

1. Introduction: Going Beyond Information Retrieval for eRulemaking

As text production and communication becomes easier, people gathering and reviewing information-intensive commentary face an increasing challenge. For news analysts, opinion pollsters, advertising planners, and, not least of all, government regulation writers, it is no longer easy to obtain a bird's-eye view of the world, given the rich, interconnected, and sometimes very subtly different texts aimed at them by the public. Thus, where straightforward information retrieval in the past might have been sufficient to enable knowledge workers to manage large amounts of textual information, increasing volumes of text, coupled with increasing expectations on the depth and complexity of analysis, make sophisticated text processing techniques and tools highly desirable, and in some cases imperative. Of particular interest are techniques to identify peoples' opinions, pinpoint groups of stakeholders such as mothers or lawyers, or segment and cross-relate comments according to various criteria of similarity. To a large extent, such tools exist only in embryonic form, and have not yet been integrated or tested in real-world settings.

The process of government regulation writing is an ideal "testbed" in which to experiment with the development, integration, and deployment of such techniques. Government regulation writers are a class of knowledge workers largely unrecognized, yet highly important to modern society. Their function is to formulate, in a tightly scripted procedure, the rules and regulations that define the details of our laws. Thousands of regulation writers employed in approximately 200 federal agencies and other entities have been granted the authority to perform an increasingly complex and vital task. After issuing proposals, agencies have to consider feedback from interest groups, commercial enterprises, and the public at large. If the regulations are to withstand legal challenges, they must incorporate substantive comments accurately.

The rule-writer's task is both enhanced and greatly complicated by a characteristic of the American system known as "Notice and Comment Rulemaking." Section 553 of the Administrative Procedures Act (APA) passed in 1946 requires that: i) regulations are published in draft form, ii) the public is allowed to comment on the proposed regulations, and iii) the agency considers the comments before issuing final regulations. When the general public gets interested in an issue, the result is a deluge of public comments that must be analyzed relatively quickly. For example, recent regulatory efforts at the United States Forest Service attracted over 1 million public comments. Agencies must respond in the final regulation to every material issue raised in the comments; failure to do so invites a lawsuit and summary dismissal of the regulation by a federal judge.

As usual, modern technologies complicate the process. In the strictly paper-based era, agencies swamped with public comments assumed that most comments were in fact easy-to-spot form letters generated by interest groups. Agency personnel could identify form letters by sight and sort them by hand (often into piles or boxes), greatly reducing the volume of unique comments requiring analysis, but risking error. In recent years agencies have begun accepting comments via the Web or email. Electronic submission eliminates many costs associated with processing paper, which is an advantage. However, electronic submission also makes it easy to edit form letters. The public has quickly taken advantage of this ability to personalize generic form letters. Each change, especially each addition, to a form letter must be checked to determine whether it introduces a new issue that must be addressed by the final rule.

Modern technologies can provide new tools to help organize, analyze, and manage large volumes of public commentary; such tools, and their effects on the rule-writing process, are the focus of this proposal. Our preliminary work with government regulation writers, for whom the stakes are often high, indicates clearly that they could benefit significantly from appropriate IT tools. The writing of regulations requires decision-making in a politically charged, information-intensive environment (Kerwin, 2003; Lubbers, 2003). One study estimates that regulations "aimed at protecting health, safety, and the environment alone cost over \$200 billion annually or about 2% of GDP" (Hahn & Litan, 2003, p. 2).

However, the requisite IT either is nonexistent, or has not yet been integrated, tested, or formally evaluated. Our proposal combines basic research on text analysis and empirical social science research on the effects of new technologies on the rule-writing process. Better text management tools have the potential to: lower the costs, and improve the consistency and quality, of processing public comments; produce more durable regulations; and provide better methods of communicating information to (and from) the public and other stakeholders. Our goals are both to improve the state-of-the-art in human language technologies and to assist materially the processes of public administration so that these potentials can be realized.

We recognize that the task is challenging. The PIs have studied the problem, individually and collectively, for a number of years. We have published papers on the policy implications of accepting public comment in electronic form (e.g., Shulman, Thrane & Shelley, 2004); organized focus groups and workshops with rule-writers and a diverse group of stakeholders (e.g., environmental and business groups, as well as labor unions) (Thrane, 2003); created a testbed of public comments for use by research communities; begun initial research on text analysis tools that might be of use for these problems; and made presentations about the underlying technology and potential solution paths to regulatory agencies and at professional meetings. We understand the problem, we have the technological and social science skills to complete the proposed research, and we have engaged partners at a range of agencies and in a variety of public interest groups to help us evaluate our research.

The focus of this proposal is research leading to and evaluating a **Rule-Writer's Workbench**, in which a variety of text analysis tools can be applied to help analysts, singly or jointly, master a mountain of incoming text to produce comprehensive, detailed, and densely cross-indexed analyses. Our initial work with the DOT and EPA suggest that the following capabilities are important:

- **basic information retrieval**, for gathering relevant texts both within and outside the docket;
- **text classification**, for channeling comments to the appropriate rule writer's desk;
- **overall text characterization using word frequency counts**, for identifying key issues;
- **duplicate detection**, for quickly identifying form letters;
- **near-duplicate detection**, for identifying and extracting text changes to form letters;
- **text summarization**, for creating a first rough cut through the data;
- **author typing**, for stakeholder analyses during and after a public comment period;
- **opinion/affect determination**, for determining what stakeholder concerns exist; and
- **document partitioning and cross-indexing**, for connecting comments to sections of regulations.

Although our focus in this proposal is on a specific problem, i.e., the regulation writer's tasks, our larger goals in information retrieval and human language technologies are a set of text access and analysis tools that are generally applicable across a wide range of language processing tasks. In particular, the areas of automated opinion analysis and author typing are almost completely new in the field; to our knowledge, there has been no workshop or book devoted to author typing, and only one public workshop.

Text access and analysis tools are becoming increasingly common. However, their social impact—in this case, their impact on the quality and validity of the government regulations that are produced using them—is unknown and barely studied. The social science component of our research will establish research methodologies for measuring these impacts.

