Inferring missing metadata from environmental policy texts

Steven Bethard, Egoitz Laparra, Sophia Wang, Yiyun Zhao
Ragheb Al-Ghezi, Aaron Lien, Laura López-Hoffman
University of Arizona
{bethard,laparra,rxnsp689,yiyunzhao,raghebalghezi,alien,lauralh}
@email.arizona.edu

Abstract

The National Environmental Policy Act (NEPA) provides a trove of data on how environmental policy decisions have been made in the United States over the last 50 years. Unfortunately, there is no central database for this information and it is too voluminous to assess manually. We describe our efforts to enable systematic research over US environmental policy by extracting and organizing metadata from the text of NEPA documents. Our contributions include collecting more than 40,000 NEPA-related documents, and evaluating rule-based baselines that establish the difficulty of three important tasks: identifying lead agencies, aligning document versions, and detecting reused text.

1 Introduction

Hurricanes inundating low-income neighborhoods. Air and noise pollution delaying learning in children. Raging wildfires displacing communities. These are *wicked problems* (Rittel and Webber, 1973) that span jurisdictions and disciplines; have multiple, complex causes; and undergo rapid change with high uncertainty. Solutions to such problems must integrate scientific information about causes, consequences, and uncertainties, with social and political information about public values, concerns, and needs.

In the United States, the National Environmental Policy Act (NEPA), passed by a near-unanimous US congress almost 50 years ago (91st Congress, 1970), is intended as a tool for such problems. NEPA is elegant in the simplicity of its vision: that science results in more informed decisions, and that a democratic process that engages the public results in better environmental and social outcomes. The heart of NEPA is the environmental impact statement (EIS), a detailed, scientific analysis of the expected impacts of federal actions (plans, projects, and activities) and an assessment of possible alternative actions. EISs are developed by the federal government with participation from the public in determining the scope and commenting on draft documents. Since 1970, some 37,000 EISs have analyzed the impacts of federal actions such as construction of transportation infrastructure; permit approvals for oil, gas, and mineral extraction; management of public lands; and proposed regulations.

Unfortunately, congress did not mandate the organized storage of the scientific data NEPA generates, nor the evaluation of its outcomes or of the public engagement processes it requires. There is no central database for this information and it is too voluminous to assess manually. As a result, scientists are able only to support decision-making about specific actions and to assess the outcomes only of specific projects. But systematic analysis across projects is stymied.

We describe a project that aims to enable such systematic research by using natural language processing (NLP) techniques to extract and organize metadata from the text of NEPA documents. Our main contributions are:

- Collecting a large set of environmental policy documents in need of NLP solutions.
- Implementing baseline NLP models for some of the high-priority text normalization tasks.
- Analyzing model performance and illustrating some of the remaining challenges.

2 Data collection

There is no single repository of NEPA documents, and each governmental department or agency chooses its own way to make the documents available to the public. We have thus begun a large-scale web-crawling effort to collect NEPA documents from across the many governmental websites. This means creating a custom scraping tool for each de-

Proc. of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 46–51
Minneapolis, MN, USA, June 7, 2019.

Source of download	Documents
EPA	9238
DOI	13450
DOE	19484

Table 1: Documents collected so far from the different department or agency[1]websites.

		Document type	
		EIS	Other
Version type	Draft	777	4305
	Final	709	3055
	Other	3	40

Table 2: Breakdown of documents collected so far from the EPA. We could not recover version type or document type meta-data for 349 of the 9238 documents.

partment or agency, as none of the sites except for regulations.gov have any programmatic APIs. We have primarily focused on collecting EISs, but have also collected other related documents when they are available. Table 1 shows the progress of our collection efforts so far, and Table 2 shows a breakdown of just the epa.gov documents by whether the files are part of a draft or final version of an EIS.

Each EIS "document" downloaded from these sites is typically a zip archive many PDFs, with the different chapters and appendices of a each EIS broken out into separate PDFs. This is convenient for the distributing agency, but inconvenient for automated analysis. Since there is no standardized naming convention or organization, there is no simple way to automatically combine the various PDFs into a properly ordered single text for the entire EIS. Thus, in the analyses of the current paper, we often treat each PDF separately, but we acknowledge that future work will need a better solution to this PDF ordering and concatenation problem.

Most of the websites hosting these documents contain little or no metadata about them. Some critical metadata that is needed for all documents includes: Which governmental departments or agencies contributed to which documents? Which documents should be linked to each other (e.g., because one is a draft and one is a final version of the same EIS)? Which fine-grained locations (cities, mountains, rivers, etc.) are involved?

