Text Annotation for Political Science Research

Digitization is dramatically altering research demands and opportunities in political science, and in the social sciences more generally. To cite just a few examples, the advent of e-government has challenged governments to keep pace with rapidly expanding opportunities for public commenting via e-mail or Web portals during the development of government policy (Balla & Daniels, 2007); the creation of online media has dramatically increased the amount of accessible digital political content and altered the pace and dynamics of political campaigns (Hopkins & King, 2007); governments around the world now release huge volumes of digitized data on a daily basis (e.g., the U.S. Federal Register—http://www.gpoaccess.gov/fr/Index.html), while national projects are digitally scanning vast numbers of historical documents. For example, British parliamentary debates from the 17th century to the present are now accessible online, and ongoing research will extend their availability back to 1066.[1]

These developments in data accessibility are creating unprecedented opportunities both to reinvestigate longstanding questions in political science and to embark on the study of new questions. However, a central challenge of working with data of any sort is that it must be organized and classified so that the researcher can use it for the task at hand. In this volume, the data of interest is text. A government agency that receives tens of thousands of comments on a proposed regulation, for example, needs to be able to cull, categorize, and summarize the substantive information contained in those comments in a useful way. A scholar studying campaign coverage on the Internet needs to analyze and organize that coverage to test specific questions about its character.

Manual approaches to extracting information from textual data can be challenging for large tasks where resources are limited (as they usually are). Computer-assisted approaches seem to be an attractive alternative: They can enable researchers to complete certain tasks with much greater speed. Nonetheless, it is also important to recognize that faster methods are not necessarily better methods. A computer program might be able to sort public comments by zip code more quickly than, and as accurately as, humans; but humans might be substantially better, albeit slower, at classifying public comments by topic. Ultimately, each manual, automated, or semi-automated method for analyzing textual data has its own set of benefits and costs that vary depending on the task at hand.

This special volume of *JITP* includes eight articles investigating a diverse set of political science tasks, from e-government to political speeches to campaign coverage. The articles nicely illustrate a range of methodological challenges where extracting information from text is concerned. In the process, they also demonstrate the strengths and limitations of alternative text analysis methodologies.

Text has always been an important source of data in political science. Text annotation

The editors would like to sincerely thank the authors, the many anonymous reviewers, *JITP*'s Emily Huisman, and especially *JITP* executive editor Stuart Shulman, without whom this special issue would not have been possible.

Journal of Information Technology & Politics, Vol. 5(1) 2008
Available online at http://jitp.haworthpress.com
© 2008 by The Haworth Press. All rights reserved.
doi:10.1080/19331680802149590

methods have also been used within political science for many years (see Hillard, Purpura, & Wilkerson, 2008). What has changed, in our view, is (a) the increased availability and accessibility of text in a digital world, and (b) the subsequent increased interest in, and need for, new text annotation methods. We proposed this volume because we thought that many social scientists would like to learn more about the rapidly expanding options for automated and partially automated annotation of textual data. We also hoped that computer and information scientists would like to learn more about the research questions that interest political scientists and the range of technical challenges involved in extracting information from political text. Finally, we hoped that the special issue would make clear the many possibilities for interdisciplinary collaboration between political and computer scientists.

A RANGE OF TEXT ANNOTATION TECHNIQUES

In general, text annotation methods can be thought of as varying according to the extent to which humans are involved in the annotation process. At one extreme are *manual methods*, where humans do all of the annotation. At the other extreme are *unsupervised* (computer-based) *learning algorithms* that search for patterns in text and require no external input. In between are *supervised learning algorithms* that are trained via a text corpus that has been manually annotated to replicate the human's annotation decisions. Like the supervised methods, *weakly (or semi-) supervised learning algorithms* are trained to replicate human annotation decisions, but require far fewer manual annotations to reach the same levels of performance.[2]

Within each of these general approaches there exists a range of techniques to handle different types of annotation tasks, each with its own set of advantages and disadvantages over the available alternatives. Unsupervised learning algorithms (e.g., agglomerative clustering, factor analysis) are relatively easy to implement and allow the researcher to explore the data in fairly flexibly ways. But the methods can sometimes be quite slow, and they are generally not appropriate for tasks where the categories of interest are predefined, or where stable results across related data sets are essential. Alternatively, annotation instructions to guide the manual categorization of, or information extraction from, text can be designed. These manual annotation methods are slow to apply, but for some tasks, they might be the best option for ensuring validity and reliability (depending on the size and complexity of the task).

