of how much content has been added to the original document.

There are two possible ways to operationalize this intuition. Let the original version od the document be document X and the edited version be document Y. The first way to operationalize our intuition is to calculate the number of unique unigrams in X that are also in Y ($u(X) \in u(Y)$), and vice versa ($u(Y) \in u(X)$). We can then form a measure of topical subtraction editing as the additive inverse of the ratio of $u(X) \in u(Y)$ to the number of unique terms in X.

$$T^{(-)}_{\text{unigrams}} = 1 - \frac{||u(X) \in u(Y)||}{||u(X)||} \tag{5}$$

were $||\cdot||$ is the norm (or count) operator. $T^{(-)}_{\text{unigrams}}$ takes its maximum value (1) when X and Y share no unigrams in common, and its minimum value (1) when they share all unigrams in common. Similarly, we can define a measure of topical addition editing as the additive inverse of the ratio of $u(Y) \in u(X)$ to the number of unique terms in X.

$$T^{(+)}_{\text{unigrams}} = 1 - \frac{||u(Y) \in u(X)||}{||u(Y)||} \tag{6}$$

$T^{(+)}_{\text{unigrams}}$ has analogous properties to $T^{(-)}_{\text{unigrams}}$.

The draw of this method is that differences unigram vocabularies can arguably tell us important information about how the editing effects the topical content of a document. If a large proportion of the vocabulary from X is removed, then it is likely that the edits removed some of the topical content in X. Alternatively, if a large portion of the vocabulary in Y was not in X, then it is likely that the edits have introduced new topical content into the document. Both of these cases may also be true at the same time, indicating the topical content in X was removed and replaced with new content in Y.

