Segmentation of Rulemaking Documents for Public Notice-and-Comment Process Analysis

Anna Belova*
abelova@alumni.cmu.edu
Carnegie Mellon University
Pittsburgh, PA

Matthias Grabmair
mgrabmai@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA

Eric Nyberg
ehn@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA

ABSTRACT

We evaluate feasibility of automated identification of comment discussion passages and comment-driven proposed rule revisions in the US Environmental Protection Agency's (EPA's) rulemaking documents. We have annotated a dataset of final rule documents to identify all spans in which EPA discusses and evaluates the merits of public comments received on its proposed rules, and present lessons learned from the annotation process. We implement several baseline supervised discourse segmentation models that combine classic linear learners with sentence representations using handcrafted features as well as Bidirectional Encoder Representations from Transformers (BERT). We observe good agreement on annotation comment discussions and our models achieve a classification F1 of 0.73. Public comment dismissals and rule revisions are substantially harder to annotate and predict, leading to lower agreement and model performance. Our work contributes a dataset and a baseline for a novel discourse segmentation task of identifying public comment discussion and evaluation by the receiving agency.

1 INTRODUCTION

Government agencies are created by the legislatures worldwide to regulate social, economic, and political aspects of people's lives. These agencies belong to the executive branch of the government, yet they create legally enforceable regulations and rules that implement broad legislation. In the US, public notice-and-comment processes have become an important venue for influencing social and economic policy. In that, US agencies publish proposed rules in the Federal Register (FR) and all interested parties are given an opportunity to comment. Agency regulatory proposals receive public feedback from individuals, businesses, organized groups (of individuals or businesses), and other agencies. Comments represent heterogeneous interests in particular regulatory outcomes. The agency is not obliged to react to each individual received comment. However, it has to respond to comments that raise significant issues with the proposed rule and, if the points raised have merit, may substantively revise of the rulemaking document. The final rule document is published in the FR and contains the discussion of submitted comments, or points to other documents in the docket that address concerns raised in the comments.

The online forum for the US public notice-and-comment process—`regulations.gov`—was launched in January 2003, as part of the US eRulemaking program established as a cross-agency E-Gov initiative under Section 206 of the 2002 E-Government Act (H.R. 2458/S. 803). In this collection, all documents pertaining to the development of a particular rule are compiled in a regulatory docket. A typical docket contains a proposed rule document, many public comment documents, and a final/revised rule document.[1] As such, `regulations.gov` provides a testbed for study of the public notice-and-comment discourse in the US.

In this work, we focus on (1) identifying spans in the final rule documents that contain the agency's discussion of the public comments it received, and (2) classifying those spans as being either dismissals of the commenter claims or revisions of the proposed regulations prompted by the comment. In that, we analyze 353 US Environmental Protection Agency (EPA) regulations proposed in January 2003 or later, and finalized as of March 2018.[2]

Our work contributes a dataset[3] and a baseline for a novel discourse segmentation task of identifying public comment discussion and evaluation by the receiving agency. Automatic detection of comment discussion passages in the rulemaking documents could improve the efficiency of regulatory review conducted by experts at a number of organizations, including the US Office of Information and Regulatory Affairs, regulatory agencies, and other stakeholders of the regulatory process. In addition, segmentation of regulatory discourse is the first step bringing agency's narrative deliberations in the study of bureaucratic politics and decision making (e.g., regulatory capture theory) by economists and political scientists [37], which to date has relied on structured data generated by surveys and administrative record-keeping (e.g. permitting, inspections).

2 RELATED WORK

In the peer-reviewed literature, discussion of e-rulemaking benefits, challenges, and related artificial intelligence (AI) methods began in the early 2000s [9]. Over a decade later, surveys by [7] and [37] describe several e-rulemaking initiatives that involved successful applications of AI. One line of e-rulemaking research has focused on tasks relevant to management of massive amount of public comments received by agencies (e.g., [58], [31], [52]). Another line of

*Corresponding author

In: Proceedings of the Workshop on Artificial Intelligence and the Administrative State (AIAS 2019), June 17, 2019, Montreal, QC, Canada.

Published at http://ceur-ws.org

[1] Other documents, such as transcripts of public hearings, technical support documents, detailed comment response documents, copies of pertinent scientific papers, e-mails and other correspondence, may also be included. Finally, a docket may also contain tabular data and software source code used to produce analytical results.

[2] We have chosen to focus on EPA because this agency published the most rules (~ 20% of all rule documents) and received the most comment submissions (~ 10% of all comment documents) in `regulations.gov` during the studied time period.

[3] The data and code are available at https://github.com/mug31416/PubAdmin-Discourse.git

research, conducted as part of Cornell University's RegulationRoom project, has focused on tools to improve the quality of public discourse around rulemaking (e.g., [41], [46]). Research on the text of rules developed by agencies has mostly focused on the search for similar rules in the FR [33], rather than segmentation of the comment-related discourse in the rule documents.

Prior to launch of `regulations.gov`, work on e-rulemaking used several rule-specific comment collections that were either shared by the agencies—EPA, Fish and Wildlife Service (FWS)—or gathered as part of the RegulationRoom experiments in collaboration with the US Department of Transportation (DOT). The tasks have included near duplicate detection to address mass comment campaigns [58], comment topic modeling [5, 8, 30, 51, 59], stakeholder attitude identification [1, 31], and presence of substantive points in public comments [2, 44, 45, 57]. The RegulationRoom project has generated a number or papers on argument mining and conflict detection within comments [29, 34, 43]. These research efforts have focused on examining only a few regulatory proceedings at a time, whereas we evaluate a signifcantly larger dataset containing hundreds of rule documents.

More recent work on e-regulation has analyzed public comment data collected by `regulations.gov` [13, 14, 35, 37, 50, 52], rule-specific data from the Canadian government [53], and data from the White House e-petition platform [15, 19–21]. The tasks addressed in this body of work are topic modeling [15, 20, 21, 35, 37, 52, 53], sentiment analysis [13, 14, 37, 50], named entity recognition [20], and social network analysis [19].