The development of IT techniques and evaluation measures, and their deployment and evaluation in a workbench, will be useful not only to government regulation writers, but also to academics, interest groups, individual commentators, news analysts, historians, federal judges, business trend watchers, and even college and high school students. The underlying technologies would be of interest to anyone who must analyze large volumes of text. These tools will change substantially the manner in which public commentary is parsed and incorporated into the basic rules that govern American society.

1.1. Background: The World of Regulations

Controversial regulatory actions invariably result in massive amounts of public comment. Electronic public comment opens the floodgates wider, but in ways that can inadvertently undermine the intent of public participation. In one instance—the USDA’s national organic standard—a small team of rule writers in the late 1990s faced the task of sorting manually over a quarter million public comments (Shulman, 2003). The number of comments for such controversial rules is likely only to increase.

Interest groups have implemented their own information technology (IT) applications to: educate and mobilize supporters, bombard agencies with comments, delay regulatory action, gain favorable publicity, and recruit members. However, such efforts may dilute the voice of the public as agencies face statutory and administrative deadlines to incorporate public input into a defensible final rule. The tools we propose to develop and evaluate may increase the likelihood that all public input is analyzed in a way that does not compromise the importance of public involvement. End users in and out of government will have a rare chance to beta test the technologies that will help make sense of public commentary.

Under the eRulemaking Initiative, the federal government has developed a three-stage plan for technology development to support rule-writing. Module 1, the first step, is a simple portal that allows access to all federal rules open for public comment and provides a Web-based form for submitting comments to the appropriate federal agency. A new federal portal, at www.regulations.gov, opened on January 23, 2003, with the EPA acting as the managing partner (Skrzycki, 2003). Later, the eRulemaking Initiative will move all federal eRulemaking to a single, unified docket system (Module 2) and unveil a set of tools (Module 3) for the rule writers. The research proposed here is directly relevant to Module 3 of the federal eRulemaking Initiative.

1.2. Current Technology in Use

The ad hoc shift to eRulemaking at a number of agencies has resulted in a somewhat uncoordinated mixture of in-house and external commercial uses of technology. Primarily, the track record to date is that IT has been deployed to alleviate some of the mundane text collection and storage tasks. For example, several agencies now routinely digitize their collected public comments into PDF format, but employ no technology to process this electronic form further. Almost all still require paper archiving of all materials; some are still all paper-based. A review of the commercial off-the-shelf content analysis tools currently employed by agencies and contractors reveals that no firm or agency has effectively integrated human language technologies into the content analysis procedure. At the DOT, which was an early and successful adopter of an agency-wide Docket Management System (DMS), there is no method for conducting full-text search within a docket. At the EPA, where a newer EDOCKET system provides the model for the new federal eRulemaking architecture, navigation through the docket is difficult at best.

1.3. Current State of eRulemaking Theory and Knowledge

Although some milestones have been passed in recent years, the eRulemaking research domain is in its infancy. An interdisciplinary research workshop was held about 13 months ago at Harvard’s Kennedy School of Government (Coglianese, 2003a). Some scholarly eRulemaking research plans have been funded on different occasions, by different divisions in the NSF, and have been published in academic journals (Garson, 2003; Lubbers 2002; Shulman et al., 2003). A growing body of eRulemaking scholarship focuses on whether Internet-enhanced public participation results in better rules (Carlitz & Gunn, 2002) or in a process characterized by informed deliberation (Brandon & Carlitz, 2002). Democratic theorists and administrative law scholars have begun to take notice that the 2,500-year-old quest for meaningful participation in decision-making may be entering a completely uncharted phase (Garson, 2004; Johnson, 1997; Zavestoski & Shulman, 2002). Key players in the development of the federal architecture are welcoming all the help they can get building and assessing new tools that will shape citizen-government interaction via eRulemaking (Heyman, 2003).

Building upon these studies, this proposal represents the next major milestone in the development of multi-disciplinary, inter-agency collaboration. In Section 2, we describe the proposed IT, and in Section 3, the concomitant IT evaluation plans. Two of the toughest tasks over the brief history of the Digital Government research domain have been: (1) the integration of computer and social science research and (2) the development of a realistic plan for technology transfer from research to agency practice. In Section 4 we suggest an approach that outlines solutions to both these obstacles. Finally, Section 5 lists the potential project contributions for both IT and social science research domains.

1.4. IT Research in Process or Completed

Our preliminary funding (“SGER Collaborative: A Testbed for eRulemaking Data”) produced several new resources that support our research, and are available for use by others studying text mining, text analysis, and/or the policy and political impact large, public comment datasets.

The **eRulemaking Testbed** web site¹ distributes public comment process data for use by the research community. As of this writing it contains data for eight regulations proposed by the Department of Transportation, the Environmental Protection Agency, and the Department of Agriculture; we constantly solicit new datasets. Proposed regulations, public comments, and final regulations (if any were issued) are available on the site. The site also provides search access to the datasets (requested by some of the agencies that provided the comments) and several simple text analysis tools. Access is controlled, to avoid privacy problems caused by search engines indexing the public comments, but access is granted to any researcher who requests it.

The **eRulemaking Research** web site² is a clearinghouse for conference papers, manuscripts, workshop information, and presentations related to the Project Director’s eRulemaking research.

An **eRulemaking Workshop** in September 2003 with government rule writers (first day) and interest groups (second day) produced a pair of evaluation reports disseminated via the Internet (Thrane, 2003). Other workshops have focused on the government perspective; we believe this is the first time anyone solicited the opinions of the activists and interest groups (that generate most of the form letters).

In addition to these tangible results, Callan and Hovy have done preliminary analysis and feasibility studies with the public comment data collected so far. Callan’s research has focused primarily on simple statistical analyses for identifying important topics and stakeholders represented in a set of comments. Hovy has embedded comments from a DOT rulemaking initiative in C*ST*RD (Lin et al., 2003), a testbed environment that supports clustering, summarization, and visualization. Hovy also has analyzed the feasibility of applying opinion-detection technology to typical comments.

