Network Construction based on Structured and Unstructured Text Data in ConText

Jana Diesner (jdiesner@illinois.edu)

Submission to NETCOM 2015, ICA

Abstract:

We introduce methodological innovations for constructing network data based on various text data and meta data based features and compare the comparability and agreement of these techniques to ground truth data. We also provide examples for using these techniques for applied research in the context of impact assessment of issue-focused information products.

Paper:

Natural language text data can serve as a stand-alone or supplementary source for enhancing or constructing network data. In computing, the process of locating and classifying nodes and links based on unstructured text data as accurately and efficiently as possible is referred to as relation extraction (Agichtein & Gravano, 2000; Bunescu & Mooney, 2006; Culotta & Sorensen, 2004; McCallum, 2005; Mihalcea & Radev, 2011). In the computational social sciences and digital humanities, such methods have been used to construct, for example, social networks from archived narrative texts (Abello, Broadwell, & Tangherlini, 2012), news data (Johnson & Krempel, 2004; Kleinnijenhuis, de Ridder, & Rietberg, 1997; Roberts, 1997) and communication logs (Corman, Kuhn, McPhee, & Dooley, 2002; Dabbish, Towne, Diesner, & Herbsleb, 2011; Diesner, Aleyasen, Mishra, Schecter, & Contractor, 2014), actor-event networks (Gerner, Schrodt, Francisco, & Weddle, 1994; Leetaru & Schrodt, 2013), geopolitically contextualized social networks (Diesner, Carley, & Tambayong, 2012) and meta-data enriched semantic networks (Van Atteveldt, 2008).

Going from words to networks involves a plethora of methodological choices: First, text based networks can be constructed from both a) meta data, such as key words and index terms, and/ or b) the content of text bodies (Diesner, 2013). Approach a) is basically a database operation and therefore fairly straightforward and highly efficient. Approach b) requires natural language processing techniques for identifying nodes and edges. Prior research has begun to shed light on how the outcomes of these techniques compare, but our knowledge about this issue is insufficient. Second, people have been using text mining techniques to extract the structure of social systems from text data sources (Diesner et al., 2014; Roth & Cointet, 2010). Again, our knowledge on how the resulting graphs compare to those built based on non-text based sources and by using more classic methods, e.g. surveys and observations, is incomplete. Finally, relation extraction can be based on a variety of text-based features, e.g. semantic, syntactic, proximity-based and probabilistic information (Diesner & Carley, 2011; Mihalcea & Radev, 2011). For this

case, we also have insufficient knowledge on the complementarity and convergence of methods in terms of a) comparing to each other and b) to ground truth data.

In order to answer these open and practically relevant questions and to also enable others to work on these problems, we have been building ConText (http://context.lis.illinois.edu/). ConText is a publicly available software suite for constructing network data from text data and pertinent meta data using a variety of off-the-shelf as well as novel and unique techniques. The latter group includes entity extractors for node classes and sentiment types relevant in the social sciences but not covered by classic computing based coding schemas and tools and the automated construction of social network data based on the content of communication log data.

In this talk, we introduce the methodological innovations available through ConText and provide empirically based answers to some of the abovementioned questions. We also provide examples for using these techniques for applied research in the context of impact assessment of issue-focused information products.

Acknowledgement: This work is supported by the FORD Foundation, grant 0145-0558. I am grateful to the following graduate students from UIUC for their help with this work: Amir Hossein Aleyasen, Julian Chin, Ming Jiang, Shubhanshu Mishra, Kiumars Soltani and Liang Tao.

References:

Abello, J., Broadwell, P., & Tangherlini, T. R. (2012). Computational folkloristics. *Communications of the ACM, 55*(7), 60-70.

Agichtein, E., & Gravano, L. (2000). *Snowball: extracting relations from large plain-text collections.* Paper presented at the Fifth ACM International Conference on Digital Libraries, San Antion, TX, USA.

Bunescu, R., & Mooney, R. (2006). Subsequence kernels for relation extraction. *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 18*, 171.

Corman, S. R., Kuhn, T., McPhee, R. D., & Dooley, K. J. (2002). Studying Complex Discursive Systems: Centering Resonance Analysis of Communication. *Human Communication Research, 28*(2), 157-206.

Culotta, A., & Sorensen, J. (2004). *Dependency tree kernels for relation extraction*.

Dabbish, L., Towne, B., Diesner, J., & Herbsleb, J. (2011). Construction of association networks from communication in teams working on complex projects. *Statistical Analysis and Data Mining, 4*(5), 547-563.

Diesner, J. (2013). From Texts to Networks: Detecting and Managing the Impact of Methodological Choices for Extracting Network Data from Text Data. *Künstliche Intelligenz/ Artificial Intelligence, 27*(1), 75-78. doi: 10.1007/s13218-012-0225-0

Diesner, J., Aleyasen, A., Mishra, S., Schecter, A., & Contractor, N. (2014). *Comparison of Communication Networks built from explicit and implicit data*. Paper presented at the Computational Approaches to Social Modeling (CHASM 2014), ACM WebScience Conferenc, Blomington, IN.

Diesner, J., & Carley, K. M. (2011). Words and Networks. In G. Barnett & J. G. Golson (Eds.), *Encyclopedia of Social Networking* (pp. 958-961): Sage.

Diesner, J., Carley, K. M., & Tambayong, L. (2012). Extracting socio-cultural networks of the Sudan from open-source, large-scale text data. *Computational & Mathematical Organization Theory, 18*(3), 328-339.

Gerner, D., Schrodt, P., Francisco, R., & Weddle, J. (1994). Machine Coding of Event Data Using Regional and International Sources. *International Studies Quarterly, 38*(1), 91-119.

Johnson, J. C., & Krempel, L. (2004). Network Visualization: The" Bush Team" in Reuters News Ticker 9/11-11/15/01. *Journal of social structure, 5*.

Kleinnijenhuis, J., de Ridder, J., & Rietberg, E. (1997). Reasoning in Economic Discourse: An Application of the Network Approach to the Dutch Press. In C. W. Roberts (Ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts* (pp. 191–208). Mahwah, NJ: Lawrence Erlbaum Associates.

Leetaru, K., & Schrodt, P. A. (2013). *GDELT: Global data on events, location, and tone, 1979–2012.* Paper presented at the ISA Annual Convention.

McCallum, A. (2005). Information extraction: distilling structured data from unstructured text. *ACM Queue, 3*(9), 48-57.

Mihalcea, R. F., & Radev, D. R. (2011). *Graph-based Natural Language Processing and Information Retrieval*: Cambridge University Press.

Roberts, C. W. (1997). A Generic Semantic Grammar for Quantitative Text Analysis: Applications to East and West Berlin Radio News Content from 1979. *Sociological methodology, 27*, 89-129.

Roth, C., & Cointet, J. (2010). Social and semantic coevolution in knowledge networks. *Social Networks, 32*(1), 16-29.

Van Atteveldt, W. (2008). *Semantic network analysis: Techniques for extracting, representing, and querying media content*. Charleston, SC: BookSurge Publishers.