Corpus Development and Network Extraction for Comparative Analysis of Networks

Markus Dickinson, Armando Razo
Indiana University, Bloomington, IN USA
md7@indiana.edu, arazo@indiana.edu

1 Introduction and Motivation

We present the start of a project whose end goal is to systematize information about networks in political science research, creating a framework to facilitate comparative analysis in political science. To do this requires, at this stage: 1) curating a corpus of political science articles; 2) extracting entities from unstructured text; and 3) extracting network events and relations.

2 Setting the Goals

Domain of study: Networks and International Development Although *institutions* play an important role in opening up opportunities to overcome poverty and inequality (e.g., Acemoglu et al., 2001), scholars have long recognized that developing countries rely heavily on relation-based mechanisms embedded in a variety of political and economic *networks* (e.g., Hydén, 2006). For example, public policy in developing societies is affected by various types of citizen-politician linkages that involve either vote-buying or some indirect exchange of political and economic benefits (Kitschelt and Wilkinson, 2007). Although networks are evidently relevant, there is much confusion about their actual impact. On the one hand, networks underlie most arguments about the nature and negative impact of corruption on development (Campos, 2002). On the other hand, many scholars and policymakers see networks as the solution to various social problems (e.g., Bank, 2003). To help resolve this confusion, we need to address the lack of systematic characterizations of the diversity of relevant networks in the development literature.

Overarching goals Research on networks and international development relies on diverse disciplinary and subdisciplinary languages along with incomplete descriptions, both of which lead to haphazard theoretical and empirical analysis. Our proposed solution to this problem starts with developing a corpus which will serve to delineate the scope of research on networks and development, facilitating applied research and policy formulation. With our corpus as a focal point, we then wish to build an ontology of networks, offering: (1) a controlled vocabulary; (2) a universal repository of related theoretical and empirical network studies; and (3) workflow templates to assist scholars in formulating research designs for empirical studies. While the characterization of the ontology can be done by hand, at least in part, to generate information for an ontology (e.g., point (2) above) requires extracting information about network relations from relatively unstructured data, namely academic articles.

Relation to Previous Relation Extraction The goals relate in several ways to previous work. O'Connor et al. (2013), for example, extract events between political actors from news data,

similar in purpose to the TABARI system,[1] which seeks to predict conflicts and other political changes. We focus on: a) academic texts, which: b) discuss network relations in a theoretical context, and c) are focused on development, not conflict. Likewise, there is work on extracting information from scientific or academic articles (e.g., Peng and McCallum, 2004), but this often focuses on meta-data extraction, not on specific relations within the domain. The specificity of what we are trying to extract—across different styles of articles (e.g., mathematical formulations of networks vs. qualitative analyses)—makes the re-use of existing tools our first challenge.

3 Exploring the Data

We start by examining a handful of political science articles, a fraction of our corpus (section 4).

Key terms Many entities, relations, and events exist in simple keyword or key phrase form. The various terms in (1a) (Collins, 2002), for instance, indicate that this network concerns FORMAL entities, and that these are ORGANIZATIONS (cf. *institutions*). Such terms occur in key patterns: in (1b) (Collins, 2002) the simple key phrase *a(n) X is a(n) Y*, where *X* is a keyword, identifies the properties (in *Y*) which define this type of network. We will exploit this by using a set of key terms, together with dependency parses, to obtain larger patterns.

(1) a. **clans, pacts** ... **clans** and **tribal divisions**
b. A **clan**$_X$ is an **informal social institution in which actual or notional kinship based on blood or marriage forms the central bond among members**$_Y$.

Diversity of data The data covers a range of topics, written in different styles, from various academic fields. Fowler and Jeon (2008), for example, discusses networks covering appellate courts, while Collins (2002) deals with ethnonational divisions; networks can be within any domain, and even within a particular domain the potential set of relevant words is enormous. Low-frequency terms like *indebtedness* are used, and one may want to know that such a word indicates something consistent with monetary transactions. Our approach will be to use a small set of initial seed terms and patterns to bootstrap domain- or article-specific terms.

Ambiguity A problem for consistently identifying networks and properties is that of context-dependent definitions. In (2) (Atieno, 2001), for instance, the phrase in bold may indicate a network, but only if the supply of credit is contingent upon personal relations. We plan on using document structure and collocational information within a section or paragraph to gain confidence in terms being network-related.

(2) ...one question that arises is the extent to which **credit can be offered to the rural poor** to facilitate their taking advantage of the developing entrepreneurial activities.

Shifting reference Each document may reference several networks, shifting between them. The example in (3) (Atieno, 2001) illustrates how an author can contrast different networks, shifting from an ORGANIZATION-TO-ORGANIZATION to a PEOPLE-TO-PEOPLE network.

(3) There are a number of credit institutions that support small and microenterprise activities ...There are also a number of financial transactions ... outside these institutions, like those between **relatives and friends, traders, and welfare groups**.

