

Better Public Policy through Natural Language Information Access

Boris Katz Roger Hurwitz Jimmy J. Lin Ozlem Uzuner
MIT Artificial Intelligence Laboratory
Cambridge, MA 02139
{boris,rhhu,jimmylin,ozlem}@ai.mit.edu
www.ai.mit.edu/projects/infolab

Abstract

Federal agencies implement laws passed by the Congress by creating rules and regulations that can be applied in practice. During this process, staffs at the various agencies may review past and current regulations and receive comments from stakeholders and the public regarding the proposed regulations.

Putting rulemaking online can increase the public's awareness of the proposed rules and its participation in the process. It can also facilitate staff work. A key factor in realizing these benefits will be the availability of simple, intuitive, and timely access to the empowering legislation, the proposed rules and information regarding them. We propose to provide such access through an information architecture that allows members of the public as well as staff and stakeholders to obtain the texts and information they desire by using everyday language. Over the past decade, we have developed the START and Omnibase systems for natural language question answering and have applied them in a variety of domains. We plan to use these systems and our experience to support online rule making.

We note that besides providing information access, these systems can function more proactively, by soliciting feedback from targeted parties or by sending out notifications and information in response to standing queries submitted by the users.

1. Introduction

Federal agencies implement laws passed by the Congress by creating rules and regulations that can be applied in practice. During this rule-making process, staffs at the various agencies may review past and current regulations, and receive comments from stakeholders and the public regarding the proposed regulations. Efforts to put this process online have several goals: First, they extend the movement to make more government information available in electronic form. Second, this availability can facilitate staff reviews of current and proposed rules. Third, it has the potential to increase the public's awareness of new proposals and its participation in the rule-making process. When this last goal is realized, public acceptance of the rules and trust in the process will increase.

The availability online of the texts does not itself assure the realization of these goals. Ordinary people, stakeholders and even agency staff are interested in the rules for items that affect them, not as objects of study. In effect, they want to know how or why a (proposed) rule affects something specific they are doing or care about, and they need to know it in time to consent, complain or suggest a change. Consequently, all parties will need simple, intuitive, timely and fine-grained means of access to the information generated

in rulemaking. The recent creation at www.regulations.gov of a site that collects topically indexed links to electronic dockets is certainly a step toward simplifying access, especially when there are about two hundred agencies that can make rules. Nevertheless, someone who browses this site to find out about a specific issue will have to look through all the dockets returned for a keyword that subsumes his issue. Moreover, if the user also want to know whether Congress had anything to say about the issue when it created the empowering legislation or what other people have to say about the rule, he will have to search elsewhere and further.

Note, the ordinary member of the public remains at a disadvantage vis a vis the agencies and stakeholders, such as business and advocacy groups. The agencies and stakeholders have information specialists working for them, with knowledge and time to track the rules, understand their contexts and grasp their possible implications. Furthermore, the stakeholders are typically aware of a process well before an agency notifies the public of its intent to issue new rules and invites comments. Some groups may have lobbied to get the empowering legislation to meet their interests; while others may have influenced an agency's initial deliberations over whether and what new rules were needed to meet the legislative intent.

While we have no illusions that information technology can eliminate these gaps, we believe the appropriate information access architecture can reduce them and improve in several respects the efficacy and experience of all parties to rulemaking. This architecture enables people to access information by using ordinary language to ask the questions they want the information to answer. It leverages the natural language question answering technology that we have developed over the past decade, per our belief that natural language is the best information access mechanism for people, since it is intuitive, easy to use, requires no specialized training and is arbitrarily specific.

The particular payoffs for this architecture are the increased relevance for the user of the information returned, a potentially enhanced voice for ordinary members of the public and the possibility of a member being notified early about rulemaking that concerned him. First, in this architecture all the currently available information that addresses the user's question would be retrieved, that is, the relevant segments of rules, proposals, comments and other documents. Second, agency officials could use it to aggregate opinions concerning various parts of the proposed rules. Third, users could submit their questions as standing queries, against which new information, such as early discussions in an agency regarding the empowering legislation, would be measured. A user would be notified, when the information addressed his question.