2. IT Research: Tasks, Techniques, Technologies, and Glass-Box Evaluation

In this section we describe the proposed IT research. We proceed from work with relatively immediate applicability to increasingly challenging functionality. As described in Section 4, as the work progresses, we will deploy it in a testbed Rule-Writer’s Workbench for annual performance evaluations with our government partners, and will use the feedback to drive subsequent research.

2.1. Important Concepts

When comments are received, one of the first tasks is to identify the issues that they (as a group) address. Usually an agency has expectations about what will be commented on. A first priority is to discover whether those expectations are correct. Agencies often are particularly interested in comments that address unexpected issues. Agencies are also interested when they don’t receive comments on issues that were expected to be important, because it may indicate that key constituencies are missing from the comment pool; in such cases agencies often solicit comments from the missing constituencies.

¹ <http://www.cs.cmu.edu/~callan/Data/eRulemaking/>

² <http://www.drake.edu/artsci/faculty/sshulman/eRulemaking/>

How best to identify important concepts is an open question. A common approach is to count the frequencies of bigrams (two-word sequences), sort by frequency, and display the most frequent. Another common approach is to assign part-of-speech tags, for example using the Brill tagger (Brill, 1999), count the frequencies of noun sequences, sort by frequency, and display the most frequent. Both approaches are good at identifying high-profile concepts, but agencies are likely to anticipate these concepts already. Unexpected concepts are more likely to occur with medium or low frequency, mixed in with many useless low frequency concepts. A research priority is identifying meaningful medium and low frequency concepts, based upon, for example, collocation with other, more frequent concepts or emotional language.

The proposed regulations themselves can be mined in the same manner, for example, to find concepts that might be expected to occur in the comments. Knowing how well the comments cover the regulations might be another way of identifying areas in which greater comment should be solicited.

Once an issue, for example “sewage sludge,” is identified as being a frequent topic of public comment, the question is how to drill down into that part of the corpus to see what people are saying about it. There are a variety of approaches that can be employed, but it is impossible to say a priori which will prove most effective. Possibilities include:

- Applying frequent concept analysis recursively within the set of comments that mention the concept, to generate more fine-grained analysis that links eventually to individual comments;
- Comparing statistical language models (word histograms) of the comments that contain the concept and the corpus as a whole, displaying as a summary the areas in which the divergence is greatest; and
- Determining co-occurrence relationships (e.g., using a mutual information measure) between individual concepts and individual stakeholder types (discussed in Section 2.4) to identify which issues are important to which stakeholder groups.

Other techniques that might also be appropriate to this task—for example, clustering, summarization and opinion detection—are discussed in later sections.

2.2. Near-Duplicate Detection, Difference Extraction, and Clustering

Form letters (“exact duplicates”) and edited form letters (“near duplicates”) are a major problem for the regulatory agencies. The problem is exacerbated by agency processes, such as stamping dates across the “content” part of paper copies, merging metadata (sender, date, etc.) information with the content itself, or converting electronic text to PDF in ways that make the original ASCII difficult to extract. Our current testbed contains all of these problems. Regular contact with people doing automatic analysis of the text, over time, would encourage the agencies to redesign their processes to clear up the simpler problems. However, at its core this is a difficult issue for which no off-the-shelf solution exists.

The first step is to identify exact duplicates in electronic text, for example by stripping off header (e.g., address or metadata, salutation) and footer (e.g., signature block) information that is embedded in the text, normalizing white space, clustering by length, and comparing word sequence “fingerprints.” Frequency is likely to indicate the reference copy of a form letter with relatively high reliability.

Some agencies use a +/- 10% length rule to identify potential near duplicates. Our preliminary research indicates this heuristic is highly unreliable. Most form letters are 1–2 pages long. People routinely add and delete paragraphs in form letters, often changing length by far more than 10%. Instead, multiple fingerprints must be created for each document to identify its near duplicates. Edit-distance is unreliable at the beginnings and ends of comments, where people routinely make large changes, but it may be more useful in the middle of comments. Once near-duplicates are detected, sentence-level differencing is sufficient to identify additions, deletions, and modifications. Regulators say that they aren’t interested in finer-grained analysis; if a sentence was changed, they want to see the whole sentence.

Some agencies, such as the Department of Transportation, routinely scan and OCR comments submitted on paper. OCR errors make exact- and near-duplicate detection much more challenging. More imaginative solutions will be required to address those problems, for example, using combinations of shorter fingerprints to identify exact- and near-duplicate candidates, and then applying noisy channel models and statistical language models (Brown et al., 1993) to estimate the probability that two candidate passages differ only due to OCR errors. Such models can be trained by generating documents electronically, printing them, scanning them, aligning the original and OCR texts using correctly-recognized words, and using current parallel-corpus techniques for statistical machine translation to learn the probabilities of different types of errors. This approach isn't perfectly accurate, but perfection is not required. Today this task is done manually; any degree of automation will be an improvement.

2.3. Comment Decomposition and Cross-Indexing

Anyone trying to analyze responses to several issues simultaneously faces the problem of unscrambling the commentary and cross-referencing each comment to the relevant topic(s). Regulation writers today all indicate that finding all comments applicable to a specific proposed regulation change is a significant problem.

Automatically segmenting text into useful segments, for cross-referencing and a variety of other purposes, has long been a goal of human language technologies research. Unfortunately, as shown in numerous studies, there are multiple, not always corresponding, cues that indicate segment boundaries, including discourse structure markers (Mann & Thompson, 1988); syntactic features (Polanyi, 1988), speaker intentions (Grosz & Sidner, 1986), and prosodic contour (Grosz & Hirschberg, 1992). In addition, systems such as the groundbreaking TextTiling (Hearst, 1997) and C99 (Choi, 2000), which work quite well in some domains/genres, rely not on these cues but on over-threshold changes of word usage frequencies across paragraph boundaries. The most relevant work has been done by Law and colleagues at Stanford, under DG research (Lau et al., 2003). In a very ambitious project, they outline the in-depth semantic parsing of regulations and the cross-reference of related clauses, using first-order predicate calculus formulations of the regulations, to find correspondences and anomalies. While we applaud this work, we are looking for robust technology that can be applied at large scale.