[1] `http://eventdata.parusanalytics.com`

Features of networks Similar to shifting reference, specific subparts of a network can be discussed. In (4) (Fowler and Jeon, 2008), for instance, the network of citations referred to is composed of *judicial* citations. In both cases, dependency structures and document proximity (cf. word upweighting (Manning et al., 2008)) can help with proper extraction.

(4) **Each judicial citation** contained in an opinion is essentially a latent judgment about the case cited. ... We use the complete **network of citations** in all 30,288 majority opinions contained in the U.S. Reports from 1754 to 2002 ...

4 Proposed Plan

As much as possible, we hope to re-use existing tools, adapting them to our context as necessary, mainly with respect to which key terms are of interest.

1. **Document Preparation and Annotation**. Given the new domain, we need data upon which to evaluate systems. We have started adding high-level annotations to an existing database of 100+ previously classified works to create an evaluation set. For example, a document may be marked as containing a `bonding` network, between `governmental groups` and `non-governmental groups`; such tuples, in addition to network features (e.g., `self-organizing`), form the core of the annotation. We are additionally collecting and codifying a random sample of 100 articles and books from the extant literature. It should be noted that the annotation is a work in progress, evolving as we map sentential properties to document properties.

2. **Linguistic Annotation**. We are using a dependency parser (de Marneffe et al., 2006) to generate relationships between words. Together with a small set of seed terms (e.g., *actor*, *relationship*), we isolate: a. modifiers of key terms to generate a list of possible features/relations (e.g., noting that a `relationship` is `bonding` and `reciprocal`); b. tuples involving the key term, generally corresponding to events (e.g., `maximize(relationship,credibility)`. We are currently investigating how to isolate the most important connections—by utilizing document structure, such as weighting abstract information or using text zoning to identify the purpose of each sentence in the article (Poibeau et al., 2014)—and how to map individual tuples to document-wide properties.

Additionally, we are exploring information gained from SuperSense tags in the extracted tuples (Ciaramita and Altun, 2006). Note that, although we originally intended to detect named entities, the entities we seek to detect are not *named*. In these articles, authors generally abstract away from the specific names and discuss the ramifications of types of networks.

3. **Relation and Event Extraction**. In parallel with using a relatively simple dependency-based approach (in #2), we are also taking a lightly-supervised approach to event and relation extraction to determine network relations and properties, building from an existing system, OLLIE (Mausam et al., 2012). We plan to bootstrap terms from a small seed set of patterns and from our corpus. As we expect to overgenerate events and relations—i.e., we only need to extract a small number of events from a document—we also plan to filter out unwanted events by placing linguistic restrictions on extracted information (cf. Cybulska and Vossen, 2011).

Future: **Network Classification**. After these steps, we will identify which network-focused entities are co-referential and attempt to classify the networks into particular types (e.g., `individual-to-individual network`).

References

Acemoglu, D., Johnson, S., and Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5):1369–1401.

Atieno, R. (2001). Formal and informal institutions' lending policies and access to credit by small-scale enterprises in kenya: An empirical assessment. Technical report, The African Economic Research Consortium.

Bank, W. (2003). *World Development Report 2004 Making Services Work for Poor People*. World Bank, Washington, DC.

Campos, J. E. (2002). *Corruption: The Boom and Bust of East Asia*. Ateneo de Manila University Press, Manila.

Ciaramita, M. and Altun, Y. (2006). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 594–602, Sydney.

Collins, K. (2002). Clans, pacts, and politics in central Asia. *Journal of Democracy*, 13(3):137.

Cybulska, A. K. and Vossen, P. (2011). Historical event extraction from text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 39–43, Portland, OR.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*.

Fowler, J. H. and Jeon, S. (2008). The authority of supreme court precedent. *Social Networks*, 30:16–30.

Hydén, G. (2006). *African politics in comparative perspective*. Cambridge University Press, Cambridge, UK; New York.

Kitschelt, H. and Wilkinson, S. (2007). Citizen-politician linkages: an introduction. In Kitschelt, H. and Wilkinson, S., editors, *Patrons, clients, and policies: patterns of democratic accountability and political competition*, pages 1–49. Cambridge University Press, Cambridge, UK; New York.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mausam, Schmitz, M., Soderland, S., Bart, R., and Etzioni, O. (2012). Open language learning for information extraction. In *Proceedings of the EMNLP-CoNLL-12*, pages 523–534, Jeju Island.

O'Connor, B., Stewart, B. M., and Smith, N. A. (2013). Learning to extract international relations from political context. In *Proceedings of ACL-13*, pages 1094–1104, Sofia, Bulgaria.

Peng, F. and McCallum, A. (2004). Accurate information extraction from research papers using conditional random fields. In *Proceedings of HLT-NAACL 2004*, pages 329–336, Boston.

Poibeau, T., Omodei, E., Cointet, J.-P., and Guo, Y. (2014). Social and semantic diversity: Socio-semantic representation of a scientific corpus. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 71–79, Gothenburg, Sweden.