2. Natural Language Annotations: Knowledge about Knowledge

Our approach to providing the user with "just the right information" is based on the idea of teaching the computer *where* and *how* to find the right pieces of knowledge, by giving it *knowledge about the knowledge*. This idea has been implemented in START (Katz, 1988; Katz, 1997), the first natural language question answering system available on the World Wide Web. Since it came on-line in December, 1993, START (<http://www.ai.mit.edu/projects/infolab>) has engaged in exchanges with hundreds of thousands of users all over the world, supplying them with useful knowledge.

To answer natural language questions, START uses natural language annotations (Katz, 1997), which are machine-parseable sentences and phrases that describe the content of various information segments and help identify the right piece of information. The annotations serve as metadata describing the types of questions that a particular piece of knowledge is capable of answering. For example, a web page about the Consumer Broadband and Digital Television Promotion Act—legislative background information on rulemaking by the Federal Communications Commission— might be annotated as follows:

CBDTPA anti-copy bill.

Sen. Fritz Hollings introduced CBDTPA on March 21, 2002.

CBDTPA was previously known as SSSCA.

CBDTPA prohibits the sale of technology that does not include copy-protection standards.

START parses these annotations and stores the parsed structures (embedded ternary expressions (Katz, 1988)) with pointers back to the original information segment. When a question is asked, the user query is compared against the annotations stored in the knowledge base. This match occurs at the level of syntactic structures and linguistically sophisticated machinery, such as synonymy/hyponymy, ontologies, and structural transformation rules operate in the matching process and allows the system to handle complex syntactic alternations involving verb arguments. When a match is found between ternary expressions derived from annotations and those derived from the query, the segment corresponding to the annotations is returned to the user as the answer. For example, given the annotations above START can answer the following questions:

When was CBDTPA introduced?

The CBDTPA was introduced by whom?

What was the CBDTPA previously called?

Because START can understand user questions, it can provide precise answers, even when there are syntactic or lexical mismatches between queries and their answers; for example, the annotation "... previously known as. . ." can match a question phrased with "... previously called. . .".

Another important feature of the annotation concept is that any information segment can be annotated, including images, videos and even procedures (see Figure 1). This feature is the basis of another aspect of the START system: providing natural language information access to structured and semi-structured data.

3. Annotating Structured and Semi-structured Knowledge

A useful extension of our annotation technology is the ability to easily access the enormous quantities of structured and semi-structured data available online. Structured information refers to collections of records, often stored in a database, that have fixed fields. A registry of motor vehicles is a good example: every car has a license plate, a registered owner, and additional properties. Semi-structured information is a set of records that is often stored as loosely connected web pages. Some records may have missing or extra fields, or the properties may be inconsistently formatted across the records. To provide uniform access to variously structured and semi-structured online sources, we developed Omnibase, a "virtual" database that serves as an "abstraction layer" over these diverse resources.



Figure 1: Sample responses from START.

Omnibase acts like you probably would, if someone asked you "When was Rutherford Hayes president of the United States?" You would find a resource with the answer, e.g., a biographical dictionary or a web site about presidents, find the entry for Rutherford B. Hayes, and see the date of his inauguration. Millions of questions can be answered by following this same recipe: extract an *object* (Rutherford Hayes) and a *property* (presidential term) from the question, find a data source (e.g., the Internet Public Library web site) for that type of object, look up the object's web page, and extract the *value* for the answer. START and Omnibase cooperate to answer natural language questions in much the same way. From its knowledge base, START figures out that the above question can be answered by a page from the Internet Public Library. Accordingly, it constructs an object–property–value (OPV) query for Omnibase:

```
(get "ipl" "Rutherford Hayes" "presidential-term")
```

Omnibase looks up the data source and property to find an associated wrapper script and applies the script to the object in order to retrieve the property value for the object:

```
(get "ipl" "Rutherford Hayes" "presidential-term") =>
("March 4, 1877 -- March 3, 1881")
```

START then assembles the answer and presents it to the user either as a fragment of HTML or couched in natural language (see Figure 2).