All the above studies have been done on rather formal text. We are interested in determining which cues work reliably in less formal genres such as email. The emails submitted as rulemaking commentary provide a somewhat unusual, but intriguing, opportunity to study segmentation of such less formal text. Our challenge is to segment comments, in particular emails, so that each segment can be aligned exactly and only with those portions of the proposed regulation changes to which they pertain. To do so, we can employ the longer, more carefully written comments, usually submitted by lawyers and relevant industry groups, which tend to refer explicitly to relevant portions of the regulations by their subsection numbers. By finding segment cues in the longer texts, and then finding the incorporated argumentation reflected in the emails, we can map the segment boundaries over to the emails as well. We propose to address this challenge as follows:

- As baseline, we will first deploy our implementation of TextTiling upon the test corpora that explicitly contain reference (by number) to the regulation under consideration. We will use these references to calibrate the algorithm's segmentation threshold parameters, to maximize the confluence of segment breaks between references to different portions of the regulation.
- Next, we will experiment with various types of language models to characterize both the portions of the regulations under discussion and the corresponding comment segments. We will use traditional ngram language models as well as topic signatures (Lin & Hovy, 2000) to determine the required sensitivity thresholds at which segment boundaries should be introduced, and at which regulation section and comment segment may be aligned.

- Combining these parameterizations, we will proceed to map the email contents into the longer, more formal comments, and induce the segmentations in the email.
- Finally, we will deploy machine learning algorithms, such as C4.5, SVM, and others, upon the features we can identify in the emails as likely to be relevant, including keywords, paragraph breaks, discourse markers such as “also,” and so on.

We hope this work will be useful both to the regulation writers and to other human language technologies researchers interested in handling email. Fortunately, for regulation writers, we can err on the side of recall, because their current situation, in which regulation writers have to scan *everything*, is the worst possible case: we can only make things better by suggesting an ordering on what they read first.

2.4. Author and Stakeholder Typing

Comments written by individuals often identify a person’s role relative to the proposed rule. For example, comments on the 1997 National Organic Program included “...I have the right to know as a **mother**,” “As a **biologist** I know the hazards of heavy metal toxins...,” “I am a **biologist** and the impact ...,” “My husband and I are both **wildlife biologists**...,” and “As a **chemist**, my only comment...” Such information provides more detail to regulators about the constituencies represented in a set of comments. Regulators sometimes seek this information manually when it is expected to be important—for example, to make sure that flight attendants and pilots are both represented in comments about flight safety standards—but in general the cost of acquiring it discourages any consistent use. Regulators report that Members of Congress frequently seek such information.

Simply scanning for lists of occupations and expected roles doesn’t work, because the public routinely mentions other occupations, for example “Organic means ‘from nature without the interference of the **chemist**’,” “I am only a **consumer** and not a **chemist**...,” and “...present day cancers are purported by **medical doctors** to be....”

Preliminary research indicates that machine learning algorithms can be used to create pattern-based rules that detect and extract stakeholder expressions. The more difficult problem is generalizing them and organizing them hierarchically so that interesting groups are revealed. The head nouns of phrases can be recognized relatively easily, enabling expressions such as “consumer of organic foods” and “longtime consumer of organic produce” to be recognized as two different specializations of “consumer”. WordNet (Fellbaum, 1998) can positively identify and indicate relationships among some types of careers and roles. Statistical techniques designed for creating subsumption hierarchies (e.g., Sanderson & Croft, 1999) can identify generalization/specialization relations based on word overlap and co-occurrence information, albeit with less accuracy. Rule-writers can improve the accuracy of these techniques over time by interactively providing feedback about which candidate stakeholders are “real” and which are artifacts of the recognition process (“false positives”), thus “tuning” it over time.

Just knowing which stakeholders are represented in the comments is valuable information that rule writers don’t have now. However, stakeholder information can be combined with other analyses to provide a more detailed view. Significant co-occurrence relationships between stakeholder groups and frequent concepts (Section 2.1) can be identified, or stakeholders can be clustered by their opinions (Section 2.5), to quickly provide an informative picture of what different groups of stakeholders believe.

2.5. Opinion Detection and Clustering by Opinion

Over the past 18 months, the challenge of automatically identifying expressions of opinion in text, and reliably classifying texts accordingly, has surfaced in several communities, including the web search community (who want to identify and possibly separate out blatantly opinionated material), the intelligence/homeland security community (who want to be able to find people with troublesome opinions), and the eRulemaking community (who want to be able to identify the range of opinions

expressed with regard to candidate regulations, to determine the relative strengths of each, and to then adjust their regulations appropriately).

What exactly an opinion is remains a topic for philosophy and rhetoric. Our preliminary reading of a small selection of the available literature (Aristotle, 1954; Toulmin, 2003; Toulmin et al., 1979), as well as our own text analysis, indicates that a profitable approach to opinion requires a system to know and/or identify at least the following elements: the topic (T), the opinion holder (H), the belief (B), and the opinion valence (V). For the purposes of the various interested communities, neutral-valence opinions (such as: *we believe the sun will rise tomorrow*; *Susan believes that John has three children*) is of less interest; more relevant are opinions in which the valence is positive or negative. Such valence usually falls together with the actual belief, as in *going to Mars is a waste of money*; in which the word *waste* signifies both the belief *a lot [of money]* and the valence *bad/undesirable*.

Recent work in human language technologies (Kim & Hovy, in prep.; Turney & Littman, 2003; Wiebe et al., 2002; Yu & Hatzivassiloglou, 2003), indicates that systems can be trained to recognize opinion, subjectivity, and even valence, with some accuracy. The 2003 NIST-sponsored TREC evaluation competition featured one track (Soboroff & Harman, 2003) that included the following test: Given 50 newspaper texts for each of 25 controversial topics, can the system reliably identify all and only those sentences that express an opinion regarding the topic? Our best-scoring system (Kim & Hovy, in prep.), with an F-score of 0.597 (precision 53%, recall 83%) scored second-highest of the 55 systems submitted by 14 groups from around the world. Based on this promising result, we have been conducting experiments with both opinion detection and valence classification. Our approach is to train a unigram-based classifier on various training corpora (one derived manually; a second extended by reference to WordNet; another derived from the Wall Street Journal, which, following Yu and Hatzivassiloglou, we split into 7,053 opinion-bearing documents (editorials, letters) and 166,025 non-opinion ones (news, etc.)). Even with such crude seed material, we obtain intuitively reasonable characterizations of words, as shown in Table 1.

