Advancing Social Science Research by Applying Computational Linguistics

An-Shou Cheng
College of Information Studies, University of Maryland, College Park, MD 20742, ascheng@umd.edu

Kenneth R. Fleischmann
College of Information Studies, University of Maryland, College Park, MD 20742, kfleisch@umd.edu

Ping Wang
College of Information Studies, University of Maryland, College Park, MD 20742, pwang@umd.edu

Douglas W. Oard
College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, oard@umd.edu

Abstract
This paper discusses the growing trend of applying computational thinking and linguistic approaches to social science research, arguing that computational linguistics is a useful but underutilized approach that may potentially be able to make significantly contributions to research in a wide range of social science domains. The paper begins with an introduction to computational thinking and argues that this approach can be applied not only in the sciences but also in the social sciences. Next, the paper discusses the linguistic turn in the social sciences and provides an overview of research on manual content analysis. The following sections describe how automatic content analysis evolved from manual content analysis, and describe how a computational approach can support content analysis. The paper then describes principles and techniques for applying automatic content analysis. Next, the paper gives examples of domains where automatic content analysis is already being applied and domains where it could be applied in the future. Finally, the paper calls attention to the need for additional work in this area.

Introduction

Recently, there has been a growing trend to apply computational thinking to a wide variety of problems and domains. Simultaneously, the social sciences have increasingly turned to language as a means for understanding social dynamics. At the intersection of these two broader trends, computational linguistics is increasingly being applied to domains such as intelligence gathering, machine translation, automatic content analysis, and indexing and retrieval in full-text databases. However, the application of computational linguistics to the social sciences remains an important yet underutilized approach. This paper reviews an exemplar application of computational linguistics – automatic content analysis – to the social sciences, lays out useful principles and approaches for such applications, and calls attention to the need for additional work in this area.

Computational thinking and how it can be applied to social science

As coined by Wing (2006), "computational thinking" is an approach to "solving problems, designing systems, and understanding human behavior that draws on concepts fundamental to computer science" (Wing, 2006: 33). Computational thinking is an understanding of how computer scientists think and solve problems; it is not trying to get humans to think like computers. The implication of computational thinking is that thinking like computer scientists can be beneficial for scholars in many fields, not just computer science. Wing argues that in the twenty-first century, computational thinking should be a fundamental skill, like reading, writing, and arithmetic. Computational thinking involves thinking recursively, abstractly, procedurally, logically, and concurrently. Computational thinking has primarily been applied in computer science and is increasing being applied in science and engineering including chemistry and physics (Wing, 2006), but this approach has relevance not only for the physical sciences, but also for the social sciences. In the realm of the social sciences, computational thinking may provide insights for how to deal with large corpora of data by automating text analysis in order to produce a quantifiable result. Statistical machine learning, which can be used to refine algorithms over time, and reputation services, which can be used to provide recommendations, can also be used to solve research problems in the social sciences.

Content analysis as a linguistic approach in the social sciences

At the same time, social scientists have been paying increasing attention to the importance of language for studying human and social dynamics. Language serves as the primary vehicle by which people communicate and record information. It has the potential for expressing an enormous range of ideas, and for conveying complex thoughts succinctly. It is used to convey knowledge and to understand the knowledge conveyed by others. Researchers are starting to rethink and reconsider some aspects of social science research from a linguistic point of view (Alvesson & Karreman, 2000). The use of content analysis as a research methodology is an

example of the use of language to study human cognition and communication. The methodology is based on the assumption that the analysis of text is a way for researchers to understand how people make sense of the world around them (McKee, 2003). Scholarly treatises, corporate reports, and political documents all use language to represent a part of reality.