Segmentation of text into discourse units [38] is a core natural language task. Many downstream tasks, such as information extraction [27], sentiment analysis [3], information retrieval [16], and summarization [4, 36], can benefit from discourse segmentation. Because lexical and syntactic text properties form important discourse clues [6], many segmentation methods rely on hand-crafted features to capture them [17, 26]. Classic learning frameworks that have been used for discourse segmentation are linear Support Vector Machines (SVM) [11] and linear-chain Conditional Random Fields (CRF) [32].

One of the key challenges in discourse segmentation development is the dearth of annotated data, which, until recently, prevented the use of neural architectures. Effective neural discourse segmentation methods [22, 56] have relied on word representations obtained from an external neural model trained to perform a related task using a large corpus [39, 49]. The state-of-the art neural discourse segmentation framework [18, 56] has employed a Bidirectional Long-Short-Term Memory-CRF architecture (BiLSTM-CRF) [25] with an attention mechanism [55].

For our baseline model development, we have combined several classic learning methods with hand-crafted, as well as neural sentence representations, from Bidirectional Encoder Representations from Transformers (BERT) [12], which were trained on English Wikipedia (2,500 million words) and BooksCorpus (800 million words) [60] using masked language and next sentence prediction objectives. BERT representations have demonstrated to perform well on a wide range of natural language processing tasks. We also explore whether fine-tuning of BERT on the unlabeled documents in our corpus improves performance.

3 DATA

3.1 Rule-Making Documents

We work with the EPA's final rule documents that are part of the FR. Along with a summary, each of our documents can contain one or more of the following sections: regulatory background, scope of the regulation, rationale for action, technical material describing the regulatory requirements, responses to public comments on the proposed regulation, statutory and executive order review, and legal references. We are interested in automated identification of all passages where the agency discusses public comments, which could occur throughout the document and are not necessarily confined to the comment response section.

We note that the structure of the final rule documents can vary significantly depending on whether it has been produced by the EPA headquarters or a regional office, as well as depending on the specific EPA office (e.g., Office of Water, Office of Air and Radiation). For example, rule documents produced by the headquarters offices are usually major federal regulations that tend to be long and receive significant public feedback. On the other hand, rule documents produced by regional offices tend to be shorter.[4]

It should be noted that our dataset only contains final rule documents as published in the FR. It does not include submitted comment documents, technical support documents, or detailed, dedicated comment response documents that are part of the docket but extraneous to the register.

3.1.1 *Task 1: Detecting Comment Discussions.*

In the first task, we want to identify the spans in the document where the EPA discusses submitted public comments. Examples of a comment discussion include:

- Descriptions of comments received by the agency. For example, "EPA received comments suggesting that the definition of clean alternative fuel conversion should be limited to a group of fuels with proven emission benefits.";
- Descriptions of the agency's responses to the comments it receives. For example, " EPA believes however that the public interest is better served by a broader definition that allows for future introduction of innovative and as-yet unknown fuel conversion systems. EPA is therefore finalizing the proposed definition of clean alternative fuel conversion...".

By distinction, we are not interested in:

- Summarized feedback from petitions (as opposed to public comments) to the agency;
- Descriptions of the public comments on another rule;
- Statements such as "we received no comments";
- Passages discussing revisions of a regulatory standard rather than revisions of the proposed rule;
- Referrals to another document in the docket with detailed responses to comments.

3.1.2 *Task 2: Classification of Comment Merit.*

In the second task, we want to classify each comment discussion span as to whether the discussed comment prompted a change in the final rule from the proposed rule. As such, we are considering three categories:

[4] With the possible exception of the regional air quality rules that still tend to attract considerable public attention

passages in which the agency indicates a revision of the rule based on a public comment, passages in which the agency dismisses a comment, and neutral comment discussion passages (i.e., the passages in which the agency neither dismisses the comment nor indicates a revision).

Examples of formulations reflecting comment-based regulatory change are rule revisions and rule withdrawals:

- "To address concerns about space limitations, EPA will allow the label information to be logically split between two labels that are both placed as close as possible to the original Vehicle Emission Control Information (VECI) or engine label."
- "EPA agrees and is including use of this procedure in the OBD demonstration requirement for intermediate age vehicles."
- "The EPA has reviewed the new data submitted by the commenter and used these data to determine the revised MACT floor for continuous process vents at existing sources."
- "EPA received one adverse comment from a single Commenter on the aforementioned rule. As a result of the comment received, EPA is withdrawing the direct final rule approving the aforementioned changes to the Alabama SIPs."

Examples of comment dismissals without a subsequent regulatory change are:

- "We disagree that our action to approve California's mobile source regulations that have been waived or authorized by the EPA under CAA section 209 is inconsistent with the Ninth Circuit's decision..."
- "EPA is finalizing the conversion manufacturer definition as proposed."
- "While we agree with the commenter that pressure release from a PRD constitutes a violation, we will address this in a separate rulemaking..."
- "In the final rule we will clarify our position..."
- "EPA appreciates support from the commenters for this initiative and agrees that the rule makes it possible for EPA to process the TRI data more quickly."
- "EPA believes that no further response to the comment is necessary..."

We observe that this task requires considerably more complex inference, potentially spanning multiple sections of the document. As seen in the examples above, comment dismissals range from very obvious to rather subtle. In turn, determinations of whether a rule was materially revised based on the public comments may also require a clear understanding of what was proposed in the first place.

An extreme example of this can be seen from the following comment dismissal sentence:

"Certain aspects of good engineering judgment described in the exhaust control system, evaporate control system, and fuel delivery control system sections may be approached differently than described above, but EPA expects that test data demonstrating compliance is required rather than optional in such cases."