Adjectives	Valence score	Verbs	Valence score
careless	0.638	harm	0.617
wasteful	0.500	hate	0.539
degraded	0.428	spoil	0.500
unpleasant	0.153	yearn	0.500
southern	-0.275	enter	-0.487
vertical	-0.500	crack	-0.500
Scored	-0.587	combine	-0.585
Initial	-0.624	purchase	-0.664

Table 1. Words with opinion valence (+1 = undesirable; 0 = neutral; -1 = desirable).

In current work, we are experimenting with different valence prediction models (including, for example, summing the valence scores of all words in a sentence; comparing just the most positive and most negative valence scores; considering the valence median; etc.). The best of our current models provides Precision and Recall scores (against human judges) of 63.5% and 75.5% respectively, which compares well with inter-human agreement.

Encouraged by these preliminary results, we propose in this work to identify and classify the opinions expressed in comments submitted to regulation writers. Since these documents are inherently opinionated, we expect this to be a rich source of data. We will proceed as follows:

- Perform initial opinion detection, using our existing methods. Perform initial evaluations with the regulation writers to determine how satisfactory this is, and to gather suggestions for further lines of inquiry.

- We expect that long comments will contain many, sometimes opposing, opinions. If needed, and in conjunction with comment decomposition (see Section 2.3), we will develop methods to decompose comments into zones of homogenous topic and opinion.
- Armed with human judgments of these zones, train new classifiers (wordlists, and, if needed, others, such as decision lists, decision trees, Bayesian classifiers, etc.), as appropriate in the regulation writer domain. By the nature of the enterprise, we expect to find much less ambiguity of potential opinion-bearing words than in general text. For example, “expensive” might carry positive connotations in the domain of luxury goods (furs or automobiles), but probably never will do so when applied to the effects of regulations.
- Apply these classifiers to new comments, to partition them into at least three categories: For, Neutral, and Opposed. Should subsequent evaluation studies with regulation writers (see Section 4) indicate the need for more fine-grained classes, we will experiment to find the appropriate thresholds of the valence scores.

Given the novelty of automated opinion and valence detection, almost any research in this area will be welcomed by the human language technologies community. Its applicability to a wide range of areas makes it an exciting and potentially valuable component of the proposed work.

2.6. Text Summarization for Draft Rule Writer Response

Research on automated text summarization was conducted first in the 1950s and 1960s, and, after a hiatus of some 30 years, has received new impetus. However, the vast majority of recent summarization research was applied to news. Inherently, given the stylized conventions of newspaper writing, this made automated summarization less challenging than it should be.

Regulation comments, in contrast, have no general structural conventions. They range from very brief—perhaps just 50 words—to quite extensive—perhaps 500 pages of study. They range from informal, often barely literate, emails, to exhortatory bulletins, to carefully reasoned scientific or sociopolitical treatises. The challenge for automated summarization is significant: Can a system reliably identify and compress all important material, taking into account different modes of expression when removing redundancies?

We do not believe that three years of work in a small project can solve this problem. But we believe that even current-day summarization techniques can be of use to regulation writers, and that these techniques provide a springboard from which to investigate the deeper issues of recognizing alternative phrasings and of compressing them. Even limited-functionality summarization can help with two different parts of the regulation writer’s task:

- during the comment analysis phase: rapidly identifying the portions of current interest in one or more comments (or any length); and
- during the summary/response writing phase: gathering all currently relevant portions of comments so as to suggest suitable phrasing to use in the summary.

Since 1995, we have gathered considerable experience with text summarization systems at USC/ISI. SUMMARIST could accept English, Spanish, Korean, Japanese, Bahasa Indonesia, and Italian newspaper articles (Hovy & Lin, 2000). It served as the testbed for a wide variety of summarization strategies (Hovy & Lin, 1998). NeATS is a multi-document summarizer that introduced a novel technique for ensuring rhetorical cohesion (Lin & Hovy, 2003). NeATS has been a consistent top-scorer (first or second place) in NIST’s annual DUC summarization evaluation tests (Harman & Over, 2003). GOSP is an experimental summarization system that produces short (headline-length) summaries characterizing one or more texts devoted to a topic (Zhou & Hovy, 2003). In the first DUC headline evaluation test, GOSP was one of the top performers, using the length-normalized score (Harman & Over, 2003).

We plan to experiment with summarizing comments during Years 2 and 3 of the contract. As training material, we will use the (publicly available) responses provided to comments to earlier regulation writing efforts. We already have these comments (as described in Section 2.3) and will obtain additional ones from our government partners, as needed. We will focus on two principal research questions: *summarizing across genres* and *summarizing across extreme ranges of source length*.

- With respect to genres, first we will separate the major types of commentary (individuals' emails, exhortatory (form) letters and near-duplicates (see Section 2.2), scientific studies, etc.). Then we will apply the NeATS sentence scoring heuristics to each class separately, since it is likely that different heuristics will dominate in different genres (for example, keywords/phrases in form letters, the position policy OPP (Lin & Hovy, 1997) in studies, etc.). If needed, we will develop additional heuristics pertinent to this domain, such as rewarding sentences or sections that refer explicitly (by number) to portions of the regulations under discussion, or that explicitly cite recognized authorities on relevant questions. Throughout, we will train the heuristic combination function (Lin, 1999) by reference to the regulation writers' summaries. In the evaluations (Section 4), we will present variations, as suggested by our research, to regulation writers, to solicit their opinions.
- With respect to source length, we will investigate the automatic grouping of source material into a so-called topic structure, in which the major topic heads are fleshed out incrementally by details. This work builds on interesting experiments by Moens and colleagues in Belgium (Moens & De Busser, 2001), and relates to attempts to use summarization to deliver lots of material on small interfaces such as PDAs through incremental expansion. We will investigate the assumption that when a short email and a long study on the same topic actually say the same thing, the former will find correlates in certain recognizable portions of the latter (its abstract, headings, conclusion, etc.). If so, this will allow us to try to merge short texts into long ones at the appropriate positions, the result of which—an even longer text—then will be input into the single-document summarizer for normal summarization. (The duplication of certain portions, of course, will sway the content of the summary, as indeed it should.) Also in this area, we will be guided by past regulation writers' summaries, and by the comments of our government collaborators.