Dille and Young (2000) code the utterances of political leaders and other individuals based on the understanding that the more frequently a specified theme or variable appears, the more important it is to the speaker, which might also have implications for policy. Zhou and Moy (2007) employ a content analysis of 206 online posts and 114 news reports regarding a sociopolitical incident in China to test the associations and causal relationships between the salience of opinion and media. Meijer and Kleinnijenhuis (2006) analyze corporate reputation by measuring the amount of news about specific issues. Wang and Ramiller (2004) analyze trade press articles on enterprise systems software published over the course of a decade and find that different types of actors contribute different types of knowledge about the software and that the dominant knowledge types and actors change over time. These examples illustrate the trend of the ‘linguistic turn’ in the social sciences. However, this research method, which involves drawing representative samples of content and training human coders raises the problematic trade-off between breadth and depth in research design (Riffe, Lacy, & Fico, 1998).

Content analysis

Content analysis is an established research methodology for systematic examination of textual materials that has been adopted by a wide range of academic disciplines, including communications, psychology, sociology, political science, and organizational research, and which incorporates a wide range of theoretical frameworks, methods, and analytical techniques (Denzin & Lincoln, 2000). Berelson (1952) defines content analysis as “a research technique for the objective, systematic and quantitative description of the manifest content of communication” (p.18). It limited the scope of content analysis to quantitative studies of the manifest characteristics of messages. Holsti (1969), however, provides no restriction on the quantitative description of manifest content. He defines content analysis as “any technique for making inferences by objectively and systematically identifying specified characteristics of messages” (p.2). Woodrum (1984) also contended that content analysis provides methods for measuring the characteristics of both manifest and latent communications. Shapiro and Markoff (1997) review six major definitions from various sources in the social sciences. They propose a minimal definition of content analysis as “any methodical measurement applied to text for social sciences purposes” (p.14). From these points of view, “content analysis is a research technique for making replicable and valid inferences from data to their context” (Krippendorff, 1980: 21). It is a research

technique, which involves specialized procedures for processing scientific data. It is an instrument that is reliable and replicable.

Krippendorff (1980) proposes a framework for content analysis that includes principles that should be followed by researchers when conducting a content analysis. First, researchers should be clear about which data will be analyzed, how they are defined, and from which population they are drawn. Second, the context relative to which data are analyzed should be made explicit. Any research must define the boundaries beyond which its analysis does not extend. Third, the aim or target of the inferences should be clearly stated. In content analysis, the task is to make inferences from data and to justify these inferences. It is by this process that data become recognized as symbolic or are rendered informative about something of interest to the analyst. Fourth, content analysis should be validatable in principle. This is to prevent analysts from pursuing research questions that do not allow empirical validation.

Content analyses are most successful when they focus on facts constituted in language (Krippendorff, 2004). Social scientists, especially political scientists and communication researchers, have a tradition of analyzing text in various media to understand the different purpose, focus, and techniques employed. Several advantages of content analysis are addressed by researchers. One of which is that content analysis provides a replicable methodology to access deep individual or collective structure (Carley, 1997; Kabanoff, 1997). In this sense, people's values, intentions, attitudes, and cognitions can be access and analyzed by using content analysis method (Duriau, Reger, & Pfarrer, 2007). Another advantage is the analytical flexibility allowed. For instance, researchers may focus on the manifest content of the text, which can be captured and revealed in a number of text statistics; they may also be interested in the latent content and deeper meaning embodied in the text, which may require more interpretation (Woodrum, 1984). Third, content analysis is an unobtrusive technique. It can facilitate empirical study without disrupting the research subject. Since the data is pre-existing, it is less likely to be biased, which would jeoparize the validity of the research (Krippendorff, 1980). Finally, content analysis can handle unstructured data, making it useful for dealing with texts in diverse formats associated with different purposes. Multiple sources of data, therefore, can serve as inputs to content analysis. The advantage of the unstructured data is that "it preserves the conceptions of the data's sources, which structured methods largely ignore" (Krippendorff,2004: 41).