Supervised learning systems (see Mitchell, 1997 and Russell & Norvig, 2002 for general introductions to supervised learning) seek to balance the speed benefits of automated annotation approaches with the validity and reliability benefits of human-centered annotation. They are, in general, more difficult to build than many unsupervised algorithms, but off-the-shelf implementations of the most used supervised learning algorithms are available in the public domain for a wide variety of text analysis tasks, including: text categorization (e.g., support vector machines (SVMs), naïve Bayes), classification (e.g., decision trees, SVMs, rule learners), regression (e.g., neural networks), sequence tagging (e.g., conditional random fields (CRFs); Lafferty, McCallum, & Pereira, 2001; McCallum, 2002), text-based pattern learning (e.g., CRFs, Autoslog; Riloff, 1996), topic segmentation (e.g., Choi, 2000), semantic labeling (e.g., Punyakanok, Roth, & Yih, 2005), syntactic parsing (e.g., Klein & Manning, 2004), and pronoun and general coreference resolution (e.g., Ng & Cardie, 2002). The holy grail of text annotation is an automated system that accurately and reliably annotates very large numbers of cases using relatively small amounts of manually annotated training data. This is the goal of semisupervised learning, a relatively new area of research in machine learning and natural language processing (NLP).

For those new to the area of NLP, two very good, general introductions to the field are Jurafsky and Martin (2008) and Manning and Schütze (1999). The Web pages associated with each textbook also provide pointers to computer programs for many text-analysis tasks.

EVALUATING PERFORMANCE

A central challenge of text annotation is that there is usually no objective standard for assessing the success of the process of converting data to information. The *gold standard* of text annotation research is usually work performed by human coders (although different standards are sometimes used). In other words, the assessment is not whether the system accurately classifies events, but the extent to which the system agrees with humans where those classifications are concerned.

But what level of agreement is acceptable? If humans are bound to make mistakes, then a perfect classifier should produce less than 100% agreement. At a minimum, any system should perform better than one would expect by chance. If there are just two categories (e.g., positive or negative tone of a newspaper article), then 50% agreement is unimpressive. But with 200 categories (e.g., policy topics that are each used an equal number of times), 50% agreement may be acceptable or even impressive. Thus, statistics commonly used to assess inter-annotator agreement (between humans or between a human and a computer), such as Cohen's Kappa, make adjustments for what is expected by chance.

When a manually annotated gold standard is available for a task, researchers commonly report *precision* and *recall* when evaluating the performance of their automated annotation systems. Precision asks, "What percentage of the annotations proposed by the system is correct (when compared to the gold standard)?" Recall asks, "What percentage of the annotations in the gold standard is correctly identified by the system?" For many tasks, researchers are interested in obtaining high levels of precision in conjunction with high, or at least reasonable, levels of recall. In these cases, the *F-score* provides useful information: The F-score is the harmonic mean of recall and precision, that is, it is an average that rewards precision and recall values that are close together. (Thus, a system that has a precision of .60 and recall of .60 will have a higher F-score than a system with a precision of .80 and recall of .40, even though the average of precision and recall for both systems is .60.) Precision and recall with respect to particular categories of annotation are also used diagnostically to identify and address areas of weakness in system performance.

However, performance scores alone should not be used as the sole basis for evaluating whether a system performs well. Some datasets—such as one with a large number of duplicate records—may be easier to annotate than others. As a result, performance should always be compared to one or more baseline systems (e.g., for a categorization task, this might be a system that always guesses the most frequent category, that is the *majority class*) and to previously reported state-of-the-art approaches to the task, if they exist. Finally, differences in performance should be tested for statistical significance.

It is our opinion that much research remains to be done in the area of system evaluation and validation of automated text annotation techniques in the context of political science questions. As you read the papers in this issue, therefore, keep in mind the issues of evaluation described above. What aspects of the system evaluation are strong? Where might the researchers have done better? In addition, keep in mind a different, but equally important question with respect to evaluation: Were the methods adequately evaluated for their appropriateness to the political science question being investigated? After all, the common goal of the techniques presented here is to advance research in political science.

TEXT ANNOTATION IN THE CURRENT ISSUE

Table 1 provides an overview of the papers in this special issue. Along with the name, authors, and political science question addressed in each paper, the final column lists the primary text annotation technique(s) employed as well as the types of linguistic knowledge that play an important role in the annotation process.