Neither summarization across genres nor summarization across extreme length variations has received much attention in the literature. We expect that this work can help form a foundation for increased research in non-news summarization in the international community, which currently is writing a roadmap document to survey the likely future avenues of research.

3. Political Science and Sociological Research: Overall Evaluation

With staff support from the University Center for Social and Urban Research at the University of Pittsburgh, and external project evaluation and data analysis from experts at Iowa State University, PIs Shulman and Zavestoski will employ several methods to evaluate the effectiveness of the tools described in Section 2. The first step will be to work with the members of our federal agency Advisory Group to recruit a wider group of content analysts and rule writers. We will interview and work closely with personnel at various federal agencies, to gather public comments from open rulemakings and incorporate them into the testbed. This will enable rule writers in our partner agencies to begin employing the tools available at the testbed.

Every user of the testbed would register (with anonymity, if desired) as either: (a) agency personnel, (b) interest group member or representative, or (c) member of the general public. A short web-based exit survey would ask a small number of Likert-type questions about testbed functionality, and open-ended questions to solicit further feedback about the future shape and impact of the testbed tools. It is anticipated that these pretest surveys and focus groups could begin this summer, under the auspices of existing funding.

Structured and semi-structured feedback from testbed users will be collected at various points during the life of the project. Initial evaluation will use focus groups, workshops, and web-based surveys. Focus groups will be used to attempt to identify what works and what does not. Group techniques for converging to a consensus solution, such as the Delphi method, will be employed to benchmark the metrics that the rule-writers themselves identify as most useful for determining the effectiveness of these tools. Workshop participants will be debriefed thoroughly, to ascertain the maximum achievable “deep structure” of thought processes and potential problems underlying testbed applications, and workshop results will be checked closely for alignment with the intent of testbed implementation.

These data will be central to the development of a pretest-posttest quasi-experimental design. The posttest will include self-reported satisfaction with the testbed tools, as measured by a more detailed Web-based questionnaire including items addressed in the pretest instrument, as well as objective measures of the speed and accuracy of testbed performance. Rule writers working on the same or similar rules will serve as a control group to compare with the satisfaction and performance of the rule writers employing the testbed tools, adjusting for relevant covariates that may differ across the two groups. The follow-up web-based surveys will combine demographic, attitudinal, and behavioral variables, using mostly Likert-type fixed-response questions and a small number of open-ended items. Construct validity will be examined through factor analysis methods, and reliability of resulting summated rating scales will be established through Cronbach’s standardized item alpha coefficients.

Finally, we will collect click-through data about users of the eRulemaking testbed. These “session cookie” data will provide informal measures of the extent to which each tool/technique is used, in what combinations and orders they are used, and where there may be problems due to usability or interface issues. Ultimately, these results will be combined to provide a detailed picture of the efficacy of the IT for the regulation writer task, and by extension to other commentary analysis tasks. Introducing this kind of evaluation into social science research, where it is somewhat novel, may help ensure that innovative social science research both informs the development of federal eRulemaking practices and becomes more widespread in the disciplines of political science and sociology.

Analysis of the collected data will be undertaken using SPSS and SAS statistical software, applying methods appropriate to the sample size of the dataset and measurement properties of the variables. It is anticipated that a combination of statistical methods for categorical data (e.g., logistic regression models, general loglinear models, and crosstabulation methods) and linear models (multiple linear regression, analysis of variance and analysis of covariance, and structural equation models) will be used, as appropriate.

4. Project Structure and Government Collaboration

This proposal is the culmination of three years of consensus and team building by PI Shulman, who acts as the overall Project Director. His networking and current research targets personnel involved in the development of eRulemaking research and practice (Shulman, 2004b). Two Digital Government-sponsored small grants for exploratory research (described in Section 6) have resulted in the development of an ad hoc advisory group comprised of federal agency personnel. In this section, we discuss current and recent collaboration with federal agencies and summarize the project structure and management plan.

4.1. Foundation: Past Collaboration with Government Agencies

The original collaboration began in the fall of 1999, when the USDA’s National Organic Program (NOP) shared a dataset of over 20,000 public comments received via the Internet during their rulemaking. Subsequently, under the auspices of an SGER (EIA 0089892), a first annual workshop was held in May 2001 at the Council for Excellence in Government. A second workshop was held in June 2002 at National Defense University.

In January 2003, the eRulemaking Research Group (ERG) was formed consisting of the four PIs—Callan, Hovy, Shulman, and Zavestoski. In February 2003, the ERG submitted a successful Collaborative

SGER to the Digital Government program, with Oscar Morales, Director of the eRulemaking Initiative, as the government partner. Subsequently, Mr. Morales participated in regular phone calls, attended the dg.o2003 conference in Boston, and supplemented the Collaborative SGER with a \$30,000 contribution. A third workshop was held at the National Science Foundation in, Arlington, VA, in September 2003. The ERG presented its plans for future text analysis research, and conducted focus groups with agency personnel one day and interest groups the second day. Subsequently, the ERG was invited to planning meetings at the DOT, EPA, U.S. Forest Service's Content Analysis Team (USFS-CAT), and the federal eRulemaking Initiative. The focus group reports from the fall 2003 workshop were used as background documents for a January 2004 eRulemaking workshop held at American University.