The sentence responds to technical objections to a regulation by conceding that alternatives are valid ("may be approached differently") but goes on to state the substantive decision in domain terminology ("compliance is required rather than optional", suggesting that the comment had advocated for the "optional" alternative). Without context, it is unclear whether this sentence has anything to do with comments at all, let alone whether required vs. optional compliance results in it agreeing with, or dismissing, the comment's arguments.

3.2 Acquisition and Sampling

We have created our corpus from `regulations.gov` data by selecting EPA regulatory dockets for rules proposed in January 2003 or later and finalized as of March 2018. Our selection has been constrained to dockets containing at least one proposed rule document, at least one final rule document, and at least one comment document. Our corpus contains 1,566 EPA dockets (meta-data 8.8 MB), 2,645 final rule documents (HTML, 376 MB), 2,531 proposed rule documents (HTML, 400 MB), and 282,655 comment documents (85% PDF, 36 GB; 15% plain text, 836 MB).

For the purposes of exhaustive rule document annotation, we have used stratified random sampling at the docket level to select two development docket sets (dev1 and dev2) and one test docket set. The sampling procedure has ensured that the docket sets are a representative mix of EPA program offices and regions.[5] As such, we have obtained 75 dev1-set dockets (116 documents), 76 dev2-set dockets (136 documents), and 73 test-set dockets (99 documents).

In our qualitative examination of the regulatory documents, we have found that the section headers of the rule documents are often informative about whether a section contains a discussion of public comments. To make use of this additional information, we have applied the same random sampling procedure to the remaining dockets to obtain 211 training dockets (817 training documents) and 103 validation dockets (197 validation documents) for the section header annotation.

3.3 Preprocessing

The rule documents were processed in two steps. First, we have applied a rule-based rule document parsing procedure to delete tables, split the text into sections, and retrieve section titles of the first and second level super-sections. This procedure exploits the regular structure of documents to create heuristics applicable to roughly 90% of documents.[6] When exceptions to the standard structure are detected, we manually fixed irregularities to enable automatic parsing. Second, the section text has been split into sentences, tokenized, and lemmatized using SpaCy [24][7].

3.4 Annotation

3.4.1 Rule Documents. We hired ten students from Carnegie Mellon University and the University of Pittsburgh to perform the annotation tasks during the period of February 2019–April 2019. All annotators are at least second year undergraduate students. Five of the annotators are masters students in fields including computer science, public health, product management, and international relations. The other five are undergraduate students in civil engineering, creative writing, business, and human computer interaction.

[5] For example, Office of Water/Headquarters, Office of Air and Radiation/Region 1 – Boston.

[6] For example, the first and the second level sections are numbered consecutively in Roman numbers and Latin letters, respectively.

[7] Version 2.0.18 (model en_core_web_sm)

The annotators were trained to perform the two tasks described in Section 3.1.1 and Section 3.1.2. For the first task, each annotator received an hour-long in-person training as well as individualized feedback on a set of four training documents. For the second task, the guidelines were delivered via a video. Each annotator received 50 documents on average, including reliability annotations. The documents were allocated such that each annotator worked on a balanced mix of documents from different EPA offices, regions, and dev1/dev2/test set dockets. The annotations were performed using an online tool developed by a collaborating group at the University of Pittsburgh called *Gloss*.

Finally, we note that some annotators did not complete all assignments for the segmentation task, leading to some redistribution of work. The comment response classification task was completed by eight annotators of the initial ten annotators.

3.4.2 Section Headers. Annotation of the section headers was performed by a sole expert annotator (the first author). To this end, all unique section titles were extracted along with three samples of the first paragraph following the section title. These examples are used to judge whether a section contains comment discussion: If all three sample paragraphs include comment discussions, the section title is flagged as the comment-discussion-indicative title.[8]

4 METHODS

To generate baseline results, we use a classic linear SVM[9] and linear-chain CRF[10] learners to segment the rule documents into spans that contain public comment discussion and merit evaluation by the agency.[11] The benefit of the CRF over the SVM is that, when predicting a sentence label, it takes into account the label of the prior and subsequent sentence in addition to the focal sentence's feature vector. In addition, to understand the impact of incorporating feature interactions, we conduct experiments with the Multi-Layer-Perceptron (MLP)[23].[12]

We estimate three binary sentence-level models predicting whether a given sentence contains: (i) a public comment discussion, (ii) a dismissal of a public comment by the agency, and (iii) an agency decision to revise the proposed rule based on the public comments. For the CRF modeling, a training instance is a sequence of sentences within the rule document section boundaries. To address the label sparsity for the comment dismissal/revision classification, we explore the utility of training models only on data that is known to contain comment discussion (i.e. on the non-ignorable sentences) and then composing a two tiered model to first detect comment discussions, and then then classify their polarity. The hyperparameters have been tuned by fitting the models to the dev1-set and evaluating results on the dev2-set.

4.1 Handcrafted Features

For sentence representation we concatenate three categories of handcrafted features. First, we featurized the text of the sentence for which the prediction needs to be made, as well as the text of the preceding sentence, and concatenate the feature vectors. We use original tokens (including stop words, but excluding punctuation), modified tokens with attached POS tags, bigrams of modified tokens, and bigrams of POS tags.[13] We apply feature hashing [40] to reduce dimensionality. This results in a feature set of size 2,001.

Second, we featurized the text of the section header containing the sentence in question. In that, we apply the same feature generation process used for sentences to the text of the sentence-bearing section header and the header that precedes it. The dimension of this feature set is 101.

Third, we also add a binary flag equal to one if a header of the section in which the sentence occurs has been predicted to contain a comment discussion. We generate these predictions through instance-based learning on the unique section headers from the training set of dockets set aside for this purpose (see Section 3.4.2). Based on the unique headers from the associated validation docket set, this signal mining procedure has a recall of 0.54 and a precision of 0.88.