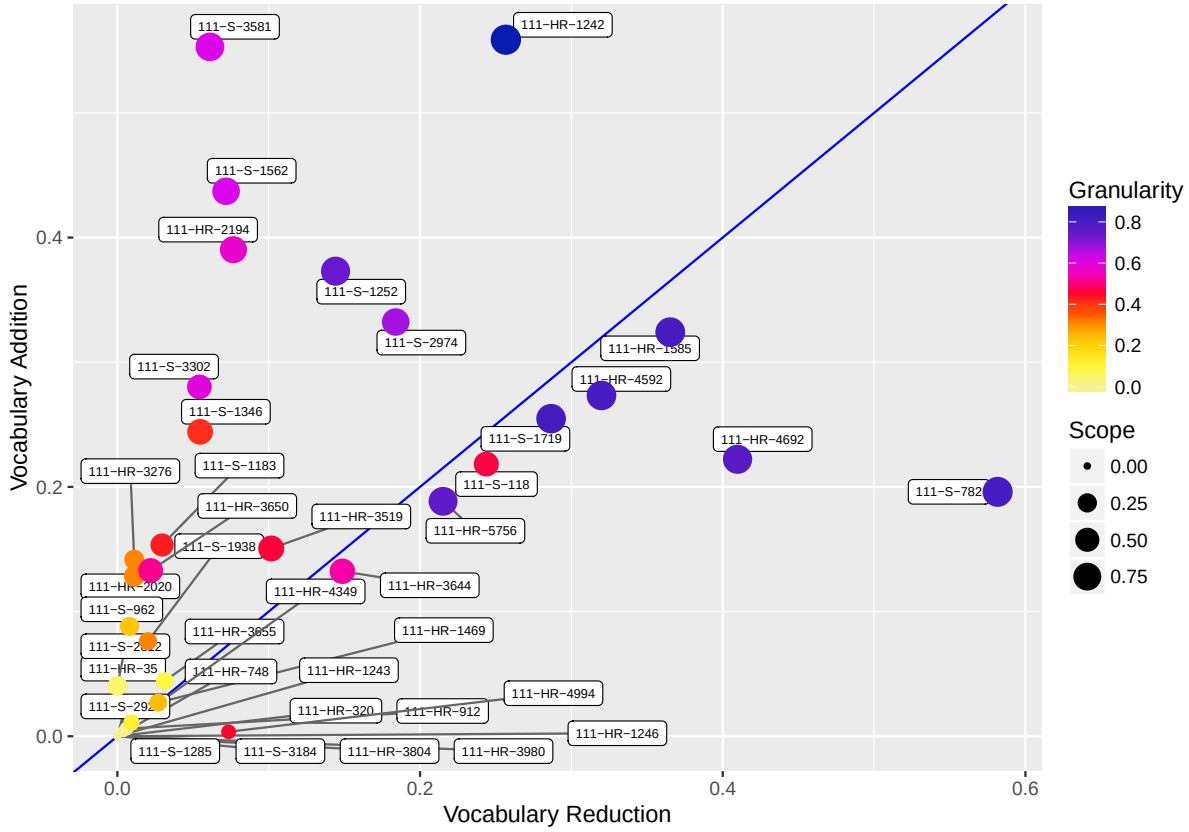

Figure 2: Scatter plot of vocabulary additions and subtractions (unigrams only) for 39 example bill pairs from the 111th Congress. Points are sized according to their edit scope, and colored according to their edit granularity.

We can plot unigram subtractions and additions, along with edit granularity and scope in a scatter plot. We do so for 40 example bill version pairs in Figure 2. We can see that in general, as the magnitude of the vocabulary additions and reductions increase, both the scope and the granularity of the edits increases as well. However there are some interesting cases of edits with very low scope that have a relatively high granularity (111-HR-4994, bottom of figure). These bills are likely edited in very specific places in the style of a book editor making corrections and cutting unnecessary bits fo text, as opposed to a bill like 111-HR-1242 (top middle of figure), which was essentially gutted and replaced with something else. These two examples highlight an inherent issue with this approach, which is that it has trouble distinguishing between edits that occur at a very fine scale and edits that simply replace the entire bill with a new one (both of which can lead to a high granularity score). The primary usefulness in this approach over the sequence based edit metrics described in the next section is that it allows us to characterize changes in topicality through changes in the vocabulary, which a sequence-based approach does not do so easily.

2 Sequence Based Edit Characterization

An alternative approach to characterizing the structure of edits between documents is to focus on the distribution of n-grams that match an n-gram across document versions. Instead of leveraging ensembles of Dice coefficients, this approach focusses on sequences of overlapping n-grams. The intuition behind this approach is that we can gain leverage on the distribution and granularity of edits by seeing how they clump within documents. For example, we might see a document where most of the n-grams match between versions, but there are numerous short sequences of overlapping n-grams that do not match. This would be indicative of an editing process where individual words or phrases were changed here and there, thus affecting small clusters of contiguous n-grams. Alternatively, we might come across a document that has the same proportion of non-matching n-grams between versions, but now these non-matching n-grams occur all on one large contiguous sequence. Such a pattern would indicate to us that a large block of text was inserted or removed. The major advantage of this approach is that it can tell us both about where deletions were made in the original document, but also where additions were made in the edited version, and provides the basis for a nearly infinite set of statistics to be calculated on the sequences of matching and non-matching n-grams in the two versions, thus providing a large amount of information to characterize the structure of edits from a relatively simple foundation.

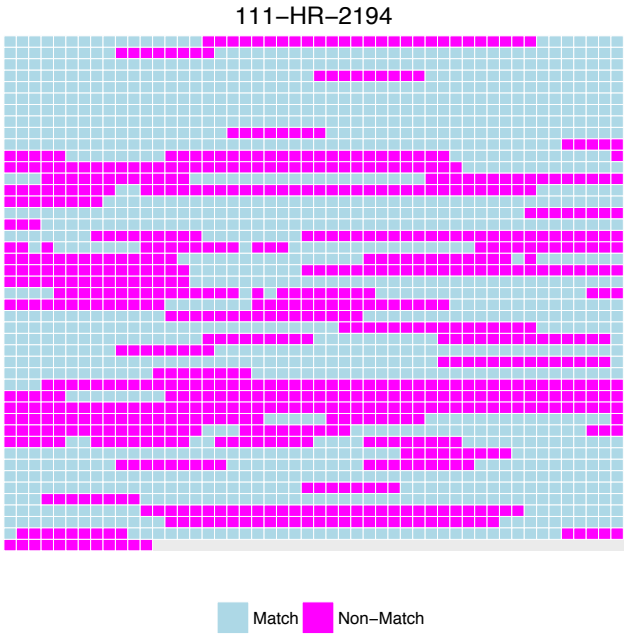

Figure 3: The sequence of 10-grams in the first version of HR-2194 (11th Congress) generated using a rolling window (1-10,2-11, etc.) colored by whether they exactly matched a 10-gram in the second version of the bill. Sequence goes from left to right and then by rows