On 14 December 2018, NEPA.gov released a spreadsheet of additional metadata on 1161 EISs for which a a final EIS was published between January 1, 2010, and December 31, 2017. This spreadsheet contains several useful things: a canonical title, the dates of all the versions of the EIS, and the lead department and agency for the EIS. Table 3 shows the number of EISs for each of the top agencies in this spreadsheet. Note that the spreadsheet does not link directly to any PDF documents, so work is required to match the metadata to the documents it is describing. Nonetheless, the spreadsheet provides an initial set of annotations that can enable NLP analysis of NEPA documents.

Agency	Count
USFS	276
BLM	128
FHWA	114
USACE	89
NPS	77

Table 3: Distribution of EISs for the top 5 agencies (out of 51 agencies and 1161 EISs in the data), according to the metadata released by NEPA on 14 Dec 2018.

3 Challenge: Identifying lead agencies

A simple but critical piece of metadata needed for analyzing EISs is which governmental agency led the development of the EIS. US agencies are organized in a hierarchy, where, for example, the Forest Service (USFS) and the Animal and Plant Health Inspection Service (APHIS) are under the Department of Agriculture (USDA). Documents usually identify their lead agency in the first few pages, but how they do this varies widely from document to document. For instance, the leading agency may be identified by a logo, as text on the title page, on a later page with "leading agency" nearby, etc.

Note that the task of identifying lead agencies differs from the classic NLP task of named entity recognition in two important ways: not all organizations mentioned in a document are the lead agency (most organizations are not), and agency names must also be standardized (i.e., it is an *entity-linking* problem Shen et al., 2015).

3.1 Baseline model

To judge how sophisticated of an NLP system would be necessary for this task, we first applied a simple rule-based baseline. First, all phrases in the first 15 pages of the document that exactly match a department or agency name[1] were identified and sorted by their position in the document. Any agency in the sorted list that was followed by one of its children (according to the agency hierarchy) was discarded. The first name in the sorted, filtered list was then predicted as the lead agency. For an EIS "document" that consisted of multiple PDFs, we applied this rule-based model to each of the PDFs, and selected the most frequently predicted agency. If there was a tie, the rule-based model predicted no lead agency for this EIS.

We evaluated the performance of this baseline on 107 project folders (730 files), achieving an accuracy of 86%.

3.2 Remaining challenges

This baseline fails when the lead agency does not appear as the earliest agency in the majority of the PDFs representing the EIS "document". For example, in a document where *National Marine Fisheries Service* was specifically indicated as the leading agency, the model incorrectly predicted *National Oceanic and Atmospheric Administration* because it occurred earlier in the text where an National Oceanic and Atmospheric Administration Award was mentioned. As another example, the model correctly found the lead agency in the main PDF of one EIS, but supplementary documents of that EIS never mentioned the correct lead agency, and instead mentioned a few other agencies, so the final prediction after voting was incorrect.

In the future, we expect to achieve better performance on this task by training a machine learning classifier that considers the context of each candidate for useful trigger words like *lead* and *award*.

4 Challenge: Aligning document versions

Understanding an EIS means understanding the process of its creation, from draft EIS, through the public comment period, and on to the final EIS. Sometimes draft and final versions of an EIS are explicitly linked together on the governmental agency's website, but most of the time the documents are delivered separately, with no metadata explicitly linking them.

4.1 Baseline model

We applied a few simple rule-based baselines to establish how difficult of a task it would be to link draft and final versions of an EIS. The first baseline, TITLE, only matches a draft document with a final document when they have exactly the same title. The second baseline, DATE+AGENCY+STATE, uses the 14 December 2018 metadata release to establish how much additional metadata beyond the title would help. It takes a metadata entry, which gives a draft EIS date, a final EIS date, an agency, and a state, and finds all (draft, final) document pairs that are consistent with that entry. The final baseline, TITLE|DATE+AGENCY+STATE performs both of the above matching strategies.

Matching model	Precision	Recall
TITLE	1.000	0.403
DATE+AGENCY+STATE	1.000	0.516
TITLE|DATE+AGENCY+STATE	1.000	0.674

Table 4: Performance of baseline models on matching draft and final versions of the same EIS in 1161 EISs in the 14 December 2018 metadata release.

