BillSum: A Corpus for Automatic Summarization of US Legislation

Anastassia Kornilova  
FiscalNote Research  
Washington, DC  
anastassia@fiscalnote.com

Vlad Eidelman  
FiscalNote Research  
Washington, DC  
vlad@fiscalnote.com

Abstract

Automatic summarization methods have been studied on a variety of domains, including news and scientific articles. Yet, legislation has not previously been considered for this task, despite US Congress and state governments releasing tens of thousands of bills every year. In this paper, we introduce BillSum, the first dataset for summarization of US Congressional and California state bills (https://github.com/FiscalNote/BillSum). We explain the properties of the dataset that make it more challenging to process than other domains. Then, we benchmark extractive methods that consider neural sentence representations and traditional contextual features. Finally, we demonstrate that models built on Congressional bills can be used to summarize California bills, thus, showing that methods developed on this dataset can transfer to states without human-written summaries.

1 Introduction

The growing number of publicly available documents produced in the legal domain has led political scientists, legal scholars, politicians, lawyers, and citizens alike to increasingly adopt computational tools to discover and digest relevant information. In the US Congress, over 10,000 bills are introduced each year, with state legislatures introducing tens of thousands of additional bills. Individuals need to quickly process them, but these documents are often long and technical, making it difficult to identify the key details. While each US bill comes with a human-written summary from the Congressional Research Service (CRS),¹ similar summaries are not available in most state and local legislatures.

¹http://www.loc.gov/crsinfo/

Automatic summarization methods aim to condense an input document into a shorter text while retaining the salient information of the original. To encourage research into automatic legislative summarization, we introduce the BillSum dataset, which contains a primary corpus of 33,422 US Congressional bills and reference summaries split into a train and a test set. Since the motivation for this task is to apply models to new legislatures, the corpus contains an additional test set of 1237 California bills and reference summaries. We establish several benchmarks and show that there is ample room for new methods that are better suited to summarize technical legislative language.

2 Background

Research into automatic summarization has been conducted in a variety of domains, such as news articles (Hermann et al., 2015), emails (Nenkova and Bagga, 2004), scientific papers (Teufel and Moens, 2002; Collins et al., 2017), and court proceedings (Grover et al., 2004; Saravanan et al., 2008; Kim et al., 2013). The later area is most similar to BillSum in terms of subject matter. However, the studies in that area either apply traditional domain-agnostic techniques or take advantage of the unique structures that are consistently present in legal proceedings (e.g precedent, law, background).²

While automatic summarization methods have not been applied to legislative text, previous works have used the text to automatically predict bill passage and legislators’ voting behavior (Gerrish and Blei, 2011; Yano et al., 2012; Eidelman et al., 2018; Kornilova et al., 2018). However, these studies treated the document as a “bag-of-words” and did not consider the importance of individual

²Kanapala et al. (2017) provide a comprehensive overview of the works in legal summarization.
sentences. Recently, documents from state governments have been subject to syntactic parsing for knowledge graph construction (Kalouli et al., 2018) and textual similarity analysis (Linder et al., 2018). Yet, to the best of our knowledge, BillSum is the first corpus designed, specifically for summarization of legislation.

3 Data

The BillSum dataset consists of three parts: US training bills, US test bills and California test bills. The US bills were collected from the Govinfo service provided by the United States Government Publishing Office (GPO). Our corpus consists of bills from the 103rd-115th (1993-2016) sessions of Congress. The data was randomly split into 28,408 train bills and 5014 test bills. For California, bills from the 2015-2016 session were scraped directly from the legislature’s website; the summaries were written by their Legislative Counsel.

The BillSum corpus focuses on mid-length legislation from 5000 to 20,000 character in length. We chose to measure the text length in characters, instead of words or sentences, because the texts have complex structure that makes it difficult to consistently measure words. The range was chosen because on one side, short bills introduce minor changes and do not require summaries. While the CRS produces summaries for them, they often contain most of the text of the bill. On the other side, very long legislation is often composed of several large sections. The summarization problem thus becomes more akin in its formulation to multi-document summarization, a more challenging task that we leave to future work. The resulting corpus includes about 28% of all US bills, with a majority of the removed bills being shorter than 5000 characters.