For each document version we can construct a sequence of n-grams of length k using a rolling window. For example, if we consider the case of 10-grams, we would construct them using the 1st-10th unigram tokens, 2nd-11th, 3rd-12th, and so on. Each of these 10-grams could then be compared against all 10-gram tokens in a different document version to look for matches, as in Figure 3. In the case illustrated in Figure 3, the non-matching sequences of 10-grams indicate changes or deletions from the first to the second version of the document. More specifically, the sequence matching process proceeds as follows:

1. We begin by selecting two document versions. We then select an n-gram size on which to compare these documents. Our goal in selecting he n-gram size is to choose one such that we limit the number of false positive matches, while maximizing our resolution. Intuitively, if we were to select unigrams as our measure of comparison, we would find instances where one unigram was present in one version but not present in another, and mark these as instances of an addition or deletion (depending

on the document version). However, taking such an approach would lead us to underestimate the prevalence of edits in these documents, as the same set of unigrams can be combined in different sequences to produce different meanings. Thus, we need to use a longer n-gram length in order to reduce the probability of false positive matches between document versions. On the other hand, if we select an n-gram length that is too long (for example, 50-grams), then we may miss shorter sequences of text that stayed the same between versions because edits were made at a granularity that was below the resolution of the n-gram size. For example, if we immagine a document where one word was changed every 49 words, then a 50-gram measure would find no matches between versions, even though the vast majority of the text did in fact match. This leads us to select for a moderate n-gram length (in the neighborhood of 7-10 grams) to balance resolution and the false positive rate.

Future work could instead take an approach where multiple n-gram lengths are combined to eliminate the tradeoff between resolution and false postive rate, but such an approach would be exceedingly computationally expensive, and is thus put asside in the current investigation. One other factor that can affect the appropriate n-gram length is the choice of whether to remove domain stopwords before performing the analysis. If domain stopwords are removed, the sequence based approach may still be applied so long as the order of non-stopwords is preserved. However, the effective n-gram length is thus increased, so a shorter actual n-gram length may be selected. In the application in this study, an n-gram length of 5 was selected.

2. Having selected an n-gram length on which to compare document versions, the raw text of each document is read in and tokenized. At this point, domain stopwords and other special characters or markup may be removed, as long as the sequence of the remaining tokens is preserved. Next, for each document version a moving window of length k is used to construct overlapping n-grams. These n-grams will then be compared against each other.

3. Having constructed sequences of overlapping n-grams of length k for each document version, comparisons between these sequences are performed as follows. For each n-gram in each versions' sequence, we look for a match between that n-gram and any n-gram in the other version. If we find a match we record it a vector indicating matches/non-matches for each sequence. The reason we compare each n-gram against all n-grams in the other document version is to account for the possibility that the particular part of the text may be been moved between versions, or that text before or after the current segment of text in the current version may have been added or deleted. The obvious issue with this approach is that we might double count matches. For example if the n-gram "increase funding for this program by 1% over the next two years" occurs ten times in the first version of a document but only once in the second version, we would count all ten occurrences in the first version as matches, thus overestimating the degree of matching. To some extent this issue is ameliorated by selecting a long enough n-gram length, as it reduces the likelihood of such false positive based on two short and generic an n-gram. On the other hand, we do not expect that large block of text will literally be repeated verbatim in a document version, so the removal of nine instances of the above phrase will also likely be accompanied by the removal of the text around that phrase, the signalling a edit. One potential solution to this problem would be to adopt a rule such that each time an n-gram is matched from one version to the other, it is removed from the second version to prevent multiple matches. Of course, this might also lead to bias because the true match for an n-gram in the second version could have been with its second occurence in the first version, thus giving an incorrect characterization of the location of edits between versions. All things considered, we feel that simply matching against all n-grams in the other version is most appropriate.

4. Once the vectors indicating matches and non-matches in the sequences of overlapping n-grams for each document version are constructed, we can then calculate a number of sufficient statistics on these vectors and use these statistics to characterize the structure of the edits.