The first two articles of this special issue (by Dyson and by Guerini, Strapparava, and Stock) do not rely primarily on learning-based text analysis methods, in contrast to the remaining articles. They tackle very different types of political science questions by statistically analyzing

TABLE 1. Political Science Tasks and Text Analysis Techniques for Articles in the Special Issue on Text Annotation in Political Science

Article	Political Science Task	Technique(s) and Linguistic Knowledge Employed
Text Annotation and the Cognitive Architecture of Political Leaders: British Prime Ministers from 1945–2007 (Dyson)	Characterization and comparison of the information-processing styles of political leaders	Ratios of word counts Dictionary of high- vs. low-complexity words
CORPS: A Corpus of Tagged Political Speeches for Persuasive Communication Processing (Guerini, Strapparava, & Stock)	Creation of a corpus of political speeches annotated with audience responses; automatic identification of persuasive expressions in political speeches	Word-based statistical analysis Lemmatization; part-of-speech tagging; audience reaction information; named entity identification; dictionary of sentimentbearing words
Classifying Party Affiliation from Political Speech (Yu, Kaufmann, & Diermeier)	Classification of U.S. Congressional speeches according to the party affiliation of the speaker	Supervised machine learning; text categorization. Bag-of-words text representation
Recognizing Citations in Public Comments (Arguello, Callan, & Shulman)	Classification of sentences from public comments with respect to whether or not they contain references to external sources of information	Supervised machine learning; sentence-level text categorization; classifier ensembles Named entity detection; bag-of-words; dependency tree parsing; frame semantics
Good News or Bad News? Conducting Sentiment Analysis on Dutch Text to Distinguish Between Positive and Negative Relations (van Atteveldt, Kleinnijenhuis, Ruigrok, & Schlobach)	Classification of the polarity of relations between actors and issues in political newspaper articles	Supervised machine learning; sentence-level sentiment analysis; word co-occurrence-based clustering; clustering by distributional similarity of syntactic relations Lemmatization; part-of-speech tagging; dependency parsing; word bigrams; syntactic dependency bigrams; word co-occurrence statistics; distributional similarity statistics; conjunction patterns
Automatic Annotation of Semantic Fields for Political Science Research (Beigman Klebanov, Diermeier, & Beigman)	Semantic annotation of text for political science research	Clustering; semantic dictionary-based word counting; supervised machine learning Bag-of-words; word frequencies; distance between words; morphology; semantic distance; co-occurrence statistics
Workbench Notes		
Natural Language Processing for Comparing the Editorial Slant of Online Media: A Case Study on the U.S. Presidential Elections (Scharl & Weichselbraun)	Ingestion and annotation with respect to editorial slant of online coverage of U.S. presidential candidates	Keyword extraction; pattern-matching; semantic orientation computation Dictionary of sentiment-bearing words; bag-of-words
Media Monitoring by Means of Speech and Language Indexing for Political Analysis (Demiros, Papageorgiou, Antonopoulos, Pipis, & Skoulariki)	Environment to support (a) the manual annotation of speech and text and (b) the automatic retrieval of speech, text, and images	Vector-space text retrieval; weighted Boolean retrieval based on metadata fields; manual annotation Speech recognition; bag-of-words

just the *lexical* (i.e., word-level) information in the documents under study. The Guerini et al. article introduces a number of concepts and techniques from computational linguistics that will reappear in subsequent articles—in particular, concepts from the area of lexical syntax (e.g., part-of-speech tagging) and phrase-level syntax and semantics (e.g., identification and categorization of *named entities* such as person names, locations, countries, dates, book titles). Guerini et al. also describe the creation of a corpus that is annotated automatically with both

lexical information as well as with limited *pragmatic* knowledge (i.e., knowledge about the use of language) in the form of audience reactions to political speeches.

The next two articles rely primarily on text categorization techniques. Yu, Kaufmann, and Diermeier make binary document-level decisions regarding the political party of the speaker (i.e., Democrat or Republican), while Arguello, Callan, and Shulman address a problem from electronic rulemaking that requires sentence-level categorization. Both employ machine learning algorithms for their text categorization components.