According to Duriau et al. (2007), several additional methodological and practical benefits have been noted in implementing content analysis. First, content analysis is a robust methodology because the coding scheme can be corrected if flaws are detected as the study proceeds. If the

researcher determines that a portion of necessary information was missed or incorrectly coded, it is feasible to return to the texts and supplement the original data collection (Woodrum, 1984). Second, when done correctly, content analysis entails the specification of category criteria for reliability and validity checks that foster the creation of a replicable database. Category criteria reduce the concept ambiguity and help analysts to generate replicable research suitable for hypothesis testing (Woodrum, 1984). Third, content analysis can be used in conjunction with other methods. Historical, longitudinal, time series, and comparative research often can be accomplished economically with content analysis (Woodrum, 1984).

How automatic content analysis evolved from content analysis

Content analysis is particularly well suited to answering the classic communication research questions: “Who says what, to whom, why, how, and with what effect” (Babbie, 2004). The Internet opens up new areas of research by making new sources of data available and reducing the costs of data collection and analysis. As such, Internet-based research can be useful for addressing several limitations of traditional content analysis.

First, content analysts now have access to a much larger volume of data than ever before. The Web may have become one of the largest repositories of information ever known (Weare & Lin, 2000). Researchers have shifted their interest from small collections of printed messages to electronic full-text databases such as newspapers, journals, court decisions, and Web materials. Second, the large volumes of electronically available data call for different research techniques. A vast amount of text is being generated daily in digital format, representing almost every topic of interest to social scientists. The coding of large volumes of text is, however, a very labor intensive, time consuming, tiresome, and costly task. This limits the volume of text that researchers can examine. If computational methods can be applied to perform sophisticated coding with a degree of reliability similar to that expected of human coders, analyses of large amount of text could be done in a fraction of the time and with significantly lower costs than traditional human coding (Nacos, Shapiro, Young, Fan, Kjellstrand, & McCaa, 1991). Computational approaches are thus moving content analysis toward a promising future.

Computational technologies offer significant benefits for undertaking language-based social science research. They can facilitate individual research, allowing social scientists to gather information more quickly and analyze large bodies of data more efficiently. Social scientists have recently begun to apply automatic content analysis to enable reliable and detailed content analysis of various documents, including newspapers, magazines, journals, and legal/political documents (Gerner, Schrodt, Francisco, & Weddle, 1994; Hall & Wright, 2006; Evans et al., 2007).

How a computational approach can support content analysis

There are several ways that a computational approach can benefit content analysis. The most important reason for using a computational approach to content analysis is that computers can process large volumes of text in a relatively short period of time (Krippendorff, 2004; Duriau et al., 2007). When conducting research in electronic full-text and online databases such as newspapers, magazines, journals, newsgroups, online discussion forums, and Web materials, researchers who seek to balance breadth and depth may benefit from computational methods by analyzing large corpora of text automatically. Computer-aided text analysis (CATA) offers features for organizing, searching, retrieving, and linking text that renders the process of handling a large project much more manageable and productive (Kabanoff, 1997).

Second, a computational approach can potentially provide more uniformity. Ambiguities and uncertainties can be reduced through a standardized and automated approach to content analysis (Krippendorff, 2004). CATA can help researchers to overcome difficulties related to coding reliability in content analysis (Evans et al., 2007). It is important to make coding rules explicit in order to ensure the reliability and comparability of results across texts (Duriau et al., 2007). Morris (1994) tested the validity and reliability of manual and computerized approach. She found that the results from computerized coding and human coders agreed at an acceptable level and that computerized coding yielded an acceptable level of semantic validity.

Third, computers can reduce the time and cost of undertaking content-analysis projects (Nacos et al., 1991). They are most appropriate for recurrent and repetitive tasks. The savings of time and costs also stem from reductions of coding tasks, coder training, and inter-coder checks, all of which can greatly facilitate the ease of running multiple analyses (Carley, 1997).