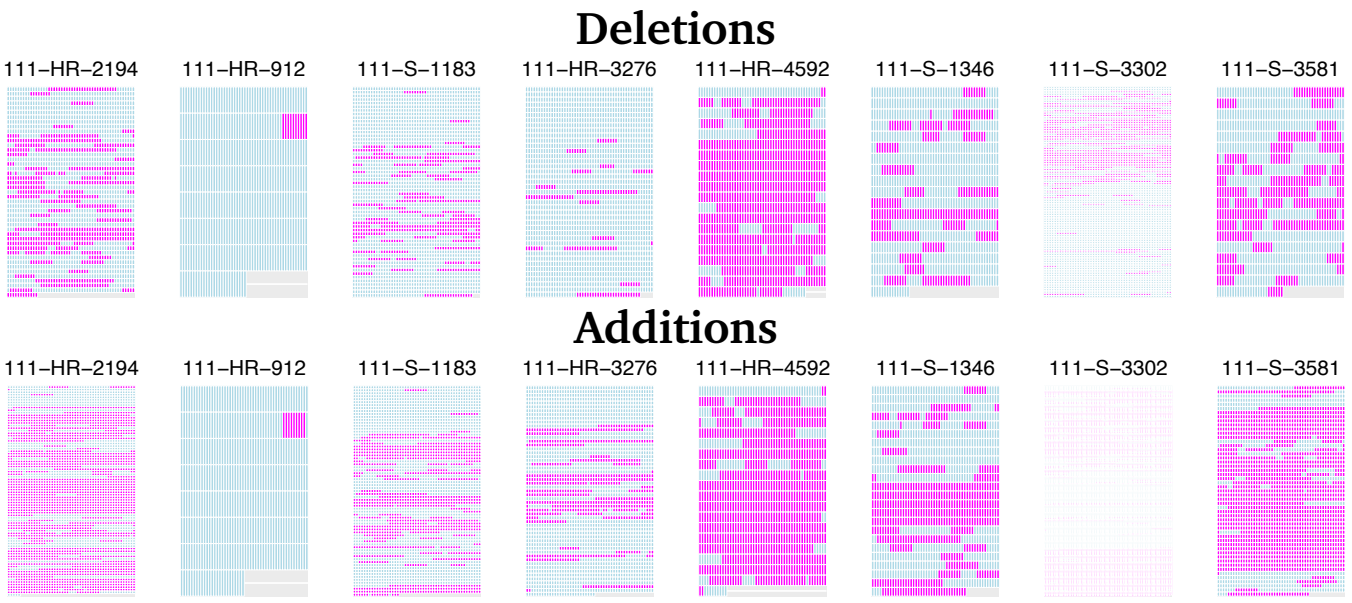

Figure 4: Example deletions and additions to 8 congressional bills using a 10-gram rolling window. Each cell is colored by whether it exactly matched a 10-gram in the other version of the bill, with blue cells indicating a match and magenta cells indicating no match. Sequences go from left to right and then by rows.

Examples of sequences of matching and non-matching sequences of n-grams in earlier (top) and later (bottom) versions of 8 congressional bills are illustrated in Figure 4. These sequences were constructed using 10-grams, and without removing any stopwords from the text. Each cell represents an n-gram, matches are colored light blue and non-matches are colored in magenta. As we can see, some bills like 111-HR-912 only had minor edits made to a specific part of the text, as is evidenced by the short sequences of non-matches between versions that occur in the same place in each version. However, other bills, such as 111-HR-4592 are subject to much more extensive editing that spans almost the entire length of the document. On the other hand, some bills (like 111-S-3581) have rather targeted deletions throughout the first version, but are then subject to very large block additions in the second version. Inspection of these plots can thus provide a very nuanced graphical summary of the structure of edits between documents.

2.1 Summarizing Edit Sequences

While graphical inspection may be a valuable tool for understanding the structure of edits between two version of a document, it is not well suited to characterizing the structure of edits across an entire corpus. Therefore, we introduce four summary measures on which to compare documents. These measures are plotted for 40 example congressional bill version pairs in Figure 5.

1. **Scope of Deletions**: the scope of deletions is just the proportion of n-grams in the first version of the document that do not match an n-gram in the second version. This gives us a measure of how much of the original text survived to the second version fo the text.