If any of the above baselines would have matched more than two documents (one draft and one final), we marked such a prediction as incorrect. We applied this restriction because there should be only two versions of each document, draft and final, so finding more than two suggests that we were finding versions from more than one EIS.[2]

Table 4 shows the performance of these baselines on the 1161 EISs in the 14 December 2018 metadata release. Though all the baselines are highly precise, even the baseline that uses the manually curated metadata is unable to find a draft and final version of the EIS for more than 30% of the EISs in the metadata release.

4.2 Remaining challenges

The baselines fail when there is no exact match between the titles; when any of the information of date, state or agency is imprecise; or when multiple projects occur with the same date, state and leading agency. We found that unmatched titles may differ in only tiny ways (e.g., spelling errors) or in major ways (e.g., major reprhrasing). For example, in one project, the only difference was that the word *mccone* in the title was misspelled as *mccore*, whereas in another project, the title *entry control*

[1] The full hierarchy of department and agency acronyms is at https://www.loc.gov/rr/news/fedgov.html

[2] As we have further explored the data, it appears that there are occiasionally more than two versions of the same EIS (e.g., some have a *supplemental draft* version). We are thus in the process of manually annotating sets of similar titles allowing for more than two possible drafts.

reconfiguration area at wright-patterson air force base, ohio was changed to *base perimeter fence relocation in area a fairborn oh*. There are also agency/date/state metadata errors. For example, in one project, the agency is sometimes labeled as *NGB* but sometimes labeled as *DOD*.

It's also worth noting that the baselines that include dates are more oracles than baselines, since they assume that there is a metadata entry somewhere that gives draft and final dates of a single EIS. Such information is unavailable outside of the 1161 entries manually curated by NEPA.gov.

In the future, we expect to achieve better performance on this task by applying techniques that are more robust to word variations, such as measuring title similarity through cosines over word TF-IDF vectors, or more modern approaches like the Universal Sentence Encoder (Cer et al., 2018).

5 Challenge: Detecting reused text

An important research question about NEPA is the degree to which public comments result in changes to the proposed actions. One way of measuring such changes is to look at how much an EIS changes between its draft (pre-comments) version and its final (post-comments) version.

5.1 Baseline model

We apply the baseline from the PAN Plagiarism Detection shared task (Potthast et al., 2012), which partitions texts into 50-character chunks after ignoring non-alphanumeric characters and spaces. Then, it intersects the set of source chunks with the set of target chunks to determine the overlapping text between them. This baseline is representative of the other approaches to that task, which vary primarily on the size of chunks selected and under what conditions chunks were merged. We selected this baseline because it is more conservative, suggesting only very confident matches. We applied this model to 37 draft/final document pairs that we curated from 10 EIS "documents" (138 PDF files), where we, for example, manually confirmed that the draft file `SEP-HCP Draft EIS 10-10-2014` corresponded to the final file `SEP-HCP Final EIS 11-18-15 w app`.

For each draft/final pair, we calculated a DRAFT-REUSE score: the fraction of the text in the final version that was identified as being reused from the text in the draft version. Figure 1 plots the histogram of DRAFT-REUSE scores. The majority of final documents in our sample reused 90% or more of the text from their draft versions. That is, in most cases, less than 10% of the document changed as a result of the public comments.

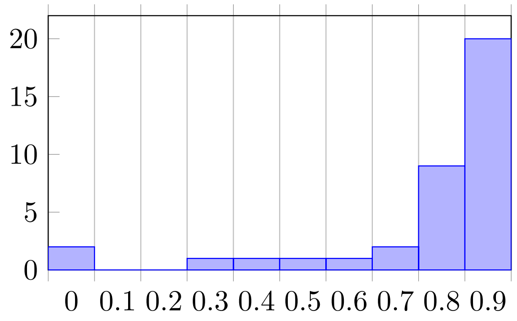

Figure 1: Distribution of EIS DRAFT-REUSE scores in a sample of 37 draft/final document pairs.

5.2 Remaining challenges

The baseline model fails when text is reused with many small changes, and when there are failures in the PDF-to-text process. An example of many small changes is that the word *Draft* typically gets globally replaced with *Final*, so many near-copy-pastes are not detected since they mismatch at each point where *Draft* was previously in the text. An example of PDF-to-text failures is `ACP SHP FEIS Volume II part 3` and `ACP SHP DEIS Volume II part 3`, where the DRAFT-REUSE score was only 0.5 because the volumes are primarily diagrams and images, and even captions that should match do not because the PDF-to-text process produces many partial or weirdly segmented words when they are in captions.