The Project Director has been in regular communication with members of the project's Agency Advisory Group, which includes Oscar Morales (eRulemaking Initiative), Neil Eisner (DOT), Andrew Malone (USDA-APHIS), Sharon Whitt (NARA), Stuart Miles-McClean (EPA), Jody Sutton (USFS-CAT), and Carl Zulick (BLM). The Advisory Group met February 18, 2004 to finalize arrangements for collaboration and to shape the final proposal. Members of the Advisory Group will meet twice annually in Washington, DC and communicate regularly with the Project Director regarding the status of the project.

4.2. Government Collaborators

As shown in the attached Memoranda of Understanding and letters of intent, we have been obtaining assistance and commitments for future collaboration from several federal government agencies. The collaboration includes: help with proposal preparation, data sharing, access to personnel, participation in evaluation, technical support, commitments to seek out year funding, supplements for the current SGER, and advisory group participation. We anticipate a pending agency supplement to the SGER Collaborative will result in a summer 2004 workshop, which will allow the team to demo the existing testbed tools, setting the stage for a coordinated start of the four-year project during fall 2004.

4.3. Project Structure and Management

This project integrates social and computer science research. It calls for the cyclic and incremental development, testing, and refinement of IT to perform increasingly in-depth and sophisticated analysis of public commentary, to be deployed and evaluated annually as an application for eRulemaking research and practice. We will follow a coordinated research effort across the ERG's four PIs:

- Human Language Technologies research: Callan (CMU) and Hovy (USC/ISI): technology prototypes and annual integration and deployment (Sections 3.1 to 3.6)
- Political Science and Sociological research: Zavestoski (USF) and Shulman (Pitt): annual overall evaluation of new IT with regulation writers; studies of effects on their practices (Section 3)

The four PIs have been in regular communication since a January 2003 DG-sponsored eRulemaking workshop held at Harvard's Kennedy School of Government. We since have developed a close, collegial relationship. Each PI will manage the budget and the work of students and/or programmers at his institution. In addition to annual meetings at dg.o and the continuing eRulemaking workshops, the ERG will meet twice a year with the members of the Agency Advisory Group.

Overall project management will be the responsibility of the Project Director. Shulman will coordinate all interaction and collaboration between the ERG and federal agency partners. The Project Director is responsible for continuing and enhancing the workshops, supervising the preparation of yearly reports to NSF, and facilitating regular telephone, email, and face-to-face communication between the group members. The Project Director will manage a budget that includes modest funding for regular stakeholder workshops in the DC area and he will solicit agency sponsorship for more frequent meetings.

Our cyclic research-deployment-evaluation development plan, listed in the table below, calls for a new, refined and extended, version of the testbed toolkit, and subsequent evaluation in context every year.

The findings of the evaluation will be fed back into the research of the subsequent year. To start this process, we have already put in place a zero-order prototype, which will be the subject of the first evaluation.

4.4. Technology Transfer

Having participated in DG grants before, the PIs are well aware of the difficulties of technology transfer. Our strategy has three thrusts. First we will develop and deploy the new technology, starting with the most immediate and retaining it in the Rule-Writers Workbench throughout, thereby allowing the more stable tools and their interface to mature and allowing the regulation writers to become thoroughly familiar with them. Second, we will seek to establish an understanding with the federal eRulemaking Initiative under which we would transition to them the more mature and desired technology, as deployed, tested, and evaluated in the Workbench, during the last six months of the project. Third, we will work with Lockheed Martin, the contractors for Modules 2 & 3 of the government’s eRulemaking Initiative, to explore methods of their licensing our technology and taking it further into Module 3. Exploratory discussions along these lines, held in October 2003, have been positive.

4.5. Project Timeline

Date	Task
Fall 2004	Develop and deploy Round I Tools based on past work Recruit and finalize agency participants; collect and post appropriate data to testbed Conduct interviews and “pre-test” subjects
Spring 2005	Analysis of Round I Feedback from rule-writers, public & interest groups Further refine IT tools
Summer 2005	Hold eRulemaking workshop and conduct focus groups Deliver interim report to agency partners
Fall 2005	Develop Round II Tools with prototype IT Refine IT testbed based on evaluation feedback
Spring 2006	Analysis of Round II Feedback from rule-writers, public & interest groups Further refine IT tools
Summer 2006	Hold eRulemaking workshop, administer survey, and conduct focus groups Deliver interim report to agency partners
Fall 2006	Develop Round III Tools with prototype IT Refine IT testbed based on evaluation feedback
Spring 2007	Analysis of Round III Feedback from rule-writers, public & interest groups Further refine IT tools
Summer 2007	Hold eRulemaking workshop, administer survey, and conduct focus groups Deliver interim report to agency partners
Fall 2007	Transfer technology to the federal eRulemaking Initiative
Spring 2008	Prepare a manuscript for an eRulemaking book
Summer 2008	Complete final report to agency partners and the NSF

5. Results: Dissemination, Impact, and Broader Social Benefits

All four PIs are highly experienced in submission of successful peer-review publications. The ERG expects to make contributions both to disciplinary and interdisciplinary conferences and journals based on the proposed research.

5.1. Dissemination Plan

Research results will be disseminated using two major channels. Results will be published in high quality research-oriented conferences and journals within Computer Science and Political Science, as well as in interdisciplinary conferences and journals, including, of course, the dg.o Digital Government conference and peer-reviewed eGovernment journals.

We will take pains to disseminate datasets, tools, bibliographies, and publications via the project's eRulemaking Testbed and eRulemaking Research web sites and the E-Rulemaking Clearinghouse (www.e-rulemaking.org) established by the Kennedy School. We see Web dissemination as crucial to our long-term research mission. It will make our research reproducible, allow Computer Scientists and Political Scientists to explore directions we could not, and broaden the research community by reducing the barriers to entry for new researchers.