Principles and techniques for applying automatic content analysis

Content analysis relies on several procedures for handling texts. According to Neuendorf (2002), the procedures include problem identification, conceptualization decisions, operationalization, development of coding schemes, sampling, coding (applying statistical procedures), and interpret and report results. Problem identification points out what content will be examined and why. It is the statement of research objectives. Researchable problems may come from direct observation or may be suggested by previous studies or theory. Conceptualization considers what variables will be used in the study, and how do researchers define them conceptually? Operationalization is the process by which researchers move from the conceptual level to the operational level, describing abstract or theoretical variables by using actual measurement procedures. A coding

scheme is a system trying to classify observable features in texts into content categories of interest. It is content analysis protocols that explain how the variables in the study are to be measured. The coding scheme is the heart of content analysis. The first step in developing a coding scheme is to define content categories. The second step is to define the basic unit of text to be classified. Individual words, phases, sentences, paragraphs, or whole texts may be used as the unit for analysis. The third step is to develop lists of words and phrases associated with each of the content categories. These words and phrases serve as indicators of the concepts of interest. After developing a coding scheme, researchers need to decide what population of content units will be examined, and how a subset of the content will be sampled. Then researchers need to apply dictionaries to the sample through statistical procedures to generate results. Interpreting and reporting the results is the final phase. The results can be evaluated in terms of how well it answers the research questions and fulfills the study's purpose.

Popping (2000) identified three approaches to quantitative text analysis. The first approach is thematic text analysis. In this approach texts are quantified as counts of words and phrases that were classified according to a set of content categories. Thematic text analysis allows the investigator to determine what, and how frequently, concepts occur in texts. The second approach is semantic text analysis. This approach involves not only the identification of concepts but also the relationships among these concepts. In semantic text analysis an encoding process is needed to acquire a semantic grammar that specifies the relations among themes and then the texts' theme are encoded according to the relations specified in the semantic grammar (Roberts, 1997). After the encoding process, the semantically encoded data can be used to make inferences from the texts (Popping, 2000). The third approach is network text analysis. It is derived from the semantic links among concepts. As Popping (2000) explains, after one has encoded semantic links among concepts, one can proceed to construct networks of semantically linked concepts. "When concepts are depicted as networks, one is afforded more information than the frequency at which specific concepts are linked in each block of text; one is also able to characterize concepts and linkages according to their position within the network" (Popping, 2000: 30). Thematic text analysis is based on analyzing words, phrases, sentences or paragraphs that are literally contained in a body of texts. The semantic network approach, however, needs to acknowledge the relational properties of the connections between concepts. Therefore, a semantic network can depict the relationships between text units, even when they occur in different parts of a text (Krippendorff, 2004).

Range of social science problems where automatic content analysis can be applied

Automatic content analysis has already been applied to domains such as communication, political science, and organizational research. In political science, pronouncements and speeches of

public officials, public opinions and legal documents can be analyzed by automatic content analysis. Coffey (2005) measured state governors' ideologies by analyzing the state of the state speeches given by governors in 2000 and 2001. He first coded sentences as liberal or conservative if a governor expressed a clear ideological position on a policy. Each liberal or conservative sentence was assigned to one of ten policy categories. Then he used automatic content analysis software to code sentences in the speeches based on the appearance of words from a predefined dictionary. He manually reviewed the output of this process to exclude sentences coded by computer software that had no ideological meaning and to include sentences omitted by computer software that had a clear ideological meaning. This research uses a thematic approach to measure governors' preferences, values, and ideology. The ideological content of these speeches clearly distinguished governors on the basis of their party affiliation. Coffey concluded that the computer-aided content analysis of speeches and other public pronouncements was a useful means for assessing the views of governors and other public officials.