4.2 Neural Features

We employ BERT[12] to create embedded vector representations for sentences and section headers. BERT is a state of the art neural network language model trained on a large collection of English text in a quasi-unsupervised fashion by having it learn to predict masked words in a sentence, or to classify whether one sentence follows another, or not. By doing so, BERT learns to maintain a neural representation of language context. These vector representations of English text can then be used as for various natural language processing tasks and have been shown to yield significantly better performance than context-independent word embeddings.

As in case of the hand-crafted features, we concatenate both the vectors of the sentence/header in question as well as the context represented by the preceding sentence/header to form a final feature vector. We explore performance of the available pretrained BERT model as well as a BERT model that has been fine-tuned on approximately 6,000 rule documents from our corpus that have not been included in the annotated document sets. To this end, we rely on a PyTorch[47] implementation of BERT.[14] The size of the generated sentence/header embedding is 728. The fine-tuned model was trained for seven epochs.

5 EVALUATION

We evaluate the quality of the rule document annotation using Cohen's kappa coefficient [10], as well as qualitatively. Performance of our baseline text segmentation models is evaluated on the test set at the sentence level using area under the ROC curve (AUC), F1-score, precision, and recall. We found a sentence to be the most meaningful operational definition of a passage, because comment-discussing

[8] For example, there were several first level section titles "What comments did EPA receive?".

[9] We use scikit-learn version 0.20.2 SVC implementation [48] with an error term penalty parameter of 1, and 1,500 as the maximum number of iterations.

[10] We use PyStruct 0.3.2 implementation [42] of margin re-scaled structural SVM using the 1-slack formulation and cutting plane method [28]. We used regularization parameter of 0.1 and 1,500 as the maximum number of iterations.

[11] We have been unable to fit kernelized polynomial and RBF SVMs to our data because these methods do not scale well to the size of our dataset.

[12] We use a scikit-learn version 0.20.2 MLP implementation [48] with one hidden layer of 100 units optimized for at most 100 epochs at the default settings.

[13] We do not use a TFIDF feature representation because it has not performed as well as a simple count-based featurizer in our preliminary experiments.

[14] PyTorch Pretrained BERT: The Big and Extending Repository of pretrained Transformers from https://github.com/huggingface/pytorch-pretrained-BERT. We used the bert-base-uncased version of the model.

sentences are often interspersed with ignorable sentences of a section or a paragraph. For each model, the classification cutoff has been determined using a threshold that maximizes the F1-score on the training data.

6 RESULTS

6.1 Annotation

Table 1 summarizes the key properties of the annotated dataset. For this summary, we have converted span-level annotations into sentence-level annotations. To this end, we have assigned a label to a sentence if an annotator has marked 80% of tokens that make up that sentence. For documents that have been annotated by multiple individuals, we assign a label to a sentence if at least one individual has labeled the sentence. This approach has been motivated by a qualitative examination of annotations, which revealed low recall issues for some annotators. Depending on the dataset, non-ignorable content (i.e. text labeled as discussing comments) comprises 21% to 33% of all sentences, comment dismissals comprise 4% to 5% of all sentences, and comment-based revisions comprise 2% to 3% of all sentences. Approximately half of all labeled sentences have been annotated by two individuals. Due to the annotator attrition, reliability annotations for a more refined labeling task (i.e., identification of comment dismissals and comment-based rule revisions) are available for 73% to 79% of all double-annotated sentences.

Table 1 also reports the inter-annotator agreement statistics, while Table 2 summarizes agreement with the expert annotator on four final rule documents used as part of the annotator training. (Expert annotations have been produced by the first author, who has 10 years of professional experience in supporting EPA's regulatory proposal development.) For the non-ignorable content, inter-annotator agreement scores range from 0.38 to 0.67 (depending on the dataset), whereas agreement with the expert is 0.74 on average (range: 0.35–0.95). We note that agreement on this task appears to improve from the dev1 set to the test set, which may reflect that the annotators learned to do the task better over time, given the order in which the documents have been assigned. Inter-annotator agreement for the comment dismissal labeling task ranges from 0.18 to 0.32, while agreement on the comment-based rule revisions is very low, ranging between 0.086 and 0.19. Agreement with the expert on these tasks is also low: 0.33 (range: 0–0.54) for the comment dismissals and 0.38 (range: 0–0.75) for the comment-based rule revisions.

We have reviewed the annotator errors vis-a-vis the expert annotator. False negatives tend to occur most commonly when:

- The annotator captures only the initial part of the comment discussion that contains typical lexical cues (e.g., "EPA received comments suggesting...", "Commenters noted...", "EPA agrees with the commenters...") but fails to include the entire—usually technical—comment discussion that can span multiple subsequent paragraphs;
- A passage with comment discussion is "buried" in the middle of a longer paragraph, as often happens when comments are discussed in the background section;
- For the more difficult annotation task of identifying comment-based rule revisions and comment dismissals, we have noted that false negatives tend to occur when the evaluation of the passage requires complex inference. As such, the annotators tended to be conservative about assigning these labels for less obvious examples.

Table 1: Characteristics of the Annotated Data

Characteristic	Dev1-set	Dev2-set	Test-set
Number of the Data Set Elements			
Dockets	75	76	73
Documents	116	136	99
Sections	2,197	2,123	1,766
Sentences	72,969	61,837	61,042
Words	1,820,619	1,583,518	1,430,134
Number of the Annotated Sentences			
Non-ignorable content	19,465	20,105	12,979
Comment dismissals	3,527	3,225	2,202
Comment-based regulatory change	2,092	1,015	1,088
Number of the Double-Annotated Sentences			
Non-ignorable content*	42,296	25,300	41,572
Refined content**	33,595	18,561	32,331
Annotator Agreement (Kappa)			
Non-ignorable sentences*	0.42	0.52	0.67
Non-ignorable sentences**	0.38	0.43	0.64
Neutral comment discussion	0.39	0.44	0.66
Comment dismissals	0.32	0.18	0.29
Comment-based regulatory change	0.086	0.19	0.16
Multi-class	0.33	0.38	0.56

Notes: * Sentences for which double annotation of non-ignorable content is available. ** Sentences for which double annotation of content is also available.