5.2. Computer Science Research Impact

Recognized as important problems in Computer Science, text analysis and text mining have enjoyed a measure of success in recent years, with the advent of web search and mining companies. However, despite its obvious desirability for a large number of applications and areas, sophisticated text analysis technology, such as opinion analysis, author/stakeholder typing, document segmentation and cross-reference, and the others described in Section 2, are still a long way from being thoroughly understood. Objectively assessing the quality of performance of such tools is not trivial, either. The proposed research is an opportunity to deploy text analysis techniques in operational environments and to measure their effectiveness. It also will collect and package relatively large-scale public comment datasets for use by other researchers. These corpora will be significantly different from the other corpora used most commonly in the research community—for example, the TREC corpora distributed by NIST.

We believe that a carefully assembled sequence of techniques, deployed and evaluated in a real-world context, not only will open the way to further innovations, but also will help set a standard for the way in which IT can be introduced gradually into government agencies, eased by regular tests administered by political and social scientists.

5.3. Social Science Research Impact

Political scientists and sociologists have been slow to take up the interdisciplinary challenges inherent in digital government research. For instance, membership in the IT & Politics section of the American Political Science Association is about 210, despite a total association membership of over 15,000. As a result, leading disciplinary journals and textbooks too often do not reflect important social, governmental, and technological change (indeed, the top-tier journals in political science have yet to publish research on the impact of IT on American politics). We believe this project, as a close coupling of political scientists and IT researchers, working hand in glove with regulation writers and their managers, has a chance to help focus the attention of the American Political Science Association on IT. Our Project Director, Shulman, as President-Elect of the IT & Politics section, is well-positioned to work with other DG-funded political scientists (e.g., Dawes, LaPorte, Fountain, and Coglianese) to put the study of digital government in the core of disciplinary research and education. Already, these political scientists jointly presented a roundtable at the 2003 meetings of the American Political Science Association focusing on digital government research and its impact on the state of the discipline.

5.4. Interdisciplinary Research Impact

Collaboration between computational and social scientists raises unique challenges, as does work between academic and governmental personnel. Nonetheless, significant inroads have been made toward attaining all these goals over the past two years. Digital government requires interdisciplinary collaboration, to ensure that technological innovation meets the requirements of democratic institutions and traditions. This project will foster collaboration not only through research and publications, but also through regular presentations and workshops that bring together advisory group members, academic researchers, private sector technology designers, and federal agency personnel.

Evaluating technical systems for social applications such as rulemaking requires that agencies and university researchers collaborate across the traditional “stovepipe” barriers, whether they lie between or within agencies, or amongst the academic disciplines. According to a National Research Council report, *Making IT Better*, “[n]ontraditional research mechanisms may be needed that will encourage the

participation of end user organizations in research, broaden the outlook of IT researchers, and/or overcome disciplinary boundaries in universities” (NRC, 2000, p. 168). Our proposed research represents precisely such a mechanism. We intend to continue on this path, recognizing from prior experience the importance of devoting time not only to IT development and evaluation, but to constant and thorough liaison with government partners.

5.5. Societal Impact

Ultimately, the value of this research will be tested in two arenas. First, how much easier is it for the public to interact with the government’s cadre of regulation writers, and for the regulation writers to handle the public’s comments? Second, which of the newly developed techniques are used not only by regulation writers, but by others, such as news analysts, opinion pollsters, and so on, to analyze in detail the richness and complexity of public discussion?

It will take a long time before such questions can be answered. If just two of the technology functionalities we outlined in Section 2 actually become used regularly as part of Module 3 of the government’s eRulemaking Initiative software, we will have succeeded. There already have been signs of interest in our plans from the public watchdog organizations. If we are even partly successful, we hope that the primary benefit to society of this research will be better and more durable federal rules that citizens and groups will shape, understand, and accept as part of a transparent, efficient, and deliberative democratic process.

6. Prior NSF-Funded Research by PIs (Only Relevant to this Project)

A) Shulman, EIA-0089892 (\$33,428) 09.01.00-08.31.01, “Digital Government: SGER: Citizen Agenda-Setting in the Regulatory Process: Electronic Collection and Synthesis of Public Commentary”

Results: Representatives from seven agencies and 12 undergraduate students participated in a day-long eRulemaking Workshop at the Council for Excellence in Government (May 2001) and National Defense University (June 2002).

Publications: Stuart W. Shulman, “An Experiment in Digital Government at the United States National Organic Program,” *Agriculture and Human Values* 20, 3 (Fall 2003), 253-265; Stuart W. Shulman, David Schlosberg, Stephen Zavestoski, and David Courard-Hauri, “Electronic Rulemaking: New Frontiers in Public Participation,” *Social Science Computer Review* 21, 2 (Summer 2003); Stephen Zavestoski and Stuart W. Shulman, “The Internet and Environmental Decision-Making,” *Organization and Environment* 15, 3 (Fall 2002), 323-327.

B) Callan, Hovy, Shulman, & Zavestoski, EIA-0327979, 0328175, 0328914 & 0328618 (\$129,587) 05.01.03-04.30.04, “SGER Collaborative: A Testbed for eRulemaking Data”

Results: 14 agency representatives and 13 interest group representatives attend day-long workshops and focus group sessions at the NSF (September 2003); a testbed of eRulemaking data was compiled and made available to researchers; presentations were made to the EPA, DoT, USFS, and the federal eRulemaking Initiative Advisory Board.

Publications: Stuart W. Shulman, Lisa Thrane, and Mack C. Shelley, “eRulemaking,” forthcoming in the *Handbook of Public Information Systems* (2nd Edition); Stuart W. Shulman, “eRulemaking: Issues in Current Research and Practice,” forthcoming in the *International Journal of Public Administration*.

C) Callan, EIA-9983253. (\$481,437), “A Language Modeling Approach to Metadata for Cross-Database Linkage and Search.”

Results: New federated search techniques for acquiring compact descriptions of uncooperative text search engines, selecting the best search engines for a given query, and merging results returned from different search engines. A research prototype is being developed for the FedStats.gov web site.

Publications: See Callan’s homepage for a complete list. Two examples are listed here. L. Si and J. Callan. “Relevant document distribution estimation method for resource selection.” In *Proceedings of the Twenty Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. 2003; L. Si and J. Callan. “A semi-supervised learning method to merge search engine results.” *ACM Transactions on Information Systems*, 24(4) (pp. 457-491). ACM. 2003.