2. **Scope of Additions**: the scope of additions is similar to the scope of deletions, but is instead defined as the proportion of n-grams in the second version of the document that do not match an n-gram in the first version. This gives us a measure of how much of the edited version of the text was not present in the original version of the text.

3. **Average Edit Size**: the average edit size is defined as the average length of sequences of non-matching n-grams across both versions of a document. The shorter the average edits size, the shorter the length of edits made to a document. We do not normalize this measure by document length because this length has a natural interpretation. If we were to change one unigram in a document between versions to a unigram that did not exist in the first version, then we will generate a number

of mismatches that is equal to the n-gram length. For example, with an n-gram length of 10, changing one unigram would result in ten mismatched 10-grams in each version. Then for each contiguous token we add to the sequence of mismatched text, we increase the number mismatches by one. Thus sequence of mismatches of length 60 would indicate a change of roughly 50 tokens. In this way, we can gain a sense of whether the average edit was around the size of single word, a sentence, or even an entire paragraph.

4. **Proportion of Possible Changes:** this metric captures the number of unique sequences of mismatches in both versions of the text relative to the maximum number of sequences of mismatches possible given the n-gram length used for comparison. Thus is can range between zero, indicating that no changes were made at all, to one, indicating that a maximal number of unigram changes were made to the document (relative to the resolution limit implied by the choice of n-gram size). The numerator is just the count of distinct sequences of mismatches across both versions, and the denominator is the number of n-grams in both versions divided by one plus the n-gram length. This accounts for editing of one word every k + 1 tokens, resulting in k mismatches followed by one match and then another block of k mismatches and so on.