Table 2: Annotator Agreement* with Expert

Kappa	Mean	Min	Max
Non-ignorable content	0.74	0.35	0.95
Comment dismissals**	0.33	0	0.54
Comment-based revisions**	0.38	0	0.75

Notes: * Agreement is calculated at the sentence level for four final rule documents. A total of 4,105 sentences are available for this evaluation. ** These statistics are calculated for the eight annotators who performed the task.

For the false positives, we have observed the following tendencies:

- EPA regulations are typically incremental, in that they often tend to modify older, preexisting rules. Therefore, the final and proposal rule document discuss changes/ revisions of the prior regulatory standard. This has been a significant source of confusion for the annotators, who found it difficult to separate comment-based revisions of the proposed regulation from the revisions of the regulatory standard on the regulatory agenda, leading to false positives.

- Another challenge for the annotators has been the decision of when the discussion switches from comment-related to the general topics, also leading to false positives.
- Specifically for the comment-based rule revisions, some annotators found it challenging to distinguish between revisions of the proposed rule that were based on comments from revisions that occurred for other reasons. For example, the EPA may implement revisions based on new evidence that emerges after the proposed rule is submitted for public review.

6.2 Classification Results

Table 3 and Table 4 show the test set evaluation performance results for each binary classification task divided by learning framework and feature set. The models have produced better than random predictions, with largest AUC of 0.937 noted for the non-ignorable content prediction and smallest AUC of 0.677 noted for the comment-based rule change prediction. These patterns largely reflect the differences in the quality of annotations obtained for our prediction tasks, with the segmentation task being significantly easier than the comment response classification task.

For the non-ignorable content prediction, the models produce recall in the range of 0.636–0.708 and precision in the range of 0.688–0.798. Unsurprisingly, for the more complex annotation tasks with low annotator agreement, classification quality is poor. For the comment dismissal prediction, recall is 0.085–0.537 and precision is 0.091–0.249, whereas for the comment-based rule change prediction, recall is 0.065–0.490 and precision is 0.056–0.189.

6.2.1 Linear Model Analysis. CRF model results do not appear to be materially different from those generated by the SVM model on the same handcrafted feature set, even through they take into account the labels of neighboring sentences. We note, however, that the CRF models have produced consistently higher precision scores, compared to the SVM models estimated on the same feature set. Because we experienced some convergence problems with CRF models, we have fit them to only one feature set.

Table 3 also shows that neural BERT features on average tend to generate higher AUC, precision, and recall. We note that the two-tiered models perform better for the comment dismissal prediction, but not for the comment-based revision prediction. In the latter case, the gains in precision are minor and do not offset the significant losses in recall.

We also observe that neural features based on the fine-tuned BERT can perform better than those using out-of-the-box BERT (e.g. best AUC and precision on non-ignorable content prediction). Interestingly, combining neural and handcrafted feature sets generally does not produce synergy performance increases, which could be due to the substantial increase in the overall feature dimension, or the lack of feature interaction capacity in linear models.

6.2.2 Multi-Layer Perceptron Results. In a second set of experiments we assessed whether classification performance increases with models that allow for feature interactions. To this end, we trained a series of Multi-Layer-Perceptron models (i.e. a neural network with one hidden layer of size 100 and a two-class softmaxed output) on our tasks and feature sets. Table 4 contains the results we obtained on for the MLP with an identity transformation (MLP-Id) before the final softmax.[15]

Table 3: Baseline Test Set Results

Model	**AUC**	**F1**	**Prec.**	**Recall**
All Non-ignorable Content				
Random	0.501	0.200	0.164	0.256
CRF+HCF	n.a	**0.717**	0.750	0.687
SVM+HCF	0.911	0.716	0.734	**0.698**
SVM+BERT (as is)	0.921	0.695	0.721	0.672
SVM+HCF+BERT (as is)	0.915	0.689	0.753	0.636
SVM+BERT (tuned)	**0.928**	0.703	**0.764**	0.651
SVM+HCF+BERT (tuned)	0.913	0.693	0.709	0.677
Comment Dismissals				
Random	0.502	0.020	0.017	0.023
Semi-Random	0.811	0.152	0.162	0.144
CRF+HCF	n.a.	0.209	0.225	0.195
SVM+HCF	0.760	0.258	0.177	0.478
SVM+BERT (as is)	0.869	0.277	0.194	0.484
SVM+HCF+BERT (as is)	0.862	0.258	0.170	**0.537**
SVM+BERT (tuned)	0.869	0.278	0.196	0.478
SVM+HCF+BERT (tuned)	0.768	0.257	0.191	0.393
2-SVM+HCF	0.872	0.281	0.202	0.460
2-SVM+BERT (as is)	0.874	0.286	0.214	0.432
2-SVM+HCF+BERT (as is)	0.881	0.257	0.214	0.322
2-SVM+BERT (tuned)	**0.888**	**0.318**	**0.249**	0.441
2-SVM+HCF+BERT (tuned)	0.830	0.271	0.212	0.375
Comment-based Regulatory Change				
Random	0.503	0.038	0.032	0.046
Semi-Random	0.678	0.050	0.053	0.048
CRF+HCF	n.a.	0.088	0.091	0.085
SVM+HCF	0.677	0.092	0.056	0.273
SVM+BERT (as is)	0.802	**0.126**	0.074	0.420
SVM+HCF+BERT (as is)	0.736	0.099	0.058	0.335
SVM+BERT (tuned)	**0.815**	0.125	0.077	0.337
SVM+HCF+BERT (tuned)	0.754	0.091	0.051	**0.446**
2-SVM+HCF	0.724	0.081	0.091	0.073
2-SVM+BERT (as is)	0.796	0.104	0.112	0.097
2-SVM+HCF+BERT (as is)	0.745	0.078	0.075	0.081
2-SVM+BERT (tuned)	0.808	0.086	**0.128**	0.065
2-SVM+HCF+BERT (tuned)	0.744	0.108	0.088	0.138

Notes: Random – predictions are draws from a Bernoulli distribution with probability set to the target class prior. Semi-Random – predictions are generated by first applying the best-performing non-ignorable content classifier and then drawing from a Bernoulli distribution with probability set to the target class conditional prior. 2-SVM – a two-tiered SVM model. HCF – hand crafted features. AUC – area under the ROC curve. CRF model does not produce confidence scores, hence AUC estimation was not possible. The classification cutoff was chosen to maximize F1 score for each model.