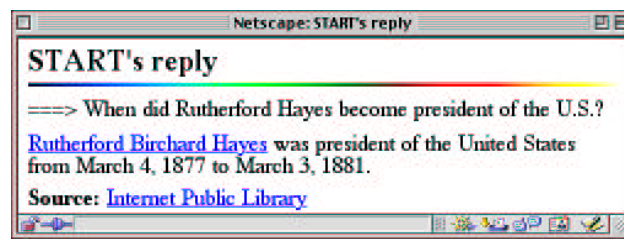


Figure 2: START answering a question with data from Omnibase, presented within a generated sentence.

Using our OPV data model, START can currently answer millions of natural language questions about places (e.g., cities, countries, lakes, coordinates, weather, maps, actors, directors), people (e.g., birth dates, biographies), dictionary definitions, and much, much more.

We believe that our technology for organizing semi-structured data will prove to be very valuable in the domain of public policy. Many government agencies and offices already maintain semi-structured knowledge sources that could be seamlessly integrated into Omnibase to provide the public with natural language information access to a wealth of information. For example, if a database of bills were structured using our object–property–value data model and integrated into Omnibase, users could then ask a variety of interesting questions, including questions that require some computation:

- What bills has Fritz Hollings introduced?
- By what margin did the Digital Millennium Copyright Act pass in the House?
- Were more bills introduced in 1995 than in 1994?
- Which senators have supported more than one anti-tobacco bills?
- What bill passed by the largest margin in 2000?
- What's the longest bill ever considered by Congress?

In many cases, however, appropriate databases in policy-related domains do not exist. We believe that our experiences will help set up the infrastructure necessary for constructing such databases. Specifically,

our OPV data model serves as a guide for organizing data; because our model is grounded in real user queries, we can better ensure that newly constructed databases will be easy and intuitive to understand.

4. Leveraging Information Access Technologies

As for rule-making, our discussion of Omnibase suggests that public comments would be more useful and accessible to regulators if they were collected as semi-structured information. This can be accomplished with comment interfaces that capture the section(s) of a proposed rule returned in response to a query and commented on. The results would give officials and the public alike a first-cut basis for assessing the distribution of opinion regarding the various sections and for identifying the controversial parts.

We recognize that short, parsable response might not adequately articulate the opinions regarding the various parts of a proposal or the proposal as a whole. To solicit more extensive public commentary and participation in the rulemaking process, we plan to provide a distributed framework where ordinary users can express their comments and provide natural language annotations that epitomize these comments. As in the Open Meeting System (Hurwitz and Mallery, 1996), users could be asked to indicate explicitly the thrust of their comment, e.g., an argument for or against the proposal, an alternative suggestion. These annotations could then be used to answer questions like "Give me the arguments against CBDTPA." Annotations could also include interpretations and explanations of the information in question and even links to other material.

Annotations collected from a variety of different people would be very useful in broadening the coverage of an information access system. Finally, comments can be automatically collected and summarized to give an overall picture of public approval or disapproval regarding a particular rule or policy. Such information could then be automatically circulated to all relevant agencies and identified stakeholders.

5. Ongoing Work

Because START performs sophisticated syntactic and semantic processing of questions to pinpoint the exact information need of a user, questions can be answered with remarkable precision. In the period from January, 2002 to December, 2002, START and Omnibase replied to over 342 thousand queries from users all over the world. Of those, 69% were answered successfully by our system (54% of the questions answered were handled by Omnibase).

Although START and Omnibase provide access to a wealth of information, we are aware that there exist on the Web and elsewhere vast amounts of unstructured documents that cannot be handled by our semi-structured database technology, and are too numerous to be annotated manually. Some of the governmental agency documents are in this category. To address this issue, we are developing two separate technologies: large-scale syntactic indexing and use of the Semantic Web.