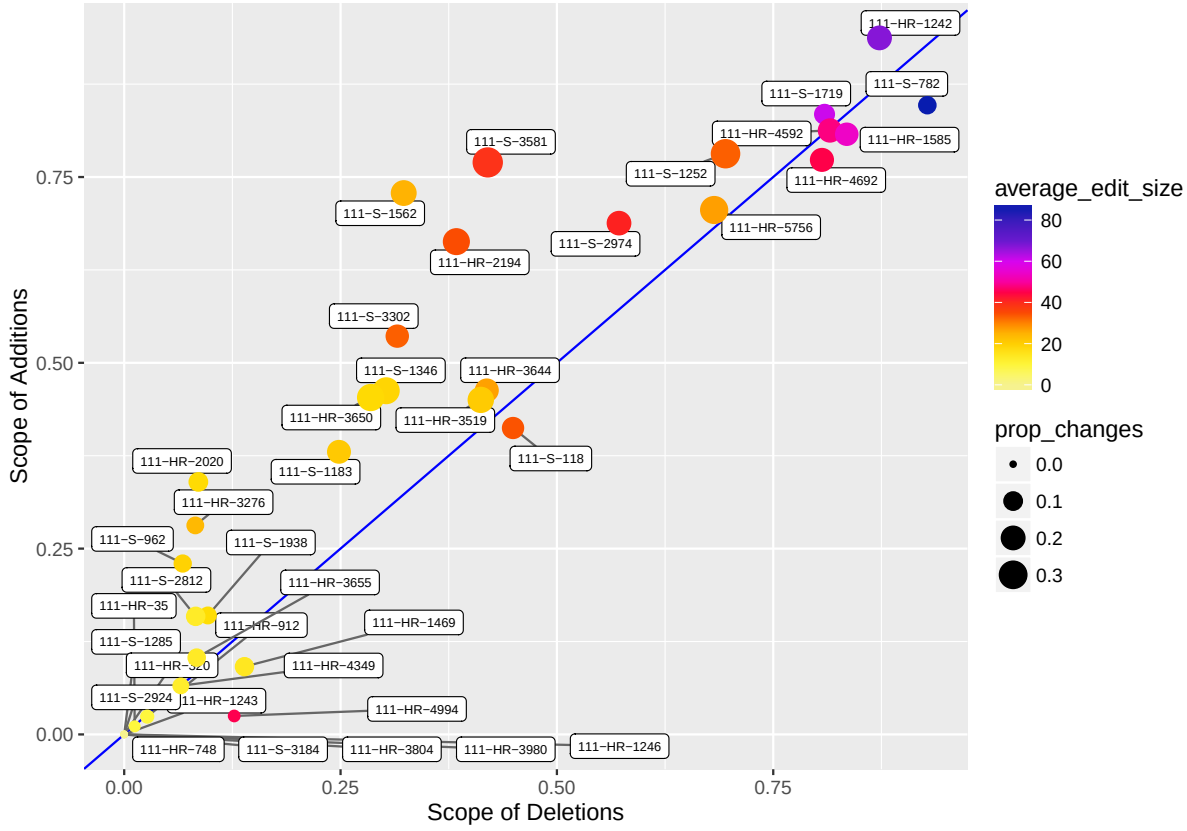

Figure 5: Scatter plot of the scope of additions and deletions for 39 example bill pairs from the 111th Congress. Points are sized according to the proportion of the maximum possible non-contiguous edits that were made to the document (with a higher proportion indicating more granular edits), and colored according to the average edit length (the length of non-matching sequences).

There are of course a great deal more potential statistics we might calculate on the sequences of matches and mismatches between documents to characterize the editing process, yet the four measures described above provide a good overall picture of the editing process. The scope of deletions and additions tells us about the degree to which content was edited out and added in to the document respectively, and their ratio can tell us whether the document on a whole got bigger or smaller. Together, these two measures also tell us about the overall scope of the edits, and give at least some information about the

degree to which the edits stayed on topic or were off topic. The average edit size and proportion of possible changes combine to give us information about the granularity of edits to the document. The larger the proportion of possible changes, the more granular the edits, conditional on the the average edit size, and vice versa. The average edit size may be a better overall summary of the edit granularity, but it is sensitive to document length and can be swayed by a single huge edit, so it is likely better to rely on a combination of these measures.

Figure 5 plots deletion scope against addition scope, with dots size by the proportion of possible changes and colored by the average edit size. The most striking feature of this plot is that the edits made to the sample of example bill pairs tend to be more heavily focussed on additions as opposed to deletions. This make some sense if we think about the process of logrolling in Congress whereby legislators support legislation they do not care so much about in exchange for support on legislation they do care about. The general preference for additions may just be a consequence of legislators adding unrelated content to a bill in order to secure support from a colleague. We also see that the scope of additions and deletions is generally positively related to the average edit size, indicating that most of the large edits to bills tend to take the form of block insertions and deletions.

One other potentially relevant statistic to calculate on the edit sequences might be the variance in the average mismatching n-gram sequence length. A high variance would indicate a more heterogeneous editing process where small and large changes are combined, thus potentially giving some indication about the combination of motives involved in the editing process. If document tend to be characterized by high variance edit length, then the median edit length may also be more appropriate than the mean for understanding the granularity of edits to document. It is important to note that none of the summary statistics proposed above take into account the lengths or distributions of matching sequences of n-grams. Similar statistics could be defined based on matching sequences that might reveal further relevant dimensions of the editing process. For example, the ratio of average edit length to the average length of n-gram sequences that were present in both documents might give further information about how the editors chose to preserve text. These alternate statistics should be explored in greater detail in future work.

3 Edits To Legislation In The 111th Congress

It is exceedingly rare that a piece of legislation will be introduced in the United States Congress, passed by both chambers, and signed into law without any changes being made along the way. The nature of these changes reflects the power dynamics and strategic interactions between the different branches of government, chambers of congress, parties, and even individual members themselves. Therefore it is a domain where characterizing the structure of edits is likely to be valuable. In this pilot analysis, we focus on edits made between different version of bills introduced durring the 11th Congress. There were a total of 5,970 bill version pairs produced across both chambers durring this period, which we examine in this analysis. This of course excludes the thousands of bills for which no additional versions were recorded (because they died in committee), and does not account for the similarities that might exist between different bills (parts of a failed bill were later included in a different bill), but does allow us to focus on the bill as a vehicle, and how its content changes as it progresses through the legislative process.

Figure 6 plots deletion scope against addition scope, with dots size by the proportion of possible changes and colored by the log of the average edit size, for all bill version pairs from the 111th Congress. The log of average edit size was utilized because a small number of bills had an extremely large average edit size (approximately 3,000). The measures were also calculated on tokenized text with a list of standard english stopwords removed, as well as all special characters. Because of the relatively heavy preprocessing, we used an n-gram length of 5 for this analysis. Inspection of the figure reveals a great deal of variation in the characteristics of edits between document versions, with some edits strictly removing or adding

single blocks of text, all the way through some documents where the editors made numerous small edits throughout the document. There are also a number of cases where the bill did not change at all, or a bill was completely replaced with unrelated text as part of the editing process. Each of these cases illustrates an importan power dynamic in Congress.