We observe that nonlinear models using BERT features can achieve somewhat higher AUC and F1 scores than the linear models shown in Table 3. We also see that adding handcrafted features to

[15] We have also obtained results for the MLP with an a Rectified Linear Unit (ReLU) activation function before the final softmax MLP-ReLU. The practical difference is that a ReLU activation will truncate all incoming negative activation values to 0 and leave positive ones unchanged. We do not report these results because they were largely inferior to those obtained for the MLP-Id variant.

a model can occasionally yield some performance synergy. From this we infer that nonlinear models could potentially produce better results on our dataset, and hence we plan to experiment with recurrent or dilated convolutional models for sequence tagging to leverage the document context in future work.

Table 4: Auxiliary Test Set Results

Model	AUC	F1	Prec.	Recall
All Non-ignorable Content				
Random	0.501	0.200	0.164	0.256
MLP-Id+HCF	0.911	0.711	0.776	0.656
MLP-Id+BERT (as is)	0.917	0.678	0.688	0.669
MLP-Id+HCF+BERT (as is)	0.930	0.731	**0.798**	0.674
MLP-Id+BERT (tuned)	**0.937**	0.705	0.772	0.648
MLP-Id+HCF+BERT (tuned)	0.930	**0.732**	0.759	**0.708**
Comment Dismissals				
Random	0.502	0.020	0.017	0.023
Semi-Random	0.811	0.152	0.162	0.144
MLP-Id+HCF	0.851	0.289	0.208	0.476
MLP-Id+BERT (as is)	0.875	0.273	0.204	0.410
MLP-Id+HCF+BERT (as is)	0.871	0.297	0.213	**0.492**
MLP-Id+BERT (tuned)	**0.893**	0.284	0.209	0.442
MLP-Id+HCF+BERT (tuned)	0.825	0.291	0.212	0.460
2-MLP-Id+HCF	0.850	0.284	0.229	0.374
2-MLP-Id+BERT (as is)	0.859	0.281	0.218	0.394
2-MLP-Id+HCF+BERT (as is)	0.887	0.301	**0.243**	0.395
2-MLP-Id+BERT (tuned)	0.890	**0.309**	0.240	0.432
2-MLP-Id+HCF+BERT (tuned)	0.882	0.294	0.239	0.383
Comment-based Regulatory Change				
Random	0.503	0.038	0.032	0.046
Semi-Random	0.678	0.050	0.053	0.048
MLP-Id+HCF	0.718	0.103	0.061	0.329
MLP-Id+BERT (as is)	0.818	0.121	0.072	0.384
MLP-Id+HCF+BERT (as is)	0.766	0.113	0.068	0.335
MLP-Id+BERT (tuned)	**0.837**	**0.138**	0.086	0.343
MLP-Id+HCF+BERT (tuned)	0.806	0.114	0.065	**0.490**
2-MLP-Id+HCF	0.757	0.078	0.093	0.068
2-MLP-Id+BERT (as is)	0.723	0.092	0.077	0.114
2-MLP-Id+HCF+BERT (as is)	0.770	0.092	0.112	0.078
2-MLP-Id+BERT (tuned)	0.766	0.123	**0.189**	0.091
2-MLP-Id+HCF+BERT (tuned)	0.789	0.130	0.113	0.154

Notes: Random – predictions are draws from a Bernoulli distribution with probability set to the target class prior. Semi-Random – predictions are generated by first applying the best-performing non-ignorable content classifier and then drawing from a Bernoulli distribution with probability set to the target class conditional prior. 2-MLP – a two-tiered MLP model. HCF – hand crafted features. AUC – area under the ROC curve. MLP-Id – a multi-layer perceptron with one hidden layer with 100 units and an identity non-linearity followed by a Softmax; this model is equivalent to a generalized linear regression model with interaction terms. The classification cutoff was chosen to maximize F1 score for each model.

6.2.3 Error Analysis. For our best-performing models we have generated and examined five random examples for each type of error. Our findings are as follows:

False Positives: The models tend to produce false positives when sentences contain certain trigger words (such as “response”, “revision”, “finalizing the rule as proposed”) yet the overall context of the passage is not related to the discussion of public comments. For example, these trigger words have been observed in passages discussing petitions and revisions of the regulatory standard that are not based on comments, similar to mistakes made by human annotators. There is also a fair share of label noise: As noted earlier, the annotators have been challenged by longer comment discussions and occasionally failed to capture the entire relevant span. We also conjecture that in this case the models have been guided by the section-header related signal.

False Negatives: The false negatives tend to occur in sections that do not commonly contain comment discussion (e.g., “Background”, “Executive Order Review”). Sentences that lack the boilerplate language (e.g., “response”, “EPA”, “comment”) also tend to be missed more often. As with the false positives, we observed some amount of label noise, often in cases when the annotators mislabeled discussions of regulatory revisions that have not been driven by public feedback or when annotators have failed to determine an appropriate boundaries for the technical discussion of comments.

Label Confusion: We have observed several cases of the models being confused about the polarity of EPA assessment, particularly when the sentence has included trigger words such as “agree” and “disagree” together.

