Word Networks

Natural language text data can serve as a single or complementary source for collecting, creating and enriching network data. Ultimately, the integration of text analysis and network analysis contributes to a comprehensive understanding of the form and function of networks, and facilitates the investigation of the interplay and co-evolution of language and other types of social interaction. This entry describes the main use cases and respective approaches for going from texts to networks, and points out current limitations and challenges.

Relation Extraction from Texts

Sometimes, text data are the only source of information about a network. Most of the respective cases are instances of one or more of the following types of networks:

1. Networks which are inaccessible or unobservable for the data collector. Prominent examples are covert networks, such as groups engaged in organized crime and secret societies, and cognitive models, which are representations of the knowledge and information that individuals hold in their minds.
2. Networks that have ceased to exist at the time of data collection. Examples are former cultures and bankrupt companies.
3. Large-scale networks for which gathering data within the appropriate network boundaries by using traditional network data collection methods, such as surveys, is prohibitively expensive, and which do not allow for sampling due to the skewed distribution of the number of links per node. Examples are sizeable communities of practice, geopolitical entities, and the diffusion of behavior through segments of society.
4. Virtual networks that do not necessarily feature an underlying real-world social network and that are confined to the traces of behavioral data generated by the members of the network. Examples are open collaboration initiatives.

Text data that potentially contain relevant information about such networks include documents authored by members of these networks, such as diaries, narratives, interpersonal communication, mission statements, and annual reports, and material originating from outside the network, such as news wire data, reports from subject matter experts, and transcripts of court hearings. In these cases, Relation Extraction methods can be employed in order to identify the relevant pieces of information and their connections as they are explicitly or implicitly represented in the text data, and converting these information into the nodes and edges of a network. Across many Relation Extraction methods, triples comprising the subject, action, and object of an event or phenomena form the smallest structured unit in network of words. Depending on the method and user’s needs, these data can be enhanced, for instance, with spatial and temporal information, attributes, and weights.

Here are two examples for the usage of Relation Extraction: First, information about the who, what, when, where, why, and how of an event can be converted into nodes of the type agent, task or event, date, location, motivation or sentiment, and means or resources, respectively. Connecting the dots results in a structural representation of a single event, i.e. a network of words comprising nodes and edges. Applying this process to data on many events, such as longterm and large-scale news feeds, allows analysts to generate network data that can be used to investigate the structure, properties, evolution and behavior of complex, dynamic, and sizable socio-technical networks. Another example are transcripts of narrations, interviews, and conversations, to which anthropologists, cognitive scientists, and social scientists, among others, apply Relation Extraction methods in order to identify the themes, intensions, and emotions addressed by the authors. Linking up these concepts according to how the authors of the documents had connected them can result in cognitive models. Such models are used for instance to study how students and members of teams coincide and differ in their understanding and perception of certain phenomena.

Zooming out from these examples to the general application of extracting network data from texts suggests that the resulting output is often used as input to traditional network analysis. Beyond that, Relation Extraction methods are used in empirical user studies, as input to visualizations, simulations, statistical methods such as multi-dimensional scaling and principal component analysis, subsequent computational processes, for populating relational databases, and as an underlying mechanism for search engines and question answering systems.

The various Relation Extraction methods differ in their terminology, underlying theories and assumptions, methods for finding and classifying nodes and edges, degree of automation, evaluation methods, and appropriate applications. Many of these differences are due to the emergence of Relation Extraction from multiple disciplines with little cross-disciplinary syntheses. The next section focuses on one of these dimensions, namely the methods for identifying nodes and edges:

Performing qualitative text analysis requires humans to assign codes to words or text passages, to establish typed links between the codes, and to document their coding choices; typically in the form of memos. Codes are relevant concepts derived from theory or by examining the data. Software programs assist humans in associating portions of text data with codes, retrieving all data that share a code, aggregating codes into variables, arranging codes and variables into structural models, and visualizing and analyzing these models. Linking up codes and variables according to how the researcher perceives their interdependencies is part of Grounded Theory methodology, which ultimately serves the development of structural models of social phenomena that help to explore the data, generate hypotheses subject to further investigation, and gain an in-depth understanding of corpora of moderate size.