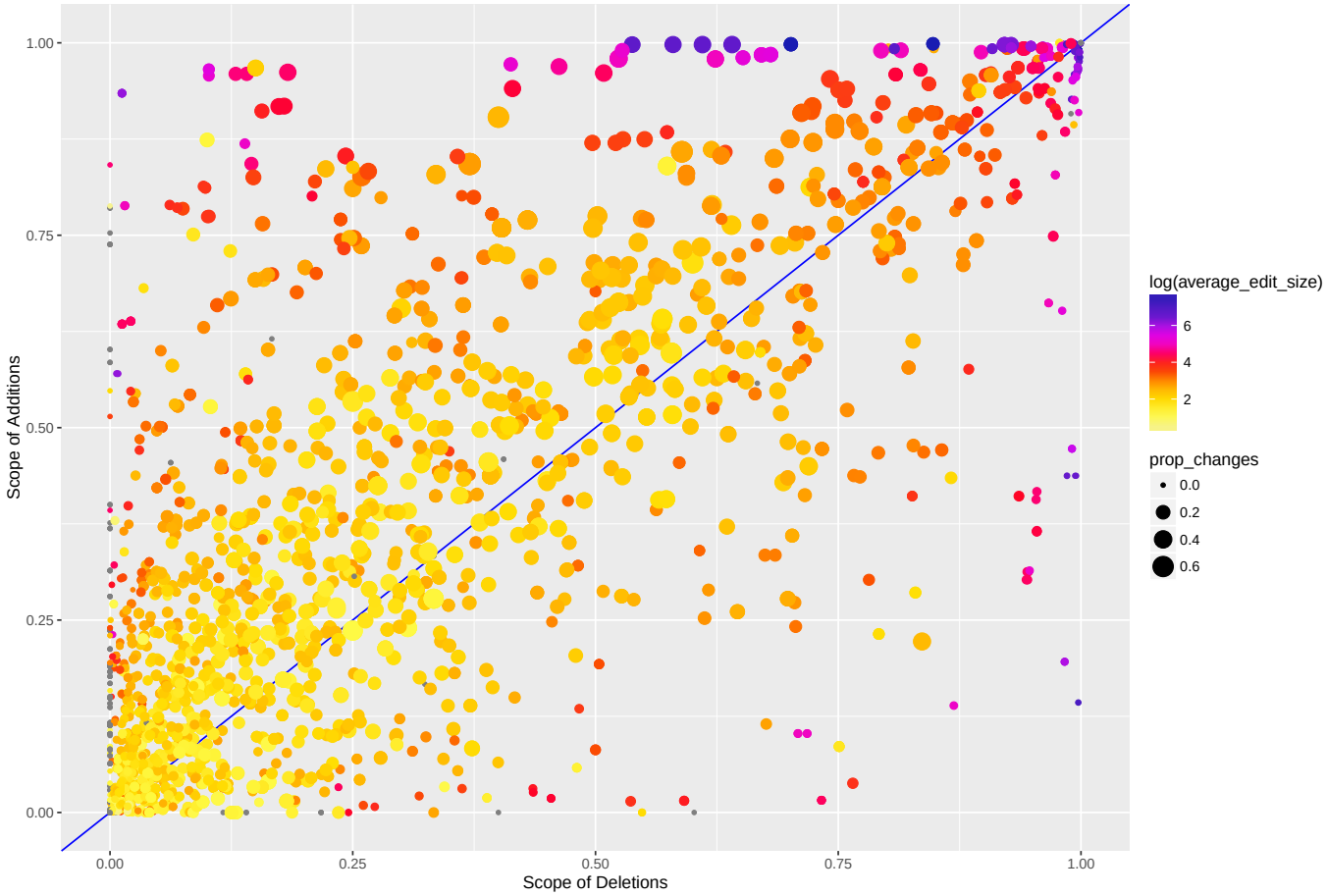

Figure 6: Scatter plot of the scope of additions and deletions for all bill pairs from the 111th Congress. Points are sized according to the proportion of the maximum possible non-contiguous edits that were made to the document, and colored according to the natural log of the average edit length.