For the summaries, we chose a 2000 character limit as 90% of summaries are of this length or shorter; the limit here is, also, set in characters to be consistent with our document length cut-offs. The distribution of both text and summary lengths is shown in Figure 1. Interestingly, there is little correlation between the bill and human summary length, with most summaries ranging from 1000 to 2000 characters.

For a closer comparison to other datasets, Table 1 provides statistics on the number of words in the texts, after we simplify the structure of the texts.

![Figure 1: Bill Lengths](https://example.com/bill-lengths.png)

Stylistically, the BillSum dataset differs from other summarization corpora. Figure 2 presents an example Congressional bill. The nested, bulleted structure is common to most bills, where each bullet can represent a sentence or a phrase. Yet, content-wise, this is a straightforward example that states key details about the proposed grant in the outer bullets. In more challenging cases, the bill may state edits to an existing law, without whose context the change is hard to interpret, such as:

*Section 4 of the Endangered Species Act of 1973 (16 U.S.C. 1533) is amended in subsection (a) in paragraph (1), by inserting “with the consent of the Governor of each State in which the endangered species or threatened species is present”*

The average bill will contain both types of language, encouraging the study of both domain-specific and general summarization methods on this dataset.

4 Benchmark Methods

To establish benchmarks on summarization performance, we evaluate several extractive summarization approaches by first scoring individual sentences, then using a selection strategy to pick the best subset (Carbonell and Goldstein, 1998). While we briefly considered abstractive summarization models (Chopra et al., 2016), we found that the existing models trained on news and Wikipedia data produced ungrammatical results, and that the size of dataset is insufficient for the
Table 1: Text length distributions on preprocessed texts.

<table>
<thead>
<tr>
<th></th>
<th>Words</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>mean 1382</td>
<td>min 245</td>
</tr>
<tr>
<td></td>
<td>25th 922</td>
<td>50th 1254</td>
</tr>
<tr>
<td></td>
<td>75th 1773</td>
<td>max 8463</td>
</tr>
<tr>
<td>CA</td>
<td>mean 1684</td>
<td>min 561</td>
</tr>
<tr>
<td></td>
<td>25th 1123</td>
<td>50th 1498</td>
</tr>
<tr>
<td></td>
<td>75th 2113</td>
<td>max 3795</td>
</tr>
<tr>
<td>US</td>
<td>mean 46</td>
<td>min 7</td>
</tr>
<tr>
<td></td>
<td>25th 31</td>
<td>50th 41</td>
</tr>
<tr>
<td></td>
<td>75th 58</td>
<td>max 372</td>
</tr>
<tr>
<td>CA</td>
<td>mean 47</td>
<td>min 12</td>
</tr>
<tr>
<td></td>
<td>25th 31</td>
<td>50th 42</td>
</tr>
<tr>
<td></td>
<td>75th 59</td>
<td>max 137</td>
</tr>
</tbody>
</table>

50

5 As noted in Section 3, it is difficult to define sentence boundaries for this task due to the bulletted structure of the documents. We simplify the text with the following heuristic: if a bullet is shorter than 10 words, we treat it as a part of the previous sentence; otherwise, we treat it as a full sentence. This cut-off was chosen by manually analyzing a sample of sentences. A more sophisticated strategy would be to check if each bullet is a sentence fragment with a syntactic parser and a Rouge-2 Precision score of a sentence relative to the reference summary and simplify it to a binary value based on whether it is above or below 0.1 (Lin, 2004; Zopf et al., 2018). As an example, the sentences in the positive class are highlighted in green in Figure 2.

4.1 Document Context Model (DOC)

A good summary sentence contains the main ideas mentioned in the document. Thus, researchers have designed a multitude of features to capture this property. We evaluate how several common ones transfer to our task:

The position of the sentence can determine how informative the sentence is (Seki, 2002). We encode this feature as a fraction of 'sentence position / total sentence count', to restrict this feature to the 0−1 range regardless of the particular document's length. In addition, we include a binary feature for whether the sentence is near a section header.