The efficiency of node identification can be increased by applying lists of relevant terms as a positive filter to texts such that only matches between list entries and the text data are kept. The remaining terms can be subsequently tested for their appropriateness for being converted into nodes. In the context of this article, terms are single words or meaningful multi-word units (n-grams), and concepts are more abstract representations of terms. For example, the term and bigram “Georg Simmel” could be associated with the concept “sociologist”. The relationship between terms and concepts can be specified in ontologies that are predefined or inferred from the data. The effectiveness of node identification can be increased by deploying highly accurate and fully-automated Natural Language Processing techniques. Examples include the reduction of words to their morpheme (stemming), n-gram detection, parts of speech tagging, and the identification of the most descriptive terms and topics in a given document or corpus. Thesauri, i.e. collections of pairs of terms and concepts, can be applied in the same fashion as the aforementioned lists, but additionally allow for translating terms into concepts. Thereby, thesauri support the expansion of acronyms, normalization of spelling variations and typos, and context-dependent term disambiguation. However, lists and thesauri can be incomplete, outdated and erroneous. A more flexible approach to node and also edge identification is offered by rule-based techniques, such as regular expressions, which search text data for patterns without constraining the search to particular terms. However, rules – as well as lists and dictionaries - are deterministic; meaning that unspecified and deviating terms and patterns cannot be found. This limitation can be overcome by applying probabilistic techniques. An example for probabilistic techniques are machine learning methods, which exploit the experience that the learner is initially provided with in order to improve their performance with respect to a performance metric, e.g. the accuracy of locating and classifying nodes and edges in new and unseen data.

Common methods for establishing links between terms are based on one or more of the following features:

- Collocation, i.e. the proximity of words within user-defined text units.
- Logical relations between terms, which are used in methods originating from the Artificial Intelligence.
- Grammars, including syntactic dependencies and semantic grammars, which comprise lists and relations tailored for specific content domains.
- Semantic relations between words, which are used for instance in Discourse Representation Theory, and which can be specified in ontologies.

Relation Extraction methods are typically evaluated by comparing the resulting networks against a ground truth if available, and by testing the performance of a method on new data and domains. If humans are involved in data coding, the consistency in coding per person (intra-coder reliability) and across people (inter-coder reliability) is also assessed.

Combining text data and network data

Text data can help to contextualize and enrich network data. An example for the joint availability of text data and network data are surveys that gather not only information about relationships between entities, but also answers to open-ended questions. These answers can help to interpret the nature of relationships in a network. Another example are email data, where the explicit specification of communication partners in the email headers can be converted into a traditional social network. In such networks, link weights typically indicate the cumulative number of emails exchanged between any pair of nodes. While this approach takes the occurrence and frequency of communication into account, it does not consider the substance of communication. This limitation can be overcome by also analyzing the content that is explicitly and implicitly contained in the email bodies.

Generalizing from these two examples to the wider domain of communication data it can be stated that the joint consideration of text data and network data allows users to not only comprehend networks by examining the occurrence, strength, likelihood and type of relations, but by also analyzing the substance of social interaction as expressed through natural language. The increasing popularity and related commercial relevance of virtual collaboration environments and the participatory web have led to a new momentum in the demand for methods, metrics, and tools for combining network data with the content, opinions, and sentiments that the people in these networks express. Examples for socio-technical systems in which social entities interact and contribute expressions in the form of natural language include offline and online collaboration teams, including software development initiatives, virtual worlds, and various instances of social media, such as wikis, blogs, chat rooms, question answering sites, and customer review services.