Along these lines, we can disaggregate the bill version pairs to understand how edits are made to these bills at different stages in the legislative process. Figure 7 illustrates three examples of the scope of additions and deletions for bill pairs from the 111th Congress, broken down by the versions of each bill in the pair. These Plots record changes as a bill goes from the version that was introduced in the House of Representative, through versions that were passed by the House, the Senate and finally sent to the President's desk. IH → EAS records changes between the version of a bill introduced in the House (IH) and passed by the Senate (EAS). IH → EH records changes between the version of a bill introduced in the House and passed by the House (EH). IH → ENR records changes between the version of a bill introduced in the House and the version that is sent to the president's desk for a signature. Starting with IH → EH, we see that both bills that received lots of deletions and few additions, and vice versa were advanced to passage in the chamber. However, when we look at IH → EAS edits, we see that those bills which also pass the Senate are primarily edited through additions to the bill as opposed to deletions. However, once we look at bills which were then re-passed in identical form by both chambers and sent to the president's desk (IH → ENR), we see more of an emphasis on deletions than for those bills passed by the Senate. Therefore, we might draw the preliminary conclusion that the House is more focussed on removing content and reducing the overall scope of legislation, while the Senate is more focussed on

expanding the scope of legislation.

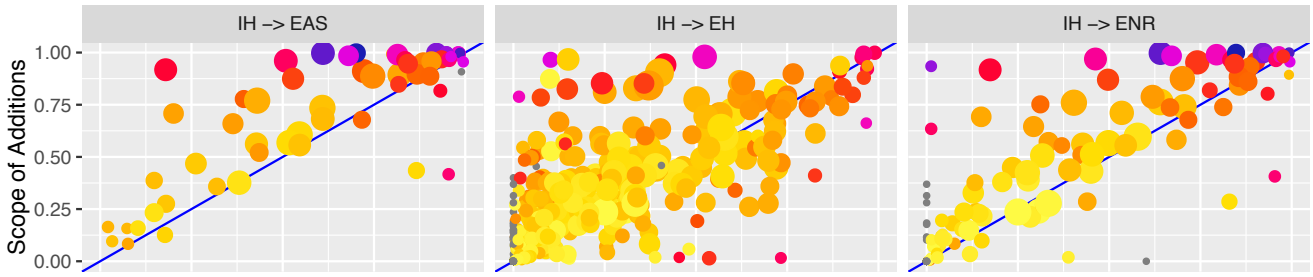

Figure 7: Scatter plots of the scope of additions and deletions for bill pairs from the 111th Congress, broken down by the pairing of bill versions. IH → EAS records changes between the version of a bill introduced in the House (IH) and passed by the Senate (EAS). IH → EH records changes between the version of a bill introduced in the House and passed by the House (EH). IH → ENR records changes between the version of a bill introduced in the House and the version that is sent to the president's desk for a signature. Points are sized according to the proportion of the maximum possible non-contiguous edits that were made to the document, and colored according to the natural log of the average edit length.

4 Conclusion

This study introduces two classes of metrics for characterizing the nature of the editing process between versions of a document. One of these metrics is primarily based on the similarity between text segments in document version at multiple scales, while the other is primarily focussed on the structure of sequences of mismatches between documents. While each approach has its merits, the sequence based approach is generally preferred based on the analyses conducted in this study and its reduced computational complexity relative to the Dice based approach. In the future, we intend to extend this study to look at a broader range of legislation, and to conduct further ground truth validations of the measures we derive here. The measures could also be profitably applied across a number of other domains which we intend to explore in future work.

References

Handler, Abram, Matthew J. Denny, Hanna Wallach, and Brendan O'Connor. Bag of What? Simple Noun Phrase Extraction for Text Analysis. In *Proceedings of the Workshop on Natural Language Processing and Computational Social Science at the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016. https://brenocon.com/handler2016phrases.pdf.

Manning, Christopher D. and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999. http://nlp.stanford.edu/fsnlp/.

O'Connor, Brendan. Learning Frames from Text with an Unsupervised Latent Variable Model. 2013. http://arxiv.org/abs/1307.7382.

Wilkerson, John, David Smith, and Nicholas Stramp. Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach. *American Journal of Political Science*, 59(4):943–956, 2015.