An informative sentences will contain words that are important to a given document relative to others. Following a large percentage of previous works, we capture this property using TF-IDF features (Seki, 2002; Ramos et al., 2003). First, we calculate a document-level TF-IDF weight for each word, then take the average and the max of these weights for a sentence as features. To relate language between sentences, "sentence-level" TF-IDF features are created using each sentence as a document for the background corpus; the average and max of the sentence's word weights are used as features.

We train a random forest ensemble model over these features, including the three document-level and sentence-level features, with 50 estimators (Breiman, 2001).

We encode this feature as a fraction of the sentence / total sentence count’, to restrict this feature to the 0−1 range regardless of the particular document’s length. In addition, we include a binary feature for whether the sentence is near a section header.

An informative sentences will contain words that are important to a given document relative to others. Following a large percentage of previous works, we capture this property using TF-IDF features (Seki, 2002; Ramos et al., 2003). First, we calculate a document-level TF-IDF weight for each word, then take the average and the maximum of these weights for a sentence as features. To relate language between sentences, “sentence-level” TF-IDF features are created using each sentence as a document for the background corpus; the average and max of the sentence’s word weights are used as features.

We train a random forest ensemble model over these features with 50 estimators (Breiman, 2001). This method was chosen because it best then reconstruct full sentences; however, the former approach is sufficient for most documents.
captured the interactions between the small number of features.

4.2 Summary Language Model (SUM)

We hypothesize that certain language is more common in summaries than in bill texts. Specifically, that summaries primarily contain general effects of the bill (e.g. awarding a grant) while language detailing the administrative changes will only appear in the text (e.g. inserting or modifying relatively minor language to an existing statute). Thus, a good summary should contain only the major actions.

Hong and Nenkova (2014) quantify this aspect using hand-engineered features based on the likelihood of words appearing in summaries as opposed to the text. Later, Cao et al. (2015) built a Convolutional Neural Network (CNN) to predict if a sentence belongs in the summary and showed that this straightforward network outperforms engineered features. We follow their approach, using the BERT model as our classifier (Devlin et al., 2018). BERT can be adapted for and has achieved state-of-the-art performance on a number of NLP tasks, including binary sentiment classification.

To adapt the model to our domain, we pretrain the Bert-Large Uncased model on the “next-sentence prediction” task using the US training dataset for 20,000 steps with a batch size of 32. The pretraining strategy has been successfully applied to tune BERT for tasks in the biomedical domain (Lee et al., 2019). Using the pretrained model, the classification setup for BERT is trained on sentences and binary labels for 3 epochs over the training data.

4.3 Ensemble and Sentence Selection

To combine the signals from the DOC and SUM models, we create an ensemble averaging the two probability outputs.

Next, we applied our US model to CA bills. Overall, the performance is lower than on US bills

5 Results

To estimate the upper bound on our approach, an oracle summarizer is created by using the true Rouge-2 Precision scores with the MMR selection strategy. In addition, we evaluate the following unsupervised baselines: SumBasic (Nenkova and Vanderwende, 2005), Latent Semantic Analysis (LSA) (Gong and Liu, 2001) and TextRank (Mihalcea and Tarau, 2004). The final results are shown in Table 2. The Rouge F-Score is used because it considers both the completeness and conciseness of the summary method.

We evaluated the DOC, SUM, and ensemble classifiers separately. All three of our models outperform the other baselines, demonstrating that there is a “summary-like” signal in the language across bills. The SUM model outperforms the DOC model showing that a strong language model can capture general summary-like features; this result is in line with Cao et al. (2015) and Collins et al. (2017) sentence level neural network performance. However, in those studies incorporating several contextual features improved the performance, while DOC+SUM performs similarly to DOC. In future work we plan to incorporate contextual features into the neural network directly; Collins et al. (2017) showed that this strategy is effective for scientific article summarization. In addition, we plan to explore additional sentence selection strategies instead of always adding sentences to the 2000 character limit.