Parsing: We have noted several instances of erroneous sentence parsing (e.g., a citation “40 CFR 51.1010(b).” has been isolated as a sentence) that lead to classification errors. This issue could be remedied by a sentence boundary detector oriented towards processing legal text [54].

7 DISCUSSION

It is likely possible to automatically identify certain type of content in regulatory documents with irregular structure. Our baseline segmentation performance for detecting comment discussion sentences with recall in the range of 0.636–0.708 and precision in the range of 0.688–0.798. While we have focused on identifying comment discussion by the receiving agency, we believe that there are other types of content (e.g., regulatory requirements) automated segmentation of which may be both, desired and feasible. Detecting specific comment discussions that either dismiss comments or announce rule revision turns out to be a harder task for both annotators and, consequently, for models. Moving forward, this begs the question of which information need the model caters to. If value is added by quickly pointing an expert to comment discussion passages, then a well-performing model is within reach given good training data. On the other hand, an automated analysis of topics for which comments have been influential remains a hard problem.

We also note that our dataset has been compiled using highly educated non-expert annotators. We have found that this type of background is sufficient for producing relatively coarse annotations (e.g., identifying parts of the document that contain comment discussion). We have measured the annotator-expert agreement of 0.74 for the comment discussion identification task. However, more refined annotation tasks, such as the ones determining the agency's responses to public feedback, would likely require expert-level understanding of the domain.

We believe that our baseline modeling results can be further improved by developing a fully neural sequence tagging model, such as the one developed for the standard discourse segmentation corpus [56]. However, even with access to the sequence encoders such as BERT, the limited size of our corpus may still present a modeling challenge.

8 CONCLUSIONS

We have produced a dataset and baseline for a novel discourse segmentation task of identifying public comment discussion and evaluation by regulatory agencies. In doing so we presented evidence that detecting comment discussions automatically using mainstream NLP techniques is feasible given good training data. Classifying discussions of a particular type is harder both because of data sparsity and low annotator agreement. While good general detection performance will add value in some practical settings, we see opportunity for further improvement in the use of neural sequence tagging models, albeit subject to the limitations of data quality as a function of annotator expertise, training, and type system design.

9 ACKNOWLEDGMENTS

The authors thank University of Pittsburgh Intelligent Systems Program student Jaromir Savelka for permission to use the *Gloss* annotation tool.

REFERENCES

[1] Jaime Arguello and Jamie Callan. 2007. A bootstrapping approach for identifying stakeholders in public-comment corpora. In *Proceedings of the 8th annual international conference on Digital government research: bridging disciplines & domains*. Digital Government Society of North America, 92–101.

[2] Jaime Arguello, Jamie Callan, and Stuart Shulman. 2008. Recognizing citations in public comments. *Journal of Information Technology & Politics* 5, 1 (2008), 49–71.

[3] Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from rst discourse parsing. *arXiv preprint arXiv:1509.01599* (2015).

[4] Mohammad Hadi Bokaei, Hossein Sameti, and Yang Liu. 2016. Extractive summarization of multi-party meetings through discourse segmentation. *Natural Language Engineering* 22, 1 (2016), 41–72.

[5] Claire Cardie, Cynthia R Farina, Matt Rawding, and Adil Aijaz. 2008. An erulemaking corpus: Identifying substantive issues in public comments. (2008).

[6] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*. Springer, 85–112.

[7] Nuno Carvalho and Rui Pedro Lourenço. 2018. E-Rulemaking: Lessons from the Literature. *International Journal of Technology and Human Interaction (IJTHI)* 14, 2 (2018), 35–53.

[8] Lijun Chen. 2007. Summaritive digest for large document repositories with application to e-rulemaking. (2007).

[9] Cary Coglianese. 2004. E-Rulemaking: Information technology and the regulatory process. *Administrative Law Review* (2004), 353–402.

[10] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[11] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[13] Tao Ding and Shimei Pan. 2016. How Reliable Is Sentiment Analysis? A Multi-domain Empirical Investigation. In *International Conference on Web Information Systems and Technologies*. Springer, 37–57.

[14] Lauren M Dinour and Antoinette Pole. 2017. Potato Chips, Cookies, and Candy Oh My! Public Commentary on Proposed Rules Regulating Competitive Foods. *Health Education & Behavior* 44, 6 (2017), 867–875.

[15] Catherine Dumas, Teresa M Harrison, Loni Hagen, and Xiaoyi Zhao. 2017. What Do the People Think?: E-Petitioning and Policy Decision Making. In *Beyond Bureaucracy*. Springer, 187–207.

[16] Yixing Fan, Jiafeng Guo, Yanyan Lan, Jun Xu, Chengxiang Zhai, and Xueqi Cheng. 2018. Modeling diverse relevance patterns in ad-hoc retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 375–384.

[17] Vanessa Wei Feng and Graeme Hirst. 2014. Two-pass discourse segmentation with pairing and global features. *arXiv preprint arXiv:1407.8215* (2014).

[18] Elisa Ferracane, Titan Page, Junyi Jessy Li, and Katrin Erk. 2019. From News to Medical: Cross-domain Discourse Segmentation. *arXiv preprint arXiv:1904.06682* (2019).

[19] Loni Hagen, Teresa M Harrison, and Catherine L Dumas. 2018. Data Analytics for Policy Informatics: The Case of E-Petitioning. In *Policy Analytics, Modelling, and Informatics*. Springer, 205–224.

[20] Loni Hagen, Teresa M Harrison, Özlem Uzuner, Tim Fake, Dan Lamanna, and Christopher Kotfila. 2015. Introducing textual analysis tools for policy informatics: a case study of e-petitions. In *Proceedings of the 16th annual international conference on digital government research*. ACM, 10–19.