In the future, we expect to achieve better performance on this task by incorporating some of the merging rules applied by the other systems in the PAN Plagiarism Detection shared task (Potthast et al., 2012). But we will first need to acquire at least a small set of examples where NEPA experts have annotated snippets of document reuse. This will allow us to fairly evaluate the performance of different models.

6 Related Work

There have been some previous projects that gathered, organized and extracted metadata from collections of political and social science documents, such as newswire sources (Sönmez et al., 2016) or historical archives (Zervanou et al., 2011). However, to the best of our knowledge, ours is the first project to consider the large number of environmental policy documents produced within the NEPA framework. Our project is also the first to look at extracting metadata fields specific to such documents, such as the lead federal agency. Though there is some relation between extracting lead agencies and extracting other organizational information like affiliations (Jonnalagadda and Topham, 2010) or science funding bodies (Kayal et al., 2017), the different role that lead agencies play in drafting environmental policy documents yields a different information extraction problem.

There is some prior work on automatically analyzing edits between document versions. Some have focused on classifying edits in Wikipedia articles (Bronner and Monz, 2012; Daxenberger and Gurevych, 2013), and Goyal et al. (2017) measured the importance of different kinds of changes between versions of news articles. The EIS documents we analyze have a very different semantics to their versioning. The NEPA process specifies that a public comment period must come between the draft and final EIS, and it is expected that the changes between versions will address issues raised during this period. Thus, our data yields a unique possibility of investigating how external comments influence document versions.

7 Discussion

We have presented our first steps toward extracting and organizing metadata from the texts of environmental policy documents produced under the National Environmental Policy Act (NEPA). We believe this data presents an interesting and challenging opportunity for the NLP community to support research on environmental policy. The current work has established baselines for three important tasks (identifying lead agencies, aligning document versions, and detecting reused text) and our analysis of the places where the baselines have failed should make an excellent starting point for the application of modern NLP techniques (e.g., deep learning models) to solve these challenges.

It is an explicit goal of our project to make avaialble for future research all documents we have collected and all metadata we have inferred. As all documents are generated and publicly released by the United States government, there are no copyright issues in providing access to such a collection. We are currently in the process of setting up a server and designing an application programming interface (API) to provide access to researchers and other interested parties. The server and API will be hosted at `http://nepaccess.org/`.

8 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. (1831551). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Amit Bronner and Christof Monz. 2012. User edits classification using document revision histories. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 356–366. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

91st Congress. 1970. An act to establish a national policy for the environment; to authorize studies, surveys, and research relating to ecological systems, natural resources, and the quality of the human environment; and to establish a board of environmental quality advisers. Public Law 91-190. 83 Stat. 852.

Johannes Daxenberger and Iryna Gurevych. 2013. Automatically classifying edit categories in wikipedia revisions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 578–589. Association for Computational Linguistics.

Tanya Goyal, Sachin Kelkar, Manas Agarwal, and Jeenu Grover. 2017. An empirical analysis of edit importance between document versions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2784. Association for Computational Linguistics.

S. R. Jonnalagadda and P. Topham. 2010. NEMO: Extraction and normalization of organization names from PubMed affiliations. *J Biomed Discov Collab*, 5:50–75.

Subhradeep Kayal, Zubair Afzal, George Tsatsaronis, Sophia Katrenko, Pascal Coupet, Marius Doornenbal, and Michelle Gregory. 2017. Tagging funding agencies and grants in scientific articles using sequential learning models. In *BioNLP 2017*, pages 216–221. Association for Computational Linguistics.

Martin Potthast, Tim Gollub, Matthias Hagen, Jan Graßlegger, Johannes Kiesel, Maximilian Michel, Arnd Oberländer, Martin Tippmann, Alberto Barrón-Cedeõ, Parth Gupta, Paolo Rosso, and Benno Stein. 2012. Overview of the 4th international competition on plagiarism detection. In *CLEF2012 Working Notes*.

Horst W. J. Rittel and Melvin M. Webber. 1973. Dilemmas in a general theory of planning. *Policy Sciences*, 4(2):155–169.

W. Shen, J. Wang, and J. Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Çağıl Sönmez, Arzucan Özgür, and Erdem Yörük. 2016. Towards building a political protest database to explain changes in the welfare state. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 106–110. Association for Computational Linguistics.

Kalliopi Zervanou, Ioannis Korkontzelos, Antal van den Bosch, and Sophia Ananiadou. 2011. Enrichment and structuring of archival description metadata. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 44–53. Association for Computational Linguistics.