In organizational research, Meijer and Lleinnijenhuis (2006) analyzed corporate reputation by measuring the amount of news about specific issues. The content analysis they used for analyzing business news articles was based on network analysis for evaluative texts, one of the methods used for relational content analysis (Popping, 2000). This paper not only coded the appearance of an issue, but also its relationship with actors and other issues. By matching media data with survey data on the salient beliefs and reputations held by news consumers, researchers found a positive relationship between issue salience and corporate reputation.
Although automatic content analysis has already been applied to several disciplines, work that applies computational approaches to enhance social science research is still limited. Take organizational research for example: all sources of textual information such as trade magazines, scholarly journals, annual reports, internal company documents and other publicly available documents could serve as data for content analysis using computational methods. Despite the advantages of automatic content analysis, Duriau et al. (2007) found that only about 25% of the articles they reviewed in the field of organizational studies reported using computers for content-analytic task. Therefore, more work needs to be done to maximize the potential usefulness of applying insights and techniques from computational linguistics to social science research.

Content analysis has been found to be an appropriate methodology for the study of online support groups (Burnett & Buerkle, 2004). Klemm, Hurst, Dearholt, and Trone (1999) use content analysis to identify four categories of online cancer support groups, including information giving and

seeking, statements of encouragement and support, statements of personal opinion, and statements of personal experience. Cassell and Tversky (2005) examined how linguistic interaction patterns changed over time among a geographically and ethnically diverse group of young people in an online virtual community. For research focusing on electronic full-text databases in virtual communities such as newsgroups, online discussion forums, and Web materials, automating the process of content analysis can be an effective way to conduct data analysis by processing larger volumes of textual data with higher speed and lower cost.

Applying computational methods to benefit social science research

It is possible to use computational technologies to pinpoint useful information, summarize vast amounts of information, discover concepts and themes, and analyze the language of documents. In research using electronic full-text databases such as newspapers, journals, court decisions, and Web materials, social scientists who seek to balance breadth and depth may benefit from computational linguistics by analyzing large corpora of text automatically. In addition to the advantage of achieving more detailed and extensive context in a research, computational linguistics also allows researchers to overcome difficulties related to coding reliability in content analysis (Evans et al., 2007; Hopkins & King, 2007). With most current publications originating in the electronic medium, it is technically relatively easy to obtain large amounts of online text in various domains (Hausser, 1999).

Use of automated content analysis can result in significant benefits to social scientists interested in extracting systematic meaning from text in two ways. First, social scientists can transform their problems from manual analysis of a small amount of data to automated analysis of a large amount of data. Second, through the process of machine learning, analysts can at least partially mitigate some problems related to noise and ambiguity. Finally, the application of computational linguistics to the social sciences may work best when it is executed in interdisciplinary research teams, including social scientists, computer scientists, and bridging researchers whose expertise spans these fields and facilitates collaboration across disciplinary boundaries.

References

Alvesson, M., & Karreman, D. (2000). Taking the linguistic turn in organizational research. *The Journal of Applied Behavioral Science*, *36*(2), 136-158.

Babbie, E. (2004). *The Practice of Social Research* (10th ed.). Belmont, CA: Thompson Wadsworth.

Berelson, B. (1952). *Content Analysis in Communication Research*. New York: Free Press.

Burnett, G., & Buerkle, H. (2004). Information exchange in virtual communities: A comparative study. *Journal of Computer-Mediated Communication*, *9*(2), 00–00.

Carley, K. (1997). Extracting team mental models through textual analysis. *Journal of Organizational Behavior*, *18*, 533–558.

Cassell, J., & Tversky, D. (2005). The language of online intercultural community formation. *Journal of Computer-Mediated Communication*, *10*(2), 00–00.

Coffey, D. (2005). Measuring gubernatorial ideology: A content analysis of state of the State speeches. *State Politics & Policy Quarterly, 5*(1), 88-103.

Denzin, N. J., & Lincoln, Y. S. (Eds.). (2000). *Handbook of Qualitative Research* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Dille, B., & Young, M. D. (2000). The Conceptual complexity of Presidents Carter and Clinton: An automated content analysis of temporal stability and source bias. *Political Psychology*, *21*(3), 587-596.

Duriau, V. J., Reger, R. K., & Pfarrer, M. D. (2007). A content analysis of the content analysis literature in organization studies: Research themes, data Sources, and methodological refinements. *Organizational Research Methods*. *10*(1), 5-34.