[21] Loni Hagen, Özlem Uzuner, Christopher Kotfila, Teresa M Harrison, and Dan Lamanna. 2015. Understanding Citizens' Direct Policy Suggestions to the Federal Government: A Natural Language Processing and Topic Modeling Approach. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on*. IEEE, 2134–2143.

[22] Mehedi Hasan, A Kotov, S Naar, GL Alexander, and A Idalski Carcone. 2019. Deep neural architectures for discourse segmentation in e-mail based behavioral interventions. In *American Medical Informatics Association (AMIA)*.

[23] Geoffrey E Hinton. 1990. Connectionist learning procedures. In *Machine learning*. Elsevier, 555–610.

[24] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear* (2017).

[25] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).

[26] Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 13–24.

[27] Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-Level N-ary Relation Extraction with Multiscale Representation Learning. *arXiv preprint arXiv:1904.02347* (2019).

[28] Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. 2009. Cutting-plane training of structural SVMs. *Machine Learning* 77, 1 (2009), 27–59.

[29] Barbara Konat, John Lawrence, Joonsuk Park, Katarzyna Budzynska, and Chris Reed. 2016. A Corpus of Argument Networks: Using Graph Properties to Analyse Divisive Issues.. In *LREC*.

[30] Namhee Kwon, Stuart W Shulman, and Eduard Hovy. 2006. Multidimensional text analysis for eRulemaking. In *Proceedings of the 2006 international conference on Digital government research*. Digital Government Society of North America, 157–166.

[31] Namhee Kwon, Liang Zhou, Eduard Hovy, and Stuart W Shulman. 2007. Identifying and classifying subjective claims. In *Proceedings of the 8th annual international conference on Digital government research: bridging disciplines & domains*. Digital Government Society of North America, 76–81.

[32] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).

[33] Gloria T Lau. 2004. *A comparative analysis framework for semi-structured documents, with applications to government regulations*. Stanford University.

[34] John Lawrence, Joonsuk Park, Katarzyna Budzynska, Claire Cardie, Barbara Konat, and Chris Reed. 2017. Using argumentative structure to interpret debates in online deliberative democracy and eRulemaking. *ACM Transactions on Internet Technology (TOIT)* 17, 3 (2017), 25.

[35] Karen EC Levy and Michael Franklin. 2014. Driving regulation: using topic models to examine political contention in the US trucking industry. *Social Science Computer Review* 32, 2 (2014), 182–194.

[36] Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 137–147.

[37] Michael A Livermore, Vladimir Eidelman, and Brian Grom. 2017. Computationally assisted regulatory participation. *Notre Dame L. Rev.* 93 (2017), 977.

[38] Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT press.

[39] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[40] John E. Moody. 1988. Fast Learning in Multi-Resolution Hierarchies. In *Advances in Neural Information Processing Systems 1, [NIPS Conference, Denver, Colorado, USA, 1988]*. 29–39. http://papers.nips.cc/paper/

175-fast-learning-in-multi-resolution-hierarchies

[41] Peter Muhlberger, Nick Webb, and Jennifer Stromer-Galley. 2008. The Deliberative E-Rulemaking project (DeER): improving federal agency rulemaking via natural language processing and citizen dialogue. In *Proceedings of the 2008 international conference on Digital government research*. Digital Government Society of North America, 403–404.

[42] Andreas C. Müller and Sven Behnke. 2014. pystruct - Learning Structured Prediction in Python. *Journal of Machine Learning Research* 15 (2014), 2055–2060. http://jmlr.org/papers/v15/mueller14a.html

[43] Joonsuk Park. 2016. *Mining and evaluating argumentative structures in user comments in eRulemaking*. Cornell University.

[44] Joonsuk Park, Cheryl Blake, and Claire Cardie. 2015. Toward machine-assisted participation in eRulemaking: An argumentation model of evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*. ACM, 206–210.

[45] Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*. 29–38.

[46] Joonsuk Park, Sally Klingel, Claire Cardie, Mary Newhart, Cynthia Farina, and Joan-Josep Vallbé. 2012. Facilitative moderation for online participation in eRulemaking. In *Proceedings of the 13th Annual International Conference on Digital Government Research*. ACM, 173–182.

[47] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[49] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[50] Rachel A Potter. 2017. More than spam? Lobbying the EPA through public comment campaigns. In *Brookings Series on Regulatory Process and Perspective*. https://www.brookings.edu/research/more-than-spam-lobbying-the-epa-through-public-comment-campaigns

[51] Stephen Purpura, Claire Cardie, and Jesse Simons. 2008. Active learning for e-rulemaking: Public comment categorization. In *Proceedings of the 2008 international conference on Digital government research*. Digital Government Society of North America, 234–243.

[52] Reza Rajabiun. 2015. Beyond Transparency: The Semantics of Rulemaking for an Open Internet. *Ind. LJ Supp.* 91 (2015), 33.

[53] Reza Rajabiun and Catherine Middleton. 2015. Public Interest in the Regulation of Competition: Evidence from Wholesale Internet Access Consultations in Canada. *Journal of Information Policy* 5 (2015), 32–66.

[54] Jaromir Savelka, Vern R Walker, Matthias Grabmair, and Kevin D Ashley. 2017. Sentence boundary detection in adjudicatory decisions in the united states. *Traitement automatique des langues* 58, 2 (2017), 21–45.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[56] Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward Fast and Accurate Neural Discourse Segmentation. *arXiv preprint arXiv:1808.09147* (2018).

[57] Antje Witting. 2015. Measuring the use of knowledge in policy development. *Central European Journal of Public Policy* 9, 2 (2015), 54–62.

[58] Hui Yang and Jamie Callan. 2005. Near-duplicate detection for eRulemaking. In *Proceedings of the 2005 national conference on Digital government research*. Digital Government Society of North America, 78–86.

[59] Hui Yang and Jamie Callan. 2008. Ontology generation for large email collections. In *Proceedings of the 2008 international conference on Digital government research*. Digital Government Society of North America, 254–261.

[60] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*. 19–27.