CPS 430/590.06 **Prof. John Reif**
Design and Analysis of Algorithms **Fall 2013**
Project Abstract **Matt Dickenson**

Project Title: Machine Learning Algorithms with Application to Political Event Data

Project Description This project will describe two machine learning algorithms and discuss their application to automated processing of political event data. The first algorithm, Hidden Markov Modeling (HMM), uses observed data (e.g. newspaper records) to make inferences about an unobserved latent state (e.g. whether two countries are involved in a dispute). The second method, Hierarchical Association Rule Modeling (HARM) selects a subset of decision rules from a set of candidates, again to allow classification into a category such as peace or war. Applying machine learning to the classification of political event data can greatly reduce the cost in human effort, time, and money.

Social science research relying on event data has become increasingly prominent in both academic research and policy discussions. These event data are used to represent a wide variety of events, included terrorist attacks (GTD, WITS) and political instability (ICEWS). The first major research using event data was conducted by McClelland (1961), whose studied international crises. A more recent example is the Militarized Interstate Disputes (MID) dataset, produced by the Correlates of War project. The MID dataset has been widely used in political research over the past three decades and is increasingly used in policy applications. Although statistical methods for studying event data have advanced greatly since then, methods of data collection and coding have not. Virtually all event datasets are processed in the same manner as McClelland's: with humans reading textual records (often from journalistic accounts) and manually entering the classification of the event according to a defined schema (Gerner et al., 1994; Grimmer and Stewart, 2013).

Reliance on human labor to process text into event data is undesirable in several respects. First, the accuracy of humans is often much lower than that of machines. Several studies have estimated that humans classify events about 55 percent as accurately as automated methods for major categories (King and Lowe, 2003) and substantially lower for detailed classifications (Mikhaylov et al., 2012; Ruggeri et al., 2011). Second, the use of human coders introduces delays on the magnitude of years. For example, the most recent version of MID data was released in 2004 and contains data through 2001. An update through 2010 was expected last summer but is delayed indefinitely. Third, the cost of paying humans to process data is on the order of millions of dollars for datasets such as MID.

This project will explore whether HMM and HARM can help automate the processing of political event indicators to update the MID dataset. The final papers will include a write-up of the algorithms used for HMM and HARM, a comparison in cost and complexity to extant methods for processing event data, and preliminary results used to automated the production of MID data.

References

Gerner, D. J., P. A. Schrodt, R. A. Francisco, and J. L. Weddle (1994). The analysis of political events using machine coded data. *International Studies Quarterly 38*(1), 91–119.

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 1–31.

King, G. and W. Lowe (2003). An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization 57*(03), 617–642.

McClelland, C. A. (1961). The acute international crisis. *World Politics 14*(1), 182–204.

Mikhaylov, S., M. Laver, and K. R. Benoit (2012). Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis 20*(1), 78–91.

Ruggeri, A., T.-I. Gizelis, and H. Dorussen (2011). Events data as bismarck's sausages? intercoder reliability, coders' selection, and data quality. *International Interactions 37*(3), 340–361.