On the most general level, there are two approaches for combining text data and network data: First, existing network data can be augmented with additional structured data derived from texts. This approach is commonly put into practice by identifying the relevant concepts from a document or corpus, converting them into nodes, and linking these nodes to the social entities that they are associated with. Applying this approach to communication data such as emails or group discussions results in a two-mode network that features connections among and between two node classes: social entities, typically individuals or groups, and information. Integrating text data and network data in this fashion allows analysts to move beyond reducing communication to the fact and frequency of message transmission, and has been used to help answering questions like: who is talking to whom about what? Which topics connect or separate which individuals or groups? What social groups emerge when performing topic-based clustering, and how do these groups compare to those identified by clustering the social network only? Other applications of combining network data and text data include science mapping projects like analyses of patents, funding allocation, and publications that portrait the landscape of associations between researchers and their ideas or focus areas. Terms that best represent the content of text data can be identified, for instance, by performing topic modeling; an unsupervised machine learning technique that reduces the dimensionality of text data to a few salient topics. When these topics are unstructured collections of terms, the agent by topic matrix can be transposed in order to obtain a topic by topic network, which sometimes is called a semantic network. Alternatively, relation extraction techniques as described earlier in this entry can be used to extract not only concept nodes, but actual networks from texts. These networks can also be merged with other network data, which has been used for example to concurrently represent social structure and mental models.

The second approach for combining text data and network data is to enhance, refine, and post-process existing network data based on text analysis, but without adding nodes or edges that represent information from the texts. Two application domains for this approach are common:

First, text analysis can be used to classify nodes and links. Here, classifying means to assign roles, types or labels to nodes and links. These additions are often represented as attributes. Text-based classification has been used to detect malicious behavior from email data by separating nodes representing regular individuals from spammers, and to categorizing links as regular versus suspicious communication. Inferring the roles of nodes and the nature of relationships from index terms such as keywords and email subject lines has shown to be less accurate and informative than exploiting larger portions of text data such as email bodies.

Second, text data can help to perform disambiguation and entity resolution on network data. For example, in social networks, different individuals who share the same first and last name might have been lumped together into a single node. Analyzing texts written by or about these people can help to disambiguate and split up nodes accordingly such that each node represents one unique individual. The reverse of this scenario are situations in which unique entities occur as multiple nodes in a single network. This is typically due to typos, variations in spelling, usage of acronyms and pseudonyms, and references to social entities by their name, role and title. Again, analyzing texts pertaining to the network can assist in identifying clusters of nodes, where each cluster contains the different node names that are used to refer to a single entity. The nodes per cluster can then be merged into a single node.

Summary

Relation Extraction methods can be used to collect network data from text data. These methods serve as a remedy when other network data collection methods are not applicable. The resulting network data can be directly used as input to network analysis, or can be used to enhance other network data. The joint utilization of text data and network data facilitates the consideration of the substance of communication as a form of social interaction in networks. The integration of words and networks is achieved by using interdisciplinary methodologies, which may also combine elements from qualitative and quantitative research. Cutting-edge methodologies and tools in this domain are expected to concurrently support the automated, efficient, and in-depth analysis of large datasets and networks. Computational methods, especially machine learning techniques, that incorporate components from other disciplines, such as linguistics, social science, and communication science, are a promising strategy for tackling this challenge.

References and further readings:

Carley, Kathleen M. and Michael Palmquist (1991). "Extracting, Representing, and Analyzing Mental Models," Social Forces, 70(3): 601 – 636.

Corman, Steven R., Timothy Kuhn, Robert D. McPhee and Kevin J. Dooley (2002). “Studying Complex Discursive Systems: Centering Resonance Analysis of Communication”. Human Communication Research, 28(2), 157-206.

JA Danowski (1993) “Network analysis of message content”. In G. Barnett & W. Richards. Progress in communication sciences. Ablex Publishing Corporation. 12, 197-222.

Diesner Jana and Kathleen M. Carley KM (2008) “Conditional Random Fields for Entity Extraction and Ontological Text Coding.” Journal of Computational and Mathematical Organization Theory (CMOT), 14(3), 248 – 262.

McCallum, Andrew (2005). “Information Extraction: Distilling Structured Data from Unstructured Text.” ACM Queue 3: 48-57.

Milroy, Lesley (1987). “Language and social networks” (2nd edition). Oxford: Blackwell.

Roth, Camille and Jean-Philippe Cointet (2009). “Social and Semantic Coevolution in Knowledge Networks", Social Networks, 32(1): 16-29.

SEE ALSO: Semantic Networks, Communication Networks, Cognitive Networks, Methods of Data Collection

Jana Diesner, Carnegie Mellon University

Kathleen M. Carley, Carnegie Mellon University