Next, we applied our US model to CA bills. Overall, the performance is lower than on US bills
(Table 2b), but all three supervised methods perform better than the unsupervised baselines, suggesting that models built using the language of US Bills can transfer to other states. Interestingly, DOC+SUM outperforms DOC in the CA dataset, suggesting that the BERT model may have overfit to the US language. An additional reason for the drop in the performance is the difference in the structure of the summaries: In California the provided summaries state not only the proposed changes, but the relevant pieces of the existing law, as well (see Appendix B.3 for a more in-depth discussion). We hypothesize that a model trained on multi-state data would transfer better, thus we plan to expand the dataset to include all twenty-three states with human-written summaries.

Table 2: ROUGE F-scores (%) of different methods.

<table>
<thead>
<tr>
<th></th>
<th>Rouge-1</th>
<th>Rouge-2</th>
<th>Rouge-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>47.97</td>
<td>32.14</td>
<td>40.89</td>
</tr>
<tr>
<td>SumBasic</td>
<td>30.56</td>
<td>15.33</td>
<td>23.75</td>
</tr>
<tr>
<td>LSA</td>
<td>32.24</td>
<td>14.02</td>
<td>23.75</td>
</tr>
<tr>
<td>TextRank</td>
<td>34.10</td>
<td>17.45</td>
<td>27.57</td>
</tr>
<tr>
<td>DOC</td>
<td>38.18</td>
<td>21.22</td>
<td>31.02</td>
</tr>
<tr>
<td>SUM</td>
<td>41.29</td>
<td>24.47</td>
<td>34.07</td>
</tr>
<tr>
<td>DOC + SUM</td>
<td>41.28</td>
<td>24.31</td>
<td>34.15</td>
</tr>
</tbody>
</table>

(a) Congressional Bills

<table>
<thead>
<tr>
<th></th>
<th>Rouge-1</th>
<th>Rouge-2</th>
<th>Rouge-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>51.24</td>
<td>34.86</td>
<td>45.91</td>
</tr>
<tr>
<td>SumBasic</td>
<td>35.47</td>
<td>16.16</td>
<td>30.10</td>
</tr>
<tr>
<td>LSA</td>
<td>35.05</td>
<td>16.34</td>
<td>30.10</td>
</tr>
<tr>
<td>TextRank</td>
<td>35.81</td>
<td>18.10</td>
<td>30.10</td>
</tr>
<tr>
<td>DOC</td>
<td>37.32</td>
<td>18.72</td>
<td>31.87</td>
</tr>
<tr>
<td>SUM</td>
<td>38.67</td>
<td>20.59</td>
<td>33.11</td>
</tr>
<tr>
<td>DOC + SUM</td>
<td>39.26</td>
<td>21.16</td>
<td>33.77</td>
</tr>
</tbody>
</table>

(b) CA Bills

5.1 Summary Language Analysis

The success of the SUM model suggests that certain language is more summary-like. Following a study by Hong and Nenkova (2014) on news summarization, we apply KL-divergence based metrics to quantify which words were more summary-like. The metrics are calculated by:

1. Calculate the probability of unigrams appearing in the bill text and in the summaries ($P_t(w)$ and $P_s(w)$ respectively).

2. Calculate $KL$ scores as: $KL(w|T) = P_s(w) \log \frac{P_s(w)}{P_t(w)}$ and the opposite.

A large value of $KL(S|T)$ indicates that the word is summary-like and $KL(T|S)$ indicates a text-like word. Table 3 shows the most summary-like and text-like words in bills and resolutions. For both document types, the summary-like words tend to be verbs or department names; the text-like words mostly refer to types of edits or background content (e.g., “reporting the rise of…”). This follows our intuition about summaries being more action driven. While a complex model, like BERT, may capture these signals internally; understanding the significant language explicitly is important both for interpretability and for guiding future models.

Table 3: Examples of summary and text like words

| Summary-like       | prohibit, DOD, VA, allow, penalty, prohibit, EPA, eliminate, implement, require |
| Text-like          | estimate, average, report, rise, section, finish, percent, debate |

6 Conclusion

In this paper, we introduced BillSum, the first corpus for legislative summarization. This is a challenging summarization dataset due to the technical nature and complex structure of the bills. We have established several baselines and demonstrated that there is a large gap in performance relative to the oracle, showing that the problem has ample room for further development. We have also shown that summarization methods trained on US Bills transfer to California bills - thus, the summarization methods developed on this dataset could be used for legislatures without human written summaries.