Although full syntactic and semantic analysis of open-domain natural language text is beyond current technology, we believe we can augment START's manual-annotation-based approach with automatically built annotations, by extracting a limited subset of relations from unstructured text. This would automatically provide us with high precision question answering to vast numbers of documents. By employing robust natural language processing technology, we can extract syntactic relations from free text and automatically construct large-scale knowledge bases. These knowledge bases capture important relations in unstructured text, and provide information access capabilities far beyond standard "bag-of-words" information retrieval systems. We have implemented this technology in a system called Sapere (Katz and Lin, 2003), and demonstrated its effectiveness in providing high precision knowledge access to vast amounts of unstructured sources by making it possible to distinguish between two questions such as "What regulations does the Federal Railroad Administration control?" and "What regulations control the Federal Railroad Administration?" Since both these questions contain the same keywords, traditional keyword search engines

would not be able to answer them correctly.

The vision of the Semantic Web (Berners-Lee, et al., 2001) is to convert existing Web information into a more machine-readable form, to make the Web more effective for users. Because future policy-related documents might be encoded as part of the Semantic Web, we believe that Semantic Web technologies can be leveraged to provide intuitive information access. Our current work centers around integrating natural language annotations technology with the Resource Description Framework (RDF) (Lassila and Swick, 1999; Brickley and Guha, 2002), the foundation of the Semantic Web. By integrating formal ontologies with natural language, we can achieve both human accessibility and computer readability to more webpages. We have concretely described three separate mechanisms for marrying natural language annotations and RDF, and have built initial prototypes to verify our ideas (Katz, et al., 2002; Karger, et al., 2003).

6. Conclusion

We have described a variety of information access strategies suitable for different types of information. Our technologies can handle data ranging in complexity from fully structured databases to free text. Our technologies can also engage in various modes of operation ranging from manual-crafting of high-quality annotations to automatic creation of knowledge structures that describe content. The domain of public policy offers a rich and fertile ground for the application of these technologies. By providing the public intuitive and timely access to relevant rules and policies, we can take a step towards creating informed citizens and encouraging participation in the rulemaking process.

Acknowledgements

This research is funded by DARPA under contract number F30602-00-1-0545 and administered by the Air Force Research Laboratory. Additional funding is provided by the Oxygen Project.

References

- Berners-Lee, T., Hendler, J. and Lassila, O. 2001. The Semantic Web. *Scientific American*, 284(5):34-43.
- Brickley, D. and Guha, R.V. 2002. RDF vocabulary description language 1.0: RDF Schema. W3C Working Draft, World Wide Web Consortium.
- Hurwitz, R. and Mallery, J. 1996. The Open Meeting: A Web-based system for conferencing and collaboration. *World Wide Web Journal: The Fourth International WWW Conference Proceedings*, 1(1):19-46.
- Karger, D., Katz, B., Lin, J. and Quan, D. 2003. Sticky notes for the Semantic Web. In *Proceedings of the 2003 International Conference on Intelligent User Interfaces (IUI 2003)*.
- Katz, B.. 1988. Using English for indexing and retrieving. In *Proceedings of the 1st RIAO Conference on User-Oriented Content-Based Text and Image Handling (RIAO '88)*.
- Katz, B.. 1997. Annotating the World Wide Web using natural language. In *Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO '97)*.
- Katz, B., Lin, J. and Quan, D. 2002. Natural language annotations for the Semantic Web. In *Proc. of the International Conference on Ontologies, Databases, and Application of Semantics (ODBASE 2002)*.
- Katz, B. and Lin, J. 2003. Selectively using relations to improve precision in question answering. In *Proceedings of the EACL-2003 Workshop on Natural Language Processing for Question Answering*.
- Lassila, O. and Swick, R. R. 1999. Resource Description Framework (RDF) model and syntax specification. W3C Recommendation, World Wide Web Consortium.