Readers who are unfamiliar with the standard process for training and testing text categorization systems (or supervised learning algorithms, in general) are referred to the Yu et al. article, which provides a nice introduction. The Arguello et al. article, on the other hand, shows how to go beyond the standard *bag-of-words*[3] representation for textual data: They incorporate named entities, syntactic parse information, and knowledge from one type of *semantic lexicon* (i.e., dictionary).

The van Atteveldt, Kleinnijenhuis, Ruigrok, and Schlobach article also employs supervised machine-learning methods for classifying sentences, but, in addition, introduces two unsupervised techniques for clustering words. It is a good article to look at for additional background on the general framework for supervised inductive learning and especially for examples of the kinds of linguistic features that can be employed by the learning algorithms. Very generally, a *feature* is information in the training examples that is relevant to what is being studied. For a text categorization task, the presence or absence of particular words (or phrases) are the features that enable the algorithm to make distinctions among categories.

The Beigman Klebanov, Diermeier, and Beigman article is unique in the special issue in that it compares and contrasts three very different text annotation techniques—an unsupervised approach (clustering), a word-counting approach (that makes use of a dictionary), and a supervised machine learning technique (classification).

We end the special issue with two Workbench Notes that describe end-to-end systems for the analysis of political text. The system for analyzing media coverage of U.S. presidential candidates (by Scharl and Weichselbraun) employs keyword extraction and pattern-matching techniques to compute the semantic orientation of an article. In contrast, Demiros, Papageorgiou, Antonopoulos, Pipis, and Skoulariki present a system that aids the manual annotation of articles as well as the automatic retrieval of articles using information retrieval methods and metadata-based Boolean search.

Together the articles show how a variety of methods from natural language processing can be applied to text annotation tasks in the service of political science research. They represent, however, only a small part of the spectrum of what is possible given the current state-of-the-art in NLP and political science research. We hope that this special issue encourages further communication between the political science and NLP communities in the area of text annotation.

Claire Cardie
Professor, Faculty of Computing and Information Science, Cornell University
John Wilkerson
Associate Professor, Department of Political Science
University of Washington

NOTES

1. http://www.parliament.uk/parliamentary_publications_and_archives/parliamentary_archives/archives_electronic.cfm#public

2. Alas, this special issue contains on articles on semi-supervised learning; for examples of its use on political texts, see Hillard, Purpura, and Wilkerson (2008) and Purpura, Cardie, and Simons (in press).

3. A bag-of-words representation encodes a document as an *unordered* set of (the counts of) the words that comprise the document.

REFERENCES

Balla, S., & Daniels, B. (2007). Information technology and public commenting on agency regulations. *Regulation & Governance, 1*, 46–67.

Choi, F. (2000). Advances in domain independent linear text segmentation. *Proceedings of the Conference for the North American Chapter of the ACL (NAACL-2000), 4*, 26–33. New York: Association for Computing Machinery.

Hillard, D., Purpura, S., & Wilkerson, J. (2008). Computer assisted topic classification for mixed methods social science research. *Journal of Information Technology and Politics, 4*(4).

Hopkins, D., & King, G. (2007). Extracting systematic social science meaning from text. Unpublished manuscript, Harvard University. Available online at http://gking.harvard.edu/files/abs/words-abs.shtml.

Jurafsky, D., & Martin, J. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.

Klein, D., & Manning, C. (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL), no. 478*. New York: Association for Computing Machinery.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning*. Available online at http://www.aladdin.cs.cmu.edu/papers/pdfs/y2001/crf.pdf.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. Available online at http://mallet.cs.umass.edu.

Mitchell, T. (1997). *Machine learning*. New York: McGraw-Hill.

Ng, V. & Cardie, C. (2002). Improving machine learning approaches to coreference resolution. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 104–111. New York: Association for Computing Machinery.

Punyakanok, V., Roth, D., & Yih, W. (2005). The necessity of syntactic parsing for semantic role labeling. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. Available online at http://cllt.osu.edu/mclc/paper/necessity_punyakanok.pdf.

Purpura, S., Cardie, C., & Simons, J. (in press). Active learning for e-rulemaking: Public comment categorization. *Proceedings of the Ninth International Conference on Digital Government Research*, 234–243. Montreal, Canada.

Riloff, E. (1996). Automatically generating extraction patterns from untagged text. *Proceedings of the Thirteenth National Conference on Artifical Intelligence*, 1044–1049. New York: Association for Computing Management.

Russell, S., & Norvig, P. (2002). *Artificial intelligence: A modern approach* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.