References


Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering docu-


Fridolin Linder, Bruce A. Desmarais, Matthew Burgess, and Eugenia Giraudy. 2018. Text as policy: Measuring policy similarity through bill text reuse. SSRN.


A  Additional ROUGE Scores

As discussed in the Results section, F-Scores encourage a balance between comprehensiveness and conciseness. However, as it is useful to analyze the precision and recall scores separately, both are presented in Table 4 for US Bills and in Table 5 for CA Bills. All tested methods favor recall, since they consistently generate a 2000 character summary, instead of stopping early when a concise summary may be sufficient. For both datasets, the difference in Recall between the Oracle and DOC+SUM summarizer is a lot smaller than for Recall; which suggests that a lot of useful summary content can be found with an extractive method. In future work, we will focus on extracting more granular snippets to improve precision.

B  Additional Bill Examples

We highlight several example bills to showcase the different types of bills found in the dataset.

B.1 Complex Structure Example

In the Data section, we discussed some of the challenges with processing bills: complex formatting and technical language. Figure 3 is an excerpt from a particularly difficult example:

- The text interleaves several layers of bullets. Lines 3, 15, 27 represent the same level (points (3) and (4) omitted for space); lines 16, 17, 19 and 21 go together, as well. These multiple levels need to be handled carefully,
Figure 3: US H.R.1680 (115th)

or the summarizer will extract snippets that cannot be interpreted without context.

- Lines 22-26 both introduce new language for the law and use the bulleted structure.

- Line 27 states that the existing “subsection (f)” is being removed and replaced. While lines 28 onward state the new text, the meaning of the change relative to the current text is not clear.

The human-written summary for this bill was:

(Sec. 4) “Women’s business center” shall mean a project conducted by any of the following eligible entities:

- a private nonprofit organization;
- a state, regional, or local economic development organization;
- a state-chartered development, credit, or finance corporation;
- a junior or community college; or
- any combination of these entities.

The SBA may award up to $250,000 of financial assistance to eligible entities per project year to conduct projects designed to provide training and counseling meeting the needs of women, especially socially and economically disadvantaged women.

Most of the relevant details are captured in the text between lines 8-14 and 20-24. For examples similar to this one, the summary language is extracted almost directly from the text, but, parsing them correctly from the original structure is a non-trivial task.

B.2 Paraphrase Example

For a subset of the bills, the CRS will paraphrase the technical language. In these cases, extractive summarization methods are particularly limited. Consider the example in Figure 4 and its summary:

This bill amends the Endangered Species Act of 1973 to revise the process by which the Department of the Interior or the Department of Commerce, as appropriate, reviews petitions to list a species on the endangered or threatened species list. Specifically, the bill establishes a process for the appropriate department to declare a petition backlog and discontinue the petitions when there is a backlog.

Figure 4: US H.R.6355 (115th)

While the bill elaborates of the “process”, the summary states that one was created. This type of summary would be hard to construct by a purely extractive method.

B.3 California Example

The California bills follow the same general patterns as US bills, but the format of some summaries is different. In Figure 5: the summary,
first, explains the existing law, then explains the change. The additional context is useful, and in the future we may build a system that references the existing law to create better summaries.

LEGISLATIVE COUNSEL'S DIGEST
Existing law requires that a bicycle operated during darkness upon a highway, a sidewalk where bicycle operation is not prohibited by the local jurisdiction, or a bikeway, as defined, be equipped with a red reflector on the rear that is visible from a distance of 500 feet to the rear when directly in front of lawful upper beams of headlamps on a motor vehicle. A violation of this requirement is an infraction.

This bill would instead require that a bicycle operated under those circumstances be equipped with a red reflector or a solid or flashing red light with a built-in reflector on the rear that is visible from a distance of 500 feet to the rear when directly in front of lawful upper beams of headlamps on a motor vehicle. By revising the definition of a crime, the bill would impose a state-mandated local program. The bill would also include a statement of legislative findings and declarations. The California Constitution requires the state to reimburse local agencies and school districts for certain costs mandated by the state. Statutory provisions establish procedures for making that reimbursement.

Figure 5: California Bill Summary