Evans, M., McIntosh, W., Lin, J., & Cates, C. L. (2007). Recounting the courts? Applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies*, *4*(4), 1007-1039.

Gerner, D. J., Schrodt, P. A., Francisco, R. A., & Weddle, J. L. (1994). Machine coding of event data using regional and international sources. *International Studies Quarterly*, *38*(1), 91-119.

Hall, M. A., & Wright, R. F. (2006). Systematic content analysis of judicial opinions. Wake Forest University Legal Studies Paper No. 913336. Available at SSRN: http://ssrn.com/abstract=913336

Hausser, R. (1999). *Foundations of Computational Linguistics: Man-Machine Communication in*

Natural Language. New York: Springer.

Hopkins, D., & King, G. (2007). Extracting systematic social science meaning from text, Working Paper, Retrieved February 20, 2008, from http://gking.harvard.edu/files/words.pdf

Kabanoff, B. (1997). Computers can read as well as count: Computer-aided text analysis in organizational research. *Journal of Organizational Behavior*, *18*(5), 507-511.

Klemm, P., Hurst, M., Dearholt, S. L., & Trone, S. R. (1999). Gender differences on Internet cancer support groups. *Computer Nursing*, *17*(2), 65–72.

Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology*. Beverly Hills, CA: Sage.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology* (2nd ed.). Thousand Oaks, CA: Sage.

McKee, A. (2003). *Textual Analysis: A Beginner's Guide*. London: Sage.

Meijer, M., & Kleinnijenhuis, J. (2006). Issue news and corporate reputation: Applying the theories of agenda setting and issue ownership in the field of business communication. *Journal of Communication*, *56*(3), 543-559.

Morris, R. (1994). Computerized content analysis in management research: A demonstration of advantages and limitations. *Journal of Management*, *20*(4), 903-931.

Nacos, B. L., Shapiro, R. Y., Young, J. T., Fan, D. P., Kjellstrand, T, & McCaa, C. (1991). Content analysis of news reports: Comparing human coding and a computer assisted method. *Communication*, *12*(1), 111-128.

Neuendorf, K. A. (2002). *The Content Analysis Guidebook*. Thousand Oaks, CA: sage.

Popping, R. (2000). *Computer-assisted Text Analysis*. Thousand Oaks, CA: Sage.

Riffe, D., Lacy, S., & Fico, F. (1998). *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Roberts, C. W. (1997). Semantic text analysis: On the structure of linguistic ambiguity in ordinary discourse. In C. W. Roberts (Ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*, 55-77. Mahwah, NJ: Lawrence Erlbaum Associates.

Shapiro, G., & Markoff, J. (1997). A matter of definition. In C. W. Roberts (Ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*, 9-31. Mahwah, NJ: Lawrence Erlbaum Associates.

Stone, P. J. (1997). Thematic text analysis: New agendas for analyzing text content. In C. W. Roberts (Ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*, 35-54. Mahwah, NJ: Lawrence Erlbaum Associates.

Wang, P., & Ramiller, N. C. (2004). Community learning in information technology fashion. *Proceedings of the Twenty-Fifth International Conference on Information Systems (ICIS)*, Washington DC, 11-24.

Weare, C., & Lin, W. Y. (2000). Content analysis of the World Wide Web: Opportunities and challenges. *Social Science Computer Review*, *18*(3), 272-292.

Wing, J. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33-35.

Woodrum, E. (1984). Mainstreaming content analysis in social science: Methodological advantages, obstacles, and solutions. *Social Science Research*, *13*, 1-19.

Zhou, Y., & Moy, P. (2007). Parsing framing processes: The interplay between online public opinion and media coverage. *Journal of Communication*, *57*(1), 79-98.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant IIS-0729459. The authors would also like to thank Chia-jung Tsui, Lidan Wang, and Yejun Wu for their helpful comments.